RESONANCE-75

Promoting Science Education

Volume I
# Contents

## Volume I

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Author</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Foreword</td>
<td>ix</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Preface</td>
<td>xi</td>
<td></td>
</tr>
</tbody>
</table>

**Mathematical Sciences**

1. The Football
   - A R Rao
   - 5

2. On Ancient Babylonian Algebra and Geometry
   - Rahul Roy
   - 15

3. (K)not So Abstract!
   - Swatee Naik
   - 27

4. Hilbert’s Nullstellensatz and the beginning of Algebraic Geometry
   - Vishwambhar Pati
   - 35

5. A Pizza Saga
   - Shailesh A Shirali
   - 51

6. On Shapes of Algebraic Loci
   - Subhashis Nag
   - 59

7. From Calculus to Wavelets: A New Mathematical Technique
   - Gerald B Folland
   - 67

8. How to Count – An Exposition of Polya’s Theory of Enumeration
   - Shriya Anand
   - 79

9. Congruent Numbers, Elliptic Curves, and the Passage from the Local
to the Global
   - Chandan Singh Dalawat
   - 91

10. The Unexpected Appearance of Pi in Diverse Problems
    - Rajendra Bhatia
    - 117

11. Investigating the Primes
    - Kaneenika Sinha
    - 125
12. Balanced Number System
   Shobha Bagai
   139

13. Some G Pólya Gems from Complex Analysis
   Shobha Madan
   151

14. Expander Codes
   Priti Shankar
   161

15. A Sketch of Modern Cryptology
   Palash Sarkar
   173

16. Petri Nets
   Y Narahari
   187

17. Bootstrap Methods in Statistics
   Peter Hall and Abhinanda Sarkar
   197

18. Opinion Polls and Statistical Surveys: What They Really Tell Us
   Rajeeva L Karandikar and Ayanendranath Basu
   203

Physical Sciences
   213

19. Origin (?) of the Universe
   Jayant V Narlikar
   217

20. Gravitational Collapse and Structure Formation in an Expanding Universe
   J S Bagla and Pritpal Kaur Sandhu
   223

21. LIGO-India
   Tarun Souradeep
   235

22. Recent Progress on the Black Hole Information Paradox
   Raghu Mahajan
   241

23. The Story of the Photon
   N Mukunda
   251

24. The Interference of Polarised Light
   Rajaram Nityananda
   267

25. The EPR Paradox
   Arvind
   279

26. The Quantum Zeno Effect
   Anu Venugopalan
   287
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>27.</td>
<td>Pollen Grains, Random Walks and Einstein</td>
<td>Sriram Ramaswamy</td>
</tr>
<tr>
<td>28.</td>
<td>Snippets of Physics</td>
<td>T Padmanabhan</td>
</tr>
<tr>
<td>29.</td>
<td>Swinging in Imaginary Time</td>
<td>Cihan Saclioglu</td>
</tr>
<tr>
<td>30.</td>
<td>Mirrors and Merry-Go-Rounds</td>
<td>Joseph Samuel</td>
</tr>
<tr>
<td>31.</td>
<td>Electrons in Condensed Matter</td>
<td>T V Ramakrishnan</td>
</tr>
<tr>
<td>32.</td>
<td>Emerging Trends in Topological Insulators and Topological Superconductors</td>
<td>Arijit Saha and Arun M Jayannavar</td>
</tr>
<tr>
<td>33.</td>
<td>An Experiment that Shaped the Physics of the Century</td>
<td>Rohini Godbole</td>
</tr>
<tr>
<td>34.</td>
<td>A Hymn to Agni – The God of Fire</td>
<td>P K Kaw</td>
</tr>
<tr>
<td>35.</td>
<td>Quadrupole Ion Traps</td>
<td>Pushpa M Rao, Richard D’Souza and S A Ahmad</td>
</tr>
<tr>
<td>36.</td>
<td>The Challenge of Fluid Flow</td>
<td>Roddam Narasimha</td>
</tr>
<tr>
<td>37.</td>
<td>The Indian Monsoon</td>
<td>Sulochana Gadgil and M Rajeevan</td>
</tr>
<tr>
<td>38.</td>
<td>Continent Deformation Due to Earthquake Cycle and Landscape Evolution</td>
<td>Vineet K Gahalaut</td>
</tr>
</tbody>
</table>
Volume II

Chemical Sciences

39. From Matter to Life: Chemistry?!
   Jean-Marie Lehn

40. Chemistry of Materials
   C N R Rao

41. Emerging Solar Technologies: Perovskite Solar Cell
   Amruta Matalikdesai and Sheela K Ramasesha

42. On Benzene and Aromaticity
   M V Bhatt

43. Coordination Chemistry of Life Processes: Bioinorganic Chemistry
   R N Mukherjee

44. Chemistry Triggered the First Civil Disobedience Movement in India
   Gopalpur Nagendrappa

45. Electrons in Molecules
   B M Deb

46. Molecule Matters
   K C Kumara Swamy

47. Breaking Rules – Making Bonds
   A G Samuelson

48. Arrows in Chemistry
   Abirami Lakshminarayanan

49. Experimental Determination of the Avogadro Constant
   S Ramasesha

50. Chemical Research of Sir Prafulla Chandra Rāy
   Sreebrata Goswami and Samaresh Bhattacharya

51. The Ziegler Catalysts
   S Sivaram

52. The Explosive Chemistry of Nitrogen
   Dheeraj Kumar and Anil J Elias

53. Essential Oils and Fragrances from Natural Sources
   Padma S Vankar
54. Single-Molecule Spectroscopy
   Kankan Bhattacharyya

55. Chemistry at the Nanoscale
   Ashwin B R Kumar and Ram Ramaswamy

**Biological Sciences**

56. Cambrian Explosion of Life: The Big Bang in Metazoan Evolution
   P V Sukumaran

57. Body Size Matters in the Lives of Organisms
   T Ramakrishna Rao

58. The Handicap Principle
   Laasya Samhita

59. Evolutionary Stable Strategy
   Jayanti Ray-Mukherjee and Shomen Mukherjee

60. Nature Watch
   Sindhu Radhakrishna

61. Molecular Tools For Biodiversity Conservation
   Shomita Mukherjee and Uma Ramakrishnan

62. The Honeybee Dance-Language Controversy
   Raghavendra Gadagkar

63. C H Waddington: Canalisation, Genetic Assimilation and the
   Epigenetic Landscape
   Vidyanand Nanjundiah

64. The Price Equation and the Mathematics of Selection
   Amitabh Joshi

65. Koinophilia and Human Facial Attractiveness
   Aishwariya Iyengar, Rutvij Kulkarni and T N C Vidya

66. Gender in Plants
   Renee M Borges

67. Defense Signaling in Plants
   G Sivakumar Swamy

68. Biodiversity and Biological Degradation of Soil
   Upasana Mishra and Dolly Wattal Dhar
69. Sydney Brenner: The Tamer of an Elegant Worm
   Kaling Danggen and Varsha Singh

70. The Immune System and Bodily Defence
   Vineeta Bal and Satyajit Rath

71. Stem Cells: A Dormant Volcano Within Our Body?
   Devaveena Dey and Annapoorni Rangarajan

72. Glial Cells: The Other Cells of the Nervous System
   Medha S Rajadhyaksha and Yasmin Khan

73. Switches in the Brain?
   Dilawar Singh

74. Genome Editing Revolution in Life Sciences
   Alok Kumar Singh, Sivapra kash Ramalingam, Desirazu N Rao and
   Srinivasan Chandrasegaran

75. The Power of Small
   Suvasini Ramaswamy and Anirban Mitra
This volume is a collection of 75 articles selected from all past issues of Resonance – *journal of science education*, to celebrate 75 years of India’s independence. It contains articles in all the major areas of science in roughly equal proportions – 18 in mathematical sciences, 20 in physical sciences, 17 in chemical sciences and 20 in biological sciences.

*Resonance*, published monthly by the Indian Academy of Sciences, has been appearing since January 1996, so it is about a quarter century old. It grew out of recommendations made in 1994 to the Academy Council by a specially constituted Panel to examine University Science Education in India, and to suggest steps that could be taken to help improve the quality of education. Publication of *Resonance* is one of several activities initiated by the Academy in the 1990’s, arising out of the Panel’s recommendations.

Over time, *Resonance* has achieved a special and unique status worldwide as a journal devoted to college and university level science education in all the sciences. Each year several tens of thousands of articles are downloaded from its pages by readers in various parts of the world. Some special features are that each issue generally contains articles in each area of science, and it also highlights the life and work of a chosen outstanding figure from the past in mathematics or the natural sciences. In this way so far about 300 scientists have been featured in its pages. Thus, in time *Resonance* has become a rich source of material for practical use by both teachers and students.

In fact, on several occasions and in many ways, *Resonance* has been used to create useful compact collections of articles. There have been some theme-based collections in mathematics such as Analysis, Geometry, Topology, and Probability; and more recently one on the Darwinian Theory of Evolution by Natural Selection. Another series of Master Class volumes are author based and bring together articles by a chosen contributor to *Resonance* over a few years. There have been such volumes in mathematics, chemistry, physics as well as biology.

The year 2005, the centenary of Albert Einstein’s *Annus Mirabilis*, was celebrated internationally as the World Year of Physics. *Resonance* joined this effort by publishing in December of that year a special issue titled, ‘A Celebration of Physics’, a 250 pages collection of 58 chosen articles in physics that had appeared in its pages over the preceding ten years.

The present volume thus continues a valuable tradition at *Resonance*, this time to mark an important event in India’s history. We may confidently expect that like the other volumes, this one too will become a valuable and treasured collection that
With a wide-ranging collection such as this, perhaps a reader primarily interested in one area will by chance be led to something in a different area new to her and feel she has learnt something really new. In each of the four major areas there is an impressive list of topics that are covered. Here is an indication of the range in this selection. In mathematics: Euclidean and algebraic geometry, number theory and primes and pi, opinion polls and statistical surveys. In physics: the expanding universe, gravity and gravitational waves, the history of photons, entanglement in quantum mechanics, statistical physics of pollen grains and random walks, plasma physics and condensed matter physics, studies of earthquakes and monsoons. In chemistry: catalysis, chemistry in life processes, materials, spectroscopy, nano scale chemistry. In biology: the Cambrian explosion of life, signaling in plants, the mathematics behind natural selection.

With a wide-ranging collection such as this, perhaps a reader primarily interested in one area will by chance be led to something in a different area new to her, and feel she has learnt something really new. With this hope we offer this commemoration of 75 years of Indian independence to students and teachers both in India and all over the world and wish you happy reading.

N. Mukunda

Founding Editor, Resonance

Bengaluru

October 2021
Resonance – Journal of Science Education was started in 1996, based on a recommendation in a white paper commissioned by the Indian Academy of Sciences, Bengaluru, on the status of higher education in India. The journal has completed 25 years of contributing to pedagogy in the sciences in India, particularly at the undergraduate level. The articles published in Resonance are written by expert teachers and scholars, and many of the articles have been used as complementary reading for undergraduate courses in India and abroad. In addition to explaining and illuminating outstanding contributions to science as well as basic concepts, the journal devotes a specific section to articles providing aids to classroom teaching. Collections of Resonance articles written on a general theme, or by a specific author, are published as books in the Masterclass series from time to time, and these can be freely downloaded from the website of the Academy (https://www.ias.ac.in/Journals/Resonance--Journal_of_Science_Education/). As a part of the celebrations of 75 years of India's Independence, the Academy decided to collect and republish in book form, a set of 75 articles from various areas of science, mathematics, and engineering, that have appeared in Resonance, but have not been republished as part of a Masterclass volume.

Science knows no boundaries. Humans have looked at the sky and beyond, recorded the movement of planets and stars and galaxies over the years, and inferred the origin of the universe. They looked around, at flora and fauna, the seasons, the nurturing Sun and probed the mysterious workings of Nature at first hand. The disciplines of science, as we know them today, were founded on these observations. Matter, in the most complex structures, is held together by chemical bonds that are formed or broken in less than a billionth of a second. The universe and life, on the other hand, have evolved over time scales of billions of years, and the mysteries of these are still being decoded with all the tools at our disposal. Natural sciences, which explore phenomena over vast ranges of time scales and varying complexities of matter, can be classified for convenience into physical sciences, chemical sciences, and biological sciences. Mathematics describes objects and actions in a world populated by abstract notions that we have developed from our observation of nature. The language of mathematics has been found to be remarkably useful in analyzing data, constructing models, and making predictions in almost all branches of science. Mathematics has led to the development of computer science and allied subjects too. Here, the 75 selected articles published over the years in Resonance are listed under four headings: mathematical sciences, physical sciences, chemical sciences, and biological sciences.

The idea underlying this collection of articles is to communicate to the readers, the vibrancy of science, the excitement of how science has developed over the years (history of science), the concepts and models that have evolved over time, where we are now and where we are headed in our understanding of phenomena—in and around us and abstracted by us.
The present collection of articles covers a wide spectrum of topics ranging from the smallest particle known, to the size of the universe and its past and present; the living and the nonliving; and the laws of nature that govern them all.

are now and where we are headed in our understanding of phenomena—in and around us and abstracted by us.

Selection of these 75 articles, from the large number published over the last 25 years has not been easy. It has been undertaken with the help of our present and former colleagues in the editorial team of Resonance. Since each article is scrutinized by experts in the area before it is accepted for publication in Resonance, the quality of an article alone would not have much discriminatory power. The colleagues examined all the articles published over the last 25+ years and selected 75 of them across the length and breadth of science and its history, with an attempt to accommodate sub-disciplinary diversity, as well as diversity among authors.

The present collection of articles covers a wide spectrum of topics ranging from the smallest particle known, to the size of the universe and its past and present; the living and the nonliving; and the laws of nature that govern them all. For the convenience of the reader, we have grouped the articles under four subheadings as mentioned above. We hope that the readers will enjoy reading them and feel motivated to contribute to Resonance in the years to come.

We take this opportunity to thank Professor P. Majumder, the past President of the Indian Academy of Sciences, for assigning the task to us and Professor U. Waghmare, the President of the Academy for continuing the support. It has been a learning experience for all of us. The benefit accrued to each one of us during the process cannot be quantified.

We have included in these two volumes, some of the cartoons drawn exclusively for Resonance by the inimitable cartoonist R K Laxman. We remember him fondly. We are grateful to Ayan Guha for contributing special cartoons under the heading ‘Science Smiles’ for each issue of Resonance over the last several years, and for volunteering to identify some of them for inclusion in this special edition. We are grateful to Subhankar Biswas for designing the cover page of Resonance-75.

We are grateful to Ms Srimathi and the Publications Team at the Academy for bringing out this volume in time, and we especially thank Ms Pushpa and Ms Geetha from the Resonance team for facilitating the production of these volumes.

Amber Habib, Amitabh Joshi, Anil J Elias, B Sury,
Jasjeet Singh Bagla, K Indulekha, Kaneenika Sinha,
Kusala Rajendran, Madhavan Mukund, N Sathyamurthy,
Nandini Nagarajan, Patrick Dasgupta, Susanta Mahapatra,
Shobhana Narasimhan, T N C Vidy, Varsha Singh, V S Borkar
Bengaluru
January 2022
Mathematical Sciences
You mean you didn’t think of horizontal flight when you invented vertical take-off?
A football is a 3-dimensional convex polyhedron with each face a regular pentagon or a regular hexagon and with at least one hexagonal face.

This article is in two parts. In this first part, we will prove that a football exists and is unique and in the second, we identify its group of symmetries. (We will incidentally do similar things for the platonic solids to some of which the football is closely related.) I heard of this problem from Amit Roy of TIFR, Mumbai. The ideas used in the proof of the existence and uniqueness are also his. Most of the other proofs presented here can be found in Gallian (1999) and Coxeter (1948).

A convex set is a set \( C \subseteq \mathbb{R}^3 \) such that \( A, B \in C \Rightarrow AB \subseteq C \). (Here \( AB \) denotes the segment joining \( A \) and \( B \) and we will study only convex subsets of \( \mathbb{R}^3 \).) Intersection of any family of convex sets is convex. There are plenty of examples: the empty set, a point, a line, a line segment, a plane, a half plane, a quadrant, a disc, an elliptic region, a half space (i.e., points lying on one side of a plane), a ball, a pyramid and a prism with a convex base (right or not). See V S Sunder’s articles [3] for a discussion on various aspects of convexity. The five platonic solids are convex. The two figures in Figure 1 are not. Note that the hexagon shown in the figure is equilateral but not equiangular. By a regular polygon we mean a plane polygon which is both equilateral and equiangular.

A convex polyhedron is a finite intersection of closed half-spaces. The disc and the cylinder are not convex polyhedra.

A convex polytope is a convex polyhedron which is bounded.

By a regular solid we mean a convex polyhedron such that the faces are all regular, equal polygons and the same number of faces occurs at each vertex. It was already known 2400 years ago that there are exactly five such solids, viz. the platonic solids, see Box 1. The Greeks associated the tetrahedron (this means a solid bounded by four faces) with fire, the cube with earth, the octahedron with air, the icosahedron with water and the dodecahedron with universe or cosmos. The study of dodecahedron was considered dangerous and restricted during some period. On the other hand, the dodecahedron was used as a toy at least 2500 years ago.

Apparently Theaetetus “first wrote on the ‘five solids’ as they are called” around 380 B.C. and probably knew that there are exactly five regular solids. Around 320 BC,
Box 1. Introduction to Geometry by H S M Coxeter, 1961.

- Tetrahedron \{3,3\}
- Cube \{4,3\}
- Octahedron \{3,3\}
- Dodecahedron \{5,3\}
- Icosahedron \{3,3\}
Aristaeus (known as ’the elder’) wrote a book called *Comparison of the five regular solids*. Euclid wrote his *Elements* around 300 BC.

In the diagrams in *Box 1*, the symbol \( \{p, q\} \), known as a Schläfli symbol, means that each face is a regular \( p \)-gon and that there are \( q \) faces at each vertex. Of the five regular solids, the cube and octahedron are duals of each other, the dodecahedron and the icosahedron are duals of each other and the tetrahedron is self-dual in the following sense: if we start with the cube and form a new solid by taking a new vertex at the centre of each face of the cube and joining two new vertices by an edge iff they are centres of adjacent faces of the cube, we get the octahedron. If we do the same starting from the octahedron we get back the cube; similarly for the dodecahedron and the icosahedron. (This duality is the same as that used for planar maps in graph theory.)

Incidentally, the tetrahedron, cube and octahedron are the crystal structures of sodium sulphanitrimoniate, sodium chloride (common salt) and chrome alum, respectively. The skeletons of certain microscopic sea animals called *Circorrhegma dodecahedra* and *Circogonia icosahedra* (and some other viruses) are in the shape of a dodecahedron and an icosahedron, respectively, see Gallian (1999) and Coxeter (1948). In 1985, Robert Curl, Richard Smalley and Harold Kroto created a form of carbon by using a laser beam to vapourize graphite. The resulting molecule has 60 carbon atoms arranged in the shape of a football. Curl, Smalley and Kroto received the Nobel Prize for this discovery in 1996.

We now show briefly how vertices, edges and faces, which we all understand intuitively, can be defined formally. An extreme subset of a convex set \( C \) is a convex set \( D \subseteq C \) such that \( C \in D, C \in AB, C \neq A, C \neq B \) and \( A, B \in C \Rightarrow A, B \in D \).

Such an extreme subset is called a vertex, edge or face accordingly as it is of dimension 0, 1 or 2. The dimension of a non-empty proper subset of \( \mathbb{R}^3 \) is 0 if it is a singleton, 1 if it is contained in a line and is not a singleton and 2 if it is contained in a plane and is not contained in any line. Note that a cube has 8 vertices, 12 edges and 6 faces.

Recall the Krein–Milman theorem which was discussed in V S Sunder’s article [3]. A simple consequence of the theorem is: *A convex polytope has finitely many vertices and is their convex hull.* Conversely, the convex hull of finitely many points is a convex polytope.

Every extreme subset of a convex polytope \( C \) is the intersection of \( C \) with a plane \( P \) such that \( C \) is contained in a half-space corresponding to \( P \). An extreme subset of an extreme subset is an extreme subset. Each edge of \( C \) is the line segment joining two vertices of \( C \) and is on the boundary of exactly two faces. Each face of \( C \) is a convex polygon formed by some edges of \( C \).

Since each face of a football is bounded, it can be proved that the football is bounded and, so, is a convex polytope. We omit this proof.
From now on, we consider only convex polytopes with dimension 3. Also, whenever we talk of $\angle ABC$, we shall mean that angle which is between $0^\circ$ and $180^\circ$. We prove the uniqueness of a football first assuming its existence and later prove the existence. We start with a simple result which is intuitively obvious.

**Lemma 1.** There are at least 3 edges at every vertex of a convex polytope.

**Lemma 2.** (Euclid, XI.20) Suppose $A, B, C$ and $O$ are not coplanar. Then $\angle AOB + \angle BOC > \angle AOC$.

**Proof:** We may assume that $\angle AOC > \angle AOB$, for, otherwise the result is trivial. Let $D$ be a point in the plane $AOC$ such that $\angle AOD = \angle AOB$ and $OD = OB$. Refer to Figure 2. We may take $C$ to lie on $AD$ extended. Now triangles $AOB$ and $AOD$ are congruent, so $AB = AD$. Since $AB + BC > AC$, we get $BC > AC - AD = DC$. So comparing triangles $DOC$ and $BOC$, we get $\angle BOC > \angle DOC$. So $\angle AOB + \angle BOC > \angle AOD + \angle DOC = \angle AOC$. □

**Lemma 3.** (Euclid, XI.21) The sum of the angles in all the faces at any vertex $u$ of a convex polytope with dimension 3 is less than $360^\circ$.

**Proof:** We may take the faces at $u$ to be $u_1u_{i+1}, i = 1, 2, \ldots, k$ where $u_1u_2 \ldots u_k$ is a convex polygon in a plane $P$ and $u \notin P$. See Figure 3. Let us call the angles of the type $uu_{i-1}$ or $uu_{i+1}$ base angles, angles of the type $u_{i-1}u_iu_{i+1}$ polygonal angles and angles of the type $u_1u_{i+1}$ vertical angles. Using the result that the sum of the angles in a triangle equals $\pi$, we see that the sum of the base angles and the vertical angles is $k\pi$. Using the fact that the sum of the two base angles at $u_i$ is greater than the polygonal angle at $u_i$, we see that the sum of all the base angles is greater than the sum of all the polygonal angles which is $(k - 2)\pi$. So the sum of all the vertical angles is less than $2\pi$. This proves the lemma. □

It may be worth noting here that the following simple ‘proof’ for the preceding lemma does not work always: let $v$ be the foot of the perpendicular from $u$ to the plane $P$. We may assume that $v$ lies inside the convex polygon $u_1u_2 \ldots u_k$. Then, it is perhaps natural to guess that angle $uu_{i-1}u_i < \angle u_{i+1}u_{i+1}$ for each $i$, and so the lemma would follow. But, the inequality stated can be false if one of the angles $uv_{i+1}$ and $v_{i+1}u_i$ is greater than a right angle (to get a counter-example, take angle $uv_{i+1}u_{i+1}$ close to $180^\circ$ and length $uv$ moderately large).

Next, we can single out an observation about the football.

**Theorem 1.** At every vertex of a football, there are exactly three faces and so three edges.

**Proof:** Since the angles in a regular pentagon are $108^\circ$ each and the angles in a regular hexagon are $120^\circ$ each, there cannot be more than three faces at any vertex by lemma 3. So the theorem follows from lemma 1. □

**Lemma 4.** If two regular polygons $\ldots ABCD \ldots$ and $\ldots XBCY \ldots$ in $\mathbb{R}^3$ have a
common edge \( BC \) (see Figure 4), then \( \angle ABX = \angle DCY \).

**Proof:** We first clarify that a regular polygon is, by definition, planar. Now triangles \( ABX \) and \( DCY \) are congruent since each is the reflection of the other in the plane \( P \) perpendicularly bisecting \( BC \). Thus, the lemma follows. \( \square \)

Here is another observation about the football.

**Theorem 2.** At every vertex of a football, there is exactly one pentagonal face (and so there are two hexagonal faces).

**Proof:** Suppose that at a vertex \( u \) there are three pentagonal faces \( P_1, P_2 \) and \( P_3 \). See Figure 5. Then, lemma 4 applied to \( P_1 \) and \( P_2 \) gives \( F_1 \) is a pentagon. Similarly, \( F_2 \) and \( F_3 \) are pentagons. Since we can go from \( P_1 \) to any face by passing along adjacent faces, it follows that all faces are pentagons, a contradiction. (Incidentally, there is a convex polytope called dodecahedron with 12 faces all of which are regular pentagons.)

Suppose next that at a vertex \( u \) there are two pentagonal faces \( P_1 \) and \( P_2 \) and a hexagonal face \( H_1 \) (see Figure 6). Then, lemma 4 applied to \( P_1 \) and \( F_1 \), gives \( F_2 \) is a pentagon. This gives a contradiction to lemma 4 when applied to \( P_1 \) and \( F_3 \).

Thus at any vertex there is at most one pentagonal face. By lemma 3, all the three faces at a vertex cannot be hexagonal, so the theorem follows. \( \square \)

**Lemma 5.** Suppose \( u_1u_2u_3u_4u_5 \) is a regular pentagon in some plane in \( \mathbb{R}^3 \). Then there is a unique way in which two regular hexagons can be attached at \( u_1u_2 \) and \( u_2u_3 \) so that they have a common edge \( u_2v \) and lie above the plane of \( u_1u_2 \ldots u_5 \). (See Figure 7).

**Proof:** Take \( u_2 = (0,0,0), v = (-1,0,0) \) and \( u_3 = (\alpha, \beta, 0) \). Refer to Figure 8.

Since \( \angle u_2u_3 = 120^\circ \), we have \( -\alpha = \langle v, u_3 \rangle = \cos 120^\circ = -1/2 \). Here \( \langle u, v \rangle \) denotes the dot product of \( u \) and \( v \). Since \( u_2u_3 = 1 \), we may take \( \beta = -\sqrt{3}/2 \). Thus \( u_3 = (1/2, -\sqrt{3}/2, 0) \).

Let \( u_1 = (\gamma, \delta, \epsilon) \). Since \( \angle u_2u_1 = 120^\circ \), we get \( \gamma = 1/2 \) as above. So \( u_1 = (1/2, \delta, \epsilon) \) where \( \delta^2 + \epsilon^2 = 3/4 \). Now \( \angle u_1u_2u_3 = 108^\circ \) and \( \cos 108^\circ = (1 - \sqrt{5})/4 \). So

\[
\frac{1 - \sqrt{5}}{4} = \langle u_1, u_3 \rangle = \frac{1}{4} - \sqrt{3} \delta.
\]

Hence \( \delta = \sqrt{5}/(2\sqrt{3}) \) and \( \epsilon = \pm 1/\sqrt{3} \). Assuming that \( u_1 \) lies above the \( x-y \) plane, we get \( \epsilon = 1/\sqrt{3} \). Thus \( u_1 = (1/2, \sqrt{5}/(2\sqrt{3}), 1/\sqrt{3}) \). Since a regular polygon is determined by three consecutive vertices, it follows that the relative positions of the three polygons at \( u_2 \) are uniquely determined and the lemma follows. \( \square \)

It is easy to write down the equations of the three planes at \( u_2 \) and so their normals at \( u_2 \). Using these, we can find the angle between the planes of the two hexagons to be \( \cos^{-1}(\sqrt{5}/3) \approx 41.81^\circ \) and the angle between the planes of the pentagon and each of the hexagons to be \( \tan^{-1}(3 - \sqrt{5}) \approx 37.38^\circ \).

At every vertex of a football, there is exactly one pentagonal face and so there are two hexagonal faces.
Figure 9.

We can now prove that there is at most one football.

**Theorem 3.** Given a regular pentagon $P$ in some plane in $\mathbb{R}^3$, a football with $P$ as a face and lying on a given side of the plane of $P$ is unique if it exists.

**Proof:** The positions of the five hexagons around $P$ are unique by the preceding lemma. Imagine attaching regular pentagons and regular hexagons (with the side same as that of $P$) in the order shown in Figure 9. At each stage, the type of face to be used is unique by theorem 2 and its position is unique since three or four consecutive vertices of a regular polygon determine the polygon. Hence the uniqueness follows. □

Here is a result which provides the key to actually assembling the football thereby proving its existence.

**Lemma 6.** Let $\ldots ABCD \ldots$ be a regular polygon. Let $BE$ and $CF$ be such that $\angle ABE = \angle DCF$ and $\angle EBC = \angle FCB$ (see Figure 10). If $E$ and $F$ are both on the same side of the plane of $ABCD$, then $E, B, C$ and $F$ are coplanar.

**Proof:** We may take $B = (-1/2, 0, 0), C = (1/2, 0, 0)$ and $D = (a, b, 0)$. Then clearly $A = (-a, b, 0)$. Now let $F = (\alpha, \beta, \gamma)$ and $E = (\delta, \epsilon, \phi)$. We may also suppose that $BE = CF = BC$. Then

$$\cos \angle EBC = \langle (\delta + \frac{1}{2}, \epsilon, \phi), (1, 0, 0) \rangle = \delta + \frac{1}{2}$$

and $\cos \angle FCB = -(\alpha - \frac{1}{2})$. So $\delta = -\alpha$. Now $\cos \angle ABE = \cos \angle DCF$ gives $b\epsilon = b\beta$ and so $\epsilon = \beta$. Now $BE^2 = CF^2$ gives $\phi^2 = \gamma^2$. Since $\gamma$ and $\phi$ have the same sign, they are equal. Thus $E = (-\alpha, \beta, \gamma)$ and $E, B, C$ and $F$ lie on the plane $\gamma y - \beta z = 0$. □

We now prove that a football exists. Why, one may wonder, because all of us have

---

The footballs we see are not supposed to be footballs as defined here because nobody wants to play football with a solid with sharp edges and corners.
seen footballs. Well, there is a problem here. Firstly, the footballs we see are not supposed to be footballs as defined here because nobody wants to play football with a solid with sharp edges and corners. (The edges of the football we see are geodesics and the faces are spherical regions.) Secondly it is possible that a football as defined here does not exist and the footballs we see are only approximations.

Theorem 4. The football referred to in theorem 3 exists.

Proof: We show that the football can be assembled as shown in Figure 9 used in the proof of theorem 3. Refer also to Figure 11. We start with faces 1, 2 and 3. This is possible by lemma 5. By lemma 4, \( \angle ABC = 120^\circ \), so we can attach face 4 at \( ABC \). Similarly we can attach face 5 also. Then by lemma 6, \( D, E, F \) and \( G \) are coplanar, so face 6 can be fitted there. Then clearly faces 7 and through 11 can be fitted. Then, again by lemma 6, \( H, I, J \) and \( K \) are coplanar, so face 12 can be fitted there. Next we can fit faces 13 through 16. Now \( Q, R, S \) and \( T \) are coplanar, so face 17 can be fitted there. Proceeding thus we fit faces 18-21, then 22-26, then 27-31 and finally face 32. This proves that the football can be assembled and so exists.

We next see how symmetric the football is. We start by showing that the vertices lie on a sphere. Note that the following analogue in two dimensions is false: the vertices of a convex polygon with all sides equal lie on a circle. The polygon shown in Figure 12 is far from equi-angular and can be perturbed further.

Theorem 5. The normals to any three mutually adjacent faces of a football are concurrent at a point \( O \) which is the centre of a sphere on which the vertices lie.

Proof: By the normal to a face we mean the line passing through its centre and perpendicular to its plane. Refer to Figure 13. Let \( AB \) be the common edge between two faces and \( C \) and \( D \) the centres of the two faces. Then it is easy to see that the plane perpendicularly bisecting \( AB \) will contain the normals to the two faces. So these normals are coplanar. Since the faces are not parallel, these normals intersect at, say, \( O \). Since \( O \) lies on the normal to face \( ABF \), we have \( OA = OB = OF \). Since \( O \) lies on the normal to face \( ABG, OB = OG \). Thus \( OF = OB = OG \). So \( O \) lies on the planes perpendicularly bisecting \( BF \) and \( BG \). Hence \( O \) lies on the normal to the face \( FBG \). Thus the normals to the three faces are concurrent at \( O \) and \( OA = OB = OF = OG \). By proceeding through adjacent faces, we can see that \( O \) lies on the normal to every face. This proves the theorem.

It will be an interesting exercise to determine the radius of the sphere on which the vertices of the football lie, given the length of an edge.

We now determine the numbers of vertices, edges and faces on the football. This will be needed later in part II where we determine its group of symmetries. Though these numbers can be counted from a drawing of the football, we will use a bit of graph theory to find these (partly explaining my interest in the topic).

A (finite) graph \( G \) consists of a finite non-empty set \( V \) whose elements are called
The numbers of vertices, edges and faces on the football can be counted from a drawing of the football but a bit of graph theory can be used too to find these.

**Lemma 7.** (Euler’s formula): For any plane graph $G$,

$$v - e + \gamma = 1 + p$$

where $v$ is the number of vertices, $e$ is the number of edges, $\gamma$ is the number of faces (including the unbounded face) and $p$ is the number of components.

**Proof:** We prove the result by induction on $e$. If $e = 0$, then $\gamma = 1$ and $p = v$, so the result follows. So assume the result for plane graphs with less than $e$ edges and let $G$ have $e$ edges. If an edge belongs to a ‘cycle’, then by deleting this edge, we get a plane graph with $v$ vertices, $e - 1$ edges, $\gamma - 1$ faces and $p$ components, so by induction hypothesis, we are done. If an edge $uv$ does not belong to any cycle, then by deleting this edge, we get a plane graph with $v$ vertices, $e - 1$ edges, $\gamma$ faces and $p + 1$ components, so by induction hypothesis, we are again done. □

**Theorem 6.** A football has 60 vertices, 90 edges and 32 faces of which 12 are pentagons and 20 are hexagons.

**Proof:** Any football can be represented by its *Schlegel diagram* which is what the football (assumed to be transparent except for the edges) appears like when seen from a position just outside the centre of one face. This is like stereographic projection from the top (assumed to be not a vertex and not lying on any edge) of the sphere on which the vertices of the football lie, onto a horizontal plane below the sphere. The Schlegel diagram is a plane graph $G$, vertices, edges and faces of the football corresponding naturally to those of $G$, the face of the football nearest to the viewer corresponding to the unbounded face. Note that $G$ has only one component, so Euler’s formula reduces to $v - e + \gamma = 2$. By theorem 1, there are exactly three edges at every vertex and every edge is incident with exactly two vertices. Thus $2e = 3v$. Since every vertex is incident with exactly one pentagon and each pentagon is incident with exactly 5 vertices, it follows that the number of pentagons is $v/5$. Since every vertex is incident with exactly two hexagons and each hexagon is incident with exactly 6 vertices, it follows that the number of hexagons is $v/3$. Substituting these in Euler’s formula we get

$$v - \frac{3v}{2} + \frac{v}{5} + \frac{v}{3} = 2,$$

so, $v = 60$. Now the theorem follows easily. □
Finally, we show how Euler’s polyhedral formula can be used to show that there are only five regular solids. In a plane graph, a $k$-cycle refers to a sequence of edges of the type $v_1v_2, v_2v_3, \ldots, v_{k-1}v_k, v_kv_1$ where the vertices $v_1, v_2, \ldots, v_k$ are all distinct and $k \geq 1$.

**Lemma 8.** If each face of a connected plane graph $G$ is a $p$-cycle for a fixed $p \geq 3$ and if there are $q \geq 3$ edges at every vertex of $G$, then $(p, q) = (3, 3), (3, 4), (3, 5), (4, 3)$ or $(5, 3)$.

**Proof:** We first note that every edge joins two distinct vertices since if there is a self-loop, the face just inside it cannot be a $p$-cycle with $p \geq 3$. So, as in the proof of theorem 6, we get $qv = 2e = p\gamma$. Now, by Euler’s formula, $v - e + \gamma = 2$. So

$$
\frac{v}{q} + \frac{e}{\gamma} = \frac{2}{q} = \frac{4pq}{2p - pq + 2q},
$$

So $2p - pq + 2q > 0$ or $(p - 2)(q - 2) < 4$. It follows easily that $p \leq 5$. Moreover, if $p = 3$, then $q$ can take only the values 3, 4 and 5. If $p = 4$ or 5, then $q$ can take only the value 3. □

**Lemma 9.** (Euclid’s Comment at the end of Book XIII): The Schlafli symbol of any regular solid (i.e., a 3-dimensional polytope with each face a regular $p$-gon and with exactly $q$ faces at each vertex) is $(3, 3), (3, 4), (3, 5), (4, 3)$ or $(5, 3)$.

**Proof:** We will give two proofs of this result, the first using Euler’s formula. The Schlegel diagram of any regular solid with Schlafli symbol $(p, q)$ is a plane graph $G$ satisfying the hypothesis of lemma 8, so $(p, q)$ can take only one of the five values mentioned in that lemma. This proves lemma 9.

We now give a second proof, essentially due to Euclid, which is applicable only to regular solids and which does not use Euler’s formula. Since each angle in a regular $p$-gon is $(1 - 2/p)\pi$ and there are $q$ such faces at any vertex, it follows from lemma 3 that $q(1 - 2/p)\pi < 2\pi$ which, on simplification, becomes exactly $2p - pq + 2q > 0$. Now the conclusion follows as in lemma 8. □

Now, if there is a regular solid with Schlafli symbol $(p, q)$, then $(p, q)$ can take only the five values stated in the preceding lemma. Using the equalities displayed in the proof of lemma 8 (or by direct counting), it is easy to find the numbers of vertices, edges and faces in each of the above five cases. Finally it can be shown, as for a football, that a regular solid with Schlafli symbol any one of the five referred to above, is unique. This shows that the platonic solids are the only ‘regular solids’ as stated by Euclid at the end of Book XIII.

Incidentally, a plane **tessellation** is a covering of the plane with nonoverlapping (except for the edges between) polygons. A tessellation is a regular **tessellation** if the polygons are all regular $p$-gons for some $p$. If a regular plane tessellation exists with $p$-gons and if there are $q$ such polygons at some vertex, then we have $q(1 - 2/p)\pi = 2\pi$, so $(p - 2)(q - 2) = 4$. It follows that $(p, q) = (3, 6), (4, 4)$ or $(6, 3)$. Each of these is
actually possible as the tessellations in Figure 15 show. (Note that, now, $p$ determines $q$ and the same number of polygons occurs at every vertex; this was not assumed in the definition).

We mention in passing that a beehive looks quite like a 3-dimensional convex polytope with every face a regular hexagon and with three faces meeting at every vertex. However, this cannot really be, since the sum of the angles at any vertex will then be $360^\circ$ and the polytope has to be planar by lemma 3. Thus the polytope with the stated properties does not exist and a beehive is only a clever approximation. This shows the need for proving the existence of a football.

We end the first part here. In the next part, we shall identify the group of symmetries of a football. That will also contain a discussion on the groups of symmetries of some other objects in 3-space. *

Suggested Reading


---

* A draft of the second part of the article was with the Editors, but was never published as Professor Rao passed away.
On Ancient Babylonian Algebra and Geometry *

Rahul Ray

1. Introduction

In an earlier article [1] we had discussed some aspects of ancient Babylonian mathematics as deciphered from various clay tablets excavated from modern Iraq, viz. the Pythagoras theorem and also the sexagesimal number system prevalent during the ancient Mesopotamian civilization. In this article, we study the exciting new approach of the last decade in the decipherment and interpretation of the Babylonian mathematical tablets.

In the 1930’s, following the discovery of a clay tablet pertaining to mathematics and dating to the Old Babylonian period (2000-1600 BCE), there was considerable activity in Germany and France in the interpretation of the tablets. The work of this period is to be found in the compendiums of Neugebauer [2] and Thureau-Dangin [3] whose analyses of the tablets are understandably influenced deeply by the mathematics current at our times.

However mathematics, like any other subject, is not culture-free; instead it is subject to the socially prevalent mores and conventions. Thus an understanding of the culture, language and history of the Mesopotamian civilization provides a better insight into the thought processes of the ancient Babylonian mathematicians. In this context, consider the following two examples given by Robson [4].

If asked to draw a triangle, most of us would draw a triangle with a horizontal base. However, a typical triangle drawn on the tablets of ancient Babylon has a vertical edge and the other two edges lying to the right of it, and none of them horizontal (see Figure 1).

Also, if asked to give the formula for the area of a circle, we would immediately say \( \pi r^2 \). Even in a situation where the radius \( r \) and circumference \( c \) were given to us, we would not give the formula \( \frac{c^2}{4\pi} \). The Babylonians, however, preferred the latter, as has been attested in almost all tablets dealing with areas of circles. Indeed, as Robson [4] writes “in modern mathematics the circle is conceptualised as the area generated by a rotating line, the radius. In ancient Mesopotamia, by contrast, a circle was the shape contained within an equidistant circumference. Even when the diameter of a circle was known, its area was calculated by means of the circumference. We also see this conceptualisation in the language used: the word kippatum, literally ‘thing that curves’, means both the two-dimensional disc and the one-dimensional

---


Keywords
Babylonian mathematics, sexagesimal system
In modern mathematics the circle is conceptualised as the area generated by a rotating line, the radius. In ancient Mesopotamia, by contrast, a circle was the shape contained within an equidistant circumference. Even when the diameter of a circle was known, its area was calculated by means of the circumference.

Nearly all the problems involving ‘quadratic equations’ are stated in terms of length, width, square, surface, height and volume. The solutions given in the tablets also use these terms. The earlier translations took these terms to be generic for the variables $x$ (length), $y$ (width), $x^2$ (square), $xy$ (surface), $z$ (height) and $xyz$ (volume). The translations from the tablets were as described by Neugebauer himself as being “substantially accurate” in the sense that the mathematical substance of the text was retained. Jens Hoyrup rereads the tablets and presents a “conformal translation”, which he describes as a translation which maps the original structure (i.e., subject, object, verb construction) and retains the etymological meaning of the Mesopotamian words. This research of Jens Hoyrup over the past decade (presented in Hoyrup [5]) provides a geometric understanding of the Babylonian methods. The geometric understanding, however, does not preclude the algebraic structure lying behind these solutions.

In Section 2 we consider three algebraic problems of the old Babylonian period. In Section 3 we discuss two geometric problems from this same period, and here we see the difference in the geometry used to study the algebraic and the geometric problems.

The usefulness of the Babylonian study of mathematics is to be seen in the development of astronomy. The study of astronomy in the Babylonian civilization dates from 1800 BCE to 300 BCE – although the most important work was done during the latter half of this period. Indeed, the ‘Astronomical Diaries’, which is the record of systematic observations of the heavenly bodies from 800/700 BCE to 100 BCE, is probably the longest uninterrupted stretch of scientific research in the history of any civilization. The Babylonian astronomers were indeed renowned for their work in their times as may be read from the book ‘Geography’ by the Greek geographer Strabo of Amasia: “In Babylon a settlement is set apart for the local philosopher, the Chaldaeans, as they are called, who are concerned mostly with astronomy; but some of these who are not approved of by the others profess to be writers of horoscopes. There are also several tribes of the Chaldaean astronomers. For example, some are called Orcheni, others Borsippeni, and several others by different names, as though divided into different sects which hold to various different dogmas about the same subjects. And the mathematicians make mention of some of these men: as, for example, Cidenas, Naburianus and Sudines”.

The influence of this work on Greek astronomical work is apparent. Not only was the Babylonian sexagesimal number system (Box 1) adopted by the Greeks for their astronomical work, the Babylonian development of mathematics for their astronomical circumference that defines it. The conceptual and linguistic identification of a plane figure and one of its external lines is a key feature of Mesopotamian mathematics. For instance, the word mithartum¹ (“thing that is equal and opposite to itself”) means both “square” and “side of square”. We run into big interpretational problems if we ignore these crucial terminological differences between ancient Mesopotamian and our own mathematics”.

¹ This word will be translated as ‘confrontation’ in the subsequent sections.

The ‘Astronomical Diaries’, which is the record of systematic observations of the heavenly bodies from 800/700 BCE to 100 BCE, is probably the longest uninterrupted stretch of scientific research in the history of any civilization.
Box 1. An Aside on the Sexagesimal Number System

To understand the sexagesimal number system, we first look at the familiar decimal number system, which is a positional system with base 10. Here the number 237 is understood to be \((2 \times 10^2) + (3 \times 10^1) + (7 \times 10^0)\), while the number 0.237 is \((2 \times \frac{1}{10}) + (3 \times \frac{1}{10^1}) + (7 \times \frac{1}{10^2})\). We need 10 different digits 0, 1, ..., 9 to express any number in this system. Computers however use a binary number system based on the two digits 0, 1. Thus the decimal number 27 = \((1 \times 2^4) + (1 \times 2^3) + (0 \times 2^2) + (1 \times 2^1) + (1 \times 2^0)\) will be translated as 11011 by the computer for its calculations, while the binary number 0.1011 is equivalent to the decimal number \((1 \times \frac{1}{2}) + (0 \times \frac{1}{2^2}) + (1 \times \frac{1}{2^3}) + (1 \times \frac{1}{2^4}) = 0.6875\). The sexagesimal system needs 60 digits, which for the present purpose and rather unimaginatively we denote by 0, 1, ..., 58, 59. Here the sexagesimal number 24\(\bar{9}\) equals \((24 \times 60^1) + (9 \times 60^0) = 1449\) in the decimal system, while the sexagesimal fraction 0;24\(\bar{9}\) equals the decimal number \((24 \times \frac{1}{60}) + (9 \times \frac{1}{60^2}) = 0.4025\). In modern mathematics, whenever a possibility of ambiguity arises, we write 237\(\mod 10\) to express the number 237 in the decimal system, 1011\(\mod 2\) for the binary number 1011 and 0.24\(\bar{9}\)\(\mod 60\) for the sexagesimal number 249. Note however that when just one number system is in use, one does not need to have a mathematical understanding of decomposition of a number in terms of its base ..., \(10^2\), \(10^1\), \(10^0\), \(10^{-1}\), ..., or ..., \(60^2\), \(60^1\), \(60^0\), \(60^{-1}\), ..., for day-to-day use.

The representation of the Babylonian number system was rather cumbersome. They had a symbol for 1, a symbol for 10 and a symbol for 60. The digits 1 to 9 were expressed by writing the requisite number of 1’s, either consecutively or bunched together in groups of three with one group on top of another. Similarly the ‘digits’ 10, 20, ..., 50 were expressed by the requisite number of 10’s, etc. Thus the number 147 would be represented by two symbols of 60, two of 10 and seven of 1. The digit 0 was not used and presumed to be understood from the context, although in the later Babylonian period it was denoted by a wedge.

The Babylonian development of mathematics for their astronomical studies contributed both the arithmetic and the geometric methods of the Greek studies. Studies contributed both the arithmetic and the geometric methods of the Greek studies, in addition to supplying the empirical observations which were used to build these mathematical theories. The Greeks were to further this study by their theories of arcs and chords to determine positions of celestial bodies. It is the latter which found full expression in the trigonometric works of Aryabhata and other Indian mathematicians. The Babylonian work also influenced Indian astronomy, as is shown in Neugebauer [1969], where he shows through various examples how whole sections of Varahamira’s *Pancha Siddhantika* may be explained by means of the Babylonian planetary texts.

This evidence of a Babylonian influence on Greek and Indian mathematics reveals the exaggeration in the assertions of the ‘greatest contributions’ of India in mathematics (e.g.[6] “the invention of the decimal notation and creation of modern arithmetic; the invention of the sine and cosine functions leading to the creation of modern trigonometry and creation of algebra”) and places the post-Vedic Indian contribution to mathematics and astronomy in the proper historical context. As in modern academics, in the ancient Chinese, Greek and Indian studies we find a continuity which is built on earlier works from other civilizations and other cultures.
2. Algebra

In this section we discuss the problems from three tablets of the Old Babylonian period.

Consider the problem\(^2\) from the tablet BM 13901 #1:

\[
\text{I totalled the area and (the side of) my square: it is } 0;\bar{45}.
\]

Clearly the problem may be written as \(x^2 + x = 0;\bar{45}\), where \(x\) is the (unknown) length of the side of a square. This equation we wrote involving the symbol \(x\) is, of course, a modern transcription of the problem.

To understand the above problem in the Babylonian cultural milieu, we see the ‘conformal’ translation in [5] of the problem together with its solution as given on the tablet\(^3\):

The surface and my confrontation I have accumulated: 0;\bar{45} is it. \(\bar{1}\), the projection, you posit. The moiety of \(\bar{1}\) you break, 0; 30 and 0; 30 you make hold. 0; 15 and 0; 45 you append: by \(\bar{1}\), \(\bar{1}\) is the equalside. 0; 30 which you have made hold in the inside of \(\bar{1}\) you tear out: 0; 30 the confrontation.

The first sentence says that the surface of the square and its edge are combined to give an area 0;\bar{45}. Since a 1-dimensional object and a two dimensional object cannot be added ‘geometrically’, the second sentence of the problem says that the 1-dimensional line is transformed into a 2-dimensional surface by giving it a thickness 1. Thus geometrically we have Figure 2.

The next two sentences ask us to bisect the newly projected surface and let the two pieces hold together. Next it asks us to append the square of area 0; 15 (i.e., of sides 0; 30) to the figure constructed earlier and obtain a square of area 0; 15 + 0; 45 = \(\bar{1}\).

These steps are shown in Figure 3.

In this larger square of area 1 (Figure 3c), the dotted part of the edge has length 0; \(\bar{30}\) and thus the length of the edge of the unknown square is 0; \(\bar{30}\).

In the language of modern mathematics, the first and the second lines produce the equation \(x^2 + x \times 1 = 0;\bar{45}\) as given in Figure 2. In the next few lines, the rectangle \(x \times 1\) is broken into \(x \times \frac{1}{2} + x \times \frac{1}{2}\) (Figure 3a and 3b) and then we ‘complete the square’.
by adding $0;30 \times 0;30$ (Figure 3c) to obtain $x^2 + x \times \frac{1}{2} + x \times \frac{1}{2} + 0;30^2 = 0;45 + 0;30^2$, i.e. $(x + 0;30)^2 = 1$; from which we obtain that $x = 0;30$.

Høyrup [5] analyses the texts of many such tablets and arrives at similar geometric solutions. To show that behind all these geometry there is an underlying algebraic structure we consider the bi-quadratic problem from the tablet $4$ BM 13901 # 12:

The surfaces of my two confrontations I have accumulated: $0;21\overline{40}$.
My confrontations I have made hold: $0;10$.

This problem asks us to find the lengths of the sides of two squares, given that the sum of their areas is $0;21\overline{40}$ and the product of the two lengths is $0;10$. The solution as presented on the tablet is:

The moiety of $0;21\overline{40}$ you break: $0;10\overline{50}$ and $0;10\overline{50}$ you make hold, $0;10\overline{50}$ inside $0;15\overline{72}\overline{140}$ you tear out: by $0;6\overline{172}\overline{140}$, $0;40$ is equalside. $0;40$ to one $0;1050$ you append: by $0;15$, $0;30$ is equalside. $0;30$ the first confrontation. $0;40$ inside the second $0;10\overline{50}$ you tear out: by $0;640$, $0;20$ is equalside. $0;20$ the second confrontation.

In this problem we understand the full import of the algebraic methods used in these geometric solutions.

The first sentence asks us to bisect a line of length $0;21\overline{40}$ and then make each of the equal parts form a square of area $0;15\overline{72}\overline{140}$.

Taking $u$ and $v$ to be the unknown lengths of the two squares we have, from Figure 4, $AB = u^2$, $BC = v^2$, $G$ is the mid-point of $AC = AB + BC = 0;21\overline{40}$ and $AGA'D$ is the square of sides $(u^2 + v^2)/2 = 0;15\overline{72}\overline{140}$ each. In Figure 4, the point $B'$ is such that $GB' = GB$.

The next sentence asks us to tear out an area $0;140 = 0;10 \times 0;10$ from $AGA'D$ to leave an area $0;6\overline{172}\overline{140}$, which is the area of a square of sides $0;4\overline{10}$ each.

Here although the area $0;140 = 0;10 \times 0;10$ is expressed as that formed by ‘holding’ two sides of length $0;10$ each, experts believe that this area is obtained as the area of a rectangle with sides of length $u^2$ and $v^2$. In Figure 5, this is depicted by the rectangle $EF'D'F$. Now the rectangles $ABFD$ and $A'B'F'D'$ are congruent. Thus the region $ABE'B'A'D$ has area $0;140$, and so after ‘tearing’ this region out from the square $AGA'D$ we are left with a square $BGB'E$ of area $0;6\overline{172}\overline{140}$. Now $BG = (u^2 - v^2)/2 = 0;4\overline{10}$.
The next few sentences obtain $u$ and $v$ from the relation $u^2 = ((u^2 + v^2)/2) + ((u^2 - v^2)/2)$ and $v^2 = ((u^2 + v^2)/2) - ((u^2 - v^2)/2)$.

Finally, we look at a problem from the tablet YBC 6504. This tablet has four problems together with their solutions. In all of these there is a rectangle of size $l \times w$ from which we are asked to ‘tear out’ a square of size $(l - w) \times (l - w)$ to leave a resultant area of 0; 820. In each of the four problems we have to obtain $l$ and $w$, when (i) $l - w$ or (ii) $l + w$ or (iii) $l$ or (iv) $w$ is given. We discuss the second problem of this tablet, i.e. when $l + w$ is given. As we will see in the next section, a similar construction is made in a geometry problem.

We explain the solution with the help of Figure 6, which is the construction suggested by the solution.
In Figure 6, as suggested in the first sentence on the tablet, $ABCD$ is the rectangle, from which the square $EFGC$ is ‘torn out’ leaving an area $0;820$. The second sentence asks us to ‘accumulate’ the length $l$ and width $w$ to form the line $AH$ of length $0;50$. The third sentence asks us to ‘make hold’, i.e. make the square $AHKL$ from the line $AH$ of area $0;50 \times 0;50 = 0;4140$. The fourth sentence “$0;4140$ to $0;820$ you append: $0;50$ you posit” suggests that we add the area of the square $AHKL$ to that of the region $ABEFGD$. Since the shaded smaller square inside $AHKL$ is congruent to the ‘torn’ square $EFGC$, we have that $0;50$ is equal to the area of five of the original rectangles $ABCD$. In the fifth sentence we are asked to “a 5th part you detach: $0;12$ you posit” which translates to $1/5 = 0;12$ and the next sentence says that $0;12 \times 0;50 = 0;10$. Now having obtained $lw = 0;10$ and knowing that $l + w = 0;50$, the next few sentences sets up the equations $(1/2)(l + w) = 0;25$ and $(1/2)(l - w) = 0;5$ to obtain $l = 0;30$ and $w = 0;20$.

It should be noted here that regarding this geometric solution no clay tablet has been found that contains these figures. Thus evidence as in a ‘smoking gun’ is not present to validate Hoyrup’s method based on his conformal translation. Although it may appear that an absence of a confirmation of Hoyrup’s analysis, in terms of figures on clay tablets, undermines the validity of his geometric solutions, they are in fact in complete and faithful accord with the texts of the tablets.\(^5\)

Moreover there is incidental evidence to support Hoyrup’s methods. First, as observed from tablets connected with land measurements, there are drawings of plots of land where the lengths, widths and angles are not in accord with those given numerically, i.e. the drawings do not respect scale or angles (e.g., see Figure 7 from the tablet IM 55357 given in this article). Instead these are\(^6\) “structure diagrams ... to identify and summarize the role of measured segments”. This suggests that a lot of the mathematics may have been carried out as ‘mental geometry’ (akin to the ‘mental arithmetic’ of our times). Second it has been suggested that cuneiform writing was practised on

\(^5\) In a personal communication SG Dani points out that “avoiding to draw figures especially in formal communications is to be found even today the reasons are presumably different but the parallel is intriguing”.

\(^6\) ibid. Hoyrup [5].

Figure 7. IM 55357.
YBC 7289: The YBC 7289 tablet from the Yale Babylonian Collection. The number on the horizontal line in the middle of the tablet (which is the diagonal of the square) is an approximation of \( \sqrt{2} \) accurate up to 6 decimal places. This is needed to obtain the length of the diagonal using Pythagoras' theorem. (Photo Credit: Yale Babylonian Collection)

IM 55357: The IM 55357 tablet from the collection of the Iraqi Museum in Baghdad. Unfortunately it disappeared after the invasion of Iraq in 2003. For more details of this tablet see Figure 7 and the accompanying text. https://www.irem.univ-mrs.fr/expo2013/english/Mesurer-A1-babylone-en.pdf

sand of the school yard; in which case these figures could also be drawn on sand (again similar to calculations on ‘rough paper’ of our times). Indeed the use of sand or dustboard in ancient Greece is also supported by familiar anecdotes of Archimedes drawing figures on sand. The Greek word for the dustboard is ‘abacus’, which is etymologically related to the Semitic word ‘abaq’ suggesting that the dustboard may have been originally from the Syro-Phoenician area. The close connections in ancient times of Mesopotamia and the Syro-Phoenician region strengthen the contention that the dustboard may have been available to the Babylonians.

To conclude this section we quote from [5] “Old Babylonian ‘algebra’ remained an art, not a science, if this is understood as an Aristotelian episteme whose aim is principles. On this account, however, any supposed algebra before Viète forsakes, however deep its insights. If we accept to speak of (say) Indian, Islamic, or Latin/Italian medieval ‘algebra’ as algebra, then we may safely drop the quotation marks and speak of Old Babylonian algebra without reserve”.

RESONANCE-75 | Promoting Science Education
3. Geometry

Regarding the geometry of the Babylonians, we had earlier [1] discussed some tablets related to the Pythagoras' theorem. As far as calculations of areas are concerned, the Babylonians knew methods to calculate the areas of right-angled triangles, rectangles and trapezium. For nearly rectangular quadrilaterals they would use the ‘surveyor’s formula’ which is the product of the average of the lengths and the average of the widths of the quadrilateral. The error inherent in this formula was also realized and so occasionally quadrilaterals were decomposed into smaller pieces so as to get a good approximation of their areas.

In this section we present problems from two tablets. The first tablet we discuss is IM 55357. Figure 7 is a reproduction with identifying letters of the figure on the clay tablet of the problem.

A triangle $10$ the length, $115$ the long length, $45$ the upper width.
$2230$ the complete surface. In $2230$ the complete surface, $86$ the upper surface.
$511;224$ the next surface, $319;356936$ the 3rd surface.
$553;5395024$ the lower surface.
The upper length, the shoulder length, the lower length and the descendant what?
You, to know the proceeding, igi $10$, the length detach, to $45$ raise,
$0;45$ you see. $0;45$ to $2$ raise, $1;30$ you see, to $86$ the upper surface raise, $129$ you see. By $129$, what is equalside? $27$ is equalside.
$27$ the width, $27$ break, $13;30$ you see. Igi $13;30$ detach, to $86$ the upper surface raise, $36$ you see, the length which is the counterpart of the length $45$, the width.

The text breaks off at this point.

The first five lines of the text sets up the problem and specifies the areas of the various triangles as in Figure 7. The sixth line obtains the ratio $AB/AC$ by multiplying $45$ and the reciprocal (igi) of $10$. In lines 7 and 8, this ratio, $0;45$ is multiplied with $2$ to obtain $1;30$, which is then multiplied with $86$ to obtain $129$, the square of $27$. Here the similarity of the triangle $ABD$ and $ABC$ is used to obtain $AB/AC = BD/AD = 0;45$ and then from the relation $(1/2)BD.AD = 86$ one obtains $BD^2 = 2(BD/AD)86 = 129$. Line 9 says that $BD = 27$ and then using the area of the triangle obtains $AD = 36$. The text proceeds in a similar fashion to obtain the length of the other unknown sides of the inscribed triangles, although the text breaks off before $DE$ and $EF$ are obtained.

One should note here that the text implicitly assumes that the triangle $ABC$ is right-
angled – an observation which the text setter probably had in mind because of the proportion $3 : 4 : 5$ of the sides of the triangle $ABC$. Also, the similarities of the triangles $ABD$ and $ABC$, $ADE$ and $ADC$, and $EDF$ and $EFC$ are assumed, though never stated. On the contrary, the lines $AD$ and $EF$ are not drawn perpendicular to $BC$ in the figure on the tablet. As to why this problem was solved by using the similarity properties of the triangles and not by using Pythagoras’ theorem, one can only speculate that the solution was illustrative of the use of the ‘surveyor’s formula’ in calculating areas of polygons by decomposing it into smaller triangles/rectangles which were similar, though not congruent.

Our next tablet DB$_2$-146 from the old Babylonian period exemplifies the use of Pythagoras’ theorem. Here the accompanying figure from the text is reproduced in Figure 8.

If, about a rectangle with diagonal, somebody asks you thus, $1;15$ the diagonal, $0;45$ the surface; length and width corresponding to what? You, by your proceeding, $1;15$, your diagonal, its counterpart lay down: make them hold: $1;3345$ come up, $1;3345$ may your hand hold $0;45$ your surface to two bring: $1;30$ comes up. From $1;3345$ cut off: ... $0;3345$ the remainder The equalside of $0;345$ take: $0;15$ comes up. Its half-part, $0;730$ comes up, to $0;730$ raise: $0;5615$ comes up $0;5615$ your hand. $0;45$ your surface over your hand, $0;455615$ comes up. The equalside of $0;455615$ take: $0;5230$ comes up, $0;5230$ its counterpart lay down, $0;730$ which you have made hold to one append: from one cut off. $1$ your length, $0;45$ the width. If $1$ the length, $0;45$ the width, the surface and the diagonal corresponding to what? You by your making, the length make hold: $1$ comes up ... may your head hold. ... $0;45$, the width make hold: $0;3345$ comes up. To your length append: $1;3345$ comes up. The equalside of $1;3345$ take: $1;15$ comes up. $1;15$ your diagonal. Your length to the width raise, $0;45$ your surface. Thus the procedure.

Here the first three lines sets out the problem: given a rectangle with diagonal $1;15$ and area $0;45$ what are its length and width? The fourth and fifth lines require that the diagonal of the rectangle be made to hold a square, i.e. a square with sides of length $1;15$. This square has area $1;3345$, which we are asked to keep at hand in line six. In lines seven to nine, two of the original rectangles are now ‘cut off’ from the square to yield a square of area $0;345$ and sides of length $0;15$. Comparing with the texts of other tablets, as well as that of the tablet YBC 6504, Hoyrup concludes that bringing the surface ‘to two’ and then to ‘cut off’ from the square results in the diagram given in Figure 9.

8 Note the ‘typo’ $0;3345$ in the text of the tablet.
Figure 9. The diagram which follows from the text of Db$_2$-146.

Figure 10. The Hsuan-thu diagram.
Once the middle square is known i.e. \((1/2)(l - w) = 0;7\overline{3}0\), where \(l\) and \(w\) are the unknown length and width of the rectangle, lines 10 to 15 establish \((1/2)(l + w) = \sqrt{[(1/2)(l - w)]^2 + lw} = \sqrt{0;0561} + 0;45 = 0;5230\). Thus, line 16 concludes that \(l = \overline{1}\) and \(w = 0;45\). The remainder of the text verifies that the solution is correct, by doing a ‘back calculation’.

As an aside compare the figure constructed for the previous problem with the Hsuan-thu\(^9\) diagram from the ancient Chinese text *Chou Pei Suan Ching* (The arithmetic classic of the Gnomon and the circular paths of the heavens) from around 1000 BCE. This diagram was used in the book to illustrate the *kou-ku*\(^{10}\) theorem.

**Suggested Reading**


---


\(^{10}\) The ancient Chinese nomenclature for the Pythagoras’ theorem.
This article is intended to give the reader a flavor of three-dimensional topology in an informal setting.

1. What is Topology and Why Study it?

Topology is a branch of mathematics, sometimes referred to as ‘rubber-sheet geometry’ based on the understanding that objects that can be made to look exactly like each other by stretching, bending, or repositioning without cutting, tearing, or poking holes, are ‘topologically equivalent’. Truth be told, while establishing topological equivalence, sometimes even cutting or tearing is allowed as long as the tear is properly repaired, as we shall soon see.

Topology is not only relevant in several areas of mathematics, but it also has applications in physical sciences. Biologists call on topological methods to examine the effects of certain enzymes on DNA and to study structures of neural networks; in modern chemistry, altering the chemical and physical properties of compounds may be achieved through the synthesis of topologically different molecules; configuration spaces used in robotics incorporate topology; topological techniques for dimension reduction and robustness against noise have proved extremely useful in analysis of large data sets that we nowadays routinely encounter; quantum field theories and gauge theories that originated in physics are closely connected to three and four-dimensional manifolds (special type of topological spaces to be defined later), and their studies have immensely benefited both mathematics and physics.

In the Science Smiles Section of March 2018 issue of Resonance [1], it was mentioned that a coffee cup and a doughnut are the same for a topologist. In essence, what topology captures is not the exact shape or size of an object but deeper structural properties, one such being the number of ‘handles’. Both a coffee cup and a traditional doughnut or a medhuwada have one handle each, but the double doughnut seen in Figure 1 has two, making it topologically inequivalent to the coffee cup.

A topological equivalence between objects or (topological) spaces is a continuous, one-to-one, onto function from one to the other with a continuous inverse. Such a

---

DOI: https://doi.org/10.1007/s12045-019-0785-5

This material is based upon work supported by and while serving at the National Science Foundation. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.
Knot theory studies knotted one-dimensional circles in the three-dimensional space. Two knots $K_1$ and $K_2$ are considered equivalent, or isotopic if there is a homeomorphism from the entire space to itself that maps one knot to the other.

function is called a ‘homeomorphism’, and equivalent spaces are said to be homeomorphic. We need mathematical structure to make sense of continuity, but as our examples are subsets of Euclidean spaces, continuous functions can be intuitively understood as ones that take ‘nearby points to nearby points’.

2. What Knot?

The curves shown in Figure 2 are all homeomorphic to each other. The homeomorphism can be seen in steps: First, snip the thread in any one place, then untangle the knot if there is one, and finally glue the snipped ends back together to repair the cut. In this process, points on the thread that were nearby have remained so, preserving continuity. Clearly, this can be done to any knot. But as we shall see, there is more to the story. The knot on the right in Figure 2 can be transformed into the one in the middle without any snipping as shown in Figure 3. Conclusion? These two knots are
equivalent to each other through a relationship that is stronger than a homeomorphism between knots. This is what ‘knot theory’, a subbranch of topology is built on.

Knot theory studies knotted one-dimensional circles in the three-dimensional space. Two knots $K_1$ and $K_2$ are considered equivalent, or isotopic if there is a homeomorphism from the entire space to itself that maps one knot to the other. A cut causes damage to the surrounding space as well as to the knot, and the damage to the space is not repaired when we make a local repair on the knot. If a knotted circle is transformed into another without cutting, that sequence of moves corresponds to a homeomorphism of the surrounding space to itself.

The knot in Figure 4 is not the same as (i.e. not isotopic to) the right handed trefoil, depicted on the right in Figure 2. Can you see the difference?

The pictures we have drawn are called the ‘knot diagrams’. These are two-dimensional projections with under-over crossing information that tells us about positioning in space. Two diagrams represent isotopic knots, if and only if, one can be transformed into the other as in Figure 3 without cutting.

Scientists originally became interested in knots because of theories about the physical universe, which have since become outdated. Mathematical interest in knot theory continued through their applications in algebraic geometry and topological studies of 3- and 4-dimensional manifolds, eventually returning to a myriad of new real-world applications. Mathematicians have proved that by removing a thickened knot (a knotted solid torus) from a three-dimensional space and rolling in the hole differently, we can obtain different kinds of three-dimensional manifolds. We will return to this in Section 5.

### 3. Managing the Surroundings

So far, we have vaguely referred to the space surrounding a knot as three-dimensional, giving the impression that we are working within the Euclidean space $\mathbb{R}^3$ with $x$, $y$, $z$ axes. Euclidean spaces extend to infinity in all directions. It is easier to work within more manageable spaces called the ‘compact manifolds’ that we describe below.

An $n$-dimensional manifold is a subset of some Euclidean space $\mathbb{R}^m$ (typically $m > n$) with the property that near any point, it ‘looks like’ an $n$-dimensional Euclidean space. For subsets of $\mathbb{R}^m$, compact means ‘closed’ and ‘bounded’, where closedness is characterized by every convergent sequence of points in the set having a limit inside the set itself, and boundedness indicates that the entire set lies within a finite distance from the origin.

For example, the circle $S^1 = \{(x, y) \in \mathbb{R}^2 | x^2 + y^2 = 1\}$ is a compact, 1-dimensional manifold: around any point of $S^1$ one can draw an ‘open interval’, just like on the number line!
At times it may be convenient to rely on formulas from algebra and geometry, but shapes are not fixed in topology.

Figure 5. A circle minus a point is homeomorphic to a line.

Do you see that the standard sphere $S^2$ is a 2-dimensional manifold?

$$S^2 = \{ (x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 = 1 \}.$$ 

Now, consider $f: S^1 - \{(0, 1)\} \to \mathbb{R}$ given by $f(x, y) = \left(\frac{x}{1-y}, 0\right)$. What this function does can be seen geometrically in Figure 5.

It maps a point on the circle (minus its north pole $N$) to the point where the line joining $N$ with the point on the circle intersects the $x$-axis. It is easy to show that this defines a homeomorphism from the circle with a point removed to the real line. Taking it backwards, the real line is homeomorphic to a subset of the circle that misses only one point on the circle. We can envision this phenomenon as the two sides of the real line bending up towards $(0, 1)$. The ‘point at infinity’ that both ‘ends’ of $\mathbb{R}$ approach is the point $N = (0, 1)$ in $\mathbb{R}^2$.

You may have encountered stereographic projection that maps a point, say $(p, q, r) \neq (0, 0, 1)$ on $S^2$ to the point in the $xy$-plane that the line joining $(0, 0, 1)$ and $(p, q, r)$ intersects, and provides a homeomorphism between $S^2$ minus a point and the Euclidean plane $\mathbb{R}^2$. Imagine the infinite plane bending upwards with its edges trying to reach the point at infinity, which happens to be conveniently located at $(0, 0, 1)$ in $\mathbb{R}^3$.

The general concept at play here is that an $n$-dimensional sphere is a one-point compactification of the $n$-dimensional Euclidean space. In particular, the three-dimensional sphere

$$S^3 = \{ (x, y, z, w) \in \mathbb{R}^4 \mid x^2 + y^2 + z^2 + w^2 = 1 \},$$

is a compact 3-dimensional manifold, and $S^3$ with a point removed is homeomorphic to $\mathbb{R}^3$. We interchangeably view knots as residing within $\mathbb{R}^3$ or $S^3$ and continue to use knot diagrams in the plane. If two knots are isotopic in $S^3$, they are so in $\mathbb{R}^3$, and vice versa.

This is a good place to reiterate that although at times it may be convenient to rely on formulas from algebra and geometry, shapes are not fixed in topology. Objects can be moved, stretched, and deformed by means of homeomorphisms. The radius of a topological circle can be as large or as small as the situation calls for, and $S^1$ could even be an ellipse or any other simple, closed curve. For our discussion, let’s agree to reserve the term ‘circle’ for unknotted, simple, closed curves.
4. Product Spaces and an Alternative Description of $S^3$

Any point in the Euclidean plane $\mathbb{R}^2$ can be located by its $x$ and $y$ coordinates. Moreover, any function whose codomain is $\mathbb{R}^2$ is continuous if and only if its projections onto the axes are continuous. This is an example of a product space that can be described in terms of its coordinate spaces. In case of $\mathbb{R}^2$ both coordinate spaces are $\mathbb{R}$, represented by the $x$- and the $y$-axis, respectively. Likewise, the 4-dimensional Euclidean space $\mathbb{R}^4$ can be viewed as a product space $\mathbb{R}^2 \times \mathbb{R}^2$. If the coordinate spaces $X$ and $Y$ are distinct spaces, the product space $X \times Y$ happens to be homeomorphic to $Y \times X$.

Can you see that a solid torus is spanned by a 2-dimensional disk $D^2$ rotated around a circle $S^1$? As we see in Figure 6, there is a circle corresponding to each point of the disk and the solid torus is a union of these circles.

We can write ‘coordinates’ for a point on the solid torus by specifying a point of $S^1$ and one of $D^2$. In this sense, a solid torus is (homeomorphic to) a product space $S^1 \times D^2 \subset \mathbb{R}^2 \times \mathbb{R}^2 \subset \mathbb{R}^4$. The boundary of this solid torus is a product of two circles $S^1 \times S^1$. One of these bounds a disk within the solid torus; this circle is called a ‘meridian’. The other $S^1$ that intersects a meridian once at right angles is called a ‘longitude’. A circle on the bounding torus parallel to a meridian is also a meridian and likewise for the longitude. The ‘core’ of the solid torus is the circle that passes through the center of a meridional disk and is parallel to a longitude.

Let $T_1 = \{ (x, y, z, w) \in \mathbb{R}^4 \mid x^2 + y^2 \leq 1/2, \ z^2 + w^2 = 1 - x^2 - y^2 \}$. The set $\{ (x, y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq 1/2 \}$ describes a 2-dimensional disk of radius $1/\sqrt{2}$. If we fix a point, say $(x_0, y_0)$ in this disk, then $1 - x_0^2 - y_0^2$ is a number $r_0^2$ with $1/\sqrt{2} \leq r_0 \leq 1$ and the set $\{ (z, w) \in \mathbb{R}^2 \mid z^2 + w^2 = r_0^2 \}$ is a circle of radius $r_0$. The set $T_1$ is homeomorphic to $D^2 \times S^1$. Its boundary is the product of a meridian $\{ (x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = 1/2 \}$ and a longitude $\{ (z, w) \in \mathbb{R}^2 \mid z^2 + w^2 = 1/2 \}$ each with radius $1/\sqrt{2}$. To get corresponding curves in $\mathbb{R}^4$, we have to fix a point $(x_1, y_1, z_1, w_1)$ with $x_1^2 + y_1^2 = 1/2 = z_1^2 + w_1^2$ and a meridian of $T_1$ in $\mathbb{R}^4$ is in fact:

$$\{ (x, y, z_1, w_1) \in \mathbb{R}^4 \mid x^2 + y^2 = 1/2 \}.$$
Figure 7. The 3-sphere as a union of solid tori.

Likewise, for the longitude. Finally, the core of $T_1$ is:

$$\{ (0, 0, z, w) \in \mathbb{R}^4 \mid z^2 + w^2 = 1 \}.$$

We really have to ‘stretch’ our imagination to see a solid torus for which the radius of the central circle is larger than that of a parallel circle on the boundary. Now consider another solid torus:

$$T_2 = \{ (x, y, z, w) \in \mathbb{R}^4 \mid z^2 + w^2 \leq 1/2, x^2 + y^2 = 1 - z^2 - w^2 \}.$$

Solid tori $T_1$ and $T_2$ are positioned in $\mathbb{R}^4$ so that a longitude of $T_1$ bounds a meridional disk of $T_2$, and a longitude of $T_2$ bounds a meridional disk for $T_1$. They intersect in a common boundary, and their union is:

$$T_1 \cup T_2 = \{ (x, y, z, w) \in \mathbb{R}^4 \mid x^2 + y^2 + z^2 + w^2 = 1 \} = S^3 \subset \mathbb{R}^4.$$

Figure 7 is an attempt to describe this in three dimensions. The vertical line along with the point-at-infinity represents the core of $T_2$.

5. Surgery to Create New Manifolds

In the early 1960s mathematicians Lickorish and Wallace, independently and by different methods, proved that any closed (compact without boundary), orientable, connected, 3-dimensional manifold may be obtained by removing a thickened link from $S^3$ and filling in the holes differently. This procedure is known as ‘surgery’. Links are disjoint unions of knots possibly entangled with one another. See examples in Figure 8. Links can have several component circles and these circles may themselves be knotted.
Let’s start with the core of $T_2$ as our 1-component link, and remove the interior (thickened core) of $T_2$ from $S^3$. This leaves:

$$T_1 = \{ (x, y, z, w) \in \mathbb{R}^4 | x^2 + y^2 \leq 1/2, \ z^2 + w^2 = 1 - x^2 - y^2 \} .$$

To fill the hole left behind by $T_2$ with a solid torus $T$, we now identify a longitude of $T$ with a longitude of $T_1$ (instead of a longitude with a meridian as done before to obtain $S^3$), and we identify the boundary of a meridional disk of $T$ with the boundary of a meridional disk of $T_1$.

We cannot draw the resulting space in three dimensions, but what we can see is that gluing two meridional disks along their boundaries (Figure 9), keeping interiors disjoint, results in a 2-dimensional sphere. A product of this sphere with the longitudinal circle is $S^2 \times S^1$. Techniques from algebraic topology show that this indeed is a new space, not homeomorphic to $S^3$.

The sphere $S^3$ and the product space $S^1 \times S^2$ are the two ends of a large spectrum of 3-manifolds obtained by surgery on a single unknot.

6. Classification and Invariants

Although knot and link diagrams may be easy to draw, and many of the research problems may be stated simply, a rigorous mathematical treatment takes work. For example, without additional tools, proving that the trefoil and the diagram in Figure 4 represent distinct, non-isotopic knots would amount to showing that no sequence

![Figure 8. Examples of links with 2 unknotted components.](image)

![Figure 9. Union of disks with boundary identifications.](image)
Typically, when mathematicians wish to study a complicated class or category of objects with a class of functions and an established sense of equivalence between two objects, they try to make a transition to an associated category of objects that are easier to compute, distinguish, or study.

The minimal number of crossings with which a diagram for a given knot or link can be drawn in the plane is a numerical invariant called the crossing number.

Of diagram moves can transform one into the other. How do we know that we have exhausted all possible move sequences? And for diagrams with many crossings, even if they happen to represent the same knot, showing this with diagram moves can take forever!

Typically, when mathematicians wish to study a complicated class or category of objects (e.g. topological spaces) with a class of functions (e.g. continuous) and an established sense of equivalence (e.g. homeomorphism) between two objects, they try to make a transition to an associated category of objects that are easier to compute, distinguish, or study. Such associated objects may be numbers, polynomials, more complicated algebraic structures, or something else entirely. The association is achieved in such a way that if an object in the original category is transformed into another through an equivalence, then the associated objects are equivalent in the new category as well. Such associated objects are called invariants. The minimal number of crossings with which a diagram for a given knot or link can be drawn in the plane is a numerical invariant called the crossing number. A link invariant called the Jones polynomial helped prove conjectures about crossing numbers that had been open for over a century. Its generalizations resulted in computable invariants of 3-manifolds.

Related work that was completed in 1984–87 led to Vaughan Jones and Edward Witten receiving the 1990 Fields Medal, the highest recognition in mathematics.

Knot theory continues to grow as a vibrant area of research within low dimensional topology that is rich with applications. We encourage readers to do an internet search for the latest information and free online resources for knots, links, and their scientific applications. Mathematical basics are covered in the reference books below listed in Suggested Reading: [2, 3] are intended for undergraduate students, whereas [4–6] are written for graduate students with knowledge of point-set and algebraic topology.

Suggested Reading

Hilbert’s Nullstellensatz and the beginning of Algebraic Geometry *

Vishwambhar Pati

1 A Question on polynomials

1.1 Introduction

In this article, we discuss a theorem that could be called the cornerstone of algebraic geometry over fields. In German, “Nullstellensatz” means “Zero-Set Theorem”.

The objects of study in algebraic geometry are the loci, or zero sets of polynomials. In school coordinate geometry (in 2 or 3 dimensions), one encounters things like the circle, which is the zero-set of the polynomial \( f(X, Y) = X^2 + Y^2 - 1 \) in \( \mathbb{R}^2 \), or the hyperboloid, which is the zero-set of \( f(X, Y, Z) = X^2 + Y^2 - Z^2 - 1 \) in \( \mathbb{R}^3 \). These objects are examples of “affine algebraic sets”, and were discussed in the article [1] in an earlier issue of Resonance.

1.2 Affine algebraic sets

So, more formally, let \{\( f_i(X_1, \ldots, X_n) \)\}\( _{i \in S} \) be some collection of polynomials in \( n \)-variables with real coefficients, indexed by some (finite or infinite) set \( S \). The real affine algebraic set \( V(S) \) is defined as:

\[
V(S) := \{(a_1, \ldots, a_n) \in \mathbb{R}^n : f_i(a_1, \ldots, a_n) = 0 \text{ for all } i \in S\}
\]

That is, \( V(S) \) is the set of common zeroes (or zero locus) of all the polynomials in \( S \). For brevity, it is often just called a real algebraic set. Similarly one can define a complex algebraic set as the common zero-set in \( \mathbb{C}^n \) of some collection \( S \) of polynomials in \( n \) variables \( X_1, \ldots, X_n \) with complex coefficients. (If \( S \) is a finite set, say \( S = \{f_i\}_{i=1}^k \), we customarily write \( V(S) \) as \( V(f_1, \ldots, f_k) \). It turns out, as will be discussed later, that all algebraic sets, real or complex, can be defined by only finitely many polynomials.)

It is clear that if \( f \) is a polynomial of the form:

\[
f(X_1, \ldots, X_n) = \sum_{f_i \in T} g_i(X_1, \ldots, X_n) f_i(X_1, \ldots, X_n)
\]

where \( T \) is any finite subset of \( S \), and \( g_i \) any polynomials in \( X_1, \ldots, X_n \), then \( f \) will vanish identically on the algebraic set \( V(S) \). Then one can ask the converse question:

---

Suppose one knows that some polynomial $f(X_1, \ldots, X_n)$ identically vanishes on $V(S)$. Can one assert that $f$ is somehow a combination of some $f_i$’s in $S$?

A trivial example shows that this is too much to expect. To wit, consider the degree 2 polynomial $F(X) = X^2$, and the algebraic set $V(F) \subset \mathbb{R}$. Clearly $V(F)$ is just the single point 0. The polynomial $X$ clearly vanishes on $V(F)$, but the degree 1 polynomial $X$ can never be $g(X)F(X)$, since $F(X)$ is of degree 2. This same example also works if $\mathbb{R}$ is replaced by $\mathbb{C}$. In other words, this hitch occurs over both $\mathbb{R}$ and $\mathbb{C}$.

Note, however, that $X^2$ is a multiple of (in fact equal to) $F$, so we may be tempted to make the modified

**Conjecture 1.1.** Suppose a polynomial $f$ with real (resp. complex) coefficients, in $n$-variables, vanishes identically on the algebraic set $V(S) \subset \mathbb{R}^n$, (resp. $\subset \mathbb{C}^n$). Then, we claim that there is a positive integer $r$ such that:

$$f^r = g_1f_1 + \ldots + g_kf_k$$

where $g_i$ are some polynomials with real (resp. complex) coefficients, and $f_i \in S$.

Let us first examine this conjecture in the real case, by looking at an example:

Take $V(F) \subset \mathbb{R}^2$ where $F(X, Y) = X^2 + Y^2$. Clearly, $V(F)$ is just the single point $(0, 0)$, the origin. The polynomial $X$ vanishes identically on $V(F)$. However, the reader can easily check that no power $X^r$ of $X$ can ever be a multiple of $F$.

Suppose we break out of $\mathbb{R}$, and consider this same example over the complex numbers $\mathbb{C}$. Now, $V(F) \subset \mathbb{C}^2$ becomes a large set. In fact $V(F) = \{(a, \pm ia) : a \in \mathbb{C}\}$ (where $i = \sqrt{-1}$) is a pair of (complex) lines in $\mathbb{C}^2$. Now we are in better shape because of the following:

**Exercise:** Prove that if a polynomial $f(X, Y)$ with complex coefficients vanishes identically on $V(F) \subset \mathbb{C}^2$, where $F(X, Y) = X^2 + Y^2$, then $f$ is divisible by $F$. In this example, one does not even need to raise $f$ to a power. (Hint: Change to new variables: $Z = X + iY$, $W = X - iY$, and show that any complex polynomial $f(Z, W)$ vanishing identically on both $Z$ and $W$ axes is divisible by $ZW$).

So, the conjecture above is verified at least for the above example, if we do everything over $\mathbb{C}$. Is it always true over $\mathbb{C}$? What can one say in the real case, if anything? Why are the two situations different? Can one use other fields besides $\mathbb{R}$ and $\mathbb{C}$? What are “fields”, anyway?

### 2 Some Algebra

In this section we’ll introduce some algebraic constructs that will help us to get a handle on the conjecture 1.1. The reader may wish to look at [2], [3] or [4] for further details. Fields were recently discussed in the article [5] in an earlier issue of *Resonance*. However, the account below is more or less self-contained.
The basic algebraic objects that we immediately need to get familiar with are rings, fields, and finally rings of polynomials with coefficients in a field. In the sequel, one could expend effort and censor out all mention of rings, fields, ideals and the like, and state everything in terms of polynomials, multiplication and division. But it is better to do some propaganda for the axiomatic method by showing you its effectiveness. Note that the conjecture 1.1 is a concrete question about polynomials, and we are introducing some machinery to solve this problem, and not for its own sake!

2.1 Rings and fields

A **ring** is a set $A$ together with two operations, say “$+$” and “$.$” (called addition and multiplication) which satisfy the following:

(R1) $a + b = b + a$ for all $a, b \in A$ (commutativity of $+$)

(R2) $a + (b + c) = (a + b) + c$; $a.(b.c) = (a.b).c$ for all $a, b, c \in A$ (associativity of $+$ and $.$)

(R3) There exists an element $0$ satisfying $a + 0 = a$ for all $a \in A$ (existence of additive identity).

(R4) For each $a \in A$, there exists an element $(-a)$ such that $a + (-a) = 0$ (existence of additive inverses).

(R5) $a.(b + c) = a.b + a.c$, and $(b + c).a = b.a + c.a$ for all $a, b, c \in A$ (left and right distributivity).

If in addition, we also have $a.b = b.a$ for all $a, b \in A$, we call it a **commutative ring**. If there exists an element $1 \neq 0$ and satisfying $a.1 = a = 1.a$ for all $a \in A$, we call it a **ring with identity**. A **subring** $B$ of a ring $A$ is a subset of $A$ which is also a ring with the $+$ and $.$ operations from $A$. If $A$ is a commutative ring with identity such that for each non-zero element $a \in A$, there exists an element $a^{-1}$ (called the multiplicative inverse of $a$) satisfying $a.a^{-1} = 1$, we call $A$ a **field**. So a field is an abelian group with respect to $+$, and its non-zero elements form an abelian group with respect to multiplication “$.$”.

Fields are usually denoted by the lowercase letter $k$, or the bold uppercase $F$. If a ring $A$ is a vector space over a field $k$, such that the scalar multiplication from $k$ is compatible with the ring operations (i.e. $\lambda(a + b) = \lambda a + \lambda b$ and $\lambda(a.b) = (\lambda a).b = a.(\lambda b)$ for all $\lambda \in k$ and all $a, b \in A$) then $A$ is called a $k$-**algebra**. In the particular case when $A$ contains a field $k$ as a subring, $A$ clearly becomes a $k$-algebra, with vector addition being the ring addition $+$, and scalar multiplication coming from ring multiplication “$.$” by elements of $k$.

**Example 2.1. (Examples of Rings, Fields, Algebras)**

(i) The set of integers $\mathbb{Z}$ is a commutative ring with identity, with its usual addition and multiplication operations. The **rational numbers** $\mathbb{Q}$, the **real numbers** $\mathbb{R}$ and

\[\text{(R3)} \quad a + 0 = a \text{ for all } a \in A \quad \text{(existence of additive identity).} \]

\[\text{(R4)} \quad a + (-a) = 0 \text{ for all } a \in A \quad \text{(existence of additive inverses).} \]

\[\text{(R5)} \quad a.(b + c) = a.b + a.c, \quad (b + c).a = b.a + c.a \text{ for all } a, b, c \in A \quad \text{(left and right distributivity).} \]

1. Take the definition of an $\mathbb{R}$-vector space, replace $\mathbb{R}$ by $k$ everywhere, and you have the definition of a $k$-vector space.
complex numbers $\mathbb{C}$ with their usual operations are fields. Clearly $\mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R} \subset \mathbb{C}$ is a chain in which each inclusion is one of a subring in a bigger ring. Thus $\mathbb{R}$ is a $\mathbb{Q}$-algebra (of infinite, in fact uncountable dimension as a $\mathbb{Q}$ vector space), and $\mathbb{C}$ is an $\mathbb{R}$-algebra of vector space dimension 2 (spanned by the basis $\{1, i\}$).

(ii) Let $m \geq 2$ be a natural number. The set of integers modulo $m$, (also called residue classes mod $m$) denoted $\mathbb{Z}_m$, is the set $\{0, 1, \ldots, m-1\}$. The sum $\overline{a} + \overline{b}$ is defined as $\overline{r}$ where $r$ is the remainder ($< m$) upon dividing the integer $a + b$ by $m$. Multiplication is defined similarly. These are called addition and multiplication modulo $m$, and $\mathbb{Z}_m$ becomes a commutative ring with identity under these operations. The properties R1 through R5 for $\mathbb{Z}_m$ follow from the corresponding properties for $\mathbb{Z}$.

(iii) If $m = p$ a prime, then $\mathbb{Z}_p$ becomes a field (why?), often denoted $\mathbb{F}_p$ to emphasise its "fieldhood".

(iv) The set of continuous complex valued functions on the closed interval $[0, 1]$ is a commutative ring with identity under the operations of multiplying and adding continuous functions pointwise, and is denoted $C([0, 1])$. Likewise $C(\mathbb{R})$, the commutative ring of continuous complex valued functions on $\mathbb{R}$. The ring $C(\mathbb{R})$ contains the subring $C_c(\mathbb{R})$ of continuous functions on $\mathbb{R}$ of compact support (i.e. continuous functions $f$ such that $f(x) = 0$ for $|x| > a$, for some $a$ depending on $f$). Note $C_c(\mathbb{R})$ is a ring without identity, (the constant function 1 is not of compact support!). In fact $C(\mathbb{R})$, since it contains the constant functions, is a $\mathbb{C}$-algebra. $C_c(\mathbb{R})$ is also a commutative ring (because multiplying a continuous compactly supported function with $\lambda \in \mathbb{C}$ gives a continuous compactly supported function) even though it does not contain $\mathbb{C}$ as a subalgebra! Another interesting way of making $C_c(\mathbb{R})$ a ring is to retain the old addition, but make multiplication the convolution product $f \ast g$ defined by:

$$ (f \ast g)(x) = \int_{-\infty}^{\infty} f(x-y)g(y)dy $$

With this new ring structure, $C_c(\mathbb{R})$ becomes a commutative ring without identity!

(v) The set of $n \times n$ matrices with entries in a commutative ring $A$ with identity, is a ring with identity under matrix addition and multiplication. It is denoted by $M(n, A)$, and is not a commutative ring (therefore not a field) if $n \geq 2$. More generally, for those who are aware of Hilbert spaces, the set of all bounded operators $B(\mathcal{H})$ on a Hilbert space $\mathcal{H}$ is a ring with identity.

(vi) The set of polynomials in $n$-variables with coefficients in a field $k$ is denoted $k[X_1, \ldots, X_n]$, and is a commutative ring with identity (with the operations of addition and multiplication of polynomials). It contains $k$ as the subring of degree 0 polynomials, and is therefore a $k$-algebra. It is called the polynomial ring or polynomial algebra over $k$ in $n$-variables. (Polynomials were discussed in the article [6] in an earlier issue of Resonance).

(vii) An important object is the field of rational functions $k(X_1, \ldots, X_n)$ in $n$-variables. It is defined as the set:

$$ \left\{ \frac{P(X_1, \ldots, X_n)}{Q(X_1, \ldots, X_n)} : P, Q \in k[X_1, \ldots, X_n], Q \neq 0 \right\} $$
(where common factors in \( P, Q \) can be cancelled without changing \( \frac{P}{Q} \)). **Caution:**

When we say that a polynomial:

\[
f(X_1, ..., X_n) = \sum_{i_1 + \ldots + i_n \leq d} a_{i_1 \ldots i_n} X_1^{i_1} \ldots X_n^{i_n}
\]

in \( k[X_1, ..., X_n] \) is non-zero (denoted \( f \neq 0 \)), we mean that some coefficient \( a_{i_1 \ldots i_n} \) of that polynomial is non-zero! Its values at all points of \( k^n \) may be zero. For example, the 1-variable polynomial

\[
f(X) = X^2 + X \in \mathbb{F}_2[X]
\]

gives zero when evaluated at the two points \( \{0, 1\} \) of \( \mathbb{F}_2 \), but \( f \) is not the zero polynomial. This, of course, doesn’t happen in \( \mathbb{R}[X_1, ..., X_n] \) or \( \mathbb{C}[X_1, ..., X_n] \) (Why not?).

The addition and multiplication in \( k(X_1, ..., X_n) \) are defined analogously to what we do for rational numbers, i.e. using a common denominator etc. Note that \( k(X_1, ..., X_n) \) is a field, because the inverse of a non-zero element \( P/Q \) is \( Q/P \). Also \( k(X_1, ..., X_n) \) clearly contains \( k[X_1, ..., X_n] \) as a subring, and so \( k \) is a subfield of this field, and so it is a \( k \)-algebra in a natural way. Those who have done some complex analysis will recognise the 1-variable case \( \mathbb{C}(X) \) as the ring of meromorphic functions on the complex plane having at worst a pole at \( \infty \).

### 2.2 Ideals

From this point on, all rings we consider will be assumed to be commutative, unless otherwise stated. For the sake of convenience we will write \( a \cdot b \) instead of \( ab \) for the product of the elements \( a \) and \( b \) in a ring. An **ideal** \( I \) of a ring \( A \) is a subset of \( A \) such that:

1. \( I \) is an additive group (I1)
2. \( ax \in I \) for all \( a \in A, x \in I \) (I2)

Clearly, if \( I \) contains the identity element \( 1 \), the property (I2) of an ideal would force \( I \) to be all of \( A \). **Thus the interesting ideals do not contain** \( 1 \). An ideal \( I \) in a ring \( A \) which is not equal to \( A \) is called a **proper ideal**. Again, from (I2), proper ideals do not contain any invertible element.

**Lemma 2.2.** The only proper ideal in a field \( k \) is the zero ideal \( \{0\} \). Conversely, if a commutative ring \( A \) with 1 contains no proper ideals except \( \{0\} \), then it is a field.

**Proof:** If an ideal \( I \) in a field \( k \) contains a non-zero element \( a \), then \( a^{-1}a = 1 \) is forced to lie in the ideal \( I \). Thus \( I = A \) by the above. Conversely suppose a ring \( A \) contains no non-zero ideals besides \( A \). Then for \( x \neq 0 \) in \( A \), consider the ideal \( Ax := \{ax : a \in A\} \). This is a non-zero ideal \( I \) in \( A \) since it contains \( x \), and therefore must be equal to \( A \). In particular it contains \( 1 \), so \( 1 = ax \) for some \( a \in A \), and \( a = x^{-1} \), so \( A \) is a field. \( \square \)

Let us run through some examples of ideals.
Example 2.3. (Examples of Ideals)

(i) If $A$ is any ring, and $S \subset A$ is any subset, define the subset $< S > \subset A$ by:

$$< S > := \{ \sum_{x_i \in F} a_i x_i : F \text{ a finite subset of } S, a_i \in A \}$$

$< S >$ is easily checked to be an ideal in $A$, called the **ideal generated by** $S$. Indeed, $< S >$ is the intersection of all ideals of $A$ that contain $S$ (prove!), and thus the smallest ideal of $A$ containing the subset $S$. If $S$ is a singleton $\{x\}$, the ideal $< S >$ is precisely $Ax$ introduced above, and called a **principal ideal**. If an ideal $I$ can be written as $I = < S >$ for a finite set $S$, it is said to be **finitely generated**.

(ii) The only ideals in $\mathbb{Z}$ are principal ideals, viz., $\mathbb{Z}m = < m >$ for some integer $m$. This can be seen by taking the smallest positive integer $m$ in the ideal $I \subset \mathbb{Z}$, and using the division algorithm to prove that if some $a \in I$ were not divisible by $m$, there would be a minimum strictly positive remainder $s = a - rm$ on dividing $a$ by $m$, and $s$ would (i) lie in $I$ (since $a$ and $m$ lie in it), and (ii) be strictly less than $m$, a contradiction. Since $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$ are fields, they contain no proper ideals except $\{0\}$ by lemma 2.2 above.

(iii) By the use of the result for $\mathbb{Z}$, it easily follows that the only ideals in $A = \mathbb{Z}_m$ are principal ideals $Ax$ for some $x \in A$.

(iv) Matrix rings $M(n, A)$ for $n \geq 2$ are not commutative, so left, right and 2-sided ideals have to be distinguished. For example the set of all matrices with vanishing first column is a left ideal. Left multiplying such a matrix with any matrix will give a matrix with vanishing first column, as will adding two such matrices. Similarly, those with vanishing first row will constitute a right ideal. If $A$ is commutative, and $I$ is an ideal in $A$, the subset $M(n, I)$ will be a 2-sided ideal. The set of all compact operators on a Hilbert space $\mathcal{H}$ is a two-sided ideal in $B(\mathcal{H})$.

(v) Let $Z \subset [0, 1]$. The set of all continuous functions in $C([0, 1])$ which identically vanish on $Z$ will constitute an ideal in $C([0, 1])$. One can define similar ideals in $C(\mathbb{R})$, and $C_c(\mathbb{R})$. The conjecture 1.1 can be formulated for $C([0, 1])$ as well: If a closed subset $Z \subset [0, 1]$ is the set of common zeroes of an ideal $I \subset C([0, 1])$, and $f \in C([0, 1])$ vanishes identically on $Z$, how does $f$ relate with $I$? The answer, which is the famous Gelfand-Naimark Theorem in analysis, says that $f$ lies in the closure of the ideal $I$ with respect to the sup-norm topology on $C([0, 1])$. The forthcoming Nullstellensatz can be thought of as the analogue of this theorem for polynomial rings.

(vi) Let $Z$ be *any* subset of $k^n$, where $k$ is a field. The ideal

$$I(Z) = \{ f \in k[X_1, \ldots, X_n] : f(a_1, \ldots, a_n) = 0 \text{ for all } (a_1, \ldots, a_n) \in Z \}$$

of all polynomials vanishing identically on $Z$ is an ideal in the polynomial ring $k[X_1, \ldots, X_n]$. It is called the **ideal of** $Z$. 


For any set $S \subset k[X_1, ..., X_n]$, one can define an **algebraic subset** $V(S)$ over $k$ as the following subset of $k^n$:

$$V(S) = \{(a_1, ..., a_n) \in k^n : f(a_1, ..., a_n) = 0 \text{ for all } f \in S\}$$

Note that $V(S) = V(<S>)$, so all algebraic sets are zero sets of some ideal, and we might as well only consider the algebraic sets $V(I)$ for $I \subset k[X_1, ..., X_n]$ an ideal. In fact, more is true:

**Proposition 2.4** (Hilbert Basis Theorem). If $k$ is a field, and $I \subset k[X_1, ..., X_n]$ an ideal, then $I$ is finitely generated. i.e. $I = < f_1, ..., f_m >$ for some polynomials $f_i$, $1 \leq i \leq m$.

In particular, every algebraic set is the common zero set of **finitely many polynomials**, since for any subset $S \subset k[X_1, ..., X_n]$:

$$V(S) = V(<S>) = V(<f_1, ..., f_m>) = V(f_1, ..., f_m)$$

We will not digress to prove the proposition 2.4. The case $n = 1$ is easy, because for single variable polynomial rings $k[X]$, one can divide one polynomial $f$ by another polynomial $g$ of degree $\deg g = d$ and get a remainder $r$ such that either $r = 0$ or $\deg r < d$. The proof given in (ii) of example 2.3 above to show that all ideals in $\mathbb{Z}$ are principal (singly generated) works also for $k[X]$, by choosing the polynomial of least degree in an ideal $I \subset k[X]$. Unfortunately, this “division algorithm” fails for $n \geq 2$. In fact, convince yourself that the ideal $< X_1, X_2 >$ in $k[X_1, X_2]$ cannot be singly generated. If you are unwilling to take the above proposition on faith, or look up the (one page) proof in any of the books listed above, be assured that you can still read on by mentally substituting “finitely generated ideal” whenever you read “ideal”, and be none the worse for it. Incidentally, finding minimal sets of generators for ideals in polynomial rings is an old and still very active area of algebra.

Now we can reformulate the conjecture 1.1 for any field $k$. Clearly, if $Z = V(I)$, i.e. the common zero set of some ideal $I \subset k[X_1, ..., X_n]$, then the ideal $I$ is certainly contained in $I(Z)$ (defined in (vi) of example 2.3 above). The conjecture 1.1 may be rephrased as:

**Conjecture 2.5.** Let $k$ be a field, and $I \subset k[X_1, ..., X_n]$ be an ideal. If $f \in I(V(I))$, is it true that $f^r \in I$ for some $r$ (depending on $f$)? Characterise the fields $k$ for which it is true. (We saw in §1.2 above that it is false for $k = \mathbb{R}$).

Why did we introduce ideals? Mainly because we can construct new rings out of old, “by going modulo” an ideal $I$. Let $I \subset A$ be an ideal in a commutative ring $A$. We define an equivalence relation in $A$ as follows: $a \sim b$ if $a - b \in I$. It is also often denoted as $a \equiv b \pmod{I}$ (“$a$ is **congruent to** $b$ **modulo** $I$”). Check that it is an equivalence relation. The set of equivalence classes $\{\overline{a} : a \in A\}$ is denoted $A/I$. It
also becomes a ring by defining \( \overline{a + b} := \overline{a} + \overline{b} \) and \( \overline{ab} := \overline{a} \overline{b} \). To check that these operations are well defined, i.e. if \( \overline{a} = \overline{a_1}, \overline{b} = \overline{b_1} \), then by definition, \( a - a_1 = x \in I \) and \( b - b_1 = y \in I \). Thus \( (a + b) - (a_1 + b_1) = x + y \in I \), i.e. \( \overline{a + b} = \overline{a_1 + b_1} \).

Similarly check that \( \overline{ab} = \overline{a} \overline{b} \). It is precisely here that all the properties of an ideal are used. \( A/I \) is called the quotient of \( A \) by \( I \), and as a familiar example, if we take \( A = \mathbb{Z}, I = \langle m \rangle \), then \( A/I = \mathbb{Z}_m \) discussed earlier. Historically, ideals were created by Kummer to handle the breakdown of unique prime factorisation of elements in rings like

\[
\mathbb{Z}[\sqrt{5}] := \{m + n\sqrt{5} \in \mathbb{R} : m, n \in \mathbb{Z}\}
\]

(a subring of \( \mathbb{R} \)) where \( 2.2 = (\sqrt{5} - 1)(\sqrt{5} + 1) \) are two distinct irreducible factorisations of 4), by replacing it with factorisation of ideals.

### 2.3 Coordinate rings, \( k \)-algebras of finite type

We can now give a geometric meaning to some quotients of polynomial rings. Let \( Z \) be an algebraic set in \( k^n \). If we take any polynomial \( f(X_1, \ldots, X_n) \), it can be regarded as a \( k \)-valued function on \( k^n \), by evaluating it at the point \( p = (a_1, \ldots, a_n) \). We can restrict such a polynomial function \( f \) to the subset \( Z \subset k^n \), and get a function on \( Z \). But different polynomials \( f, g \in k[X_1, \ldots, X_n] \) could end up restricting to the same function on \( Z \). This will happen if and only if \( f - g \) vanishes identically on \( Z \). That is, if and only if \( f - g \in I(Z) \). That is, if and only if \( f = g \) in \( k[X_1, \ldots, X_n]/I(Z) \). Thus \( k[X_1, \ldots, X_n]/I(Z) \) is just the ring of functions on \( Z \) which are restrictions of polynomial functions on \( k^n \), and is called the ring of \( k \)-regular functions or \( k \)-coordinate ring of \( Z \), and denoted \( k[Z] \) (not to be confused with the polynomial ring in the variable \( Z \)). The reformulated conjecture 2.5 asks to what extent an ideal \( I \) can be gleaned from its zero-set \( Z = V(I) \). Another way of formulating it would be to ask to what extent can the algebraic object \( k[X_1, \ldots, X_n]/I \) be gleaned from the geometric \( k \)-coordinate ring \( k[V(I)] = k[X_1, \ldots, X_n]/I(V(I)) \).

Since we are in the business of formalising everything, we note that if \( I \subset k[X_1, \ldots, X_n] \) is a proper ideal, its intersection with the degree 0 polynomials \( k \subset k[X_1, \ldots, X_n] \) is \( \{0\} \). Thus the quotient ring \( k[X_1, \ldots, X_n]/I \) also contains \( k \) as a subring, and thus becomes a \( k \)-algebra. Quotient rings of the form \( B = k[X_1, \ldots, X_n]/I \), where \( I \) is a proper ideal, are called \( k \)-algebras of finite type. The equivalence classes \( \overline{X_i} \) in \( B \) are usually denoted \( x_i \) for convenience, and called coordinate functions. Clearly, \( f(X_1, \ldots, X_n) = f(x_1, \ldots, x_n) \) by the definition of the ring operations in \( B \), and thus these \( x_i \)'s generate \( B \) as a \( k \)-algebra, i.e. every element of \( B \) can be written as a polynomial in the \( x_i \)'s with coefficients from \( k \). Unlike the polynomial ring however, some of these polynomials in the \( x_i \)'s maybe zero. For, if \( f(x_1, \ldots, x_n) \in I \), then by definition, the equivalence class \( f(\overline{X_1}, \ldots, \overline{X_n}) = 0 \) in \( B \), i.e. \( f(x_1, \ldots, x_n) = 0 \). Conversely, if \( f(x_1, \ldots, x_n) = 0 \), then \( f \in I \). Thus \( I \) is precisely the set of “constraints” or “relations” that measures the departure of the ring \( B \) from the polynomial ring \( k[X_1, \ldots, X_n] \). The ring \( B \) is denoted \( k[x_1, \ldots, x_n] \), the lower case letters reminding us that, unlike in the
polynomial ring $k[X_1, \ldots, X_n]$, there may be non-trivial relations between the $x_i$'s in $k[x_1, \ldots, x_n]$. For example, if we take the ideal $I = \langle X^2 - 2 \rangle$ in $Q[X]$, then in the ring $Q[X]/I = Q[x]$, we have $x^2 = 2$. This ring (which turns out to be a field, why?) is therefore also written as $Q[\sqrt{2}]$, and is in fact the smallest subfield of $R$ containing $Q$ and $\sqrt{2}$. Similarly, in the $R$-algebra of finite type $R[X]/\langle X^2 + 1 \rangle$ (which is $R[i] = C$), we have $i^2 + 1 = 0$. On the other hand, since $\pi$ is transcendental over $Q$, i.e. does not satisfy any non-zero polynomial with rational coefficients, the ring $Q[\pi]$ (defined as the smallest $Q$-subalgebra of $R$ containing $\pi$) is identical (isomorphic) to the polynomial ring $Q[X]$.

Here is a little lemma which comes in handy later on:

**Lemma 2.6.** If $B$ is a $k$-algebra of finite type, then the dimension of $B$ as a $k$-vector space (denoted $\dim_k B$) is countable.

**Proof:** By definition, $B = k[X_1, \ldots, X_n]$ for some $n$. Since every element in $B$ is a polynomial in $x_i$ with coefficients in $k$, it is a (not necessarily unique) finite $k$-linear combination of the monomials

$$\{x_1^{i_1}x_2^{i_2}\ldots x_n^{i_n} : i_j \in Z_+\}$$

which is a countable set $S$. By standard linear algebra $S$ contains a basis, which is therefore also a countable set. $\square$

3 Maximal ideals, field extensions

3.1 Maximal ideals

*From now on, all rings are assumed to be commutative, with 1.*

An ideal $I$ in a commutative ring $A$ is called a **maximal ideal** if it is a proper ideal, and contained in no proper ideal other than itself.

Let us prove that every proper ideal $I$ of a ring $A$ is contained in some maximal ideal. For, consider the family $\Sigma$ of all proper ideals which contain $I$, which is non-empty since $I \in \Sigma$. Partially order $\Sigma$ by inclusion, and check that if $\{J_\alpha\}_{\alpha \in \Gamma}$ is any totally ordered subset of $\Sigma$, then $\bigcup_{\alpha \in \Gamma} J_\alpha$ is a proper ideal, and in fact an element of $\Sigma$ which is an upper bound for this subset. Thus, by Zorn’s Lemma, (see [7] for a discussion of this all-pervasive lemma that is crucial to all mathematics!) $\Sigma$ contains maximal elements, which are just maximal ideals containing $I$. Since $\{0\}$ is a proper ideal in every ring, every ring has maximal ideals.

There is an easy way to determine if $I$ is maximal or not by looking at the quotient ring $A/I$. To wit,

**Lemma 3.1.** A proper ideal $I$ in a ring $A$ is maximal iff the quotient ring $A/I$ is a field.

\[A\] A deep theorem due to Lindemann and Weierstrass.
A polynomial \( f \in k[X] \) is said to be irreducible if it is not a product \( f = gh \) with \( g, h \in k[X] \) and \( \deg g < \deg f, \deg h < \deg f \).

**Proof:** Clearly if \( I \subset J \) for some ideal \( J \) of \( A \), then the set \( J/I := \{ \overline{a} : a \in J \} \) becomes an ideal in \( A/I \). Further, \( J \) is a proper ideal of \( A \) iff \( J/I \) is a proper ideal of \( A/I \). Finally, \( J = I \) iff \( J/I = \{0\} \). Thus \( I \) is a maximal ideal iff \( A/I \) contains no proper ideals other than the zero ideal \( \{0\} \). We saw in the lemma 2.2 of the last subsection that this happens iff \( A/I \) is a field. 

**Example 3.2.** We have seen that if \( A = \mathbb{Z} \), or the single variable polynomial ring \( k[X] \), all ideals in \( A \) are principal, i.e. of the form \( < a > \) for some \( a \in A \). Further, such ideals will be proper iff \( a \) is not an invertible element, i.e. \( a \neq \pm 1 \) if \( A = \mathbb{Z} \), and \( a \notin k \setminus \{0\} \) if \( A = k[X] \). If \( a > < b > \) are proper ideals in either of these rings, then \( a > < b > \) iff \( a \) is a multiple of \( b \). Thus \( a > \) is a maximal ideal iff \( a \) is not invertible, and divisible by nothing except itself, or 1, or other invertible elements in \( A \). In the case of \( \mathbb{Z} \), this happens iff \( a \) is \pm prime and \( \neq \pm 1 \). (Thus the lemma 3.1 explains (iii) of example 2.1). If \( A = k[X] \), this happens iff \( a \) is an irreducible polynomial\(^3\) in \( k[X] \) and not in \( k \setminus 0 \), i.e. irreducible of degree \( \geq 1 \). For example, \( X^2 - 2 \) is irreducible in \( \mathbb{Q}[X] \), and \( X^2 + 1 \) is irreducible in \( \mathbb{R}[X] \) (and the above lemma 3.1 explains why \( \mathbb{Q}[\sqrt{2}] \) and \( \mathbb{R}[i] = \mathbb{C} \) are fields). What are the irreducible polynomials in \( \mathbb{C}[X] \)? Before we answer this, let us remark that for the purposes of ideal generation in any polynomial ring \( k[X] \) with \( k \) a field, the ideal \( < \alpha f > \) for \( \alpha \neq 0 \) in \( k \), and \( f \in k[X] \) any polynomial. Thus, when talking of non-zero ideals \( < f > \), we can always assume that the leading coefficient (coefficient of the highest degree term) of \( f \) is 1. Such polynomials are called **monic polynomials**.

To see what the irreducible polynomials in \( \mathbb{C}[X] \) are, we need the:

**Theorem 3.3 (Fundamental Theorem of Algebra).** Every polynomial of degree \( \geq 1 \) in \( \mathbb{C}[X] \) has a root in \( \mathbb{C} \).

This theorem is the key to why complex numbers have such magical powers. A proof of this result can be obtained by using the Liouville theorem in complex analysis, or some elementary topology. See, for example [8] or [9], or [10].

**Corollary 3.4.** Every polynomial of degree \( \geq 1 \) in \( \mathbb{C}[X] \) is a product of polynomials of degree 1, i.e. linear polynomials.

**Proof:** If \( \deg f(X) = 1 \), there is nothing to prove. Otherwise, note that \( f \) has a root \( \alpha \in \mathbb{C} \), and is therefore divisible by \( (X - \alpha) \) (prove!). Thus \( f = (X - \alpha)g(X) \) where \( \deg g = \deg f - 1 \). Induct on degree. \( \square \)

Thus the only irreducible polynomials in \( \mathbb{C}[X] \) which are of degree \( \geq 1 \) are linear polynomials, and the above discussion of example 3.2 clearly yields:

**Corollary 3.5.** The only maximal ideals in \( \mathbb{C}[X] \) are the ideals \( < X - \alpha > \), for some \( \alpha \in \mathbb{C} \). Thus maximal ideals in \( \mathbb{C}[X] \) are in 1-1 correspondence with \( \mathbb{C} \).
Contrast this with $\mathbb{R}[X]$, where aside from the linear polynomials, quadratic polynomials like $x^2 + 1$ are also irreducible. In fact any quadratic polynomial $x^2 + ax + b$ with real coefficients $a$ and $b$ satisfying $a^2 < 4b$ will be irreducible in $\mathbb{R}[X]$. These, together with the linear polynomials $(X - a)$, exhaust all the irreducible monic polynomials in $\mathbb{R}[X]$ of degree $\geq 1$. (Why?)

The preceding discussion leads to the following definition-cum-lemma:

**Lemma 3.6.** For a field $k$, following conditions are equivalent:

(i) Every polynomial $f(X)$ of degree $\geq 1$ in $k[X]$ has a root in $k$, i.e. there exists an $\alpha \in k$ such that $f(\alpha) = 0$.

(ii) Every polynomial $f(X)$ of degree $\geq 1$ is a product of linear factors. Hence every root of $f(X)$ is in $k$.

(iii) The only irreducible polynomials of degree $\geq 1$ are linear polynomials.

(iv) $I$ is a maximal ideal in $k[X]$ iff $I = \langle X - \alpha \rangle$ for some $\alpha \in k$.

A field $k$ will be called an **algebraically closed field** iff it satisfies any of the above four equivalent conditions.

### 3.2 Field extensions

An inclusion $k \subset K$ of fields such that $k$ is a subring of $K$ is called a field extension. For example, $\mathbb{Q} \subset \mathbb{R}$, $\mathbb{R} \subset \mathbb{C}$, or $k \subset k(X_1, \ldots, X_n)$ where $k$ is any field, are all field extensions. If $k \subset K$ is a field extension, an element $\alpha \in K$ is said to be **algebraic over** $k$ if there exists a polynomial $f(X) \in k[X]$ such that $f(\alpha) = 0$. Otherwise it is said to be **transcendental** over $k$. For example $\sqrt{2} \in \mathbb{R}$ is algebraic over $\mathbb{Q}$. However, by the theorem of Lindemann-Weierstrass cited earlier, $\pi \in \mathbb{R}$ is transcendental over $\mathbb{Q}$. Similarly $1 + \sqrt{-5} \in \mathbb{C}$ is algebraic over $\mathbb{R}$, in fact even over $\mathbb{Q}$. Clearly every element of a field $k$ is algebraic over $k$. Further, Theorems 3.4, 3.5 of [5] are easily seen to imply that the subset of elements of $K$ which are algebraic over $k$ form a subfield of $K$. Thus, the elements of $\mathbb{C}$ which are algebraic over $\mathbb{Q}$ form the field of **algebraic numbers** denoted $\overline{\mathbb{Q}}$. If $k \subset K$ is a field extension in which every element of $K$ is algebraic over $k$, we call it an **algebraic field extension**. For example $\mathbb{R} \subset \mathbb{C}$ is an algebraic extension (why?), as is $\mathbb{Q} \subset \overline{\mathbb{Q}}$, but $\mathbb{Q} \subset \mathbb{R}$ is not, since e.g. $\pi$ is not algebraic over $\mathbb{Q}$.

There are many other algebraically closed fields besides $\mathbb{C}$. For example, the field $\overline{\mathbb{Q}}$ is algebraically closed! In fact, there is a result (proved using Zorn’s Lemma) which says that for any field $k$, there is a unique “smallest” algebraically closed field $\overline{k}$ containing it, called the **algebraic closure** of $k$. Since we won’t be needing this result, we will skip the proof. (see [3] for a proof). Clearly, theorem 3.3 implies $\overline{\mathbb{R}} = \mathbb{C}$. Note that no finite field $k$ can be algebraically closed, because the polynomial
An element \( x \) in a ring \( A \) is said to be prime if \( x \) divides \( ab \) implies \( x \) divides \( a \) or \( x \) divides \( b \).

An element \( x \) in a ring \( A \) is said to be prime if \( x \) divides \( ab \) implies \( x \) divides \( a \) or \( x \) divides \( b \). However, there is an algebraically closed field \( \overline{\mathbb{F}_p} \) in which \( p \) times every element is 0 (here \( p \) is a prime). The field of rational functions \( k(X_1, \ldots, X_n) \) is not algebraically closed. (Check that no rational function \( a = P/Q \) can satisfy the polynomial \( f(X) = X^2 - X_1 \), for example. You just mimic the proof that \( \sqrt{2} \) is irrational, and use the fact that unique factorisation into irreducibles (which are also primes\(^4\)) holds in \( k[X_1, \ldots, X_n] \). For a more detailed discussion of algebraic extensions, see the article [5].

In the light of (ii) of 3.6, we clearly have the following remark:

**Remark 3.7.** If \( k \subset K \) is an algebraic field extension, and \( k \) is algebraically closed, then \( k = K \).

Now we have an innocuous looking proposition which is actually equivalent to the nullstellensatz. We prove it for fields of uncountable cardinality, though it is true for all fields \( k \). The proof in general requires the Noether Normalisation Lemma. However, with the cardinality assumption on \( k \), we can give a simple proof which goes back to Krull and van der Waerden.

**Proposition 3.8.** Let \( k \subset K \) be a field extension such that \( K \) is a \( k \)-algebra of finite type. Assume that \( k \) is an uncountable set. Then \( K \) is an algebraic extension of \( k \).

**Proof:** As noted in the lemma 2.6, the vector space dimension \( \dim_k K \) is countable, by the hypothesis on \( K \). Every element of \( k \) is certainly algebraic over \( k \), so let \( \alpha \) be any element of \( K \setminus k \). Consider the set:

\[
S = \{ (\alpha - a)^{-1} : a \in k \}
\]

which makes sense since \( \alpha - a \neq 0 \) for \( a \in k \). Also, all of these elements are distinct. Thus the cardinality of \( S \) is the cardinality of \( k \), that is, \( S \) is an uncountable subset of \( K \). Thus all elements of \( S \) cannot be \( k \)-linearly independent, otherwise it could be enlarged to a \( k \)-vector space basis of \( K \), which would also be uncountable, and contradict that \( \dim_k K \) is countable. Thus there exist some non-zero elements \( \{\lambda_i\}_{i=1}^n \in k \) such that:

\[
\lambda_1(\alpha - a_1)^{-1} + \ldots + \lambda_n(\alpha - a_n)^{-1} = 0
\]

Now multiply this relation by \( \prod_{i=1}^n (\alpha - a_i) \) to get a polynomial relation \( f(\alpha) = 0 \) with coefficients in \( k \). (Check that \( f \) is not the zero polynomial!) Thus \( \alpha \) is algebraic over \( k \). \( \square \)

From the remark 3.7 and the proposition 3.8 above, we have the immediate:

**Corollary 3.9.** If \( k \subset K \) is a field extension such that (i) \( K \) is a \( k \)-algebra of finite type, (ii) \( k \) is uncountable and algebraically closed, then \( K = k \).
4 Hilbert’s Nullstellensatz

In the sequel, let $k$ be an uncountable algebraically closed field. (e.g. $k = \mathbb{C}$)

4.1 Maximal ideals in polynomial rings

As remarked before, the cardinality assumption on $k$ can be dropped, but since we need the corollary 3.9, which we have proved only with this assumption, we retain it. Note that $\mathbb{C}$ is an uncountable algebraically closed field, whereas $\mathbb{Q}$ and $\mathbb{F}_p$ are not (Why? Just enumerate all polynomials of all degrees with coefficients in $k = \mathbb{Q}$ or $\mathbb{F}_p$!)

We are now ready to analyse maximal ideals in $k[X_1, ..., X_n]$. We have the following proposition, which is often called the “small nullstellensatz”, even though it is equivalent to the later “big nullstellensatz”

**Proposition 4.1** (Hilbert’s Nullstellensatz I). Let $k$ be as above. Then an ideal $I \subset k[X_1, ..., X_n]$ is maximal iff $I = \langle X_1 - a_1, ..., X_n - a_n \rangle$ for some $a_i \in k$. In other words, maximal ideals are in 1-1 correspondence with points of $k^n$.

**Proof:** We remark that for the single variable case of $n = 1$, we have already seen this result in (iv) of lemma 3.6.

First, let us convince ourselves that all the ideals $I = \langle X_1 - a_1, ..., X_n - a_n \rangle$ are indeed maximal. For this we need to show that $A/I$ is a field (by lemma 3.1), where $A = k[X_1, ..., X_n]$. If $f(X_1, ..., X_n)$ is any polynomial in $A$, one can rewrite it as $f((X_1 - a_1) + a_1, ..., (X_n - a_n) + a_n)$. Now any power $((X_i - a_i) + a_i)^{n_i}$ can be written in the form $(X_i - a_i)h_i + a_i^{n_i}$ (by binomial expansion), where $h_i$ is a polynomial in $X_i$. Substituting these in $f$, we can “Taylor expand” $f$:

$$f(X_1, ..., X_n) = f(a_1, ..., a_n) + \Sigma_{i=1}^n g_i(X_1, ..., X_n)(X_i - a_i)$$

where $g_i \in A$ are some polynomials. The second term on the right hand side is clearly in the ideal $I = \langle X_1 - a_1, ..., X_n - a_n \rangle$. This shows that every element in $A = k[X_1, ..., X_n]$ is congruent to the element $f(a_1, ..., a_n) \in k$ (mod $I$). Also the element $a \in k \subset A$ is clearly congruent only to itself (mod $I$), and no other element of $k$ (since $I \cap k = \{0\}$). Thus $A/I$ is just the field $k$, and $I$ is maximal.

It is in the converse that the algebraic assumptions on $k$ are required. Let $I \subset A = k[X_1, ..., X_n]$ be a maximal ideal. Then by definition $A/I$ is a $k$-algebra of finite type, contains $k$, and is a field by lemma 3.1. Denote it by $K$. Thus $k \subset K$ is a field extension. By the corollary 3.9, $K = k$. Denoting the equivalence classes of $X_i$ by $\overline{X}_i \in K = A/I$, it follows that $\overline{X}_i = a_i \in k$ for all $i = 1, ..., n$. Saying that $\overline{X}_i = a_i$ in $K$, by definition, means $X_i - a_i \in I$ for all $i$. Thus the ideal $\langle X_1 - a_1, ..., X_n - a_n \rangle \subset I$. However, in the last para we saw that the ideal $\langle X_1 - a_1, ..., X_n - a_n \rangle$ is maximal. Thus it must be equal to $I$, which is a proper ideal (being a maximal ideal!). This proves the proposition. \(\square\)
Exercise 4.2. Prove that for \( k \) as above, a maximal ideal in \( A = k[X_1, \ldots, X_n] \) is precisely the ideal of all polynomials vanishing at some point \((a_1, \ldots, a_n) \in k^n\).

More generally, show that for any ideal \( I \subset A \), the points of \( V(I) \subset k^n \) are in 1-1 correspondence with all the maximal ideals in \( A \) which contain \( I \).

Corollary 4.3. Let \( k \) be as above, and \( \{f_i\}_{i=1}^m \) be some set of polynomials in \( A = k[X_1, \ldots, X_n] \). Then the algebraic set \( V(f_1, \ldots, f_m) \) is empty if and only if there exist polynomials \( h_i \in A \) for \( i = 1, \ldots, m \) such that \( \sum_{i=1}^m h_if_i = 1 \).

Proof: Consider the ideal \( J = \langle f_1, \ldots, f_m \rangle \). Then either \( J \) is a proper ideal or \( J = A \). If \( J \) were a proper ideal, by the first para of §3.1, it would follow that \( J \subset I \) for some maximal ideal \( I \). By the small nullstellensatz I above, \( I = \langle X_1-a_1, \ldots, X_n-a_n \rangle \). Thus every element of \( J \) would be of the form \( \sum_{i=1}^m g_i(X_i - a_i) \), and would thus vanish at the point \((a_1, \ldots, a_n)\). In particular \( V(J) = V(f_1, \ldots, f_m) \) would contain \((a_1, \ldots, a_n)\), and be non-empty. Thus \( V(f_1, \ldots, f_m) = \phi \Rightarrow J = A \), and hence \( 1 \in J = \langle f_1, \ldots, f_m \rangle \).

Note how crucial it is for \( k \) to be algebraically closed for the corollary above because, for example, \( V(X^2+1) \subset \mathbb{R} \) is empty, but you cannot multiply \( X^2+1 \) by any polynomial \( h \in \mathbb{R}[X] \) to get 1.

4.2 The big Nullstellensatz

Now, for the answer to the conjectures 1.1, 2.5 posed earlier!

Theorem 4.4 (Hilbert’s Nullstellensatz II). Let \( k \) be as assumed at the outset of this section (i.e. uncountable and algebraically closed). If a polynomial \( f \in A = k[X_1, \ldots, X_n] \) vanishes identically at all points of \( V(I) \) for some ideal \( I \subset A \), then \( f^n \in I \) for some \( n \).

Proof: By 2.4, write \( I = \langle f_1, \ldots, f_m \rangle \). If \( f = 0 \), there is nothing to prove, so assume \( f \neq 0 \). The trick is to add an extra variable, and “invert \( f^n \)” (called the Rabinowitch trick). Indeed, all the polynomials in \( A \) can be regarded as elements of the bigger ring \( B = k[X_1, \ldots, X_{n+1}] \). Consider the ideal \( J \subset B \) generated by the elements \( f_i(X_1, \ldots, X_n) \) for \( 1 \leq i \leq m \) and the extra element \( X_{n+1}f(X_1, \ldots, X_n) - 1 \). Claim that \( V(J) \subset k^{n+1} \) is empty. If not, there would be a point \((a_1, \ldots, a_{n+1}) \in V(J)\). Since all the \( f_i = f_i(X_1, \ldots, X_n) \) would have to vanish at this point, it would follow that \((a_1, \ldots, a_n) \in V(I)\). Also since \( X_{n+1}f - 1 \) would have to vanish at this point, we would have \( a_{n+1}f(a_1, \ldots, a_n) = 1 \). But then since \( f \) vanishes identically on \( V(I) \), and \((a_1, \ldots, a_n) \in V(I)\), we have \( f(a_1, \ldots, a_n) = 0 \). Thus \( a_{n+1} = 1 \), i.e. \( 0 = 1 \), a contradiction. This proves the claim.

So, by the corollary 4.3 above (applied to \( J \subset B \)), there must be polynomials...
\[ h_t(X_1, \ldots, X_{n+1}) \in B \text{ such that:} \]
\[ 1 = h_{m+1}(X_1, \ldots, X_{n+1})(X_{n+1}f(X_1, \ldots, X_n) - 1) + \sum_{i=1}^{m} h_i(X_1, \ldots, X_{n+1})f_i(X_1, \ldots, X_n) \]

The above identity holds in the polynomial ring \( B = k[X_1, \ldots, X_{n+1}] \). Substituting \( X_i = X_i \) for \( 1 \leq i \leq n \) and \( X_{n+1} = 1/f \) in this identity gives us an identity in the field \( k(X_1, \ldots, X_n) \), since \( f \neq 0 \). This substitution kills the first term, and we get:
\[ 1 = \sum_{i=1}^{m} h_i(X_1, \ldots, X_n, \frac{1}{f})f_i(X_1, \ldots, X_n) \]
as an identity in \( k(X_1, \ldots, X_n) \). Clearly, by using high enough power \( f^r \) as the common denominator on the right hand side, (e.g. \( r = \text{maximum of the } X_{n+1} \text{ degrees of all the } h_i \)), and cross-multiplying, we have:
\[ f^r = \sum_{i=1}^{m} P_i(X_1, \ldots, X_n)f_i(X_1, \ldots, X_n) \]
where \( P_i \) are some polynomials. This last identity holds in the field \( k(X_1, \ldots, X_n) \), and both sides of the identity are in \( A = k[X_1, \ldots, X_n] \). Since \( A \) sits as a subring in \( k(X_1, \ldots, X_n) \), the identity holds in \( A \). Thus \( f^r = \sum_i P_if_i \in I \) as asserted. □

5 Concluding remarks

5.1 Radical ideals and reduced algebras

If \( I \) is any ideal in a ring \( A \), one can define a new ideal called the radical of \( I \), denoted \( \sqrt{I} \) by:
\[ \sqrt{I} = \{ a \in A : a^r \in I \text{ for some } r \} \]
It is easy to check that this is an ideal, and it contains \( I \) by definition. An ideal \( I \) is said to be a radical ideal if \( \sqrt{I} = I \). Again it is easy to verify that the radical of any ideal is a radical ideal. The condition for an ideal \( I \) to be radical can be reformulated in terms of the quotient ring \( A/I \). Clearly, \( I \) is a radical ideal if \( A/I \) has no nilpotent elements other than 0 (a nilpotent element \( x \) in a ring is an element such that some power of it is 0). Rings which have no nilpotents other than 0 are called reduced.

The main example of a radical ideal is \( I(Z) \) where \( Z \) is an algebraic set, for clearly if \( f^r \) vanishes identically on \( Z \) for a polynomial \( f \), then \( f \) must also do so. Thus the \( k \)-coordinate ring \( k[Z] \) of an algebraic set \( Z \) is a reduced \( k \)-algebra of finite type.

Thus the nullstellensatz II (Theorem 4.4) implies:

Let \( k \) be an (uncountable) algebraically closed field, and \( A = k[X_1, \ldots, X_n] \) the \( n \)-variable polynomial ring over it:

(i) For \( I \subset A \) any ideal, \( I(V(I)) = \sqrt{I} \). If \( I \) is a radical ideal \( I(V(I)) = I \).

Thus, this reformulation has established a complete back-and-forth passage between the geometric objects called algebraic sets, and the algebraic objects called reduced \( k \)-algebras of finite type, when \( k \) is algebraically closed. This has beautiful and far reaching consequences not only for algebra and geometry, but all of mathematics!
(ii) The maps: $I \mapsto V(I)$, and $Z \mapsto I(Z)$ are inverses of each other from the set of radical ideals in $A$ to the set of algebraic sets in $k^n$.

(iii) $B$ is a reduced $k$-algebra of finite type iff it is the $k$-coordinate ring of some algebraic set in some $k^n$. (By definition, $B = k[X_1, ..., X_n]/I$ for some ideal $I$. $B$ reduced implies $I$ is a radical ideal. Now use (ii))

The reader may wonder what happens if we take some subset $S \subset k^n$, not necessarily an algebraic one, consider the ideal $I(S)$ of all polynomials vanishing identically on $S$ (which we have seen to be an ideal in (vi) of example 2.3), and then take the algebraic set $V(I(S))$. The answer is that one gets the smallest algebraic set containing $S$, called the “Zariski closure of $S$”. The algebraic subsets of $k^n$ define the closed sets of a topology on $k^n$, called the Zariski topology, and Zariski closure is closure in this topology. (Exercise: What is the Zariski topology on $C$? How does it compare with the usual topology?)

Thus, this reformulation has established a complete back-and-forth passage between the geometric objects called algebraic sets, and the algebraic objects called reduced $k$-algebras of finite type, when $k$ is algebraically closed. This has beautiful and far reaching consequences not only for algebra and geometry, but all of mathematics!

Suggested Reading

A Pizza Saga *

Five Delicacies of Modern Geometry

Shailesh A Shirali

In this article we introduce the reader to some pretty results concerning a circle and regions within it defined by various line segments. We have framed the results in terms of a pizza, so we could well call this an introduction to Pizza Geometry! The novel aspect is the variety of approaches used to prove these results.

1. First Pizza Theorem

Shown in Figure 1 is a circular region representing a pizza. Through an arbitrary point P within it, four chords AE, BF, CG and DH are drawn such that the angle between each pair of adjacent chords is $45^\circ$. This creates eight wedge shaped regions PAB, PBC, PCD, . . . (“eight wedges of the pizza’), of which, typically, no two will be congruent to each other (unless P lies at the centre of the circle). Let the regions be coloured alternately red and green, as shown in Figure 1.

The first pizza theorem states: The total area of the green regions equals the total area of the red regions, and this is so regardless of where P is located within the circle.

We shall give a delightful calculus based proof of this result, drawing on results from circle geometry; the proof is from [2]. The basic lemma used is the following: Let $P$ be a point within a circle $(O, R)$, and let $AC$ and $BD$ be a pair of perpendicular chords of the circle, intersecting at $P$. Then $PA^2 + PB^2 + PC^2 + PD^2 = 4R^2$. (See Figure 2.)

To see why, note that $PA^2 + PB^2 + PC^2 + PD^2 = \frac{1}{2}(AB^2 + BC^2 + CD^2 + DA^2)$. Next, from $AC \perp BD$ we get $\measuredangle AOB + \measuredangle COD = 180^\circ$; i.e., $\measuredangle AOB$ and $\measuredangle COD$ are supplementary angles. Since $AB = 2R \sin \frac{1}{2} \measuredangle AOB$, and $CD = 2R \sin \frac{1}{2} \measuredangle COD = 2R \cos \frac{1}{2} \measuredangle AOB,$ we get $AB^2 + CD^2 = 4R^2$. In the same way we get $BC^2 + AD^2 = 4R^2$, and the stated result follows.

Referring to Figure 1, let $P$ be the pole of a polar coordinate system, with ray PA as the axis; let the equation of the circle be $r = f(\theta)$. Then the above lemma may be expressed in terms of $f$ as follows:

$$f(\theta)^2 + f\left(\theta + \frac{\pi}{4}\right)^2 + f\left(\theta + \frac{\pi}{2}\right)^2 + f\left(\theta + \frac{3\pi}{4}\right)^2 = 4R^2,$$ for all $\theta$. \hspace{1cm} (1)

With the vertices labelled as shown in Figure 1 (i.e., counter-clockwise), the total area

---


Keywords

Pizza theorem, Euler–Poincaré formula, Ptolemy’s theorem, vector formula for area.
We give a calculus-based proof of the theorem.

The second pizza theorem reads the same as the first one.

We give a calculus-based proof of the theorem.

The second pizza theorem reads the same as the first one.

Figure 1. (left) Figure for the first pizza theorem.
Figure 2. (right) Lemma concerning two perpendicular chords (for the first pizza theorem).

The second pizza theorem reads the same as the first one.

We give a calculus-based proof of the theorem.

The second pizza theorem reads the same as the first one.

2. Second Pizza Theorem

For the second theorem, we start once again with a circular region representing a pizza, and select an arbitrary point $P$ within it. This time we mark eight points $A, B, C, D, E, F, G, H$ on the periphery of the circle at regular intervals (so that angles $AOB, BOC, COE, \ldots, HOA$ are all equal to $45^\circ$). We now join $P$ to each of these points, thereby creating eight wedge shaped regions $PAB, PBC, PCD, \ldots$ (“eight wedges of the pizza”), of which, typically, no two will be congruent to each other (unless $P$ lies at the centre of the circle). As earlier, we colour these regions alternately red and green, as shown in Figure 3.

The second pizza theorem, remarkably, reads the same as the first one: The total area of the green regions equals the total area of the red regions, and this is so regardless of where $P$ is located within the circle.

It is clearly sufficient to show that this equality of areas is true for the regular polygon $ABCDEFGH$ rather than the circle (see Figure 4), and this is what we shall do, using vector methods. With $O$ as the origin, let the position vectors of the various points be indicated using boldface symbols, thus: $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}, \mathbf{E}, \mathbf{F}, \mathbf{G}, \mathbf{H}$; these vectors have

\[
\int_0^{\pi/4} \frac{1}{2} f(\theta)^2 d\theta + \int_{\pi/2}^{3\pi/4} \frac{1}{2} f(\theta)^2 d\theta + \int_{\pi}^{5\pi/4} \frac{1}{2} f(\theta)^2 d\theta + \int_{3\pi/2}^{7\pi/4} \frac{1}{2} f(\theta)^2 d\theta
\]

Using relation (1) this reduces to

\[
\int_0^{\pi/4} 2R^2 d\theta = \frac{1}{2} \pi R^2 = \text{half the area of the circle (as claimed)}.
\]
equal magnitude, and the angles between A and B, between B and C, between C and D, . . . are equal to one another. We use the following well known formula: Given three points K, L, M, with position vectors K, L, M, respectively, the area of triangle KLM is the magnitude of the vector $\frac{1}{2}(L - K) \times (M - K)$. More symmetrically expressed, the area is the magnitude of $\frac{1}{2}(K \times L + L \times M + M \times K)$.

So the area of the red region is half the magnitude of the following vector:

$$(P \times A + A \times B + B \times P) + (P \times C + C \times D + D \times P) + (P \times E + E \times F + F \times P) + (P \times G + G \times H + H \times P).$$

The above expression simplifies to

$$P \times (A - B + C - D + E - F + G - H) + (A \times B + C \times D + E \times F + G \times H).$$

This further simplifies to

$$A \times B + C \times D + E \times F + G \times H,$$

because the vector sums A + E, B + F, C + G, D + H all vanish (by symmetry), so that A - B + C - D + E - F + G - H vanishes as well. Hence the area of the red region is independent of P.

In the same way we see that the area of the green region is half the magnitude of the vector

$$B \times C + D \times E + F \times G + H \times A,$$

and it is clear that (4) and (5) represent the same vector (since $A \times B = B \times C, \ldots$). So the total areas of the red and green regions are equal to each other, as claimed.

**Remark**

The two theorems proved above have featured the number 8, but this number can be substituted by any multiple of 8, and the corresponding result will continue to be true.
The obvious guess for $f(n)$ turns out to be wrong.

3. Third Pizza Theorem

For the third theorem, we mark $n$ points on the periphery of the pizza, and run straight cuts through the pizza along the lines joining all possible pairs of points; Figure 5 shows the cases $n = 2, 3, 4, 5$. Our intention is to divide the pizza into as many pieces as possible (the pieces are not required to be of equal size, so sympathies need to be offered to those who get the small pieces), so we must position the $n$ points in such a way that no three of the cuts concur. What is the largest number of pieces we can get? Let $f(n)$ denote this number. The values of $f(n)$ for $n = 1, 2, 3, \ldots$ may be found by experimentation (see Figure 5), and we get the following table of values:

<table>
<thead>
<tr>
<th>$n$</th>
<th>$1$</th>
<th>$2$</th>
<th>$3$</th>
<th>$4$</th>
<th>$5$</th>
<th>$\cdots$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f(n)$</td>
<td>$1$</td>
<td>$2$</td>
<td>$4$</td>
<td>$8$</td>
<td>$16$</td>
<td>$\cdots$</td>
</tr>
</tbody>
</table>

From this data we readily guess that $f(n) = 2^{n-1}$. But this guess turns out to be wrong. In fact, $f(6) < 32$.

Instead we have the following remarkably compact formula:

$$f(n) = \binom{n}{4} + \binom{n}{2} + \binom{n}{0}. \quad (6)$$

For example, $f(4) = \binom{4}{4} + \binom{4}{2} + \binom{4}{0} = 8$, and $f(5) = \binom{5}{4} + \binom{5}{2} + \binom{5}{0} = 16$. These agree with the figures given above.

Using the identity $\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}$, we may write (6) in the form

$$f(n) = \binom{n-1}{4} + \binom{n-1}{3} + \binom{n-1}{2} + \binom{n-1}{1} + \binom{n-1}{0}. \quad (7)$$

from which we can see why $f(n) = 2^{n-1}$ for $n - 1 \leq 4$, i.e., for $n \leq 5$. (Recall that for each $n$ the sum $\binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{n}$ equals $2^n$.)

To prove formula (6) we shall use the well known formula $v - e + f = 2$ which holds for any connected planar graph, with $v$, $e$, and $f$ representing the numbers of vertices, edges, and faces of the graph respectively. For example, for the case $n = 4$ in Figure 5, we have $v = 5$, $e = 12$, $f = 9$, and $5 - 12 + 9 = 2$, as we should expect. (Note that $v = 5$, and not 4, because the intersecting chords have created an additional vertex within the circle. Also, $f = 9$, and not 8, because there is the “outside” region too to be counted.)

---

1. The formula was first proved by Leonhard Euler in middle of the eighteenth century; but it had been proved earlier by René Descartes, in the seventeenth century, though in a more geometric form (it was with reference to the solid and face angles of a convex polyhedron). See [4].
To apply this to our setting, we need the values of $v$ and $e$. We first compute $v$. To start with, there are $n$ vertices on the circle. Each vertex within the circle is the intersection of the diagonals of a quadrilateral formed by four vertices on the circle. The number of such quadrilaterals is precisely the number of ways of choosing 4 vertices out of the set of $n$ vertices, hence it is equal to $\binom{n}{4}$. It follows that

$$v = \binom{n}{4} + n. \quad (8)$$

Now we compute $e$. Visit each vertex of the graph and count the edges emanating from it; then the total count we get in the end will be $2e$, since each edge gets counted twice (once from each end). Since the number of edges emanating from each vertex on the circle is $n - 1$, while the number of edges emanating from each interior vertex is 4, we get the following equality:

$$2e = n(n - 1) + 4 \binom{n}{4}. \quad (9)$$

It follows that

$$e = \binom{n}{2} + 2 \binom{n}{4}. \quad (9)$$

We are now in a position to compute $f$, using the relation $v - e + f = 2$:

$$f = \binom{n}{2} + 2 \binom{n}{4} - \binom{n}{4} - n + 2 = \binom{n}{4} + \binom{n}{2} + 2.$$

The number of regions within the circle is $f - 1$, and hence is equal to

$$\binom{n}{4} + \binom{n}{2} + 1.$$

This may be written in the more pleasing form $\binom{n}{4} + \binom{n}{2} + \binom{n}{0}$, as noted earlier.

For those who get the smaller pieces of the pizza in this highly unequal division, we trust that this formula offers more than enough compensation.

### 4. Fourth Pizza Theorem

Yet again we divide the pizza into wedges through a point not at its centre, but this time we divide it into 6 pieces, the angle between adjacent cuts being 60° (see Figure 6). Let the three cuts be AD, BE, and CF. Then we have the following pleasing relation between the lengths of the six line segments so created:

$$PA + PC + PE = PB + PD + PF. \quad (10)$$

Different proofs are possible, but we opt to give one using Ptolemy’s theorem which states that for a cyclic quadrilateral ABCD we have: $AC \cdot BD = AB \cdot CD + AD \cdot BC$.
To prove the fourth theorem, we use Ptolemy’s theorem of circle geometry.

Let the perpendicular bisectors \( l_1, l_2, l_3 \) of chords AD, BE, CF be drawn (see Figure 7). The same three lines are obtained no matter where \( P \) is located within the circle, and they divide the interior of the circle into 6 congruent sectors. With each point in the circle we associate a 3-tuple of signs, namely,

\[
\text{Sign} (PA - PD), \text{Sign} (PB - PE), \text{Sign} (PC - PF),
\]

where ‘Sign \( (x) \)’ refers to the sign of the real number \( x \). We shall call this the ‘signature’ of the point \( P \). There are 6 such tuples, one corresponding to each of the 6 sectors.

The argument to prove the theorem has to be phrased slightly differently depending on the sector in which \( P \) lies; but the essential logic is the same, so we need to give it for any one sector; there is no loss of generality in doing so. We shall suppose that \( P \) lies in the sector with signature \([-+,+]\).

Let \( L, M, N \) denote the midpoints of chords AD, BE, CF, respectively (see Figure 8). Then \( L, M, N \) lie on the circle \( w_1 \) with \( OP \) as diameter, and triangle LMN is equilateral. Since \( P \) lies in the sector with signature \([-+,+]\), the picture is as shown in Figure 8, with \( P \) on the minor arc LM of circle \( w_1 \). Ptolemy’s theorem applied to quadrilateral PMNL tells us that \( PN \cdot LM = PL \cdot MN + PM \cdot NL \), and since triangle LMN is equilateral, this implies that \( PN = PL + PM \). Hence we have:

\[
PC - PF = (PD - PA) + (PB - PE),
\]

and therefore:

\[
PA + PC + PE = PB + PD + PF.
\]

5. And A Fifth Pizza Theorem . . .

As observed by several writers, there is a fifth pizza theorem — and it is the oldest of them all, predating the others by two millennia; maybe predating pizzas as well!
This is the result that the volume of a pizza of radius \( z \) and thickness \( a \) is given by the expression

\[
\pi za.
\]

Happy eating!

**Postscript (added by December 2021)**

Closely related to the fourth pizza theorem (Section IV) are the following two beautiful results presented and proved in [5]:

**Theorem 1.** Suppose an odd number \( n \) \((n > 1)\) of chords of a circle meets successively at angles of \( \pi/n \) radians at a point \( P \) inside a circle. If \( c_1, c_2, \ldots, c_{2n} \) are the lengths of the segments of these chords taken counter-clockwise from \( P \) to the circle, then

\[
c_1 + c_3 + c_5 + \cdots + c_{2n-1} = c_2 + c_4 + c_6 + \cdots + c_{2n}.
\]

\[c_1 + c_3 + c_5 + c_7 + c_9 = c_2 + c_4 + c_6 + c_8 + c_{10}\] (the case \( n = 5 \) of Theorem 1).
**Theorem 2.** The sum of the alternate extensions of the sides of a convex equilateral polygon inside a circle is equal to the sum of the remaining alternate extensions.

The author proves these results using the intersecting chords theorem (the statement that if two chords $AB$ and $CD$ of a circle intersect in a point $P$, then $AP \cdot PB = CP \cdot PD$). By allowing the polygon in Theorem 2 to shrink to a point and appealing to considerations of continuity, we obtain the fourth pizza theorem.

For an elegant proof without words of the first pizza theorem, see [6].

**Suggested Reading**


The article [3] by Mabry & Deiermann extends the (first) pizza theorem in a pretty and non-obvious way, and is highly recommended.
On Shapes of Algebraic Loci
Wild-life in the Forests of Geometry

Subhashis Nag

It is of great interest to determine the topological shape of the locus of solutions of any given set of polynomial equations. Answer to this geometrical question leads us to the interaction of topology, algebraic/analytic geometry and differential geometry.

We have all surely studied something about certain simple loci defined by algebraic (namely polynomial) equations in the space of several real variables. Familiar examples are:

- \( \{(x, y) \in \mathbb{R}^2 : y = x\} \) defines a straight line - topologically the locus is the real line \( \mathbb{R} \).
- \( \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = 1\} \) defines the unit circle, \( S^1 \).
- \( \{(x, y) \in \mathbb{R}^2 : x^2 - y^2 = 1\} \) defines a hyperbola; this one is, topologically, the disjoint union of two copies of \( \mathbb{R} \).
- \( \{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 + z^2 = 1\} \) defines a 2-dimensional surface – the (unit) sphere, \( S^2 \).

Notice how just a little change of sign in the polynomial from the second to the third example above completely changes the topological shape of the locus. Whereas, if we were to replace the equation in the second example by say \( 1000x^2 + 100y^2 = 1 \) we would obtain an ellipse which is topologically no different from the circle \( S^1 \).

Thus the question arises, given a set of polynomial equations in some real or complex variables, can we determine the topological shape of the locus of solutions? Indeed what shapes can arise? Since solving equations is a fundamental matter in mathematics, the questions at hand are pretty basic. (Note that the solutions form a subset of the appropriate real (or complex) Euclidean space, according as we use polynomials with real (complex) variables.)

In this article I want to give you just a little hint of how various areas of Topology, Algebraic/Analytic Geometry and Differential Geometry interact to answer such geometrical questions about algebraic loci.

We will concentrate on 2-dimensional loci (surfaces) – since these are not very difficult to visualize. In general we will be looking at one real equation in three real variables, or two real equations in four real variables to define the surfaces. Notice that the latter

---

*Based on an invited talk at the Mathematical Carnival. Indian Institute of Technology, Chennai, March 10, 1996.


†Deceased
case arises naturally if we take one complex polynomial equation to determine the locus in the space of two complex (i.e., four real) variables. We shall explain these through some examples.

### 1. Topology of ‘Manifolds’

Firstly then, some topology (= ‘rubber sheet geometry’). We want some list of ‘good’ topological spaces which could be candidates for the shapes of our loci. Therefore we introduce the concept of \( n \)-dimensional manifold \( X \), as a space which is locally (topologically) equivalent to the \( n \)-dimensional Euclidean space \( \mathbb{R}^n \). We often indicate the dimension of a manifold by the superscript \( n \). Some examples are shown in the figures.

There are many ways of building up new, and more complicated, manifolds from old ones. (‘New lamps for old. . .!’). We consider the following operations:

- **Connected sums** – here we cut out little \( n \)-dimensional discs from two \( n \)-dimensional manifolds and paste the resulting surfaces together along the boundaries of the excised discs. See Figure 2. We get a (two-sided) surface of genus \( g \) by making a connected sum of \( g \) copies of the torus \( T^2 \).

- **Cartesian products** – here the dimension of the resulting manifold is the sum of the dimensions of the factors. Thus the cylinder \( \mathbb{R} \times S^1 \) as well as the torus \( T^2 = S^1 \times S^1 \) appear this way. The Möbius strip is a more general type of ‘twisted product’ (called a fibre bundle) of the real line with the circle. See Figure 1.

- **Quotient spaces** – these are identification spaces obtained by appropriate gluing operations. See Figure 3 for some interesting examples.

We can obtain the following three compact 2-dimensional manifolds (Figure 3), by identifying the two pairs of opposite sides of a square rubber sheet (make either a ‘parallel’ or ‘anti-parallel’ identification for each pair).
(i) torus $T^2 = S^1 \times S^1$

(ii) projective plane $\mathbb{P}^2$. This is the $2$-sphere $S^2$ with diametrically opposite points identified. Do you see that from the picture?

(iii) Klein bottle $K^2$.

The cases (ii) and (iii) have no faithful pictures in three dimensional Euclidean space — so they are a bit difficult to visualize. If you cut out a hole in $K^2$, then you can faithfully embed $K^2$ – disc in $\mathbb{R}^3$; see Figure 3. The figure shown for $\mathbb{P}^2$ is an ‘immersion’ in $\mathbb{R}^3$.

One-sided (non-orientable) manifolds: Note that the manifolds (ii) and (iii) are ‘one-sided’ surfaces (like the Möbius strip). (You can paint the entire surface in one colour without ever lifting your paint brush!) Question: Think about whether the connected sums of various copies of (i), (ii) and (iii) amongst themselves will be 1-sided or not.

By making connected sums of these basic objects we get a very interesting list of compact (namely, without any boundary and closing up on itself) $2$-dimensional manifolds. We get the following complete lists (see Siefert and Threlfall in Suggested Reading):

Orientable (two-sided) compact surfaces: The sphere $S^2$ (genus = 0), the torus $T^2$. (genus = 1), . . . , the $g$-holed torus (genus $g = 2, 3, 4, \ldots$). It is an exhaustive list of compact two-sided surfaces.

Non-orientable (one-sided) compact surfaces: We, of course, get a whole lot of compact one-sided surfaces by performing connected sums of copies of $\mathbb{P}^2$ and $K^2$. But here again, the entire family of non-orientable compact surfaces is known to arise by taking connected sums of copies of $\mathbb{P}^2$ alone.

Question: Can you see that the Klein bottle $K^2$ is itself the connected sum of two copies of the projective plane $\mathbb{P}^2$?

2. Algebraic Manifolds

The upshot of the above foray into the realm of topology is that we see a whole lot of interesting manifolds even when we restrict to two dimensions. We ask whether algebraic loci can realize these shapes. (Warning: It is not necessary that an algebraic locus should always look like a manifold. What about $xy = 0$, or curves with nodes, that you must have met in elementary coordinate geometry? The point is that at certain locations algebraic loci can develop ‘singularities’. At present, however, we will think only about the ‘non-singular’ or ‘smooth’ algebraic loci — i.e., those that represent manifolds.)

We will now work with solutions in complex space. Since the complex numbers are algebraically closed, questions about solution sets of polynomial equations become actually easier to analyze over $\mathbb{C}$ than over $\mathbb{R}$. (The real locus can be obtained from
the corresponding complex locus by taking the intersection with the subspace where
the imaginary parts vanish.)

Notice that one complex equation represents a pair of real equations (real part = 0, and imaginary part = 0). So one certainly expects to get 2-dimensional surfaces as the loci of solutions for one complex polynomial in two complex variables.

The surface determined as the locus of solutions:

\[ \mathcal{L} = \{(z, w) \in \mathbb{C}^2 : P(z, w) = 0\} \]

where \( P \) is a polynomial in two variables, must be unbounded, and therefore non-compact, in \( \mathbb{C}^2 \). In fact, by the algebraic completeness of the complex field, for any large complex value of \( z \) there must be solutions \( w \) satisfying \( P \). We wish to compare the surface \( \mathcal{L} \) with the compact manifolds we talked about earlier. Because of the non-compactness of \( \mathcal{L} \) it turns out that, in non-singular situations, \( \mathcal{L} \) must be topologically identical with one of our list of compact 2-dimensional surfaces minus some holes or punctures.

3. Examples

\[ \mathcal{L}_1 = \{(z, w) \in \mathbb{C}^2 : w^2 + z^2 = 1\} \]
\[ \mathcal{L}_2 = \{(z, w) \in \mathbb{C}^2 : w^2 + z^3 = 1\} \]
\[ \mathcal{L}_3 = \{(z, w) \in \mathbb{C}^2 : w^2 + z^5 = 1\} \]

I will not keep the answers a secret from you!

\( \mathcal{L}_1 \) is \( S^2 \) minus 2 points
\( \mathcal{L}_2 \) is \( T^2 \) minus 1 point
\( \mathcal{L}_3 \) is \( T^2 \# T^2 \) minus 1 point.

Compactifying the locus: By obvious intuitive process of ‘filling in the punctures’ it should be clear to you that there is indeed a compact surface associated to the locus \( \mathcal{L} \) given above. In fact, this intuitive process can be made mathematically rigorous by homogenizing the polynomial \( P \) and looking for the locus of solutions of the homogenized creature in complex projective space, \( \mathbb{C}P^2 \), of two complex dimensions, (instead of in \( \mathbb{C}^2 \)). See Walker in Suggested Reading for understanding this. (Example: For the second case \( \mathcal{L}_2 \) above, the homogenized polynomial would be \( w^2 t + z^3 - t^3 = 0 \), where \( t \) is introduced as the ‘homogenizing variable’. The locus in \( \mathbb{C}P^2 \) now indeed becomes the compact (unpunctured!) torus \( T^2 \), as you might expect.) Finally then, it transpires that there is a neat and natural way to associate a compact surface to any given polynomial locus \( \mathcal{L} \) as above. The resulting objects – which surely figure among some of Nature’s loveliest gifts to us – are called compact.
Riemann surfaces, or alternatively, complex algebraic curves. See Farkas and Kra, Siegel, Walker in Suggested Reading for more details.

If you had powerful computer graphics you could actually try to plot the locus of solutions and attempt to see the answers I assert! Can you produce a more general polynomial so that the locus surface is of any desired genus $g$, given an arbitrary $g \geq 0$?

Orientability of complex loci: Notice that all the loci seen from the complex equations above are two-sided surfaces. In fact, it turns out that since complex equations defines loci with ‘complex structure’ it is not hard to prove that all these loci must be two-sided, (i.e., orientable), surfaces. So $\mathbb{P}^2$, $\mathbb{K}^2$, or their connected sums can never appear as these shapes! Thus there are certain restrictions on the topological type of the shapes that can be realized by complex polynomial loci. That is a sample of the type of results we are after.

On the other hand, one-sided manifolds can still be realized as real algebraic loci. For example:

Projective plane $\mathbb{P}^2$ in $\mathbb{R}^4$: Verify that the mapping

$$q : \text{The unit sphere } S^2 \rightarrow \mathbb{R}^4$$

$$(u, v, w) \mapsto (uv, uw, vw, u^2 - v^2)$$

takes equal values only at diametrically opposite pairs of points. It follows that the image locus is nothing other than the projective plane $\mathbb{P}^2$. By eliminating $(u, v, w)$ you should find the algebraic equations in $\mathbb{R}^4$ representing the image. Thus $\mathbb{P}^2$ can be indeed embedded as a real algebraic locus in four dimensional real space!

4. Differential Geometry on Manifolds

I would like to close with a method from differential geometry – involving the curvatures of surfaces – that can be used to deduce the topological type.

Gaussian curvature of surfaces: Suppose that the locus at hand is a smooth, compact, two-sided surface, $X$. According to Gauss, we can calculate the curvature of $X$ at each point $x \in X$ as follows. Recall that the curvature of a plane curve at a given point on it is the reciprocal of the radius of the circle of closest fit (the osculating = kissing circle) at that point. Now we can look at a whole family of plane curves lying on $X$ and passing through $x$, by considering the sections of $X$ by normal planes (namely, the planes containing the normal direction at $x$). (We are thinking of $X$ as embedded in Euclidean space $\mathbb{R}^3$.) As the plane containing the normal rotates in space, one obtains this family of curves, each of which has a certain value for its curvature at $x$. The curvatures of these curves come with sign. Indeed, choose a positive direction for the normal vector at $x$. Then the curves for which the osculating circle has its centre on
Gaussian curvature is a real-valued function and is positive at ‘mountain-top-type’ (or ‘lake-bottom-type’) points, whereas it takes negative values at ‘saddle-type’ points.

the positive side of the normal will be assigned positive curvature, whereas the curves for which that circle lies on the negative side will be assigned negative curvature.

Now let \( k_{\text{max}} \) and \( k_{\text{min}} \) denote the maximum and minimum values attained by the curvature as the plane of section rotates. Gauss prescribes the product of \( k_{\text{max}} \) and \( k_{\text{min}} \) as the curvature of the surface itself at \( x \):

\[
\text{curvature}(x) = k_{\text{max}}k_{\text{min}}
\]

(Notice that the sign of this product is independent of the choice made for the orientation of the normal vector.) Thus, Gaussian curvature is a real-valued function on \( X \); it is positive at ‘mountain-top-type’ (or ‘lake-bottom-type’) points, whereas it takes negative values at ‘saddle-type’ points. (At ‘saddle-type’ points the curves representing the maximum and minimum values of curvature are concave towards opposite sides of \( X \). The two osculating circles lie on opposite sides of the tangent plane at \( x \). On the other hand, at places of positive curvature these two curves are concave towards the same side of the surface). It is a remarkable fact that the curves of maximum and minimum curvature, at each point of \( x \), are perpendicular to each other! I recommend the very interesting book by Hilbert and Cohn-Vossen (Suggested Reading) to learn more about the curvature of surfaces.

It is a great discovery of Gauss that the Gaussian curvature we just defined is an ‘intrinsic’ notion – depending only on the local geometry of \( X \) in a neighbourhood of \( x \); it does not really depend on the actual embedding of \( X \) in an ambient space, (even though the definition we gave above clearly requires such an embedding).

The Gauss-Bonnet Theorem: Now consider the total integral of the curvature function over the whole space. The Gauss-Bonnet theorem asserts:

\[
\int\int_{X} (\text{curvature}) d(\text{area element}) = 4\pi(1 - g)
\]

where \( g \) is the genus of \( X \). So this topological invariant – the genus – can be calculated by working out the integral on the left side! To calculate that integral, remember that one is doing calculus (or ‘differential geometry’) on the given locus. Isn’t that fascinating! See Dubrovin, Fomenko and Novikov in Suggested Reading for a discussion.

If \( X \) is the round sphere of radius \( r \), then the curvature on it is everywhere constant at \( 1/r^2 \). Do you see that the Gauss-Bonnet theorem is satisfied? On a smooth torus \((g = 1)\), for example, the theorem says that the total contribution from places of positive curvature must exactly balance with that from the places of negative curvature. (Can you have a torus whose curvature is everywhere zero? A ‘flat’ tyre!?)

A wild-life hunt: All the examples and methods that I have indicated above have vast generalizations organized into comprehensive theories. Manifolds are indeed certain species of the fascinating wild-life roaming in the forests of geometry. The hunting horns are blaring, I wish you a happy hunt!
Suggested Reading


From Calculus to Wavelets:
A New Mathematical Technique *

Wavelet Analysis

Gerald B Folland

A ‘wavelet’ is a function that exhibits oscillatory behaviour in some interval and then decays rapidly to zero outside this interval. A remarkable discovery of recent years is that the translations and dilations of certain wavelets can be used to form sets of ‘basic’ functions for expanding general functions into infinite series, and these expansions have many theoretical and practical applications.

The expansion of functions into infinite series is one of the most powerful techniques of mathematical analysis. The general idea is to write a function $f(t)$ (defined on the real line or an interval in the real line) as

$$f(t) = \sum c_n \phi_n(t),$$

where the $\phi_n$’s are a family of ‘basic’ functions with properties that are significant for the problem at hand and the coefficients $c_n$ are real or complex numbers depending on $f$. (As we shall see, sometimes it is more convenient to choose an index set for the $\phi$’s other than the natural numbers $0, 1, 2, \ldots$) For such an expansion to be useful, the coefficients $c_n$ must be computable in terms of $f$ in a reasonably straightforward way, and the series $\sum c_n \phi_n(t)$ must encode important information about the function $f$ in an easily accessible form.

Usually the first type of series expansion that one meets as a mathematics student is power series, that is,

$$f(t) = \sum_{n=0}^{\infty} c_n (t - a)^n$$

(2)

The coefficients of such a series are given by Taylor’s formula,

$$c_n = \frac{f^{(n)}(a)}{n!}.$$ 

In particular, a function cannot be represented as a power series unless it possesses derivatives of all orders (and even this is not a guarantee of the validity of (2)). However, if $f$ has derivatives up to order $N$, one can still use Taylor’s formula with remainder,

$$f(t) = \sum_{n=0}^{N} \frac{f^{(n)}(a)}{n!} (t - a)^n + R_N(t),$$

(3)

in which the remainder \( R_N \) vanishes more rapidly than \( |t - a|^N \) as \( t \to a \). This formula provides a very precise description of \( f(t) \) for \( t \) near \( a \) but is usually not of much use when \( t \) is far away from \( a \).

Another very useful type of series is the trigonometric series or Fourier series. These series come in several closely related forms, of which the simplest for analytical purposes is the one involving the complex exponentials \( e^{2\pi int} = \cos 2\pi nt + i \sin 2\pi nt \):

\[
f(t) = \sum_{-\infty}^{\infty} b_n e^{2\pi int}. \tag{4}
\]

It was Fourier’s great discovery that such series can be used to represent more or less arbitrary functions on the real line that are periodic with period 1 (that is, \( f(t+1) = f(t) \) for all \( t \)). More precisely, if \( f \) is any periodic function – even one with lots of discontinuities and infinite singularities – for which the area under the graph of \( |f|^2 \) over one period is finite,

\[
\int_{-1/2}^{1/2} |f(t)|^2 \, dt < \infty, \tag{5}
\]

then \( f \) admits an expansion of the form (4). (If \( f \) is a rather ‘rough’ function, there may be some exceptional points where the series does not converge to \( f(t) \), but it converges to \( f(t) \) ‘almost everywhere.’) The Fourier coefficients \( b_n \) are given by a simple integral formula, whose meaning is explained in Box 1:

\[
b_n = \int_{-1/2}^{1/2} f(t) e^{-2\pi int} \, dt. \tag{6}
\]

One of the virtues of the Fourier representation (4) is that it converts differentiation into a simple algebraic operation. Indeed, term-by-term differentiation of (4) yields

\[
f'(t) = \sum_{-\infty}^{\infty} 2\pi in b_n e^{2\pi int}.
\]

This needs some justification, but the result is correct: Suppose \( f \) is differentiable and \( f' \) satisfies condition (5). If the Fourier coefficients of \( f \) are denoted by \( b_n \), then the Fourier coefficients of \( f' \) are \( 2\pi in b_n \). Similarly for higher derivatives: the Fourier coefficients of \( f^{(k)} \) are \((2\pi in)^k b_n \).

As a consequence, it is easy to read off the order of differentiability of \( f \) from its Fourier coefficients. If \( f \) has derivatives up to order \( k \) that satisfy (5), then the series \( \sum (2\pi in)^k b_n e^{2\pi int} \) must converge, and hence \( n^k b_n \) must tend to zero as \( |n| \to \infty \). In other words, \( b_n \) must tend to zero more rapidly than \( |n|^{-k} \). Conversely, if \( |b_n| \leq C|n|^{-k} \), the series (4) can be differentiated at least \( k - 1 \) times (the \( k - 1 \)th derivative may exist only ‘almost everywhere’). In short: The order of smoothness of a function is reflected in the rate of decay of its Fourier coefficients.
Box 1. The Meaning of Orthonormality

A sequence of complex-valued functions \(\phi_n\) is said to be orthonormal on an interval \(I\) (which might be the whole real line) if

\[
\int_I \phi_n(t)\phi_m(t) \, dt = 0 \quad \text{for } n \neq m \quad \text{(orthogonality), (a)}
\]
\[
\int_I |\phi_n(t)|^2 \, dt = 1 \quad \text{(normalization). (b)}
\]

(The terminology comes from the analogy with an orthonormal or unit-perpendicular set of vectors \(v_n\); the integrals \(\int_I \phi_n \phi_m\) replace the dot products \(v_n \cdot v_m\).) For example, it is an easy exercise to check that the functions \(e^{2\pi i nt}\) used in Fourier analysis are orthonormal on any interval of length 1. If the functions \(\phi_n\) are orthonormal and \(f\) can be expanded in terms of them,

\[
f(t) = \sum c_n \phi_n(t) \quad (t \in I), \quad \text{(c)}
\]

the coefficients \(c_n\) must be given by

\[
c_m = \int_I f(t)\phi_m(t) \, dt. \quad \text{(d)}
\]

To see this, multiply both sides of (c) by \(\overline{\phi_m(t)}\) and integrate over \(I\):

\[
\int_I f(t)\phi_m(t) \, dt = \sum_n c_n \int_I \phi_n(t)\phi_m(t) \, dt.
\]

By (a), the integrals on the right all vanish except the one with \(n = m\), and that one equals 1 by (b), so the sum on the right reduces to \(c_m\). Conversely, if (c) always implies (d), the sequence \(\phi_n\) must be orthonormal. Indeed, the identity

\[
\phi_m = \cdots + 0 \cdot \phi_{m-1} + 1 \cdot \phi_m + 0 \cdot \phi_{m+1} + 0 \cdot \phi_{m+2} + \cdots
\]

is an instance of (c), and the corresponding instance of (d) is just (a) together with (b).

The bad news is that it is hard to tell from a Fourier series where the non-smooth behavior occurs, or what it looks like. For example, the function

\[
g_1(t) = \sum_{n \neq 0} e^{\frac{2\pi i nt}{2\ln}} = \sum_{n=1}^{\infty} \frac{\sin 2\pi nt}{n} \quad \text{(7)}
\]

has jump discontinuities at the points 0, \(\pm 1, \pm 2, \ldots\) but is perfectly smooth (in fact, linear) elsewhere. Its close relative

\[
g_2(t) = \sum_{n \neq 0} (-1)^n e^{\frac{2\pi i nt}{2\ln}} = \sum_{n=1}^{\infty} (-1)^n \frac{\sin 2\pi nt}{n} \quad \text{(8)}
\]

has the same sort of behavior, but with discontinuities at \(\pm \frac{1}{2}, \pm \frac{3}{2}, \pm \frac{5}{2}, \ldots\) And the
Taylor series provide precise local information about a function, but only if the function has many derivatives. Fourier series or integrals can be used for functions with no smoothness properties, and they yield lots of information about the global properties of the function, but they are inefficient for analyzing the detailed behavior of a function near a point. (How could you ever tell where the discontinuities of the functions in (7) and (8) occur by examining their Fourier coefficients?) What is missing is a method for analyzing the local irregular behavior of functions that aren’t smooth – and this is where wavelets come into play.

What is a ‘wavelet’? This term is loosely used to denote a function that exhibits oscillatory behavior in some interval $I$, then decays rapidly to zero (or perhaps vanishes identically) outside $I$. If we have one such function $\psi(t)$, we can obtain others by dilations or translations – that is, by replacing the variable $t$ by $at$ ($a > 0$) or $t - b$ ($-\infty < b < \infty$). The dilation $\psi(t) \rightarrow \psi(at)$ changes the frequency of the oscillations of $\psi$ by a factor of $a$, and simultaneously changes the length of the
interval on which $\psi$ lives by a factor of $1/a$. The translation $\psi(t) \to \psi(t - b)$ simply shifts $\psi$ over by the amount $b$. We can also combine these operations to get $\psi(at - b)$, which is the result of first translating by the amount $b$ and then dilating by the factor $a$. (If we do it in the reverse order, we get something different: $\psi(t) \to \psi(at) \to \psi(a[t - b]) = \psi(at - ab)$.) See Figure 4.

Now, here is the basic problem of wavelet analysis: Can we use the translations and dilations of a wavelet $\psi$ to form a set of ‘basic’ functions for expanding general functions into infinite series as in (1)? Let us make a formal definition: An orthonormal wavelet is a real- or complex-valued function $\psi(t)$ defined on the real line, with the following properties:

1. $\psi(t)$ tends to zero faster than any power of $t$ as $t \to \pm \infty$.
2. $\psi$ possesses continuous derivatives up to order $N$, for some positive integer $N$.
3. For all integers $j$ and $n$, let
   \[ \psi_{jn}(t) = 2^{j/2} \psi(2^j t - n). \] (12)
   Then every function $f$ on the real line that satisfies (11) can be expanded uniquely in a series of the $\psi_{jn}$:
   \[ f(t) = \sum_{j,n=-\infty}^{\infty} c_{jn} \psi_{jn}(t). \] (13)
4. The coefficients $c_{jn}$ in (13) are given by
   \[ c_{jn} = \int_{-\infty}^{\infty} f(t) \overline{\psi_{jn}(t)} \, dt, \] (14)
   where $\overline{\psi_{jn}(t)}$ is the complex conjugate of $\psi_{jn}(t)$.

As explained in Box 1, properties [3] and [4] imply that the functions $\psi_{jn}$ are orthonormal; hence the name orthonormal wavelet. The factor of $2^{j/2}$ in (12) is there to make $\int_{-\infty}^{\infty} |\psi_{jn}(t)|^2 \, dt$ independent of $j$ and $n$.

The existence of orthonormal wavelets is not at all obvious. If we omit condition [2], there is a simple example that has been known since 1910, the Haar wavelet

\[ \psi_{\text{Haar}}(t) = \begin{cases} 1 & \text{if } 0 < t < \frac{1}{2}, \\ -1 & \text{if } \frac{1}{2} < t < 1, \\ 0 & \text{otherwise.} \end{cases} \]

See Figure 5. However, the Haar wavelet is too crude to be very useful. Finding orthonormal wavelets that satisfy [2] is much trickier, and it was not done until 1986–87 when a group of mathematicians led by Yves Meyer took some ideas that had been used in signal processing and other areas of applied mathematics and turned them into a rigorous mathematical theory. Meyer constructed an orthonormal wavelet that has
derivatives of all orders, but its decay at infinity is relatively slow (a computational disadvantage). See Figure 6. Guy Battle and Pierre-Gilles Lemarié constructed a family of orthonormal wavelets, one for each order of smoothness $N$ in [2], with exponential decay at infinity. Finally, by an ingenious adaptation of some algorithms developed by engineers for signal processing, Ingrid Daubechies constructed a family of orthonormal wavelets, with all possible finite orders of smoothness, that actually vanish outside of finite intervals. See Figure 7.

Let us pause to say a few more words about these constructions. Meyer’s method is Fourier-analytic; he gives an explicit formula not for $\psi$ but for its Fourier transform. The Battle-Lemarié wavelets are splines, that is, piecewise polynomial functions. Both these types of functions are well known in mathematical analysis. But Daubechies’s wavelets are something new. They are given not by analytic formulas but as limits of recursive transformations. That is, a Daubechies wavelet $\psi$ is obtained as the limit of a sequence of functions $\phi_k$, where $\phi_0$ is a simple piecewise linear function and $\phi_k$ is defined in terms of its predecessor $\phi_{k-1}$ by a recursion formula of the form

$$\phi_k(t) = \sum_{n=-N}^{N} C_n \phi_{k-1}(2t - n), \quad (15)$$

and the whole secret is to find the appropriate coefficients $C_n$ to make this process work. The resulting limit function $\psi$ is a rather peculiar beast from the standpoint of classical mathematics, but it is lovely from the computational point of view, because the formula (15) involves only simple arithmetic which a computer is quite happy to perform over and over again!

We have suggested that wavelets should have some oscillatory behavior, and the accompanying graphs also indicate this, but there seems to be nothing in properties [1]–[4] about oscillation. In fact, it is not too hard to show that properties [1]–[4] automatically imply the following additional property of wavelets:

[5] Let $N$ be the integer in property [2]. Then

$$\int_{-\infty}^{\infty} \psi_{jn}(t) P(t) \, dt = 0 \quad (16)$$

for all integers $j$ and $n$ and all polynomials $P$ of degree $\leq N$. This property implies that $\psi$ must be quite oscillatory in order to produce the cancellations that make all the integrals in (16) vanish. For example, by taking $P(t) \equiv 1$ (and $j = n = 0$) we see that $\int_{-\infty}^{\infty} \psi(t) \, dt = 0$, which means that positive values of $\psi$ in one region must be balanced by negative values in another.

We now examine the meaning and applications of wavelet expansions of the form (13). To simplify the discussion, we shall assume that we are working with a Daubechies wavelet $\psi$ with $N$ continuous derivatives that vanishes outside a bounded interval $I$. (Almost everything we say remains true for other wavelets if one modifies the statements to take account of the rapidly decaying tails of the wavelet.)
The function $\psi_{jn}$ defined by (12) then vanishes outside the interval $I_{jn}$ obtained by translating $I$ by the amount $n$ and then dilating by the factor $2^{-j}$; thus, if $I$ has length $l$ then $I_{jn}$ has length $2^{-j}l$. At the same time, we may think of the frequency of the oscillations of $\psi$ as being on the order of magnitude of 1. (This is just a manner of speaking, as the oscillations of $\psi$ are irregular and don’t have a definite frequency.) Then the frequency of the oscillations of $\psi_{jn}$ is on the order of magnitude of $2^{-j}$. In short, for $j \ll 0$, $\psi_{jn}$ represents low-frequency oscillations over long intervals, and for $j \gg 0$, $\psi_{jn}$ represents high-frequency oscillations over short intervals.

We can now see how to interpret (13) and (14):

$$f = \sum_{j,n} c_{jn}\psi_{jn} \quad \text{and} \quad c_{jn} = \int_{-\infty}^{\infty} f(t)\overline{\psi_{jn}(t)} \, dt = \int_{I_{jn}} f(t)\overline{\psi_{jn}(t)} \, dt.$$  

Let us rewrite the expansion of $f$ as

$$f = \sum_{j=-\infty}^{\infty} S_j, \quad \text{where} \quad S_j = \sum_{n=-\infty}^{\infty} c_{jn}\psi_{jn}. \quad (17)$$

The sum $\sum_{j<0} S_j$ represents a ‘blurred’ version of $f$ that shows its large features but omits the fine details of size 1 or less. The sum $S_0$ adds in oscillations of frequency roughly 1 on intervals of length roughly 1; the sum $S_1$ adds in oscillations of frequency roughly 2 on intervals of length roughly $1/2$; and so forth. Each additional sum $S_j$ adds another level of detail at the length scale $2^{-j}$.

Moreover, the individual terms in these sums are local in nature. The coefficient $c_{jn}$ depends only on the values of $f$ on the interval $I_{jn}$, and the term $c_{jn}\psi_{jn}$ lives on this same interval. Since these intervals only overlap to a finite extent as $n$ varies (with $j$ fixed), the sum $S_j$ provides an efficient description of the local details of the behavior of $f$ at length scale $2^{-j}$. As $j$ increases, it is like looking at pieces of $f$ under increasingly strong microscopes.

(On the other hand, when $j \ll 0$, $S_j$ might be said to represent a view of $f$ through the wrong end of a telescope. If $f$ is negligibly small outside an interval of length $L$, it is easy to see that $S_j$ will also be negligibly small for $2^{-j} \gg L$, because things on the scale of $L$ are ‘invisible’ to the long-range waves $\psi_{jn}$. Hence the terms in (13) with $j \ll 0$ are not of much importance for most purposes.)

We mentioned earlier that smoothness of a periodic function is reflected in the decay of its Fourier coefficients. The same is true for wavelet expansions, but now on a local level: The smoothness of $f$ in the neighbourhood of a point $t = a$ is reflected in the decay as $j \to +\infty$ of those wavelet coefficients $c_{jn}$ for which the interval $I_{jn}$ contains $a$. The reason for this fact (of which there are several much more precise versions) can be briefly explained as follows. If $f$ has $N$ continuous derivatives near $t = a$, we can use Taylor’s formula (3) to write $f = P_N + R_N$ where $P_k$ is the Taylor polynomial.
Wavelets have been found to be particularly useful in problems that require separating out the parts of a signal that pertain to different scales of time or length, and in processing signals or images whose most significant features lie in the regions where rapid variations occur.

There is considerable evidence that the human brain uses something akin to wavelet analysis in processing the information it receives from the eyes and ears.

of degree $N$ about $t = a$ and $R_N(t)$ is very small for $t$ near $a$. Hence, by property [5],

$$c_{jn} = \int_{I_{jn}} f(t)\bar{\psi}_{jn}(t) \, dt = \int_{I_{jn}} [P_N(t) + R_N(t)]\bar{\psi}_{jn}(t) \, dt = \int_{I_{jn}} R_N(t)\bar{\psi}_{jn}(t) \, dt$$

But if $I_{jn}$ contains $a$ and $j$ is large (so that the length of $I_{jn}$ is small), $I_{jn}$ will be contained in the region where $R_N$ is very small, and hence $c_{jn}$ will be very small!

For example, let us return to the functions $g_1$, $g_2$, and $g_3$ defined by (7), (8), and (9). The Fourier series for $g_1$ and $g_2$ look almost alike, but their wavelet expansions are very different because the sets of indices $(j, n)$ for which $I_{jn}$ contains one of the discontinuities of $g_1$ or $g_2$ are very different. It is only for those values of $j$ and $n$ that the coefficients $c_{jn}$ for $g_1$ or $g_2$ will be significant for $j$ large. (In fact, if $I_{jn}$ does not contain a discontinuity of $g_1$ or $g_2$, $c_{jn}$ will actually be zero by property [5], since $g_1$ or $g_2$ is linear on $I_{jn}$!) On the other hand, $g_3$ is more or less equally rough at all points, so its coefficients $c_{jn}$ will all decay slowly as $j \to +\infty$.

Wavelet expansions have significant applications in a number of areas, particularly in the analysis and processing of electric and acoustic signals and of two-dimensional images such as photographs. (The latter require wavelets in two variables, whose theory is similar to the one-dimensional theory sketched above.) Wavelets have been found to be particularly useful in problems that require separating out the parts of a signal that pertain to different scales of time or length, and in processing signals or images whose most significant features lie in the regions where rapid variations occur. (For example, in an ordinary photograph the most important features may be the edges that indicate the boundaries of the objects in the photograph; in an astronomical photograph it is important to distinguish between distant stars and even more distant galaxies, whose details show up at different length scales.) There is considerable evidence that the human brain uses something akin to wavelet analysis in processing the information it receives from the eyes and ears, and researchers in human vision were using wavelets in an informal way long before the present mathematical theory was developed.

This is not to say that wavelet expansions of the sort we have discussed are good for everything, however. For example, they are not particularly efficient for encoding musical signals because they do not provide enough frequency resolution. (Roughly speaking, the expansion (17) only provides frequency resolution to within the nearest octave, because the frequency scales specified by the index $j$ differ by factors of 2.) However, the discovery of orthonormal wavelets has stimulated the development of a variety of related techniques for expanding functions (some of which go under the names of bi-orthogonal wavelets, wavelet packets, local Fourier bases, etc.) that offer much greater flexibility in adapting the expansion to the problem at hand. For example, one can sacrifice some precision in spatial or temporal resolution to obtain better frequency resolution, or sacrifice the perfect efficiency of the orthogonal wavelet expansion for expansions that have some built-in redundancy and hence are less sensitive to encoding errors. But this is another story.
Suggested Reading


The crow said “Yes, I understand the steps needed to solve a problem”.

How to Count – An Exposition of Polya’s Theory of Enumeration

Shriya Anand

Consider the following problem: Mr. A has a cube whose faces he wants to paint either red or green. He wants to know how many such distinct coloured cubes he can make.

Now, since the cube has 6 faces, and he has 2 colours to choose from, the total number of possible coloured cubes is $2^6$. But, painting the top face red and all the other faces green produces the same result (aesthetically speaking), as painting the bottom face red and all the other faces green. That is why Mr. A is so confused!

Trial and error is not the best way to solve this problem. We want to find a general method. Consider the set of all possible coloured cubes (in this case, these are $2^6$ in number). The rotational symmetries transform the cube and, evidently, we would consider two colouring patterns to be different only if either cannot be obtained from the other by a rotation. In fact, we consider two coloured cubes to be equivalent precisely if a rotation is all that distinguishes them. To find the various possible colour patterns which are inequivalent, we shall exploit the fact that the rotational symmetries of the cube have the structure of a group.

Let us explain the above in precise terms. Let $D$ denote a set of objects to be coloured (in our case, the 6 faces of the cube) and $R$ denote the range of colours (in the above case {red, green}). By a colouring of $D$, one means a mapping $\phi : D \rightarrow R$.

Let $X$ be the set of colourings. If $G$ denotes a group of permutations of $D$, we can define a relation on the set of colourings as follows:

$\phi_1 \sim \phi_2$ if, and only if, there exists some $g \in G$ such that $\phi_1 g = \phi_2$.

By using the fact that $G$ is a group, it is easy to prove that $\sim$ is an equivalence relation on $X$, and so it partitions $X$ into disjoint equivalence classes.

Now for each $g \in G$, consider the map $\pi_g : X \rightarrow X$ defined as $\pi_g(\phi) = \phi g^{-1}$; it is a bijection from $X$ to itself. In other words, for each $g \in G$, we have $\pi_g \in \text{Sym } X$, where $\text{Sym } X = \text{the group of all permutations on } X$.

Let us define $f : G \rightarrow \text{Sym } X$ as $f(g) = \pi_g$.

Now,

$\pi_{g_1g_2}(\phi) = (g_1g_2)^{-1} = \phi g_2^{-1} g_1^{-1} = \pi_{g_1}(\phi g_2^{-1}) = \pi_{g_1}(\pi_{g_2}(\phi)) = \pi_{g_1}\pi_{g_2}(\phi)$.

Keywords
Rotational symmetry, Polya’s theorem, Burnside’s lemma, cycle index, isomers.

We consider two coloured cubes to be equivalent precisely if a rotation is all that distinguishes them. To find the various possible colour patterns which are inequivalent, we shall exploit the fact that the rotational symmetries of the cube have the structure of a group.

Therefore \( f \) is a homomorphism from \( G \) to the group of permutations on \( X \) i.e., \( G \) can be regarded as a group of permutations of \( X \).

Recall that one says that a group \( G \) acts on a set \( X \) if there is a homomorphism from \( G \) to the group of all permutations of the set \( X \). It is clear that the orbits of the action described above are precisely the different colour patterns i.e., the equivalence classes under \( \sim \). Therefore, we need to find the number of inequivalent colourings, i.e. the number of equivalence classes of \( \sim \), i.e. the number of orbits of the action of \( G \) on \( X \). Note that, like in the example of the cube we shall consider only finite sets \( D, R \).

The answer will be provided by a famous theorem of Polya. Polya’s theorem was published first in a paper of J H Redőeld in 1927 and, apparently no one understood this paper until it was explained by F Harary in 1960. Polya’s theorem is considered one of the most significant papers in 20th century mathematics. The article contained one theorem and 100 pages of applications. Before stating this theorem we will recall what has come to be generally known as Burnside’s lemma and which will be needed in the proof. Apparently, it is due to Cauchy but attributed to Burnside by Frobenius (see [2] for this bit of history).

**Burnside’s Lemma**

Let \( G \) be a group of permutations of a set \( X \). Then, number of orbits = \( \frac{1}{|G|} \sum_{x \in G} |X^g| \), where \( X^g = \{ x \in X | x \cdot g = x \} \), the set of points of \( X \) fixed under \( g \).

**Proof.**

Consider the subset \( S \) of \( X \times G \) consisting of elements \( (x, g) \) such that \( x \cdot g = x \). Then, \( |S| = \sum_{g \in G} |X^g| \) as is apparent from counting over the various \( x \)’s corresponding to a particular \( g \) and then summing over the \( g \)’s. Also, counting over the \( g \)’s corresponding to a particular \( x \) and then summing over \( x \) gives us \( |S| = \sum_{x \in X} |G_x| \), where \( G_x = \{ g \in G | g \cdot x = x \} \), the so-called stabiliser of \( x \). Note that each \( G_x \) is a subgroup. Let the orbits in \( X \) be \( X_1, X_2, \ldots X_k \). But, the stabilizers of elements in the same orbit have the same cardinality as they are conjugate subgroups.

Therefore \( |S| = \sum_{i=1}^{k} \sum_{x \in X_i} |G_{x_i}| \).

The assertion on stabilisers holds because, if \( y = xg \), then

\[
G_y = \{ h \in G : yh = y \} = \{ h \in G : xgh = xg \} = \{ h \in G : xghg^{-1} = x \} = \{ h \in G : ghg^{-1} \in G_x \} = g^{-1}G_xg.
\]
Equating the two expressions for $S$, we get

$$k = \frac{1}{|G|} \sum_{g \in G} |X^g|.$$ 

To use this lemma for permutation groups, we need the notion of a cycle index. First, recall that any permutation $\sigma$ in $S_n$ has a disjoint cycle decomposition viz.,

$$\sigma = (i_1, \cdots, i_{r_1}) \cdots (i_{r_d}, \cdots, i_{r_{d+1}}),$$

where the cycles have no index common.

For instance, in $S_6$, the permutation which interchanges 1 and 3 and interchanges 2 and 4 can be written as $(1, 3)(2, 4)(5)(6)$.

**Definition: The Cycle Index**

Let $G$ be a group of permutations on a set of $n$ elements. Let $s_1, s_2, \ldots, s_n$ be variables. For $g \in G$, let $\lambda_i(g)$ denote the number of $i$-cycles in the disjoint cycle decomposition of $g$. Then, the cycle index of $G$, denoted by \( z(G; s_1, s_2, \ldots, s_n) \) is defined as the polynomial expression

$$z(G; s_1, s_2, \ldots, s_n) = \frac{1}{|G|} \sum_{g \in G} s_1^{\lambda_1(g)} s_2^{\lambda_2(g)} \cdots s_n^{\lambda_n(g)}$$

**Examples.**

1. $G = \{e, (1 \ 2), (3 \ 4), (1 \ 2)(3 \ 4)\}$. Then,

$$z(G; s_1, s_2, s_3, s_4) = \frac{s_1^4 + 2s_1^2s_2 + s_2^2}{4}$$

2. $G = S_3 = \{e, (1 \ 2), (1 \ 3), (2 \ 3)(1 \ 2 \ 3), (1 \ 3 \ 2)\}$

$$z(G; s_1, s_2, s_3) = \frac{1}{6} (s_1^3 + 3s_1s_2 + 2s_3)$$

In fact,

$$z(S_n; s_1, s_2, \ldots, s_n) = \sum_{\lambda_1+2\lambda_2+\cdots+k\lambda_k=n} \frac{s_1^{\lambda_1}s_2^{\lambda_2} \cdots s_k^{\lambda_k}}{\lambda_1!\lambda_2!\cdots\lambda_k!}$$

To see this, look at the number of permutations in $S_n$ of the type $(\lambda_1, \lambda_2, \cdots, \lambda_k)$.

The $i$-cycles can be arranged amongst themselves in $\lambda_i!$ ways giving rise to the same permutation. Also, in each $i$-cycle, one can write any one of the $i$ symbols first and, therefore, we must also divide by $i^{\lambda_i}$.

3. In our example of the cube, $G$ is the group of rotations of a cube induced on the set of 6 faces. The rotations of the cube which leave it invariant are (see Figure 1):

\[\text{Figure 1. Cube and its rotations}\]
(i) 90 degree (clockwise or anti-clockwise) rotations about the axes joining the centres of the opposite faces – there are 6 such;

(ii) 180 degree rotations about each of the above axes – there are 3 such;

(iii) 120 degree (clockwise or anti-clockwise) rotations about the axes joining the opposite vertices – there are 8 such;

(iv) 180 degree rotations about the axes joining the midpoints of the opposite edges and;

(v) the identity.

The permutations of the 6 faces induced by these rotations are as follows.

The rotations of type (i) are permutations of the form \((1, 2, 3, 4)\) etc. where we have numbered the faces from 1 to 6. The 6 permutations of this type give the term \(6s_1^2 s_4\) in the cycle index of \(G\).

Similarly, the types (ii),(iii),(iv) and (v) give the terms \(3s_1^2 s_2^2, 8s_3^2, 6s_2^3\) and \(s_6^6\), respectively. Therefore, the cycle index of \(G\) is

\[
z(G; s_1, \cdots, s_6) = \frac{1}{24} (6s_1^2 s_4 + 3s_1^2 s_2^2 + 8s_3^2 + 6s_2^3 + s_6^6).
\]

4. Let \(G = C_n = \text{cyclic group of order } n\).

The cyclic group \(C_n\) is regarded as the group of permutations of the vertices of a regular \(n\)-gon. That is, it is the subgroup of \(S_n\) generated by an \(n\)-cycle \((1, 2, \cdots, n)\). Note that, for a generator \(g\) of \(S_n\), the element \(g^i\) has the same cycle structure as that of \(g^{(i,n)}\). Therefore, the cycle index is

\[
z(C_n; s_1, s_2, \ldots, s_n) = \frac{1}{n} \sum_{d|n} \phi \left( \frac{n}{d} \right) s_d^{n/d} = \frac{1}{n} \sum_{d|n} \phi(d) s_d^{n/d}.
\]

Here, \(\phi\) is Euler’s totient function defined by \(\phi(n)\) being the number of \(m\) up to \(n\) which are coprime to \(n\).

5. For \(n > 2\), the dihedral group \(D_n\) is defined as the group of rotations of the regular \(n\)-gon given by \(n\) rotations about an \(n\)-fold axis perpendicular to the plane of the \(n\)-gon and reflections about the \(n\) two-fold axes in the plane of the \(n\)-gon like the spokes of a wheel, where the angle between consecutive spokes is \(\frac{2\pi}{n}\) or \(\frac{\pi}{n}\) according as \(n\) is odd or even. It has order \(2n\).

It can be regarded as a subgroup of \(S_n\) as follows. The \(n\) rotations corresponding to the powers of \(\sigma = (1, 2, \cdots, n)\) and the group \(D_n\) is the subgroup

\[
\{I_d, \sigma, \cdots, \sigma^{n-1}, \tau, \tau \sigma, \cdots, \tau \sigma^{n-1}\},
\]

where \(\tau = (2, n)(3, n - 1)\cdots\) The cycle index of \(D_n\) is

\[
z(D_n; s_1, \cdots, s_n) = \frac{1}{2n} \sum_{d|n} \phi \left( \frac{n}{d} \right) s_d^{n/d} + \frac{1}{4} \left( s_1^2 s_2^{n-1} + s_2^{\frac{n}{2}} \right).
\]
if \( n \) is even and
\[
z(D_n; s_1, \ldots, s_n) = \frac{1}{2n} \sum_{d|n} \phi\left(\frac{n}{d}\right) s_{n/d}^d + \frac{1}{2} s_{s_1}^{n/2} s_2^{n/2}
\]
if \( n \) is odd.

So, the dihedral group \( D_6 \) is the symmetry group of the hexagon. One can represent it as the subgroup of \( S_6 \) generated by \((1, 6)(2, 5)(3, 4)\) and \((1, 2, 3, 4, 5, 6)\). Thus,
\[
z(D_6) = \frac{1}{12} (s_1^6 + 3s_1^2s_2^2 + 4s_3^2 + 2s_3^2 + 2s_6).
\]

We shall need this later.

6. The cycle index of the group of symmetries of the vertices of the regular octahedron can be obtained as for the cube and, is
\[
\frac{1}{24} (6s_1^2s_4 + 3s_1^2s_2^2 + 8s_3^2 + 6s_3^2 + s_1^6).
\]

Note that this is the same as that of the group of symmetries of the faces of the cube.

Now, we are in a position to state Polya's theorem.

**Polya’s Theorem**

*Suppose \( D \) is a set of \( m \) objects to be coloured using a range \( R \) of \( k \) colours. Let \( G \) be the group of symmetries of \( D \). Then, the number of colour patterns = \( \frac{1}{|G|} z(G; k, k, \ldots, k) \).*

**Proof.**

As explained before, \( G \) acts on \( X \), the set of all possible colourings. Clearly, \( |X| = k^m \). The number of colour patterns is computed using Burnside’s lemma as the number of orbits of this action. This equals
\[
\frac{1}{|G|} \sum_{g \in G} |X^g| \text{ where } X^g = \{ \phi \in X | \phi g = \phi \}.
\]

So, now we need to find the number of colourings fixed by \( g \). But, a colouring is fixed by \( g \) precisely when it is fixed by all the cycles in the disjoint cycle representation of \( g \). Therefore, number of colourings fixed by \( g \) equals
\[
k^{\lambda_1(g)} k^{\lambda_2(g)} \ldots k^{\lambda_m(g)}.
\]

This is evidently equal to \( k^{n(g)} \) where \( n(g) \) is the number of cycles in \( g \).

Therefore, the number of patterns
\[
\frac{1}{|G|} \sum_{g \in G} k^{n(g)} = \frac{1}{|G|} z(G; k, k, \ldots, k).
\]
This proves the theorem.

So, in our example of the cube, the number of distinct coloured cubes

\[
\frac{1}{24} [2^6 + 6 \cdot 2^3 + 8 \cdot 2^2 + 3 \cdot 2^2 \cdot 2^2 + 6 \cdot 2^2 \cdot 2]
\]

\[
= \frac{1}{24} \times 240 = 10.
\]

There are 10 distinct cubes in all.

Now, our problem of equivalence of colourings has been disposed of. But, a second problem often encountered in counting is that sometimes not all objects are counted with same weight. So, for instance, if Mr. A did not merely wish to know how many cubes he could paint, but how many would have precisely 2 red faces and 4 green faces, then the above is not good enough. So we will proceed to state and explain a more general form of Polya’s theorem which can handle both the above problems.

For that, we will make use of the following concepts: Consider all maps from \(D\) to \(R\) as before. But, now each \(r \in R\) has a weight \(w(r)\) attached to it. The \(w(r)’s\) can be thought of as independent variables and polynomial expressions in them with \(Q\)-coefficients can be manipulated formally like polynomials. (In other words, they are elements from a commutative algebra over \(Q\)). The weight of a colouring \(\phi : D \rightarrow R\) is defined as

\[
w(\phi) = \prod_{d \in D} w(\phi(d)).
\]

\[\sum w(r)\] is called the inventory of \(R\) and \(\{\sum w(\phi) : \phi \in X\}\) is called the inventory of \(X\).

Now, we notice some useful facts about weights, viz.:

**Proposition.**

(i) \(\sum_{\phi \in \delta} w(\phi) = \left[\sum w(r)\right]^{\lvert D \rvert}\)

(ii) If \(D_1, D_2, \ldots, D_k\) partition \(D\) and,

\[\text{if } S = \{\phi \in X \mid \phi(d) = \text{constant } \forall d \in D_i, \forall i\},\]

then, \(\sum_{\phi \in S} w(\phi) = \prod_{i=1}^{k} \sum_{r \in R} w(r)^{\lvert D_i \rvert}\).

**Proof.**

(i) Let \(D = \{d_1, d_2, \ldots, d_n\}\), and \(R = \{r_1, r_2, \ldots, r_m\}\).

Then, the right hand side is \((w(r_1) + w(r_2) + \ldots + w(r_m))^n\).

Any term here is of the form \(w(r_{i_1})w(r_{i_2}) \ldots w(r_{i_n})\). This is equal to \(w(\phi)\) for that map \(\phi\) which takes \(d_1\) to \(r_{i_1}\), \(d_2\) to \(r_{i_2}\), and so on. Conversely, any \(w(\phi)\) from the left side is of the form \(w(r_{j_1})w(r_{j_2}) \ldots w(r_{j_n})\) which gives a unique term of the right side. This proves (i).

We prove (ii) now. A term of the right hand side has the form \(w(r_{i_1})^{\lvert D_1 \rvert}w(r_{i_2})^{\lvert D_2 \rvert} \ldots\)
which is precisely the weight of a function which assumes the value \( r_1 \) on \( D_1 \), \( r_2 \) on \( D_2 \) and so on. Conversely, every function has such a weight and the result follows.

Along with these concepts, we will also use the following generalisation of Burnside’s lemma known as the **weighted form of Burnside’s lemma**:

**Suppose** \( G \) **is a finite group acting on a finite set** \( S \). **Let us write** \( s_1 \approx s_2 \) **if, and only if,** \( \exists g \in G \) **such that** \( s_1 \cdot g = s_2 \). **Let** \( \xi \) **be the set of classes of** \( S \). **Let** \( \bar{S} \) **denote the weight of any element in the equivalence class** \( \bar{s} \). **Then,**

\[
\sum_{\bar{s} \in \xi} w(\bar{s}) = \frac{1}{|G|} \sum_{g \in G} \sum_{s \in S^g} w(s).
\]

Note that, by putting \( w(s) = 1 \forall s \in S \) we get the statement of the earlier form of Burnside’s lemma.

**Proof.**

The proof is very similar to the proof of Burnside’s lemma when we consider the subset \( Y \) of \( S \times G \), consisting of elements \((s, g)\) such that \( s \cdot g = s \). Instead of finding the cardinality of \( Y \), we find \( \sum_{(g, s) \in Y} w(s) \) proceeding in the same way as the earlier proof and, the asserted result follows.

Our next aim is to obtain a weighted version of Polya’s theorem. Again, suppose \( D \) is a finite set of objects to be coloured using a finite range \( R \) of colours. As before, let \( X = \{\phi : D \rightarrow R\} \) be the set of all colourings. Then, the group \( G \) of permutations of \( D \) acts on \( X \) in the same way as explained before. Suppose now that each \( r \in R \) is given a weight \( w(r) \) with the property that equivalent colourings have the same weight (here, as before, the weight of any colouring \( \phi \) is \( w(\phi) = \prod_{d \in D} w(\phi(d)) \)). Let us write \( w(\Phi) \) for the weight of any colouring belonging to a particular pattern \( \Phi \).

Then, the weighted version of Polya’s theorem is:

**Polya’s Theorem (weighted form)**

The inventory of patterns is given by

\[
\sum_{\Phi} w(\Phi) = \xi \left( G; \sum w(r), \sum (w(r))^2, \ldots \right).
\]

**Proof.**

Using the weighted form of Burnside’s lemma, we get

\[
\sum_{\Phi} w(\Phi) = \frac{1}{|G|} \sum_{g \in G} \sum_{\phi \in X_g^g} w(\phi), \quad \text{where} \quad X_g^g = \{\phi \in X|\phi g = \phi\}.
\]
So, we need to find $\phi$’s that are fixed by $g$. Let $g$ split $D$ into cycles $D_1, D_2, \ldots, D_n$. These are clearly disjoint and partition $D$. An element $g$ fixes $\phi$ precisely if all the cycles of $g$ fix $\phi$, i.e., $\phi(d)$ is constant $\forall d \in D_i, \forall i = 1, 2, \ldots, n$. Therefore, $X^g, D_1, D_2, \ldots, D_n$ satisfy the conditions of proposition (ii). In fact, we note that

$$|D_i| = 1 \text{ for } 1 \leq i \leq \lambda_1;$$

$$|D_i| = 2 \text{ for } \lambda_1 < i \leq \lambda_1 + \lambda_2, \text{ etc.}$$

By proposition (ii)

$$\sum_{\phi \in X^g} w(\phi) = \prod_{i=1}^{n} \sum_{r \in R} w(r)^{|D_i|} = \left( \sum w(r) \right)^{\lambda_1(g)} \left( \sum (w(r))^2 \right)^{\lambda_2} \ldots \left( \sum (w(r))^n \right)^{\lambda_n}$$

Therefore

$$\sum_{\Phi} w(\Phi) = \frac{1}{|G|} \sum_{g \in G} \prod_{i=1}^{n} \left[ \sum_{r \in R} (w(r))^i \right]^{\lambda_i(g)} = \frac{1}{|G|} \zeta(G; \sum w(r), \sum (w(r))^2, \ldots)$$

which completes the proof.

To illustrate the above, let us come back to the same example of the cubes. Let weight (red) = $r$, w (green) = $g$. Then,

$$\sum w(r) = r + g$$

$$\sum (w(r))^2 = r^2 + g^2$$

$$\vdots$$

$$\sum (w(r))^k = r^k + g^k.$$ 

Also, we saw in example 3 that

$$z(G; s_1, s_2, \ldots, s_6) = \frac{1}{24} [s_1^6 + 6s_1^2s_2^2 + 8s_1^2s_3^2 + 3s_1^2s_2^2 + 6s_1^2s_4]$$

Using Polya’s theorem, $\sum w(\Phi)$ is

$$\frac{1}{24} [(r + g)^6 + 6(r^2 + g^2)^3 + 8(r^3 + g^3)^2 + 3(r + g)^2(r^2 + g^2)^2 + 6(r + g)^2(r^4 + g^4)] = r^6 + r^5g + 2r^4g^2 + 2r^3g^3 + 2r^2g^4 + rg^5 + g^6.$$ 

So, from the above inventory of patterns, it is easy to see that there are exactly 2 patterns with precisely 2 red faces and 4 green faces (the coefficient of $r^2g^4$).

We also note that on putting $r = 1 = g$, we get 10, i.e. the total number of patterns. Thus, in the weighted form of Polya’s theorem, by putting $w(r) = 1 \forall r \in R$, we get the total number of patterns.
Another example: How many distinct circular necklace patterns are possible with \( n \) beads, these being available in \( k \) different colours?

So, we need to find out how many of the \( k^n \) possible necklaces are distinct. Clearly, the group \( G \) of rotational symmetries here is \( C_n \), the cyclic group of order \( n \).

We have already computed

\[
z(C_n, s_1, s_2, \ldots, s_n) = \frac{1}{n} \sum_{d|n} \phi(d) s_d^{n/d}.
\]

Special case: \( n \) is prime.

Then, number of patterns = \( k + \frac{k^n-k}{n} \).

Let us consider the case when only white and black beads are allowed (i.e. \( k = 2 \)) and \( n \) is prime, say \( n = 5 \).

\[
\sum w(\phi) = z(C_5; w + b, w^2 + b^2, \ldots, w^5 + b^5) = \frac{1}{5} [(w + b)^5 + (w^5 + b^5)] = w^5 + w^4b + 2w^3b^2 + 2w^2b^3 + wb^4 + b^5.
\]

In fact, these patterns are shown in the Figure 2 here.

Yet another very useful form of Polya’s theorem uses the concept of ‘content’. Here, it is convenient to think of \( R \) not as a range of colours but of ‘figures’. Maps from \( D \) to \( R \) are called configurations. Especially, this is useful in counting isomers of chemical compounds as we shall see. Every figure in \( R \) has a ‘content’ which is a non-negative integer. The figure counting series is

\[
c(x) = c_0 + c_1x + c_2x^2 + \ldots + c_kx^k + \ldots
\]

where \( c_k \) is the number of figures in \( R \) with content \( k \).

Content of a configuration is the sum of contents of figures which occur as images (taking into account the multiplicity) i.e., content of \( \phi \) equals \( \sum_{d|D} c(\phi(d)) \).

So, if we introduce some equivalence of maps by the action of a group \( G \) on \( D \), (which induces, therefore, an action of \( G \) on configurations), then equivalent configurations have the same content.

The generating function to count all configurations is defined as the formal power series

\[
F(x) = 1 + F_1x + F_2x^2 + \ldots + F_kx^k + \ldots,
\]

where \( F_k \) = number of configurations with content \( k \).
Now, let a figure \( r \in R \) with content \( k \) be considered as having weight \( x^k \).

\[
\sum_r w(r) = c_0 + c_1 x + c_2 x^2 + \ldots = c(x)
\]
\[
\sum_r [w(r)]^2 = c_0 + c_1 x^2 + c_2 x^4 + \ldots = c(x^2)
\]
\[
\vdots
\]
\[
\sum_r [w(r)]^k = c_0 + c_1 x^k + c_2 x^{2k} + \ldots = c(x^k).
\]

Then, given \( \phi : D \rightarrow R \),

\[
w(\phi) = \prod_{d \in D} w(\phi(d)) = \prod_{d \in D} x^{\text{content of } \phi(d)} = \sum_{x} \text{content of } \phi\text{'s}(d) = \sum_{x} \text{content of } \phi.
\]

Therefore the configuration counting series is

\[
F(x) = 1 + F_1 x + F_2 x^2 + \ldots + F_k x^k + \ldots = \sum_{\phi} w(\phi).
\]

We call \( F(x) \), the inventory of all configurations.

Therefore a group \( G \) acts on that and, hence on the set of configurations, the inventory of inequivalent configurations,

\[
\sum w(\Phi) = 1 + \Phi_1 x + \Phi_2 x^2 + \ldots + \Phi_k x^k + \ldots
\]

where \( \Phi_k \) = number of inequivalent configurations with content \( k \).

\[
\sum w(\Phi) = z(G; \sum w(r), \sum [w(r)]^2, \ldots) = z(G; c(x), c(x^2), \ldots)
\]

In other words, we have proved the **‘content’ version of Polyá’s theorem**:

*Let there be a permutation group \( G \) acting on the domain \( D \), and hence, on the set of maps into the set \( R \) of ‘figures’. Let the figure counting series be \( c(x) \). Then, the inequivalent configuration counting series \( \Phi(x) \) is obtained by substituting \( c(x^r) \) for \( s_r \) in the cycle index of \( G \) i.e., \( \Phi(x) = z(G; c(x)) \).*

**Examples.**

One of the important applications of the content version of Polyá’s theorem is the finding of different possible isomers of a chemical compound. Recall that isomers are
chemical compounds with the same chemical formula with a different arrangement of the atoms.

1. Let us find the number of benzene rings with Cl substituted in the place of H.

The symmetry group of the benzene ring is $D_6$ (i.e., the symmetries of a regular hexagon).

Now, $z(D_6) = \frac{1}{12} [s_1^6 + 4s_2^3 + 2s_3^2 + 3s_1^2s_2^2 + 2s_6]$. Here, $D = \{1, 2, 3, 4, 5, 6\}$ $R= \{H, Cl\}$.

Let content (H)=0, content (Cl)=1. So $c(x) = 1 + x$.

$$\sum \Phi(x) = \frac{1}{12} \left[ (1 + x)^6 + 4(1 + x^2)^3 + 2(1 + x^3)^2 + 3(1 + x)^2(1 + x^2)^2 + 2(1 + x^6) \right]$$

$$= 1 + x + 3x^2 + 3x^3 + 3x^4 + x^5 + x^6.$$

Therefore there are 13 chemical compounds obtained in this manner (see the Figure 3).

2. Similarly, there are two isomers of the octahedral molecule PtBr₄Cl₂ with Pt at the centre and Br and Cl at the vertices. This is proved by using the cycle index in example 6, of the vertices of a regular octahedron.

3. To find the number of (simple, undirected) graphs upto isomorphism on a set of $n$ vertices.

![Figure 3. Isomers of benzene with Cl.](image-url)
Here $D$ is the set of $\binom{n}{2}$ pairs of vertices. If there is an edge between a pair of vertices, let it have content 1, if it has no edge, then let it have content 0.

So, a configuration is a graph, and its content is the number of edges.

Then, $c(x) = 1 + x$.

Now, two labelled graphs are isomorphic if there exists a bijection between the vertices that preserves adjacency (therefore, two graphs are equivalent if they are isomorphic). Now, the group $G$ of symmetries of the set of pairs of vertices is called $S_n^{(2)}$ and its cycle index is

$$z(S_n^{(2)}) = \frac{1}{24}[s_6^3 + 9s_1^2s_2^2 + 8s_3^2 + 6s_2s_4].$$

Counting series for such graphs

$$= \frac{1}{24}[(1 + x)^6 + 9(1 + x)^2(1 + x^2)^2 + 8(1 + x^3)^2 + 6(1 + x^2)(1 + x^4)]$$
$$= 1 + x + 2x^2 + 3x^3 + 2x^4 + x^5 + x^6.$$

Therefore total number of unlabelled graphs on 4 vertices = 11. These are shown in the Figure 4.

For a wealth of information on Polya’s theory, the interested reader is referred to [1] and [3].

**Suggested Reading**


The ancient unsolved problem of congruent numbers has been reduced to one of
the major questions of contemporary arithmetic: the finiteness of the number
of curves over $\mathbb{Q}$ which become isomorphic at every place to a given curve. We
give an elementary introduction to congruent numbers and their conjectural
characterisation, discuss local-to-global issues leading to the finiteness problem,
and list a few results and conjectures in the arithmetic theory of elliptic curves.

The area $\alpha$ of a right triangle with sides $a, b, c$ (so that $a^2 + b^2 = c^2$) is given by
$2\alpha = ab$. If $a, b, c$ are rational, then so is $\alpha$. Conversely, which rational numbers $\alpha$
arise as the area of a rational right triangle $a, b, c$? This problem of characterising
“congruent numbers” – areas of rational right triangles – is perhaps the oldest unsolved
problem in all of Mathematics. It dates back to more than a thousand years and has
been variously attributed to the Arabs, the Chinese, and the Indians.

Three excellent accounts of the problem are available on the Web: Right triangles and
elliptic curves by Karl Rubin, Le problème des nombres congruents by Pierre Colmez,
which also appears in the October 2006 issue of the Gazette des mathématiciens, and
Franz Lemmermeyer’s translation Congruent numbers, elliptic curves, and modular
forms of an article in French by Guy Henniart. A more elementary introduction is
provided by the notes of a lecture in Hong Kong by John Coates, which have appeared
in the August 2005 issue of the Quarterly journal of pure and applied mathematics. A
detailed account is to be found in Introduction to elliptic curves and modular forms
(Springer, 1984) by Neal Koblitz. None of these sources goes beyond the theorems
of John Coates and Andrew Wiles [1] (see Theorem 14) and of Jerrold Tunnell [2]
(see Theorem 25).

In 1991, Rubin [3] (see Theorem 15) reduced the congruent number problem to a
natural finiteness question in the arithmetic of elliptic curves (with “complex multi-
plications”). An excellent survey of such finiteness questions can be found in Barry
Mazur’s article [4].

This article consist of three parts of quite different nature. The first part is an
elementary presentation of the problem of congruent numbers ($\S 1$) and its conjectural
solution ($\S 2$); the material here is borrowed from the accounts which have been cited.
The second part introduces local number fields ($\S 3$) and discusses the local-to-global
principle – its validity in the case of conics ($\S 4$) and its failure in the case of cubics
($\S 5$) – in a language which can be understood by bright undergraduates. The last


Keywords
Congruent numbers, elliptic curves, Hasse principle, Birch
and Swinnerton-Dyer conjecture, Shafarevich-Tate conjecture.
part, which requires greater mathematical maturity, is a catalogue of results – some old, some new – and conjectures in the arithmetic theory of elliptic curves in general (§6) and those without complex multiplications in particular (§7); it ends with a word about the role played by modular forms (§8).

1. Congruent Numbers

If a rational number $\alpha$ is the area of a right triangle with rational sides, then so is $\alpha\beta^2$ for every rational $\beta \in \mathbb{Q}^\times$. Indeed, if $\alpha$ is the area of a rational right triangle with sides $a, b, c$, then $\alpha\beta^2$ is the area of the rational right triangle with sides $a|\beta|, b|\beta|, c|\beta|$. So, up to replacing $\alpha$ by $\alpha\beta^2$ for a suitable $\beta$, we may assume that $\alpha$ is an integer, and moreover that $\alpha$ is not divisible by the square of any prime number. In other words, we assume that $\alpha$ is a positive squarefree integer.

**Definition 1.** A squarefree integer $\alpha > 0$ is said to be a congruent number if there exist $a, b, c \in \mathbb{Q}$ such that $a^2 + b^2 = c^2$ and $ab = 2\alpha$.

The terminology is classical and comes from the fact that $\alpha$ is congruent if and only if it is the common difference (congruum, in Latin) of a three-term arithmetic progression of rational squares. For if $\alpha$ is the area of a rational right triangle with sides $a < b < c$, then, putting $d = (c/2)^2$, the arithmetic progression $d - \alpha, d, d + \alpha$ consists of rational squares. Conversely, if there is a rational number $d$ such that $d - \alpha, d, d + \alpha$ are all three squares, then $\alpha$ is the area of the rational right triangle with sides $\sqrt{d + \alpha} - \sqrt{d - \alpha}, \sqrt{d + \alpha} + \sqrt{d - \alpha}$ and $2\sqrt{d}$.

The problem we address is that of deciding which numbers are congruent: we are asking for an easily-checked criterion which would tell us whether a given number is congruent or not.

Let us first study the single equation $a^2 + b^2 = c^2$ in strictly positive rational numbers; such a triple $(a, b, c)$ will be called a rational solution for short. Two rational solutions $(a, b, c), (a', b', c')$ of this equation are called equivalent if $a = \lambda a', b = \lambda b', c = \lambda c'$ for some $\lambda \in \mathbb{Q}^\times$ (necessarily positive). A rational solution is called primitive if $a, b, c \in \mathbb{Z}$, and if they have no common prime divisor. Every rational solution is equivalent to a primitive one, and no two primitive solutions are equivalent.

Reducing a primitive solution modulo 4, we see that precisely one of $a, b$ is even.

**Proposition 2.** Let $(a, b, c)$ be a primitive solution of $a^2 + b^2 = c^2$, with $a = 2t$ even (and $b, c$ odd). Then there exist integers $m > n > 0$, $\gcd(m, n) = 1$, $m \not\equiv n \pmod{2}$, such that

$$a = 2mn, \quad b = m^2 - n^2, \quad c = m^2 + n^2. \quad (1)$$

Conversely, every such pair $m, n$ of integers ($m > n > 0$, mutually prime, not of the same parity) gives rise via (1) to a primitive triangle $a, b, c$ with $a$ even (and $b, c$ odd).
Proof: As $b$ is odd, so is $c$. Hence $c + b$ and $c - b$ are even; write $c + b = 2u$ and $c - b = 2v$. If a number divides both $u$ and $v$, it would divide their sum $u + v = c$ and their difference $u - v = b$. But $\gcd(b, c) = 1$, so we have $\gcd(u, v) = 1$. The relation $a^2 + b^2 = c^2$ implies that $t^2 = uv$, which shows that each of $u, v$ must be a square. Let $m > n > 0$ be such that $u = m^2, v = n^2$; clearly, $\gcd(m, n) = 1$. Also, $m \neq n \pmod{2}$ because $b = m^2 - n^2$ is odd.

Conversely, we have to show that the triangle $a, b, c$ obtained from such a pair $m, n$ is primitive. Now, 2 divides neither $b$ nor $c$; if a prime $p \neq 2$ divides both $b$ and $c$, it would divide $c + b$ and $c - b$, hence $u$ and $v$, and hence $m$ and $n$.

Let $C$ be a projective conic with a rational point $O$, for example the one defined by $a^2 + b^2 = c^2$, with $O = (1 : 0 : 1)$. Denoting by $D$ the projective line of lines through $O$, the morphism $f$ which sends a point $P \in C$ to the line $f(P) \in D$ passing through $O$ and $P$ – the tangent to $C$ at $O$ if $P = O$ – is an isomorphism.

This result allows us to generate a list which will eventually contain any given congruent number: it suffices to go through the list of all such pairs $(m, n)$, compute the area $mn(m^2 - n^2)$ of the triangle (1), and take the ‘squarefree part’. Thus the pair $(2, 1)$ shows that the number $6 = 2 \cdot 3 \cdot (2^2 - 1^2)$ is congruent.

Retaining only the squarefree parts of the numbers produced by this procedure, the first few congruent numbers which show up are

$$5, 6, 7, 13, 14, 15, 21, 22, 23, 29, 30, 31, 34, 37, 38, 39, 41, \ldots$$  \hspace{1cm} (2)

Note that we have not proved that the numbers 1, 2, 3 are not congruent; it may simply be that they haven’t yet shown up on the list! Indeed, Leonardo of Pisa (called Fibonacci) (1175–1240) was challenged to find a rational right triangle of area 5 (he succeeded) and he conjectured that 1 is not congruent; this was settled much later by Pierre Fermat (1601–1665).

How can we determine if a specific number such as 157 is congruent? The na"ive approach, suggested by the discussion just after Definition 1, would be to go through a ‘list’ of squares $d$ of rational numbers and to see if both $d - 157$ and $d + 157$ are squares. There is indeed such a ‘list’: first we go through the squares of the finitely many rational numbers whose numerator and denominator have just one digit, then through the squares of those – again finitely many – whose numerator and denominator have at most two digits, and so on. It turns out that the first square which works for 157, according to Don Zagier, is

$$d = \left(\frac{22440351770436969924557513090674863160948472041}{2 \times 8912332268928859588025535178967163570016480830}\right)^2.$$

Clearly, this number could not have been found by the na"ive approach; some theory is needed. Also, as before, this approach cannot prove that the given number, for example 1, is not congruent.
Theorem 3 (P Fermat ~ 1640). The number 1 is not congruent.

Proof: We have to show that there is no rational right triangle whose area is a square. If there is such a triangle, we may assume, as before, that its sides are integers not all divisible by any prime number. Fermat’s idea of infinite descent consists in showing that if there were such a ‘primitive’ triangle whose area is a square, then there would be a smaller primitive triangle whose area is also a square. Clearly, this cannot go on for ever.

Let \((a, b, c)\) be a primitive triangle whose area is a square. Assume that \(a\) is even (and \(b, c\) odd). Write \(a = 2mn, b = m^2 - n^2, c = m^2 + n^2\), with \(m > n > 0\), \(\gcd(m, n) = 1\) and \(m \not\equiv n \pmod{2}\) (Proposition 2). As the area \(mn(m + n)(m - n)\) is a square, and as no two of the four factors have a common prime divisor, all four must be squares: \(m = x^2, n = y^2, m + n = u^2, m - n = v^2\). Both \(u, v\) are odd because \(m + n, m - n\) are odd, and \(\gcd(u, v) = 1\). We also have \(u^2 - v^2 = 2y^2\), which we rewrite as

\[
2y^2 = (u + v)(u - v).
\] (3)

As \(u, v\) are odd and \(\gcd(u, v) = 1\), we have \(\gcd(u + v, u - v) = 2\). So one of the two factors on the right in (3) must be of the form \(2r^2\) and the other of the form \(4s^2\). In any case, the sum of their squares is \(16s^4 + 4r^4\). At the same time, \((u + v)^2 + (u - v)^2 = 2(u^2 + v^2) = 4m = 4x^2\). Comparing these two results, we get \(4s^4 + r^4 = x^2\), which means that \((2s^2, r^2, x)\) is also an integral right triangle whose area \((rs)^2\) is a square. This triangle is smaller than our original triangle \((a, b, c)\) because \(x^2 = m^2 < m^2 + n^2 = c\); it may not be primitive, but the corresponding primitive triangle is even smaller.

The passage from the triple \((a, b, c)\) to the triple \((2s^2, r^2, x)\) can be construed as division by \(\pm 2\) on the elliptic curve \(C_1: y^2 = x^3 - x\); cf. the discussion before Exercise 7, and the beginning of § 6. The idea of the size of a triple leads to the notion of the height of a rational point on an elliptic curve.

Corollary 4. The equation \(x^4 - y^4 = z^2\) has no solutions in integers with \(xyz \neq 0\).

Proof: If there were a solution, the integral triangle \((2x^2y^2, x^4 - y^4, x^4 + y^4)\) would have square area \((xyz)^2\).

Corollary 5. The equation \(x^4 + y^4 = z^4\) has no solutions in integers with \(xyz \neq 0\).

Proof: If there were a solution, we would have \(z^4 - y^4 = (x^2)^2\).

The system of equations \((a^2 + b^2 = c^2; ab = 2\alpha)\) whose solvability in rational numbers characterises \(\alpha\) as a congruent numbers can be changed into a single, more familiar, equation.

Proposition 6. The integer \(\alpha\) is congruent if and only if the equation

\[
C_\alpha : \alpha y^2 = x^3 - x
\] (4)

has a solution \(x, y \in \mathbb{Q}\) with \(y \neq 0\).
**Proof**: If \((x, y)\) is such a solution, then the area of the rational right triangle \((2|x|, |x^2 - 1|, |x^2 + 1|)\) is \(\alpha\), up to a rational square \((y^2)\). Conversely, let \((a, b, c)\) be a rational right triangle and write

\[
a = \lambda \cdot 2mn, \quad b = \lambda \cdot (m^2 - n^2), \quad c = \lambda \cdot (m^2 + n^2)
\]

for some \(\lambda \in \mathbb{Q}\) and integers \(m, n\) as in Proposition 2. If the area of this triangle is \(\alpha\), we have \(\alpha = \lambda^2 \cdot mn(m^2 - n^2)\), which means that (4) has the solution \(x = m/n\) and \(y = 1/\lambda n\).

From a given rational point \(P = (x, y)\) \((y \neq 0)\) on \(C_\alpha\) (4) we can generate infinitely many others: the tangent to \(C_\alpha\) at the point \(P\) meets \(C_\alpha\) at another rational point \(P_1 = (x_1, y_1)\), and this process can be continued. It is not obvious, but it can be shown that this process does not terminate, essentially because it leads to points whose coordinates have more and more digits (cf. the discussion of torsion points on \(C_\alpha\), before Theorem 18).

**Exercise 7.** \((x_1, y_1) = \left( \frac{(x^2 + 1)^2}{4(x^2 - 1)}, -\frac{x^6 - 5x^4 - 5x^2 + 1}{8\alpha^2 y^3} \right)\).

This has an amusing consequence which is not at all obvious at the outset, and which goes back to Fermat:

**Corollary 8.** If \((a\text{ squarefree positive integer})\) \(\alpha\) is congruent, then it is the area of infinitely many rational right triangles.

**Corollary 9.** A squarefree positive integer \(\alpha\) is congruent if and only if the equation \(\alpha y^2 = x^3 - x\) has infinitely many solutions \(x, y \in \mathbb{Q}\).

Here are the first few rational squares \(d\) such that both \(d - 6\) and \(d + 6\) are also squares:

\[
\left( \frac{5}{2 \times 1} \right)^2, \left( \frac{1201}{2 \times 70} \right)^2, \left( \frac{7776485}{2 \times 1319901} \right)^2, \left( \frac{2094350404801}{2 \times 241717895860} \right)^2, \ldots
\]

The morphism \(y \mapsto \sqrt{\alpha} y\) shows that the curves \(C_1\) and \(C_\alpha\) become isomorphic over \(\mathbb{Q}(\sqrt{\alpha})\); this is expressed by saying that \(C_\alpha\) is a ‘quadratic twist’ of \(C_1\). The problem of congruent numbers thus consists in characterising the quadratic twists of the fixed elliptic curve \(C_1\) which have infinitely many rational points.

### 2. The Conjectural Solution

After these elementary observations, let us give the conjectural answer to the problem of characterising congruent numbers.

Recursively define the polynomial \(g_r(T) = g_{r-1}(T)(1 - T^{8r})(1 - T^{16r})\), starting with \(g_1(T) = T(1 - T^8)(1 - T^{16})\). Notice that \(g_r(T) - g_{r-1}(T)\) is of degree \(> 8r\), which means that the polynomials \(g_r\) and \(g_{r-1}\) have the same terms till degree \(8r\). This implies that as \(r \to +\infty\), the \(g_r\) tend to a formal series \(g \in \mathbb{Z}[[T]]\).
**Notation 10.** For \( j = 1, 2 \) and integer \( n > 0 \), define \( c_j(n) \) as being the coefficient of \( T^n \) in the formal series \( g(T) \theta_j(T) \), where

\[
g(T) = T \prod_{n=1}^{\infty} (1 - T^{8n})(1 - T^{16n}) \quad \text{and} \quad \theta_j(T) = 1 + 2 \sum_{n=1}^{\infty} T^{2jn^2}.
\]

Notice that the numbers \( c_j(n) \) are quite easy to compute. Here are the first few, for \( n \) odd and squarefree.

**Table 11.**

<table>
<thead>
<tr>
<th>( n )</th>
<th>1</th>
<th>3</th>
<th>5</th>
<th>7</th>
<th>11</th>
<th>13</th>
<th>15</th>
<th>17</th>
<th>19</th>
<th>21</th>
<th>23</th>
<th>\ldots</th>
</tr>
</thead>
<tbody>
<tr>
<td>( c_1(n) )</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>-2</td>
<td>0</td>
<td>0</td>
<td>-4</td>
<td>-2</td>
<td>0</td>
<td>0</td>
<td>\ldots</td>
</tr>
<tr>
<td>( c_2(n) )</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>-2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-4</td>
<td>0</td>
<td>\ldots</td>
</tr>
</tbody>
</table>

**Exercise 12.** Let \( n \) be an odd squarefree integer. If \( n \equiv 5, 7 \pmod{8} \), then \( c_1(n) = 0 \). If \( n \equiv 3 \pmod{4} \), then \( c_2(n) = 0 \).

For the remainder of this section, let \( \alpha \) be a squarefree integer > 0, and write \( \alpha = jn \), with \( j = 1, 2 \) and \( n \) odd.

**Conjecture 13.** The number \( \alpha = jn \) is congruent if and only if \( c_j(n) = 0 \).

As we shall see, this conjecture is implied by Conjecture 24 (Birch and Swinnerton-Dyer), combined with Theorem 25 (Tunnell).

The reader should marvel at how unexpected the (conjectural) characterisation is, how far-removed from rational right triangles and their areas!

The physicist Richard Feynman claims in his *Surely you are joking* that he could guess whether a mathematical statement explained to him in elementary terms was true or false. It would have been interesting to have given him Definition 1 and Notation 10, and to have asked him if Conjecture 13 is true.

We do know one of the implications in Conjecture 13:

**Theorem 14.** (J Coates and A Wiles [1]). If \( c_j(n) \neq 0 \), then \( \alpha = jn \) is not congruent.

It follows for example that the numbers 1, 2, 3, 10, 17, 19, 26 and 42 (Table 11) are not congruent (cf. Theorem 3).

If the squarefree odd integer \( n \) is \( \equiv 3 \pmod{4} \) (resp. \( \equiv 5, 7 \pmod{8} \)), then \( 2n \) (resp. \( n \)) should be congruent (Exercise 12, Conjecture 13), and the first few such \( n \) are indeed so (cf. (2)). In a paper which became influential only when it was properly understood years after its publication, K Heegner proved that this is true if \( n \) is prime [5].

However, in general, the result is only conditional. It is conditional on the finiteness of a certain set \( S_\alpha \), which will be discussed in detail in later parts of this article.
(cf. Conjecture 16). Suffice it to say here that the finiteness of the set $S_\alpha$ is equivalent to the finiteness of the group $\operatorname{III}(E_\alpha)$ which is more familiar to arithmeticians. We have chosen to formulate things in terms of the closely related set $S_\alpha$ because it can be defined in elementary terms.

**Theorem 15** (K. Rubin [3]). If $c_j(n) = 0$ and if the set $S_\alpha$ is finite, then the number $\alpha = jn$ is congruent.

Note that if $c_j(n) = 0$ and if $S_\alpha$ is finite, then Theorem 15 shows that $\alpha$ is congruent without exhibiting a rational right triangle of area $\alpha$. However, in some cases (‘rank 1’), the set $S_\alpha$ is known to be finite and there is a method (‘Heegner points’) for constructing such a triangle. Zagier’s example showing that 157 is congruent, displayed before Theorem 3, is of this type. A method for explicitly constructing solutions in the ‘higher rank’ case would be a major advance.

### 3. Local Number Fields

We have seen (Theorem 15) that the problem of deciding whether a given (squarefree) integer $\alpha > 0$ is congruent or not comes down to deciding if the set $S_\alpha$ – which we have mentioned but not yet defined – is finite or not. In order to define it, we first need to introduce K. Hensel’s local number fields.

The group $\mathbb{Q}^\times$, modulo its torsion subgroup $\{1, -1\}$, admits the set of prime numbers as a $\mathbb{Z}$-basis. For every prime number $p$, there is thus a unique homomorphism $v_p : \mathbb{Q}^\times \to \mathbb{Z}$ such that $v_p(p) = 1$ and $v_p(l) = 0$ for every prime number $l \neq p$; this defines a discrete valuation because, extending it to a map on the whole of $\mathbb{Q}$ by posing $v_p(0) = +\infty$, we have

$$v_p(x + y) \geq \inf(v_p(x), v_p(y)) \quad \text{for all } x, y \in \mathbb{Q}.$$ 

Define $| \cdot |_p : \mathbb{Q} \to \mathbb{R}$ by $|x|_p = p^{-v_p(x)}$ (convention: $p^{-\infty} = 0$). Then $|x - y|_p$ is a metric on $\mathbb{Q}$ with respect to which it can be completed to obtain a field $\mathbb{Q}_p$ much in the same way as we obtain the field $\mathbb{R}$ from $\mathbb{Q}$ by completing it with respect to the usual metric $|x - y|_{\infty} = \sup(x - y, y - x)$. For this reason, the field of real numbers is sometimes denoted $\mathbb{Q}_{\infty}$.

It can be shown that the $v_p$ ($p$ prime) are the only discrete valuations, and $| \cdot |_{\infty}$ the only archimedean absolute value, on the field $\mathbb{Q}$ (A. Ostrowski, 1918). Thus the absolute values $| \cdot |_p$ ($p$ prime or $p = \infty$) determine all the places of $\mathbb{Q}$.

The fields $\mathbb{Q}_p$ (including $p = \infty$) play a fundamental role in arithmetic. It is always a good idea to first study ‘global questions’ – questions about rational numbers – everywhere ‘locally’ in the fields $\mathbb{Q}_p$, before trying to answer the original question. It is a good idea because questions become simpler over local number fields and can often be further reduced to questions over finite fields. Sometimes it suffices to consider just one local field, sometimes finitely many, sometimes all of them. In most
cases, the local study is easy at almost all places. We discuss a basic example in the next section.

For $p$ prime, the (locally compact) field $\mathbb{Q}_p$ comes equipped with a continuous discrete valuation extending $v_p$; elements of positive valuation form a (compact) subring $\mathbb{Z}_p$ (‘the ring of integers’) in which $p\mathbb{Z}_p$ is the unique maximal ideal; it consists of elements whose valuation is strictly positive. The quotient $\mathbb{Z}_p/p\mathbb{Z}_p$ (‘the residue field’) is the same as the finite field $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$ of $p$ elements.

All books on Number Theory (Artin, Hasse, Weil, Serre, Katoś-Kurokawa–Saito, ...) provide an introduction to the fields $\mathbb{Q}_p$ and their extensions.

4. The Local-to-Global Principle for Conics

To avoid speaking of curves, we use the equivalent language of a function field $F$ over a field $k$: a finitely generated extension of $k$ in which $k$ is algebraically closed; we’ll be mostly concerned with the case when $F$ has transcendence degree 1 over $k$. Concretely, if $f \in k[x, y]$ is an absolutely irreducible polynomial – one which remains irreducible over every finite extension of $k$ –, then the field of fractions $F$ of the (integral) ring $k[x, y]/fk[x, y]$ is a function field over $k$; we write $F = k(x, y)$, with the relation $f = 0$. For every extension $L$ of $k$, we then get a function field over $L$ by ‘extending the scalars’ of $F$ from $k$ to $L$: the field of fractions of $L[x, y]/fL[x, y]$.

Let us fix an algebraic closure $\overline{\mathbb{Q}}$ of $\mathbb{Q}$. Clearly, the function field $\mathbb{Q}(x)$ becomes isomorphic to $\overline{\mathbb{Q}}(x)$ over $\overline{\mathbb{Q}}$. Are there any other function fields over $\mathbb{Q}$ which do? And, is there a way to classify them all?

Fix an algebraic closure $\overline{\mathbb{Q}}_p$ of $\mathbb{Q}_p$. The corresponding local question is: find all function fields over $\mathbb{Q}_p$ which become isomorphic over $\overline{\mathbb{Q}}_p$ to $\overline{\mathbb{Q}}_p(x)$. Such function fields will be called solutions to our problem.

The trivial solution to the problem is the rational function field $\mathbb{Q}_p(x)$. It can be shown that there is precisely one other solution; let us call it $F_p$. Thus the function field $F_p$ is not the rational function field but becomes (isomorphic to) the rational function field over $\overline{\mathbb{Q}}_p$. For example, when $p = \infty$, the field $F_\infty$ is $\mathbb{Q}_\infty(x, y)$ with the relation $x^2 + y^2 + 1 = 0$. When $p$ is an odd prime, choosing an integer $u \notin p\mathbb{Z}$ which does not become a square in $\mathbb{F}_p^\times$, we may take $F_p$ to be the function field over $\mathbb{Q}_p$ defined by the relation $ux^2 + py^2 − 1 = 0$. Over $\mathbb{Q}_2$, we may take $F_2$ to be the function field defined by $ux^2 + 2y^2 − 1 = 0$, where $u$ is any odd integer such that $(u^2 − 1)/8$ is also odd. Moreover, for every place $p$, it is an easy matter to decide if a given ‘local solution’ is isomorphic to $\mathbb{Q}_p(x)$ or to $F_p$.

Now, if $F$ is ‘global solution’ to our problem, then it is a ‘local solution’ everywhere. In other words, if $F$ is a function field over $\mathbb{Q}$ which becomes the rational function field over $\overline{\mathbb{Q}}$, then $F$ becomes isomorphic to one of $\mathbb{Q}_p(x)$, $F_p$ over every completion $\mathbb{Q}_p$ of $\mathbb{Q}$, including $p = \infty$. 

Chandan Singh Dalawat
What can be shown is that every such $F$ becomes isomorphic to $\mathbb{Q}_p(x)$ for almost every $p$, the places where it doesn’t—there are thus only finitely many of them—are even in number, and, given any finite set $\Sigma$ of places, even in number, there is a unique global solution which becomes isomorphic to $F_p$ for all $p \in \Sigma$ and to $\mathbb{Q}_p(x)$ for all $p \not\in \Sigma$.

There are many equivalent ways—curves of genus 0, quadratic forms in three variables, quaternion algebras—of expressing this principle.

It follows that if two global solutions $F$, $F'$ are ‘everywhere locally isomorphic’ (become isomorphic to each other at every place $p$, including $p = \infty$), then they are $\mathbb{Q}$-isomorphic. This happy circumstance is expressed by saying that such function fields obey the local-to-global principle. (In fact, in the case at hand, it is sufficient to demand that $F$, $F'$ be isomorphic at all places but one; they are then automatically isomorphic at the remaining place.)

The best accounts of this circle of ideas, in the equivalent language of quadratic forms, are to be found in Serre’s *Course in arithmetic* and in *Number theory 1, Fermat’s dream* by Kato, Kurokawa and Saito. A theorem of Adrien-Marie Legendre can be considered to be a precursor of local-to-global considerations, see Weil’s *Number theory, an approach through history*.

I don’t know of any classification of function fields over $\mathbb{Q}$ which become the 2-variable rational function field over every completion.

### 5. The Failure of the Local-to-Global Principle

In the last section we saw that the local-to-global principle holds for function fields over $\mathbb{Q}$ which become isomorphic over $\mathbb{Q}$ to the rational function field. Such function fields are of the form $\mathbb{Q}(x, y)$, $ax^2 + by^2 = 1$, for some $a, b \in \mathbb{Q}^\times$, and it is easy to decide when this field is isomorphic to the one defined by $a'x^2 + b'y^2 = 1$ ($a', b' \in \mathbb{Q}^\times$), because it suffices to check that they are isomorphic everywhere locally. This is a finite amount of computation because for any odd prime $p$ where all four numbers $a$, $b$, $a'$, $b'$ have valuation 0, the two function fields are automatically isomorphic to the rational function field $\mathbb{Q}_p(x)$.

In the early 1940s, Carl-Erik Lind and Hans Reichardt found the first examples of function fields which violate the local-to-global principle. Equivalently, Reichardt showed that $2y^2 = 1 - 17x^4$ has solutions in every completion of $\mathbb{Q}$ but no rational solutions—not even on the geometers’ ‘line at infinity in $\mathbb{P}^2$’ (not to be confused with our ‘place at infinity’ $\infty$ of $\mathbb{Q}$).

Lind’s thesis was reviewed by André Weil in the *Mathematical Reviews*, and it is amazing to note that he does not mention this discovery. Nor does the reviewer of Reichardt’s paper, in spite of the explicit title: *Einige im Kleinen überall lösbare, im Grossen unlösbare diophantische Gleichungen*. It must be said that the first instance
of the failure of a local-to-global principle, due to Hasse, was discovered by him after he had proved its validity for quadratic forms.

The most commonly cited example, originating with Ernst Selmer, is that of the function field \( \mathbb{Q}(x,y) \), \( 3x^3 + 4y^3 + 5 = 0 \). Cf. Example 27.

Let \( \alpha \) be a squarefree integer \( > 0 \) and consider the function field \( \mathbb{Q}(C_\alpha) \) defined by the equation \( C_\alpha : ay^2 = x^3 - x \). It may happen that there are many function fields \( F \) over \( \mathbb{Q} \) which become isomorphic to \( \mathbb{Q}(C_\alpha) \) at every place \( p \) of \( \mathbb{Q} \). In other words, \( \mathbb{Q}(C_\alpha) \) may have ‘twisted forms’ \( F \) which become isomorphic to it when we extend scalars of \( F \) and \( \mathbb{Q}(C_\alpha) \) from \( \mathbb{Q} \) to \( \mathbb{Q}_p \). Let us denote the set of \( \mathbb{Q} \)-isomorphism classes of such \( F \) by \( S_\alpha \). This is the set which appears in Theorem 15.

Thus the problem of congruent numbers would be solved if we could settle the following conjecture, whose generalisation Conjecture 26 is a major open question in contemporary arithmetic.

**Conjecture 16** (I Shafarevich and J Tate). For every \( \alpha \), the set \( S_\alpha \) of \( \mathbb{Q} \)-isomorphism classes of function field \( \mathbb{Q}(C_\alpha) \) at every place \( p \) is finite.

The more standard version of this conjecture asserts the finiteness of the group \( \text{III}(E_\alpha) \), whose definition is more advanced. The reader who knows it should be able to prove that \( S_\alpha \) is finite if and only if \( \text{III}(E_\alpha) \) is finite [4]. The same remark applies to Conjecture 26.

We have seen that the congruent number problem amounts to the arithmetic study of the equation \( \alpha y^2 = x^3 - x \), which can be rewritten as \( y^2 = x^3 - \alpha^2 x \). The rest of this report is devoted to a rapid survey of the arithmetic of equations of the type \( y^2 = f(x) \), where \( f \in \mathbb{Q}[x] \) is a monic cubic polynomial with distinct roots (in \( \overline{\mathbb{Q}} \)).

### 6. Elliptic Curves: Results and Conjectures

In the next two sections, we enumerate some arithmetic properties of elliptic curves. For the sake of simplicity, we work over the field \( \mathbb{Q} \); the only exceptions being a result over finite fields, one over \( \mathbb{Q}_p \), and an example over \( \mathbb{Q}(\sqrt{-1}) \).

An elliptic curve \( E \) over a field \( k \) is a curve defined in the projective plane by (the homogenous version of) an equation of the type

\[
  f(x, y) = y^2 + a_1 xy + a_3 y - x^3 - a_2 x^2 - a_4 x - a_6 = 0 \quad (a_i \in k)
\]  

(5)

without singularities, a condition which says that the discriminant \( \Delta \) – a certain polynomial in the \( a_i \) – is \( \neq 0 \), or equivalently that the corresponding function field is of ‘genus 1’, unlike the function fields which become isomorphic to \( \overline{\mathbb{Q}}(x) \), which are of genus 0.

More precisely, the discriminant of \( f \) – the result of eliminating \( x, y \) from \( f, f_x, f_y \) – is

\[
  \Delta = -b_2^2 b_8 - 23 b_4^3 - 3^3 b_6^2 + 3^2 b_2 b_4 b_6,
\]
where
\[ b_2 = a_1^2 + 2^2 a_2, \quad b_4 = a_1 a_3 + 2 a_4, \quad b_6 = a_3^2 + 2^2 a_6, \]
and
\[ b_8 = b_2 a_6 - a_1 a_3 a_4 + a_2 a_3^2 - a_4^2. \]

The curve \( E \) has a ‘point at infinity’ \( O \); for any extension \( L \) of \( k \), there is a natural group law on the set \( E(L) \) consisting of \( O \) and the solutions of (5) in \( L \), uniquely determined by the requirement that \( O \) be the neutral element and that the sum of the three points (counted with multiplicity) in which \( E \) intersects a given line be \( O \); the groups \( E(L) \) are commutative. Associativity is not obvious, but follows from a classical result in plane projective geometry.

Elements of \( E(L) \) can be identified with triples \((x, y, z) \neq (0, 0, 0) (x, y, z \in L)\) satisfying the homogenised version of (5); two such triples being considered the same if each is a multiple of the other by an element of \( L^\times \). The point \( O \) is the one with homogenous coordinates \((0 : 1 : 0)\). Two elliptic curves are said to be isomorphic if the corresponding function fields over \( k \) are \( k \)-isomorphic.

Let \( C \) be a smooth proper absolutely connected \( k \)-curve and let \( J \) be its jacobian – an abelian \( k \)-variety. If \( C \) has a \( k \)-rational point \( O \), there is a unique \( k \)-morphism \( C \to J \) sending a point \( P \) to the class of the divisor \( P - O \). If moreover \( C \) has genus 1, then \( C \to J \) is an isomorphism. Conversely, if (a smooth, proper, absolutely connected curve) \( k \)-curve \( C \) admits a group law, then \( C \) has genus 1 (and carries a rational point).

For surveys of arithmetic on elliptic curves, see Cassels [6] and Tate [7].

**Theorem 17** (L. Mordell, 1922). *For every elliptic curve \( E \) over \( \mathbb{Q} \), the group \( E(\mathbb{Q}) \) is finitely generated.*

This result was conjectured by Henri Poincaré around 1900. Mordell’s proof is a generalisation of Fermat’s method of infinite descent – employed in the proof of Theorem 3 – ; its modern renditions consist of two parts.

The first part shows that the group \( E(\mathbb{Q})/2E(\mathbb{Q}) \) is finite. The second part studies a canonical real-valued ‘height’ function \( h \) on \( E(\mathbb{Q}) \), coming from the various absolute values of \( \mathbb{Q} \), which measure how ‘big’ the coordinates of a point are. The method of infinite descent is distilled in the statement that a commutative group \( \Gamma \), endowed with such a function \( h \), for which \( \Gamma/2\Gamma \) is finite, is necessarily finitely generated.

Accounts of the proof can be found in the books by Weil and Kato–Kurokawa–Saito cited above, as well as in Silverman–Tate, *Rational points on elliptic curves*.

Notice that the additive group \( \mathbb{Q} \), the multiplicative group \( \mathbb{Q}^\times \), and the ‘elliptic’ group \( E(\mathbb{Q}) \) differ from each other greatly in their structure. By contrast, the corresponding local result says that for an elliptic curve \( E \) over \( \mathbb{R} \), the group \( E(\mathbb{R}) \) has a subgroup of index at most 2 isomorphic to \( \mathbb{R}/\mathbb{Z} \), and, for an elliptic curve \( E \) over \( \mathbb{Q}_p \) (\( p \) prime), \( E(\mathbb{Q}_p) \) has a subgroup of finite index isomorphic to \( \mathbb{Z}_p \). Thus, for an elliptic curve \( E \)
over \( \mathbb{Q} \), although the three groups \( \mathbb{Q}, \mathbb{Q}^\times, E(\mathbb{Q}) \) have very different structures, they are ‘almost the same’ everywhere locally.

For a given \( E \) over \( \mathbb{Q} \), the torsion subgroup of \( E(\mathbb{Q}) \) is easy to determine (Ét Lutz); for example, the torsion subgroup of \( C_\alpha(\mathbb{Q}) \) consists of \( O \) and the three points \((-1, 0), (0, 0), (1, 0)\) of order 2.

**Theorem 18** (B Mazur [8]). Let \( E \) be an elliptic curve over \( \mathbb{Q} \). The torsion subgroup of \( E(\mathbb{Q}) \) is isomorphic to one of the fifteen groups

\[
\mathbb{Z}/m\mathbb{Z} \quad (m = 1, 2, \ldots, 10, 12), \quad \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z} \quad (\nu = 1, 2, 3, 4).
\]

and each of these groups occurs as the torsion subgroup of \( E(\mathbb{Q}) \) for infinitely many \( E \).

No unconditional method is known, however, for determining the rank of \( E(\mathbb{Q}) \) for a given \( E \). The set of possible ranks for variable \( E \) over \( \mathbb{Q} \) is not known either, but N Elkies has recently produced examples where \( \text{rk} E(\mathbb{Q}) \) is at least 28. We shall mostly concentrate on the question of deciding if the rank is 0 or > 0.

Let \( p \) be a prime number and let \( E \) be an elliptic curve over \( \mathbb{Q}_p \), given by (5). We may assume by a change of variables that \( a_i \in \mathbb{Z}_p \); the discriminant \( \Delta \) is then in \( \mathbb{Z}_p \). If the \( a_i \in \mathbb{Z}_p \) can be so chosen that \( \Delta \in \mathbb{Z}_p^\times \), we say that \( E \) has good reduction at \( p \); if so, equation (5), read modulo \( p \), defines an elliptic curve \( E_p \) – uniquely determined by \( E \) and \( p \) – over the finite field \( \mathbb{F}_p \), and there is a homomorphism \( E(\mathbb{Q}_p) \to E_p(\mathbb{F}_p) \) which sends a point to the reduction modulo \( p \) of any of its representatives \((x : y : z)\) with coordinates in \( \mathbb{Z}_p \) and at least one coordinate in \( \mathbb{Z}_p^\times \). Any given elliptic curve \( E \) over \( \mathbb{Q} \) has good reduction at almost all primes because the defining equation (5) can be taken to have coefficients in \( \mathbb{Z} \) and because the discriminant \( \Delta \) has only finitely many prime factors.

There is a criterion for good reduction (‘Néron-Ogg-Shafarevich’). Let \( \bar{\mathbb{Q}}_p \) be an algebraic closure of \( \mathbb{Q}_p \). There is a unique extension of \( v_p \) to a valuation \( v_p : \bar{\mathbb{Q}}_p \to \mathbb{Q} \) of which the residue field \( \bar{\mathbb{F}}_p \) is an algebraic closure of \( \mathbb{F}_p \). The inertia group is the kernel of the natural surjection \( \text{Gal}(\bar{\mathbb{Q}}_p/\mathbb{Q}_p) \to \text{Gal}(\bar{\mathbb{F}}_p/\mathbb{F}_p) \); it acts on the \( m \)-torsion \( mE(\bar{\mathbb{Q}}_p) \) – the kernel of the multiplication by \( m \) – for every integer \( m \).

**Theorem 19** (J-P Serre and J Tate [9]). An elliptic curve \( E \) over \( \mathbb{Q}_p \) has good reduction if and only if the the action of the inertia group on \( mE(\mathbb{Q}_p) \) is trivial for every \( m \) prime to \( p \).

One might ask to what extent \( E \) is determined by the the number \( |E_p(\mathbb{F}_p)| \) of points modulo \( p \) for varying \( p \). We say that two elliptic curves are isogenous if their function fields can be embedded into each other.

**Theorem 20** (G Faltings [10]). If \( E' \) is an elliptic curve over \( \mathbb{Q} \) such that \( |E'_p(\mathbb{F}_p)| = |E_p(\mathbb{F}_p)| \) for almost all primes \( p \), then \( E' \) is isogenous to \( E \).

If two elliptic curves are isogenous, they have good reduction at the same primes...
We might wish to fix a finite set $T$ of primes and ask for a characterisation of all elliptic curves which have good reduction outside $T$—at every prime $p \notin T$. The first step is the following result:

**Theorem 21** (I Shafarevich, 1962). Given a finite set $T$ of primes, there are only finitely many elliptic curves over $\mathbb{Q}$ having good reduction at every prime $p \notin T$.

Another result of Shafarevich states that there are no elliptic curves over $\mathbb{Q}$ which have good reduction everywhere. This is an analogue of Minkowski’s theorem according to which there is no finite extension of $\mathbb{Q}$ (other than $\mathbb{Q}$ itself) which is everywhere unramified: the discriminant cannot be $\pm 1$.

There is a sense in which the more fundamental quantity is not $|E_p(F_p)|$ but $a_p(E)$, defined by $|E_p(F_p)| = 1 - a_p(E) + p$, and there is sense in which the following theorem is the analogue, for function fields of elliptic curves over finite fields, of the famous Riemann Hypothesis: ‘The zeros in the critical strip $0 < \text{Re}(s) < 1$ of the zeta function $\zeta$ of $\mathbb{Q}$ have real part $\frac{1}{2}$’.

**Theorem 22** (H Hasse, 1933). Let $A$ be an elliptic curve over a finite field $k$ of $q$ elements, and define $a$ by $|A(k)| = 1 - a + q$. Then $|a| \leq 2\sqrt{q}$.

Retuming to our $E$ over $\mathbb{Q}$, Birch and Swinnerton-Dyer argued that if $E(\mathbb{Q})$ is infinite, the groups $E_p(F_p)$ (for $p$ a prime of good reduction for $E$) should have more elements ‘on the average’ than if $E(\mathbb{Q})$ is finite. In view of Hasse’s theorem, the product $\prod_p \frac{1}{1 - a_p(E)/p - s + p^{-2s}}$ should diverge to $0$ if the rank is $> 0$, and converge to a limit $\neq 0$ if the rank is $0$. This is made precise in terms of the $L$-function of $E$.

For a prime $p$ of good reduction for $E$, we have the number $a_p(E)$; for ‘cohomological’ reasons, consider the infinite product (for $s \in \mathbb{C}$)

$$L(E, s) = \prod_p \frac{1}{1 - a_p(E)/p - s + p^{-2s}}.$$

Theorem 22 implies that this converges for $\text{Re}(s) > \frac{3}{2}$, but more is true:

**Theorem 23** (A Wiles, R Taylor, F Diamond, B Conrad, C Breuil, 1995–2000). The function $L(E, s)$ admits an analytic continuation to the whole of $\mathbb{C}$.

For the congruent number elliptic curves $C_\alpha$, this is due to André Weil. There is a way of introducing factors in $L(E, s)$ corresponding to the primes which divide $\Delta$, and indeed to the place $\infty$. This ‘completed’ $L$-function $\Lambda(E, s)$ has a ‘functional equation’ for $s \mapsto 2 - s$, just as the $\zeta$-function, when ‘completed’, has a functional equation for $s \mapsto 1 - s$.

Note that the product $\prod_p \frac{1}{1 - a_p(E)/p - s + p^{-2s}}$ is formally equal to $L(E, 1)$. The above heuristic considerations and extensive calculations on one of the first electronic computers at Cambridge led to the following conjecture.

**Conjecture 24** (B Birch and P Swinnerton-Dyer, 1965). The group $E(\mathbb{Q})$ is infinite.
if and only if \( L(E, 1) = 0 \). More precisely, its rank equals the order of vanishing of \( L(E, s) \) at \( s = 1 \).

The order of vanishing of the completed \( L \)-function \( \Lambda(E, s) \) is the same as that of \( L(E, s) \) at \( s = 1 \). There is a refined version of Conjecture 24 which gives the leading coefficient of \( \Lambda(E, s) \) at \( s = 1 \) in terms of the local and global arithmetic invariants of the curve \( E \); its formulation is subject to the truth of Conjecture 26.

Conjecture 13 follows from Conjecture 24, thanks to the following criterion:

**Theorem 25.** (J Tunnell [2]). For a squarefree integer \( \alpha = jn \) (\( j = 1, 2 \) and \( n \) odd), one has \( L(C_\alpha, 1) = 0 \) if and only if \( c_j(n) = 0 \).

The elliptic curve \( E \) has the function field \( \mathbb{Q}_{l}(E) \) at the various places \( l \) of \( \mathbb{Q} \). Just as we did in the case of the congruent number elliptic curves \( C_\alpha \), we now consider the set \( S_E \) of (isomorphism classes of) all function fields over \( \mathbb{Q} \) which becomes isomorphic to \( \mathbb{Q}_{l}(E) \) at every place \( l \); of course, \( \mathbb{Q}(E) \) belongs to \( S_E \).

**Conjecture 26** (I Shafarevich and J Tate). For every elliptic curve \( E \) over \( \mathbb{Q} \), the set \( S_E \) is finite.

The original conjecture asserts the finiteness, for every \( E \) over \( \mathbb{Q} \), of the group \( \text{III}(E) \) of ‘torsors’ under \( E \) which are ‘everywhere locally trivial’. This is equivalent to the finiteness of \( S_E \).

Yuri Manin has introduced an ‘obstruction’ to explain the failure of the local-to-global principle for the function field \( \mathbb{Q}(E) \) of an elliptic curve \( E \) over \( \mathbb{Q} \). He shows that the finiteness of \( S_E \) is equivalent to his obstruction being the only one.

The equation \( x^3 + y^3 + 60 = 0 \) can be put in the form (5) by a change of variables; it therefore defines an elliptic curve.

**Example 27** (B Mazur [4]). For \( E \) defined by \( x^3 + y^3 + 60 = 0 \), the set \( S_E \) consists of \( \mathbb{Q}(E) \) and the function fields defined by

\[
3x^3 + 4y^3 + 5, \; 12x^3 + y^3 + 5, \; 15x^3 + 4y^3 + 1, \; 3x^3 + 20y^3 + 1.
\]

The best available result in the direction of Conjectures 24 and 26 to date, the fruit of a succession of papers by numerous mathematicians, is a theorem of Victor Kolyvagin, of which the theorem of Coates and Wiles (Theorem 14) is a particular case, and which subsumes some of the results of Benedict Gross and Don Zagier [11].

**Theorem 28** (V Kolyvagin [12]). If \( L(E, 1) \neq 0 \), then \( E(\mathbb{Q}) \) is finite. If \( L(E, s) \) has a simple zero at \( s = 1 \), then \( E(\mathbb{Q}) \) has rank 1. In both these cases, the set \( S_E \) is finite.

If the zero at \( s = 1 \) has multiplicity \( > 1 \), Conjecture 26 is needed (Cf. Theorem 15):

**Theorem 29** (C Skinner and É Urban [13]). Suppose that \( L(E, 1) = 0 \) and that the set \( S_E \) is finite. Then the group \( E(\mathbb{Q}) \) is infinite.

There is a parallel theory of elliptic curves \( E \) over function fields \( F \) over finite fields. The analogue of Mordell’s theorem (Theorem 17) is true: the group \( E(F) \) is finitely
generated. K Kato and F Trihan [14] have proved the analogue of (the refined version of) the Birch and Swinnerton-Dyer conjecture (Conjecture 24), subject to the truth of the analogue of the Shafarevich-Tate conjecture (Conjecture 26).

The study of ‘special values’ of $L$-functions, of which the refined conjecture of Birch and Swinnerton-Dyer is the prototype, is one of the major themes of contemporary arithmetic. This is a very active area in which P Deligne, S Bloch, A Beilinson, K Kato, J-M Fontaine, B Perrin-Riou, among others, have made seminal contributions.

7. Complex Multiplications

Let $E$ be an elliptic curve over $\overline{\mathbb{Q}}$, defined by an equation $f(x, y) = 0$. Because $E$ has a group law, there are many embeddings of the function field $\overline{\mathbb{Q}}(E) = \overline{\mathbb{Q}}(x, y)$ into itself: for every integer $n \neq 0$, there is an embedding $[n]_E$ which sends $x, y$ to $x_n, y_n$, the coordinates of the multiple $nP$ of the point $P = (x, y)$; the embedding $[n]_E$ is of degree $n^2$. For example, when $E$ is the congruent number elliptic curve $C_\alpha$ (4) and $n = -1$, it is the automorphism $x \mapsto x, y \mapsto -y$ of the function field; when $n = -2$, it is the degree-4 embedding given in Exercise 7.

In a sense, for most elliptic curves, these are the only embeddings of the function field into itself. But there are some elliptic curves for which there are more embeddings, for example the automorphism $x \mapsto -x, y \mapsto iy$ ($i$ being a chosen square root of $-1$) of $\overline{\mathbb{Q}}(C_\alpha)$ whose square is $[-1]_{C_\alpha}$. In such a case we say that the elliptic curve $E$ has ‘complex multiplications’; it then determines an imaginary quadratic field $K$, the field of fractions of the ring of $\overline{\mathbb{Q}}$-endomorphisms of $E$. In the case of the curves $C_\alpha$, it is $\mathbb{Q}(i)$. We say that $E$ has complex multiplications by $K$.

The arithmetic properties of elliptic curves differ vastly according as they have complex multiplications or not. For example, for elliptic curves having complex multiplications, the theorem ‘$L(E, 1) \neq 0 \Rightarrow E(\mathbb{Q})$ is finite’ was proved by Coates-Wiles (cf. Theorem 14) a good eleven years before Kolyvagin’s general result (cf. Theorem 28); the analytic continuation of $L(E, s)$ was proved by Weil and Max Deuring in 1953–1957, much before the general result of Wiles and his school in 1995–2000 (Theorem 23); the implication ‘$L(E, 1) = 0$ and $S_E$ finite $\Rightarrow E(\mathbb{Q})$ infinite’ was proved by Rubin (cf. Theorem 15) some fifteen years before the general result of Skinner–Urban (Theorem 29).

We illustrate the differences by three examples. For the first, recall that an elliptic curve $A$ over $\mathbb{F}_p$ is called supersingular if the $p$-torsion $pA(\overline{\mathbb{F}}_p)$ is reduced to $\{O\}$, or, equivalently for $p \neq 2, 3$, if $|A(\mathbb{F}_p)| = 1 + p$ (equivalently, $a = 0$, in the notation of Theorem 22). Returning to our $E$ over $\mathbb{Q}$, we ask: How often is $E_p$ supersingular? Deuring showed if $E$ has complex multiplications, then this happens for half the primes $p$ (cf. Example 33); if not, Serre proved that the set of supersingular primes has density 0. That it is infinite is a relatively recent result.
For the second example, recall that for every prime \( p \), if we adjoin the \( p \)-torsion of the multiplicative group \( \mathbb{Q}^\times \), which consists of \( p \)th roots of 1, to \( \mathbb{Q} \), we get a galoisian extension \( \mathbb{Q}(\mu_p) \) whose group of automorphisms is \( \text{Gal}(\mathbb{Q}(\mu_p)|\mathbb{Q}) = GL_1(\mathbb{F}_p) \). For an elliptic curve \( E \) over \( \mathbb{Q} \), the \( p \)-torsion of \( E(\mathbb{Q}) \) is a 2-dimensional vector \( \mathbb{F}_p \)-space; if we adjoin it to \( \mathbb{Q} \), we get a galoisian extension \( \mathbb{Q}(\mu_p,E) \). What is \( \text{Gal}(\mathbb{Q}(\mu_p,E)|\mathbb{Q}) \)?

**Theorem 31** (J-P Serre [16]). Suppose that \( E \) does not have complex multiplications. Then the group of automorphisms of \( \mathbb{Q}(\mu_p,E) \) is \( GL_2(\mathbb{F}_p) \) for almost all \( p \) but finitely many \( p \) – primes.

The corresponding local result for \( E \) over \( \mathbb{Q} \) says, at least in the case of good reduction, that \( \text{Gal}(\mathbb{Q}(\mu_p,E)|\mathbb{Q}) \) is cyclic for \( l \neq p \) (cf. Theorem 19).

If \( E \) (over \( \mathbb{Q} \)) has complex multiplications, the group of automorphisms is much smaller: if \( K \) – an imaginary quadratic field – is the field of complex multiplications, then \( K(\mu_p,E) \) is an abelian extension of \( K \). However, such \( K \) serve a different, if related, purpose.

Recall that the theorem of Kronecker–Weber asserts that if we adjoin the entire torsion subgroup of \( \mathbb{Q}^\times \) – all roots of 1 – to \( \mathbb{Q} \), we get its maximal abelian extension.

Generating the maximal abelian extension of other number fields is a major open problem (Kronecker’s Jugendtraum, Hilbert’s Problem 12); the theory of complex multiplications provides the answer in the case of imaginary quadratic fields, as in the next example.

**Example 32.** Let \( E \) be the elliptic curve \( y^2 = x^3 + x \), which has complex multiplications by \( \mathbb{Q}(i) \) \((i = \sqrt{-1})\). If we adjoin the entire torsion subgroup of \( E(\mathbb{Q}) \) to \( \mathbb{Q}(i) \), we get its maximal abelian extension.

Our third example concerns a ‘formula’ for \( a_p(E) \) for a fixed \( E \) and varying \( p \). There is indeed such a formula if \( E \) has complex multiplications, as illustrated by a theorem of Gauss about the curve \( x^3 + y^3 + 1 = 0 \) (which can be put in the canonical form (5), and has complex multiplications by \( \mathbb{Q}(\zeta_3) \), where \( \zeta_3 \) is a primitive cube root of 1). It uses the fact that for a prime \( p \equiv 1 \pmod{3} \), there is a pair of integers \((c_p,d_p)\), unique up to signs, such that \( 4p = c_p^2 + 27d_p^2 \); to fix the sign of \( c_p \), assume that \( c_p \equiv -1 \pmod{3} \).

**Example 33** (C Gauss, 1801). Let \( E \) be the elliptic curve \( x^3 + y^3 + 1 = 0 \) and \( p \) a prime. If \( p \equiv 1 \pmod{3} \), then \( a_p(E) = c_p \). If \( p \equiv -1 \pmod{3} \), then \( a_p(E) = 0 \).

See Silverman–Tate for a proof. Note that this implies Theorem 22 for \( E \).

By contrast, if \( E \) does not have complex multiplications, the behaviour of the \( a_p(E) \) is entirely different. Mikio Sato and John Tate independently arrived at a conjectural distribution law for \( \gamma_p(E) = a_p(E)/2\sqrt{p} \), which lies between \(-1\) and \(+1\) for every \( p \) (cf. Theorem 22). How often does it lie in \([\beta, \delta] \subset [-1, +1]\)?
Conjecture 34 (M Sato and J Tate, 1960). Suppose that $E$ does not have complex multiplications, and let $[\beta, \delta] \subset [-1, +1]$ be an interval. Then the proportion of primes $p$ for which $\gamma_p(E) \in [\beta, \delta]$ is given by

$$\frac{2}{\pi} \int_{\beta}^{\delta} \sqrt{1-x^2} \, dx.$$ 

This conjecture has been proved, subject to a mild technical hypothesis on $E$, by Laurent Clozel, Michael Harris, Nicholas Shepherd-Barron and Richard Taylor in a series of three papers in early 2006. The technical hypothesis demands that $E$ have ‘multiplicative reduction’ at some prime $p$, which means roughly that the best possible reduction at $p$ is not an elliptic curve $E_p$ as in the case of good reduction, but the multiplicative group (and not the additive group – the third possibility). An algorithm due to Tate allows one to determine the type of reduction at any given $p$ in terms of the coefficients $a_i$ (5) defining $E$. Concretely, although we cannot choose $a_i \in \mathbb{Z}_p$ with minimal $v_p(\Delta)$ so as to have $v_p(\Delta) = 0$, they can be so chosen as to have $v_p(c_4) = 0$, where $c_4 = b_2^2 - 2^3.3.b_4$, and the $b_i$ are displayed after equation (5).

It is only a matter of time before this hypothesis is removed.

Theorem 35 (L Clozel, M Harris, N Shepherd-Barron and R Taylor [17]). Conjecture 34 is true if $E$ has multiplicative reduction at some prime $p$.

8. Modular Forms

We have not mentioned them, although they have appeared in these notes without being named. If we evoke them here, it is only to say that most of the spectacular recent results which we have enumerated would not have been possible without their help. Take the analytic continuation of $L(E, s)$ (Theorem 23): the crucial result (Wiles and others) is to show that the sequence $(a_p(E))_p$ defines a modular form.

Results of Gross–Zagier and of Kolyvagin (Theorem 28), which predate Wiles, were enunciated only for those elliptic curves whose $L$-functions have this modularity property; thanks to Wiles and his successors, we now know that they all have.

Mazur’s determination of the possible torsion subgroups (Theorem 18) involves the study of modular curves, which are intimately related to modular forms.

Tunnell’s criterion (Theorem 25) is actually an expression for $L(C_\alpha, 1)$ in terms of (the ‘real period’ of $C_\alpha$ and) the coefficients $c_j(n)$ of certain modular forms of half-integral weight (cf. Notation 10).

The role of automorphic forms – a generalisation of modular forms – is even greater in the results of Skinner–Urban (Theorem 29) and in the proof of the Sato–Tate conjecture (Theorem 34). It is unlikely to diminish in the future: more and more $L$-functions are going to become automorphic, fulfilling the prophetic vision of Robert Langlands [18].
For a first introduction, apart from Serre’s *Course*, see the book by Koblitz and Knapp’s *Elliptic curves*.

**Acknowledgements**

I thank Pete Clark for a very careful reading of the manuscript, and for his suggestions for improvement.

**Postscript (added in December 2021)**

*Number-theory is not standing still.*

— André Weil, *Two lectures on Number Theory*, 1974

Since the appearance of my popular article on *Congruent numbers, elliptic curves, and the passage from the local to the global*, published in *Resonance* in December 2009, a number of new results have been obtained on the topics discussed there. We review a few of these, using the same notation and terminology. We also continue with the numbering of the statements and bibliographic items.

Three major advances in the arithmetic of elliptic curves have taken place in the last decade. The Goldfeld Conjecture (recalled below) has been proved for the family of elliptic curves arising from the congruent number problem (and for many other families besides), a converse to the theorems of Coates & Wiles, Rubin, Gross & Zagier, Kolyvagin, and Kato has been proved, and a number of striking statistical results has been obtained about curves whose group of rational points has rank 0 or 1, about the average rank of elliptic curves over $\mathbb{Q}$, and about the proportion which satisfy the Birch & Swinnerton-Dyer Conjecture. There has also been much speculation about the boundedness of the rank in various families of elliptic curves, or about the number of elliptic curves of given rank. These are the four topics we touch upon in this update. We also briefly mention that the Parity Conjecture (a very special case of the Birch & Swinnerton-Dyer Conjecture) follows from the Shafarevich-Tate conjecture.

Before we begin, I should warn the reader that some of the material covered in these notes has not yet appeared in print and therefore has not yet been fully verified by the mathematical community. My role here is that of a reporter, not an endorser.

First of all, note that the technical hypothesis in Theorem 35 requiring the elliptic curve to have at least one place of multiplicative reduction has been removed:

**Theorem 36** (T. Barnet-Lamb, D. Geraghty, M. Harris, & R. Taylor [19]). — *The Sato-Tate conjecture (Conjecture 34) is true for every elliptic curve over $\mathbb{Q}$ (without complex multiplications).*

An elementary introduction to the problem of congruent numbers and elliptic curves leading up to the Sato-Tate conjecture can be found in the inaugural lecture by Toby
9. **The sign of the functional equation**

We have seen that a squarefree integer \( \alpha > 0 \) is congruent if and only if there exist \( x, y \in \mathbb{Q} \) such that \( y \neq 0 \) and \( \alpha y^2 = x^3 - x \) (Proposition 6). In other words, the group \( C_\alpha(\mathbb{Q}) \) of rational points on the elliptic curve

\[
C_\alpha : \alpha y^2 = x^3 - x
\]

is infinite if and only if \( \alpha \) is a congruent number. The conjecture of Birch & Swinnerton-Dyer (Conjecture 24) predicts that the group \( C_\alpha(\mathbb{Q}) \) is infinite if and only if \( L(C_\alpha, 1) = 0 \) (the \( L \)-function \( L(C_\alpha, s) \) vanishes at \( s = 1 \), or equivalently the order of vanishing \( \text{ord}_{s=1} L(C_\alpha, s) > 0 \)). By computing the “global root number” or the “sign of the functional equation” of \( C_\alpha \) [22], one can check that \( \text{ord}_{s=1} L(C_\alpha, s) \) is even if \( \alpha \equiv 1, 2, 3 \) (mod. 8) and odd if \( \alpha \equiv -3, -2, -1 \) (mod. 8).

If the order of vanishing is odd, then certainly \( L(C_\alpha, 1) = 0 \). We are thus led by Birch & Swinnerton-Dyer to the following:

**Conjecture 37.** — If \( \alpha > 0 \) is squarefree and \( \equiv -3, -2, -1 \) (mod. 8), then it is a congruent number.

The converse is not true: 34 is the area of the right-triangle with sides \( 225/30, 272/20, \) and \( 353/30 \). Nearly a decade before Birch & Swinnerton-Dyer came up with their conjecture, Heegner had already taken the first important step towards proving Conjecture 37 in a paper [23] which went unnoticed at the time:

**Theorem 38 (K. Heegner, 1952).** — Let \( \alpha > 0 \) be a squarefree integer such that \( \alpha \equiv -3, -2, -1 \) (mod. 8). If \( \alpha \) has precisely one odd prime divisor, then \( \alpha \) is a congruent number.

This was generalised by Monsky (1990) to allow \( \alpha \) to have two odd prime divisors. Ye Tian [24] took a major step forward by proving the following important result:

**Theorem 39 (Y. Tian, 2014).** — For every \( j \equiv -3, -2, -1 \) (mod. 8) and every \( n > 0 \), there are infinitely many squarefree integers \( \alpha > 0 \) having precisely \( n \) odd prime divisors and such that \( \alpha \equiv j \) (mod. 8) which are congruent numbers.

For an introduction to Tian’s work, see the article [25] by John Coates.

How about squarefree integers \( \alpha > 0 \) such that \( \alpha \equiv 1, 2, 3 \) (mod. 8), when the order of vanishing \( \text{ord}_{s=1} L(C_\alpha, s) \) is even? If moreover \( \alpha \) has a unique odd prime divisor \( p \), there is the following result which goes back to the 19th century in part:
**Theorem 40** (A. Genocchi, 1874 ; M. Razar, 1974). — *Let $p$ be a prime number. If $p \equiv 3 \pmod{8}$, then $p$ is not congruent. If $p \equiv 5 \pmod{8}$, then $2p$ is not congruent.*

What happens if $p \equiv 1 \pmod{8}$, or if $\alpha$ has several odd prime divisors? The following analogue of Theorem 39 says something about these cases:

**Theorem 41** (K. Feng, 1996 ; D. Li & Y. Tian, 2000 ; C. Zhao, 2001). — *For every $j \equiv 1, 2, 3 \pmod{8}$ and every $n > 0$, there are infinitely many squarefree integers $\alpha > 0$ having precisely $n$ odd prime divisors and such that $\alpha \equiv j \pmod{8}$ which are not congruent numbers.*

See for example the expository article [26] by Shou-Wu Zhang and his lecture [27] at Luminy.

**10. The minimalist philosophy**

Before we state Goldfeld’s conjecture in the special case of the family of elliptic curves $C_{\alpha}$ (where $\alpha > 0$ is a squarefree integer), we need to recall that if $S$ is a set of positive integers and $T$ is a subset of $S$, then the *density*, or the *natural density*, of $T$ in $S$ is the limit (if it exists)

$$\delta_{T,S} = \lim_{x \to +\infty} \frac{t_x}{s_x},$$

where $t_x$ (resp. $s_x$) is the number of elements in $T$ (resp. $S$) which are $< x$. We also say that $100\delta_{T,S}\%$ of elements of $S$ are in $T$. For example, when $S = \mathbb{N}$ is the set of all positive integers and $T = 2\mathbb{N}$ is the set of even numbers, then $\delta_{S,T} = 1/2$, or $50\%$ of positive integers are even. We also say that the *proportion* of elements of $S$ which are in $T$ is $\delta_{T,S}$.

Note that $\delta_{T,S} = 0$ does not imply that $T$ is empty, nor does $\delta_{T,S} = 1$ imply that $T = S$. For example, $0\%$ of positive integers are prime numbers, yet there are infinitely many of them.

**Conjecture 42** (D. Goldfeld, 1979). — *Among the squarefree integers $\alpha > 0$ such that $\alpha \equiv 1, 2, 3 \pmod{8}$, the percentage of those for which $\text{ord}_{s-1} L(C_{\alpha},s) = 0$ is $100\%$.*

This is sometimes called Goldfeld’s even-parity conjecture. Combined with the theorem of Coates & Wiles, this means that the percentage of such $\alpha$ for which $C_{\alpha}(\mathbb{Q})$ is infinite should be $0\%$. As $\alpha$ is a congruent number if and only if $C_{\alpha}(\mathbb{Q})$ is infinite, Conjecture 42 implies the following

**Conjecture 43** (D. Goldfeld, 1979). — *The percentage of squarefree $\alpha > 0$ such that $\alpha \equiv 1, 2, 3 \pmod{8}$ which are congruent is $0\%$.*

Similarly, he conjectured the following odd-parity version:
**Conjecture 44** (D. Goldfeld, 1979). — *Among the squarefree integers* \( \alpha > 0 \) *such that* \( \alpha \equiv -3, -2, -1 \) (mod. 8), *the percentage of those for which* \( \operatorname{ord}_s L(C_\alpha, s) = 1 \) *is 100%.*

By Kolyvagin’s theorem (Theorem 28), this would imply that the percentage of squarefree \( \alpha > 0 \) such that \( \alpha \equiv -3, -2, -1 \) (mod. 8) which are congruent numbers is 100%.

Alexander Smith [28] and Daniel Kriz [31] have succeeded in proving these three conjectures but the proofs have not yet been published. In fact, their result applies to the family of “quadratic twists” of any elliptic curve satisfying a certain technical hypothesis (which \( C_1 \) does).

**Theorem 45** (A. Smith, 2017). — *The percentage of squarefree* \( \alpha > 0 \) *such that* \( \alpha \equiv 1, 2, 3 \) (mod. 8) *which are congruent is 0%.*

One can listen to Smith’s talks at the Institute for Advanced Study [32] and at the Institut Henri Poincaré [33]. For an expository account, see [34].

**Theorem 46** (D. Kriz, 2020). — *The percentage of squarefree* \( \alpha > 0 \) *such that* \( \alpha \equiv -3, -2, -1 \) (mod. 8) *which are congruent is 100%.*

Note that the Birch & Swinnerton-Dyer Conjecture predicts that all squarefree \( \alpha > 0 \) such that \( \alpha \equiv -3, -2, -1 \) (mod. 8) are congruent, because \( \operatorname{ord}_s L(C_\alpha, s) \) is odd, as explained above (Conjecture 37).

Recall (Notation 10) that we had defined \( c_j(n) \), for \( j = 1, 2 \) and \( n > 0 \), as the coefficient of \( T^n \) in the formal series \( g(T)\theta_j(T) \), where

\[
g(T) = T \prod_{n>0} (1 - T^{3n})(1 - T^{16n}) \quad \text{and} \quad \theta_j(T) = 1 + 2 \sum_{n>0} T^{2jn^2}.
\]

Recall also that if \( \alpha = jn \) with \( j = 1, 2 \) and \( n > 0 \) squarefree and odd, then \( \alpha \) is not a congruent number if \( c_j(n) \neq 0 \), by the theorems of Coates & Wiles (Theorem 14) and Tunnell (Theorem 25). In view of this observation, Theorem 45 follows from the following recent result:

**Theorem 47** (A. Burungale & Y. Tian [35]). — *The percentage of squarefree integers* \( \alpha = jn \) *(with* \( j = 1, 2 \) *and* \( n > 0 \) *odd) such that* \( \alpha \equiv 1, 2, 3 \) (mod. 8) *for which* \( c_j(n) = 0 \) *is 0%.*

**11. Statistics of elliptic curves**

We have just come across conjectures about the percentage of elliptic curves in the family \( C_\alpha \) (\( \alpha > 0 \) running through squarefree integers) for which the group \( C_\alpha(\mathbb{Q}) \) has rank 0, rank 1, etc. The curves \( C_\alpha \) are “quadratic twists” of the elliptic curve \( C_1 \).
There are similar conjectures about the family of all elliptic curves over $\mathbb{Q}$, giving the percentage of curves for which the group of rational points has rank 0, rank 1, etc. These percentages are defined by first computing the percentage, among the finitely many elliptic curves of bounded "height", of those which have the given rank, and then taking the limit as the bound goes to $+\infty$. More precisely, every elliptic curve $E$ over $\mathbb{Q}$ admits a unique equation

$$E : \quad y^2 = x^3 + Ax + B, \quad (A, B) \in \mathbb{Z}; \quad 4A^3 + 27B^2 \neq 0$$

such that there is no prime number $p$ for which $p^{12}$ divides $\gcd(A_4, A_6)$; the height of $E$ is defined to be $h(E) = \text{Sup}(|4A_4|, |27A_6^2|)$. Clearly, for every $c > 0$, there are only finitely many elliptic curves $E$ over $\mathbb{Q}$ such that $h(E) < c$.

One can similarly define the average rank of elliptic curves over $\mathbb{Q}$ as the limit, when $c \to +\infty$, of the average of $\text{rk}_{\mathbb{Z}} E(\mathbb{Q})$ for the finitely many elliptic curves $E$ over $\mathbb{Q}$ such that $h(E) < c$.

**Conjecture 48** (D. Goldfeld, 1979; N. Katz & P. Sarnak [36]). For 50% of elliptic curves $E$ over $\mathbb{Q}$, the group $E(\mathbb{Q})$ has rank 0; for 50% of them, it has rank 1; and for 0% of them, it has rank $> 1$. Consequently, the average rank of elliptic curves over $\mathbb{Q}$ is $\frac{1}{2}$.

When this conjecture was made, it was not known that the percentage of elliptic curves of rank 0 or 1 is $> 0\%$, let alone $= 100\%$, or that the average rank of elliptic curves is $< +\infty$, let alone $= \frac{1}{2}$. Since then, remarkable progress has been made:

**Theorem 49** (M. Bhargava & A. Shankar [37]). When elliptic curves $E$ over $\mathbb{Q}$ are ordered by height, the average rank of the $\mathbb{Z}$-module $E(\mathbb{Q})$ is $< 1$.

**Theorem 50** (M. Bhargava & A. Shankar [38]). When elliptic curves $E$ over $\mathbb{Q}$ are ordered by height, the proportion for which $\text{rk}_{\mathbb{Z}} E(\mathbb{Q}) = 0$ is $> 0$, and the proportion which satisfy the Birch & Swinnerton-Dyer conjecture is $> 0$.

**Theorem 51** (M. Bhargava & C. Skinner [40]). When elliptic curves $E$ over $\mathbb{Q}$ are ordered by height, the proportion for which $\text{rk}_{\mathbb{Z}} E(\mathbb{Q}) = 1$ is $> 0$.

For a succinct account of these results when they were discovered, see Manjul Bhargava’s talk at the Clay conference at Harvard in 2011 [39]; see also the Bourbaki talk by Bjorn Poonen [43]. Since then the percentages have been steadily improved. For example, within a few years, it was proved that a majority of elliptic curves satisfy Conjecture 24:

**Theorem 52** (M. Bhargava, C. Skinner, & W. Zhang [41]). When ordered by height, more than half the elliptic curves over $\mathbb{Q}$ satisfy the Birch & Swinnerton-Dyer conjecture.
For recent oral overviews of the Birch & Swinnerton-Dyer conjecture, see the talks by Manjul Bhargava [42] and Andrew Wiles [44] at Clay conferences, by Chris Skinner at the Heilbronn Institute [45], and by Benedict Gross [46] at Caltech. Among the written sources, one may consult the description by Wiles [47] for the Clay Institute, the lectures by John Coates [48], the article by Wei Zhang [49], and the recent article by Ashay Burungale, Chris Skinner & Ye Tian [50].

Let us finally mention that there has been very important work towards proving a converse to the theorems of Coates-Wiles (cf. Theorem 14), Rubin (cf. Theorem 15), Gross & Zagier, Kolyvagin (Theorem 28), and Kato by Chris Skinner, Wei Zhang, Ashay Burungale, Ye Tian, Daniel Kriz, and Chan-Ho Kim. The proof of the following converse theorem is now complete, although all the details have not yet been published:

**Theorem 53** (A. Burungale, C. Skinner, & Y. Tian [53]). — Let $E$ be an elliptic curve over $\mathbb{Q}$. If the Shafarevich-Tate conjecture (Conjecture 26) holds for $E$ and $\text{rk}_Z(E(\mathbb{Q})) = 0, 1$, then $\text{ord}_{s=1} L(E, s) = \text{rk}_Z(E(\mathbb{Q}))$.

See also the preprint by Chan-Ho Kim [54]. The interested reader is referred to the Princeton talks of Burungale [51] and Tian [52]. Note that according to Conjecture 48, 100% of elliptic curves over $\mathbb{Q}$ have $\text{rk}_Z(E(\mathbb{Q})) = 0, 1$.

### 12. The parity of the rank

For an elliptic curve $E$ over $\mathbb{Q}$, the Parity Conjecture asserts that $\text{rk}_Z(E(\mathbb{Q})) \equiv \text{ord}_{s=1} L(E, s) \pmod{2}$. Paul Monsky [57] has proved that this easy consequence of Conjecture 24 (Birch & Swinnerton-Dyer) follows from Conjecture 26 predicting the finiteness of the Shafarevich-Tate group $\text{III}(E)$. More recently, Tim & Vladimir Dokchitser have proved that it is enough to require the finiteness of the $p$-primary component of $\text{III}(E)$ for a single arbitrary prime $p$. See the nice survey [58].

### 13. The boundedness of the rank

For a long time it was believed that there are only finitely many possibilities for the rank of the $\mathbb{Z}$-module $E(\mathbb{Q})$ as $E$ runs through all elliptic curves defined over $\mathbb{Q}$. Then it was thought for a while that the ranks were unbounded, by analogy with what happens over function fields over finite fields (Shafarevich & Tate, 1967; Ulmer, 2002); see for example Tate’s lecture on the arithmetic of elliptic curves at Oslo [55] or the Bourbaki talk by Colmez on the $p$-adic Birch & Swinnerton-Dyer conjecture [56]. Noam Elkies has constructed infinitely many $E$ for which the rank of $E(\mathbb{Q})$ is $> 18$, and one $E$ for which $\text{rk}_Z(E(\mathbb{Q})) > 27$.

By constructing a linear algebraic model for the rank and other algebraic invariants of elliptic curves, the consensus seems to have returned to boundedness of late:

---

**Theorem**

Let $E$ be an elliptic curve over $\mathbb{Q}$. If the Shafarevich-Tate conjecture (Conjecture 26) holds for $E$ and $\text{rk}_Z(E(\mathbb{Q})) = 0, 1$, then $\text{ord}_{s=1} L(E, s) = \text{rk}_Z(E(\mathbb{Q}))$.

See also the preprint by Chan-Ho Kim [54]. The interested reader is referred to the Princeton talks of Burungale [51] and Tian [52]. Note that according to Conjecture 48, 100% of elliptic curves over $\mathbb{Q}$ have $\text{rk}_Z(E(\mathbb{Q})) = 0, 1$.
Conjecture 54. — There are only finitely many elliptic curves defined over \( \mathbb{Q} \) for which \( \text{rk}_\mathbb{Z} E(\mathbb{Q}) > 21 \).

There is in fact a whole series of conjectures inspired by a linear-algebraic model of elliptic curve invariants. For example, although there are infinitely many elliptic curves over \( \mathbb{Q} \) such that \( E(\mathbb{Q}) \) has a point of order 12, only finitely many of these should have rank \( > 2 \), conjecturally. See for example Bjorn Poonen’s lecture at Rio de Janeiro (text [59], video [60]) and his presentation at Luminy [61]. For more speculation on the rank, see the online talk by Noam Elkies [62].

Acknowledgements

I warmly thank Daniel Kriz and Ashay Burungale for clarifying certain points, Xavier Xarles for correcting a link, and Peter Scholze for his comments.

Suggested reading and viewing


[20] T Gee, youtu.be/6eZQu120A80


[27] S Zhang, carmin.tv/s/6433


[33] A Smith, carmin.tv/s/5462


[42] M Bhargava, youtu.be/90WkXaDm-s


[44] A Wiles, youtu.be/1WYlP-B9nPI

[45] C Skinner, heilbronn.ac.uk/video-archive/

[46] B Gross, youtu.be/EXqRiaZYeNI.


[52] Y Tian, youtu.be/Ed6WjXXnuc


[55] J Tate, youtu.be/8tiVaAldqX0


[60] B Poonen, youtu.be/yxzy2K1UX9Y

[61] B Poonen, carmin.tv/s/5784


The site gdz.sub.uni-goettingen.de has the papers [1], [15], [10], [11], [5], [3], [16], [7] and [2]. The site www.numdam.org has the paper [8]. All papers by Langlands are available at his website.
The Unexpected Appearance of Pi in Diverse Problems

Rajendra Bhatia

There is a famous essay titled The Unreasonable Effectiveness of Mathematics in the Natural Sciences by the renowned physicist Eugene P Wigner. The essay opens with the paragraph:

There is a story about two friends, who were classmates in high school, talking about their jobs. One of them became a statistician and was working on population trends. He showed a reprint to his former classmate. The reprint started, as usual, with the Gaussian distribution and the statistician explained to his former classmate the meaning of the symbols for the actual population, for the average population, and so on. His classmate was a bit incredulous and was not quite sure whether the statistician was pulling his leg. “How can you know that?” was his query. “And what is this symbol here?” “Oh,” said the statistician, “this is \( \pi \).” “What is that?” “The ratio of the circumference of the circle to its diameter.” “Well, now you are pushing your joke too far,” said the classmate, “surely the population has nothing to do with the circumference of the circle.”

Wigner then goes on to discuss the surprisingly powerful role mathematics plays in the study of nature. I have quoted this para for making a small point. The number \( \pi \), the ratio of the circumference of the circle to its diameter, appears in many contexts that seem to have no connection with diameters, areas, or volumes. One such problem that I discuss here concerns properties of natural numbers.

Every student of calculus learns the Wallis product formula

\[
\frac{\pi}{2} = \frac{2 \cdot 4 \cdot 4 \cdot 6 \cdot 6 \cdot 8 \cdot 8}{1 \cdot 3 \cdot 3 \cdot 5 \cdot 5 \cdot 7 \cdot 7 \cdot 9} \cdots
\]

(1)

On the right hand side there is an infinite product and this is to be interpreted as

\[
\lim_{n \to \infty} \frac{2 \cdot 4 \cdot 4 \cdot 6 \cdot 6 \cdot 8 \cdot 8 \cdots}{1 \cdot 3 \cdot 3 \cdot 5 \cdot 5 \cdot 7 \cdot 7 \cdot 9} \cdot \frac{2n}{2n - 1} \cdot \frac{2n}{2n + 1}
\]

(2)

This formula attributed to John Wallis (1616-1703) is remarkable for several reasons. It is, perhaps, the first occurrence of an infinite product in mathematics. And it connects \( \pi \) with natural numbers. The formula has a simple proof. Let

\[
I_n = \int_0^{\pi/2} (\sin x)^n \, dx.
\]

References

* A part of a talk given at the Jesus and Mary College on 27th November, 2002 under the Delhi Mathematics Awareness Program sponsored by the National Board for Higher Mathematics. Reproduced with permission from Resonance, Vol.8, No.6, pp.34–43, 2003.

Keywords

Pi, Riemann zeta function, probability distribution, Fourier expansion.
Integrate by parts to get the recurrence formula
\[ I_n = \frac{n-1}{n} I_{n-2}. \]

The sequence \( I_n \) is a monotonically decreasing sequence of positive numbers. This and the recurrence formula show that
\[ 1 < \frac{I_n}{I_{n+1}} < 1 + \frac{1}{n}. \]

So \( I_n/I_{n+1} \) tends to 1 as \( n \to \infty \). Note that \( I_0 = \pi/2 \) and \( I_1 = 1 \). The recurrence formula can be used to get
\[ \frac{I_{2n+1}}{I_{2n}} = \frac{2 \cdot 4 \cdot 4 \cdot \ldots \cdot 2n}{1 \cdot 3 \cdot 3 \cdot \ldots \cdot 2n - 1} \frac{2}{2n + 1}. \]

Taking the limit as \( n \to \infty \) we get (1).

Many infinite sums involving natural numbers lead to \( \pi \). One that we need for our discussion is a famous formula due to Leonhard Euler (1707-1783)
\[ \frac{\pi^2}{6} = \frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \cdots \quad (3) \]

A (natural) number is said to be square-free if in its prime factoring no factor occurs more than once. Thus \( 70 = 2 \times 5 \times 7 \) is a square-free number while \( 12 = 2 \times 2 \times 3 \) is not.

Many problems in number theory are questions about the distribution of various special kinds of numbers among all numbers. Thus we may ask:

What is the proportion of square-free numbers among all numbers?

Or

If a number is picked at random what is the probability that it is square-free?

Now, randomness is a tricky notion and this question needs more careful formulation. However, let us ignore that for the time being. It is reasonable to believe that if we pick a number at random it is as likely to be odd as it is even. This is because in the list
\[ 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, \ldots \]
every alternate number is even. In the same way every third number is a multiple of 3, every fourth number is a multiple of 4, and so on. Thus the probability that a randomly picked number is a multiple of \( k \) is \( 1/k \), and the probability that it is not a multiple of \( k \) is \( 1 - 1/k \).

Let \( p_1, p_2, p_3, \ldots \) be the sequence of prime numbers. Let \( n \) be a randomly chosen number. For each prime \( p \) the probability that \( p^2 \) is not a factor of \( n \) is \( 1 - 1/p^2 \). Given two primes \( p_j \) and \( p_k \), what is the probability that neither \( p_j \) nor \( p_k \) is a factor of \( n \)? Again from probabilistic reasoning we know that the probability of the simultaneous
occurrence of two independent events is the product of their individual probabilities. (Thus the probability of getting two consecutive heads when a coin is tossed twice is 1/4.) Whether \( n \) has a factor \( p_j^2 \) has no bearing on its having \( p_k^2 \) as a factor. Thus the probability that neither \( p_j^2 \) nor \( p_k^2 \) is a factor of \( n \) is \( (1 - 1/p_j^2)(1 - 1/p_k^2) \). Extending this reasoning one sees that the probability of \( n \) being square free is the infinite product

\[
\prod_{j=1}^{\infty} \left(1 - \frac{1}{p_j^2}\right).
\]

(4)

There is a connection between this product and the series in (3). It is convenient to introduce here a famous object called the Riemann zeta function. This is defined by the series

\[
\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}.
\]

(5)

This series surely converges for all real numbers \( s > 1 \). Let us restrict ourselves to these values of \( s \), though the zeta function can be defined meaningfully for other complex numbers. The formula (3) can be written as

\[
\zeta(2) = \frac{\pi^2}{6}.
\]

(6)

The zeta function and prime numbers come together in the following theorem of Euler.

**Theorem.** For all \( s > 1 \)

\[
\zeta(s) = \prod_{n=1}^{\infty} \frac{1}{1 - p_n^{-s}}.
\]

(7)

**Proof.** Fix an \( N \), and use the geometric series expansion of \( \frac{1}{1-x} \) to get

\[
\prod_{n=1}^{N} \frac{1}{1 - p_n^{-s}} = \prod_{n=1}^{N} \sum_{m=0}^{\infty} p_n^{-ms}.
\]

(8)

The last expression is equal to

\[
\sum_{j=1}^{\infty} \frac{1}{n_j^s},
\]

where \( n_1, n_2, \ldots \) is an enumeration of those numbers that have \( p_1, p_2, \ldots, p_N \) as their only prime factors. As \( N \to \infty \), the sequence \( \{n_j\} \) expands to include all natural numbers. This proves the theorem.

As a consequence the product (4) has the value \( 6/\pi^2 \). This is the probability that a number picked at random is square-free.

This is one more situation where the number \( \pi \) has made an appearance quite unexpectedly. Our main point has been made; several interesting sidelines remain.

First note that our argument shows that if we pick a number \( n \) at random, then the probability that it has no prime factor with multiplicity \( k \) is \( 1/\zeta(k) \).

The probability that a number picked at random is square free is \( 6/\pi^2 \).
With a little thinking one can see that the probability that two numbers picked at random are coprime is $6/\pi^2$. (This problem is equivalent to the one we have been discussing.)

There is another interesting way of looking at this problem. Let $\mathbb{Z}^2$ be the collection of all points in the plane whose coordinates are integers. This is called the integer lattice. If the line segment joining the origin $(0,0)$ to a point $(m,n)$ does not pass through any other lattice point we say that the point $(m,n)$ can be seen from the origin. For example, the point $(1, -1)$ can be seen from the origin but the point $(2, -2)$ cannot be seen. Among all lattice points what is the proportion of those that can be seen from the origin? The answer, again, is $6/\pi^2$. The proof of this is left to the reader.

The argument used in proving the Theorem above can be modified to give a proof of the fact that there are infinitely many prime numbers. The probability that a randomly picked number from the set $\{1, 2, \ldots, N\}$ is 1 goes to zero as $N$ becomes large. So the product $\prod_p \left(1 - 1/p\right)$ where $p$ varies over all primes is smaller than any positive number. This would not be possible if there were only a finitely many factors in the product.

The number $\pi$ entered the picture via the formula (3). How does one prove it? Several proofs are known. The daring ‘proof’ first given by Euler goes as follows.

Let $\alpha_1, \alpha_2, \ldots$ be the roots of the polynomial equation $a_0 + a_1 x + a_2 x^2 + \cdots + a_m x^m = 0$. Then

$$\sum \frac{1}{\alpha_i} = \frac{-a_1}{a_0}.$$ 

We can write

$$\cos \sqrt{x} = 1 - \frac{x}{2} + \frac{x^2}{24} + \cdots.$$ 

This is a ‘polynomial of infinite degree’, and the roots of $\cos \sqrt{x} = 0$ are

$$\frac{(2n + 1)^2 \pi^2}{4}, \quad n = 0, 1, 2, \ldots.$$ 

Hence

$$\sum_{n=0}^{\infty} \frac{1}{(2n + 1)^2} = \frac{\pi^2}{8}. \quad (9)$$

The formula (3) follows from this easily.

Surely this argument has flaws. They can all be removed! With the notions of uniform convergence and $\epsilon - \delta$ arguments, we can prove formulas like

$$\frac{\sin x}{x} = \prod_{n=1}^{\infty} \left(1 - \frac{x^2}{n^2 \pi^2}\right), \quad (10)$$

from which the formulas (1) and (3) can be derived by simple manipulations. Finding the sum of the series (3) was one of the early major triumphs of Euler. He was aware
that the argument we have described above is open to several criticisms. So he gave another proof that goes as follows.

\[
\frac{\pi^2}{8} = \frac{(\arcsin 1)^2}{2} = \int_0^1 \frac{\arcsin x}{\sqrt{1-x^2}} dx
\]

\[
= \int_0^1 \frac{1}{\sqrt{1-x^2}} \left[ x + \sum_{n=1}^{\infty} \frac{1 \cdot 3 \cdots (2n-1)}{2 \cdot 4 \cdots 2n} \frac{x^{2n+1}}{2n+1} \right] dx
\]

\[
= 1 + \sum_{n=1}^{\infty} \frac{1 \cdot 3 \cdots (2n-1)}{2 \cdot 4 \cdots 2n} \frac{2n(2n-2) \cdots 2}{(2n+1)(2n+1)(2n-1) \cdots 3}
\]

\[
= \sum_{n=0}^{\infty} \frac{1}{(2n+1)^2}.
\]

Following the ideas of his first proof Euler showed that \(\zeta(2m)\) is \(\pi^{2m}\) multiplied by a rational number. Thus for example,

\[
\zeta(4) = \frac{\pi^4}{90}, \quad \zeta(6) = \frac{\pi^6}{945}.
\]  

(11)

Neither Euler, nor anyone else in three centuries after him, has found much about the values of \(\zeta(k)\) when \(k\) is an odd integer. In 1978 R Apéry showed that \(\zeta(3)\) is an irrational number. Even this much is not known about \(\zeta(5)\).

Another general method for finding sums like (3) and (11) goes via Fourier series. If \(f\) is a continuous function on \([-\pi, \pi]\) and \(f(x) = \sum_{n=-\infty}^{\infty} a_n e^{inx}\) its Fourier expansion, then

\[
\sum_{n=-\infty}^{\infty} |a_n|^2 = \int_{-\pi}^{\pi} |f(x)|^2 dx.
\]  

(12)

The method depends on recognising the summands of a particular series as coefficient of the Fourier series of a particular function \(f\) and then computing the integral in (12).

Having seen expression like (10) and (12) one is no longer surprised that \(\zeta(2m)\) involves \(\pi\) in some way.

Finally, let us briefly discuss some issues related to ‘picking a natural number at random’.

Two standard examples of completely random phenomena are tossing of a coin and throwing of a dice. In the first case we have two, and in the second case six, equally likely outcomes. The ‘sample space’ in the first case is the set \{1, 2\} (representing the two outcomes head and tail) and in the second case it is the set \{1, 2, \ldots, 6\}. One can imagine an experiment with \(N\) equally likely outcomes \{1, 2, \ldots, \(N\)\}.

The uniform probability distribution on the set \(X = \{1, 2, \ldots, N\}\) is the function that
assigns to each subset $E$ of $X$ values according to the following rules

$$
\begin{align*}
\mu(\{j\}) &= \mu(\{k\}) \quad \text{for all } j, k, \\
\mu(E) &= \sum_{j \in E} \mu(\{j\}), \\
\mu(X) &= 1.
\end{align*}
$$

Note that these three conditions imply that $\mu(\{j\}) = 1/N$ for all $j$. This is a model for a random phenomenon (like in some games of chance) with $N$ equally likely outcomes.

It is clear that if $X$ is replaced by the set $\mathbb{N}$ of all natural numbers, then no function satisfying the three conditions (13)-(15) exists. So, if ‘picking an element of $\mathbb{N}$ at random’ means assigning each of its elements $j$ an equal ‘probability’ we run into a problem. However, there is a way to get around this.

Let $X = \{1, 2, \ldots, N\}$ and let $E$ be the set of even numbers in $X$. If $N$ is even, then $\mu(E) = 1/2$. But if $N = 2m + 1$ is odd, then $\mu(E) = m/(2m + 1)$. This is less than $1/2$, but gets very close to $1/2$ for large $N$. In this sense a number picked at random is as likely to be even as odd.

In the same spirit we can prove the following.

For every $\varepsilon > 0$, there exists a number $N$, such that if $\mu$ is the uniform probability distribution on the set $X = \{1, 2, \ldots, N\}$ and $E$ is the set of square-free numbers in $X$, then

$$
\frac{6}{\pi^2} < \mu(E) < \frac{6}{\pi^2} + \varepsilon.
$$

The reader may prove this using the following observations. We know that

$$
\prod_{j=1}^{\infty} \left(1 - \frac{1}{p_j^2}\right) = \frac{6}{\pi^2}.
$$

The factors in this product are smaller than 1. So, the sequence

$$
\prod_{j=1}^{M} \left(1 - \frac{1}{p_j^2}\right), \quad M = 1, 2, \ldots
$$

decreases to its limit. Choose an $M$ such that

$$
\frac{6}{\pi^2} < \prod_{j=1}^{M} \left(1 - \frac{1}{p_j^2}\right) < \frac{6}{\pi^2} + \varepsilon.
$$

and let $N = \prod_{j=1}^{M} p_j^2$.

A (non-uniform) probability distribution on $X$ is a function $\mu$ that satisfies the conditions (14)-(15) but not (necessarily) the condition (13). There is nothing that
prevents the existence of such a distribution on \( \mathbb{N} \). Any series with non negative terms and with sum 1 gives such a distribution. In particular if we set

\[ \mu(\{j\}) = \frac{6}{\pi^2} \frac{1}{j^2}, \quad j = 1, 2, \ldots, \] (16)

then \( \mu \) is a probability distribution on \( \mathbb{N} \). This assigns different probabilities to different elements of \( \mathbb{N} \). The reader may like to interpret and prove the following statement.

The probability that two natural numbers picked at random have \( j \) as their greatest common divisor is \( \mu(\{j\}) \) as defined by (16).

Suggested Reading

[1] G H Hardy and E M Wright, *An Introduction to the Theory of Numbers*, Oxford University Press, 1959. See Chapter VIII, and in particular Theorems 332 and 333. The latter theorem attributed to Gegenbauer (1885) says that if \( Q(x) \) is the number of square-free numbers not exceeding \( x \), then

\[ Q(x) = \frac{6x}{\pi^2} + O(\sqrt{x}). \]

Here \( O(\sqrt{x}) \) represents a function whose absolute value is bounded by \( Ax \sqrt{x} \) for some constant \( A \).

Use this formula, with a computer program for testing whether a number is square-free, to obtain the value of \( \pi \) up to the third decimal place.

[2] P J Davis and R Hersch, *The Mathematical Experience*, Birkhauser, 1981. We have borrowed our main argument from the discussion on page 366 here. This occurs in a chapter titled *The Riemann Hypothesis* where the authors present an argument showing that this most famous open problem in mathematics has an affirmative solution with probability one.


[6] N D Baruah, B C Berndt and H H Chan, Ramanujan’s series for \( \frac{1}{\pi} \): a survey, *American Mathematical Monthly*, 116(2009), 567–587. This article explains the mysteries behind several astounding formulas connecting \( \pi \) and integers discovered by Srinivasa Ramanujan. For example,

\[ \frac{1}{\pi} = \frac{\sqrt{8}}{9801} \sum_{n=0}^{\infty} \frac{(4n)!}{(n!)^4} \frac{26390n + 1103}{396^{4n}}. \]

Investigating the Primes *

Kaneenika Sinha

The aim of this expository article is to introduce the reader to some of the fundamental milestones in the study of prime numbers across several centuries. Among the important developments in the study of prime numbers, we review the history of the prime number theorem, the Riemann zeta function (in relation to prime number theory), and some recent investigations into spacings between consecutive primes. We also present an important application of prime numbers in safe data transmission, namely the ‘RSA public key cryptosystem’.

Introduction

In 300 BCE, Euclid of Alexandria wrote a series of 13 volumes under the title *Elements*. These volumes contain a systematic presentation of several mathematical concepts through precise definitions, theorems, and their deductive proofs. They form the structural foundation of logic and mathematics as we study it today; in fact, much of what we learn in high school mathematics today goes back to the contents of these volumes. The topic of this article is a fundamental notion in Book 7 of *Elements*, which has fascinated humanity for the last 2300 years, namely prime numbers.

Euclid defined a prime number as “that which is measured by the unit alone”. In modern parlance, a prime number is a natural number \( n > 1 \) which is not divisible by any natural number other than 1 and itself. School textbooks typically make a passing reference to the notion of prime numbers, with some computational exercises on deriving prime factors of ‘large’ numbers (which, by real-world standards, are not that large!).

It is only in college that a mathematics student revisits prime numbers in some detail. We learn, for instance, the fundamental theorem of arithmetic which states that “every natural number other than 1 is either a prime or can be uniquely written as a product of primes”. We also learn that there are infinitely many primes. Both these theorems go back to Euclid’s *Elements*. At this stage, a student is naturally led to ask further questions about primes. How are primes distributed on the number line? Are they distributed in some sort of uniform pattern or not? Do they become sparse as one proceeds on the number line?

One can also ask some explicit questions. How do we recognize a prime when we see it? If we are given a large number, how long could it take to determine whether it is prime? If we need a large prime for some reason, how quickly can we generate it? If a number is not a prime, do we have nice (and efficient) methods to break a number down into its prime factors?


Keywords Prime numbers, sieve methods, Riemann zeta function, cryptography.
a number is not a prime, do we have nice (and efficient) methods to break a number down into its prime factors? Such questions and more about prime numbers have interested and inspired several seekers of knowledge over the last three millennia.

Answers to these questions are not immediate and in many cases, are known partially after several centuries of deep thought. Advancement in the study of prime numbers has less to do with answers and more to do with questions; questions that progressively reveal what prime numbers meant to seekers at every stage. The journey from Euclid’s discovery that there are infinitely many primes to the sophisticated investigation of prime numbers in the twenty-first century has many important milestones, and we hope to describe some of them in this article. This article does not claim to contain a complete history of primes. Instead, we review some key developments and provide useful references with the hope that an interested student will explore the world of primes in greater detail.

Organization of the Article

In Section 1, we summarise some classical developments that enhance our perspective of prime numbers. This section culminates in a fundamental 1859 article of Riemann which provided a ‘vision document’ to answer questions about prime numbers with the help of what is called the Riemann zeta function. Section 2 takes forward the study of distribution of primes and describes questions about spacings between prime numbers. Many interesting developments in this aspect have taken place within the last 15 years, and we describe some of them. Finally, in Section 3, we shift gears and turn towards an important application of prime numbers – security in our online transactions. These applications, developed in the 20th century, use properties of primes that were discovered more than 200 years ago.

1. Prime Counting: From Arithmetic to Analysis

In this section, we describe four important milestones, which take us from Euclid’s demonstration of the infinity of primes to a more refined study of primes using modern tools of mathematics. These also mark the conversion of arithmetic counting questions about prime numbers into the language of real and complex analysis. This is the language in which various problems about primes are currently studied.

Roughly speaking, the four developments that we touch upon in this section are as follows:

1. The sieve of Eratosthenes: This refers to an idea described by the Greek scholar Eratosthenes in the third century BCE to determine if a number is prime. This method is more efficient than the trivial method of checking divisibility by every number smaller than a given number. All the way until the 19th century, progressively larger records of prime numbers were built by several people using this sieve and its variants.
This is described in Section 1.1.

2. **Gauss and the prime counting function**: A major shift in the study of primes occurred when the German teenager Gauss, with a penchant for observing patterns in extensive data, made a conjecture in the 1790s about the asymptotic density of prime numbers. A big chunk of modern number theory was developed to understand and prove the conjecture of Gauss. This conjecture, also known as the prime number theorem, forms the content of Section 1.2.

3. **Chebyshev and the smooth analogue of prime number theorem**: In the early 1850s, Chebyshev defined some functions which ‘smoothen’ out the prime counting function and make way for the application of calculus to study what was heretofore seen as an arithmetic problem. Developments around this theme are described in Section 1.3.

4. **The zeta function of Riemann**: The introduction and study of zeta functions mark a very important development in the prime number theory. A breakthrough article written in 1859 by Riemann on this topic became the foundation for much of the number theory as we know it today. A brief overview of this function and relation to the prime number theorem is provided in Section 1.4.

### 1.1 Sieves: The Earliest Tools to Capture Primes

How does one determine if a number $n$ is prime? The first method that comes to mind to determine primality is, of course, trial division. We simply attempt division by all the numbers between 2 and $n - 1$, and if none of them divides $n$, we declare it a prime. But, in this method, the number of steps needed is equal to the size of $n$. Some time in the third century BCE, the Greek scholar Eratosthenes had a clever idea. If $n$ is not a prime number, then it must have a factor not bigger than the square root of $n$. Therefore, to test the primality of $n$, it would be sufficient to check the divisibility by all the primes less than or equal to the square root of $n$. Take, for example, $n = 101$. We just need to check if $n$ is divisible by all the primes up to 10, that is, 2, 3, 5 and 7. This reduces the number of steps to check the primality of 101 from 99 to 4. What’s more, among the numbers 2 to 100, if we cross out all the multiples of 2, 3, 5 and 7, we will be left with a table of all the prime numbers up to 100. This method is famously called the ‘sieve of Eratosthenes’. Early records of primes were created through a systematic use of this sieve. Mathematicians used physical tools like adjustable sliders and stencils to locate and eliminate multiples of primes by this method (see [1] for a detailed description of how early tables of primes were created).

The simple sieve method of Eratosthenes has now evolved into more sophisticated sieves developed in the 19th century and afterward. However, the fundamental insight of Eratosthenes remained a primary tool for tabulating primes for a long time until the advent of computers.

Some time in the third century BCE, the Greek scholar Eratosthenes had a clever idea. If $n$ is not a prime number, then it must have a factor not bigger than the square root of $n$. 
1.2 Counting the Primes: Beyond Prime Tables

Several centuries after the observation of Eratosthenes, the study of prime numbers received a new impetus. In the 1790s, a young German teenager by the name of Johann Carl Friedrich Gauss, who would later earn the title of the “prince of mathematics”, wanted to understand the distribution of primes. He made some interesting guesses by looking carefully at existing records of prime numbers. He asked a fundamental question that brought a fresh perspective to the study of primes: Can we count the number of primes up to a number \( x \)? More precisely, can we approximate the function

\[
\pi(x) = \#\{p \leq x, p \text{ prime}\}.
\]

Based on existing data, Gauss conjectured that

\[
\lim_{x \to \infty} \frac{\pi(x) \log x}{x} = 1.
\]

That is, as \( x \) takes larger and larger values, the value of the prime counting function \( \pi(x) \) comes closer to \( \frac{x}{\log x} \), and the error margin between \( \pi(x) \) and \( x/\log x \) decreases. The journey of this conjecture to become a theorem (now famously known as the prime number theorem) was to take a 100 more years, and it gave birth to a beautiful amalgamation of analysis and number theory.

1.3 Logging the Primes: Interface Between Analysis and Number Theory

Questions in the theory of numbers often attract attention from amateurs as well as professional mathematicians. Many of these problems can be explained in simple language and look appealing for this reason. Unfortunately, the ‘elementary’ approach often does not go far, and it could take decades, or even centuries of concerted efforts before new light is shed on such problems. In some cases, this new light comes simply by reinterpreting the problem in a different language, which opens it up to the use of other tools.

In the case of the conjecture of Gauss on the prime counting function \( \pi(x) \), this new interpretation came when the French mathematician Joseph Bertrand went through the table of primes up to \( 3 \times 10^6 \) and conjectured in 1845 that for any natural number \( n \geq 2 \), one can find a prime number lying between \( n \) and \( 2n \). That is, \( \pi(2n) - \pi(n) > 0 \) for any \( n \geq 2 \). This conjecture was soon proved by the Russian mathematician Pafnuty Chebyshev in 1852. While studying the primes, Chebyshev modified the prime counting function by attaching logarithmic weights to the primes. That is, he replaced the function

\[
\pi(x) = \sum_{p \leq x} 1,
\]

by what is called the first Chebyshev function,

\[
\theta(x) = \sum_{p \leq x} \log p.
\]
Bertrand’s conjecture is equivalent to saying that \( \psi(2n) - \psi(n) > 0 \) for all \( n \geq 2 \). This logarithmically weighted function yields itself to readily available tools in analysis. Therefore, it is natural to try and state the conjecture of Gauss in terms of this function. In fact, using an important tool from an analysis that was independently discovered by Leonhard Euler and Colin Maclaurin in the 1730s, one can approximate the sums \( \sum_{p \leq x} \log p \) by integrals of appropriate functions. In terms of this new function, the conjecture of Gauss is equivalent to the assertion that

\[
\lim_{x \to \infty} \frac{\psi(x)}{x} = 1.
\]

Another related function is one in which we isolate, not just the primes, but all the prime powers less than or equal to \( x \). The second Chebyshev function is defined as

\[
\psi(x) = \sum_p \sum_{m \geq 1} \log p.
\]

For large values of \( x \), the functions \( \psi(x) \) and \( \theta(x) \) are very close to each other. Moreover, for reasons described in Section 1.4, \( \psi(x) \) is better suited to study the prime numbers. In fact, the prime number theorem is also equivalent to the assertion that

\[
\lim_{x \to \infty} \frac{\psi(x)}{x} = 1,
\]

and it is in this new form that the prime counting function was henceforth studied.

### 1.4 Prime Numbers and the Zeta Function of Euler and Riemann

The fundamental theorem of arithmetic, which says that every natural number \( n \geq 2 \) can be uniquely written as a product of prime powers, can be restated in terms of infinite series. Let us consider the infinite series, also known as Euler’s zeta function,

\[
\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s},
\]

which converges for a real number \( s > 1 \). By unique factorization of every \( n \) into a product of prime powers, we see that for \( s > 1 \),

\[
\sum_{n=1}^{\infty} \frac{1}{n^s} = \prod_{p \text{ prime}} \left( 1 + \frac{1}{p^s} + \frac{1}{p^{2s}} + \ldots + \frac{1}{p^{ms}} + \ldots \right) = \prod_{p \text{ prime}} \left( 1 - \frac{1}{p^s} \right)^{-1}.
\]

This analytic restatement of the fundamental theorem is due to Leonhard Euler and is called the ‘Euler product formula’. It gives an indication that the study of primes is linked to the behaviour of the series \( \zeta(s) \), which in turn can be studied using standard tools of calculus, since it is an (absolutely) convergent series in the interval \((1, \infty)\).

In fact, since \( \frac{1}{p^s} < 1 \) for \( s > 1 \), using the Taylor series expansion:

\[
-\log(1-x) = \sum_{n=1}^{\infty} \frac{x^n}{n} \text{ for } |x| < 1,
\]

...
we deduce,

$$\log \zeta(s) = -\sum_p \log \left(1 - \frac{1}{p^s}\right) = \sum_p \sum_{m=1}^{\infty} \frac{1}{mp^{ms}}.$$ 

Since \(\lim_{s \to 1^+} \zeta(s) = \infty\), we have, \(\lim_{s \to 1^+} \log \zeta(s) = \infty\). Thus,

$$\lim_{s \to 1^+} \log \zeta(s) = \lim_{s \to 1^+} \left(\sum_p \frac{1}{p^s} + \sum_p \sum_{m \geq 2} \frac{1}{mp^{ms}}\right) = \infty.$$ 

But, for \(s \geq 1\), the second series on the right hand side,

$$\sum_p \sum_{m \geq 2} \frac{1}{mp^{ms}} \leq \sum_p \sum_{m \geq 2} \frac{1}{p^m} = \sum_p \frac{1}{p(p-1)} < \infty,$$

Therefore,

$$\lim_{s \to 1^+} \sum_p \frac{1}{p^s} = \infty.$$ 

This shows that there are infinitely many primes; otherwise, the above limit would be finite.

The above calculation can be refined as follows. Differentiating the identity,

$$\log \zeta(s) = -\sum_p \log \left(1 - \frac{1}{p^s}\right),$$

we get,

$$-\frac{\zeta'(s)}{\zeta(s)} = -\frac{d}{ds} \log \zeta(s) = \sum_p \frac{d}{ds} \log \left(1 - \frac{1}{p^s}\right).$$

Further, applying the chain rule for differentiation,

$$\sum_p \frac{d}{ds} \log \left(1 - \frac{1}{p^s}\right) = \log \frac{p}{p^s - 1}.$$ 

Thus,

$$-\frac{\zeta'(s)}{\zeta(s)} = \sum_p \log \frac{p}{p^s - 1} = \sum_p \sum_{m=1}^{\infty} \log \frac{p}{p^{ms}}.$$ 

This explains the use of the function

$$\psi(x) = \sum_p \sum_{m \geq 1, \ p^m \leq x} \log p,$$

in the study of the prime number theorem. This is a prototypical example of an important point of view in number theory first elucidated by Dirichlet – the study of arithmetic sums \(\sum_{n \leq x} a(n)\) by investigating the analytic properties of the Dirichlet series,

$$\sum_{n=1}^{\infty} a(n) n^{-s}.$$
in the region where it converges. In the case of the prime numbers, the relevant function that needs to be investigated is \( \sum_{n \leq x} \Lambda(n) \) where the von Mangoldt function \( \Lambda(n) \) is defined as:

\[
\Lambda(n) := \begin{cases} 
\log p & \text{if } n = p^m, m \geq 1 \\
0 & \text{otherwise}
\end{cases}
\]

Note that the partial sum \( \sum_{n \leq x} \Lambda(n) \) equals \( \psi(x) \) and the associated series to be investigated is:

\[
\frac{\zeta'(s)}{\zeta(s)} = \sum_{n=1}^{\infty} \frac{\Lambda(n)}{n^s}, \quad \text{Re}(s) > 1.
\]

(1)

Henceforth, the zeta function was to be a primary tool to study prime numbers.

In 1859, Bernhard Riemann built on this point of view and wrote a famous nine-page article whose title can be translated into English as 'On the Number of Primes Less Than a Given Magnitude' [2]. The fundamental innovation in this work was to view Euler’s zeta function as a function of a complex variable. That is, the zeta function of Riemann is a complex-valued function defined as:

\[
\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s},
\]

for complex numbers \( s \) with the real part \( \text{Re}(s) > 1 \). Since this series converges absolutely in the region \( \{ s \in \mathbb{C} : \text{Re}(s) > 1 \} \) and uniformly in all compact subsets therein, it can be viewed as a complex-analytic function in this region.

Riemann’s paper describes an idea that lies at the heart of what is called the analytic number theory today, namely analytic continuation. He derives a function on a larger domain of complex numbers, which is equal to \( \sum_{n=1}^{\infty} \frac{1}{n^s} \) when \( \text{Re}(s) > 1 \). This new function is complex-analytic on all points of the complex plane except \( s = 1 \). This extension of the zeta function to \( \mathbb{C} \setminus \{1\} \), is called Riemann’s zeta function.

Riemann then goes on to provide a ‘vision document’ that has guided research on this topic in the last two centuries. Owing to (1), the function \( \psi(x) \) and therefore, the prime counting function \( \pi(x) \) are inherently linked with the analytic properties of the function \( \frac{\zeta'(s)}{\zeta(s)} \), and this naturally leads to questions about the complex zeroes of the zeta function, that is, those points \( s \) on the complex plane where \( \zeta(s) = 0 \). Riemann stated all these connections precisely and one of the key statements in his paper is the explicit formula, an explicit description of the relation between \( \psi(x) \), and the location of the complex zeroes of the zeta function.

An important observation of Riemann was that the prime number theorem is equivalent to showing that all non-trivial zeroes of \( \zeta(s) \), (that is, zeroes of the form \( s = \rho + it \) with the imaginary part \( t > 0 \)) must lie inside the critical strip, that is, the region \( \{ s \in \mathbb{C} : 0 < \rho < 1 \} \). This idea was independently used by Hadamard and de la Vallée Poussin to prove the prime number theorem in 1896.
Properties of the prime counting function can be translated into explicit complex-analytic properties of the zeta function.

Even more refined estimates about $\psi(x)$ and $\pi(x)$ can be made with better knowledge about the location of the zeta zeroes inside the critical strip. In fact, in his paper, Riemann makes a conjecture that all zeta zeroes lying in the critical strip must have real part $1/2$. This conjecture is famously known as the ‘Riemann hypothesis’ and remains unproven till date. It is in the list of Millennium Prize Problems announced by the Clay Mathematics Institute that promises an award of 1 million US dollars for a correct resolution.

Further explanation of Riemann's article will take us well beyond the scope of this article. Hence, we conclude this section by referring the reader to an excellent exposition [3] of Andrew Granville on the techniques and ideas in Riemann's work. For further reading, we also refer the reader to the 1974 book *Riemann’s Zeta Function* by H M Edwards [4] which contains an English translation of Riemann’s original German paper and a chapter-wise description of its contents.

2. Gaps Between Primes

In this section, we focus on another statistical aspect of the distribution of prime numbers, namely, gaps between consecutive primes. This is an exciting topic in the study of prime numbers, and some of the most important advances in this theme have come about within the last 15 years.

How far apart can two consecutive primes be? For starters, we observe that for any natural number $n$,

$$n! + 2, n! + 3, n! + 4, \ldots n! + n,$$

is a string of consecutive composite numbers. That is, for any number $n \geq 2$, we can find consecutive primes which are apart by at least $n$ numbers. Thus, the gaps between consecutive primes can be arbitrarily large. What about small gaps? Do we have infinitely many pairs of primes with gaps below a fixed bound?


In this listing, we immediately observe that the gaps between consecutive primes seem to oscillate. For several pairs, we have a gap of 2, they gradually increase to 4, 6, 8, 10 and so on, but keep jumping back to 2. This pattern seems to repeat several times even in the limited list we have above. Similar observations for much larger sets of primes motivate the following questions.

1. As the primes grow larger and larger, do we have infinitely many pairs of consecutive primes with gap 2?
2. Does there come a stage beyond which the gaps between consecutive primes increase and do not fall back to smaller numbers?

3. Can we predict how large the gaps can be at any stage? How much can these gaps differ from the average?

The twin prime conjecture is the assertion that there are infinitely many pairs of consecutive primes with gap 2. This is a special case of a more general 1849 conjecture of the French mathematician Alphonse de Polignac. This general conjecture states that for any even number \( K \), there are infinitely many pairs of consecutive primes with gap \( K \). This conjecture, if proved, would immediately address the first two questions (yes to the first and no to the second). The third question ‘averages out’ the first two questions: What is the variation among gaps between consecutive primes with respect to the average gap? An underlying question behind the above questions is whether the prime numbers are distributed among the natural numbers in a discernible pattern.

To check the infinity of twin primes, an immediate ‘analytic’ idea that comes to mind (inspired by Euler’s use of zeta function) is to check if the series

\[
\sum_{p, \, p+2 \text{ prime}} \frac{1}{p}
\]

diverges. In 1919, Viggo Brun showed that this sum is finite, which, unfortunately, tells us nothing about the finiteness or infinity of twin primes.

Mathematicians continued to grapple with the above questions in various ways. It was only in 2005 that three mathematicians – Dan Goldston, János Pintz and Cem Yildirim – made a remarkable observation that brought us a little closer to the twin prime conjecture. They showed that there are infinitely many pairs of consecutive primes with gap arbitrarily small relative to the average gaps between consecutive primes. The question that now remained was if we could have infinitely many pairs with the absolutely smallest possible gap, namely, 2.

In 2013, Yitang Zhang made a giant stride in this direction. He announced that there is a number \( N \) less than 70 million such that there are infinitely many pairs of consecutive primes with gap \( N \). At first sight (to the layman), 70 million may come across as a ridiculously large bound. But that is not how the experts saw Zhang’s theorem; instead, they recognized it as the first instance of a finite and explicit bound for infinitely many consecutive prime gaps. This is sufficient to see that the answer to the second question listed above is no. Stunned and impressed by Zhang’s work, the mathematical community challenged itself to reduce the gap from 70 million to 2. Thus started the international Polymath Project with several contributors, some of the notable names being James Maynard and Terence Tao. As of today, the bound has been reduced to 246. Is it just a matter of time before 246 can be taken down to 2? Or does 246 represent the limit that can be obtained by current knowledge? We do...
not know. It is, however, widely believed that further reduction of gaps will require completely new insights.

We now make some comments on how the third question mentioned above is interpreted. We start by recalling the prime number theorem, which predicts that the number of primes up to \( x \) is asymptotic to

\[
\frac{x}{\log x},
\]
as \( x \to \infty \). We now arrange the primes in an ascending order and let \( p_n \) denote the \( n \)-th prime number. We leave it as an exercise for the interested reader to deduce from the prime number theorem that:

\[
p_n \sim n \log n \text{ as } n \to \infty.
\]

That is,

\[
\lim_{n \to \infty} \frac{p_n}{n \log n} = 1.
\]

Furthermore, if we list out the gaps between consecutive primes up to \( p_{N+1} \) as:

\[
p_2 - p_1, p_3 - p_2, \ldots, p_{N+1} - p_N,
\]
the prime number theorem also tells us that the ‘expected’ or the ‘average’ gap

\[
p_{n+1} - p_n
\]
is asymptotically \( \log p_n \). That is,

\[
\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \frac{p_{n+1} - p_n}{\log p_n} = 1.
\]

For large values of \( N \), one wonders whether, one can predict the proportion of values among

\[
\left\{ \frac{p_{n+1} - p_n}{\log p_n} : n \leq N \right\}
\]
that lie in a fixed interval \([a, b]\) of positive real numbers as \( N \to \infty \). The answer is believed to be yes. In fact, it is conjectured that the ‘normalized’ gaps

\[
\frac{p_{n+1} - p_n}{\log p_n}
\]
can be modelled by the Poisson distribution. That is, for any two positive real numbers \( a < b \),

\[
\lim_{N \to \infty} \frac{1}{N} \# \left\{ 1 \leq n \leq N : a \leq \frac{p_{n+1} - p_n}{\log p_n} \leq b \right\} = \int_{a}^{b} e^{-t} dt.
\]

An investigation of this conjecture and its implications is a foundational theme in the subject of probabilistic number theory. We are nowhere near a resolution of this conjecture, but once again, we see how a difficult question about twin primes has
evolved into a related question about the distribution of all consecutive prime gaps. In recent developments, building upon the ideas of Riemann, connections have also been made between the distribution of gaps among consecutive primes and gaps among consecutive zeroes $s$ of the Riemann zeta function lying on the line $\text{Re}(s) = 1/2$.

A comprehensive and detailed discussion of the probabilistic models that can predict the distribution of gaps between consecutive primes is present in the expository article [5] of K Soundararajan. We also refer the interested reader to the survey article [6] of M R Murty for an explanation of the recent developments around the twin prime problem, particularly the strategy of Zhang and how it was improved upon by Maynard and Tao.

3. Big Prime Numbers and Big Secrets

The study of prime numbers is not all about ‘data analysis’ of records of prime numbers. Prime numbers also have a very important use for us in the real world. They help us to transfer sensitive information over the internet between valid parties without being intercepted by an unauthorised third party. We describe this application below.

To begin with, imagine walking on the hour points of a clock. You may walk millions of steps, and yet, you will find yourself among the 12 points. In effect, therefore, the ‘net’ result of your walk is between 1 and 12. This is an example of a circular form of counting called ‘counting modulo 12’. One can, of course, replace 12 with any number $N$ and count modulo $N$. In this sense, each time we walk over $N$ points on an $N$-hour clock, we find ourselves at the same point where we started. That is, any multiple of $N$ steps amounts to a walk with zero change.

In October 1640, the French lawyer-mathematician Pierre de Fermat made an observation in a letter to a friend.

*Let $p$ be a prime number. If we take any number $A$ of our choice and walk $A^p$ steps on a $p$-hour clock, we will find ourselves at the same point as if we had merely walked $A$ steps. For example, on a 13-hour clock, whether you walk 11 steps or a whopping $11^{13}$ steps from the same starting point, you will find yourself at the same end point.*

In mathematical language,

$$A^p \equiv A \pmod{p}.$$


This theorem, at the heart of number theory, is famously known as ‘Fermat’s little theorem’, even though it was proved almost a hundred years later by Leonhard Euler.

**Theorem.** Let $N$ be a natural number and $A$ be an integer such that $A$ and $N$ are coprime, that is, the greatest common divisor of $A$ and $N$ is 1. Let $\phi(N)$ denote the number of integers lying between 1 and $N$ which are coprime to $N$. Prime numbers help in the safe digital transfer of information between two parties.
The RSA cryptosystem depends on the fundamental Fermat–Euler theorem discovered three centuries ago.

**coprime to \( N \). Then,**

\[
A^{\phi(N)} \equiv 1 \pmod{N}.
\]

**Thus,**

\[
A^{\phi(N)+1} \equiv A \pmod{N}.
\]

Around two hundred and forty years later, in 1978, Ron Rivest, Adi Shamir, and Leonard Adleman, three computer scientists at the Massachusetts Institute of Technology (USA), exploited the Fermat–Euler theorem and developed a method to share information between two parties safely. Typically, when a message is sent from a sender to a receiver, the sender has to disguise or ‘encrypt’ the message and the receiver has to decipher or ‘decrypt’ it. The method (or key) for the encryption and decryption is discussed privately between these parties or is shared between them through other safe sources. The RSA method (named after its creators) is set up in such a way that the key to encrypt a message is made public. While anyone can encrypt a message, sign it and send it across, only the intended receiver has the knowledge to decrypt a message and deduce whether it comes from a genuine sender. Such a system is called a public key cryptosystem.

The method is easy to understand and we attempt to present it below in a simple form.

1. Instead of a prime \( p \), an online merchant takes a semi-prime number \( N \), which is a product of two primes \( p \) and \( q \), roughly of equal size. The reader can quickly check that for this chosen \( N \), \( \phi(N) \) equals \( (p - 1)(q - 1) \). Thus, the Fermat–Euler theorem tells us that for any number \( A \),

\[
A^{(p-1)(q-1)+1} \equiv A \pmod{pq}.
\]

The merchant then chooses two numbers \( E \) and \( D \), so that \( (p - 1)(q - 1) + 1 = ED \). The merchant makes \( N \) and \( E \) public and keeps \( D \) private.

2. You, the buyer, will typically take your credit card number \( A \) (which will be smaller than \( N \)) and encrypt it as the net value of \( A^E \pmod{N} \). You don’t do it yourself, of course. You enter the number and the website does it for you since the encryption key \( E \) is public.

3. The number \( D \) is not public. It is known privately to the merchant. On receiving the encrypted message \( A^E \pmod{N} \), the merchant raises it again to the power \( D \) and calculates \( A^{ED} \pmod{N} \). By basic congruence arithmetic and by the Fermat–Euler theorem,

\[
(A^{(p-1)(q-1)+1})^D \equiv A^{ED} \equiv A^{(p-1)(q-1)+1} \pmod{N} \equiv A \pmod{N}.
\]

This gives back the number \( A \) to the merchant. The credit card number (or intended message) therefore gets decrypted back to \( A \).
But, you may ask: If $N$ and $E$ are public, can $D$ really stay secret? The answer is an emphatic yes.

The safety of the RSA algorithm lies in the fact that it is very, very difficult to break down or to factorise a large number $N$ into its prime factors $p$ and $q$. Thus, even if the ‘encryption’ key $E$ is public, to find out $D$, an unscrupulous third party will have to know the value of $\phi(N) = (p - 1)(q - 1)$, and this is not possible unless one is able to factor $N$ into its constituent primes $p$ and $q$. This makes it practically infeasible to find out the decryption key $D$ (which only the intended receiver knows) within a reasonable time.

In their 1978 paper [7] on the RSA algorithm, the authors estimated that the computing resources of those days could take, for example, 74 years to factor a number with 100 decimal digits and 38,000 000 000 years to factor a number with 200 decimal digits! Today, we have more sophisticated computing resources and methods to factor numbers of this size. These methods originate from abstract ideas in mathematics like the number field sieve and elliptic curves. But, we now use even larger numbers for encryption with unfeasible factoring times using current technology.

A lot of research has gone into factoring large numbers. For example, between 1991 and 2007, the RSA Laboratories, a company formed by Rivest, Shamir, and Adleman, ran the factoring challenge. They listed out semi-primes of various sizes with cash prizes for those who could factor them. One of the numbers in their list, RSA-768, a 232-digit number of bit length 768 was factorised as late as December 2009 after two years of work on several computers. Numbers of higher sizes on this list still remain unfactored.

The fact that it takes much longer to factorise numbers than to generate primes and multiply them keeps our information safe. In fact, the interested reader can go to secure websites (for example their email provider) and check the digital certificates. These clearly state the encryption algorithm (most likely, the RSA algorithm), the public key as well as the key sizes. The certificate also mentions its expiry dates. This is to ensure that the keys are updated before an unscrupulous party gets hold of an existing private key.

**Conclusion**

Prime numbers have fascinated us for a long time. We saw in this article that developments in this study are rather abruptly distributed over the last three millennia. Sometimes, centuries would elapse before any new development would take place. On the other hand, occasionally there would be intense periods of activity when in a matter of decades, breakthrough ideas would appear and create a major impact on prime number theory. In retrospect, some of these ideas may look simple to us, but to discover them from scratch requires great intellectual power and ingenuity in thinking. Sometimes, ideas would be expressed in letters among friends and would
motivate people with fresh points of view to jump into the subject. Last, but not the least, even though the study of some of these ideas would appear to be theoretical in the short term, a few centuries down the line, they would lead to innovations with a major impact on how the world functions today, for example, Fermat’s little theorem and its use in safe internet transactions without which many of us cannot imagine a life. We hope that we could give a small flavour of some of these ideas in this article and hope that the interested reader will find at least one of them interesting enough to pursue in greater detail.

Suggested Reading


Balanced Number System*
Application to Mathematical Puzzles

Shobha Bagai

The article explores the application of binary and ternary number systems to three classical mathematical puzzles – weight problem of Bachet de Méziriac, binary numbers magic trick, and the detection of a counterfeit coin. The article further elucidates the generalization of these puzzles using the balanced number system.

1. Introduction

We have all grown up with the decimal system of numeration that uses the digits 0, 1, 2, . . . , 9. The use of the decimal number system is sometimes attributed to us having ten fingers (counting the thumb) on our hands. But imagine if we were in the animated world of Fred Flintstone or Simpsons where each of the leading characters has four fingers on each hand. Then absurd mathematical expressions like $42 + 36 = 100$ or $42 - 14 = 26$ would not look so absurd. These expressions would be correct if they are calculated in the octal base system. Though the use of numbers in base 10 is the most common in present times, numbers in different bases have also found various applications.

The digital world of computers, mobile phones, audio players, etc., use the binary system (base 2). The ternary numeral system (base 3) finds its application in analog logic where the output of a circuit is either low (grounded state), high, or open (high impedance). The way the Cantor set is constructed, a ternary system is useful for defining it. Soviet computers in the early days of computing incorporated the balanced ternary system that uses $-1, 0, 1$ instead of $0, 1, 2$. Roman numerals are in the biquinary system having 2 and 5 as the base. They have different symbols for 1, 5, 10, 50, 100, 500 and 1000. The octal and the hexadecimal systems are also being used in the computers. In the Academy Awards nominated movie Avatar, the humanoid species Na’vi employs an octal numeral system in their language. Incidentally, they are also shown to have four fingers on each of their hands.

It is not that a numeral system with a base different from 10 has been used only in recent times. Numbers in a base different from 10 have been practiced in ancient times too. While Mayan civilization used numbers in base 20, the Babylonian’s number system was a base 60 system. The duodecimal system or the dozenal system was popular in the seventeenth century. A carpenter’s ruler has 12 inches, grocers deal in dozens, there are 12 months in a year and each day is divided into two sets of 12 hours.


Keywords
Binary number system, ternary number system, balanced number system, weight problem of Bachet de Méziriac.
Promoting Science Education

Table 1. Representing 464 in base 2.

<table>
<thead>
<tr>
<th>2</th>
<th>464</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>232</td>
</tr>
<tr>
<td>2</td>
<td>116</td>
</tr>
<tr>
<td>2</td>
<td>58</td>
</tr>
<tr>
<td>2</td>
<td>29</td>
</tr>
<tr>
<td>2</td>
<td>14</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 2. Representing 464 in base 3.

<table>
<thead>
<tr>
<th>3</th>
<th>464</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>154</td>
</tr>
<tr>
<td>3</td>
<td>51</td>
</tr>
<tr>
<td>3</td>
<td>17</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 3. Representing 464 in balanced ternary system.

<table>
<thead>
<tr>
<th>3</th>
<th>464</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>155</td>
</tr>
<tr>
<td>3</td>
<td>52</td>
</tr>
<tr>
<td>3</td>
<td>17</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

Each. F E Andrews [1] lists out a number of advantages of using the duodecimal system in place of the decimal system.

This article deals with how to convert numbers in base 10 to balanced ternary system and application of the balanced ternary system to some mathematical puzzles. The article further explores how these puzzles may be generalized by introducing the balanced $n$-nary system.

2. Balanced $n$-nary System

D E Knuth in his monograph *The Art of Computer Programming* comments on the balanced ternary system being the prettiest number system of all. As mentioned above, the balanced ternary number system employs integers $-1, 0, 1$ in place of $0, 1, 2$. It is customary to represent a number in base 2 in terms of 0’s and 1’s and in base 3 in terms of 0’s, 1’s and 2’s. Any number in base 10 can be represented in base 2 (or base 3) by using the expanded notation. For example:

$$(464)_{10} \equiv 1 \times 2^5 + 1 \times 2^7 + 1 \times 2^6 + 1 \times 2^4 = (111010000)_2$$

$$(464)_{10} \equiv 1 \times 3^5 + 2 \times 3^4 + 2 \times 3^3 + 1 \times 3 + 2 \times 1 = (122012)_3.$$  

Another form of representing it is by long division where the rightmost column represents the remainder, and the middle row represents the quotient at each step (see Tables 1 and 2).

In ternary system, a non-negative number $n$ may be represented as $3k$ (if divisible by 3) or $3k + 1$ (if it leaves a remainder 1 on dividing by 3) or $3k + 2$ (if it leaves a remainder 2 on dividing by 3), for $k = 0, 1, 2, \ldots$. The number $3k + 2$ may also be represented as $3(k + 1) - 1$. So we could say that instead of leaving a remainder 2 when divided by 3 it leaves a remainder $-1$. Using this idea and performing long division we can have the solution shown in Table 3.

Therefore, the balanced ternary representation of 464 in base 3 is

$$(464)_{10} \equiv (1 - 10 - 11 - 1 - 1)_3.$$  

In expanded form this would represent

$$(464)_{10} \equiv 1 \times 3^6 - 1 \times 3^5 - 1 \times 3^3 + 1 \times 3^2 - 1 \times 3 - 1 \times 1.$$  

Another way of getting this notation from the usual expanded form is:

$$(464)_{10} = 1 \times 3^5 + 2 \times 3^4 + 2 \times 3^3 + 1 \times 3 + 2 \times 1$$

$$(464)_{10} = 1 \times 3^5 + (3^5 - 3^4) + (3^4 - 3^3) + 1 \times 3 + (3 - 1)$$

$$(464)_{10} = 2 \times 3^3 - 1 \times 3^3 + 2 \times 3 - 1$$

$$(464)_{10} = (3^6 - 3^5) - 1 \times 3^3 + (3^2 - 3) - 1$$

$$(464)_{10} = 1 \times 3^6 - 1 \times 3^5 - 1 \times 3^3 + 1 \times 3^2 - 1 \times 3 - 1.$$
The above calculation illustrates that the number 2 in base 3 can be written as,

\[ 2 = 1 \times 3 - 1 = (1, -1)_3. \]

This implies that if the remainder used is \(-1\) in place of 2, it is carried forward to the next place value. We may represent it as shown in Table 4.

Use of \(-1\) in the representation can become quite cumbersome. Suggested reading [2] represents the integer \(-1\) by \(T\) whereas [3] represent the ternary digits by \(N\) (for \(-1\)), \(Z\) (for 0) and \(P\) (for 1). Since the article shall deal with general balanced \(n\)-nary number system, the representation used will be \(L_i\) (if the integer is \(i\) places to the left of 0), 0, and \(R_i\) (if the integer is \(i\) places to the right of 0). Therefore, under this representation:

\[ (464)_{10} \equiv (R_1, L_1, 0, L_1, R_1, L_1)_{3}. \]

The concept of balanced number system may be extended for any positive integer \(n > 1\). If \(n\) is odd, say \(2m + 1, m = 1, 2, 3, \ldots\) one can express any natural number \(N\) in base \(2m + 1\) in terms of 0, \(L_i, R_i, i = 1, 2, \ldots, m\). As an example let us take another number (Table 5):

\[ (2565)_{10} \equiv (40230)_5 = (R_1, L_1, R_1, L_2, 0)_5. \]

If \(n\) is even, say \(2m, m = 1, 2, 3, \ldots\) one can express any natural number \(N\) in base \(2m\) in terms of 0, \(L_i, R_i, R_m\) (or \(L_m\)), \(i = 1, 2, \ldots, m - 1\). We consider another example (Table 6):

\[ (891)_{10} \equiv (31323)_4 = (R_1, L_1, R_2, 0, L_1, L_1)_4. \]
The weight problem of Bachet de Méziriac requires one to find the minimum number of weights required that can be placed in *either pan* of a balance to weigh any integral number of pounds from 1 to 40. A slight variation of the problem is to find the minimum number of weights required that can be placed in *only one pan* of the balance to weigh any integral number of pounds from 1 to 40. The answer to the former question is four weights with measurements 1 lb, 3 lbs, 9 lbs and 27 lbs, whereas the answer to the latter problem is six: 1 lb, 2 lbs, 4 lbs, 8 lbs, 16 lbs and 32 lbs. It is easy to observe that the weights in the first case are powers of 3 and in the second case are powers of 2.

So if we need to weigh any integral number of weights from 1 to \( n \) we would need weights of measure 1, 3, \ldots , 3^k \) such that

\[ n \leq \frac{1}{2} (3^{k+1} - 1) \]

or 1, 2, \ldots , 2^k \) such that \( 2^k < n \), for \( k = 1, 2, 3, \ldots \). For any given weight measure how do we decide what weights are needed? One way to do is to convert the number in its binary form (for the latter case) or ternary numeral system (for the former case).

Let us first consider the case when the weights are placed in only one pan and we wish to measure a weight of 40 lbs.

\[ 40 \equiv (101000)_2, \]

which can be written in the expanded form as:

\[ 40 \equiv (101000)_2 = 1 \times 32 + 0 \times 16 + 1 \times 8 + 0 \times 4 + 0 \times 2 + 0 \times 1. \]

Hence one would need a weight of 32 lbs and 8 lbs to measure a weight of 40 lbs.

If the weights can be placed in either of the pan then we convert 40 in its ternary numeral system.

\[ 40 \equiv (1111)_3 = 1 \times 27 + 1 \times 9 + 1 \times 3 + 1 \times 1. \]

This may be interpreted as that in order to measure a weight of 40 lbs, one requires a weight each of 27 lbs, 9 lbs, 3 lbs and 1 lb – four weights in total which is exactly the answer we needed. Let us see what happens if we wish to weigh 50 lbs instead of 40 lbs. As before, we convert 50 in its ternary numeral system.

\[ 50 \equiv (1212)_3 = 1 \times 27 + 2 \times 9 + 1 \times 3 + 2 \times 1. \]

If we interpret it as before, we would need one weight of 27 lbs, two weights of 9 lbs, one weight of 3 lbs and two weights of 1 lb – a total of 6 weights. But this is not the
minimum number of weights required. We would need lesser number of weights to measure 50 lbs. If the weights can be kept in either of the two pans, we can look at the problem through the perspective of balanced number system. Let us expand the number 50 using the balanced number system.

\[
50 = 1 \times 27 + 2 \times 9 + 1 \times 3 + 2 \times 1 \\
= 1 \times 27 + (1 \times 27 - 1 \times 9) + 1 \times 3 + (1 \times 3 - 1 \times 1) \\
= 2 \times 27 - 1 \times 9 + 2 \times 3 - 1 \times 1 \\
= (1 \times 81 - 1 \times 27) - 1 \times 9 + (1 \times 9 - 1 \times 3) - 1 \times 1 \\
= 1 \times 81 - 1 \times 27 - 1 \times 3 - 1 \times 1.
\]

Therefore, one needs four weights – a weight each of 81 lbs, 27 lbs, 3 lbs and 1 lb to measure 50 lbs. In this case, 81 lbs is placed in one pan and 27 lbs, 3 lbs and 1 lb are placed in the other pan.

Similarly, Table 4 suggests that in order to measure a weight of 464 lbs, one needs a weight of each 729 lbs and 9 lbs in one pan and weights of 243 lbs, 27 lbs, 3 lbs and 1 lb in the other. Therefore, six weights would be needed. We ask a question here: can this be measured in fewer numbers of weights? Before we answer this question, recall the law of moments in physics that states:

For an object balanced in equilibrium the sum of the clockwise moments is equal to the sum of the anticlockwise moments.

If the system in Figure 1 is in equilibrium, then according to the law of moments, the system may be represented mathematically as:

\[
\text{Force}_1 \times \text{Distance}_1 = \text{Force}_2 \times \text{Distance}_2.
\]

The mechanical weighing balance uses the same principle. A mathematical representation of a number in base 3 analogous to the weighing balance is shown in Figure 2.
If instead we have a weighing balance that has four pans rather than two (as shown in the Figure 3), then in equilibrium, the law of moments will give us:

\[ W_1 \times x_1 + W_2 \times x_2 = W_3 \times x_3 + W_4 \times x_4. \]

The mathematical representation of a number in base 5 analogous to the weighing balance is shown in Figure 4.

To end the section, we again look at the example of weighing 464 lbs in balanced quinary system. Representing 464 in base 5 gives:

\[(464)_{10} \equiv (R_1, L_1, L_1, L_2, L_1)_5.\]

So, one would need five weights each of denomination 625 lbs, 125 lbs, 25 lbs, 5 lbs and 1 lb. The weights and the quantity \(W\) to be measured will have to be placed as shown in Figure 5.

We could also have used four weights of denominations 1, 6, 36 and 216 lbs because,

\[464 = 2 \times 216 + 1 \times 36 - 1 \times 6 + 2 \times 1\]

or, \[(464)_{10} \equiv (R_2, R_1, L_1, R_2)_6.\]
In fact, we can measure all integral weights up to 259 lbs using these four weights, but we shall need a weighing balance with six pans! Decrease in the number of weights being used increases the number of pans in the weighing balance.

4. Application to Binary Numbers Magic Trick

We now explore the usage of balanced number system in another popular mathematical trick – the binary numbers magic trick. The trick requires a volunteer to think of any number between 1 and 30. The person performing the trick asks the volunteer to point out the cards from the set of five cards in Table 7 (Figure 6).

How did the performer guess the number? One can notice that the sum of the top left corner number of the selected cards adds up to 27. That is exactly what the performer does. But we wish to understand why the trick works. Please note that variation of the trick exists with a different sets of cards.

The trick is very much like the weight problem where the weight is to be kept in only one pan. If one carefully looks at the number in each card and convert it to binary number, one will get a fair idea of how the trick works. The 0’s and 1’s appearing

![Binary magic card trick](image)

**Figure 6.** Binary magic card trick.
Table 7. Binary magic using cards.

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>5</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>11</td>
<td>13</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>19</td>
<td>21</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>27</td>
<td>29</td>
<td>31</td>
<td></td>
</tr>
</tbody>
</table>

(a)  

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>3</td>
<td>6</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>11</td>
<td>14</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>19</td>
<td>22</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>27</td>
<td>30</td>
<td>31</td>
<td></td>
</tr>
</tbody>
</table>

(b)  

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>21</td>
<td>22</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>29</td>
<td>30</td>
<td>31</td>
<td></td>
</tr>
</tbody>
</table>

(c)  

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>11</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>24</td>
<td>25</td>
<td>26</td>
</tr>
<tr>
<td>27</td>
<td>28</td>
<td>29</td>
</tr>
</tbody>
</table>

(d)  

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>17</td>
<td>18</td>
</tr>
<tr>
<td>19</td>
<td>20</td>
<td>21</td>
</tr>
<tr>
<td>24</td>
<td>25</td>
<td>26</td>
</tr>
<tr>
<td>27</td>
<td>28</td>
<td>29</td>
</tr>
</tbody>
</table>

(e)  

in the binary equivalent of the number allows you to know in which card the number should be placed and in which it should not be. For example if one looks at 27 whose binary representation is \((11011)_2\), then this tells us that 27 must be placed in the first, second, fourth and fifth card and should not be there in the third card.

If one wishes to generalize the trick using the ternary system, one needs to go back to the weighing problem in ternary system or to be more precise balanced ternary system. If each number from 1 to 80 is expressed in its balanced ternary notation, then the cards in Table 8 can be used to guess any number from 1 to 80. The only

Table 8. Ternary magic cards.

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>10</td>
<td>11</td>
<td>13</td>
<td>14</td>
<td>16</td>
</tr>
<tr>
<td>19</td>
<td>20</td>
<td>22</td>
<td>23</td>
<td>25</td>
</tr>
<tr>
<td>28</td>
<td>29</td>
<td>31</td>
<td>32</td>
<td>34</td>
</tr>
<tr>
<td>37</td>
<td>38</td>
<td>40</td>
<td>41</td>
<td>43</td>
</tr>
<tr>
<td>46</td>
<td>47</td>
<td>49</td>
<td>50</td>
<td>52</td>
</tr>
<tr>
<td>55</td>
<td>56</td>
<td>58</td>
<td>59</td>
<td>61</td>
</tr>
<tr>
<td>64</td>
<td>65</td>
<td>67</td>
<td>68</td>
<td>70</td>
</tr>
<tr>
<td>73</td>
<td>74</td>
<td>76</td>
<td>77</td>
<td>79</td>
</tr>
</tbody>
</table>

(a)  

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>5</td>
<td>11</td>
<td>14</td>
<td>20</td>
</tr>
<tr>
<td>12</td>
<td>15</td>
<td>18</td>
<td>19</td>
<td>23</td>
</tr>
<tr>
<td>20</td>
<td>21</td>
<td>22</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>29</td>
<td>30</td>
<td>31</td>
<td></td>
</tr>
</tbody>
</table>

(b)  

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>6</td>
<td>12</td>
<td>15</td>
<td>21</td>
</tr>
<tr>
<td>29</td>
<td>32</td>
<td>38</td>
<td>41</td>
<td>47</td>
</tr>
<tr>
<td>30</td>
<td>33</td>
<td>39</td>
<td>42</td>
<td>48</td>
</tr>
<tr>
<td>56</td>
<td>59</td>
<td>65</td>
<td>68</td>
<td>74</td>
</tr>
<tr>
<td>57</td>
<td>60</td>
<td>66</td>
<td>69</td>
<td>75</td>
</tr>
<tr>
<td>58</td>
<td>61</td>
<td>67</td>
<td>70</td>
<td>76</td>
</tr>
</tbody>
</table>

(c)  

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>14</td>
<td>32</td>
<td>41</td>
<td>59</td>
</tr>
<tr>
<td>6</td>
<td>15</td>
<td>33</td>
<td>42</td>
<td>60</td>
</tr>
<tr>
<td>7</td>
<td>16</td>
<td>34</td>
<td>43</td>
<td>61</td>
</tr>
<tr>
<td>8</td>
<td>17</td>
<td>35</td>
<td>44</td>
<td>62</td>
</tr>
<tr>
<td>9</td>
<td>18</td>
<td>36</td>
<td>45</td>
<td>63</td>
</tr>
<tr>
<td>10</td>
<td>19</td>
<td>37</td>
<td>46</td>
<td>64</td>
</tr>
<tr>
<td>11</td>
<td>20</td>
<td>38</td>
<td>47</td>
<td>65</td>
</tr>
<tr>
<td>12</td>
<td>21</td>
<td>39</td>
<td>48</td>
<td>66</td>
</tr>
<tr>
<td>13</td>
<td>22</td>
<td>40</td>
<td>49</td>
<td>67</td>
</tr>
</tbody>
</table>

(d)  

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>41</td>
<td>23</td>
<td>50</td>
<td>32</td>
</tr>
<tr>
<td>15</td>
<td>42</td>
<td>24</td>
<td>51</td>
<td>33</td>
</tr>
<tr>
<td>16</td>
<td>43</td>
<td>25</td>
<td>52</td>
<td>34</td>
</tr>
<tr>
<td>17</td>
<td>44</td>
<td>26</td>
<td>53</td>
<td>35</td>
</tr>
<tr>
<td>18</td>
<td>45</td>
<td>27</td>
<td>54</td>
<td>36</td>
</tr>
<tr>
<td>19</td>
<td>46</td>
<td>28</td>
<td>55</td>
<td>37</td>
</tr>
<tr>
<td>20</td>
<td>47</td>
<td>29</td>
<td>56</td>
<td>38</td>
</tr>
<tr>
<td>21</td>
<td>48</td>
<td>30</td>
<td>57</td>
<td>39</td>
</tr>
<tr>
<td>22</td>
<td>49</td>
<td>31</td>
<td>58</td>
<td>40</td>
</tr>
</tbody>
</table>
alteration is that now we use two colors. This is analogous to using the two pans – left and right in the weight problem. If the number is in the first card and is black add 1 and if it is red subtract 1. Similarly, if the number is in the second card and is black add 3 and if it is red subtract 3. For the third card the cue number is 9 where as for the fourth card it is 27. The resultant answer is the required number. For numbers greater than 40 the resultant sum will come out to be negative.

If we look at the number 58, for example:

\[(58)_{10} \equiv (R_1, L_1, 0, R_1)_{3}.\]

The number 58:

- is present in the first card and is black. Therefore add 1.
- is also there in the second card and is black; add 3 so that the resultant sum is 4.
- is not there in the third card but present in the fourth card in red color. Hence, subtract 27 so that the resultant is –23.

If there was a fifth card, its cue number would be 81 and 58 would be present in it in black colour. Therefore we add 81 to the resultant to get 58 = 81 – 23.

Recall that in base 5, one needed a weighing balance with four pans. So if the trick is to be extended one would need four colors (Table 9). If the number is present in the first card add 1 – the cue number, if it is black, two times 1 if it is blue, –2 times 1 if it is red and –1 times 1 if it is green. If the number is present in the second card, add 5 (the cue number) if it is black, 2 times 5 if it is blue, –2 times 5 if it is red and –1 times 5 if it is green. For the third card the procedure remains the same but the cue number is now 25. Till 62 (\(62 = 2 \times 1 + 2 \times 5 + 2 \times 25\)), one would get a positive sum. Beyond 62, till 124 (< 125), one would get a negative number. Add the negative number to 125 to get the answer.

In each set of cards a number of patterns can be observed. The reader may like to work out these patterns to generalize the puzzle. They may also refer to [8].
5. Detecting a Counterfeit Coin

A classical mathematical puzzle requires a fake coin to be detected from a pile of twelve coins that are identical in appearance. The fake coin may either be lighter or heavier. The problem is to determine the minimum number of weighings needed to detect the fake coin using a balance scale without the measuring scales. The problem is stated as follows in the literature:

There are \( m \geq 3 \) coins identical in appearance. Only one coin differs from the others in weight, though it is not known in which direction. What is the smallest number of weighings needed to find this coin and identify its type using balance scales without measuring weights?

J Sarkar and B K Sinha [9] constructed both sequential and non-sequential weighing plans for minimum number of weighings needed to detect at most one fake coin for any given number of coins. In this article, generalization of the Dyson’s method presented in [10] for a balance with four pans is presented (Figure 7).

The number of coins \( m \) that can be detected in \( n \) weighings using a weighing balance with four pans will now satisfy:

\[
m \leq \frac{1}{2} (5^n - 5).
\]

The argument is similar to that presented in [7]. We illustrate the use of balanced number system for the special case when \( n = 3 \) and \( m = 60 \). The coins are marked using the balanced numbers \(-2, -1, 0, 1, 2\) instead of \(0, 1, 2, 3, 4\). The right markers are obtained using the cyclic permutations and the left markers corresponding to each right markers are such that the sum is zero. The right markers and the left markers of each of the sixty coins is given in Table 10.

The coins are divided into five sets \( M(i, -2), M(i, -1), M(i, 0), M(i, 1) \) and \( M(i, 2) \) for \( 1 \leq i \leq 3 \). The use of the balance quinary system aligns to our understanding of integers – negative 2 implies 2 units to the left and positive 2 suggests 2 units on the right. Therefore, in the first weighing \((i = 1)\) coins in the set \( M(i, -2) \) are placed in

**Figure 7.** Detecting a fake coin using a weighing balance with four pans.
In the second weighing assume that the pan \( L_2 \) on the left side is heavier. Then the

<table>
<thead>
<tr>
<th>Coin No.</th>
<th>Left Marker</th>
<th>Right Marker</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2, 2, 1</td>
<td>-2, -2, -1</td>
</tr>
<tr>
<td>2</td>
<td>2, 2, 0</td>
<td>-2, -2, 0</td>
</tr>
<tr>
<td>3</td>
<td>2, 1, 2</td>
<td>-2, -1, -2</td>
</tr>
<tr>
<td>4</td>
<td>2, 1, 1</td>
<td>-2, -1, -1</td>
</tr>
<tr>
<td>5</td>
<td>2, 1, 0</td>
<td>-2, -1, 0</td>
</tr>
<tr>
<td>6</td>
<td>2, 1, -1</td>
<td>-2, -1, 1</td>
</tr>
<tr>
<td>7</td>
<td>2, 1, -2</td>
<td>-2, -1, 2</td>
</tr>
<tr>
<td>8</td>
<td>2, 0, 2</td>
<td>-2, 0, -2</td>
</tr>
<tr>
<td>9</td>
<td>2, 0, 1</td>
<td>-2, 0, -1</td>
</tr>
<tr>
<td>10</td>
<td>2, 0, 0</td>
<td>-2, 0, 0</td>
</tr>
<tr>
<td>11</td>
<td>2, 0, -1</td>
<td>-2, 0, 1</td>
</tr>
<tr>
<td>12</td>
<td>2, 0, -2</td>
<td>-2, 0, 2</td>
</tr>
<tr>
<td>13</td>
<td>1, 1, 0</td>
<td>-1, -1, 0</td>
</tr>
<tr>
<td>14</td>
<td>1, 1, -1</td>
<td>-1, -1, 1</td>
</tr>
<tr>
<td>15</td>
<td>1, 0, 2</td>
<td>-1, 0, -2</td>
</tr>
<tr>
<td>16</td>
<td>1, 0, 1</td>
<td>-1, 0, -1</td>
</tr>
<tr>
<td>17</td>
<td>1, 0, 0</td>
<td>-1, 0, 0</td>
</tr>
<tr>
<td>18</td>
<td>1, 0, -2</td>
<td>-1, 0, 2</td>
</tr>
<tr>
<td>19</td>
<td>1, 0, -1</td>
<td>-1, -1, 2</td>
</tr>
<tr>
<td>20</td>
<td>1, -1, 2</td>
<td>-1, 1, -2</td>
</tr>
<tr>
<td>21</td>
<td>1, -1, 1</td>
<td>-1, 1, -1</td>
</tr>
<tr>
<td>22</td>
<td>1, -1, 0</td>
<td>-1, 1, 0</td>
</tr>
<tr>
<td>23</td>
<td>1, -1, -1</td>
<td>-1, 1, 1</td>
</tr>
<tr>
<td>24</td>
<td>1, -1, -2</td>
<td>-1, 1, 1</td>
</tr>
<tr>
<td>25</td>
<td>0, 0, -1</td>
<td>0, 0, 1</td>
</tr>
<tr>
<td>26</td>
<td>0, 0, -2</td>
<td>0, 0, 2</td>
</tr>
<tr>
<td>27</td>
<td>0, -1, 2</td>
<td>0, 1, -2</td>
</tr>
<tr>
<td>28</td>
<td>0, -1, 1</td>
<td>0, 1, -1</td>
</tr>
<tr>
<td>29</td>
<td>0, -1, 0</td>
<td>0, 1, 0</td>
</tr>
<tr>
<td>30</td>
<td>0, -1, -1</td>
<td>0, 1, 1</td>
</tr>
</tbody>
</table>

Table 10. Coin markers.

the pan \( L_2 \), coins in the set \( M(i, -1) \) are placed in the pan \( L_1 \), coins in the set \( M(i, 1) \) are placed in the pan \( R_1 \) and coins in the set \( M(i, 2) \) are placed in the pan \( R_2 \). In case the left side is heavier, the weighing is denoted by the digit \(-2\) or \(-1\) depending upon whether the pan \( L_2 \) or \( L_1 \) is heavier. Similarly if the right side is heavier, the weighing is denoted by the digit \(1\) or \(2\) depending upon whether the pan \( R_1 \) or \( R_2 \) is heavier. If the pans balance, the weighing is denoted by \(0\). The combination of the digits obtained from the three weighings gives the marker of the coin.

As an example, let us assume that the pan \( R_1 \) on right side is heavier (Figure 8). Then the digit for the first weighing is \(1\).

In the second weighing assume that the pan \( L_2 \) on the left side is heavier. Then the
Figure 8. Weighing for the counterfeit coin.

digit corresponding to the second weighing is $-2$. In the third weighing if all the four pans balance, then the digit is 0. Combining the three we get the marker $1, -2, 0$ which is the right marker for coin number 39. Hence, we detect coin number 39 to be fake which is heavier than the rest (also see [11]).

Suggested Reading

Some G Pólya Gems from Complex Analysis *

Shobha Madan

George Pólya (1887–1985) was a brilliant Hungarian mathematician; and many of you may have come across his famous (bestseller) book *How to Solve it*. Elsewhere in this issue you will find a more complete biography of Pólya. Here we will talk about some of his work in complex analysis. Among Pólya’s contemporaries were mathematicians like Leopold (Lipót) Fejér, his thesis advisor, Adolf Hurwitz (1859–1919), G H Hardy (1877–1947), Gábor Szegő (1895–1985).

1. Introduction

From a first course in complex analysis, you know that every analytic function has a power series valid in its disc of convergence; this power series is unique, its coefficients being determined by the values of successive derivatives at the centre of the disc of convergence. Now if the function is entire, as soon as we know the values of all the derivatives of the function at a single point, we get a power series for the function valid everywhere. But if the function $f$ has some isolated singularities, then it will require different power series to represent the function in different regions; the coefficients for each power series always come from values of the derivatives at some point. This means then that every property of the function is, at least implicitly, contained in the set of its power series, i.e., in the sequence of its coefficients. Pólya was interested in finding the analytic character of a function from these coefficients, and one precise problem he worked on is discussed in Section 2. This concerns finding where the singularities of the function are located from the knowledge of some properties of the coefficients in the power series.

In Section 3, we will look at meromorphic functions and show a surprising connection that Pólya found between the location of the poles of a function and properties of the set of zeros of successive derivatives of the function.

In this article, it will not be possible to give proofs of any of Pólya’s theorems, but we will try to explain the content of the results, and illustrate by doing explicit computations on examples. As you will see, these examples are rather interesting.

2. On Gaps in Power Series

Consider a function defined by a power series

$$f(z) = \sum_{n=0}^{\infty} a_n z^n.$$  

Keywords
Analytic functions, power series, isolated singularities, poles and zeros, meromorphic functions.

Let the radius of convergence of this series be \( 0 < R < \infty \). We know that the above series then converges for all \( z \in D(0, R) = \{ z : |z| < R \} \) and that the series diverges for all points in the set \( \{ z : |z| > R \} \). On the circle, \( C(0, R) = \{ z : |z| = R \} \), the series may converge at some points, not at others. The standard example

\[
\frac{1}{1-z} = \sum_{0}^{\infty} z^n, \quad |z| < 1
\]

is one where, although the power series diverges if \( |z| > 1 \), the expression on the left defines an analytic function at all points except at \( z = 1 \), where it has a simple pole. We say that this function, initially defined by a power series in the disc of convergence has an analytic continuation to an analytic function in a larger region (a connected open set containing \( D(0, 1) \), namely, \( \mathbb{C} \setminus \{ 1 \} \)).

Next, let \( k \) be a fixed integer, and consider,

\[
\frac{1}{1-z^k} = \sum_{0}^{\infty} z^{kn}, \quad |z| < 1.
\]

Now there are \( k \) simple poles on the unit circle at the points \( e^{2\pi ij/k}, \ j = 0, 1, ..., k - 1 \). We notice immediately that the series has become somewhat sparse, that the number of singularities of the function on the circle of convergence has increased, and the function has an analytic continuation to a larger region, namely to \( \mathbb{C} \setminus \{ e^{2\pi ij/k}, \ j = 0, 1, ..., k - 1 \} \). We will look at aspects of Pólya’s work related to the following broad problem:

**What, if any, is the relation between the coefficients of a power series and the singularities of the function it represents?**

There must be a relation because after all, the power series does completely determine a function, even in regions where the power series does not converge, since if the function extends to an analytic function outside the disc of convergence to some region, then it is uniquely determined by the power series. (Remember that the zeros of an analytic function are isolated.) In order to really get into the spirit of what is to follow, we look at some examples.

**Example 2.1.** Let

\[
f(z) = \sum_{0}^{\infty} z^{n!}
\]

be a power series, where many coefficients are zero. Its radius of convergence is 1. (Why?) What can we say about the possible extension of this function, i.e., its analytic continuation to a region larger than the open unit disc \( D(0, 1) \)? The question is really about what happens at the boundary points. But it is easy to see that if \( z = e^{2\pi ip/q} \) for some integers \( p, q \neq 0 \), then for \( n \geq q \), the terms of the series are all equal to 1, and the series diverges to \( \infty \). So each such point is a singularity of the function defined by this power series, and these points are dense in \( C(0, 1) \); hence
there is no way of extending the function analytically beyond the unit disc, since \( \lim_{r \to 1^-} f(r e^{2\pi i p/q}) = \infty \).

**Example 2.2.** Consider

\[
f(z) = \sum_{0}^{\infty} z^{2n}.
\]

The sequence of coefficients looks like

\[
\{0, 1, 1, 0, 1, 0, 0, 1, 0, 0, 1, 0, \ldots\}.
\]

(Notice how the gaps between non-zero coefficients keep increasing, as in the previous example!) Once again, the radius of convergence equals 1, and \( z = 1 \) is clearly a singular point. Now, for every positive integer \( k \), we have

\[
f(z^{2k}) = \sum_{0}^{\infty} (z^{2k})^{2n} = \sum_{0}^{\infty} z^{2k+2n}
\]

so that

\[
f(z^{2k}) = f(z) - \sum_{0}^{k-1} z^{2n}.
\]

It follows that each point of the form \( e^{\frac{2\pi i p}{2k}} \), with \( p \) odd, is a singular point for \( f \). Such points form a dense subset of the unit circle, and again \( \lim_{r \to 1^-} f(r e^{2\pi i p/2k}) = \infty \), so no analytic continuation of \( f \) is possible.

These two examples have three things in common:

1. The power series is very sparse, in the sense that there are huge gaps between nonzero coefficients.
2. The function blows up on the boundary on a dense subset of the boundary of the region of definition.
3. The function cannot be analytically continued to any larger domain.

The reader may begin to get an idea that the fact that the power series blows up is the reason for the circle of convergence being the *natural boundary* of the function represented by it. But consider the following example:

**Example 2.3.** We modify Example 2.1 and let

\[
f(z) = \sum_{0}^{\infty} \frac{z^{n!}}{n!}.
\]

Now the power series has radius of convergence 1, and converges absolutely for all points on the unit circle, and so defines a continuous function on the closed unit disc. The question is whether this function has an analytic continuation to a larger region or not.
This problem was considered by the French mathematician Jacques Hadamard (1865–1963) in his thesis as early as in 1892. Hadamard defined the existence of adequate gaps in a quantitative way as follows:

**DEFINITION.**

A power series \( \sum_{n=0}^{\infty} a_n z^n \) is called a Hadamard lacunary series if there exists a sequence \( n_0, n_1, \ldots, n_k, \ldots \) such that

\[
n_{k+1}/n_k \geq \lambda > 1,
\]

and for which \( a_j = 0, \forall n_k < j < n_{k+1}, \) and \( a_{n_k} \neq 0. \)

All the examples above are Hadamard lacunary series.

**Theorem 2.1 (Hadamard’s Gap Theorem).** Any Hadamard lacunary series \( \sum_{n=0}^{\infty} a_n z^n \) with radius of convergence \( R > 0 \) does not have an analytic continuation to any region larger than \( D(0, R) \). Moreover, every point of the boundary \( C(0, R) \) is a singular point.

The author finds this result rather strange because on the one hand, with the large gaps, the series converges quite fast in \( D(0, R) \) and on the other, has singularities everywhere on the boundary.

In 1899, E. Fabry proved that the gap condition can be greatly improved to get the same result as in Hadamard’s theorem, which then becomes a special case.

First a definition: Given a sequence \( \{n_k\}_{k=0}^{\infty} \) of positive integers, let

\[
\rho = \liminf_{k \to \infty} \frac{k}{n_k}.
\]

Then \( \rho \) is called the lower asymptotic density of the sequence, and is also a measure of lacunarity. We see easily that every Hadamard lacunary series has \( \rho = 0. \)

**Theorem 2.2 (Fabry).** Suppose \( \{n_k\} \) is an increasing sequence of positive integers with lower asymptotic density \( \rho = 0, \) and the power series \( f(z) = \sum_{n=0}^{\infty} a_n z^{n_k} \) has radius of convergence \( R > 0, \) then \( f \) cannot be analytically continued from \( D(0, R) \).

At this point a natural question arises: Can the density condition be weakened further to get the same result as in Fabry’s theorem. Pólya said, “No”; and he completely solved the problem by proving a converse of Fabry’s theorem.

**Theorem 2.3.** Suppose \( \{n_k\} \) is an increasing sequence of positive integers. Suppose that every power series \( \sum a_n z^{n_k} \) with a finite radius of convergence \( R > 0, \) cannot be continued analytically to any larger region than \( D(0, R) \), then \( \lim_{k \to \infty} \frac{k}{n_k} = 0. \)

So Pólya’s big contribution is to say that Fabry’s theorem is the best possible.

Before ending this section we will mention another of Pólya’s insightful results; that there are many (in fact, uncountably many) series, not necessarily lacunary, which have their circle of convergence as their natural boundary. In fact, as soon as we have a power series with a finite radius of convergence, then by changing signs of the
coefficients suitably, singularities can appear in a big way. A Hurwitz and G Pólya wrote a paper [3], where the first part is authored by Hurwitz giving one proof, and the second part is authored by Pólya giving a different proof of the following result:

**Theorem 2.4 (Hurwitz–Pólya).** Let $\sum_{n=0}^{\infty} a_n z^n$ be a power series with a finite radius of convergence $R > 0$. Then the set of functions of the form

$$
\sum_{n=0}^{\infty} \epsilon_n a_n z^n, \; \epsilon_n \in \{-1, 1\},
$$

which have no analytic continuation outside the disc $D(0, R)$, is uncountable.

Observe that the theorem does not say that the set of functions of the above form which can be analytically continued beyond $D(0, R)$ is countable!

### 3. Zeros of Derivatives

Pólya thought a lot about what happens to the zeros of successive derivatives of analytic functions. In this section, we will discuss a particularly elegant result for meromorphic functions on the plane. Recall that a function $f : \mathbb{C} \to \mathbb{C}$ is called meromorphic if all its singularities are poles (hence these are isolated points). It will be useful to begin with simple examples.

**Example 3.1.** Consider the simplest example of a meromorphic function,

$$
f(z) = \frac{1}{z},
$$

with a single simple pole. Observe that $f$ does not have any zero in $\mathbb{C}$, and a simple zero at the point at infinity. As we take successive derivatives of $f$, namely $f^{(n)}$, we find that no new poles, nor new zeros appear, although the order of the pole at $z = 0$ and of the zero at $z = \infty$ change.

Next, consider any rational function with a single pole, i.e., a function such as

$$
f(z) = \frac{1}{z^k} + P(z),
$$

where $P(z)$ is any polynomial, and $k \geq 1$ an integer. Now $f$ and the first few derivatives (as many as the degree of $P(z)$) have finitely many zeros, and for large $n$, $f^{(n)}$ has one pole at $z = 0$ and a zero at the point at infinity.

We now consider the next simplest example of a function with two poles and we begin to see that two different poles begin to interact to produce new zeros of successive derivatives.

**Example 3.2.** Let,

$$
f(z) = \frac{1}{z-1} - \frac{1}{z}.
$$
Clearly $f$ has no zeros, except the point at infinity. For the derivatives,

$$f^{(n-1)}(z) = (-1)^{n-1}(n-1)! \frac{z^n - (z-1)^n}{z^n(z-1)^n}$$

we see that the set of zeros

$$Z_{f^{(n-1)}} = \{z : z^n = (z-1)^n\}$$

are the solutions of the equation

$$z = (z-1)e^{2\pi ij/n}$$

with $j = 0, 1, ..., n-1$, i.e., the $n$ points given by

$$z_j = \frac{1}{1 - e^{-2\pi ij/n}}, \quad j = 0, 1, ..., n-1.$$ 

All these points lie on the line $\{\text{Re}(z) = 1/2\}$ (points equidistant from $z = 0$ and $z = 1$). Further, the reader can compute the imaginary part of the $z_j$’s and then easily verify that $\bigcup_0^\infty Z_{f^{(n)}}$ is a dense subset of this line.

One can check easily that if we had chosen the poles to lie at points $a, b \in \mathbb{C}$, instead of at $0, 1$, the zeros of the derivatives would lie on the line consisting of points equidistant from $a$ and $b$. But what will happen if we change the residues at the poles? Consider

$$f(z) = \frac{a}{z-1} - \frac{1}{z},$$

where $a \in \mathbb{C}, \quad |a| \neq 1$. We see that the zeros of $f^{(n-1)}$ are given by

$$Z_{f^{(n-1)}} = \{z : az^n = (z-1)^n\}$$

and will not lie on the line $\text{Re}z = 1/2$ unless $|a| = 1$. But as $n \to \infty$, the zeros will condense towards this line, since $|a|^{1/n} \to 1$, so that ultimately it is only the location of the poles that had any influence.

There is an observation that can be made at this stage:

The poles of $f$ act as repulsive centres for the zeros of its successive derivatives. (*)

Let us try and check this out with another example:

**Example 3.3.** We will investigate a simple case of three simple poles. Let

$$f(z) = \frac{1}{z-i} + \frac{1}{z} + \frac{1}{z-1}.$$ 

Then

$$f^{(n-1)}(z) = (-1)^{n-1}(n-1)! \left( \frac{1}{(z-i)^n} + \frac{1}{z^n} + \frac{1}{(z-1)^n} \right).$$

We are interested in the zeros of these functions, and hence the zeros of the polynomials,

$$P_n(z) = z^n(z-1)^n + (z-i)^n(z-1)^n + z^n(z-i)^n.$$
We will not find the zeros explicitly, but will try to find their location. Let us write $a_1 = i$, $a_2 = 0$, $a_3 = 1$. In Figure 1, the complex plane is shown as the union of three half lines $r_j = r_k$ starting at the point $\frac{1}{2}(1 + i)$, and the three open regions marked as I, II and III. We will see that each point in these open regions cannot be a zero of $f^{(n)}$ for $n$ large enough. Consider first $z \in$ III, then with $r_j = |z - a_j|$, $j = 1, 2, 3$, we have $r_3 < \min(r_1, r_2)$ in this region. Suppose first that $r_3 < r_1 \leq r_2$, then

$$|P_n(z)| = |z^n(z - 1)^n + (z - i)^n(z - 1)^n + z^n(z - i)^n| \geq |z^n(z - i)^n| - |z^n(z - 1)^n| - |(z - i)^n(z - 1)^n| = r_1^n r_2^n - r_2^n r_3^n - r_1^n r_3^n = r_2^n r_3^n \left( \frac{r_1}{r_3}^n - \frac{r_1}{r_2}^n - 1 \right) \geq r_2^n r_3^n \left( \frac{r_1}{r_3}^n - 2 \right) > 0$$

if $n$ is large enough, since $\frac{r_1}{r_3} > 1$. We can do a similar estimate if $r_3 < r_2 \leq r_1$, and so it follows that for all $z \in$ III, there is an $n$ such that $f^{(n)}(z) \neq 0$. The same happens for the regions I and II.

The three examples show that indeed (*) holds. Now let us be more precise about this. We begin with two definitions.

**DEFINITION.**

Let $f$ be a meromorphic function and $a_0 \in \mathbb{C}$ a pole of $f$. The domain of $a_0$ is defined as the set

$$\text{Dom}(a_0) = \{ z \in \mathbb{C} : |z - a_0| < |z - a| \ \forall \ \text{pole} \ a \neq a_0 \text{ of } f \} .$$
Pólya shows a very interesting connection between the location of the poles of $f$ and the final set $D$.

In Example 3.2 discussed above, the domain of the pole at $z = 0$ is the half plane to the left of the line $\{z : \Re z = 1/2\}$ and the domain of $z = 1$ is the right half plane. The regions I, II and III in Example 3.3 are respectively the domains of the poles $a_1, a_2, a_3$. In general, the domain of a pole is an open convex set whose boundary consists of polygonal lines.

Next, Pólya defined a set, called the ‘final set’ $D$ of a function $f : \mathbb{C} \to \mathbb{C}$ consisting of accumulation points of the set of all zeros of all derivatives of $f$. In other words, we can say that $z \notin D$ if there exists an $r > 0$ such that no derivative $f^{(n)}$ vanishes on the set $\{w : 0 < |w - z| \leq r\}$.

Now we are ready to state Pólya’s result, proved in [4], where he shows a very interesting connection between the location of the poles of $f$ and the final set $D$. See also [5].

The final set $D$ of a meromorphic function $f$ contains no point of the domain of any of its poles, but contains all points which are on the common boundary of two or more poles.

It is remarkable that the final set $D$ here is completely determined by the location of the poles of $f$ (and does not even depend on the order of the poles, nor on their residues). Further, observe that the result does not say anything about the location of the zeros of successive derivatives of $f$, and in fact these may lie inside the domain of a pole. In Example 3.2, all the zeros of all the derivatives do lie on the common boundary of the domains of the two poles, i.e., the line $\Re z = 1/2$, if the two residues have equal size, and if not, the set of zeros still condenses towards that line. In Example 3.3, we have not found the zeros of the derivatives explicitly; it was enough to show that each point of the domain of any pole is not a zero of $f^{(n)}$ for all large enough $n$.

Our final example is rather interesting.

**Example 3.4.**

$$f(z) = \frac{1}{1 - e^z}.$$  
There are infinitely many poles of $f$, namely the set: 

$$\{2\pi in ; n \in \mathbb{Z}\}.$$  

Pólya’s result stated above tells us that the final set $D$ of $f$ is given as a union of lines, 

$$D = \bigcup_{n = 0}^{\infty} \{z : \Im z = \pi(2n + 1)\}$$

and each line is the common boundary of the domain of two consecutive poles. It turns out that in this case, all the zeros of all the derivatives also lie in $D$, which is more than what Pólya’s result says. We will show this below. For this we first need to find the $n$th derivatives. We can write our function as a composition of two functions as follows:

$$f(z) = F(\phi(z)),$$
where \( F(z) = \frac{1}{1-z} \), and \( \phi(z) = e^z \). Now it would be nice to have a Leibniz formula for the \( n \)th derivative of the composition of two functions (the reader may want to find such a formula), but in our case, the task is somewhat simpler, since \( \phi^{(n)}(z) = \phi(z) \) for all \( n \). First, let us compute the \( n \)th derivative of a function of the form \( F(e^z) \), where \( F \) is a differentiable function. Taking the first derivative, we get,

\[
\frac{d}{dz} F(e^z) = F'(e^z) e^z.
\]

Taking subsequent derivatives we will get an expression of the form

\[
\left( \frac{d}{dz} \right)^n F(e^z) = \sum_{k=1}^{n} S^n_k F(k)(e^z) e^{kz},
\]

where the coefficients satisfy

\[
S_1^1 = 1, \quad S_n^n = 1, \quad S_{k+1}^n = S_{k-1}^n + k S_k^n
\]
as is easy to verify. We put

\[
S_k^n = 0, \quad \text{unless } 1 \leq k \leq n.
\]

The notation of using \( S_n^k \) for the coefficients is because these numbers are well known as Stirling numbers of the second kind; and in fact, \( S_n^k \) stands for the number of distinct partitions of a set of \( n \) elements into \( k \) classes. With this we can now write the expression for the \( n \)th derivative of the given function \( f \) as follows

\[
\left( \frac{d}{dz} \right)^n (1 - e^z)^{-1} = \sum_{k=1}^{n} S^n_k k! \frac{e^{kz}}{(1 - e^z)^{k+1}}
\]

\[
= \frac{1}{1 - e^z} \sum_{k=1}^{n} S^n_k k! \left( \frac{e^z}{1 - e^z} \right)^k
\]

\[
= \frac{1}{1 - e^z} P_n(\omega),
\]

where we have put \( \omega = \frac{e^z}{1 - e^z} \) and the polynomial \( P_n \) is given by

\[
P_n(\omega) = \sum_{k=1}^{n} S^n_k k! \omega^k.
\]

Clearly we need to get information on the zeros of these polynomials for all \( n \geq 1 \). For this we use induction. We follow Pólya and use the notation

\[
a^n_k = k! S^n_k.
\]

(In a footnote in the paper [8], Pólya remarks that in his search for a proof, the turning point was the introduction of this notation!) Then the \( a^n_k \)'s satisfy the following recurrence relation,

\[
a_{k+1}^{n+1} = k(a_{k-1}^n + a_k^n)
\]
and so
\[ P_{n+1}(\omega) = \sum_{k=1}^{n+1} a_k^{n+1} \omega^k = \sum_{k=1}^{n+1} k(a_{k-1}^n + a_k^n)\omega^k \]
\[ = \omega \sum_{k=1}^{n+1} ((k-1)a_{k-1}^n \omega^{k-1} + a_k^n \omega^{k-1}) + \sum_{k=1}^{n} k a_k^n \omega^{k-1} \]
\[ = \omega \left[ \sum_{k=1}^{n} k a_k^n \omega^{k-1} + P_n(\omega) + P'_n(\omega) \right] \]
\[ = \omega \left[ (\omega + 1)P_n(\omega) \right] \]
\[ = \omega \left( \frac{d}{d\omega} \right) [((\omega + 1)P_n(\omega)] \]
We can now prove

Lemma 3.1. For every \( n \geq 1 \), all zeros of the polynomial \( P_n \) are real and simple, and are contained in the interval \((-1, 0]\).

Proof. We prove by induction on \( n \). We know that the degree of the polynomial \( P_n \) is \( n \). For \( n = 1 \), \( P_1(\omega) = \omega \) and the only zero is at \( \omega = 0 \). We assume the result for an integer \( n \), and we may simply restrict the polynomials to the real line, since all zeros will be found to be real. In view of the formula derived above for \( P_{n+1} \), note first that the polynomial \((\omega + 1)P_n(\omega)\) has \( n + 1 \) simple zeros in \([-1, 0]\) by the induction hypothesis; and apart from the two simple zeros at the end-points of this interval, the other \((n - 1)\) simple zeros lie in its interior and are distinct. Therefore by Rolle’s theorem, we see that there are \( n \) distinct zeros of \( P_{n+1} \) in \((-1, 0)\), and along with the zero at \( \omega = 0 \), we have determined all the zeros of \( P_{n+1} \), as required.

The author would have liked to read the proofs of these and other remarkable results due to Pólya, but feels limited by the fact that a large body of Pólya’s papers are written in Hungarian, Russian and German. It needs to be said that the same limitation determined the choice of the two themes in this article

Suggested Reading
Expander graphs are graphs in which every set of vertices has an unusually large number of neighbours. It is a remarkable fact that graphs of this kind exist. Even more remarkable is the spectrum of applications of these graphs, ranging from providing new insights in the field of computational complexity theory to the solution of problems in communication. In this article we show how expander graphs can be used for designing efficient error correcting codes which have fast decoding algorithms.

1. Introduction

The fifty year old history of error correcting codes can be characterized by a sequence of incremental results punctuated by occasional leaps forward. The most recent dramatic development, is the recognition of the fact that one can derive excellent codes from graphs which have a high degree of local connectivity but simple structural descriptions that facilitate iterative decoding. One such class consists of expander graphs which were introduced in the 1970’s and have turned out to be very versatile tools in both theory and practice. In the mid 1990’s a clever construction for efficiently decodable codes was discovered by Sipser and Spielman, and this approach has been improved and refined over the last few years. In the first part of this article we attempt to illustrate the power of this technique in constructing efficiently decodable codes.

2. Codes and Channels

A noisy communication channel is illustrated in Figure 1. The aim is to reliably transmit information through this channel. Information is transmitted as a sequence of symbols. For example if we wanted to transmit messages made up of English letters and assuming the channel can transmit only binary digits, we might encode each of the twenty six letter using five bits, and use the remaining six combinations of the 32 possible combinations of five bits for other symbols like blanks, punctuation marks and so forth. The message will now look like a sequence of 0’s and 1’s. Suppose we want to transmit this message over the unreliable communication channel so that even if the channel corrupts some of the bits we are able to recover the original message with high probability. One solution is to transmit blocks of symbols rather than individual symbols after the addition of redundant symbols for protection against errors.

Keywords
Expander graphs, error correction, efficiently decodable codes

Suppose we want to transmit a message over the unreliable communication channel so that even if the channel corrupts some of the bits we are able to recover the original message with high probability. One solution is to transmit blocks of symbols rather than individual symbols after the addition of redundant symbols for protection against errors.
Given a code and the decoding rule, it is possible to compute the probability of decoding error if a probabilistic model of the channel is assumed. One such model is the binary symmetric channel (BSC), shown in Figure 2. On the left are the channel input symbols 0 and 1 and on the right, the channel output symbols, and these are connected by edges. The labels on the edges represent transition probabilities. Thus if a 0 or a 1 was sent, the probability of receiving a 1 or a 0 respectively, is $p$. This is the probability that a bit is flipped in transmission through the channel, and is called the raw bit error probability. The channel is said to be symmetric as the probability of its flipping a 0 or a 1 are the same. In general the input alphabet set $A$ and the output alphabet set $B$ may be different and have different sizes.

Suppose now, that we wish to compute the bit error probability for the simple repetition scheme described above, where each bit is replicated three times. Let us make the simplifying assumption that the message source produces both 0’s and 1’s with equal probability. A message bit is decoded erroneously if two or three of its copies are received in error. If $P_e$ denotes the bit error probability then

$$P_e = P\{2 \text{ channel errors}\} + P\{3 \text{ channel errors}\}$$

$$= 3p^2(1-p) + p^3$$

$$= 3p^2 - 2p^3$$  (1)

Since $p < 1/2$ this is less than the raw bit error probability $p$, and if $p$ is very small, the improvement in the reliability of the channel is quite significant. It is easy to see that even higher reliability is obtained by repeating each bit more times. For this
repetition code, the error probability can be made as small as we wish but with a corresponding decrease in rate. This is not particularly interesting and we can do much better.

In practical systems as illustrated in Figure 1, a message of length \(k\) is encoded into a codeword of length \(n\), \(n > k\). This is transmitted over the noisy channel which may corrupt some of the bits. The corrupted version \(y\) of the codeword is received at the other end and is decoded to form an estimate \(c'\) of the transmitted codeword from which the original message can be recovered. A decoding error is said to occur if the estimated codeword differs from the transmitted one. The block length of the code is \(n\) and the rate \(R = k/n\) is the fraction of real information symbols in a codeword.

Let us look at another simple code, a Hamming code of block length 7 and rate 4/7. For each message block of length 4 there are three extra bits added as checks. If \(x_0, x_1, x_2, x_3\) are the message bits, let \(x_4, x_5, x_6\) be the check or parity bits. These are determined by the equations

\[
x_4 \equiv x_1 + x_2 + x_3 \pmod{2},
\]

\[
x_5 \equiv x_0 + x_2 + x_3 \pmod{2},
\]

\[
x_6 \equiv x_0 + x_1 + x_3 \pmod{2},
\]

Thus for example if \((x_0, x_1, x_2, x_3) = (0111)\) then \((x_4, x_5, x_6) = (100)\) and the codeword sent over the channel is \((0111100)\). Let us rewrite the parity check equations above in the following way:

\[
\begin{align*}
x_1 &+ x_2 + x_3 + x_4 &= 0 \\
x_0 &+ x_2 + x_3 + x_5 &= 0 \\
x_0 &+ x_1 + x_3 + x_6 &= 0
\end{align*}
\]

If we define the binary matrix \(H\) as follows:

\[
H = \begin{bmatrix}
0 & 1 & 1 & 1 & 1 & 0 & 0 \\
1 & 0 & 1 & 1 & 0 & 1 & 0 \\
1 & 1 & 0 & 1 & 0 & 0 & 1
\end{bmatrix}.
\]

For each of the 16 codewords \(c\) we have

\[
H^T c = 0,
\]

where the righthand side is the all zero vector with three components. This just says that for each codeword, all the linear constraints are satisfied. An interesting observation here is that equation (2) is a linear dependence relation among the columns of \(H\), which picks up column \(i\) if the \(i^{th}\) bit in the codeword is a 1. Thus for every linear dependence among \(w\) columns of \(H\) there is a corresponding codeword with \(w\) non-zero components, and every codeword with \(w\) non-zero components corresponds to a linear dependence among \(w\) columns of \(H\). Each codeword is said to be orthogonal to every vector that can be generated by linear combinations of rows of \(H\).
A binary linear block code of length $n$, is a set of $2^k$ vectors of length $n$ all of which satisfy a set of $n-k$ linearly independent constraints. The Hamming distance between two vectors is the number of positions in which they differ. The minimum distance of the code is the minimum of the Hamming distances between all pairs of codewords. A binary linear block code of length $n$, whose message length is $k$ and whose minimum distance is $d$ is termed an $(n,k,d)$ code. It should be clear that the all zero vector is always a member of any linear block code.

Let us compute the minimum distance of the Hamming code described above. Note that the seven columns of $H$ consist of all distinct nonzero combinations of three bits. Since no two columns are identical, the minimum distance of the code is at least two. Further we have several linear combinations of three columns that give the all 0 vector, therefore the minimum distance of the code is 3.

A common rule for decoding is the maximum-likelihood decoding rule which, given a received vector $y$ decodes into the codeword $c'$ such that the probability of $y$ given $c'$ is the largest. For a binary symmetric channel model with $p < 1/2$ a maximum likelihood decoding rule would decode a received vector into a codeword at smallest Hamming distance from the received vector.

We now describe a maximum-likelihood decoding algorithm for the Hamming code above. It turns out to be convenient to imagine that the BSC adds (mod 2) either a 0 or a 1 to each transmitted bit. If it adds a 0, the bit is received correctly; if it adds a 1 the bit is in error. The sequence of 0’s and 1’s is the error pattern denoted by $z$. A single error pattern has a single 1 and all other bits 0.

Now the receiver who knows $y$ and wants to find an estimate of $c$, computes the vector $s = (s_0, s_1, s_2)$ such that $s^T = Hy^T = H(c + z)^T = Hc^T + Hz^T = Hz^T$. Here $s$ is called the syndrome of the received vector. The syndrome of a codeword is always the zero vector. Else if the $i^{th}$ constraint is not satisfied by the received vector, then the $i^{th}$ bit of the syndrome is a 1. The syndrome thus depends only on the error pattern $z$. However if $z$ is known then $c$ can be computed as $c = y + z$. The syndrome for each single error pattern corresponds to a column of $H$. Since the columns exhaust all possible syndromes, and since for $p < 1/2$ a single error pattern is always more likely than a multiple error pattern, maximum likelihood decoding for the Hamming code consists of the following steps:

1. Compute the syndrome $s = Hy^T$ from the received vector $y$.
2. If $s$ is 0, output the received vector as the codeword estimate and exit, else identify the unique column $i$ of $H$ equal to the syndrome.
3. Form the error pattern $z$ which has a 1 in the $i^{th}$ position and zeroes elsewhere.
4. Compute $c' = y + z$.

So all single errors can be corrected. What if two errors occur? Two errors in fact will always cause the received vector to be closer to a codeword other than the one
transmitted. For example if the codeword (0111000) is transmitted and there are two errors in the last two bits, (0111011) is received. The decoder will decode to (0110011) which is at a Hamming distance of 1 from the received vector. If we draw n dimensional spheres of Hamming radius 1 around each codeword, each sphere containing vectors at Hamming distance 1 from the center, it turns out that the spheres are non intersecting (that is why we can correct all single errors), and between them they cover the whole space (which is why a double error in one codeword will give a vector in the Hamming sphere around another). It is easy to see that if \( d = 2t + 1 \) for some \( t \), then one can draw \( n \)-dimensional spheres of radius \( t \) around each codeword and not have any of them intersect. With the assumption that a small number of errors is more likely than a large number, any received vector that falls within a sphere should be decoded to the codeword at the center of the sphere. While this is a simple rule, its implementation is difficult for the general class of linear block codes. Note that the number of possible syndromes for an \((n, k, d)\) binary code is \( 2^{n-k} \), a quantity that is exponential in the code parameters. So a scheme that uses table look-up where each syndrome indexes into the most likely error pattern, is quite impractical for large values of \( n - k \).

Before we proceed further, let us compute the error probability of the Hamming code. We can compute the block error probability i.e. the probability that the transmitted vector is not equal to the decoded vector. But this does not tell the whole story as even if the two are different, some of the bits may be correct. If we denote the bit error probability \( P\{c_i \neq c'_i\} \) by \( P^i_e \) it is possible to show by a combinatorial argument which we omit here that

\[
P^i_e = 9p^2 + 26p^3 + \ldots \tag{3}
\]

where the higher order terms are negligible for small \( p \). When we compare this with the crude repetition scheme we find that for small values of \( p \) the Hamming code at rate 4/7 performs as well as the crude repetition scheme of rate 1/3. So this is some improvement. However, the curious reader may wonder what is the best that one can do, given the channel model. The answer was provided by Claude Shannon in 1948.

In his classic paper of 1948, Shannon considered a large class of probabilistic channels and showed the rather surprising result that there was a quantity called the channel capacity such that one could achieve an arbitrarily small error probability at any rate less than the channel capacity. For example for the binary symmetric channel if \( p = 0.1 \) then the channel capacity is 0.531, and according to Shannon’s theorem there should exist a code with rate \( R \geq 0.5 \) and overall error probability less than \( 10^{-1000} \). This was a counter-intuitive result and it took communication engineers some time to understand its importance.

The outstanding problem for many years was the explicit construction of such codes, as Shannon only showed their existence. In the last few years explicit constructions have appeared. One of these arose from the PhD thesis of Robert Gallager in 1961 which was finally recognized as a deeply prophetic work. Gallager had developed
Each constraint node imposes a linear constraint on the bits of the code-word. The rate of the code is at least \((n - r)/n\) with equality if all the constraints are linearly independent.

A bipartite graph is a graph with two set of nodes. Let us call these the left nodes and the right nodes. Edges exist only between left nodes and right nodes. Every linear block code can be described by a bipartite graph as follows.

Assume that the message symbols are binary digits. The \(n\) nodes on the left called codeword nodes and \(r\) nodes on the right called constraint nodes. In Gallager's original version of an LDPC the nodes on the left are in one-to-one correspondence with the positions 1, 2, ..., \(n\) of a codeword. A binary \(n\)-bit vector \(c = (c_1, c_2, \ldots, c_n)\) is a codeword if and only if for each constraint node the modulo 2 sum of the values of its adjacent message nodes is 0. Each constraint node imposes a linear constraint on the bits of the codeword. The rate of the code is at least \((n - r)/n\) with equality if all the constraints are linearly independent. Each constraint can be thought of as a linear equation in the variables corresponding to each codeword position. The code is called "low density" because the equivalent parity check matrix, is chosen to be sparse, containing at most a constant number of ones in each row and column. We illustrate with a bipartite graph for the Hamming code just described. Though the code is not a low density code, it gives an idea of how graphs for Gallager codes are constructed. Figure 3 displays the bipartite graph for the Hamming code.

**Figure 3.** Bipartite graph for the Hamming code.
For an \((n, k, d)\) code the following two problems are of interest

- The unique bounded distance decoding problem: Given a received word \(y \in B^n\) (where \(B\) is the output alphabet set), find the unique codeword \(c\) if any, that lies within a distance \((d-1)/2\) from \(y\). We saw a geometric interpretation of this problem earlier.

- The list decoding problem: Given a received word \(y\) and an error bound \(t\), find all codewords of \(C\) within distance \(t\) of the received vector. Here we draw an \(n\) dimensional sphere of radius \(t\) around the received vector and output the set of codewords inside the sphere.

Both these problems have elegant algebraic solutions for some classes of codes eg. Reed-Solomon codes, BCH-codes, and algebraic-geometry codes, though these algorithms do not run in linear time. By a linear time algorithm is meant one whose running time is linear in \(n\), the length of the code.

The question was: Can we design codes with decoding algorithms having linear complexity and good performance in terms of their error probability?

Sipser and Spielman showed in 1996 that it was possible to construct a family of asymptotically good linear error-correcting codes which could be decoded in linear time. Their construction used expander graphs and their codes are a special subclass of low density parity check codes. We now describe these graphs.

3. Expander Codes

Expanders are sparse graphs with very good connectivity properties. Suppose we have a bipartite graph which is \(d\)-regular, meaning thereby that every vertex has degree \(d\). One could have regular graphs with degree \(d\) on the left and \(c\) on the right. Such graphs are both left regular and right regular. Now given any subset of vertices of size \(N\) on the left, if we count its neighbours on the right, the maximum value this can have is \(Nd\). We say that a graph is an expander if given a not too large subset on the left the number of neighbours on the right is at least \(\gamma\) times the size of the subset. When \(\gamma\) is very close to \(d\) we say the expander is lossless. More formally:

**Definition:** A bipartite graph \(G = (X, Y, E)\) is said to be an \((n, m, d, a, \gamma)\)-expander if \(|X| = n, |Y| = m\), the degree of each node in \(X\) is \(d\) and for every \(A \subseteq X, |A| \leq an\) the set of neighbours \(N(A)\) of \(A\) in \(Y\) satisfies \(|N(A)| > \gamma|A|\).

An expander is illustrated in Figure 4.

We now show how expanders can lead to linear time decodable codes.

Given a sparse bipartite graph (i.e. a graph with a small number of edges), representing a linear block code with vertices on the left representing positions of the codeword, vertices on the right representing parity checks or constraints, with an edge between a vertex \(a\) on the left and vertex \(b\) on the right if position \(a\) participates in...
parity check $b$, let us assume that each position participates in $r$ parity checks i.e. degree of each vertex on left is $r$, where $r$ is a constant. The following result is useful.

**Lemma 1.** Let $G = (X, Y, E)$ be a bipartite $(n, m, r, \alpha, \beta)$ expander. Then every set $S$ in $X$ of size at most $\alpha n$ satisfies the Unique Neighbour Property, i.e. there exists $y \in Y$ such that $|N(y) \cap S| = 1$.

**Proof.** Consider a set $S$ of size at most $\alpha n$. Suppose the unique neighbour property does not hold for $S$ then every $y \in N(S)$ has at least two neighbours in $S$. Therefore the number of edges leaving $S$ is at least $2|N(S)| > 2\frac{\beta}{2}|S| = r|S|$, which contradicts the assumption of the left regularity of $G$.

The next theorem due to Sipser and Spielman establishes a lower bound on the minimum distance of the code, and is interesting as it is based entirely on the expansion property of the graph.

**Theorem 1.** Let $G = (X, Y, E)$ be a bipartite $(n, m, r, \alpha, \beta)$ expander graph defining a code $C(G)$. Then the minimum distance of $C(G)$, $\text{dist}(C(G)) > \alpha n$.

**Proof.** Assume that $\text{dist}(C(G)) \leq \alpha n$. Then there is a nonzero codeword whose weight is at most $\alpha n$. Let $w$ be such a codeword and let $Z$ be its support, i.e. the set of coordinates that are not zero. Let $y$ be a vertex in $N(Z)$ satisfying the unique neighbour property. The constraint imposed by constraint node $y$ is that the sum of all the variables that it checks is 0 mod 2, but in the assignment defined by $w$ only one of those variables is assigned the value 1. Therefore the constraint cannot be satisfied and $w$ cannot be a codeword, giving a contradiction.

*Figure 5* illustrates the above argument. Note that in the figure all the non-zero components of the codeword have been gathered into the set $Z$ so that all the components outside $Z$ are 0. Since there is no particular ordering of the vertices on the left, this is always possible.
Having established the lower bound on the minimum distance, let us look at the incredibly simple decoding algorithm proposed by Sipser and Spielman. We say that a constraint is satisfied by a setting of a set of variables if the sum of those variables is 0. Else it is unsatisfied. Suppose we receive some word and consider a variable (i.e., node on left hand side of G) and its constraints. If the variable has more unsatisfied than satisfied constraints and we flip the variable, then the situation is reversed, i.e. it has more satisfied than unsatisfied constraints. The goal of the algorithm is to keep flipping variables, ensuring that the total number of unsatisfied constraints keeps decreasing until no unsatisfied constraints are left, i.e. we are left with a codeword. We will show that if the graph is a good expander and the number of errors is not too large, this can be achieved. The simple sequential decoding algorithm is described as follows.

- If there is a variable that is in more unsatisfied than satisfied constraints, then flip the value of that variable.
- Repeat until no such variable remains.

We now show that the algorithm decodes in linear time. It requires an expansion greater than what is enough to prove Lemma 1 and Theorem 1. So if such an expansion is guaranteed, we can use the lemma and theorem.

**Theorem 2.** If the expansion of sets of size at most $\alpha n$ is greater than $(3/4)r$ and if the number of errors in a received word $w'$ is at most $(\alpha/2)n$ then the simple sequential algorithm will decode correctly in linear time.

**Proof.** Let $w = (w_1, w_2, \ldots, w_n)$ be the transmitted codeword and $w' = w'_1, w'_2, \ldots, w'_n$
be the received vector, and at any point in the algorithm, let \( A \) be the set of errors in \( w' \). That is \( A = \{ v : w_v \neq w'_v \} \). If \( A \) is empty then we are done. Otherwise assume that \( |A| \leq \alpha n \). A has two kinds of neighbours, satisfied neighbours \( S \) and unsatisfied neighbours \( U \) so that \( N(A) = S \cup U \). By assumption

\[
|U| + |S| > \frac{3}{4}r|A| \tag{4}
\]

Now count the edges between \( A \) and \( N(A) \). There must be an odd number of edges going into every node in \( U \) and an even number of edges going into every node in \( S \) from \( A \) as the constraints are modulo 2 sums. Therefore there are at least \( |U| \) edges leaving \( U \) and at least \( 2|S| \) edges leaving \( S \) and entering \( A \). Since this a \( r \)-regular graph on the left,

\[
|U| + 2|S| \leq r|A| \tag{5}
\]

Combining (4) and (5) we have

\[
r|A| - |U| > 2|S| > 2\left(\frac{3}{4}r|A| - |U|\right) \tag{6}
\]

Therefore

\[
|U| > \frac{r}{2}|A| \tag{7}
\]

So the total number of unsatisfied neighbours of the \( |A| \) members of \( A \) is greater than \( \frac{r}{2}|A| \) if \( |A| \leq \alpha n \). Therefore there is at least one element of \( A \) that has greater than \( r/2 \) unsatisfied neighbours. Recall \( r \) is the degree of every vertex on the left. So what this means is that as long as there are errors, there will be at least one variable in \( A \) that has more unsatisfied than satisfied neighbours provided the expansion condition is satisfied and the size of the set \( A \) at any time is \( \leq \alpha n \). However this does not mean that the decoding algorithm will decide to flip a corrupt variable. All it guarantees is that if \( |A| \leq \alpha n \) the algorithm will decide to flip some variable and thereby reduce the size of the set \( U \).

To show that the size of the set \( A \) never exceeds \( \alpha n \) during the course of the algorithm we observe that at each step we flip some variable. Now if the variable was in the error set in the first place, we have reduced the size of the error set \( A \) and therefore we stay within the bound. On the other hand if we flip a variable not originally in \( A \) then the size of \( A \) increases by 1. If at any time the size of \( A \) exceeds \( \alpha n \) then at the previous step it must have been \( \alpha n \). Therefore by the previous argument

\[
|U| > \frac{r}{2}\alpha n \tag{8}
\]

Now in the beginning

\[
|U| \leq |N(A)| \leq r\frac{\alpha}{2}n \tag{9}
\]

(The second inequality follows from the fact that this is a left \( r \)-regular graph and the size of \( A \) is initially bounded by \( \alpha n / 2 \)). Also during the course of the algorithm the
size of the set $U$ keeps decreasing by definition. (Either by flipping bits in the original error set, or bits not in the original error set). Hence at any point of the algorithm

$$|U| \leq \frac{\alpha}{2}n$$  \hspace{1cm} (10)

Thus we have a contradiction and hence the size of the set $A$ can never exceed $\alpha n$. Therefore the algorithm will eventually terminate with all constraints satisfied. We summarize the steps so far in the proof as follows:

1. The expansion property of the graph, and the upper bound on the size of $A$ guarantee that some variable in $A$ is eligible for flipping.
2. An eligible variable flipped at each iteration implies that the size of $U$ decreases.
3. The upper bound on the assumed number of errors implies an upper bound on size of $A$. Thus $U$ must finally become empty and we get the decoded codeword.

We have still to show that the algorithm decodes to the correct codeword under all the assumptions made. We can assume without loss of generality that the all-0 codeword was sent as this is a linear code. We have seen that at no point in the algorithm does the size of the set $A$ which contains all the erroneous bits exceed $\alpha n$. But there is no codeword of weight $\leq \alpha n$, so such a word can never be output by the algorithm. Since the degrees of the bipartite graph are constant, checking satisfied and unsatisfied neighbours is a constant time operation. Also since the size of the set $U$ decreases at each iteration and the number of constraints is upper bounded by the length of the code, the algorithm runs in linear time.

Sipser and Spielman also introduced a constructive family of asymptotically good linear error-correcting codes together with a simple parallel algorithm that will always correct a constant fraction of errors.

**Suggested Reading**


A Sketch of Modern Cryptology *
The Art and Science of Secrecy Systems

Palash Sarkar

1. Introduction

Those of us who are in the habit of reading detective stories will need no introduction to ciphers and code breaking. In the typical situation, the sleuth is presented with a cryptogram (remember the ‘dancing figures’ faced by Holmes) which he has to figure out. On success either a treasure trove is discovered or some vital clue is unearthed leading to the capture of the criminal. Another well-known application of secret messages is military communication. In a war, messages need to be exchanged between units of the same army in order to coordinate joint manoeuvres. Since such messages can easily fall into enemy hands, it should be ensured that none but the intended recipient can read the message. In fact, a system of exchanging secret messages was practised in the time of Julius Caesar, and the system is called ‘Caesar shift’ after him. The subject of cryptology has an ancient history. [1] covers cryptology from its initial use by the Egyptians some 4000 years ago, to the twentieth century where it played an important role in the outcomes of both the world wars.

However, in the present day, secure communication is no longer confined to the pages of a story book or to military communication. In the modern business world, vital information needs to be exchanged between parties for the successful completion of a transaction. Moreover, current business practices are becoming increasingly dependent on extensive use of computers and the internet. In fact, in the future, whole business transactions may perhaps be completed over the internet, giving rise to so called e-commerce. This possibility gives rise to various kinds of subtle security problems. For one thing, the exchange of information must be protected against unintentional eavesdropping. This however is only a basic requirement. Later we will briefly discuss some of the other interesting security problems of the modern business world.

2. Preliminaries

Before proceeding, let us briefly understand a few relevant terms. A cryptosystem is a mechanism for providing a secure means of exchanging information between two (or more) parties. A cryptographer is a person who designs cryptosystems to be used

Figure 1. Basic model of a private key cryptosystem.

by others. To complete the picture we need a cryptanalyst, whose task is to crack cryptosystems (much like our story book detectives).

In the basic model of a cryptosystem, there is a sender and a receiver who communicate with each other (see Figure 1). The channel of communication is called a public channel, i.e. all information sent over this channel can be picked up by anybody. In particular, the opponent or the cryptanalyst can pick up any or all information from the public channel. The actual message to be sent by the sender is varyingly called message text/plaintext/cleartext. The process of transforming the message into secure form is called encryption and the process of recovering the actual message from its secure form is called decryption. The encrypted form of the message is called ciphertext or simply cipher. A cryptosystem is specified by the encryption and decryption procedures. These procedures involve a secret key which is only known to the sender and the receiver. The secret key is exchanged between the sender and the receiver using a secret channel. To understand the difference between secret and public channels consider the following example: If two persons want to say something to each other such that no one else can hear, they can whisper in each other's ear. The whisperings correspond to communication over a secret channel. But one cannot whisper for very long. After some time they will have to speak loudly and everybody can hear them. This corresponds to communication over a public channel. In the basic model, the cryptanalyst has access to all the ciphertexts being exchanged over the public channel. Also the encryption and decryption procedures are known to him but not the secret key. The task of the cryptanalyst is to recover a message or the key from this information.

The encryption and decryption methods must be very fast to be useful in practice. However, the cryptanalyst can use a lot of time (even days) to crack a system. It is usually assumed that the cryptanalyst has the best possible practical computing power. Also in the current era of internet, the actual work may be distributed to many computers working simultaneously to crack a system. Moreover, the target of the cryptanalyst is also modest. The ability to crack even a small part of a cryptosystem will render such a system useless. So the situation is the following. A cryptographer designs a cryptosystem and the cryptanalyst cracks it. The cryptographer modifies the system or designs a new one and again the cryptanalyst attacks it. Thus the cryptographer and the cryptanalyst play a never ending game of trying to win against one another. Of course the word cryptology refers to this game and more generally to the whole subject of designing and cracking secrecy systems.
At this point it might be useful to point out the difference between cryptology and coding theory, which was covered in an earlier three part article in Resonance [2]. The aims of the two disciplines are different. In coding theory the basic goal is to introduce redundancy into the data so that any subsequent error may be detected and corrected. This also involves a transformation of information from one form into another and back again (usually called encoding and decoding). However, the encoded message can be decoded by any person who knows the coding scheme. This is unacceptable from a security point of view. In cryptology, it is always assumed that the cryptanalyst knows the cryptosystem (the encryption and decryption methods) and yet is incapable of knowing the message without the secret key. On the other hand, it will not be possible to properly recover an encrypted message if transmission errors creep in. Thus for secure and reliable transmission one usually encrypts a message (for security) and then encodes it using coding techniques (for reliability). On the receiver side, first the received message is decoded to remove transmission errors and then decrypted to get the original message.

For our purpose, messages and ciphers will usually be integers expressed in binary notation. Let us briefly see the rationale for this assumption. Suppose we have a message written in English. This will be stored as a computer file in the form of a sequence of bits. This sequence is divided into fixed size blocks of bits, where the length of each block is (say) 2048 bits. Then each block of bits can be considered to be the binary representation of an integer in the range 0 to $2^{2048} - 1$. Thus the original message can be represented by a sequence of integers. The encryption mechanism will encrypt one integer at a time, so that for the description of the encryption procedure it is sufficient to consider the message to be a single integer.

Let us first consider the basic problem of secure information exchange. Consider the scenario where $n$ persons want to communicate with each other and the communication between any two persons should not be intelligible to the others. Such a situation may arise in the stock market, where any pair of brokers may want to exchange information without any of the other brokers knowing what is being exchanged. A naive solution to the problem is to use a model of cryptosystem as in Figure 1. Each person maintains a list of $n - 1$ secret keys which are used for communication with the other $n - 1$ persons. When person $i$ wants to send a message to person $j$, he chooses from his list the secret key corresponding to $j$, and uses it to construct the cipher which he then sends to person $j$. When person $j$ gets the message from $i$, he uses the key corresponding to $i$ (which is the same key that person $i$ has corresponding to $j$) to decipher the message. In this scenario, for each pair of communications, one needs a secret key and thus this gives rise to a total of $\binom{n}{2}$ keys for the whole system.

Here $\binom{n}{2}$ is the binomial coefficient ‘$n$ choose 2’. So if there are 1000 brokers in a stock market each one of them will have a list of 999 secret keys and the system will
have a total of \( \binom{1000}{2} \) secret keys overall. Clearly maintaining and managing the secrecy of so many keys is a difficult administrative problem. Also a broker might need to communicate with some other broker very infrequently (or not at all). So it is not very sensible to maintain a secret key with such a person. Moreover, if a new broker enters the market, this person will have to establish a secret key with all the existing brokers, which is a time consuming and costly affair. A brilliant solution to this problem was first proposed by Diffie and Hellman in 1976 in a landmark paper entitled ‘New Directions in Cryptography’[3], where they introduced the concept of public key cryptography.

3. Public Key Cryptography

The novel idea in public key cryptography is for each user to have exactly two keys – an encryption key and a decryption key. The encryption key is made public, i.e. made known to everybody and the decryption key is kept secret. Going back to our stock market example, each broker has an encryption key and a decryption key. The encryption keys are published in a global (broker) directory and the decryption keys are kept secret by the respective brokers. Again suppose that broker A wants to send a message \( x \) (a positive integer as explained above) to broker B. Broker A chooses the encryption key \( e_B \) of broker B from the global directory and uses the publicly known encryption method to encrypt \( x \) to obtain a message \( y \), i.e., \( y = E(e_B, x) \), where \( E(\,., \,) \) is the encryption function and the key \( e_B \) and \( x \) are parameters to this function. This \( y \) is transmitted to broker B. On receiving \( y \), broker B uses the secret decryption key \( d_B \) and the publicly known decryption method to decrypt \( y \) and obtain \( x \), i.e. \( x = D(d_B, y) = D(d_B, E(e_B, x)) \). A little reflection will convince the reader that such a scheme removes all the difficulties explained in the previous section. Since the encryption key is publicly known, such cryptosystems are called public key cryptosystems (PKC) and the model described in Figure 1 is called private (or secret) key cryptosystem. Also in a PKC the encryption and decryption keys are different and hence they are sometimes called asymmetric key systems whereas secret key cryptosystems are called symmetric key systems.

Let us now consider what are the security requirements on such a system. The functions \( E(\,., \,) \) and \( D(\,., \,) \), the encryption key \( e_B \) and the cipher \( y \) are known. From these it should be infeasible to obtain either the message \( x \) or the secret decryption key \( d_B \). Viewed another way, it should be easy to obtain \( y \) from \( x \) but without the knowledge of \( d_B \) it should be difficult to obtain \( x \) from \( y \), i.e. computation in one direction is easy, while it is hard in the reverse direction. Functions satisfying such a criteria are called one-way functions. However, the encryption function used here is not exactly a one-way function, since knowledge of \( d_B \) makes it easy to go back. So \( d_B \) can be considered to be a sort of trapdoor which allows easy inversion. Hence the function \( E(\,., \,) \) is actually called a trapdoor one-way function. So to implement
a public key cryptosystem one has to design a trapdoor one-way function. The most popular and widely used system employing a trapdoor one-way function is the system proposed by Rivest, Shamir and Adleman and called the RSA system after them. To understand the workings of this system we need to know a bit of number theory, which is what we do next.

4. A Bit of Number Theory

Here we briefly cover some of the essential concepts of divisibility and related ideas. Let $m$ and $n$ be positive integers. Then it is possible to write $m = qn + r$, where $0 \leq r < n$ and $q$ and $r$ are unique integers (this is not difficult to prove). The integer $q$ is called the quotient and the integer $r$ is called the remainder. If $r = 0$, then $n$ is said to divide $m$ and is denoted by $n \mid m$. Given integers $a, b$ and $n > 0$, we write $a \equiv b \mod n$, if $n \mid (a - b)$. From these two facts it is easy to see that if we fix a positive integer $n$, then given any integer $m$, we can write $m \equiv a \mod n$ for some $0 \leq a < n$. The set of numbers $0, \ldots, n - 1$ is called a set of residues modulo $n$.

We now turn to the concept of greatest common divisor of two positive integers. Let $m$ and $n$ be two positive integers, then their greatest common divisor is denoted by $\gcd(m, n)$ and defined as the positive integer $g$ such that $g \mid m$ and $g \mid n$ and if any $y$ divides both $m$ and $n$, then $y$ divides $g$. From our school days we know of the following method to obtain $\gcd(m, n)$.

$$m = q_1n + r_1$$
$$n = q_2r_1 + r_2$$
$$r_1 = q_3r_2 + r_3$$
$$\vdots$$

$$r_{k+1} = q_{k+3}r_{k+2} + r_{k+3}.$$ 

Here $r_{k+3} = 0$ and $\gcd(m, n) = r_{k+2}$. (Can you prove that this provides the gcd as defined above?) This procedure for finding the gcd of two integers is called the Euclidean algorithm. There is another important result on the gcd of two numbers. Let $g = \gcd(n, m)$. Then it is possible to find two integers (not both positive) $\lambda$ and $\mu$, such that, $g = \lambda n + \mu m$. (Again it is a good exercise to try and prove this result from the Euclidean algorithm.)

Of special interest are integers $n$ and $m$ such that $\gcd(n, m) = 1$. In such a situation $n$ and $m$ are said to be coprime (or relatively prime) to each other. Let $n$ and $a$ be such that $\gcd(n, a) = 1$. Then it is possible to show (using the last-mentioned result on gcd) that there is an integer $b$ such that $ab \equiv 1 \mod n$. Moreover, using a modification of the Euclidean algorithm, it is also possible to find such a number $b$. This number $b$ is called the multiplicative inverse of $a$ modulo $n$. Also if we restrict only to numbers in the range $0$ to $n - 1$, then this number $b$ is unique. Given a positive integer $n > 1$, the
number of integers \( i \) in the range \( 1 \leq i \leq n - 1 \), such that \( \gcd(i, n) = 1 \) is the value of an important number theoretic function called the Euler totient and is denoted by \( \phi(n) \). In case \( n \) is a prime number, then \( \phi(n) = n - 1 \). Also if \( \gcd(n, m) = 1 \), then \( \phi(nm) = \phi(n)\phi(m) \). Now we are in a position to state two fundamental number theoretic results.

**Fermat’s Little Theorem:** Let \( p \) be a prime and \( a \) any integer coprime to \( p \). Then,

\[
a^{p-1} \equiv 1 \mod p.
\]

**Euler’s Generalisation:** Let \( n \) be any positive integer and \( a \) be coprime to \( n \). Then,

\[
a^{\phi(n)} \equiv 1 \mod n.
\]

An elementary textbook that discusses and proves the above results (and also other interesting results) is [4]. Here we do not require any more number theory and so let us get back to cryptology.

5. The RSA System

We now present an implementation of the PKC concept as proposed by Rivest, Shamir and Adleman (the RSA system). To set up the RSA system each user chooses two large (1024 bits or more) primes \( p \) and \( q \) and forms the product \( N = pq \). (For the moment let us assume that such primes can be chosen. Later we will see how this can be done.) From \( N \), find \( \phi(N) = \phi(pq) = \phi(p)\phi(q) = (p - 1)(q - 1) \). Next two positive integers \( e \) and \( d \) (of roughly the same size) are chosen using the Euclidean algorithm such that, \( 1 < e, d < \phi(N) \) and

\[
ed \equiv 1 \mod \phi(N).
\]

Once \( e \) and \( d \) are obtained, it is no longer required to preserve the individual values of \( p, q \) or \( \phi(N) \). The public key is declared to be the pair \( (e, N) \) and the private key which is kept secret is the pair \( (d, N) \). In fact only \( d \) is kept secret.

For encryption, as explained before the binary message is divided into blocks such that each block of bits is the binary representation of an integer less than \( N \) and is encrypted independently. To encrypt an integer \( x \) one uses the public key \( (e, N) \) and forms

\[
y = x^e \mod N.
\]

This \( y \) is the cipher corresponding to \( x \) and is transmitted. To decrypt, all that is required is to form,

\[
az \equiv y^d \mod N.
\]

This \( z \) is equal to \( x \) and hence the original message has been recovered. (\( z \equiv x^{ed} \mod N \equiv x^{1+k\phi(N)} \mod N \equiv x \mod N \). Note \( x^{1+k\phi(N)} \equiv x \mod N \) if and only if \( N \mid x(x^{k\phi(N)} - 1) \). Now use the fact that either \( p \mid x \) or \( q \mid x \) or \( \gcd(N, x) = 1 \).
Thus both encryption and decryption are similar operations (modulo exponentiation). Let us briefly understand how these operations can be carried out efficiently. The first thing to observe is that \( x < N \) and at no point of the computation do we allow the intermediate result to become greater than \( N \). This is achieved by reducing the intermediate results modulo \( N \) at each step and the correctness of this procedure is guaranteed from the following identity: 

\[
st \mod N = (s \mod N)(t \mod N) \mod N.
\]

The second question is how to perform \( x^a \mod N \). If we try to perform this by multiplying \( x \) to itself \( a \) times then we require \( a \) multiplications. Since the value of \( a \) is quite large (the binary representation will be several hundred bits), such a method will be impossible to carry out in practice. A practical method uses the following simple rule of exponentiation.

\[
x^a = \begin{cases} 
(x^{\frac{a}{2}})^2 & \text{if } a \text{ is even}, \\
x(x^{\frac{a-1}{2}})^2 & \text{if } a \text{ is odd}.
\end{cases}
\]

It is not difficult to verify that the above rule also holds when we are performing modulo exponentiation. Based on this rule it can be shown that the time to perform the computation \( x^a \mod N \) depends on \( \log_2 a \) instead of \( a \). The value of \( \log_2 a \) is approximately the length of the binary representation of \( a \) and hence the operation can be carried out quite fast by a computer.

Let us now briefly try to understand the security of the system. The secret key is \((d, N)\) which a cryptanalyst will try to recover. If from \( N \), one can obtain the factors \( p \) and \( q \) of \( N \), then it is easy to find \( \phi(N) \) and since \( e \) is known, one can also find \( d \) using the Euclidean algorithm. It is believed that if \( N \) is a large composite number it is difficult to obtain the factors of \( N \). (However, in February 2020 it has been possible to factor an integer \( N \) which is 829 bits long.) Thus trying to break RSA by factoring \( N \) is going to be difficult. So one might try to obtain \( d \) by other ways. However, it can be shown that if one can obtain \( d \) or \( \phi(N) \) from \( N \), then one can also find \( p \) and \( q \), i.e., factorise \( N \). It is not difficult to show that if one can obtain \( \phi(N) \) then one can also obtain \( p \) and \( q \), but showing that if one can obtain \( d \), then one can almost certainly obtain \( p \) and \( q \) is more difficult and involves certain probabilistic arguments. Since all known attacks on RSA ultimately boil down to the problem of factoring \( N \), it is generally believed (but not proved) that breaking the RSA system is as hard as factoring \( N \). In the next section we discuss the two problems of finding a large prime and factoring a large composite integer.

6. Primality Testing and Factoring

The problem of primality testing is to ascertain whether a given number is prime. At first glance, factoring and primality testing appear to be similar problems. But in fact there is a gulf of difference between the two and on this difference rests the strength of the RSA cryptosystem. The second problem has satisfactory practical
solutions but the first problem is indeed a very difficult one. In fact it has withstood determined attacks by many brilliant researchers. Of course, the ‘problems’ really become problems when we are tackling large numbers – about 615 decimal digits or roughly 2048 bits long.

The naive algorithm for primality testing is very simple. To test a number $n$, try dividing by all positive integers greater than 1 and less than or equal to $\sqrt{n}$. If none divides $n$, then $n$ is a prime (why?). However, this will require too much time when $n$ is of the size used in cryptography (see above). Instead a probabilistic algorithm is used with a very high probability of success. More precisely, if $n$ is declared to be composite then the answer is correct, but if the number is declared to be prime then it may actually be composite but the probability of such an event is very low (typically $< \frac{1}{2^{200}}$). Such probabilistic algorithms are called Monte Carlo algorithms.

The key idea in probabilistic primality testing is to find a witness to the compositeness of $n$, i.e., if $n$ is composite, we want a number $b$ which can be used to demonstrate this. For example if $n = 525$ and $b = 50$, then since $\gcd(525, 50) > 1$, 50 is a witness to the compositeness of 525. Again let $n = 21$ and $b = 4$. Now $\gcd(4, 21) = 1$, but it is still a witness to the compositeness of 21 if we use Fermat’s theorem (see Section 4). By Fermat’s theorem for any prime $p$ and integer $b$ coprime to $p$, we have $b^{p-1} \equiv 1 \pmod{p}$. However, $4^{20} \equiv 16 \pmod{21}$ (verify!), and so Fermat’s theorem fails and hence 21 is composite. Motivated by this we introduce the notion of a ‘witness function’ $W_n(b)$, which evaluates to true if $b$ is a witness to the compositeness of $n$. Note that if we fix a witness function (such as $\gcd(b, n) > 1$ or Fermat’s theorem) and a composite number $n$, then it is possible to obtain $b$ such that $W_n(b)$ is false. Such $b$’s are called nonwitnesses. For example, if the witness function is $\gcd(n, b) > 1$, then for $n = 21$, the integer 4 is a nonwitness. Thus if $W_n(b)$ evaluates to false we cannot immediately conclude that $n$ is prime, but if $W_n(b)$ is true then we know for sure that $n$ is composite. Our aim now is to randomly pick integers $b$ in the range 1 to $n-1$ and evaluate $W_n(b)$. If $n$ is composite, then we hope to obtain a $b$ such that $W_n(b)$ is true, i.e., $b$ is a witness to the compositeness of $n$. For this procedure to succeed the following condition must be satisfied. If $n$ is composite, then there must be a large number of witnesses $b$ such that $W_n(b)$ is true. (By a large number we mean that a constant fraction of the numbers in the range 1 to $n-1$ must be witnesses.) If this condition is satisfied we use the following procedure.

1. Pick several distinct $b$’s uniformly at random from the integers 1 to $n-1$ and evaluate $W_n(b)$. Typically one will choose 200 $b$’s.

2. If for any $b$ chosen in Step 1, $W_n(b)$ is true, then we declare $n$ to be composite.

3. If all the evaluations in Step 1 are false, then we declare $n$ to be prime.

If the procedure results in the answer ‘$n$ is composite’, then we are sure the answer is correct. However, if the procedure results in the answer ‘$n$ is prime’ then it is possible that the answer is wrong, but the probability that it is wrong is equal to the probability...
that \( n \) is composite and all the \( b \)'s are nonwitnesses. Since there are a large number of witnesses the probability of this event is very low.

Now we come to the question of choosing a good witness function. If our witness function is \( \gcd(n, b) \), then the number of witnesses with respect to this function can be very small. For example, if \( n \) is the product of two primes (the kind of composite numbers used in RSA), then the number of witnesses in the range 1 to \( n - 1 \) is only 2. A solution may be to use Fermat's theorem as the witness function, i.e., \( W_n(b) \) is true if \( b^{n-1} \equiv 1 \mod n \) for some \( b \) coprime to \( n \). However, this cannot be directly used since there are composite numbers \( n \) such that for all numbers \( b \) coprime to \( n \), \( b^{n-1} \equiv 1 \mod n \). Such composite numbers are called Carmichael numbers. For example \( n = 561 = 3 \times 11 \times 17 \) is the smallest Carmichael number. A way around this problem is to strengthen Fermat's theorem to build a witness function. One such witness function (called the Miller–Rabin strong pseudoprime test) is the following.

Given \( n \) and \( b \), the function \( W_n(b) \) is true if the following holds:

1. \( 1 \leq b < n \).
2. (a) \( b^{n-1} \not\equiv 1 \mod n \) or
   (b) There exists \( i \), such that \( 2^i \mid (n-1) \) and \( 1 < \gcd(b^{n-1}-1, n) < n \).

Condition 1 enforces \( b \) to be less than \( n \) and condition 2 enforces \( n \) to be composite. For if \( 2(a) \) holds then Fermat's theorem is violated and if \( 2(b) \) holds then \( n \) has a nontrivial divisor. It can be shown that the number of \( b \)'s such that \( W_n(b) \) is true is greater than \( \frac{1}{4}(n-1) \). Also \( W_n(b) \) can be evaluated very fast. So \( k \) distinct positive integers \( b_1, \ldots, b_k \) less than \( n \) are chosen and \( W_n(b_i) \) is evaluated for each \( i \). If for some \( i \), \( W_n(b_i) \) is true, then \( n \) is composite, else declare \( n \) to be a prime. Now if \( n \) is actually a prime then the result is correct since \( W_n(b) \) will always evaluate to false. On the other hand the algorithm may declare a composite number to be a prime, but the probability of this is less than \( \frac{1}{4^k} \). So if \( k \) is 200 then the probability of error is less than \( \frac{1}{4^{200}} \), which is negligible.

The above procedure gives a method for testing whether a given number is prime. But how do we actually obtain a large prime number, which is a basic requirement to set up the RSA system? One approach is to choose a random integer and test whether it is prime. The natural question to ask is what our chances are of obtaining a prime number by this method. To get the answer we must invoke the prime number theorem, a result of fundamental importance to number theory. The prime number theorem states that the number of primes not exceeding \( n \) is approximately \( \frac{n}{\ln n} \). (Here ‘\( \ln \)' denotes log to the natural base.) Hence if \( p \) is chosen at random, the probability that it is prime is about \( \frac{1}{\ln p} \). For a 512-bit modulus, we have \( \frac{1}{\ln p} = \frac{1}{177} \). So, on average, among 177 random positive integers of appropriate size one will be prime. (If we restrict to odd integers, the probability doubles to \( \frac{2}{177} \).) Thus we have a procedure to obtain large prime numbers.
Unfortunately for the factoring problem no such algorithm exists. Sophisticated techniques are being used but to date there is no ‘good’ algorithm for finding the factors of a large composite number. If this can be achieved then the RSA system becomes vulnerable. Currently several research groups across the world are trying very hard to factor large composite numbers. In February 2020, a research group successfully factored a number which is 829 bits long. So RSA systems must employ significantly larger modulus $N$ to assure security.

7. Other Security Problems

Here we briefly discuss several other security problems relevant to modern business world and e-commerce. As must be evident by now, these problems and their solutions are heavily dependent on the use of computers and the internet. In fact, the revolution in modern cryptology is an outgrowth of the proliferation of computer technology. In the following subsections, we outline only the problems. Except for digital signatures, we do not attempt to provide any solution to these problems. Though extremely interesting, these solutions require background material not covered in this article.

7.1 Digital Signatures

In future, complete business transactions are going to be conducted over the internet. As we all know, the heart of any business transaction is signed documents, which represent the signer’s commitment to the transaction. So if transactions are going to be ‘paperless’ we must be able to provide methods to sign documents digitally through computers. A conventional signature has several properties which have to be captured by digital signature schemes. One of the basic goals is to provide authentic documents, i.e., a third person can be sure that a document is genuine if it is properly signed. The actual process of this verification is called authentication. Authentication facility provides security against the following situations:

1. A ‘planted’ message.
2. A person denying the sending of a message.
3. A person claiming a message to have come from some other person when it actually did not.

Note that secrecy and authentication are independent concepts. A transmitted message may possess one of these without possessing the other. Next we briefly describe how the RSA system may be modified to obtain both secrecy and digital signature authentication.

Suppose Alice wants to send a message to Bob. Let $(e_1, N_1)$ and $(d_1, N_1)$ be the public and the private keys for Alice. Let $(e_2, N_2)$ and $(d_2, N_2)$ be the same for Bob. Suppose Alice wants to send a message $x$ (an integer) to Bob. Alice performs the following steps.
1. \textit{(Sign the message)}: \( s = x^d \mod N_1 \).

2. \textit{(Encrypt the message)}: \( y = s^{e_2} \mod N_2 \).

3. Transmit \( y \) to Bob.

Bob performs the following steps to recover and verify the message.

1. \textit{(Decrypt the message)}: \( z = y^{d_2} \mod N_2 \). If there has been no tampering, then \( z = s \).

2. \textit{(Authenticate)}: \( w = s^{e_1} \mod N_1 \). The \( w \) should be equal to \( x \). Thus Bob obtains the pair \((x, s)\) with all the properties of a signature.

### 7.2 Identification Schemes

Suppose a bank has branches all over the country. A person (say Alice) opens an account at one of these branches but should be able to withdraw money from any other branch. Of course we do not want a distant branch to actually verify Alice’s signature with the parent branch before issuing money. This would cause unacceptable delay in payment. So the parent branch issues a secret personal identification number (PIN) to Alice. When Alice wants to withdraw money from a remote branch, she enters her PIN in a computer in the remote branch and this computer then verifies the PIN with a computer in the parent branch. This verification process is completed over a computer network. The problem here is that the PIN will be transmitted over an insecure network and the secrecy of the PIN is compromised the moment Alice reveals it. So we have to design a scheme by which the branch will be convinced that Alice is the proper person but Alice will not reveal her PIN. This is a seemingly impossible problem. In fact, one of the most interesting aspects of cryptology is to provide solutions to seemingly impossible problems. In this case, the solution is for both Alice and the verifying computer to do some computation and exchange results. From these results the verifying computer should be able to decide on Alice’s identity. At no point in the protocol is the actual value of the PIN transmitted and hence it remains a secret. Since Alice is also expected to do some computation, she should also have access to a computer. In practice, Alice will be provided with a card which has a small computer chip built into it. This chip will do the computation on Alice’s behalf. Such cards with built-in computer chips are called smart cards.

### 7.3 Secret Sharing Schemes

Those of us who have used bank lockers are aware that any such locker has two keys. One is given to the customer while the other is kept with the manager. The locker cannot be opened with any one of these keys but it can be opened if both are used. Viewed another way, if a locker has been opened then it must be the case that both the customer and the manager participated in the process. From this it is easy to imagine situations where a secret has to be shared among \( n \) persons, but at least \( k \)
persons must cooperate to learn the secret. The initial distribution is done by some trusted authority who computes the ‘share’ of each of the $n$ persons. The secret to be shared is some integer $x$. So each of the persons involved in the process gets to know something about $x$, but at least $k$ persons must get together to know the value of $x$. One elegant solution to this problem was proposed by Adi Shamir in the late seventies. More general versions of this problem have been studied later. Discussion of the solutions requires some knowledge of linear algebra and Boolean functions.

7.4 Zero Knowledge Proofs

Suppose Alice wants to convince Bob that she knows some secret which she wants to sell to Bob. Before Bob pays for the secret, he must be convinced that Alice indeed knows the secret. So he asks Alice to prove that she knows the secret. Alice can prove this by simply revealing the secret. But then Bob does not need to pay Alice any more. So she must prove to Bob that she knows the secret without revealing the secret. The solution is for Alice and Bob to participate in a challenge-and-response protocol at the end of which Bob should be convinced that Alice knows the secret and Alice should be convinced that Bob has gained no knowledge (zero knowledge) of the secret. The study of this problem requires ideas from number theory as well as the modern theory of computational complexity.

8. Conclusion

In this article a bare outline of the subject of cryptology has been presented. The basic purpose is to present cryptology as an exciting and rapidly developing subject. Another aim is to highlight the role of the subject in modern computer based businesses. As our society becomes more and more dependent on computers, the role of cryptology and data security will become extremely important and lead to new challenges and ingenious solutions. A good textbook on the subject is [5].

Acknowledgements

The original motivation for this article was provided by T Krishnan quite some time back. The author is grateful to him for the original motivation as well as his recent interest in the article after the lapse of a considerable amount of time.

Postscript (dated February 7, 2022)

The original article was written more than twenty years ago. The theory and practice of cryptology has changed tremendously in the last two decades. The goal of this postscript is to say a few words on the present state of the subject.
Symmetric key cryptography forms the backbone of encryption technology and is used in conjunction with public key encryption to perform bulk encryption. An important development in the area of symmetric key cryptography has been the adoption of the advanced encryption standard (AES) by the National Institute of Standards and Technology of the USA. Modern important applications of symmetric key encryption include securing internet traffic and full disk encryption.

Elliptic curves were independently proposed by Koblitz and Miller for the construction of public key cryptographic schemes. The security of such schemes is based on the discrete logarithm problem over elliptic curve groups. In comparison to the RSA system, appropriately chosen elliptic curves have smaller key sizes and faster encryption and decryption algorithms. Over the years, elliptic curve based encryption and signature schemes have continuously replaced RSA based schemes in various applications.

Quantum computers pose a serious challenge to public key cryptography. Shor’s quantum algorithm can be used to efficiently solve both the integer factorisation and elliptic curve discrete logarithm problems. If and when quantum computers are built, the present day public key encryption schemes will become insecure. This has led to a great amount of research into designing cryptosystems which cannot be broken even with a quantum computer. While several mathematical approaches have been proposed, the present state of such post-quantum cryptography is fluid, which will hopefully become stable within the next few years. So the story of cryptology, which is a never-ending conflict between designing and breaking cryptosystems, continues...

Suggested Reading

Petri nets offer a versatile modeling framework for complex, distributed, concurrent systems and have been used in a wide range of modeling applications. The first part of this two-part article provides an overview of Petri nets and presents important conceptual underpinnings of Petri net modeling.

1. Introduction

Modeling is a central part of all activities that lead up to the design, implementation, and deployment of systems. We build models so that we can understand better the system we are developing. Models enable to communicate the desired structure and behavior of our system and provide a basis for designing high-performance systems.

Petri nets constitute a versatile modeling tool applicable to many systems. They are graphical in nature and are backed up by a sound mathematical theory. They can be used by practitioners and theoreticians alike and their applications range over a wide variety of disciplines. They have been primarily used to describe and study discrete event dynamical systems. A discrete event dynamical system is a system in which the evolution of the activities in time depends on the occurrence of discrete events. Examples of such systems include computer systems, automated manufacturing systems, communication networks, air traffic networks, power systems, office automation systems, business processes, etc. In a computer system, for example, typical discrete events include: arrival of a new job into the system; finishing of execution of a program; commencement of an I/O operation; or a disk crash. When such an event occurs, the state of the system might change; some old events may get disabled; and some new events may get enabled. In order to capture the structure and dynamics of such a system, Petri nets offer a natural and effective modeling methodology.

In Part 1 of this article, we will first briefly trace the history of Petri net modeling (Box 1). We will then understand the notation and meaning of Petri net models, using an illustrative example (that of a simple manufacturing plant with two machine centres). Following this, we will understand how Petri net models can capture the behavior or dynamics of a modeled system. Next, we will look into the representational power of Petri nets while modeling systems.

In Part 2, we will first illustrate Petri net modeling through a representative example,
Petri nets have an interesting history and have come quite a long way from the time Carl Adam Petri proposed them in his doctoral work in the early 1960s. His doctoral dissertation was submitted to the faculty of Mathematics and Physics at the Technical University of Darmstadt, West Germany in 1962. Petri currently works in an institution called GMD in Bonn, Germany. The primary motivation behind Petri’s work was to model concurrency and asynchronism in distributed systems through a formalism more powerful than finite state automata. Petri’s pathbreaking work came to the attention of A W Holt in the mid-1960s. Holt led the Information System Theory project of Applied Data Research in the United States. This project brought out a series of influential reports on Petri net theory in the mid and late 1960s. From 1970 to 1975, the Computation Structures Group at the Massachusetts Institute of Technology became a leading centre for Petri net research and from then on, Petri nets became an active research area in several universities, particularly in Europe. Starting from 1980, annual conferences have been held, initially called as the European Workshops on Applications and Theory of Petri Nets and currently called as the International Workshops on Applications and Theory of Petri Nets. These workshops are meant exclusively for Petri net related papers.

Research and development in the area of Petri nets can be categorized into several streams. The research in the 1960s and 1970s was mostly on Petri net theory with less emphasis on applications. The theory focused on issues such as Petri net languages; characterization of Petri nets as a model of computation; and qualitative analysis of systems. During the 1980s, many Petri net based packages were developed for use in modeling and analysis of concurrent systems. Significant applications of Petri net theory were explored in the 1980s, primarily in the areas of computer operating systems, distributed computer systems, computer networks, and automated manufacturing systems. Timed Petri nets or stochastic Petri nets became prominent in these applications and Petri nets emerged as a major tool for quantitative performance analysis of systems. 1990s have seen the emergence of user-friendly modeling and analysis tools based on Petri nets for both qualitative and quantitative analyses and even non-specialists are able to effectively use these powerful tools. Interestingly, to this day, several researchers, including Carl Adam Petri, continue to advance the theory of Petri nets in many interesting directions.

that of the dining philosophers problem. We will then understand how important system properties are captured by the dynamics of a Petri net model of the system. Following this, we investigate different ways in which Petri nets models can be used in system modeling.

2. Classical Petri Nets

Petri nets are bipartite graphs and provide an elegant and mathematically rigorous modeling framework for discrete event dynamical systems. In this section an overview of Petri nets is presented with the aid of several definitions and an illustrative example.
In the following, \( N \) and \( R \) denote respectively the set of non-negative integers and the set of real numbers.

We start with some elementary definitions in classical Petri nets and illustrate the definitions with some examples.

**Definition:** A Petri net is a four-tuple \((P, T, IN, OUT)\) where

\[
P = \{p_1, p_2, \ldots, p_n\} \text{ is a set of places}
\]

\[
T = \{t_1, t_2, \ldots, t_m\} \text{ is a set of transitions } P \cup T \neq \emptyset, P \cap T = \emptyset
\]

\(IN : (P \times T) \to N\) is an input function that defines directed arcs from places to transitions, and

\(OUT : (T \times P) \to N\) is an output function that defines directed arcs from transitions to places.

Pictorially, places are represented by circles and transitions by horizontal or vertical bars. If \(IN(p_i, t_j) = k\), where \(k > 1\) is an integer, a directed arc from place \(p_i\) to transition \(t_j\) is drawn with label \(k\). If \(IN(p_i, t_j) = 1\), we include an unlabeled directed arc. If \(IN(p_i, t_j) = 0\), no arc is drawn from \(p_i\) to \(t_j\). Some places have black dots or tokens within them. The significance of the tokens will be introduced soon.

Places of Petri nets usually represent conditions or resources in the system while transitions model the activities in the system. In all subsequent definitions, we assume a Petri net \((P, T, IN, OUT)\) as given in the above definition. Also, we assume that the index \(i\) takes on the values \(1, 2, \ldots, n\), while the index \(j\) takes on the values \(1, 2, \ldots, m\).

**Example 1.** Let us consider a simple manufacturing system comprising two machines \(M_1\) and \(M_2\) and processing two different types of parts. Each part type goes through one stage of operation, which can be performed on either \(M_1\) or \(M_2\). On completion of processing, the part is unloaded from the system and a fresh part of the same type is loaded into the system. *Figure 1* depicts a Petri nets model of this system and *Table 1* gives the interpretation of the places and transitions in the model. For this Petri net,

\[
P = \{p_1, p_2, \ldots, p_8\}; \quad T = \{t_1, t_2, \ldots, t_8\}
\]

The directed arcs represent the input and output functions \(IN\) and \(OUT\), respectively. For example,

\[
IN(p_1, t_1) = 1; \quad IN(p_6, t_2) = 0.
\]

\[
OUT(p_5, t_1) = 1; \quad OUT(p_6, t_6) = 0.
\]

Note in the above example that the maximum weight of each arc is 1. Such a Petri net can be adequately described by a simpler notation \((P, T, A)\) where \(P\) and \(T\) have the usual significance and \(A\) is the set of arcs such that

\[
A \subseteq (P \times T) \cup (T \times P).
\]
Figure 1. Petri net model of a simple manufacturing system.

Table 1. Places and transitions in the Petri net model of a manufacturing plant.

<table>
<thead>
<tr>
<th>Places</th>
<th>Transitions</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_1$: Raw parts of type 1</td>
<td>$t_1$: $M_1$ starts processing a part of type 1</td>
</tr>
<tr>
<td>$P_2$: Machine $M_1$ available</td>
<td>$t_2$: $M_1$ starts processing a part of type 2</td>
</tr>
<tr>
<td>$P_3$: Raw parts of type 2</td>
<td>$t_3$: $M_2$ starts processing a part of type 1</td>
</tr>
<tr>
<td>$P_4$: Machine $M_2$ available</td>
<td>$t_4$: $M_2$ starts processing a part of type 2</td>
</tr>
<tr>
<td>$P_5$: $M_1$ processing a part of type 1</td>
<td>$t_5$: $M_1$ finishes processing a part of type 1</td>
</tr>
<tr>
<td>$P_6$: $M_1$ processing a part of type 2</td>
<td>$t_6$: $M_1$ finishes processing a part of type 2</td>
</tr>
<tr>
<td>$P_7$: $M_2$ processing a part of type 1</td>
<td>$t_7$: $M_2$ finishes processing a part of type 1</td>
</tr>
<tr>
<td>$P_8$: $M_2$ processing a part of type 2</td>
<td>$t_8$: $M_2$ finishes processing a part of type 2</td>
</tr>
</tbody>
</table>

Indeed, the use of an arc weight greater than unity is only a matter of convenience since a Petri net with arc weights greater than unity can always be represented by another Petri net having maximum arc weight unity.

Example 2. For the Petri nets of example 1, the set $A$ is given by

$$
\{(p_1, t_1), (p_2, t_1), (p_2, t_2), (p_3, t_2), (p_1, t_3), (p_4, t_3), (p_5, t_4), (p_4, t_4),
(p_5, t_5), (p_6, t_6), (p_7, t_7), (p_8, t_8), (t_1, p_5), (t_2, p_6), (t_3, p_7), (t_4, p_8),
(t_5, p_1), (t_5, p_2), (t_6, p_2), (t_6, p_3), (t_7, p_1), (t_7, p_4), (t_8, p_3), (t_8, p_4)\}.
$$
Definition: Given a transition $t_j$, the set of input places of $t_j$, denoted by $IP(t_j)$ and the set of output places of $t_j$, denoted by $OP(t_j)$, are defined by

\[
IP(t_j) = \{p_i \in P : IN(p_i, t_j) \neq 0\}
\]
\[
OP(t_j) = \{p_i \in P : OUT(p_i, t_j) \neq 0\}
\]

Definition: Given a place $p_i$, the set $IT(p_i)$ of input transitions of $p_i$ and the set $OT(p_i)$ of output transitions of $p_i$ are defined by

\[
IT(p_i) = \{t_j \in P : OUT(p_i, t_j) \neq 0\}
\]
\[
OT(p_i) = \{t_j \in P : IN(p_i, t_j) \neq 0\}
\]

Example 3. For the Petri net of Figure 1, we have

\[
IP(t_1) = OP(t_5) = \{p_1, p_2\}; \quad IP(t_5) = OP(t_1) = \{p_5\}
\]

The other sets of input places and output places can be obtained similarly. Also,

\[
IT(p_1) = \{t_5, t_7\}; \quad OT(p_1) = \{t_1, t_3\}
\]

The other sets of input transitions and output transitions can be obtained similarly.

Definition: Let $T_1$ be a subset of $T$. The transitions of $T_1$ are said to be conflicting if

\[
\bigcap_{t \in T_1} IP(t) \neq \emptyset
\]

and concurrent if

\[
IP(t_j) \cap IP(t_k) = \emptyset \quad \forall t_j, t_k \in T_1.
\]

Example 4. In the Petri net of Figure 1, the sets of transitions that are conflicting are $t_1, t_3; t_1, t_2, t_4, t_5; t_3, t_4$. Some of the concurrent sets of transitions are $t_1, t_4; t_2, t_3; t_5, t_6; t_1, t_8$. Petri nets capture concurrency of activities through concurrent transitions and non-deterministic activities through conflicting transitions. Further, they can also model co-existence of concurrent and non-deterministic activities. Elegant representation of such features is an important facet of Petri net modeling.

Definition: A marking $M$ of a Petri net is a function $M : P \rightarrow N$. A marked Petri net is a Petri net with an associated marking.

A marking of a Petri net with $n$ places is an $(n \times 1)$ vector, which associates with each place a certain number of tokens represented by black dots, and represents a state of the Petri net. We always associate an initial marking $M_0$ with a given Petri net model. In the rest of the article, we use the words state and marking interchangeably. Also, unless otherwise specified, a Petri net henceforth will refer to a marked Petri net.
Example 5. In Figure 1, the marking of the Petri nets is given by

\[ M_0 = \begin{bmatrix} M_0(p_1) \\ \vdots \\ M_0(p_8) \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \]

This corresponds to a state of the system when both machines are available for use and one fresh part of each type is waiting to be processed.

3. Dynamics of Petri Nets

Petri nets model both the structure and behavior of systems. Structural modeling is accomplished through the graphical structure and the input-output relationships among the places and transitions. Behavioral modeling is achieved through execution of firing rules (also called as the token game) that capture the dynamics of the modeled system over time. First we introduce the important notion of reachability set for a Petri net.

Definition: A transition \( t_j \) of a Petri net is said to be enabled in a marking \( M \) if

\[ M(p_i) \geq IN(p_i, t_j) \quad \forall \quad p_i \in IP(t_j) \]

An enabled transition \( t_j \) can fire at any time. When a transition \( t_j \) enabled in a marking \( M \) fires, a new marking \( M' \) is reached according to the equation

\[ M'(p_i) = M(p_i) + OUT(p_i, t_j) - IN(p_i, t_j) \quad \forall \quad p_i \in P \quad (1) \]

We say marking \( M' \) is reachable from \( M \) and write \( M \xrightarrow{t_j} M' \).

We consider that every marking is trivially reachable from itself by firing no transition. Also, if some marking \( M_j \) is reachable from \( M_i \) and \( M_k \) is reachable from \( M_j \), then it is easy to see that \( M_k \) is reachable from \( M_i \). Thus reachability of markings is a reflexive and transitive relation on the set of markings.

Definition: The transitive closure of the reachability relation, which comprises all markings reachable from the initial marking \( M_0 \) by firing zero, one, or more transitions, is called the reachability set of a Petri net with initial marking \( M_0 \). It is denoted by \( R[M_0] \).

Definition: For a marked Petri net with initial marking \( M_0 \), the reachability graph is a directed graph \((V, E)\) where \( V = R[M_0] \) is the set of vertices, and \( E \), the set of directed arcs, is given by: \((M_1, M_2) \in E \) if
(i) $M_1, M_2 \in R[M_0]$ and (ii) either there exists a transition $t \in T$ such that $M_1 \xrightarrow{t} M_2$ or there exists a set, $T_1 \subseteq T$, such that $T_1$ is a set of concurrent transitions by firing all of which $M_1$ reaches $M_2$.

In the reachability graph, the nodes are labeled by the markings they represent and the directed arcs are labeled by the transition or the set of concurrent transitions whose firing takes the source node to the destination node.

Example 6. In the marked Petri net of Figure 1, the transitions $t_1, t_2, t_3,$ and $t_4$ are all enabled. When $t_1$ fires, the new marking reached is $M_1$ where $M_1 = (00111000)^T$. Thus, $M_0 \xrightarrow{t_1} M_1$. Also, $M_0 \xrightarrow{t_2} M_2$, $M_0 \xrightarrow{t_3} M_3$, and $M_0 \xrightarrow{t_4} M_4$ where $M_2 = (11000000)^T$, $M_3 = (10010100)^T$, and $M_4 = (01100010)^T$. It can be shown that $R[M_0] = \{M_0, M_1, M_2, M_3, M_4, M_5, M_6\}$ where the details of the markings are given in Table 2. Figure 2 gives the reachability graph of this Petri net.

### Table 2. Reachable markings of the Petri net of Figure 1.

<table>
<thead>
<tr>
<th>Marking</th>
<th>$P_1$</th>
<th>$P_2$</th>
<th>$P_3$</th>
<th>$P_4$</th>
<th>$P_5$</th>
<th>$P_6$</th>
<th>$P_7$</th>
<th>$P_8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_0$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$M_1$</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$M_2$</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$M_3$</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$M_4$</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$M_5$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$M_6$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
4. Representational Power

The typical structural and behavioral characteristics exhibited by the activities in a complex system, such as concurrency, decision making, synchronization, and priorities, can be modeled elegantly by Petri nets. We have already seen in Example 4 how concurrency and conflicts are represented. Here we identify Petri net constructs for representing characteristics of various features. Figure 3 depicts these constructs.

Figure 3. Petri net primitives to represent system features.
Sequential Execution: In Figure 3(a), transition $t_2$ can fire only after the firing of $t_1$. This imposes the precedence constraint ‘$t_2$ after $t_1$’. Such precedence constraints are typical of the execution of jobs in any system. Also, this Petri nets construct models the causal relationship among activities.

Conflict: Transitions $t_1$, $t_2$, and $t_3$ are in conflict in Figure 3(b). All are enabled but the firing of any leads to the disabling of the other transitions. Such a situation will arise, for example, when a resource has to choose among jobs or a job has to choose among several resources. The resulting conflict may be resolved in a purely non-deterministic way or in a probabilistic way, by assigning appropriate probabilities to the conflicting transitions. The above primitive also implies that the transitions are mutually exclusive, i.e. one and only one of them may fire at a given time. Mutual exclusion is an important feature in all systems where there are shared resources.

Concurrency: In Figure 3(c), the transitions $t_1$, $t_2$, and $t_3$ are concurrent. Concurrency is an important attribute of system interactions. Note that a necessary condition for transitions to be concurrent is the existence of a forking transition that deposits a token in two or more output places.

Synchronization: Often, jobs in a system wait for resources and resources wait for appropriate jobs to arrive (as in assembly operations). The resulting synchronization of activities can be captured by transitions of the type shown in Figure 3(d). Here, $t_1$ will be enabled only when each of its input places has a minimum appropriate number of tokens. In the present case, $p_1$ and $p_2$ have tokens but $p_3$ does not have a token. Therefore, for $t_1$ to be enabled, we will have to wait for a token to arrive into $p_3$. The arrival of a token into this place could be the result of a possibly complex sequence of operations elsewhere in the rest of the Petri net model.

Merging: When parts from several streams arrive for service at the same resource, the resulting situation can be depicted as in Figure 3(e). Another example is the arrival of several jobs from several sources to a centralized location.

Confusion: Confusion is a situation where concurrency and conflicts co-exist. An example is depicted in Figure 3(f). Both $t_1$ and $t_3$ are concurrent while $t_1$ and $t_2$ are in conflict, and $t_2$ and $t_3$ are also in conflict.

Priorities: The classical Petri nets discussed so far have no mechanism to represent priorities. Inhibitor nets include special arcs called inhibitor arcs to model priorities. A portion of an inhibitor net is shown in Figure 3(g). $p_2$ is called an inhibitor place of $t_2$. An inhibitor arc from $p_2$ to $t_2$ is drawn as shown in Figure 3(g). The transition $t_2$ is enabled only if $p_1$ has a token and $p_2$ does not have a token. This enables a higher priority to be given to $t_1$ over $t_2$. In Figure 3(g), for example, $t_1$ is enabled but not $t_2$ because $p_2$ has a token. It is to be noted that the reachability set of an inhibitor net is a subset of the same net but with the inhibitor arcs removed. Inhibitor arcs enhance the modeling power of the Petri net model and indeed, it has been shown that Petri nets with inhibitor arcs are equivalent in power to the Turing machines. It has been
proved that the classical Petri nets, i.e., without inhibitor arcs, are less powerful than
Turing machines (see Article-in-a-box, Resonance, Vol. 2, No. 7, July 1997) and can
only generate a proper subset of context-sensitive languages.

In this article thus far, we have looked into the history of Petri nets; important notation
and semantics of Petri net models; and their modeling power. In the second part of
the article [6], we will first present the example of the dining philosophers problem to
illustrate Petri net modeling. Then we present features of Petri nets that make them an
attractive and versatile modeling tool and provide an overview of their applications.

Suggested Reading

Hall, Englewood Cliffs, New Jersey, 1992 (also available from Prentice Hall of India, New Delhi, 1998).
article).
1. Introduction and History

Several areas of science have ‘bootstrap methods’. They arise, for example, as techniques for numerical solution of differential equations. Often the name conveys the impression of getting something for nothing, like pulling oneself up by one’s bootstraps. Thus, a bootstrap method attempts to solve a problem without, on the surface, appearing to bring adequate information or expertise to bear on it.

Consider, for example, the proposal that we obtain information about the distribution of the population from which a sample was drawn, by sampling from the sample. On the surface, this appears to be a pointless exercise; we can’t get any more statistical information out of a sample simply by sampling from it. Bootstrap methods in statistics involve just this sort of operation – ‘resampling’, as it has come to be called, from the sample. It turns out to be very useful indeed, but of course it does not provide any new statistical information. What it can do is help us estimate parameters, or assess the variability of parameter estimates, using information that is already in the sample.

The name ‘bootstrap’ was conferred upon the corresponding class of statistical methods by Efron [1], in the seminal paper in which he introduced, developed and popularised this approach to inference. In addition to making a great many particular methodological contributions, in numerous papers over the next two decades, Efron had two critical insights: he saw that the notion of a ‘parameter’ could be interpreted very widely when using bootstrap methods, and he recognised that in cases where a bootstrap estimator was so complex that it could not be computed directly, it could be approximated to arbitrary accuracy by simulating, on a computer, the operation of resampling from the sample. Such a computer-intensive approach was not possible, for most of us, much before the publication of Efron’s work in the late 1970s. Thus, the time of bootstrap methods in statistics had come; their general application is inexorably linked to the development of modern computers and software.

Nevertheless, several of the ideas behind the bootstrap, for example, of simulating the variability of sampling by drawing resamples from the sample, had been discussed well before Efron developed them so successfully. Pre-eminent among early proponents...
was the remarkable Indian statistical scientist, P C Mahalanobis. More than half a century ago [2], he used bootstrap methods to estimate the variances of estimates associated with field trials. McCarthy [3] and Hartigan [4] [5] employed similar techniques; Barnard [6] suggested a related Monte Carlo approach to hypothesis testing; and Simon [7] proposed using resampling-based techniques as ‘basic research methods’ for the social sciences. Jackknife methods, introduced by Quenouille [8], should also be mentioned at this juncture. They played a major role in development of bootstrap methods; see particularly Efron [9, 10].

2. Statistics in the Bootstrap World

In their most rudimentary form, bootstrap methods use the relationship between the sample, and resamples drawn from the sample, to approximate the relationship between the population and samples drawn from it. To introduce bootstrap methods, suppose \( \hat{\theta} \) is an estimator of a parameter \( \theta \) computed from a random sample \( \mathcal{X} = \{X_1, \ldots, X_n\} \). A resample \( \mathcal{X}^* = \{X_1^*, \ldots, X_n^*\} \) is generally of the same size as \( \mathcal{X} \), and drawn by resampling as follows:

\[
P(X_i^* = X_j|\mathcal{X}) = \frac{1}{n}, \quad \text{for all} \quad 1 \leq i, j \leq n.
\]

The notation ‘\( P(\cdot|\mathcal{X}) \)’ denotes probability conditional on the dataset \( \mathcal{X} \). Likewise, ‘\( \text{var}(\cdot|\mathcal{X}) \)’ will represent the variance of a quantity, indicated by the argument, conditional on the data. We take the resample values \( X_i^* \) to be independent, given the data.

*Figure 1*, adapted from [11], shows the relationship between the ‘real world’ and the ‘bootstrap world’. A simplistic, though useful, interpretation of the bootstrap is
suggested by the figure. Nature has given us only one sample. From this sample, an estimator is computed. We are attempting to find out the properties of the estimator that would be discernible if nature were to provide similar samples under identical conditions. The problem is that we do not have other samples. So we generate them from the one that we do have. We then pretend that the random fluctuations in the bootstrap values of the estimator, demonstrated by its resampled versions, are good substitutes for the random fluctuations over the many samples Nature would have provided. A great deal of theoretical effort, surveyed for example, by Hall [12], has been expended to show that this intuition can be made mathematically tenable under suitable conditions.

The estimate of the unknown probability model that is used in the resampling scheme (1) is said to be the empirical estimate of the probability model. Other estimates can be used, which lead to other versions of the bootstrap. In particular, one does not need to resample in a uniform way, as suggested in (1). Weights $p_i$ can be used instead:

$$P(X_i^* = X_j | \mathcal{X}) = p_j, \quad \text{for all } 1 \leq i, j \leq n \quad (2).$$

The $p_i$'s define a multinomial distribution on the integers $1, \ldots, n$; that is, each $p_i \geq 0$ and $\sum_i p_i = 1$. By choosing this distribution appropriately, the bias of bootstrap methods can often be reduced, or the resulting estimator can be constrained in almost any desired manner (for example, by producing a bootstrap distribution estimator with its variance constrained to equal a predetermined quantity). See [13] for further discussion.

Estimating the variance of an estimator, by using the variance of the same quantity computed for a resample, is arguably the simplest example of bootstrap methods. We would like to know the value of $\sigma^2 = \text{var}(\hat{\theta})$, i.e. the variance of $\hat{\theta}$, but we cannot work it out because we don’t know the distribution of the population. So, we look instead for an estimate of the variance – that is, for a function of the data which represents an approximation to $\sigma^2$. Let $\hat{\theta}^*$ denote the version of $\hat{\theta}$ computed from the resample instead of the sample. Then $\hat{\sigma}^2 = \text{var}(\hat{\theta}^* | \mathcal{X})$ is our bootstrap estimator of $\sigma^2$. Sometimes we can work out the value of $\hat{\sigma}^2$ exactly, but in other instances it is prohibitively difficult to compute.

For an example of the case where $\hat{\sigma}^2$ is easily computed, note that if $\theta = \mu$ is the population mean, and $\hat{\theta} = \bar{X} = n^{-1} \sum_i X_i$ is the sample mean of $\mathcal{X}$, then a direct calculation shows that $\hat{\sigma}^2 = n^{-1} s^2$, where $s^2 = n^{-1} \sum_i (X_i - \bar{X})^2$ is the usual estimator of the population variance (but with denominator $n$ rather than $n - 1$).

For more complex statistics $\hat{\theta}$, $\text{var}(\hat{\theta}^* | \mathcal{X})$ is too complicated to be computed by hand or even on a personal computer. For example, suppose $\theta = |\mu|^{1/2}$, the square root of the absolute value of the population mean. We would generally take our estimator of $\theta$ to be $\hat{\theta} = |\bar{X}|^{1/2}$, but the variance of the latter could not be expressed simply in terms of the population distribution even if the latter were known. Likewise, a simple, exact formula for $\hat{\sigma}^2 = \text{var}(|\bar{X}^*|^{1/2} | \mathcal{X})$ does not exist, except for very small samples;
it requires us to work out all possible values of $X_1^* + \ldots + X_n^*$, for all resamples $X^*$, and weight them by their respective probabilities of arising in the resampling algorithm (1). The presence of the absolute value and square root adds further complications.

However, the Monte Carlo approach to computation is almost always tractable. It involves drawing independent realisations $X_1^*, \ldots, X_B^*$ of $X^*$, and calculating the respective values $\bar{X}_b^*$ of $\bar{X}^*$, and thence $\hat{\theta}_b^* = |\bar{X}_b^*|^{1/2}$, for $1 \leq b \leq B$. Then we take as our estimator of $\sigma^2$ the value of

$$\hat{\sigma}_B^2 = \frac{1}{\hat{\theta}_B^\text{mean}} \sum_{b=1}^{B} (\hat{\theta}_b^* - \hat{\theta}_B^\text{mean})^2$$

where $\hat{\theta}_B^\text{mean} = B^{-1} \sum_{b=1}^{B} \hat{\theta}_b^*$. This is the ‘sample variance’ of the sequence of values $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$. For any given sample $X$ it converges to $\hat{\sigma}^2$ as $B \to \infty$. Of course, this result is true quite generally; it does not require the particular example of $\hat{\theta}$ introduced in the previous paragraph. Formula (3), defining a Monte Carlo approximation $\hat{\sigma}_B^2$ to $\hat{\sigma}^2$, is used very widely in statistics today.

3. Extending the Bootstrap

The sampling regime given by (1) is usually referred to as ‘with replacement’ sampling, since it can be thought of as the result of drawing a sample value $X_i^*$ by sampling at random from $X$, then placing the observed datum back into the sample and mixing up the values again before drawing the next one, $X_{i+1}^*$, independently of all the preceding draws. Additionally, our definition of bootstrap methods also involves drawing a resample of the same size, $n$, as the original sample. Both these stipulations can be generalised. Indeed, the early techniques suggested by Mahalanobis [2], McCarthy [3] and Hartigan [4, 5] involved using resamples of smaller size than the original sample – typically half the size – obtained by sampling ‘without replacement’ from the sample. That sampling regime requires drawing $X_i^*$ randomly, and uniformly, from among the $n - i$ data that are left after the previous $i - 1$ sampling steps, which produced $X_1^*, \ldots, X_{i-1}^*$. Equivalently, a ‘random half’ of the data is the resample, if the resample is half the size of the original sample.

Clearly, using without-replacement sampling to obtain a resample $X^*$ of the same size as the original sample $X$ will always give $X^* = X$; it is degenerate. However, both with-replacement and without-replacement sampling, for obtaining samples of smaller size than the original one, are of theoretical and practical interest. For example, they are useful in one of the most recent applications of bootstrap methods, known as ‘bagging’ (short for ‘bootstrap aggregation’). Bagging also serves to illustrate the use of the bootstrap as an estimation procedure in itself, rather than as a device to judge the efficacy of conventional estimation procedures.

Bagging methods were introduced by Breiman [14] as a means of improving the accuracy of a variety of statistical procedures, particularly complex ones based on
multivariate data. In our context, they involve replacing an estimator by the average
or mode of the values it takes when computed from $B$ resamples. In the case of a
linear estimator (i.e. one which is a linear function of the data), there is virtually no
change when the estimator is ‘bagged’. For complex, highly nonlinear estimators the
improvements offered by bagging can be substantial, however. For example, recent
evidence suggests that so-called neural network algorithms can be bagged to give
improved performance. Neural networks are essentially prediction rules that generate
predictions given inputs. The rules are ‘learned’ from data previously collected.
Different resamples of these data yield different predictions. The bagged prediction
rule is the one obtained by averaging over the resampled prediction rules. See, for
example, [15] for details. Note that we have generalised the notion of an estimator to
that of a general function of the data – prediction rules are such functions.

Essentially, bagging works by smoothing out random fluctuations in conventional
functions of the data, by averaging over bootstrap resamples. Sometimes the manner of
averaging, and its effects, are rather subtle. For example, many statistical procedures
involve defining an estimator implicitly as a global extremum of a function of the
data, like the local minimum of an estimate of loss or risk, or a global maximum of a
likelihood. In complex problems there can be many local extrema, and distinguishing
the global one can be a particularly difficult task. The ‘averaging’ that is effected by
bagging smooths out the surface whose extremum is sought, eliminating many of the
relatively small bumps and making it easier to find the one that is associated with the
global extremum.

Even a superficial account of the many applications of the bootstrap, and its variants,
is beyond the scope of this short article. Fortunately, a range of monograph-length
treatments is available. It includes general introductions by Efron and Tibshirani[11]
and Davison and Hinkley [16], a theoretical account by Hall [12], and discussion of
the weighted bootstrap by Barbe and Bertail [17].

Acknowledgements

The authors are grateful to Prakesh Patil for helpful comments.

Suggested Reading

1946.
[3] P J McCarthy, Replication (an approach to the analysis of data from complex surveys), National Center


Opinion Polls and Statistical Surveys: What They Really Tell Us *

Rajeeva L Karandikar and Ayanendranath Basu

In recent times we seem to be having frequent parliamentary elections in addition to lots of assembly elections. Even as this article is being written, we are heading for another parliamentary election. At the same time, exit polls and opinion polls are gaining popularity. The media often reports the results of opinion polls predicting the seat shares of major parties before the actual elections; even exit polls are now commonplace. All major national and state elections in the recent times have been covered by exit polls. The agencies conducting these polls now recruit the services of highly qualified experts for this purpose.

At the same time, however, the general public of our country are largely unaware of the scientific issues involved in a statistically planned opinion survey (or sometimes even what it means). This leads to fantastic claims and counter-claims from politicians (depending on whether their parties are projected to win or lose); some statements go as far as asserting that opinion polls mean nothing and there is no science behind them. While the analysis of exit polls at the Doordarshan Channel was going on after the 1998 parliamentary elections, an otherwise respected political leader claimed that any opinion based on a sample of size 26000 in a country of 26 crore voters cannot give any meaningful information. It is true that no survey, not even the most optimally designed one, can predict the true state of nature with complete confidence, and there is always a degree of uncertainty involved. However, when the confusion reaches such a stage where a layman freely questions the findings of a scientific study, it becomes the duty of the scientific community to justify the merits and scope of such techniques. Can a sample of 26000 really lead to any meaningful conclusions in a country so large and diverse as India?

The first step in an exit/opinion poll (or for that matter any other statistical survey) is the estimation of some unknown parameter. (See [1] for some simple examples on parameter estimation.) While in the context of elections one can consider the unknown parameter to be the proportion of voters who intend to vote for a particular party (or alliance), indeed this is also the case for other social or economic parameters such as literacy ratios, unemployment rates, proportion of people below the poverty line, proportion of vehicle, house, TV and telephone connection owners and many such other things. If we were to simply estimate a population proportion (the proportion of votes for a given party in the election context), it can be done fairly accurately based on the information of a subset of the population – called the sample, provided the sample was large enough and ‘properly’ chosen (discussed in detail later). In an

When confusion reaches such a stage where a layman freely questions the findings of a scientific study, it becomes the duty of the scientific community to justify the merits and scope of such techniques.

election, however, the main interest of the populace (and hence the media) is not in the percentage of votes for a given party (or alliance), but rather in its seat share. This makes the problem complicated, and we need further analysis than the simple statistical estimation of a population proportion.

If the entire population size is small, we can probably enumerate all the individuals and thereby determine the exact proportion of individuals of each different type making up the population. Often, however, the population of interest will be so big that complete enumeration will be totally impractical – perhaps even impossible – due to constraints on essential resources such as money, time, manpower, etc. It is in situations such as this that the idea of a statistical survey becomes relevant. In these cases, statisticians must base their estimate of the unknown proportion on a smaller fraction of the actual population – the so called sample.

In the context of the parliamentary election, if one wants to predict the winner in each constituency of Lok Sabha with a high degree of confidence, one has to sample a reasonably large number, say 500, of individuals from each of our 544 constituencies. This will make the overall sample size prohibitively large, and the entire process will become an unmanageable and time-consuming exercise even if one had adequate financial resources. It becomes necessary therefore to restrict attention to a set of selected constituencies in some optimum way. The data obtained by sampling from these constituencies is then combined with some other information such as past voting trends and records, and the seat shares are then predicted based on a suitable mathematical model.

Since we get the actual information only from a subset of the total number of constituencies, the selection of the constituencies to be sampled is of great importance. In the United Kingdom, where the procedures for general elections and installation of new governments are very similar to our system, the idea of ‘safe seats’ is often used in the determination of the constituencies to be sampled. Both the Conservative and the Labour parties have large committed vote banks; voting patterns in UK change slowly, so that the constituencies with overwhelming support for one of the two parties can often be considered safe for the given party at elections in the not too distant future. For prediction purposes, one can therefore concentrate on the remaining marginal seats.

In India, the situation is quite different. Our election process involves a very large number of national and regional parties. Politics in India is largely personality-based rather than issue-based. Voter moods appear to be more easily swayed here. The political parties keep splitting and regrouping in different formations; old alliances are broken up and new alliances are created. All this causes voter loyalties to shift often and by wide margins, and there are very few safe seats. In addition, major events which have large impacts on voter's perception often have a very regional nature in a diverse country like India. All these make the use of the safe seat idea a rather shaky one in the Indian context.
We now discuss one by one the two stages of estimating the seat share of a particular party; these are (i) estimating the proportion of votes for a given party in a given region and (ii) estimating the corresponding number of seats based on a suitable model.

This problem may also be described in terms of urns and balls – as there being an urn with a very large number of balls of \( k \) different colours, and we have to draw a sample of appropriate size to estimate the actual proportion of balls of each colour in the urn. As the statistician will not have complete information about the population, it is improbable that the estimate offered on the basis of the sample observations (whatever it may be, usually it is the observed proportion based on the sample) will actually be exactly equal to the desired target – unless that happens to be a lucky coincidence. Allowing for a margin of error, one must therefore try to minimize the error in some average sense. One of the keys to this generally lies in the technique that is employed in choosing the sample on which the statistician’s inference is to be based – it must be ‘representative’ of the pattern of the true population. In the simplest case this is done through ‘simple random sampling’ (equal probability sampling) where one chooses a sample of a given size from the population in such a way that each individual in the population has the same chance of being included in the sample. This is necessary to diminish the possibility of the sample drawn being biased in the sense of being concentrated on a particular segment of the entire population. This randomness is a property of the sample collection method and not of the actual sample. It is not possible to say just by looking at the sample whether the particular sample has been chosen randomly or through some systematic procedure. In fact even the random sampling procedure may occasionally result in a biased sample, the conclusions based on which may be far off from the truth; however the randomness of the procedure ensures that this is unlikely.

We can illustrate this with the following. Consider a population of \( N \) individuals. Let \( B \) be the set

\[
B = \{1, 2, \ldots, N\}
\]

and suppose for each \( i \in B \), we have \( a_i \in \{0, 1\} \). Let \( M = \sum_{i=1}^{N} a_i \) and \( p = M/N \). Then \( p \) is the proportion of \( i \) for which \( a_i = 1, i = 1, 2, \ldots, N \). We want to estimate this unknown proportion \( p \) based on a sample of size \( n \) from \( B \).

Let \( S = \{(i_1, i_2, \ldots, i_n) : i_1, i_2, \ldots, i_n \in B\} \). For each \((i_1, i_2, \ldots, i_n) \in S\), let

\[
X((i_1, i_2, \ldots, i_n)) = \sum_{j=1}^{n} a_{i_j} \quad \text{and} \quad \hat{p}((i_1, i_2, \ldots, i_n)) = \frac{X((i_1, i_2, \ldots, i_n))}{n}.
\]

Each element in \( S \) is a sample and \( X \) is the number of \( i \) in the sample for which \( a_i = 1 \) and \( \hat{p} \) is the proportion of \( i \) in the sample for which \( a_i = 1, i = 1, 2, \ldots, n \). In sampling with replacement, where any element of the population can appear more than once in

As the statistician will not have complete information about the population, it is improbable that the estimate offered on the basis of the sample observations will actually be exactly equal to the desired target.

It is not possible to say just by looking at the sample whether the particular sample has been chosen randomly or through some systematic procedure.
the sample, the total number of samples (number of elements in $S$) is $N^n$ (since for each of the $n$ draws there is a choice of $N$ elements to choose from).

Here we present the following identities with brief explanations:

$$\sum_S [X((i_1, i_2, \ldots, i_n))] = nMN^{n-1}. \quad (1)$$

This follows from the fact that when we consider all possible samples, each of the $N$ members of the population will appear the same number of times among them. Since there are $N^n$ samples in all, and there are $n$ individuals in each sample, this number will be $nN^{n-1}$. Since there are $M$ elements in $B$ for which $a_i = 1$, the above identity follows.

$$\sum_S [X((i_1, i_2, \ldots, i_n))^2] = nMN^{n-1} + n(n-1)M^2N^{n-2}. \quad (2)$$

For this identity, notice that $X((i_1, i_2, \ldots, i_n))^2 = \sum_{j=1}^n a_{i_j} + \sum_{j=1}^n \sum_{k=1}^n a_{i_j}a_{i_k}$, where $j \neq k$ in the second (double) summation. Since $a_{i_j} \in \{0, 1\}$,

$$\sum_S [X((i_1, i_2, \ldots, i_n))^2] = nMN^{n-1} + n(n-1) \sum_S a_{i_1}a_{i_2}.$$

Since there are $N^n$ samples, and $M/N$ of them have 1 in the first position, and $M/N$ of them have 1 in the second position, identity (2) follows.

Hence

$$\frac{1}{N^n} \sum_S [\hat{\rho}((i_1, i_2, \ldots, i_n))] = p$$

$$\frac{1}{N^n} \sum_S [\hat{\rho}((i_1, i_2, \ldots, i_n))^2] = \frac{p}{n} + \frac{n-1}{n}p^2. \quad (3)$$

All these lead to the result

$$\frac{1}{N^n} \sum_S [(\hat{\rho}((i_1, i_2, \ldots, i_n)) - p)^2]$$

$$= \frac{1}{N^n} \sum_S [\hat{\rho}((i_1, i_2, \ldots, i_n))^2] - p^2 = \frac{p(1-p)}{n}. \quad (3)$$

As a result of (3), we have (by applying the well-known Chebyshev’s inequality)

$$\frac{1}{N^n} \{((i_1, i_2, \ldots, i_n) \in S : |\hat{\rho}((i_1, i_2, \ldots, i_n)) - p| > \epsilon\}$$

$$\leq \frac{1}{\epsilon^2} \frac{p(1-p)}{n}. \quad (4)$$

By choosing $n$ large enough, we can make the right hand side in (4) small, say less than 0.05 and thus ensure that in more than 95% of the samples the difference between observed proportion $\hat{\rho}$ and true proportion $p$ is less than (a pre-chosen) $\epsilon > 0$. 

RESONANCE-75 | Promoting Science Education
So, if we can find a mechanism to choose one of the $N^n$ possible samples in such a manner that each sample has equal probability of being chosen, we will get a representative sample with more than 95% probability. This is random sampling with replacement.

Using central limit theorem it can be shown that for large $n$ (greater than 100 say)

$$P\{|\hat{p}((i_1, i_2, \ldots, i_n)) - p| > \frac{1}{\sqrt{n}}\} \leq 0.05,$$

where the probability refers to the probability allocation corresponding to random sampling with replacement. (See [2,3,4] for some discussion on basics of probability theory and limit theorems.)

In sampling without replacement a particular object is allowed to appear at most once in the sample. However, when $n, N$ are such that $n/N$ is small (say less than 0.01), then the computations given above for the sampling with replacement case continue to be valid for practical purposes if we consider sampling without replacement. Here the set of all samples of size $n$ is the set of all subsets of $B$ with exactly $n$ elements and we choose one of these as our sample with equal probability. This is random sampling without replacement. We will refer to the ratio $n/N$ as the sample fraction.

Normally in problems such as estimating the percentage of voters favouring a particular party, the sample size is generally a very small (practically negligible) fraction of the true population size so that the results of sampling with replacement are applicable even when the actual sampling is without replacement. It is commonly believed that the sample size must reach a certain proportion of the true population size for a given degree of accuracy in estimation to be attained. This is incorrect. While it is true that the precision of the estimate increases with the size of the sample (see equation 4), given that the sample fraction is small, the accuracy of estimation will not depend on the population size or the value of the sample fraction. Besides the sample size $n$, the accuracy of the estimation process depends on the actual value of the true proportion, with true proportions closer to 1/2 being more difficult to estimate accurately than more extreme values closer to 0 or 1. In the election context, for two different constituencies where the proportions of voters favouring the candidate of a particular party are roughly the same, one will require the same sample size to estimate these proportions with a given degree of accuracy – even if the total population sizes are quite different (one may be several times the size of the other).

Thus estimation of vote proportions can be done more or less equally efficiently with the same sample size, be it in a small city or in a large state like Bihar or for the whole of India! And so constituency-wise prediction and estimation of number of seats in parliament are mammoth tasks. The implementation procedure of the data collection for the purpose of estimating the proportion of voters favouring a particular party can then be described in the following steps: first, depending upon the resources and desired accuracy, one determines the nationwide target sample. Next, the proportion of constituencies that should be sampled for the opinion poll is determined.
cided upon (say 15–20%). That many constituencies are then selected from all the constituencies via random sampling. A selected number (say 8–12) of polling booths are then randomly selected from each constituency. The nationwide sample size is divided into target sample size for each constituency. Then, from the voters’ list of these polling booths, the required number of individuals are chosen at random. This then forms the random sample on which our proportion estimation methods are to be based.

A somewhat controversial issue often associated with the question of sampling is whether or not to use a quota system for different subgroups of the population. The idea here is to get fixed representations from genders, age categories, socio-economic classes, castes, etc. In particular such quota surveys are popular in market researches and as a result have some application in opinion polls as well. In the election context, however, our intention is to determine the vote share based on the voting expectations of the entire population, and quota samples can possibly increase the bias of the method. We maintain that it is more appropriate to do overall random sampling than quota sampling in this context. A properly conducted opinion poll based on random selection of constituencies in each region followed by random selection of respondents should give appropriate representation to each group – this was the case in the CSDS-India Today opinion poll published in February 1998 (one of the authors, R L Karandikar, was associated with it).

Next we have to deal with the second issue, that of estimating the number of seats for each party. To do this in a meaningful way it becomes necessary to use some appropriate mathematical model. One such model assumes that the change in voting intention from the previous election to the present for a particular party (or alliance) is uniform over small homogeneous regions – such as states, parts of states, or other suitably defined geographical regions. Then, by using random sampling we can estimate this change and thus estimate the votes for each party/alliance in each state/region by applying this uniform change over all the individual seats of this region.

Once we have estimated the proportion of votes for each party, we can assign a probability of winning a given seat for each party. If the estimated proportions of votes for the top two parties in a given constituency differ by a big margin, we can assign probability of 1 for the leading party. If the top two are equal or only marginally separated, we can assign a winning probability of 0.5 for each of the top two contenders. For the intermediate cases, we can assign probabilities in between the above two extremes keeping in mind the sample size on which the estimate of proportion of votes is based. Eventually, the sum of probabilities of winning the seats represents the expected number of seats the party is likely to win.

The statistical part of the problem is thus to estimate the percentage of votes for each party in each state. Clearly to achieve reliable figures for each state (or region) it will be necessary to have a reasonably sized sample from the state. We suggest that to
achieve this, a sample size of approximately 2000 be selected from each state so that a certain level of accuracy is assured for the state. Nationwide, this will translate to a sample size of about 50000.

It is also necessary to make a distinction between opinion polls and exit polls. While the opinion poll can give an estimate of the possible seat share based on the voter mood on that particular day, there is no way to predict, even among those who identify themselves as likely voters, as to who will go out and actually vote on the voting day. The random sampling technique provides a representative sample for the entire population, not necessarily for those who will actually go out to vote. In addition, people do change their minds between the poll date and the voting date. The opinion poll itself provides no objective way to assess the change in the voter’s mood between the day of the poll and the day of voting. This problem is avoided in the case of the exit poll. Here the problem of who will come out to vote is taken care of since we sample only from those who have actually voted. Besides the interviewer gets all the subjects at one place rather than having to do a door to door survey and hence the costs are expected to be lower. However since the interviewer has to go by some thumb rule in this case (such as every 20th voter or so) and cannot be done on the basis of a random sample drawn from the voters list, the sampling error may be higher.

Finally, we come back to the question proposed earlier. Can a sample of 26000 individuals generate any meaningful information for a country as large as India? On hindsight it may be noted that the Doordarshan exit poll for the parliamentary elections in 1998 predicted the nationwide seat share of the different parties quite accurately (as well as most of the state-wise figures), but the individual predictions for the states of Tamil Nadu, Maharashtra and Rajasthan were not quite consistent with the final tallies – although when added up the inconsistencies were evened out. This is because while a sample size of 26000 is quite reasonable if one is interested in only the nationwide prediction of vote share, it will not necessarily give good predictions for each state or subregion where the sample size is much smaller. In our opinion, a nationwide sample of 50000 or more, with sample sizes of 2000 or more for the individual states will result in a good nationwide prediction and reasonable predictions in each of the states.

This of course requires that the sampling is done in a manner consistent with statistical theory and a reasonable method is used to translate the estimated vote percentages to estimated seats. It is interesting to observe that in spite of the large number of parties and regional diversity, it is still possible to come up with a meaningful prediction at the national level with a sample of around 50,000.

Suggested Reading


Excuse me, Sir, did you by any chance leave around your home address when you explored the South Pole?
Physical Sciences
The upper region must be getting highly polluted! First it was the small birds, then the pigeons, then crows and now the poor Superman himself!
Origin (?) of the Universe *

Historical Background

Jayant V Narlikar

The first part of this series covers the historical background to the subject of cosmology – the study of the structure and evolution of the whole universe. Ancient ideas, such as those of the Greeks, already show the beginnings of attempts to account for observations by natural laws, and to prove or disprove these by other observations. It needed the invention of the telescope and studies by scientists like Herschel and Hubble to reach the current understanding of our place in our galaxy, and its place as only one member of a far larger collection of galaxies which fill the observable universe.

1. Primitive Notions of the Universe

An assessment of our present understanding of the cosmos is best carried out with a historical perspective. The written history available today covers a very tiny fraction of the time span of human existence on earth and an even smaller fraction of the age of our planet estimated at some 4.6 billion years. Based even on such limited documentation we find that our ancient forefathers were indeed as curious about nature and the cosmos as we are today.

It is against this background that we should view the attempts our ancient forefathers made to understand the universe around them. They added conjectures and speculations to what they could observe directly. They used fertile imagination to extrapolate from the known to the unknown. Naturally the differing cultural traditions led to different cosmic perspectives in different parts of the world.

I am always impressed by the depth of ideas in our Vedas and Upanishads. Those who wrote them had a questioning mind. They perceived the complexity of the cosmological problem. The following lines from the Nasadiya Sukta are quite eloquent:

“Our ancient forefathers were indeed as curious about nature and the cosmos as we are today. I am always impressed by the depth of ideas in our Vedas and Upanishads. Those who wrote them had a questioning mind.

“Then (in the beginning) there was neither existence nor non-existence. There was no space nor was there anything beyond. (In such a situation) what should encompass (what)? For whose benefit? Was there the dense and deep water?”

“Who will tell in detail how and from where came the expanse of the existing? Who knows for sure? Even Gods came after creation. So who would know wherefrom the creation came?”

*Reproduced with permission from Resonance, Vol.1, No.1, pp.7–13, 1996.
These are fundamental questions which are being asked even by present day cosmologists. Humans however are not satisfied by only asking questions. One must have answers too – and if one cannot get them one tries to concoct some. So out of questions like these arose answers that were believed by many to be right. There were no scientific proofs for them but nevertheless they became part of the mythology and gained intellectual acceptance.

It was during the Greek civilization a few centuries before Christ, that such speculations began to be viewed somewhat scientifically. The Pythagoreans – the followers of the Greek mathematician and philosopher Pythagoras – were worried about the sun-earth relationship. They refused to accept that the earth goes around the sun (or even vice versa!). Instead they believed that the earth goes around a central fire located elsewhere. The theory predictably ran into difficulty because of the obvious question: “Why don’t we see this fire?” To answer this question, the Pythagoreans invented a ‘counter-earth’ that went around the central fire but in a smaller orbit. This orbit, they said, synchronized with the earth’s orbit in such a way that it always managed to block the view of the central fire from anywhere on earth.

The symptom of a wrong scientific theory is that to keep its prediction intact additional assumptions have to be made. Subsequently even these assumptions become untenable. The Pythagorean theory was of this type. First there was the difficulty of the central fire not being seen. Next came the problem of why the counter-earth is not seen... and so on. However in spite of our criticism of the theory from hindsight it had the merit that it was a disprovable hypothesis.

Karl Popper, the philosopher of science, has laid down this criterion for a scientific theory: it should be testable and in principle disprovable. In other words we should be able to think of a test whose outcome could rule out the theory. If the outcome does not disprove it the theory survives – until somebody can think of another more stringent test. Popper’s criterion provides us with a way of distinguishing between philosophical speculation and a scientific theory.

2. Aristotle’s Universe

Aristotle, another Greek philosopher, provided a series of principles that in today’s parlance could be called a physical theory. He was a pupil of the famous philosopher Plato and the teacher of Alexander the Great. Today Aristotle’s ideas are known to be wrong. Yet we should look upon them as man’s first attempt at quantifying the laws that govern observed phenomena. The key to Aristotle’s ideas lies in his classification of different types of motion.

Aristotle distinguished two types of motion seen in the Universe: *natural motion* which he supposed always to be in circles and *violent motion* which was a departure from circular motion and implied the existence of a disturbing agency. Why circles? Because Aristotle was fascinated by a beautiful property of circles which no other

---

**Figure 1.** A hierarchical cosmos: One of our many ancient speculations in India described the earth as resting on elephants, standing on a giant tortoise that was carried by a snake eating its own tail.
curve seemed to possess. Take any portion of a circle (what we usually call a ‘circular arc’) and move it anywhere along the circumference: that portion will coincide exactly with the part of the circle underneath it (The straight line also has this property but it can be considered a circle of infinite radius).

In the jargon of modern theoretical physics the above property is one of rotational symmetry. A one-dimensional creature moving along the circumference of a circle will find all locations on it exactly similar, there being no privileged position. As we shall find in the second part of this series, present-day cosmologists employ similar symmetry arguments about the large-scale structure of the universe.

Although the heavenly bodies, especially planets, did not appear to move (naturally) in circles the Aristotelians brought in more complicated geometrical constructions involving a series of circles called epicycles. Thus a planet may move on one epicycle whose centre moves on another epicycle whose centre moves on a third epicycle and so on leading ultimately to a fixed earth in the midst of all these moving real and imaginary points in space.

The epicycle theory was thus no different from the kind of parameter-fitting exercise that goes on in modern times when resolution of apparent conflicts between observations and a favoured theory is sought by introducing adjustable parameters into the theoretical framework. Such an exercise tells us more about the freshly introduced parameters than it does about the basic hypothesis of the original theory. In fact, as with the Greek epicyclic theory a theory, requiring too much patchwork of this sort eventually has to be abandoned.

While it is easy to deride Aristotle and welcome Copernicus, Kepler, Galileo and Newton we must acknowledge that the Greek philosopher originated the notion that natural phenomena follow certain basic rules. Aristotle’s perception of such rules turned out to be incorrect but the idea that they exist was carried over and has been the guiding light of theoretical physicists to this day.

3. The Advent of Telescopes

The major experimental input to astronomy as a science came in the seventeenth century with the discovery of the telescope. It was Galileo who first used the telescope for astronomical purposes and who first appreciated its value in observing remote heavenly bodies. Today we would not be discussing the subject of cosmology had there been no telescopes to give us a view of the universe.

No one appreciated the usefulness of the telescope more than William Herschel. A busy music master at Bath in England, Herschel was known for his organ recitals and his huge orchestras. At the age of thirty-five he decided to become an astronomer largely as a result of night-time reading of books on mathematics and astronomy. Herschel’s interest was in observational astronomy and starting with a small telescope...
The telescopic investigations of William Herschel and his son John led them to the first crude picture of our galaxy as a disc-like system of stars encompassed by the white band known as the Milky Way. By examining the distribution of stars away from the Sun in all directions the Herschels concluded that the Sun was at the centre of the galaxy. Thus although it was known in the nineteenth century that the Sun is just a common star which appears to be the brightest object in the sky only because it is the nearest, it still retained the special status of being at the centre of the galaxy.

4. Our Galaxy

This picture of the galaxy so methodically built up by the Herschels still had two defects which were not corrected until much later at the beginning of the present century. But even in the eighteenth and nineteenth centuries there were those who suspected that something was wrong and whose perceptions came remarkably close to the truth as we now know it. The mathematician J M Lambert suggested for example that the stars in the Milky Way are in motion around a common centre and that the Sun along with the planets also moves around this galactic centre.

Lambert also suggested that not all visible objects are confined to our galaxy. In addition to stars and planets astronomers had also found diffuse nebulae whose nature was not clear. Were they faraway clusters of stars or were they nearby clouds of luminous gas? Lambert argued that the nebulae were indeed very distant objects far beyond the galaxy.

Even as late as 1910-20 astronomers clung to the picture of our galaxy as developed by Herschel. For instance J C Kapteyn used the new technique of photography which proved to be a boon to astronomy and arrived at a model of our galaxy as a flattened spheroidal system about five times larger along the galactic plane than in the direction perpendicular to it. In this model commonly known as the Kapteyn Universe the Sun was located slightly out of the galactic plane at a distance of some 2000 light-years from the centre (one light-year is the distance travelled by light in one year and this is approximately $10^{13}$ kilometres). The Sun was thus not too far from the galactic centre just as Herschel had proposed.

Figure 4. This map of our galaxy as proposed by William Herschel had the Sun ‘S’ at the centre.
When Kapteyn’s work was published in 1920-22 it was already being challenged by Harlow Shapley. In a series of papers published during 1915–19, Shapley studied the distribution of large dense collections of stars called globular clusters. A globular cluster may contain up to a million stars and can be identified from a distance because of its brightness and distinctive appearance. Shapley found that the number of globular clusters falls off as one moves perpendicularly away from the galactic plane. Along the plane they seemed concentrated in the direction of the constellation of Sagittarius. Shapley therefore assumed that the galactic centre lay in that direction well away from the sun and estimated that the sun’s distance from the centre was 50000 light years. The modern estimate of this distance is only about 60 percent of this value but the sun does go around the galactic centre as guessed correctly by Lambert. The total diameter of the galaxy is about 100000 light years and it contains some 100-200 billion stars.

While Shapley was right in dethroning the sun from its presumed privileged position at the centre of the galaxy his distance estimates were too large because he ignored the effects of interstellar absorption. Nor did Shapley agree with Lambert’s view that most of the diffuse nebulae lay outside the galaxy. But by the 1920s the obscuring role of the dust began to be understood and the picture of our galaxy underwent a drastic change. Many stars which were earlier believed to be far away because they looked faint were discovered to be much nearer, their faintness being due to absorption by the interstellar dust. Even more important was the conclusion that many of the diffuse nebulae lay far away, well outside the galaxy. Indeed it soon became apparent, thanks
largely to the work of Edwin Hubble, that these nebulae were galaxies in their own right as large as our own which are moving away from our galaxy at very large speeds. It was Hubble who found an empirical law governing their motion that was to become the foundation for modern cosmology.

Suggested Reading

Gravitational Collapse and Structure Formation in an Expanding Universe *

J S Bagla and Pritpal Kaur Sandhu

We use Newtonian formalism to motivate the form of Friedmann equations that describe the expansion of the universe in the standard cosmological model. We use the same formalism to study the evolution of density perturbations in the universe. We show that a simple model like spherical collapse can be used to estimate the characteristics of halos of galaxies and clusters of galaxies.

1. The Friedmann Equations

Observations show that, at very large scales, there is no special direction in the sky [1], nor is there any special region or location within the universe [2]. At an early stage in the development of cosmology, the notion of an isotropic and homogeneous universe was raised to the level of a guiding principle known popularly as the cosmological principle [3]. Most models of the universe take this into account and assume that the distribution of mass is homogeneous and isotropic, i.e., the density of matter is constant and not a function of position though it may be a function of time if the universe undergoes expansion or contraction. It is also clear that in such a model, there can be no bulk motions that may lead to departures from homogeneity and isotropy at a later stage. An important implication of the cosmological principle is that the choice of origin and the orientation of axes of the coordinate system used to describe the universe are completely arbitrary.

If we consider the universe to consist mainly of slowly moving particles (particles whose motions are non-relativistic), then the total energy density in the universe is dominated by the rest mass energy of these particles and the kinetic energy is a small and sub-dominant contribution. In such a case, the energy density is proportional to the mass density $\rho$. It is possible to construct a Newtonian model to describe the evolution of such a universe.

Let us consider an arbitrary origin in the smooth matter distribution, and then study the motion of a thin spherical shell at a distance $r$ from the centre. The equation of motion for the shell is:

$$\ddot{r} = -\frac{GM}{r^2}. \tag{1}$$

The shell may expand or contract, but other types of motion such as rotation or shear are not allowed as these lead to the violation of homogeneity and isotropy. The motion of the shell is then conveniently described in co-moving coordinates that expand or


Keywords
Cosmology, gravitational collapse, galaxy formation.
contract with the universe as a whole. The physical coordinates \( r \) are related to the co-moving coordinates as \( r(t) = a(t)x \), where the co-moving coordinates \( x \) are fixed for a given particle even as the universe expands or contracts. Using this, we can recast (1) as

\[
\frac{\ddot{a}}{a} = -\frac{4}{3}\pi G \rho, \tag{2}
\]

where \( \rho \) is the matter density and is related to the mass enclosed inside the shell at radius \( r \) as \( M = 4\pi \rho r^3/3 \). The expansion or contraction of the universe is described by the scale factor \( a(t) \) that increases during expansion and decreases during contraction.

The mass enclosed within the shell remains constant as the matter within the shell remains inside even as the universe expands or contracts. This implies that \( \rho \propto a^{-3} \) and we may write:

\[
\rho(t) = \rho(t_0) \frac{a^3(t_0)}{a^3(t)} = \rho_0 \frac{a^3_0}{a^3}.
\]

We can also use the constancy of mass within the shell to find the first integral of motion from (1):

\[
\left(\frac{\dot{a}}{a}\right)^2 + k \frac{a^2}{a^2} = \frac{8\pi G}{3} \rho.	ag{3}
\]

Here, \( k \) is a constant and may be considered equivalent to energy. It is clear that this equation admits three classes of solutions for non-relativistic matter:

- **Open universe**: \( k < 0 \). The sign of \( \dot{a} \) does not change through the evolution of the universe. An expanding universe continues to expand forever and a contracting universe contracts for ever. In the case of an expanding universe, the scale factor \( a(t) \) increases forever with a rate \( \dot{a} \) that eventually approaches a constant value.

- **Flat universe**: \( k = 0 \). The comments relating to the unchanging sign of \( \dot{a} \) for the open universe apply here as well. In the case of an expanding universe, the scale factor \( a(t) \) increases forever with a rate that keeps on decelerating.

- **Closed universe**: \( k > 0 \). In this case, it is possible for the sign of \( \dot{a} \) to change during the course of evolution. The expansion can slow down to a halt and start a collapsing phase. For the converse, we require \( \ddot{a} > 0 \) and that is not possible with normal matter (see (2)).

Even though we derived (3) and (2) using Newtonian arguments, it turns out that for non-relativistic matter these are also the equations we get in Einstein’s general theory of relativity. Equation (3) holds for arbitrary constituents of the universe, even for relativistic constituents, whereas (2) requires a correction in case the constituent has non-negligible pressure. The relativistic form of the equation is:

\[
\frac{\ddot{a}}{a} = -\frac{4}{3}\pi G (\rho + 3p), \tag{4}
\]

where \( p \) is pressure and we have taken \( c = 1 \). The set of Friedmann equations (3) and (4) contains three unknowns and, in general, we need to add another equation in order
to solve for all the quantities. The relevant equation here is one that describes the relation between density and pressure for the constituents of the universe. However, here we are concerned mainly with non-relativistic matter and hence, we choose to ignore the pressure term as well as its relation with density.

Observations show that the universe is spatially flat, i.e., \( k = 0 \) is favoured by observations. In this case, the scale factor \( a(t) \propto t^{2/3} \) if the universe contains only non-relativistic matter\(^1\). Such a universe is referred to as the Einstein–deSitter model. Observations made towards the end of the twentieth century show that the expansion of the universe is accelerating, which cannot happen in the framework we are using. This is attributed to a form of matter known as ‘dark energy’. Thus, the Einstein–deSitter model is not realistic in the light of present day knowledge. We will go ahead and analyse it anyway, since some of the aspects covered are unchanged. We will, indicate briefly the changes that dark energy make to the results, later on.

2. The Clumpy Universe

The universe is not homogeneous and isotropic at small scales. We observe large fluctuations in density – indeed the Earth has an average density of 5 g/cc, whereas, the average density of the universe is close to \( 10^{-30} \) g/cc (see Box 1, (ii)). Our solar system and the Galaxy also represent very large over-densities. In the following discussion we would like to understand how large over-densities can form in a universe that was very smooth at early times.

Our derivation of Friedmann equations using Newtonian mechanics can be generalised to the study of density perturbations if we assume that the perturbations are spherically symmetric [4]. The introduction of such a perturbation is easier to analyse as it can be shown that matter exterior to the perturbation does not affect its evolution in any way. We note here that the introduction of a spherically symmetric perturbation retains isotropy but removes homogeneity.

2.1 Spherical Collapse Model

Consider a spherical region with a constant over-density that is expanding along with the rest of the universe at early times. We assume that the initial over-density is small in magnitude and that the size of the over-dense region is small. Now consider a shell, initially at radius \( r_i \), which encloses a mass \( M \). It can be shown that mass will remain conserved as the system evolves. The equation of motion is:

\[
\ddot{r} = -\frac{GM}{r^2}.
\]

Using the constancy of mass enclosed we get the first integral of motion:

\[
\frac{r^2}{2} = \frac{GM}{r} + E.
\]
Box 1. Density of the Universe and Hubble’s Law

The relation between the co-moving and physical coordinates can be used to derive a relation between the expansion velocity and distance \[ \text{[A]}, \] i.e., \( V = H_0 r \), where \( H_0 \), the Hubble’s constant, is the present value of \( \dot{a}/a \).

Writing the relation (3) at the present time, we find:

\[
H_0^2 + \frac{k}{a_0^2} = \frac{8\pi G}{3} \rho_0. \tag{i}
\]

We see that there is a special value of the present day density \( \rho_0 \) for which \( k \) vanishes. This is known as the critical density: \( \rho_c = \frac{3H_0^2}{8\pi G} \). Observations favour values of \( H_0 \approx 70 \text{ km/s/Mpc} \) and we get the critical density:

\[
\rho_c = \frac{3H_0^2}{8\pi G} = 9.2 \times 10^{-25} \text{kg/m}^3 \left( \frac{H_0}{70 \text{km/s/Mpc}} \right)^2 \\
\sim 10^{-30} \text{g/cc}. \tag{ii}
\]

The energy \( E \) has to be negative for this to be a bound perturbation and we write it explicitly as \( E = -|E| \). Further, as mentioned above, we assume that the perturbation is taken to be situated in an Einstein–deSitter universe. The energy can be written in terms of initial conditions:

\[
|E| = -\frac{1}{2} \dot{r}_i^2 + \frac{GM}{r_i} \\
= -\frac{1}{2} H_i^2 r_i^2 + \frac{4\pi G}{3} \rho_i r_i^2 \\
= -\frac{1}{2} H_i^2 r_i^2 + \frac{4\pi G}{3} \bar{\rho}_i (1 + \delta_i) r_i^2 \\
= \frac{1}{2} H_i^2 r_i^2 \delta_i. \tag{6}
\]

Here, \( r_i \) and \( \dot{r}_i \) are the initial radius and velocity of the shell. We assume that these are related through the Hubble’s law \( \dot{r}_i = H_i r_i \) with \( H_i \) as the value of Hubble’s parameter at the initial time. \( \rho_i = \bar{\rho}_i (1 + \delta_i) \) is the density of the perturbation and is related to the average density \( \bar{\rho}_i \) and density contrast \( \delta_i \). Qualitatively, we can see that as compared to the Einstein–deSitter universe, the expansion rate of the shell within the

over-dense region slows down at a faster rate until it comes to a halt and then begins to recollapse. The slower expansion leads to a relative increase in the density within the shell as compared to the background universe. At the stage when its radius is maximum, we have $|E| = GM/r_{\text{max}}$ and hence, $r_{\text{max}} = r_i (1 + \delta_i) / \delta_i \approx r_i / \delta_i$.

The solution to (5) can be written in a parametric form:

$$
\begin{align*}
  r &= \frac{r_{\text{max}}}{2} (1 - \cos \theta), \\
  t &= \frac{1}{2H_i \delta_i^{3/2}} (\theta - \sin \theta),
\end{align*}
$$

where we have assumed that $\delta_i \ll 1$. This is obtained by integrating (5) and substituting for $r$. We note that the characteristic collapse time for the perturbation is $H_i / \delta_i^{3/2}$ and is smaller for perturbations with a larger initial density contrast. This implies that if we have a perturbation where density contrast decreases as we go to larger radii, the inner shells collapse first and outer regions collapse later.

It is interesting to note that the characteristic collapse time is fairly long – indeed it is of the same order as the age of the universe. In contrast, the collapse of a perturbation of similar density in a static universe is about $\delta_i^{3/2}$ times smaller, i.e., for a $\delta_i \approx 10^{-2}$, the difference is three orders of magnitude if we consider perturbations with the same density. Thus, the expansion of the universe slows down the collapse of perturbations by a significant amount and as a result, the growth of density perturbations in an expanding universe is a very slow process as compared to its counterpart in a static background.

The solution given above is cyclic in nature where shells collapse to the centre and then rebound to the same amplitude in the other direction. However, it is obvious that most perturbations do not have spherical symmetry. It can also be shown that any departures from spherical symmetry grow rapidly during collapse. Hence, our estimation is likely to deviate more and more from the collapse of real perturbations at late times. We can capture the physics of key stages during further collapse by realising that in the collapsing phase where deviations from a symmetric initial condition grow rapidly, violent relaxation [5] is likely to play an important role to bring the system close to an equilibrium state in a very short time.

In equilibrium state for self-gravitating systems, the virial theorem dictates that we must have $2K + U = 0$, where $K$ is the kinetic energy and $U$ is the potential energy of the system. This can be rephrased as $2K + U = 2E - U = 0$, and therefore $|E| = GM/2r_{\text{vir}}$, where $r_{\text{vir}}$ is the radius of the system after it has reached virial equilibrium. Comparing this with the value of energy at the time when the system is at its maximum radius $r_{\text{max}}$, we find that $r_{\text{vir}} = r_{\text{max}}/2$. Thus, the system relaxes to equilibrium at half the radius of maximum expansion and after this we do not expect its size to change, whereas, the universe will continue to expand.

The density contrast for the matter contained within the shell can be written in the
Crossing time is the time taken by a typical constituent to cross the system under consideration. In the present context, this is equal to the time taken for the shell to go from maximum displacement to the origin.

The perturbation expands from an initial value of \( r_i \) to \( r_{\text{max}} = r_i / \delta_i \). As \( \delta_i \ll 1 \), it is clear that \( r_{\text{max}} \gg r_i \). On the other hand, the shell collapses only by a factor of 2 from \( r_{\text{max}} \) to \( r_{\text{vir}} \). Overall, the effect during gravitational collapse of the system is an expansion. Therefore, we do not expect heating due to compression in gravitational collapse.

The perturbation has a density contrast of \( 9 \pi^2 / 16 - 1 = 4.55 \) at the time of maximum expansion.

As discussed above, the perturbation is expected to take one crossing time beyond the stage of maximum expansion to reach close to dynamical equilibrium. Thus, it requires time \( t(\theta = \pi) \) to reach virial equilibrium from the stage of maximum expansion. We have already found that the radius at virial equilibrium is half of the radius at maximum expansion. Using these in (8) and (9), we find that the density contrast at the time it reaches equilibrium is \( 168 \). Thus, the typical over-density of a collapsed virialised halo is 168, i.e., average density inside virialised halos is 168 times the average density of the universe at the time of collapse. Note that for this second calculation we do not use the last equality, as the parametric solution for \( r \) is invalid in this regime. The parametric solution and the expected behaviour after virialization is illustrated in Figure 1.

The density of the halo remains constant once it reaches virial equilibrium but the density contrast continues to increase as the average density of the universe continues to decrease.

The discussion so far has focused on the evolution of the perturbation due to gravitational interaction. It is interesting to ask as to what happens to gases during the evolution of the perturbation. As the perturbation expands with the universe, clearly gases must cool, and there should be some heating as the perturbation collapses after reaching maximum radius. An implicit assumption here is that the entire evolution is adiabatic and that the gases do not gain or lose energy to any external source or sink. As the expansion of the perturbation is the dominant scaling as compared to compression by a factor of two during collapse, it is clear that the temperature of gases is not very high.

If we assume that after collapse, gas inside the virialised halo is supported by thermal pressure, then we can estimate the temperature of gas using the expression for virial pressure:

\[
1 + \delta = \frac{\rho}{\bar{\rho}} = \frac{3M/4\pi r^3}{\bar{\rho}}.
\] (8)

For the Einstein–deSitter universe, we have:

\[
H^2 = \frac{8\pi G}{3} \bar{\rho},
\]

where, \( H = \dot{a}/a \) is the Hubble parameter. Using the fact that \( a \propto t^{2/3} \) in this case, we have \( H = 2/3t \). This gives:

\[
\bar{\rho} = \frac{1}{6\pi G t^2}.
\] (9)

Therefore,

\[
1 + \delta = \frac{9GMt^2}{2r^3} = \frac{9 (\theta - \sin \theta)^2}{2 (1 - \cos \theta)^3}
\]

The density contrast of \( 9\pi^2 / 16 - 1 = 4.55 \) at the time of maximum expansion.

\[\text{Crossing time is the time taken by a typical constituent to cross the system under consideration. In the present context, this is equal to the time taken for the shell to go from maximum displacement to the origin.}\]

\[\text{The perturbation expands from an initial value of } r_i \text{ to } r_{\text{max}} = r_i / \delta_i. \text{ As } \delta_i \ll 1, \text{ it is clear that } r_{\text{max}} \gg r_i. \text{ On the other hand, the shell collapses only by a factor of } 2 \text{ from } r_{\text{max}} \text{ to } r_{\text{vir}}. \text{ Overall, the effect during gravitational collapse of the system is an expansion. Therefore, we do not expect heating due to compression in gravitational collapse.}\]
equilibrium, and by recognising that the kinetic energy of gas is related to its temperature. If the gas temperature is lower than this virial temperature $T_{\text{vir}}$, then the gas must collapse towards the centre of the halo and heat up until the pressure is strong enough to provide the required support. We have:

$$K = \langle \frac{1}{2} m v^2 \rangle = \frac{3}{2} k_B T_{\text{vir}} = \frac{GMm}{2r_{\text{vir}}}$$

with $k_B$ the Boltzmann constant and $T_{\text{vir}}$ as the temperature in virial equilibrium for a halo of mass $M$ and radius $r_{\text{vir}}$. Thus,

$$T_{\text{vir}} = \frac{3GMm}{r_{\text{vir}} k_B} \approx 10^7 K \left( \frac{M}{10^{12} M_\odot} \right) \left( \frac{r_{\text{vir}}}{100 \text{kpc}} \right)^{-1},$$

where $M_\odot = 2 \times 10^{30}$ is the mass of the Sun and 1 kpc = $3.08 \times 10^{19}$ m. Thus, the gas in the halo of a collapsed object like our Galaxy is expected to be very hot.

The question that naturally arises is how does the gas heat up to such a high temperature when we expect the infalling gas to be cold? The answer to this puzzle lies in the formation of shocks near the virial radius. The speed of the infalling gas as it nears the virial radius is $\dot{r} = \sqrt{GM/r_{\text{vir}}}$ and for large masses this is significantly higher than the speed of sound. Thus, the infalling gas shocks as it collapses onto material that collapsed earlier and is already in equilibrium. It can be shown that shock heating takes the temperature of the gas fairly close to $T_{\text{vir}}$. Using the jump conditions\(^6\) for

---

\(^6\) Macroscopic quantities such as density, pressure and temperature are discontinuous across a shock. We can use conservation of mass, momentum and energy as matter crosses the shock to relate the values of density, pressure and temperature across a shock. Such relations are known as junction or jump conditions.
Box 2. Isothermal Sphere

An isothermal sphere is a halo with density profile \( \varrho(r) \propto r^{-2} \). Halos with this density profile are of interest as this leads to flat rotation curves. Isothermal halos are also an equilibrium configuration [A, B, C]. Lastly, the comparison with universal density profile derived from cosmological N-body simulations suggests that halos have this density profile in the vicinity of the virial radius [D], and therefore, we may use it as a simple and useful approximation for analytic estimates.

We may formally write the density of the (singular) isothermal sphere as:

\[
\varrho(r) = \varrho(r_{\text{vir}}) \left( \frac{r_{\text{vir}}}{r} \right)^2.
\]

Here, \( r_{\text{vir}} \) is the virial radius and \( \varrho(r_{\text{vir}}) \) is the density at this scale. The mass enclosed within the virial radius is:

\[
M_{\text{vir}} = \int_0^{r_{\text{vir}}} 4\pi r^2 \varrho(r) dr = 4\pi \varrho(r_{\text{vir}}) r_{\text{vir}}^3.
\]

The average density within the virial radius is then:

\[
\bar{\varrho} = \frac{3M_{\text{vir}}}{4\pi r_{\text{vir}}^3} = 3 \varrho(r_{\text{vir}}).
\]

We have seen in the discussion on virialised halos that the average over-density at the virial radius must be equal to \( \Delta = 18\pi^2 \) at the time of virialisation in an Einstein–deSitter universe. Thus,

\[
\bar{\varrho} = \frac{3M_{\text{vir}}}{4\pi r_{\text{vir}}^3} = 3 \varrho(r_{\text{vir}}) = \Delta \varrho_c(z_c) = \Delta \varrho_c(1 + z_c)^3.
\]

Here, \( \varrho_c(z_c) \) is the critical density at the redshift \( z_c \) when the halo collapsed and \( \varrho_c \) is the critical density at present time. Thus, \( M_{\text{vir}} = 4\pi \Delta \varrho_c(1 + z_c)^3 r_{\text{vir}}^3 / 3 \), and \( M(r) = 4\pi \Delta \varrho_c(1 + z_c)^3 r_{\text{vir}}^2 r / 3 \).

From here, we get the circular velocity as:

\[
\nu_c^2 = \frac{GM}{r} = 4G\pi \Delta \varrho_c(1 + z_c)^3 r_{\text{vir}}^2 / 3 = \frac{H_0^2}{2} \Delta (1 + z_c)^3 r_{\text{vir}}^2.
\]

Halos of a given mass that collapse at higher redshift are more compact and have a higher circular velocity.


shock we get [6]

\[ T_2 = \frac{3}{16} T_{\text{vir}} + O(T_1). \]

Here, \( T_1 \) is the temperature of infalling gas before it reaches the shock and \( T_2 \) is the post-shock temperature. If \( T_{\text{vir}} \gg T_1 \) then we get \( T_2 = \frac{3}{16} T_{\text{vir}} \) and hence the infalling gas is heated almost to the virial temperature of the halo by shock heating. Other processes may lead to a further increase in the temperature of gas in a halo.

### 3. Summary

In this article, we have reviewed the spherical collapse model and shown how we may model non-linear gravitational collapse in an expanding universe. We have highlighted the fact that the expansion of the universe slows down the collapse of perturbations by a large factor. Further, we have shown that at least in the Einstein–de Sitter universe, the over-density of the collapsing perturbation has characteristic values at the time when it reaches maximum radius and the time when it reaches virial equilibrium.

Using junction conditions for shock heating in the last stages of collapse, we were able to demonstrate that the gas is indeed heated to temperatures in the vicinity of the virial temperature of halos.

As an application of the spherical collapse model, we calculated quantities of interest for an isothermal halo and discussed its implications for the formation of galaxies (see Box 2).

Processes follow the same pattern in case of a universe with dark energy, with a well-defined epoch of maximum radius and virialisation. Shock heating is the main mechanism for heating the infalling gas to temperatures near the virial temperature. There are subtle differences however, and we will present a detailed discussion in a follow-up article.

### Suggested Reading


“We are not only going to be seeing the universe, we are going to be listening to it.”
– Gabriela Gonzalez (spokesperson for the LIGO team)
Many scientists from India have been working in the area of gravitational wave detection as part of the LIGO international collaboration. This article recounts how these smaller scale efforts grew into a proposal for locating an advanced detector in India, which brings very significant scientific advantages. There is now in-principle approval for funding the Indian contribution to building such a facility. The article brings out the opportunities as well as major challenges which now open up to a much wider community of young scientists and engineers, in building and running such a unique laboratory.

The Indian Union Cabinet’s in-principle approval on Feb 17, 2016, that brought the gestating LIGO-India mega-science proposal to life, has swiftly followed the historic announcement on Feb 11, 2016, of the first detection of a gravitational wave (GW) event by Laser Interferometer Gravitational-Wave Observatory (LIGO), USA. This formally embarked a set of determined scientists of the pan-Indian IndIGO consortium and the LIGO-India teams on a fantastic adventure in a blossoming big-science frontier. What lies ahead for LIGO-India is a tough and challenging path of multi-institutional scientific collaboration together with major industries in India to build an advanced Laser Interferometer Gravitational-Wave Observatory on Indian soil over the next 8 years and then operate it at its design sensitivity for the next decade, or two. Many of the young science and technology enthusiasts reading this article would have, before then, attained sufficient learning and training to allow them to join in this amazing adventure in India. This is, especially, a note of welcome for them.

The direct detection of GW has opened a new window into the dark, or shrouded, core of astrophysical phenomena in the universe, and also allows an experimental probe of gravitation theory in extreme situations such as two black holes merging. The weakness of interaction of GW with matter that has made them so difficult to detect till recently, allows them to emerge unscathed from the very core of violent astrophysical phenomena such as a supernova explosion, gamma-ray burst from coalescing neutron stars or an invisible merger of black holes. In contrast, in dense, hot environments, such as the interior of our Sun or at the core of energetic phenomena, electromagnetic radiation is strongly scattered within the dense material and it is only possible to map the exterior boundary, generally referred to as the photosphere. GW astronomy is, hence, privy to details not accessible to astronomy using the entire spectrum of electromagnetic waves, ranging from radio waves to X-rays and gamma rays, all of which can map only up to a dense photosphere enshrouding the central engine of

---

The weakness of interaction of GW with matter that has made them so difficult to detect till recently, allows them to emerge unscathed from the very core of violent astrophysical phenomena.

energetic phenomena. The central energy source is a deep gravitational potential well into which matter falls at relativistic speeds, and Newtonian gravity must be replaced by general relativity. This gives rise to new physical phenomena which gravitational waves will allow us to probe, no doubt bringing out new surprises.

Few years into the millennium, during the phase of operation with the Initial LIGO detectors in USA and its close cousin, VIRGO, in Italy, concrete designs of the advanced interferometer detectors achievable in a decade were being drawn up. The advanced generation of GW detectors indicated a ten-fold improvement in sensitivity. That meant a thousand-fold increase in the expected rate of cosmic events that would produce detectable GW signals on Earth, implying at least one event per year even in fairly pessimistic astrophysical scenarios. It was eminently clear that the detection of GW signal from astrophysical phenomena, in particular, merger of compact neutron stars and black-hole binaries, was guaranteed with Advanced LIGO.

The recent detection of GW presents a vast opportunity over the coming decades for furthering the frontiers of astronomy with a global network of observatories, followed by GW observatories in space. The larger the baseline, the more refined is the inferred location of any GW event; hence, there was a call for establishing additional GW observatories across the globe.

A few researchers in India recognised the great opportunity it provided to the Indian scientific community. They started to plan and prepare a decade in advance to welcome the emerging field as equal partners at the global frontier. In a radical departure from the norm, Indian scientists would then be well entrenched in an emergent frontier field well before it had blossomed and gone out of reach.

India responded fairly early with the first serious discussion taking place in late 2007 at the International Conference in Gravitation and Cosmology (ICGC) in IUCAA. In August 2009, at another meeting in IUCAA, a consortium of Indian researchers called IndIGO, that collectively had expertise in theoretical and experimental gravity, cosmology and optical metrology, was formed. This consortium sought to promote gravitational wave research in the country with a dream of realizing an advanced GW observatory in India. This multi-institutional, multi-disciplinary IndIGO consortium now consists of three lead institutions for LIGO-India and nine other nodal institutions. Starting from 14 members in 2010 it has grown to 120 members today. With 3 experimenters then, it has 56 experimenters today. The Indian membership in the LIGO Science Collaboration, from a single group at IUCAA during 2000–2010, has now grown to a pan-India group with 61 members, including 21 experimenters, from nine Indian institutions, namely, IPR, IUCAA, RRCAT, TIFR, ICTS-TIFR, CMI, IIT-Gandhinagar, IISER-Kolkata, and IISER-Thiruvananthapuram.

The Initial LIGO consisted of two observatories: one with two interferometer detectors at Hanford, Washington State and the other with one detector at Livingston in Louisiana. While all three detectors were upgraded in the Advanced LIGO project
funded by NSF, it was clear that only a geographical relocation of the second Hanford detector to a globally distant location would enable the LIGO laboratories to translate the detection of GW events into a discipline of GW astronomy. The initial destination for the third distant LIGO was Australia.

Invited to the membership of the Gravitational Wave International Committee (GWIC), IndiGO explored the Indian contribution to the International Advanced GW Network. It planned for the participation in LIGO-Australia with 20 per cent contribution. When it seemed that funding for LIGO-Australia may not be forthcoming from the Australian government, in mid 2011, preliminary discussions for locating to India (i.e., LIGO-India) were initiated. With strong encouragement from Indian science luminaries, the LIGO-India proposal was put forward by IndiGO to the mega-science committee of the Indian Planning Commission in November 2011.

The LIGO-India proposal is for the construction and operation (for 10 years) of an Advanced LIGO observatory in India in collaboration with the LIGO Laboratories, USA. With design displacement sensitivity of $4 \times 10^{-20}$ m the LIGO detectors are the most precise physics apparatus. Figure 1 shows the layout of this enormous scientific apparatus. The 4 km beam tube along each arm holding about 10 million litres of ultra-high vacuum (nanotorr) would make it an impressive tribute to technological prowess of Indian science and technology. LIGO-India is expected to improve greatly the angular resolution in the location of the gravitational-wave source by the LIGO global network. Indeed, a startling attestation to the promise of LIGO-India comes from the forecast that, for the specific event detected, in Sept 2015 by the two Advanced LIGO detectors, a hypothetical LIGO-India in operation would have made a hundred-fold improvement in the angular resolution as shown in Figure 2.

The construction, commissioning and the scientific operations of the planned LIGO-India observatory will involve three Indian lead institutions, the Institute for Plasma Research (IPR), Gandhinagar, the Inter-University Centre for Astronomy and Astrophysics (IUCAA), Pune, and the Raja Ramanna Centre for Advanced Technology (RRCAT), Indore, working in close collaboration with LIGO Laboratories, USA. Broader Indian participation in this mega-science initiative will be made possible

Figure 1. Engineering concept design of LIGO-India at one of the shortlisted sites in India. Terrain data obtained from CartoSat, SAC, ISRO. Courtesy: TCE, India.

For comparison, the sizes of atomic nuclei are measured in units of $10^{-15}$ metres. A nanotorr is roughly one millionth of one millionth of atmospheric pressure.
The immense promise of LIGO-India in enabling localisation of gravitational-wave events in the sky and launching gravitational wave astronomy. The grey banana-shaped patch spanning 2500 moons is the current uncertainty of the localisation of the first discovery event. The small dark ellipse that is 100 times smaller shows the forecast uncertainty for a similar gravitational wave signal once LIGO-India becomes operational and joins the global network.

LIGO-India teams at IUCAA, IPR, RRCAT have been at work even in the pre-approval phase. The site-selection committee at IUCAA has looked at over 22 sites and worked to shortlist a few of them based on site-selection criteria like low ‘seismicity’ (ground noise), low human generated noise, socio-environmental considerations of land acquisition, air connectivity, road connectivity and data connectivity. IUCAA has also started science team-building activity and has been steadily building up a state-of-the-art computing, data infrastructure and associated manpower in anticipation of LIGO-India. The team at IPR has prepared system requirement documents, concep-
tual drawings and engineering drawings for the sophisticated civil infrastructure, and ultra-high vacuum systems in consultation with LIGO Labs. The team at RRCAT has been finalizing plans for setting up an off-site laboratory to receive the laser systems, and optics for LIGO-India. Pre-approval preparations have included many schools and workshops at different levels and also selection of Indian UG students for summer internships at LIGO USA and other top GW research centres worldwide. With the approval from the Union Cabinet, the combined LIGO-India team is now in a high state of readiness to take the first steps which have already been planned.

The discovery of gravitational waves and the approval of LIGO-India paves the road to the possibility of observing our universe using gravitational waves as the messengers carrying information. The IndIGO consortium has also seen increasing participation from the Indian astronomy community in anticipation of this new emerging frontier of ‘electromagnetic follow-up’ of GW events. Indian facilities such as AstroSat launched by ISRO, and GMRT near Pune, and Himalayan Chandra Telescope in Hanle, Ladakh would engage in this effort.

LIGO-India has the potential to impact precision experiments and cutting-edge technology in the country. The project has interfaces with quantum metrology, laser physics and technology, vacuum technologies, optical engineering, sensor technologies, control systems, grid and cloud computing to list a few. *Figure 3* adapted from

**Figure 3.** The impact of LIGO-India is multifaceted and would push the frontiers along each direction. (LIGO-India proposal document Nov 2011.)
IndIGO looks forward to enthusiastic participation by a generation of young scientists and engineers in this adventure of exploring the sky using the latest probe – gravitational waves.

Postscript (added in December 2021)

Post Indian Union cabinet’s in-principle approval 2016, the LIGO-India project has taken significant strides. A final site, amongst the three shortlisted at the time of the original article, has been formally selected (Sept 2016), and now fully acquired in the Hingoli district of Maharashtra, together with campus land in a neighbouring town. Preliminary work of securing the site with fencing, site office, all weather station and critical site characterisation studies including extensive topographical, geotechnical and seismic survey have also been completed. The current plans indicate first operations in 2026, subject to a detailed post-pandemic reassessment.

Meanwhile, gravitational astronomy has become a booming field with 90 merger events being detected in the LIGO US and Virgo (Europe) detectors to which many Indian researchers have actively contributed. In parallel, a multi-institutional LIGO-India Training, R&D program has brought in many more Indian experimental researchers into the field. A dedicated National training facility has been constructed and commissioned at RRCAT, Indore earlier this year. A recent reassessment of the science case for LIGO-India reinforces the immense scientific promise of adding LIGO-India to the global GW network [M Saleem et al 2021, Class. Quantum Grav., 39, 025004 (2022)].
Recent Progress on the Black Hole Information Paradox

Raghu Mahajan

We give a brief overview of the black hole information problem and describe qualitatively the recent research that shows that the von Neumann entropy of Hawking radiation follows the Page curve, and hence is consistent with unitary, information-preserving time evolution.

1. Classical Black Holes

Maxwell’s electrodynamics and Einstein’s general relativity form the bedrock of modern classical physics [1]. Here the adjective classical is used as a counterpoint to quantum, we will eventually discuss quantum theories too. Classical general relativity modifies Newton’s theory of gravitation, and is experimentally extremely well tested [2]. Black holes are a bizarre but robust prediction of general relativity. Very recent experiments have confirmed the existence of black holes both via the direct imaging of the black hole at the center of the M87 galaxy [3], and the indirect observation of gravitational waves that are emitted when two black holes collide and merge [4].

Black holes are erstwhile stars that exhausted their nuclear fuel and had no other sources of outward pressure to prevent the attractive force of gravity from shrinking the material of the star to “nothing” [5]. The center of mass of the erstwhile star is replaced by a “singularity” where the force of gravity is infinite. (One of the main tasks of any quantum theory of gravity is to prevent the formation of all the various types of singularities that occur in classical general relativity.) The singularity is surrounded by a surface called the event horizon, and the radius of the event horizon is called the Schwarzschild radius. The precise relationship between the mass of the black hole ($M$) and its Schwarzschild radius ($r_S$) is given by

$$ r_S = \frac{2GM}{c^2} . $$

An object with mass equal to the mass of the earth has Schwarzschild radius equal to 9 mm. This also means that if we were to shrink the earth to a radius smaller than 9 mm, it would become a black hole. An object with mass equal to the mass of the sun has Schwarzschild radius equal to 3 km.

In the region between the singularity and the event horizon, also known as the “interior” of the black hole, the force of gravity is very strong; it is so strong that even

---

DOI: https://doi.org/10.1007/s12045-020-1103-y

Keywords
Black holes, Hawking radiation, black hole information problem, Page curve, quantum extremal surface.
Cross section of a black hole

When a massive star collapses under its own gravity, it forms a black hole that is so heavy that it captures everything that passes its event horizon. Not even light can escape. At the event horizon, time replaces space and points only forward. The flow of time carries everything towards a singularity furthest inside the black hole, where density is infinite and time ends.

The light cone shows the paths of the light rays forward and backward in time. When matter collapses and forms a black hole, the light cones that cross the black hole's event horizon will turn inward, toward the singularity. An outside observer will never really see the light rays reach the event horizon, they just nudge it. No one can see further in.

**Figure 1.** The structure of space, time and lightcones near and inside the horizon of a black hole. Image Credit: https://www.nobelprize.org/uploads/2020/10/fig2-phy-en-cross-section.pdf © Johan Jarnestad/The Royal Swedish Academy of Sciences.

light, which is the fastest anything can move, cannot escape this region. Referring to Figure 1, the thick red line is the event horizon, the thick blue line is the singularity, the black region is the interior of the black hole. The yellow cones are the forward lightcones, and every object can only move within this forward lightcone. In the black interior region, the forward lightcone is fully tipped towards the singularity, and thus everything, once it is in the interior region, must fall towards the singularity; nothing can escape.

There is another drawing of the spacetime geometry in the presence of a black hole that conveys the same information but simplifies the structure of the lightcones, at the cost of severely distorting distances between points. This is called a Penrose diagram, and it has the property that lightrays always trace out 45° lines. The Penrose diagram of a spacetime that contains a star collapsing to form a black hole is shown in Figure 2. As in Figure 1, the blue line is the singularity and the red line is the event horizon. The black region is the interior of the black hole. The orange lines are the surface of the collapsing star. The green line is a spacelike surface, and represents a candidate constant time surface that is smooth across the event horizon.
2. Adding Quantum Mechanics to the Mix

Ultimately, every classical theory must be replaced by a quantum theory. Maxwell’s electrodynamics has been successfully elevated to quantum electrodynamics [6]. The same is true of the force that binds together the atomic nucleus and the force that causes beta decay: All three of these forces have a consistent quantum mechanical description under the framework of Quantum Field Theory. The experimental culmination of this effort was the discovery of the Higgs boson in 2012.

The force of gravity is a painful exception, quantum gravity does not fall into the framework of quantum field theory and a satisfactory quantum theory of gravity evades us to date. A quantum theory of gravity must kick in as the matter of a collapsing star shrinks to length scales of order \( \sqrt{\frac{\hbar G}{c^3}} \approx 10^{-35} \text{ m} \), and quantum effects must prevent the formation of the singularity. String theory is the most promising candidate for a quantum theory of gravity, and there exist certain scenarios in which string theory prevents the formation of singularities. However, for singularities that exist in the interior of black holes, the current knowledge in string theory does not offer much insight.

At this point, the reader might complain that if quantum effects in gravity are only important near the black hole singularity, they might not be of any interest to observers who live and conduct experiments outside black holes. However, Stephen Hawking shocked the physics community when he announced his result in 1974 [7, 8] that quantum effects near the horizon of a black hole cause the radius of the event horizon to continuously decrease and eventually disappear. As we mentioned above, the horizon radius of a black hole is of macroscopic size (9 mm for a black hole with mass equal to the mass of the Earth, 3 km for a black hole with mass equal to the mass of the Sun) and we completely understand the laws of physics at these macroscopic scales. Quantum gravity does not fall into the framework of quantum field theory and a satisfactory quantum theory of gravity evades us to date.

Upto unimportant greybody factors, the spectrum of Hawking radiation is the same as the blackbody spectrum. This means that black holes behave as a hot body, but with a very small temperature, proportional to \( \hbar \).
Stephen Hawking showed in 1974 that quantum effects near the horizon of a black hole cause the radius of the event horizon to continuously decrease and eventually disappear. Hawking’s result is insensitive to the precise nature of the yet unknown theory of quantum gravity.

Let’s describe Hawking’s result in more detail. Hawking worked in an approximation where the gravitational field of a black hole is treated classically, using general relativity, while, say, the photon and neutrino fields are treated quantum mechanically. The gravitational field of a black hole acts as a time-dependent perturbation to the Hamiltonian of the quantum mechanical photon and neutrino fields. In the presence of a time-dependent perturbation, quantum systems get excited, even if they start out in their ground states. This is the essence of the Hawking effect, and the excited states of the photon and neutrino fields are nothing but the respective particles: photons and neutrinos. Thus, a black hole produces particles due to quantum effects.

Furthermore, Hawking was able to calculate that, up to mild greybody factors, the spectrum of particles obtained is the same as the blackbody spectrum. The only free parameter in the blackbody spectrum is the temperature, and the temperature of the radiation emanating from a black hole ($T_H$) is simply related to the Schwarzschild radius of the black hole ($r_S$) via

$$k_B T_H = \frac{\hbar}{r_S} \cdot \frac{c}{4\pi}.$$  \hspace{1cm} (2)

Note the multiplicative $\hbar$, this reflects the intrinsically quantum nature of the Hawking temperature. For a solar mass black hole, this corresponds to a temperature of $6 \times 10^{-8}$ K, which is roughly eight orders of magnitude smaller than the temperature of the cosmic microwave background.

Hawking’s result supports the view that black holes, when viewed from the outside are like ordinary hot bodies\(^2\) that have a temperature given by equation (2). In fact, black holes also obey versions of the laws of thermodynamics [9]. Perhaps most importantly, they also have a nonzero and finite entropy [10]

$$S_{BH} = k_B \cdot \frac{1}{4} \cdot 4\pi r_S^2 \cdot \frac{c^3}{G\hbar}.$$  \hspace{1cm} (3)

Up to universal constants that are needed to match the dimensions, the entropy is equal to one-quarter of the area of the horizon. In fact, one of the major successes of string theory is to provide a counting of the states that accounts for this entropy [11]. Note that, in the classical limit $\hbar \to 0$, the entropy is infinite.

We must now address the very important question of why the gravitational field of a black hole acts as a time-dependent perturbation to the Hamiltonian of the quantum mechanical photon and neutrino fields. To define the Hamiltonian of a quantum system we need to first define a time function on the spacetime and correspondingly, pick three-dimensional spatial slices, which are nothing but the level sets of this time function. It is extremely essential that the level sets be purely spatial. An example

---

2 Note that the temperature is inversely proportional to $r_S$ while the entropy is proportional to $r_S^2$. This is unusual. For ordinary objects, the entropy increases as temperature increases. This indeed poses a major problem for black holes in flat space, but there exist black holes in negatively curved ambient space whose entropy increases with temperature.
of such a spatial slice is the green line shown in Figure 2. Looking at the left panel of Figure 1, we see that the naive notions of space and time get interchanged as we go from the outside of a black hole to the inside. This is a hint that, in the presence of a black hole horizon, there is no possible choice of spatial slices that go smoothly through the horizon and are such that time translations are a symmetry. The lack of a time-translation symmetry is equivalent to saying that the Hamiltonian is time-dependent.

3. The Information Loss, or the Entropy Production Problem

Ok, let’s take stock. A star collapses to form a black hole, the black hole emits a thermal gas of particles due to quantum effects, its horizon radius continuously shrinks, and the black hole, the entire black region in Figure 1 eventually disappears. What is the big deal? The point is that the final state seems to be just a gas of particles at some temperature, and this gas of particles has a nonzero entropy. The initial state, containing the star, could very well have had zero entropy: We start the star out in some definite, known state. No known law of physics produces entropy, but apparently the process of black hole formation and evaporation seems to do so [12]. This is the infamous information loss problem.

It should be emphasized here that the second law of thermodynamics which says that entropy always increases is merely a statement about our ignorance of the precise state of a complex many-particle system: We are much too feeble to keep track of anything but coarse variables such as the temperature and pressure of a gas. This is the sole reason why “entropy increases”. Strictly speaking, if we kept track of all the details of every single particle in our huge system containing of order Avogadro’s number of particles, entropy would stay zero if it starts at zero.

Thus, the production of entropy in the formation and evaporation of a black hole is a huge problem. The differential equations that determine time-evolution in classical physics are reversible. In quantum mechanics, the evolution of a wave-vector in Hilbert space is defined by a unitary matrix, and this is also a reversible evolution. Are black holes really an exception, and is entropy really produced, in a fundamental sense, when a black hole forms and evaporates?

4. The Page Curve

We now need to introduce another crucial ingredient into the story: The physics of entanglement [13]. A basic fact about quantum mechanics is that the joint Hilbert space of two independent quantum systems contains states that are not factorizable into states of the individual quantum systems. Such states are called entangled. The singlet state of two spin-\(\frac{1}{2}\) particles is the quintessential example of an entangled state. When two electrons are in an entangled state, the physics of both electrons is described

No known law of physics produces entropy, but apparently the process of black hole formation and evaporation seems to do so. This is the infamous information loss problem.

Up to universal constants that are needed to match the dimensions, the entropy of a black hole is equal to one-quarter of the area of the horizon. One of the major successes of string theory is to provide a counting of the states that accounts for this entropy.

A basic fact about quantum mechanics is that the joint Hilbert space of two independent quantum systems contains states that are not factorizable into states of the individual quantum systems.
using a wave-vector, but the physics of either one of the two electrons cannot be described using a wave-vector. The individual electrons must be described using a Density Matrix. For a spin-\(\frac{1}{2}\) particle, the density matrix is a \(2 \times 2\) matrix. In general, the density matrix is a square matrix with size equal to the dimension of the Hilbert space \((\dim \mathcal{H})\) of the system. This density matrix is usually denoted by the Greek letter \(\rho\), and it is a Hermitian, positive-definite matrix with \(\text{Tr}(\rho) = 1\). The von-Neumann entropy of \(\rho\) is defined as

\[
S(\rho) = -\text{Tr}(\rho \log \rho).
\]

(4)

If the state of a quantum system can be described by a wave-vector, we say that the state is **pure**. Pure states can also be described using a density matrix, but density matrices corresponding to pure states satisfy \(S(\rho) = 0\). States that are not pure are said to be mixed and have \(S(\rho) > 0\). The largest value that \(S(\rho)\) can take is \(\log(\dim \mathcal{H})\) and a state with \(S(\rho) = \log(\dim \mathcal{H})\) is said to be maximally-mixed. The singlet state of two spin-\(\frac{1}{2}\) particles is maximally mixed.

We can now restate Hawking’s paradox. Time evolution in quantum mechanics is unitary, and this implies that pure states stay pure under time evolution. However, the process of formation and evaporation of a black hole seems to turn pure states into mixed states.

In fact, things are much more drastic. If we split a quantum system into two parts \(A\) and \(B\), and if the joint state of the system \(AB\) is pure, then one can show that \(S(\rho_A) = S(\rho_B)\). This means that entropy of a very large subsystem cannot be larger than the entropy of the complementary, small subsystem. Don Page [14] gave an argument that for a typical state in a large quantum system, say a system of \(N\) spin-\(\frac{1}{2}\) electrons, the entropy of a subsystem of \(k\) electrons grows approximately linearly as \(S(k) \approx k\), when \(k < N/2\). When \(k > N/2\), our subsystem is more than half the total system, and so the complementary system contains less than \(N/2\) electrons, and the entropy is then roughly equal to \(N - k\). This tent-shaped curve is now referred to as the Page curve, and is sketched in Figure 3.

Now imagine a large black hole that forms and starts to evaporate. We will keep track of the entropy of the emitted radiation as a function of time. Hawking’s calculations predict that the entropy of the radiation should increase linearly, and this is consistent with the initial, increasing phase of the Page curve. This build up of entropy happens because the Hawking radiation which is present outside the black hole becomes entangled with the black hole itself. However, beyond the so-called **Page time**, when the black hole has half-evaporated, the prediction from the Page curve is that the entropy of the radiation should start to **decrease**, following the solid red line in Figure 3. Hawking’s calculations say the exact opposite: the entropy of the radiation outside should keep increasing (the solid plus dashed blue curve in Figure 3) until the black hole has completely disappeared and we are left with just the Hawking radiation outside. After the Page time, the entropy of Hawking radiation as predicted from
general relativity coupled to quantum fields differs, by an order-one fraction\(^3\), from the answer expected from very general arguments in quantum mechanics. Note that the puzzle appears when the black hole is still huge and the horizon radius is still macroscopic: a large black hole that has shrunk to half its size still has a macroscopic horizon.

5. Recent Developments

The tension between Hawking’s prediction and the expectations from Page’s general arguments is dire enough that it has led people to conjecture that the space near the horizon of a black hole is not “empty”, and that perhaps there is a “firewall” [15, 16] at the horizon which cuts out the interior region, or perhaps that the entire interior region of the black hole should be replaced by a horizonless “fuzzball” [17]. See, for example, [18] for a recent pedagogical review of these scenarios and the debate around the black hole information puzzle in the last decade.

Research from last year offers a definite resolution of the above entropy puzzle without resorting to either of these dramatic possibilities [19–21]. Building on years of intuition and sharpening of information theoretic tools in AdS/CFT holography [22, 23], these papers computed the entropy of Hawking radiation using the so-called Quantum Extremal Surface (QES) prescription [24], using only the setup of general relativity coupled to quantum fields. However, a novel feature is introduced compared to Hawking’s calculation. At a technical level, the novelty is the existence of a hitherto unanticipated QES which lies very close to the slowly receding black hole horizon.

We will summarize a few key takeaways from this research, leaving the details to the review [21].

\(^3\) A fractional error of order-one means that the error is huge.
Building on years of intuition and sharpening of information theoretic tools in AdS/CFT holography, researchers in 2019 computed the entropy of Hawking radiation using the so-called Quantum Extremal Surface (QES) prescription. A novel feature is introduced compared to Hawking’s calculation: the novelty is the existence of a hitherto unanticipated QES which lies very close to the slowly receding black hole horizon.

The entropy computed using the novel QES is almost equal to the Bekenstein–Hawking entropy, or area, of the slowly evaporating black hole. Thus it is naturally decreasing and goes to zero at the end of evaporation. Thus the final state of the Hawking radiation is pure.

1. The QES prescription [24] is a generalization of the notion of the Ryu-Takayanagi prescription [25], which is a recipe for computing von Neumann entropies of subsystems of a holographic quantum field theory.

2. The computations of [19, 20] are insensitive to details of the yet unknown quantum theory of gravity. They only use the setup of general relativity coupled to quantum fields, but the use of the QES prescription to compute entropies introduces a novel feature compared to Hawking’s calculations [7, 8, 12].

3. The QES prescription involves the minimization of a certain quantity, and the tent shape of the Page curve emerges because there are two extrema that exchange dominance at the halfway evaporation point. These two extrema correspond to the naive QES (blue solid plus dashed line in Figure 3) and the new hitherto unanticipated QES (red solid plus dashed line in Figure 3).

4. The entropy computed using the novel QES is almost equal to the Bekenstein–Hawking entropy, or area, of the slowly evaporating black hole. Thus it is naturally decreasing and goes to zero at the end of evaporation. Thus the final state of the Hawking radiation is pure.

For a more pictorial description, let us refer to Figure 4, which is taken from [26]. The figure depicts the Penrose diagram of an evaporating black hole, and a constant time surface at late times, $\Sigma_{\text{Late}}$. The Hawking radiation lives in the blue region to the right of the square bracket, drawn on $\Sigma_{\text{Late}}$. The nontrivial QES first discovered in [19, 20] is denoted by the thick blue dot. The variables $t, r_0, y^\pm_e$ are coordinate labels and irrelevant to the discussion here. The region to the left of the QES was dubbed the “island” in [26], which we now discuss.

![Figure 4](image-url)
6. Strict Locality vs Unitarity

One of the upshots of the existence of island regions is that faraway regions of space in the presence of gravity are not completely independent of each other. In the usual setup of quantum field theory [6], operators are labeled by spacetime points and the fact that signals cannot travel faster than the speed of light is encoded in the fact that operators located on the same spatial slice commute with each other, unless they happen to be at coincident points. One of the implications of the existence of island regions is that this is no longer true when gravity is involved. Doing something sufficiently complex to the Hawking radiation far away can create a particle in the island region in the interior of the black hole [27–29]. This is true even for black holes in equilibrium with a thermal bath of radiation [30].

Of course, it has long been appreciated that a quantum theory of gravity must have certain degree of nonlocality. In hindsight, this is perhaps not a surprise: In a theory of quantum gravity, the geometry of spacetime fluctuates, thus the mutual causal relationships of pairs of points in spacetime also fluctuate, and it is impossible to say whether two given points are timelike or spacelike separated from each other. To put it differently, in a theory of quantum gravity, unitarity is more sacrosanct than locality. However it is very hard to see violations of locality because experiments that detect violations of locality involve extremely complex operations.

To summarize, recent research has produced a computation of the von Neumann entropy of Hawking radiation that is consistent with information-preserving, unitary time evolution. Adding quantum effects to gravity like Hawking did, while, crucially, at the same time incorporating the QES prescription as a tool to compute entropies, is enough to recover the Page curve.

7. Problems for the Future

The von Neumann entropy is only one number constructed out of the huge matrix that encodes the final state of the Hawking radiation. No computation has actually determined the final state in all its detail, it would be desirable to know as much as possible about the final state. Also, the full implications of the nonlocality implied by the existence of islands remain to be worked out. Another extremely important unanswered question is about the ultimate fate of the singularity in a complete theory of quantum gravity, and how it affects the precise state of the final Hawking radiation. These questions are likely to occupy researchers for years, likely decades, to come.

Suggested Reading


The Story of the Photon*

N Mukunda

An account of the story of the light quantum or photon is given, from its inception in 1905 to its final acceptance in 1924. Necessary background information on radiation theory and historical details are included.

1. Introduction

The photon, so named by the physical chemist Gilbert Norton Lewis in 1926, is a child of the 20th century. It is the ‘particle of light’ – or ‘light quantum’ – first hypothesized by Albert Einstein in 1905, and then used by him to explain, among other things, the photoelectric effect. The story of the photon is rich in history, development of ideas, experiment and personalities. In this account an attempt will be made to convey something of each of these aspects; the fundamental motivations and currents of ideas will be described as carefully as possible, and only selected derivations will be presented.

During the year 1905, aptly called ‘Einstein’s Miraculous Year’, he submitted five research papers for publication and also completed his Ph.D. thesis. Of the former, three have become all-time classics. In chronological sequence they are: the light quantum paper (March), the paper on the Brownian Motion (May), and the paper establishing the Special Theory of Relativity (June). Einstein himself felt that of these only the first was truly path-breaking, for he wrote in a letter of May 1905 to his friend Conrad Habicht: “I promise you four papers . . . the first of which I could send you soon . . . The paper deals with radiation and the energetic properties of light and is very revolutionary, as you will see . . .”.

2. Radiation Theory from Kirchoff to Planck – a Capsule

The study of (electromagnetic) radiation forms a glorious chapter in the history of physics. The first major step was taken in 1859 by Gustav Kirchoff (the ‘grandfather’ of the quantum theory) when he proved the following result: if radiation and material bodies are in equilibrium at a common (absolute) temperature $T$, the former being reflected, scattered, absorbed and emitted by the latter, then the energy density of the radiation per unit frequency interval is a universal function of frequency and

---


---

“I promise you four papers . . . the first of which I could send you soon . . . The paper deals with radiation and the energetic properties of light and is very revolutionary, as you will see . . .”

Einstein to Conrad Habicht,
May 1905
temperature, independent of the particular material bodies present:

\[
\rho(v, T) \Delta v = \text{energy of radiation per unit volume in the frequency range } v \text{ to } v + \Delta v, \text{ at temperature } T = (\text{universal function of } v \text{ and } T) \times \Delta v.
\]  

(1)

For the proof, Kirchoff used the Second Law of the then young science of thermodynamics; and he posed the determination and understanding of the function \( \rho(v, T) \) as a major experimental and theoretical challenge. Such radiation is variously called ‘black-body’ or ‘temperature’ or ‘thermal’ radiation.

Twenty years later, in 1879, the experimentalist Josef Stefan measured the total energy density of thermal radiation by ‘summing’ over all frequencies, and then conjectured that it was proportional to \( T^4 \):

\[
u(T) = \text{total energy density of thermal radiation} = \int_0^\infty d\nu \rho(v, T) = \sigma c T^4.
\]  

(2)

Soon after, in 1884, Ludwig Boltzmann was able to give a thermodynamic proof of this result, using Maxwell’s result that the pressure of radiation is one third of its energy density. (See Box 1.) Once again, this was an outstanding and early application of thermodynamics to radiation problems – more were to follow. The constant \( \sigma \) in (2) is named jointly after Stefan and Boltzmann.

From the 1860’s onwards many guesses were made for the form of the function \( \rho(v, T) \). In 1893 Wilhelm Wien constructed a clever thermodynamical argument and proved that the dependences of \( \rho(v, T) \) on its two arguments were correlated by a scaling law:

\[
\rho(v, T) = v^3 f(v/T),
\]  

(3)

so the original Kirchoff problem became that of finding the form of the universal function \( f(v/T) \) involving only one argument. He followed this up soon after in 1896 by offering a guess for the form of \( f(v/T) \), inspired by the Maxwell velocity distribution in a classical ideal gas: with two constants \( \alpha \) and \( \beta \) he suggested

\[
f(v/T) = \alpha e^{-\beta v/T},
\rho(v, T) = \alpha v^3 e^{-\beta v/T}.
\]  

(4)

Early experiments by Friedrich Paschen (reported in January 1897) gave support to the Wien Law (4). They were done in the near infrared part of the spectrum, with wavelengths \( \lambda \) in the range \((1 \text{ to } 8) \times 10^4 \text{ Å} \) and temperatures \( T \) in the range 400 to 1600 K; and showed the validity of the Wien Law in the high frequency limit.

Now we turn to Max Karl Ernst Ludwig Planck, successor to Kirchoff and the ‘father’ of the quantum theory. His major goal was the theoretical determination of Kirchoff’s
Consider thermal radiation, at temperature $T$, enclosed in a spatial volume $V$, and treat $T$ and $V$ as independent variables. The total energy $U = V \ u(T)$ where $u(T)$ is the energy density including all frequencies. The pressure, according to Maxwell, is one third the energy density: $p = \frac{U}{3V} = \frac{u(T)}{3}$. (In contrast, for a classical (nonrelativistic) ideal gas of $n$ particles the total energy $U = \frac{3}{2}nkT$ is volume independent; while from the ideal gas law the pressure is two-thirds the energy density, $p = \frac{2}{3}U$). The Second Law of Thermodynamics implies that the expression

$$dS = \frac{1}{T}(dU + pdV)$$

must be a perfect differential. Writing this out as

$$dS = \frac{1}{T}\left(u(T)dV + V\frac{du(T)}{dT}dT + \frac{u(T)}{3}dV\right),$$

this means that

$$\frac{\partial}{\partial T} \left(\frac{4}{3} \frac{u(T)}{T}\right) = \frac{\partial}{\partial V} \left(V\frac{du(T)}{dT}\right),$$

which simplifies to

$$T\frac{du(T)}{dT} = 4 u(T).$$

The solution is the Stefan–Boltzmann Law:

$$u(T) = \text{Constant} \times T^4.$$
exact form

\[
f(\nu/T) = \frac{8\pi k}{c^3} \cdot \frac{T}{\nu}
\]

\[
\rho(\nu, T) = \frac{8\pi \nu^2}{c^3} \cdot k T,
\]

(6)

with \(c\) the vacuum speed of light and \(k\) the Boltzmann constant. Slightly later, by October 1900, Heinrich Rubens and Ferdinand Kurlbaum did experiments in the deep infrared, \(\lambda = (3 \text{ to } 6) \times 10^5 \text{ Å}, \ T = 200 \text{ to } 1500 \text{ K},\) and found again deviations from the Wien Law (4) but agreement with the Rayleigh expression (6).

Sunday, October 7, 1900 is the birthdate of the quantum theory. On the afternoon of that day, Rubens visited Planck’s home for tea, and told him of his and Kurlbaum’s latest experimental results. After he left, Planck set to work. He realised that Wien’s Law could not be the final answer to Kirchoff’s problem. While it was obeyed at high enough frequencies, it failed at the low frequency end where the Rayleigh form was valid. What Planck achieved that evening was a mathematical interpolation between these two limiting forms. His strategy seems roundabout but was, in retrospect, fortunate. He had in earlier work related the Kirchoff function \(\rho(\nu, T)\) to the average energy \(\overline{E}(\nu, T)\) of a charged material oscillator with natural frequency \(\nu\) and at a temperature \(T\), by balancing the effect on it of incident radiation and its own emission of radiation. This ‘Planck link’ reads

\[
\rho(\nu, T) = \frac{8\pi \nu^2}{c^3} \overline{E}(\nu, T).
\]

Planck translated the limiting forms of \(\rho(\nu, T)\) in the high \(\nu\) (Wien) and low \(\nu\) (Rayleigh) limits into corresponding limiting forms for \(\overline{E}(\nu, T)\); converted this into limiting forms for the entropy \(S(\overline{E})\) of the material oscillator (written as a function of energy) at high and low \(\overline{E}\), respectively; and then by solving a simple differential equation found a formula interpolating between these limiting expressions. Translating all this back into the original problem his result for Kirchoff’s function \(\rho(\nu, T)\) is the Planck radiation law we all know so well:

\[
\rho(\nu, T) = \frac{8\pi \nu^2}{c^3} \cdot \frac{h\nu}{e^{h\nu/kT} - 1}
\]

(8)

A new fundamental constant of nature with the dimensions of action, Planck’s constant \(h\), entered his result; and there was agreement with experiment at all measured frequencies. On October 19, 1900, Planck announced his formula following a talk given by Kurlbaum. In the high frequency limit we recover the Wien result (4) from (8) with

\[
\alpha = \frac{8\pi h}{c^3}, \ \beta = \frac{h}{k}
\]

(9)

Comparing (7) and (8) it follows that Planck’s formula implies that the average energy of a material oscillator \(\overline{E}(\nu, T)\) must have a value differing from the result \(kT\) of the
The equipartition theorem:

\[ \overline{E}(\nu, T) = h\nu / \left( e^{h\nu/kT} - 1 \right) \]  

(10)

During the period October to December 1900 Planck tried very hard to find a theoretical basis for this formula. Finally, “... as an act of desperation... to obtain a positive result, under any circumstances and at whatever cost”, he invented the concept of irreducible packets or quanta of energy for matter, and in mid-December 1900 he presented the following statistical derivation of (10). He imagined a large number, \( N \), of identical (but distinguishable!) material oscillators, with a total energy \( E \) and at a temperature \( T \). Assuming that this total energy \( E \) was made up of \( P \) (indistinguishable!) packets or quanta of energy \( \epsilon_0 \) each, (so that \( E = P\epsilon_0 \) and the energy of each oscillator is an integer multiple of \( \epsilon_0 \)), he counted the number of ways \( W \) (number of micro states or complexions) in which these packets could be distributed over the \( N \) oscillators. By a simple combinatorial argument, followed by an application of the Boltzmann entropy relation \( S = k \ln W \), he computed the entropy \( S/N \) per material oscillator, connected it up to the temperature \( T \), and finally arrived at the result (10) he was after, with the identification \( \epsilon_0 = h\nu \).1

3. Einstein’s State of Preparedness

It is time now to turn to Einstein. Already since 1897 during his student days at the Eidgenössische Technische Hochschule in Zurich he had become familiar with Kirchoff’s work on thermal radiation. From his teacher Heinrich Friedrich Weber in 1899 he learnt about Wien’s theorem (3) and the resulting Wien Displacement Law. He was also familiar with Planck’s work, and while he had full faith in the experimental validity of the Planck law (8), he was acutely conscious of the absence of a proper theoretical basis for it. (See Box 2 for a brief account of Einstein’s involvement with Planck’s Law). During the period 1902–1904 he rediscovered for himself the foundations and key concepts of statistical physics, obtaining independently many of Josiah Willard Gibbs’ results. He invented on his own the concept of the canonical ensemble, derived the equipartition law for energy, found ways to use the ‘Boltzmann Principle’ \( S = k \ln W \), and found the formula for energy fluctuations for a mechanical system at a given temperature. (See later.) The empirical validity of the Planck Law (8) and the realisation that it could not be derived from the classical Maxwell theory of electromagnetic radiation convinced him that the picture of radiation given by the latter had to be modified by incorporating quantum features in some way. As he was to say much later: “Already soon after 1900, i.e., shortly after Planck’s trailblazing work, it became clear to me that neither mechanics nor thermodynamics could (except in limiting cases) claim exact validity”.

Einstein independently derived, in his March 1905 paper, the Rayleigh–Jeans Law (6): he started from the ‘Planck link’ (7) between radiation and matter, used the equipartition law to substitute \( kT \) for the average energy \( \overline{E}(\nu, T) \) of the material oscillator, and

---

1 The symbol \( k \) for Boltzmann’s constant first appeared in the Planck law (8) in 1900. The formula \( S = k \ln W \) was given the name ‘Boltzmann’s Principle’ by Einstein.
Here is a chronological list of the many occasions and ways in which Einstein ‘played’ with the Planck radiation law and ‘teased out’ its deep consequences:

1905: Examines the volume dependence of entropy of radiation in the Wien limit, abstracts the light quantum idea, applies it inter alia to the photoelectric effect.

1909: Calculates energy fluctuations for thermal radiation using the complete Planck Law; arrives at the earliest ever statement of wave-particle duality in nature; considers also momentum fluctuations of a mirror placed in thermal radiation, due to fluctuations in radiation pressure.

1916: Derives the Planck Law based on Bohr’s theory of stationary states and transitions, and processes of absorption, induced and spontaneous emission of radiation by matter. Extends the 1905 analysis to show that individual light quanta are directed in space and carry momentum.

1924-25: Extends Bose’s derivation of the Planck Law to matter, finds particle-wave duality for matter, predicts Bose–Einstein condensation.

directly obtained (6)! Thus there were two theoretically well-founded, but experimentally invalid, routes to the Rayleigh–Jeans result: one applying equipartition directly to radiation; and another using the ‘Planck link’ and then applying equipartition to the material oscillator.

Added to all this, it should be mentioned that in the course of some work on the molecular theory of gases done in 1904, Einstein had realised the importance of the volume dependence of thermodynamic quantities, in particular of the entropy. The relevance of this will become clear presently.

4. The ‘Light Quantum’ Paper of 1905

Einstein’s views, circa 1905, on the radiation problem may be summarised as follows: the Planck Law is experimentally accurate but has no proper theoretical basis; the Rayleigh–Jeans limit has a proper classical theoretical foundation but is experimentally unacceptable; the Wien limit is a guess, with no derivation from first principles or classical basis, and is experimentally valid only at high frequencies. He also declared right away that, in spite of the success of Maxwell’s wave theory in explaining typical optical phenomena, he believed it was necessary to replace it by a different picture in which radiant energy is made up of discontinuous spatially localized quanta of finite energy, which could be absorbed and emitted only as complete units.

Einstein then took a ‘phenomenological’ attitude to the radiation problem: since Wien’s Law (4) is experimentally valid in a definite domain and has no classical
underpinnings, an examination of this domain from the thermodynamical point of view – involving radiation on its own and not using the ‘Planck link’ at all – should reveal key nonclassical features of radiation.

Apart from the independent derivation mentioned above of the Rayleigh-Jeans Law, in his paper Einstein recalls some results of Wien on the entropy of radiation. He then uses this to calculate the volume dependence of the entropy of thermal radiation in the Wien limit; gives the corresponding calculation for a classical ideal gas; compares the two results; and then draws his epoch-making conclusions about the existence and nature of radiation quanta. The Wien limit calculation given by Einstein is essentially equivalent to the following.

Consider thermal radiation at temperature $T$ and between frequencies $\nu$ and $\nu + \Delta \nu$, contained in a spatial volume $V$. The total energy, $E$ say, of this radiation is given, when the Wien limit is applicable, by

$$E = V \alpha \nu^3 e^{-\beta \nu / T} \cdot \Delta \nu = N V e^{-\beta \nu / T},$$

$$N = \alpha \nu^3 \Delta \nu.$$  \hspace{1cm} (11)

Treating $E$ and $V$ as the independent thermodynamic variables, the inverse temperature is

$$\frac{1}{T} = \frac{1}{\beta \nu} (\ln N + \ln V - \ln E).$$ \hspace{1cm} (12)

The entropy $S(E, V)$ of this portion of Wien radiation is obtained by integrating the basic thermodynamic relation

$$\frac{\partial S(E, V)}{\partial E} = \frac{1}{T} = \frac{1}{\beta \nu} (\ln N + \ln V - \ln E),$$

the dependences of $S(E, V)$ on $\nu$ and $\Delta \nu$ being left implicit. This leads to

$$S(E, V) = \frac{E}{\beta \nu} (\ln N + \ln V + 1 - \ln E),$$ \hspace{1cm} (14)

(apart from a function of $V$ alone which must vanish since $S(E, V) \to 0$ as $E \to 0$). If we now compare the values of the entropy for two different volumes $V_1$ and $V_2$, keeping $E$ (and of course $\nu$ and $\Delta \nu$) fixed, we find:

$$S(E, V_1) - S(E, V_2) = \frac{E}{\beta \nu} \ln \left( \frac{V_1}{V_2} \right)$$

$$= k \ln \left( \frac{V_1}{V_2} \right)^{E/h \nu},$$ \hspace{1cm} (15)

where the value of the Wien constant $\beta$ was taken from (9).

Einstein then follows up the derivation of the result (15) by a detailed calculation of a similar entropy difference for a classical ideal gas of $n$ molecules. For this he exploits
The entropy of a classical ideal gas of \( n \) particles has the form
\[
S(E, V) = nk \left( \ln V + \frac{3}{2} \ln \left( \frac{2E}{3nk} \right) \right)
\]
While the volume dependence is similar to that in (14), the energy dependence is quite different.

The ‘Boltzmann Principle’ \( S = k \ln W \) relating entropy to statistical probability; omitting the details of his argument, he arrives at the result\(^2\)
\[
S(E, V_1) - S(E, V_2) = k \ln \left( \frac{V_1}{V_2} \right)
\]
Comparison of the two results (15) and (16) leads to his profound conclusion:

“... We (further) conclude that monochromatic radiation of low density (within the range of validity of Wien’s radiation formula) behaves thermodynamically as if it consisted of mutually independent energy quanta of magnitude \( hv \). (Einstein actually wrote \( R \beta v/N \) for this last expression, which is just \( hv \)). Note carefully the explicit mention that this refers to radiation in the Wien limit; indeed the use of the complete Planck Law does not lead to such a result! Note also the conclusion that the energy quanta are mutually independent, reflecting the comparison being made to the classical ideal gas.

Thus was the concept of ‘light quanta’ first arrived at, with its stated limitations. Nevertheless, right away Einstein abstracts the key idea and boldly extrapolates it beyond these limitations to formulate his ‘heuristic principle’:

“If monochromatic radiation (of sufficiently low density) behaves, as concerns the dependence of its entropy on volume, as though the radiation were a discontinuous medium consisting of energy quanta of magnitude \( hv \), then it seems reasonable to investigate whether the laws governing the emission and transformation of light are also constructed as if light consisted of such energy quanta”. Thus he proposes that in the processes of emission and absorption and interaction of light with matter, the same particulate nature should be seen!

Einstein concluded his paper by applying his ‘heuristic principle’ to three experimental observations: the Stokes rule in photo luminescence, the photoelectric effect, and lastly the ionization of gases by ultraviolet light. We look next briefly at some highlights of the second of these applications.

5. The Photoelectric Effect

This effect was discovered accidentally by Heinrich Hertz in 1887 while studying sparks generated by potential differences between metal surfaces. (Remember at that time the electron was not yet known!). After Joseph John Thomson discovered the electron in 1897, he turned to the photo electric effect and in 1899 could state that it was the electron that was ejected when ultraviolet light shone on a metal surface. In experiments around 1902 Philip Lenard studied the dependence of the ejected electron’s energy on the intensity and frequency of the incident radiation – independent of the former, increasing with the latter.

In his 1905 paper Einstein proposed the following ‘simplest conception’ for what happens: a light quantum transfers all its energy to a single electron, independent...
of other quanta present and disappearing in the process; the electron emerges from the metal surface carrying with it the photon’s energy except for what it has to ‘pay’ to leave the metal. He then proposed the following famous and simple equation (in modern notation) for the maximum energy of the emitted electron:

\[ E_{\text{max}} = h\nu - P, \tag{17} \]

where \( \nu \) is the frequency of incident radiation and \( P \) – the work function characteristic of the metal – the energy lost by the electron in the release process.

The most extensive series of experiments to test (17) were carried out by Robert Andrews Millikan in the decade upto 1915, even though he was extremely skeptical about the light quantum hypothesis itself. In his 1915 paper he said: “Einstein’s photoelectric equation… appears in every case to predict exactly the observed results… Yet the semicorpuscular theory by which Einstein arrived at his equation seems at present wholly untenable”. Many years later, in 1949, he reminisced in these words: “I spent ten years of my life testing that 1905 equation of Einstein’s and contrary to all my expectations, I was compelled in 1915 to assert its unambiguous verification in spite of its unreasonableness, since it seemed to violate everything we knew about the interference of light”.

We discuss reasons for the widespread opposition to the photon idea later; let us conclude this section by quoting from the 1921 Physics Nobel Award citation to Einstein: “… for his services to theoretical physics and in particular for his discovery of the law of the photoelectric effect”.

6. Wave-Particle Duality, Photon Momentum

We saw that in 1905 Einstein worked only with the Wien limit of the Planck Law, not the latter in its entirety. In 1909 he went back to the Planck Law itself. As was mentioned earlier, in 1904 he had derived on his own the energy fluctuation formula on the basis of the canonical ensemble construction:

\[ (\Delta E)^2 = \langle E^2 \rangle - \langle E \rangle^2 = kT^2 \frac{\partial}{\partial T} \langle E \rangle. \tag{18} \]

(We take temperature \( T \) and volume \( V \) as the independent variables, and leave implicit the dependences of the average energy \( \langle E \rangle \) on these). Considering thermal radiation contained in the frequency range \( \nu \) to \( \nu + \Delta \nu \) and in a unit spatial volume, at temperature \( T \), the Planck Law gives:

\[ \langle E \rangle = \frac{8\pi \nu^2}{c^3} \cdot \Delta \nu \frac{h\nu}{e^{h\nu/kT} - 1}. \]
\[(\Delta E)^2 = kT^2 \cdot \frac{8\pi h^3 v^3 \Delta v}{c^3 \cdot \frac{e^{hv/kT}}{(e^{hv/kT} - 1)^2} \cdot \frac{hv}{kT^2}}\]
\[= \frac{8\pi h^2 v^4 \Delta v}{c^3 \cdot \left(\frac{1}{(e^{hv/kT} - 1)^2} + \frac{1}{(e^{hv/kT} - 1)}\right)}\]
\[= \frac{c^3}{8\pi v^2 \Delta v} \langle E \rangle^2 + hv \langle E \rangle. \quad (19)\]

At this point the reader is encouraged to check that if \(\langle E \rangle\) had been given purely by the Rayleigh–Jeans expression (6), only the first term on the right would have been obtained; while if \(\langle E \rangle\) was given solely by the Wien expression (4) only the second term on the right would have appeared. Recalling that the Rayleigh–Jeans Law is the unambiguous result of classical Maxwell wave theory and the equipartition theorem, while the Wien Law led to the light quantum hypothesis, we see in the energy fluctuation formula (19) a synthesis or duality of wave and particle aspects of radiation. In Einstein’s words: “... It is my opinion that the next phase in the development of theoretical physics will bring us a theory of light that can be interpreted as a kind of fusion of the wave and the emission theories... (The) wave structure and (the) quantum structure... are not to be considered as mutually incompatible...”

Fourteen years later, in 1923, Prince Louis Victor de Broglie would suggest a similar particle-wave duality for the electron.

The next time Einstein turned to the Planck Law was in 1916 when he gave a new derivation of it based on Bohr’s 1913 theory of stationary states of atoms (and molecules) and transitions between them accompanied by emission or absorption of radiation. In his work, Einstein introduced the famous A and B coefficients characterising the interaction between matter and radiation, and corresponding to the three distinct processes of absorption, induced emission and spontaneous emission of radiation by matter. Planck’s radiation law was shown to be the result of equilibrium among these processes, given Bohr’s postulates and the Boltzmann distribution for the numbers of molecules in the various energy or stationary states. While we will not reproduce this beautiful work here, let us mention that at the same time Einstein completed his physical picture of the light quantum – not only was it a localized parcel of energy \(hv\), it was directed and carried a momentum \(hv/c\) in its direction of motion as well. (Initial steps in this direction had earlier been taken by Einstein in 1909, by considering the momentum fluctuations of a mirror immersed in thermal radiation, as a result of fluctuations in the radiation pressure.) This result was derived by carefully analysing both energy and momentum balances when a molecule makes a transition from one energy level to another via emission or absorption of radiation, and demanding stability of the Planck distribution for radiation on the one hand, and of the Boltzmann distribution for molecules on the other.

It is interesting to realise that it took the discoverer of special relativity from 1905 to 1916 to complete the picture of light quanta. Remember though that the creation of the General Theory of Relativity had kept him busy up to November 1915.
In any case, with this additional insight into the kinematical properties of the light quantum Einstein was fully convinced of its reality. In 1917 he wrote to Besso: “With that, (the existence of) light quanta is practically certain”. And two years later: “I do not doubt any more the reality of radiation quanta, although I still stand quite alone in this conviction”.

7. Opposition to the Light Quantum – the Compton Effect

Why was there such prolonged and widespread reluctance to accept the idea of light quanta? In the cases of the electron, proton and neutron, all of which were experimental discoveries, the concerned particles were quickly accepted into the body of physics. But it was indeed very different with the photon.

One reason may have been Einstein’s own sense of caution which he expressed in 1911 in this way: “I insist on the provisional character of this concept (light quanta) which does not seem reconcilable with the experimentally verified consequences of the wave theory”. On several occasions people like Max von Laue, Arnold Sommerfeld and Millikan misinterpreted Einstein’s statements to mean that he had gone back on his ideas! Apart from that the main reason seems to have been a near universal feeling that Maxwell’s description of radiation should be retained as far as free radiation was concerned, and the quantum features should be looked for only in the interaction between matter and radiation. Indeed Planck said in 1907: “I am not seeking the meaning of the quantum of action (light-quanta) in the vacuum but rather in places where absorption and emission occur, and (I) assume that what happens in the vacuum is rigorously described by Maxwell’s equations”. And again in 1909: “I believe one should first try to move the whole difficulty of the quantum theory to the domain of the interaction between matter and radiation”. It is also amusing to see what Planck and others said in 1913 while proposing Einstein for election to the Prussian Academy of Sciences: “In sum, one can say that there is hardly one among the great problems in which modern physics is so rich to which Einstein has not made a remarkable contribution. That he may sometimes have missed the target in his speculations, as, for example, in his hypothesis of light quanta, cannot really be held too much against him, for it is not possible to introduce really new ideas even in the most exact sciences without sometimes taking a risk”.

The situation changed decisively only after the discovery of the Compton effect by Arthur Holly Compton in 1923. This is the scattering of a photon by a (nearly) free electron; the validity of the energy and momentum conservation laws convinced most skeptics of the reality of light quanta. The relation between the change in frequency of the photon and the scattering angle is very simply calculable in the photon picture, and agrees perfectly with experiment; classical explanations do not work. (Today in the language of quantum field theory we say the incident photon is annihilated and the final photon with different frequency and momentum gets created, while the
electron continues to exist throughout). In a popular article in 1924 Einstein remarked: “The positive result of the Compton experiment proves that radiation behaves as if it consisted of discrete energy projectiles, not only in regard to energy transfer but also in regard to Stosswirkung (momentum transfer).”

Except for one lone but important dissenter – Niels Henrik David Bohr. He continued to doubt the reality of light quanta, wanted to retain the Maxwellian picture of radiation, and to relegate quantum features exclusively to matter and not to radiation. As part of this line of thinking, in an important paper in 1924, Bohr and his coauthors Hendrik Anton Kramers and John Clarke Slater proposed giving up both causality and energy – momentum conservation in individual elementary processes, but retaining them only statistically. Fortunately these two ideas were experimentally tested right away – by Walther Bothe and Hans Geiger and by Compton and A W Simon respectively – and in both respects Bohr’s proposals failed.

The light quantum idea was here to stay.

8. Bose Statistics – the Photon Spin

It was emphasized earlier that from the very beginning Einstein was conscious of the fact that there was no theoretically well founded derivation of the Planck Law (8). Even his own derivation of 1916 relied on the Bohr theory for matter and interaction processes between matter and radiation. In June 1924 Satyendra Nath Bose working at Dacca University (now Dhaka in Bangladesh) sent Einstein a four page paper containing a novel logically self-contained derivation of the Planck Law, treating thermal radiation as a statistical mechanical system on its own and taking the photon picture to its logical conclusion. Einstein immediately recognised the depth of Bose’s ideas; helped in publishing his paper after translating it into German; and then followed it up with a paper of his own applying Bose’s method to the ideal material quantum gas. The key point in Bose’s method was a new way of counting complexions or microstates for an assembly of photons, in the process giving new meaning to the concept of identity of indistinguishable particles in the quantum world. In contrast to Einstein’s conclusion drawn from the Wien Law that light quanta have a certain mutual independence, Bose statistics shows that photons – because of their identity in the quantum sense – have a tendency to clump or stick together. And basically this difference accounts exactly for the Planck Law and its difference from the Wien limit.

In his paper sent to Einstein, Bose apparently made another radical suggestion – that each photon has an intrinsic angular momentum or helicity of exactly one (quantum) unit, which could be either parallel or antiparallel to its momentum direction. But – revolutionary as he was – Einstein found this suggestion too revolutionary and removed it in the published version of Bose’s paper!
9. Conclusion

Soon after the above events, modern quantum mechanics was discovered during 1925-26; and in 1927 Paul Adrien Maurice Dirac completed the task of quantising the classical Maxwell field, something which Einstein had foreseen as early as in 1917. And with that the photon was here to stay. What better way to end this account than to turn to Einstein himself in his old age:

“All the fifty years of conscious brooding have brought me no closer to the answer to the question, ‘What are light quanta?’ Of course today every rascal thinks he knows the answer, but he is deluding himself”.

Einstein to Michele Angelo Besso, December 1951

**Suggested Reading**


See how the light beam picks up a phase in this interference experiment!

This article brings out a remarkable contribution to the understanding of light, made by S (for Sivaramakrishnan) Pancharatnam, when he was a doctoral student working on the optics of crystals with C V Raman at the Raman Research Institute (RRI) in Bangalore in the 1950s. We will first illustrate this contribution by means of specific examples, which already calls for considerable background, and then go on to place the work in a wider context.

1. Polarisation and Phase

The two main actors in this drama are the phase and polarisation of light. One normally thinks of a light beam in terms of its direction, intensity, and frequency. Given these, the next two fundamental properties of light are ‘phase’ and ‘state of polarisation’. ‘Phase’ describes the cyclic change of electric and magnetic fields during one period of the wave. It is usually expressed as an angle, going from 0° to 360° in one period. One can think of phase like the position of the hour hand of a (conventional) clock, which comes back to the same point after 12 hours. The instant of zero phase is not unique. First of all, the phase is different at points which lie at different distances from the source of the wave. And even when we have chosen one such point, we could start our clock (at 0° phase) when the light field is a maximum, or zero, or anything in between. The phase difference between two beams determines the outcome of interference experiments when they are superposed at a given point, as in the famous Young’s experiment. For example, if the phase difference is 180°, we get darkness, i.e., destructive interference. (Figure 1 illustrates the concept of phase difference).

‘State of polarisation’ (henceforth briefly, ‘polarisation’) describes the shape and orientation of the path traced by the electric field during one period (Figure 2). Since electromagnetic waves are transverse, this orbit is traced in a plane perpendicular to the direction in which the light is travelling. The most familiar case is shown in Figure 2 (a, b), a simple harmonic motion along one line naturally called ‘linear polarization’. The orientation of the line is given by an angle (say with the x-axis) varying from 0 to 180°. Note that two angles differing by 180° really describe the same state of polarisation, since the path traversed is the same. (Where the point marking the tip of the electric field vector is at a given time, is a matter of phase!).

Another standard case, also shown in Figure 2, is when the orbit is a circle, which is...
Figure 1. (a) The electric field vector of linearly polarised monochromatic light, travelling along \( z \), at a given instant of time. The entire pattern moves past any given plane, say \( z = 0 \); so the time variation would be the same curve read in reverse, with the part at positive \( z \) arriving earlier. The magnetic field \( B \) is shown only at one plane, \( z = 0 \), but is equal to the electric field (in cgs units!) but along \( y \).

(b) The time variation of the \( x \)-component of the electric field at a given point in the wave is shown in the curve A. B shows the result of delaying it in phase by 30°, C by 90°, and D by 180°. A and D would give destructive interference.

naturally called ‘circular polarisation’. This can be made by superposing two linearly polarised waves in perpendicular orientations, with a 90° phase difference. The most general case is an elliptical path. How to produce light of a desired polarisation, how to measure the polarisation of a given beam, are subjects which would take us too far afield. However, the simplest example is a so-called ‘polariser’, which only transmits the component of the electric field vector in a particular direction (the perpendicular component is absorbed). This can clearly be used to prepare linearly polarised light. Also, given a linearly polarised beam, one can find its direction by rotating the polariser until the transmitted intensity is a maximum – in this role, the same device is called an ‘anlyser’ (Figure 3). In what follows, we assume that this can be done for other states of polarisation as well, and refer the reader to standard textbooks for the details.

Figure 2. Different possible states of polarization, i.e., paths traced by the electric vector: (a) horizontally polarised, (b) vertically polarised, (c) right circular, (d) left circular, (e) right rotating ellipse with major axis at 45° to the horizontal and (f) left rotating ellipse with major axis at 135° to the horizontal.
Figure 3. A polariser is a device (made from crystal or polymeric material) with the property of absorbing that component of the electric field of any wave which lies along one axis, and transmitting the component along the perpendicular axis. Once polarised light has been produced by this device, a second one, inclined at an angle $\theta$ will only transmit a fraction of the amplitude and hence of the intensity, which will now come out polarised along the preferred axis of the second polariser. In this role, it is called the ‘analyser’.

2. Non-Interference of Polarised Light

The study of the effect of polarisation states of the two beams on interference goes back to Fresnel and Arago, two of the (many!) French pioneers of optics in the 19th century. They carried out Young’s interference experiment with two coherent beams derived from the same source, but with different polarisations emerging from the two slits. We consider the case when the two beams reaching the screen have perpendicular polarisations, say along $x$ and $y$ (Figure 4); (the angle between the beams is very small). Clearly, the path difference between the two beams varies as we move along the screen. But the total intensity does not vary and in that sense one could say there are no interference fringes. To see why the intensity remains constant, independent of phase variation in one or both the beams, let us recollect that the intensity is proportional to the square of the electric field vector, i.e. its dot product with itself. If we write $E = E_1 + E_2$ for the electric field obtained by superposing the two sources, then the resultant intensity is given by $I = E_1 \cdot E_1 + E_2 \cdot E_2 + 2E_1 \cdot E_2$, where the first two terms correspond to the intensities of the individual beams, while the third represents the interference. In our case, this term vanishes because $E_1$ is perpendicular to $E_2$ (and that is why we have not indicated the factors $\cos(\omega t + f)$ representing time dependence). Two such beams are said to be orthogonal. A simple calculation (see Box 1) shows that this property is also true for two opposite circularly polarised beams. For any elliptically polarized beam as well, it is possible to find an orthogonal one, which does not give intensity variation in a Young’s experiment.
Rajaram Nityananda

The beauty and power of the Poincaré sphere, a geometric representation of polarised light, is revealed in Pancharatnam’s work:

Figure 4. The result of superposing two equal linearly polarised waves along the x- and y-axes respectively.
(a) Both components have the time dependence \( \cos(\omega t) \) and hence result in a linearly polarised state at 45° to x.
(b) The y-component lags behind x by 45°. The result is a right-rotating ellipse with axial ratio \( \sqrt{2} - 1 = 0.4142 \).
(c) The lag increases to 90° and we get right circular light (this is the well-known result that circular motion can be made up of two simple harmonic motions with a 90° phase difference).
(d) Right-rotating ellipse with the same axial ratio as in (b) but now with the major axis at 135°.
(e) Linear polarization at 135°, produced when the phase lag is 180°. The further evolution as the lag goes via 225°, 270°, and 315° is omitted, since it gives left rotating versions of (b), (c), and (d).

We hasten to add that varying the phase of one of the orthogonal pair of waves while keeping the other fixed, certainly has consequences for the polarisation of the resulting wave. This is illustrated in Figure 4 for the case of two orthogonal linear polarisations of equal intensity. As we vary the phase of one of them, from 0 to 360°, we go all the way from a linear state at 45°, to a circular state, to a linear state at 135°, and back via the opposite circular state. In this sense, interference does occur. In fact, by placing an analyser at 45°, bisecting the x and y directions, one can now get variations in intensity. Clearly the intensity will be a maximum in the case when we have a linear state at 45°, and zero when the combination gives a linear state at 135°. In fact, the use of an analyser is the standard way of revealing interference between two orthogonal beams, because it brings them to a common state of polarisation which can then interfere.

3. The Poincaré Sphere

This account would be grossly incomplete without introducing the Poincaré sphere, a geometric representation of polarised light whose beauty and power is revealed more
A less obvious case of orthogonality occurs when two circularly polarised beams, one of which is right rotating and the other left rotating, are superposed. We can now have interference between the $x$-components of the two beams, and also the $y$-components.

Beam 1: $E_x = a \cos(\omega t)$, $E_y = a \sin(\omega t)$;

Beam 2: $E_x = b \cos(\omega t + \phi)$, $E_y = -b \sin(\omega t + \phi)$.

Note the minus sign in the second beam, which makes it left rotating; the $y$-component of the electric field attains its maximum value a quarter period after the $x$-component, for Beam 1, but a quarter period after the $x$-component, for Beam 2. We now have to add the squares of the two Cartesian components of the electric field to get the intensity, which is proportional to $E_x^2 + E_y^2$. The interference term is the one proportional to the product of the two amplitudes of Beam 1 and Beam 2, i.e., to $ab$. When we do this calculation, the minus sign (the opposite circular polarisations of Beam 1 and Beam 2) is crucial to give us $2ab \cos(2\omega t + \phi)$ which averages to zero in one cycle of the waves. The most general example of two orthogonal polarisations—i.e., two superposed polarised waves not giving any intensity changes as the path difference is varied—is not much more difficult. Two elliptic waves, with the same axial ratio, but rotating in opposite senses, and with their major axes perpendicular to each other are orthogonal, as can be checked by a similar calculation to the one done above for opposite circular waves.

in Pancharatnam’s work than that of any other author I know of. Figure 5 shows how the different possible states of polarised light are mapped onto a sphere. The orientation of the major axis is first doubled (because it goes only from 0 to $180^\circ$ !) and becomes the longitude on the sphere. The equator represents linearly polarised states. Note that with the doubling of the orientation angle, two orthogonal linear polarisations, say horizontal and vertical, lie at opposite ends of the equator. The northern hemisphere is used for right rotating ellipses, with the axis ratio increasing from zero (at the equator) to unity at the pole. The longitude now corresponds to the orientation of the major axis. The states corresponding to different orientations of the major axis merge when the ellipses become right circular, at the north pole (Figure 5). The precise rule for mapping latitude to axis ratio is $\text{lat} = 2 \arctan(b/a)$ which clearly puts the right circular polarisations, $b/a = +1$ at the north pole. The southern hemisphere is used for left rotating elliptical polarisation states, in a similar manner. Conventionally, the minor axis $b$ is taken to be negative for this set of states. Left circular polarisation corresponds to $b/a = -1$ and hence a latitude of $-90^\circ$, i.e. the south pole (Figure 5).

Henri Poincaré (the issue of Resonance, Vol.5, No.2, 2000 was devoted to him) was a towering figure of the late nineteenth century, when mathematics had not fully parted
Figure 5. (a) The Poincaré sphere, showing the same sequence of states as in Figure 4. The thin lettering and dashed line indicate that we are viewing the back of the sphere. H and V represent linearly polarised states which are at 0° and 90° to the z-axis.

(b) The states of polarisation near the north pole of the sphere. Treating this as flat, the orbits of the electric field have been ‘painted onto’ the sphere, to bring out the point that ellipses with very different major axis orientations merge into a single circular state.

(c) The full set of orbits of the electric vector on the Poincaré sphere, spread out into a plane as with the Mercator projection of geography. As with maps of the earth, all the points at 90° latitude are just one point!

from physics. His restless energy led him to lecture on light for two consecutive terms, and the sphere is just a passing remark in his discussion of polarisation! Poincaré was one of the founding fathers of topology, the branch of mathematics which captures the essence of the notion of continuity and neighbourhood. To a physicist, adding a small perturbation to a polarised wave would be regarded as taking the state to a nearby one, with an orbit differing only slightly from the earlier one. Poincaré’s mapping of the different possible states of polarisation onto a sphere ensures that any two states of polarisation that a physicist would regard as differing only slightly from each other would be placed close to each other, and at the same time, any two nearby points on the sphere would only represent states which are close in a physical sense. The sphere is more than just a faithful representation of the topology of polarisation states, it has many useful and beautiful properties.
4. Through a Crystal, Darkly: the Pancharatnam Phase

Without going through the subtleties of crystal optics, we state one standard property. A fully transparent crystal – one which absorbs no energy at all at the wavelength of interest – will split incident light into two orthogonal polarizations which travel with different speeds, and hence develop a phase difference. But as we have seen earlier, this phase difference does not manifest as interference in intensity when the modes are recombined – hence the time-honoured technique of creating intensity patterns by bringing the two to the same polarisation state by means of an analyser.

Many crystals, and certainly the beautiful ones which were in Raman’s collection (Figure 6), show colours, i.e., absorption which depends on wavelength. This also depends on polarisation, and in fact is the subject of Pancharatnam’s first few papers (written when he was twenty). Some absorbing crystals have a property which singles them out from transparent ones – incident light is split into two modes which are no longer orthogonal, and can hence give rise to interference effects even without an analyser (or indeed a polariser, since they can act as their own polarisers!). Pancharatnam was impelled to study the interference of non-orthogonal beams by experiments, including his own, on a mineral called ‘iolite’. The ‘Classics’ section of the issue of Resonance, Vol. 18, No. 4, pp 387–389 (2013), reproduces the introductory passages of Pancharatnam’s paper, and pictures of the phenomena he observed and sought to understand. This paper clearly shows his depth and facility with theory, but also shows that it was accompanied by motivation firmly rooted in experiment, characteristic of the Raman school.

For simplicity, we first consider (as Pancharatnam did, in his paper) one basic thought experiment – interference between two beams A and B in different, and importantly, Pancharatnam was impelled to study the interference of non-orthogonal beams by experiments, including his own, on a mineral called ‘iolite’.
Figure 7. Illustrating the special property of the Pancharatnam definition of phase difference between non-orthogonal polarizations.

(a) Three linearly polarised states A, B, C. The dots are chosen as described in the text – B is in phase with A (i.e., would give maximum intensity on superposition) and C is in phase with B. But to get maximum intensity on superposing C with A, one would have to choose the point indicated by the open circle, on the orbit of A. So staring with A, keeping the phase difference, as defined by Pancharatnam, zero at every step, and returning to A, we have a discrepancy of 180°, or π radians, which is half the solid angle of the ‘triangle’ ABC of the Poincaré sphere in Figure 7(c).

(b) A and B same as before, but C is now right circular light. The dots are chosen so that B is in phase with A, and C with B. The open circle shows the phase which C would need to have to be in phase with A. The discrepancy is now 60° of phase, or π/3 radians, half the solid angle subtended by the triangle ABC on the Poincaré sphere of Figure 7(d).

not orthogonal states of polarisation. By adjusting the phase of B with respect to A, one could ensure that there is constructive interference, i.e., maximum intensity of the resulting beam. For linear polarisations, this is easy to understand (refer to Figure 7 for what follows). We divide B into a component orthogonal to A and one along A. By the Fresnel–Arago principle the orthogonal component does not contribute to the intensity variation. This comes entirely from the interference of the component of B parallel to A, with A (Figure 7). We get maximum intensity when this component is in phase with A. The filled circle at the tip of the linearly polarised state A represents the electric field at a given moment. The point on the state B which would be regarded as in phase with A, is also indicated by a filled circle, and it is clear that this phase
would result in maximum intensity since the component of B along A would be in phase with A.

Taking this situation as the zero of the phase difference, one finds that if the phase of one of the beams is then varied by an angle $\delta$, the intensity varies as the cosine of this angle. Let us appreciate the care and logic with which Pancharatnam proceeds from this point: “There are two properties of $\delta$, however, which enable us to speak of it as the absolute difference of phase between the two beams themselves. In the first place, we note that if the first beam is subjected to a particular path retardation relative to the second, $\delta$ as defined above decreases by the corresponding phase angle: in the second place we note that as long as no path retardation is introduced between the two beams, any alteration of the intensities of the two beams will not change the value of $\delta$, as defined above. Hence we will be guilty of no internal inconsistency if we make the following statement by way of a definition: The phase advance of one polarised beam over another [not necessarily in the same state of polarisation] is the amount by which its phase must be retarded relative to the second, in order that the intensity resulting from their mutual interference may be a maximum.” Here is a young physicist, capable of skilled experimental work, but also laying emphasis on avoiding internal inconsistency in a manner more characteristic of the axiomatic spirit of the twentieth century mathematicians!

Having adjusted B to be in phase with A, we now bring in a third state C, and adjust it to be in phase with B. The filled circle at the tip of the double arrow which represents the polarisation C represents this situation, viz., B being in phase with C. What could then be more natural than C being in phase with A? But Pancharatnam, having given a definition, sticks to it and evaluates the phase difference between C and A, and lo and behold, it is not zero! He must have himself been surprised, since his normally sober style gives way to some excitement, reflected in two extra adjectives: “The beams 1 and 2 will have a definite phase relationship with one another, which, as we shall see, depends in a remarkable fashion on the mutual configuration of the points A, B, and C” and later “… we arrive at the following unexpected geometrical result”. (Italics are mine – RN).

Since this runs counter to one’s intuition, a detailed explanation is given in Figure 7 and its caption, for two cases. In the first, the three states are linear, with B making an angle of 60°, counterclockwise, with A, and C located 60° counterclockwise from B. In this case, we find that while the phase difference, evaluated by Pancharatnam’s prescription is zero for A with respect to B and zero for B with respect to C, it is 180° for C with respect to A. The second case is when A and B are the same as earlier, but C is a right circular state. The phase discrepancy is now seen to be 60°. These are just two elementary examples chosen for illustrative purposes. Pancharatnam himself worked out the most general case and proved that the phase equals half the solid angle made by the spherical triangle ABC at the centre of the Poincaré sphere – a beautiful result, but we will not be able to elaborate on in this article.
5. Thirty Years After

Pancharatnam’s paper was published in 1956, the first of a series called the ‘Generalised Theory of Interference’. He later went on to do outstanding work in other, more modern areas of optics (see Article-in-a-Box) but sadly, died at the age of 35 in 1969, in Oxford, UK. The phase he discovered lay dormant for nearly thirty years. In 1983, Michael Berry of the University of Bristol uncovered a remarkable property of phases in quantum systems which had been missed (with a few honourable exceptions!) over the nearly sixty years that quantum mechanics had been studied by hundreds of physicists. Quantum systems can undergo ‘adiabatic evolution’, in which some external parameters (say magnetic fields acting on spins) are slowly changed and the state also changes appropriately. When the parameters return to their original values, so does the quantum state. But there is an extra term in its phase (in addition to the well-known ones) which Berry evaluated and illustrated with many examples.

This clearly reminds us of the phase discrepancy encountered while trying to ‘synchronize’ phases around the journey, A to B to C and back to A. It was realised a few years after Berry’s work that Pancharatnam’s phase, although proposed in a classical situation, was one of the early examples (in fact going beyond adiabatic changes) of this general phenomenon which is now known as ‘geometric phase’ or ‘Berry phase’ – now a flourishing industry in the world of theoretical physics. A reference for later developments in optics is the special issue of *Current Science* brought out in August 1994 [1]. We quote an extract from Berry’s own contribution to this issue which starts with describing his own feelings on reading the collected papers of Pancharatnam on his way back from a trip to Bangalore:

“...The flight was a revelation. I learned that not only had this young fellow of twenty two created the simplest example of the geometric phase, but that he had also pointed out a feature (the definition of phase difference described below) that had not, by 1987, been perceived in any of the many papers developing my work of 1983... To understand interference, it is necessary to know not the absolute phase of a wave but the phase difference between light beams in different states of polarisation. Apparently, nobody had asked this simple question. Pancharatnam did ask it, and gave a simple answer... Ignorant of Pancharatnam’s work, I had also calculated geometric phases for the case he studied, that is, a light beam whose polarisation is changed... I was astonished (and not a little humbled) that Pancharanam had had essentially the same idea (with discrete, rather than continuous polarization changes) thirty years earlier”.

6. Changing While Trying to Remain the Same

The Pancharatnam phase is just one example of a general mathematical phenomenon which is mirrored in different physical contexts. Stated in common language, it seems that one does one’s best to see that something (the phase of a polarised beam in this case) does not change, but when one comes back to the starting point after a circuitous
journey, it has changed. One example is from a story by Jules Verne. The eccentric Englishman, Phileas Fogg travels around the world, keeping careful track of days since he has staked his entire fortune on a bet to finish the journey in eighty days. Since he arrives a few minutes too late after seeing eighty sunsets, he assumes he has lost the bet. But his resourceful companion, Passepartout, (naturally a Frenchman in Verne’s novel!) finds out that in London, only seventy nine days have passed, and Fogg is able to win his wager after all, and marry his Indian bride as well. This example is actually real, since it happened to the crew of the ship which sailed around the world for the first time. This problem is solved today by introducing a discontinuity, the international date line.

The next example, again from geometry, is the notion of parallelism on a curved surface. Anyone on Earth would have no difficulty in walking in a straight line and carrying a pole, parallel to the ground but without turning it. After all, any small piece of a sphere is like a plane and we all know what is meant by a straight line, and parallelism, on a plane. But see what happens when we try to keep the direction of a long rod fixed, while taking a long journey from the north pole to the equator, along the equator for 90° of longitude, and back (around one eighth of the world in one day!). (See Figure 8.) The rod is rotated by 90° from its original position, and, more generally, by an angle proportional to the solid angle made by the journey at the centre of the sphere. A related fact was known to nineteenth century geometers – Lobachevsky, Bolyai, Gauss, Riemann – that the sum of the angles of a triangle is no longer 180° on a curved surface. Regarded as a failure of the notion of parallelism, when one makes large excursions, this concept was put forward by Levi-Civita and Ricci at the dawn of the twentieth century. (This topic is covered in Resonance (Vol.1, No.5, p.6; No.6, p.33, 1996) and is fundamental to Einstein’s theory of gravitation).

Mathematicians have a term for this general phenomenon – viz ‘holonomy’ of a connection. Physicists view things differently – Berry proposes anholonomy as more
appropriate, the negative prefix presumably conveying the idea that something we did not expect, happened! Anholonomies are everywhere. Here are a few, just to whet the reader’s appetite. The spin of the electron in the hydrogen atom turns by a small angle when it is taken for an accelerated ride around the proton, slipping smoothly from one frame of reference to another (‘Thomas precession’). The linear polarisation of light in an optical fibre coiled in a helix does not return to its original orientation, even when the fibre itself does. A bacterium goes through a cyclic change of shape and achieves a shift in its position, i.e., it swims by anholonomy. And the cat uses anholonomy to land on its feet, i.e., change its orientation, even when not given any angular momentum; (skilled divers do the same).

7. Concluding Remarks

This article will serve its purpose if it raises awareness among readers, of optics, of the general phenomenon of anholonomy, and of the very bright, very young man who, more than half a century ago and just a few hundred metres from where Resonance is put together, looked into a coloured crystal and saw what others had not seen before.

Acknowledgement

While retaining responsibility for the remaining flaws, I thank N Andal of RRI and the staff of Resonance for detailed corrections and important improvements to this article.

Suggested Reading

The EPR Paradox *
Einstein Scrutinises Quantum Mechanics

Arvind

Einstein, Podolsky and Rosen (EPR), in their famous paper, argued that the quantum-mechanical description of physical reality is incomplete. They showed that one can envisage physical situations whereby ‘an element of physical reality’ can be located such that it does not have a counterpart in quantum theory. A simple description of the EPR paradox is presented in this article, using the example of two spin-1/2 particles.

Einstein contributed key ideas to the early development of quantum theory. However, he did not think that quantum mechanics was the final answer to the question of theoretically mapping objective physical reality. His tersely worded, yet extremely lucidly written, criticism of quantum theory is formulated in the EPR paper. We intend to give a simple description of the EPR objections to quantum theory.

Objective reality exists independent of any theory. Hence, any theory that seeks to describe reality has to operate with concepts which have a correspondence with this reality. Within this framework we can demand that, for any theory to be satisfactory, the following criteria must be fulfilled: (a) the theory must be correct, i.e. its predictions should be in agreement with experimental results and (b) the theory must be complete, i.e. “every element of physical reality must have a counterpart in the theory”. The EPR paper shows that there exist elements of reality which do not have any counterpart in the quantum theory and hence, while quantum theory may be correct, it is incomplete.

Before proceeding further, we need to be able to define elements of reality for a given physical situation. This is a tricky question and can lead to long philosophical arguments! Instead of trying to identify a complete set of elements of reality for a physical system, EPR suggested a means of identifying an element of physical reality: “If without in any way disturbing a system we can predict with certainty (i.e. probability equal to unity) the value of a physical quantity, then there exists an element of physical reality corresponding to this physical quantity”.

This prescription of identifying elements of physical reality can be applied to classical as well as quantum systems. The goal here is to use the above criterion and to come up with at least one element of physical reality which does not have a corresponding counterpart in quantum theory. If this can be achieved, then one has shown that quantum theory is incomplete.

1. Elements of Physical Reality

The fact that quantum theory predicts only the probabilities and not the outcomes of individual measurements, and that the state of the system is altered by the measuring procedure are two key features of quantum measurement theory. What happens if the state of the system is one of the eigenstates of the operator being measured? For an eigenstate $|\alpha\rangle$ of the hermitian operator $\hat{A}$, only the eigenvalue $a$ will be obtained after the measurement, with a probability equal to unity. The state too will not suffer any change after the measurement process. We can thus predict the outcome of the measurement with a probability equal to one and without disturbing the system in any way. Therefore, by the EPR criterion, for an eigenstate, the value of the corresponding observable is an element of physical reality.

2. Example of a Two-level System

Consider a spin-$\frac{1}{2}$ particle as an example of a two-level system, the two levels being characterised by the spin being in the $|\uparrow\rangle$ or the $|\downarrow\rangle$ state. The three spin components of this system are represented by the hermitian operators

$$S_x = \frac{\hbar}{2}\sigma_x, \quad S_y = \frac{\hbar}{2}\sigma_y, \quad S_z = \frac{\hbar}{2}\sigma_z,$$

with $\sigma_x, \sigma_y$ and $\sigma_z$ being the standard Pauli spin matrices. These operators are the only three independent observables in this system.

Consider the eigenstates of $S_z$: $|\uparrow\rangle_z$ and $|\downarrow\rangle_z$ with eigenvalues $+\frac{\hbar}{2}$ and $-\frac{\hbar}{2}$, respectively. The value of the observable corresponding to the $z$ component of the spin is an element of reality for these states. Further, we can construct states for which $S_x$ and $S_y$ are the elements of physical reality, which are just the eigenstates of these operators, respectively. For a general state of the spin-$\frac{1}{2}$ system,

$$|\alpha\rangle = \cos\left(\frac{\theta}{2}\right)|\uparrow\rangle_z + \sin\left(\frac{\theta}{2}\right)e^{i\phi}|\downarrow\rangle_z$$

the component of spin along $(\theta, \phi)$ is an element of physical reality.

3. Simultaneous Reality of Two Observables

Let us now consider two observables $A$ and $B$ for a quantum system.

Case 1: When $A$ and $B$ commute, i.e. when

$$[A, B] = 0,$$

the two observables have simultaneous eigenstates and hence, for these eigenstates, their values can be simultaneous elements of reality. We can predict their values with a probability equal to one, and can also measure them without disturbing the state of the system.
**Case 2:** When \( A \) and \( B \) do not commute, i.e. when

\[ [A, B] \neq 0, \]

the two observables \( A \) and \( B \) do not have simultaneous eigenstates. Each eigenstate of one of the observables will, in general, be a linear combination of the eigenstates of the other. For no state does quantum theory predict precise values of these two observables together. Thus it is expected that no experiment will be able to determine the values of these observables simultaneously. Repeated observations will in general lead to different values for at least one of the observables. Let us take another look at the example of the spin-1/2 particle. We will see here that as a direct consequence of the non-commutation of \( S_z \) and \( S_x \) \([S_z, S_x] = i\hbar S_y\), there is no quantum state for which both \( S_z \) and \( S_x \) are simultaneously elements of physical reality.

If we measure the \( x \) component of the spin for either of the eigenstates of \( S_z \) (for which \( S_z \) is an element of physical reality), both the values \( \pm \frac{\hbar}{2} \) are equally probable and the state after measurement changes to the corresponding eigenstate of the \( S_x \) operator. This can be seen directly if we expand, say, the state \( \ket{\uparrow}_z \) in terms of the eigenstates of \( S_x \)

\[ |\uparrow\rangle_z = \frac{1}{\sqrt{2}} \{ |\uparrow\rangle_x + |\downarrow\rangle_x \} \]

and apply the quantum measurement theory given in Box 1. Thus, we are neither able to predict values for \( S_x \) with probability one, nor able to measure it without disturbing the particle. Therefore, quantum mechanics does not predict \( S_x \) to be an element of physical reality simultaneously with \( S_z \).

4. **EPR Argument for Two Spin-1/2 Particles**

We now apply the above arguments to a composite system consisting of two spin-1/2 particles \( P_1 \) and \( P_2 \). In this case, the four eigenstates are given by

- \( |\uparrow\rangle_{1z} |\uparrow\rangle_{2z} : P_1 \) in state \( |\uparrow\rangle_z \) and \( P_2 \) in state \( |\uparrow\rangle_z \),
- \( |\uparrow\rangle_{1z} |\downarrow\rangle_{2z} : P_1 \) in state \( |\uparrow\rangle_z \) and \( P_2 \) in state \( |\downarrow\rangle_z \),
- \( |\downarrow\rangle_{1z} |\uparrow\rangle_{2z} : P_1 \) in state \( |\downarrow\rangle_z \) and \( P_2 \) in state \( |\uparrow\rangle_z \),
- \( |\downarrow\rangle_{1z} |\downarrow\rangle_{2z} : P_1 \) in state \( |\downarrow\rangle_z \) and \( P_2 \) in state \( |\downarrow\rangle_z \).

All these states are eigenstates of the operators \( S^1_z \) and \( S^2_z \). (The superscripts 1 and 2 refer to \( P_1 \) and \( P_2 \), respectively.) This implies that, for these states, the \( z \) component of the spin of each particle corresponds to an element of physical reality.

Now, any normalised linear combination of these eigenstates is also a valid quantum state (using the **superposition principle** of quantum mechanics). Specifically, let us consider a state which is formed by taking a linear combination of two of these states, with equal weights and a phase difference of \( \pi \)

\[ |\psi\rangle_{12} = \frac{1}{\sqrt{2}} \{ |\uparrow\rangle_{1z} |\downarrow\rangle_{2z} - |\downarrow\rangle_{1z} |\uparrow\rangle_{2z} \}. \]
Box 1. Measurement in Quantum Theory

Consider a quantum system in the quantum-mechanical state $|\alpha\rangle$. Assume that we have set up an apparatus to measure the physical observable $A$. The first task is to identify the hermitian operator $\hat{A}$, corresponding to the observable $A$. Proceeding step-wise in order to interpret the results of quantum measurements on the observable $A$:

- One has to find the eigenvalues and eigenvectors of the operator $\hat{A}$,

$$\hat{A}|a\rangle = a|a\rangle.$$ 

These states form a complete set, so that every quantum state can be expanded as a linear combination of them.

- The possible outcomes of the experiment are the set of eigenvalues $\{a\}$. In an individual experiment, only one of the eigenvalues $a \in \{a\}$ is obtained.

- After every measurement, the state of the system is transformed into an eigenstate of the operator $\hat{A}$, corresponding to the eigenvalue obtained.

- If we repeat the experiment on the state $|\alpha\rangle$ a large number of times, we can obtain the relative frequency of different eigenvalues.

- To calculate these relative frequencies from quantum theory, we expand the state $|\alpha\rangle$ in terms of the eigenfunctions of the operator $\hat{A}$,

$$|\alpha\rangle = \sum_a C_a |a\rangle$$

giving us the probability of finding the eigenvalue $a$ as

$$P_a = |\langle a|\alpha\rangle|^2 = |C_a|^2.$$ 

This state is rotationally symmetric and will therefore retain the same form if we make a change of basis. In fact, it is the well-known singlet state, with zero total angular momentum. Re-writing $|\psi\rangle_{12}$ in the $x$ basis,

$$|\psi\rangle_{12} = \frac{1}{\sqrt{2}} \{ |\uparrow\rangle_{1x} |\downarrow\rangle_{2x} - |\downarrow\rangle_{1x} |\uparrow\rangle_{2x} \}.$$ 

Here

$$|\uparrow\rangle_{jx} = \frac{1}{\sqrt{2}} \{ |\uparrow\rangle_{jz} + |\downarrow\rangle_{jz} \}, \quad |\downarrow\rangle_{jx} = \frac{1}{\sqrt{2}} \{ |\uparrow\rangle_{jz} - |\downarrow\rangle_{jz} \}$$

are the eigenstates of the operator $S^x_j$ with $j = 1, 2$. Once the two particles are prepared in such a state, they remain so even after they cease to interact as total angular momentum is conserved.
We can now perform different types of measurements on the particles. Let us think of an experimental set-up which measures $S_z^1$ for the particle $P_1$. There are only two possibilities: either we get the value $+\hbar/2$ and the state of the system changes to the corresponding eigenstate $|\uparrow\rangle_{1z}|\downarrow\rangle_{2z}$ or we get the value $-\hbar/2$ and the state changes to the eigenstate $|\downarrow\rangle_{1z}|\uparrow\rangle_{2z}$. Both these situations are equally probable.

In each case the second particle $P_2$, after the measurement on the first particle $P_1$, is forced into an eigenstate of the operator $S_z^2$. Therefore, if we make a measurement of $S_z^2$ we are bound to get a value which is predictable a priori and is $-\hbar/2$ for the first case and $+\hbar/2$ for the second case.

The particles $P_1$ and $P_2$ are non-interacting and can be very far away from each other (one could be on the moon and the other on the earth!). This ensures that the measurement on particle $P_1$, which allows us to predict the value of $S_z^2$, does not disturb the particle $P_2$ in any way. We thus conclude that $S_z^2$ is an element of reality for the state $|\psi\rangle_{12}$.

This prediction can be verified by actually measuring the value of $S_z^2$.

We have seen that the state $|\psi\rangle_{12}$ has the same form when written in terms of the eigenstates of $S_z^1$. Using identical arguments, we can conclude that $S_z^2$ is also an element of physical reality. As a matter of fact, the rotational symmetry of the state $|\psi\rangle_{12}$ allows us to establish that all spin components of particle 2 are elements of reality for this state.

Quantum mechanics however does not predict two-spin components as being elements of reality simultaneously, because of the non-commutation of different spin components. Therefore, these elements of reality we have identified do not have a correspondence in quantum theory and quantum theory seems incomplete. This is in essence, the main thrust of the EPR argument and the paradox immediately suggests the scope for a more complete theory. The original EPR paper considers positions and momenta of two particles and establishes the simultaneous reality of the position and momentum of one of them. The argument was later modified by David Bohm using spin components.

This piquant situation can be summarised in the following way: Let us assume that pairs of particles are being produced in the above-mentioned singlet state. From every pair produced, one particle comes to my lab (on earth) and the other goes to my
The EPR paper finds its origins in the discussions of Bohr and Einstein, on the foundations of quantum theory, which began in the 1920s and reached their pinnacle during the fifth and sixth Solvay Conferences. The Bohr–Einstein dialogue is one of the great scientific debates in the history of physics and ensued mostly in the unofficial sessions of the Solvay Conferences. During the latter half of 1933 Einstein joined his new position in Princeton and brought Walter Mayer from Berlin to work with him. However, Mayer soon obtained an independent position and Einstein began looking for the assistance of young scientists. Boris Podolsky, who had already worked with Einstein and Tolman in 1931 on issues dealing with the Uncertainty Principle and the distinction between the past and the present in quantum mechanics, had left shortly afterward to work with Vladimir A Fock (and Lev D Landau) on quantum electrodynamics. In 1933, he returned on a fellowship to the Princeton Institute and Einstein became interested in him. In 1934 Nathan Rosen, who had obtained his PhD in atomic physics from MIT under J C Slater, began to work at Princeton University. Venturing one day to enter Einstein’s office, Rosen was surprised by the friendliness with which Einstein inquired about his work. When on the following day he met Einstein in the yard of the Institute, Einstein said to him: “Young man, what about working together with me?” Shortly thereafter Rosen became a research fellow at Einstein’s department. This then is the story of how Podolsky and Rosen joined Einstein.

5. Bohr’s Reply and Later Developments

Soon after the EPR paper (Box 2), Niels Bohr wrote a paper as a rejoinder to the EPR argument, in which he detailed the Copenhagen interpretation of quantum mechanics. He argued that the EPR paradox does not prove the incompleteness of quantum theory and that the statement “in no way disturbing the system” has to be interpreted differently in quantum mechanics. A measurement on particle $P_1$ changes the very circumstances in which the later measurements on particle $P_2$ are being conceived. Therefore one cannot say that the values of spin components can be predicted without disturbing the particle. Though Bohr’s rejoinder came within a few months of the EPR paper, it was not widely accepted and the controversy still continues! The EPR paradox has generated much debate and a large number of research papers.
People continue to work on the resolution of this paradox. In the beginning, it was felt that one ought to develop a theory which is complete, and such theories called hidden variable theories, were attempted. The most important development in this context has been through the work of John Bell. He investigated the possibility of constructing hidden variable theories which respect locality i.e. do not allow instantaneous action-at-a-distance. He found that such theories are inconsistent with the statistical results of quantum mechanics. It is now very clear that the quantum world is indeed very strange and is beyond our naive intuition in more ways than one.

Suggested Reading

In the 5th century BC, the philosopher and logician Zeno of Elea posed several paradoxes which remained unresolved for over two thousand five hundred years. In recent times, the Zeno effect made an intriguing appearance in a rather unlikely place – a situation involving the time evolution of a quantum system which is subject to ‘observations’ over a period of time. In 1977, B Misra and E C G Sudarshan published a paper on the quantum Zeno effect, called “The Zeno’s paradox in quantum theory”

1. Introduction

The quantum Zeno effect gets its name from the Greek philosopher Zeno who lived in the 5th century BC in the Greek colony of Elea (now in southern Italy). Zeno was known to be the most brilliant disciple of Parmenides, a very prominent figure of the Eleatic School of philosophers. Aristotle is supposed to have credited Zeno with having invented the method of the ‘dialectic’, where two speakers alternately attack and defend a thesis. Zeno is also credited with inventing the argument form ‘reductio ad absurdum’. However, Zeno is most famous for his four paradoxes of motion which he developed as arguments against the then existing notions of space and time. A description of the quantum Zeno effect does not necessarily require a prior knowledge of the classical paradoxes of Zeno. However, as an interesting historical reference, we will briefly describe particular versions of the four classical paradoxes before embarking on a description of the quantum Zeno effect itself. The paradoxes popularly are called: (i) Achilles and the tortoise, (ii) the arrow, (iii) the dichotomy, and (iv) the sophisticated stadium. We will briefly describe them here.

1.1 Achilles and the Tortoise

Achilles has to race a tortoise. Since Achilles is obviously faster than his opponent, the tortoise is given a head start. Let us suppose that the race begins at } \( t = 0 \), when Achilles is located at } \( x = 0 \), and the tortoise, with its head start stands at } \( x = x_0 \) (Figure 1a). Zeno argued that the running Achilles could never catch up with the tortoise because he must first reach where the tortoise started. For instance, when

---

Keywords

Zeno paradox, unstable systems, survival probability, quantum evolution, quantum measurements, continuous measurements, quantum Zeno effect, quantum anti-Zeno effect.
Aristotle is supposed to have credited Zeno with having invented the method of the ‘dialectic’, where two speakers alternately attack and defend a thesis.

Achilles reaches $x_0$ at time $t_0$, the tortoise has already progressed to $x_1$ ($> x_0$). By the time Achilles reaches $x_1$, at time $t_1$, the tortoise has crawled up to $x_2$. Thus, whenever Achilles reaches $x_i$, the tortoise would have moved to $x_{i+1}$, and Achilles would need time $t_{i+1} - t_i$ before he gets to that point (by which time the tortoise would have moved on ahead!). Zeno argued that for Achilles to reach the tortoise (or to overtake it), he must perform an infinite sum of such time increments, $t_{i+1} - t_i$, or spatial increments, $x_{i+1} - x_i$ for all $i$ up to $\infty$. If space and time are considered to be continuous, these increments would tend to zero duration, and if they are considered discrete, the increment would be of finite duration. (Note that in the continuous case, the increments tend to zero but will not be actually zero for any finite $i$). Zeno argued that if space and time are continuous, an infinite sum of elements tending towards zero length (duration) must have a total of zero length (duration). Alternatively, if space and time are discrete, then an infinite sum of finite elements must be of infinite length (duration). Since Achilles is ‘seen’ to overtake the tortoise, the above arguments fail. Thus, the seemingly absurd conclusion that follows is that both space and time can neither be continuous nor discrete. This compels us to consider the notion that space and time are illusory – hence, so is everything we see!

1.2 The Arrow

Imagine an arrow flying through space (Figure 1b). Time is considered to be made up of ‘instants’. These instants are defined as the smallest measure and they are indivisible. At any instant of time, when the arrow is observed, it must be seen to occupy a space equal to itself. If it is ‘seen’ to ‘move’ at any instant, it means that
the observer can divide an instant into a time when the arrow was ‘here’ and a time when the arrow was ‘there’. This would mean that the instant of time consists of parts, which violates its basic definition (that of it being indivisible). Thus, Zeno asserted that there are no instants of time when the arrow does move. This, again, leads to the conclusion that the arrow is always at rest and that all motion is illusory.

1.3 The Dichotomy

Imagine that Achilles wants to run from point A to point B (Figure 1c). Before he can cover half the distance to the end, he must cover the first quarter. Before this, he must cover the first eighth, and before that the first sixteenth, and so on. Before Achilles can cover any distance at all he must cover an infinite number of smaller spatial separations. Zeno argued that covering these infinite numbers of small spatial separations in a finite time would be impossible. Thus, it can be concluded that Achilles can never get started. But Achilles is ‘seen’ to move. Thus, once again, the argument compels us to conclude that all motion is an illusion. This paradoxical argument is called ‘the dichotomy’ because it involves repeatedly splitting a distance into two parts. It contains some of the same elements as the Achilles and the Tortoise paradox, but with a more apparent conclusion of motionlessness.

1.4 The Sophisticated Stadium

This is the last of Zeno’s four paradoxes. Let us suppose that space and time are discrete in nature and that motion consists in occupying different spatial points at different times. For simplicity, consider objects moving at a minimal speed, \( v \), of one fundamental spatial distance per fundamental temporal duration. Consider nine persons moving roughly collinearly (Figure 1d). Persons \( a_i \) are all stationary, while persons \( b_i \) move past them to the left, with the velocity, \(-v\). At the same time, persons \( c_i \) move to the right past the \( a_i \)s with velocity \( v \). Say, at time \( t = 0 \) the configuration is as shown in Figure 1d(i). One fundamental unit of time later, the configuration in Figure 1d(ii) is achieved. One can see that \( c_3 \) has passed by \( b_2 \), but there was never an instance when \( c_3 \) was lined up with \( b_2 \). Thus, it can be argued that there is no time at which the actual passing occurs, and hence, it never happened! Another way to bring attention to this paradoxical situation is to note that in Figure 1d(ii), each B has passed twice as many Cs as As. Thus, one might conclude that it takes row B twice as long to pass row A as it does to pass row C. However, the time for rows B and C to reach the position of row A is the same. So, it appears that half the time is equal to twice the time!

The four paradoxes, summarized above in their barest forms, foxed and confused mathematicians and philosophers for over two millenia. It was only after the development of the calculus of infinitesimals by Leibniz and Newton, the concept of functions, limits, continuity, infinite series, and convergence, that a satisfactory resolution to Zeno’s paradoxes came about. However, even today, in spite of our familiarity with

Zeno argued that infinite numbers of small spatial separations in a finite time would be impossible.

It was only after the development of the calculus of infinitesimals by Leibniz and Newton, the concept of functions, limits, continuity, infinite series, and convergence, that a satisfactory resolution to Zeno’s paradoxes came about.
these modern mathematical ideas and concepts, there is a continuing debate about the validity of Zeno’s paradoxes and their various resolutions. We end our brief introduction to the ‘classical’ paradoxes of Zeno at this point and will describe the Quantum Zeno effect – the main subject of this article, in the next section.

2. The Quantum Zeno Effect

The Zeno paradox in quantum systems was brought into focus in the 1960s by Leonid A Khalof, working in the former USSR, and by E C G Sudarshan and Baidyanath Misra, working in USA during the 1970s. Misra and Sudarshan’s paper entitled “The Zeno’s paradox in Quantum Theory” in the Journal of Mathematical Physics first introduced the name “Zeno’s paradox” for the effect studied. Quantum Zeno Effect (QZE) is a term more commonly used these days to describe similar situations in various quantum systems. In order to understand the essence of the QZE, it is necessary to first briefly review some basic concepts of quantum mechanics and quantum measurement.

2.1 Quantum Measurement

Quantum mechanics is currently accepted as the most elegant and satisfying description of phenomena at the atomic scale. Since our larger, familiar ‘macro’ world is eventually composed of elements of the ‘micro’ world which is described by quantum mechanics, quantum theory is also, inevitably, the fundamental theory of nature. Though stunningly powerful, the quantum mechanical view of the world has compelled us to reshape and revise our ideas of reality and notions of cause, effect and measurement. Without going into too many details of the various conceptual difficulties of quantum mechanics, we will only focus on the quantum mechanical description of measurement which is of direct relevance to the description of the QZE.

The quantum mechanical description of a system is contained in its wavefunction or state vector, $|\psi\rangle$, which lives in an abstract ‘Hilbert space’. The dynamics of the wavefunction is governed by the Schrödinger equation:

$$i\hbar \frac{d}{dt} |\psi\rangle = H|\psi\rangle,$$

where $H$ is the Hamiltonian operator, and the equation is linear, deterministic and the time evolution governed by it is unitary. Unitary evolutions preserve probabilities. Examples of unitary transformations are rotations and reflections. Dynamical variables or observables are represented in quantum mechanics by linear Hermitian operators, which act on the state vector. An operator, $\hat{A}$, corresponding to a dynamical quantity, $A$, is associated with eigenvalues $a_i$ and corresponding eigenvectors, $\{|a_i\rangle\}$, which form a complete orthonormal set. Any arbitrary state vector, $|\psi\rangle$, in general, can be represented by a linear superposition of these eigenvectors, or, for that matter, a combination of any orthonormal set of basis vectors in Hilbert space. Thus, one can
write \(|\psi\rangle = \Sigma c_i|\alpha_i\rangle\). A basic postulate of quantum mechanics regarding measurement is that any measurement of the quantity \(A\) can only yield one of the eigenvalues \(a_i\), but the result is not definite in the sense that different measurements for the quantum state \(|\psi\rangle\) can yield different eigenvalues. Quantum theory predicts only that the probability of obtaining eigenvalue \(a_i\) is \(|c_i|^2\). Quantum theory defines the expectation value of the operator \(\hat{A}\) as:

\[
\langle \hat{A} \rangle = \langle \psi | \hat{A} | \psi \rangle = \Sigma a_i |c_i|^2.
\]  

(2)

In terms of the density matrix \(\hat{\rho} = |\psi\rangle \langle \psi|\), an equivalent formula for the expectation value is:

\[
\langle \hat{A} \rangle = \text{Trace}\{\hat{A} \hat{\rho}\}.
\]  

(3)

An additional postulate of quantum mechanics is that the measurement of an observable \(A\), which yields one of the eigenvalues \(a_i\) (with probability \(|c_i|^2\)) culminates with the reduction or collapse of the state vector \(|\psi\rangle\) to the eigenstate \(|\alpha_i\rangle\). This means that every term in the linear superposition vanishes, except one. This reduction is a non unitary process and hence in complete contrast to the unitary dynamics of quantum mechanics predicted by the Schrödinger equation and this is where the crux of the conceptual difficulties encountered in quantum theory lies. For now we just accept this as a basic postulate of quantum theory (also called the projection postulate) and go on to describe the QZE.

### 2.2 The Result of Misra and Sudarshan

The QZE (or paradox) was the name given by Misra and Sudarshan to the phenomenon of the inhibition of transitions between quantum states by frequent measurements. For their study, they considered the decay of an unstable state, such as an unstable particle, like a radioactive nucleus. The classic model of any system decay is an exponential function of time in most situations. We are all familiar with the radioactive decay law

\[
N(t) = N(t_0)e^{-\lambda(t-t_0)},
\]  

(4)

where \(N(t)\) is the number of nuclei that have not decayed after time \(t\), and \(\lambda\) is a constant which depends on the properties of the species of nuclei. While the decay of a quantum system is similar to this classic model of exponential decay, there have been theoretical studies that show that in certain timescales (specifically, for very short and very long times as measured from the instant of preparation of the state of the system), there can be a deviation from the familiar exponential decay law. In fact, recently this theoretical deviation from the exponential decay law has also been confirmed experimentally in quantum tunnelling experiments with ultra cold atoms by a group at the University of Texas, USA. It is in these special time regimes that we see manifestations of the QZE. Consider the decay of an unstable quantum state. Let \(\psi_0\) be the (undecayed) state of the system at \(t = 0\) and \(\psi(t)\) be the state at any later time \(t\). The evolution of the state is governed by a unitary operator, \(U(t)\), where

\[
U(t) = e^{-iHt},
\]  

(5)
Anu Venugopalan

The short-time quantum decay is not exponential in time, but quadratic. $H$ being the Hamiltonian of the system. As discussed in the previous section, any observation that the state has not decayed will cause a collapse (reduction) of the wavefunction to the undecayed state. The survival probability, $P(t)$, i.e., the probability that the system is still in the undecayed state, will be the modulus squared of the survival amplitude and can be written as:

$$P(t) = |\langle \psi_0 | U(t) | \psi_0 \rangle|^2.$$ (6)

Now, (7) can be expanded as:

$$P(t) = 1 - t^2 \left( \langle \psi_0 | H^2 | \psi_0 \rangle - \langle \psi_0 | H | \psi_0 \rangle^2 \right) + \ldots$$ (8)

If

$$\Delta H = \sqrt{\langle \psi_0 | H^2 | \psi_0 \rangle - \langle \psi_0 | H | \psi_0 \rangle^2},$$ (9)

then the survival probability in the short time limit can be rewritten as:

$$P(t) \approx 1 - t^2 (\Delta H)^2 + \ldots$$ (10)

If we define $\tau_Z = 1/\Delta H$ as the Zeno time, this gives:

$$P(t) \approx 1 - \frac{t^2}{\tau_Z^2} + \ldots$$ (11)

which, for short times, can be written as:

$$P(t) \approx \left(1 - \frac{t^2}{\tau_Z^2}\right).$$ (12)

The above expression shows that the short-time quantum decay is not exponential in time, but quadratic. Now, let us suppose that one makes $N$ equally spaced measurements over the time period $[0, T]$. If $\tau$ is the time interval between two measurements, then $T = N \tau$. Let us assume that the measurements are made at times $T/N, 2T/N, 3T/N, \ldots, (N-1)T/N$ and $T$ and are instantaneous. So, essentially this describes an alternate sequence of unitary evolutions (each lasting for a time $\tau$) followed by a collapse (the basic postulate of quantum measurement). The survival probability after $N$ measurements, or after time $T$ can be written as:

$$P^N(T) = [P(\tau)]^N = \left(1 - \frac{T^2}{N^2 \tau_Z^2}\right)^N.$$ (13)

It can be seen that in the limit of continuous measurements, i.e., when $N \to \infty$,

$$\lim_{N\to\infty} P^N(T) = 1.$$ (14)
Thus the probability that the state will survive for a time $T$ goes to 1 in the limit $N \to \infty$. This means that continuous measurements actually prevent the system from ever decaying! So, much like the motionless arrow in Zeno’s paradox, the system never decays, or a ‘watched pot never boils’. Now, can we see this happening in a real experiment? Unfortunately, in spontaneous decay this effect is very difficult to observe for reasons that we will briefly discuss in a later section. No experiment has been able to probe this regime to observe the inhibition of the decay of an unstable particle like a radioactive nucleus, as yet. However, as mentioned earlier, recently there have been experimental groups which have reported the observation of the QZE in unstable systems comprising of trapped ultra cold atoms which undergo quantum mechanical tunnelling. We will discuss these later. First, we will describe an earlier experiment carried out by Itano et al in 1990 of the experimental group headed by Wineland at the National Institute of Standards and Technology, Boulder, Colorado. This was the first instance when a manifestation of the QZE was successfully demonstrated experimentally. This is discussed in the next section.

2.3 Experimental Manifestations of the QZE

2.3.1 The Experiment of Itano et al

Following an original proposal by Cook, Itano et al, at the National Institute of Standards and Technology, Boulder, Colorado, experimentally demonstrated the occurrence of the QZE in induced transitions between two quantum states. Unlike the case studied by Misra and Sudarshan, this is a situation where there is no spontaneous decay of an unstable system but an induced transition between two states of a system. The experiment of Itano et al tested the inhibition of the induced radio frequency transition between two hyperfine levels of a Beryllium ion, caused by frequent measurements of the level population using optical pulses. The experiment can be understood as follows (Figure 2): Consider a two-level system, with the levels labelled as 1 and 2. Assume that the system can be driven from level 1 to level 2 by applying a resonant radio frequency pulse. Assume that it is possible to make instantaneous measurements of the state of the system, i.e., to ascertain whether the system is in level 1 or in level 2. In order to observe the level populations, level 1 is connected
Trapped ions and atoms are considered ‘clean’ systems that can be observed for long periods of time and isolated from noise.

The dynamics of a two-level system in the presence of resonant driving field is well-studied and understood.

Suppose the ion is in level 1 at a time \( t = 0 \). An rf field having resonance frequency \( \Omega = (E_2 - E_1)/\hbar \) is applied to the system and this creates a state which is a coherent superposition of states 1 and 2. The frequency \( \Omega \) is called the Rabi frequency. An on-resonance ‘\( \pi \) pulse’ is a pulse of duration \( T = \pi/\Omega \) and takes the ion from level 1 to level 2. If \( P_2(t) \) is the probability at time \( t \) for the ion to be at level 2, then \( P_2(T) = 1 \). This would be the situation when no ‘measurement pulses’ are applied. In the experiment, \( N \) measurement pulses are applied (which connect level 1 to level 3 through an optical pulse each time), within time \( T \), i.e., at times \( \tau_k = kT/N, k = 1, 2, 3, ..., N \). Note that the dynamics of the two-level system driven by the resonant rf field is unitary and can be described quantum mechanically using the Bloch vector representation. The nonunitary projection postulate (or the collapse induced by quantum measurement) is incorporated each time a measurement is made. It is easy to show that at the end of \( N \) measurements, i.e., at the end of the rf pulse at time \( T \), the probability \( P_2(T) \), which corresponds to the population of level 2 is given by:

\[
P_2(T) = \frac{1}{2} \left[ 1 - \cos\left(\frac{\pi}{N}\right)\right].
\]

(15)

For large \( N \), i.e., in the limit of continuous measurements, one can see that

\[
P_2(T) = \lim_{N \to \infty} \frac{1}{2} \left[ 1 - \cos\left(\frac{\pi}{N}\right)\right] \approx 0.
\]

(16)

Clearly, the continuous measurements described above inhibit the induced transition from 1 to 2, making the system ‘freeze’ in level 1. This effect showed itself up in the real experimental observations of Itano et al. Thus, although it was not seen in the decay of an unstable particle, the experiment of Itano et al was the first real demonstration of the QZE. Interestingly, the experiment was followed by a slew of papers where many issues were raised regarding the actual dynamics of the mechanism explored in the experiment by Itano et al. The experiment was critically analyzed from the point of view of quantum measurements and questions were raised regarding whether or not the collapse postulate plays any role at all in the outcome of the
Anu Venugopalan

experiment. Many physicists assert that the QZE is simply a consequence of the unitary dynamics of conventional quantum mechanics and need not involve the non-unitary collapse of quantum measurement. However, since the projection postulate of conventional quantum measurement theory also successfully described the outcome of this experiment, it is valid to see the experiment of Itano et al as a demonstration of the inhibition of transition due to frequent measurements, or the QZE. In the next section we describe one more experimental manifestation of the QZE.

2.3.2 The Experiment of Kwiat et al

In 1995, Paul Kwiat and his group at the University of Innsbruck realized a version of the QZE in the laboratory using the polarization directions of single photon states. Their experiment was based on an example first suggested by Asher Peres in 1980. Consider plane polarized light. It can have two possible polarization directions, say, ‘vertical’ and ‘horizontal’. We know that when such a beam passes through an optically active liquid (e.g., sugar solution) its plane of polarization is rotated by a small angle (which depends on the concentration of the sugar solution, for example). Consider a photon directed through a series of such “rotators” so that each slightly rotates its polarization direction so that an initially vertically polarized photon ends up horizontally polarized. At the end of this series of rotators, the photon encounters a polarizer. A polarizer is a device that transmits photons with one kind of polarization but absorbs photons with the perpendicular polarization. An ideal Nicol prism acts as a polarizer (or analyzer). Now let us suppose that an experiment is set up with six rotators, each of which rotates the plane of polarization of a vertically polarized photon by 15°. At the end of this series is a polarizer which transmits only vertically polarized light, which is then detected by a photon detector (Figure 3). It is obvious that in the above set up, the photon will never get to the detector as its polarization will have turned by 90° after passing through the six rotators and become horizontal.

To implement the Zeno effect, Paul Kwiat and his colleagues sought to inhibit this stepwise rotation of the polarization, or the evolution of the polarization state from the vertical to the horizontal, by measurements of the polarization state. Kwiat et al realized this by interspersing a vertical polarizer between each rotator (Figure 3). If the first rotator rotates the plane of polarization by an angle \( \alpha \), then the vertical polarizer kept after it will transmit the photon with a probability \( \cos^2 \alpha \).}

If the first rotator rotates the plane of polarization by an angle \( \alpha \), then the vertical polarizer kept after it will transmit the photon with a probability \( \cos^2 \alpha \).

Figure 3. Six rotators turn the polarisation by 15 degrees at each stage such that a vertically polarized photon changes to a horizontally polarized one.
then encounters the second polarizer where it will be transmitted with a probability \( \cos^2 \alpha \) and the vertical polarization will be, once again, restored. This process repeats till the photon comes to the final polarizer. If \( \alpha = 15^0 \), \( (\cos^2 \alpha)^6 = \frac{1}{4} \). Thus an incident photon has two third chance of being transmitted through all six inserted polarizers and making it to the detector. It can be easily seen that if one increases the number of stages, decreasing the rotation angle at each stage, the probability of transmitting the photon to the detector increases. If there were an infinite number of stages, the photon would always get through and hence the rotation of the plane of polarization would be completely inhibited – the Zeno effect! In the actual experiment, Kwiat et al created single photon states using a nonlinear crystal.

Thus, like the experiment of Itano et al, the experiment of Kwiat et al demonstrated the suppression of evolution in a driven two-state system through frequent measurements. What about the QZE in unstable systems?

### 2.4 The Result of Kurizki and Koffman – The Anti-Zeno Effect

In the previous section we have seen experimental evidences of the QZE in induced transitions between two quantum states. A natural question that arises, then, is whether the Zeno effect can be used (in a real experiment) some day to ‘freeze’ radioactive nuclear decay. For the past three decades it seemed that the answer to the question was a ‘yes’, provided one had the experimental technology and sophistication to perform successive, ‘frequent enough’ measurements. However, recent work by Gershon Kurizki and Abraham Koffman at the Weizmann Institute of Science, Israel, has shown that such a freezing may actually not be possible at all. Kurizki and Koffman have argued that there is an ‘Anti-Zeno Effect’ which in fact enhances the decay of unstable particles instead of inhibiting it! According to their calculations, the ability to ‘freeze’ the evolution through frequent measurements depends on the ratio between the ‘memory time’ of the decay, and the time interval between successive measurements. Every decay process has a ‘memory time’. This memory time is the time following a quantum event in which the particle can still return to its initial state. In the case of radioactive decay, for instance, the memory time is the period in which the radiation has not yet escaped from the atom, allowing the system to ‘remember’ its state prior to the decay. Typically, this memory time for an unstable particle is less than one billionth of a billionth of a second. Kurizki and Koffman argue that frequent measurements in this time scale (if it were possible) would cause more particles to be created. This would interfere with, and essentially destroy the original system, making it meaningless to ask whether the decay has frozen or not. On the other hand, if the time interval between measurements is longer than the memory time (i.e., observations are not fast enough for the ‘expected’ QZE), the rate of decay increases and one would have the Anti Zeno effect. While we will not go into the details of their work, we can state that the surprising conclusion of the research of Kurizki and Koffmann is that the Anti-Zeno effect (i.e., the increase of decay through frequent...
measurement) can occur in all processes of decay, while the Zeno effect which would slow down and even stop decay requires conditions that are much rarer.

While the predictions of Kurizki and Koffman are yet to be experimentally verified on an unstable system like a radioactive nucleus, recent experiments by Mark Raizen and his colleagues at the University of Texas, Austin have demonstrated the quantum Zeno and the quantum Anti Zeno effects in the tunnelling behaviour of cold trapped atoms. Raizen’s team trapped sodium atoms in a ‘light wave’. Such a system, if left alone, will slowly decay as individual atoms escape through quantum mechanical tunnelling through an energy barrier which would be classically insurmountable. Through ingenious experimental techniques, the team showed that the tunnelling rate slowed down dramatically when they ‘measured’ the system every millionth of a second – the QZE! When they measured the system every five millionth of a second, the tunnelling rate increased – the quantum Anti Zeno effect! It is interesting to notice the happy coincidence that this spectacular experimental test of the QZE which is closest in spirit to the original proposal of Misra and Sudarshan was performed at the University of Texas, Austin – the very same place from where Misra and Sudarshan published their work, almost three decades ago.

3. Conclusions

Much water has flown under the bridge since Zeno wondered about Achilles and the tortoise at the dawn of civilization. Though mathematicians have solved the classical paradoxes of Zeno long ago by introducing the concept of real numbers, limits, continuity and calculus to describe quantities of duration and distance, the notion of freezing motion by continuous observation has turned out to be a very real effect in the strange world of quantum physics. Spectacular experiments bear testimony to the reality of this effect in the quantum domain and the fields of atom optics and cold trapped ions continue to spring up tantalizing new surprises every day. While in the classical world, Achilles overtakes the tortoise and all is well with the world, in the mysterious land of the quantum, watched pots stop boiling (or boil faster, maybe!) and the ghost of Zeno continues to make its presence felt in unimaginably interesting ways.

Acknowledgements

The author would like to thank Ragothaman Yennamalli and Vivek for help with the figures.
Suggested Reading

The topics touched upon in this article cover several references. The interested reader may look at some of the following:

This article reviews the history of the phenomenon of Brownian motion, how it confirmed the molecular view of matter, and how it led to the invention of nonequilibrium statistical mechanics.

1. Introduction

1.1 What is Brownian Motion?

Look through a microscope at very still dirty water or curdled milk, and you'll see tiny particles in a state of constant, rapid, erratic motion. Figure 1 is the trajectory of such a particle, as seen in the 1916 experiments of the French physicist Jean Perrin [1].

In 1828, a botanist named Robert Brown saw this movement in pollen grains dispersed in water, and achieved permanent fame, because it has been known as Brownian motion ever since. What he observed, to be precise, was that pollen, dropped into water, disintegrated into a very large number of tiny particles, a few μm in size, which were seen to be dancing about ceaselessly. He thought at first that the movement was a sign of life, and that the tiny particles were the fundamental constituents of living matter, since pollen came from plants. He later found that inorganic materials did this as well, and so speculated that all matter was made of these 'primitive molecules'. People seem to have realised quite soon that these tiny but visible particles weren’t molecules, just finely divided matter, and the focus shifted to what was making them dance around like that. Many explanations were offered: perhaps the particles were being carried around by flows produced by small temperature differences in the water; maybe illuminating the material was itself causing the movement; maybe electrical forces were involved; and so forth. Experiments and speculations on the phenomenon continued through the 1880s. A particularly careful set of experiments by Gouy showed that the fluctuations became more rapid if the viscosity of the fluid was decreased, and that neither changes in the intensity of the light source used to illuminate the sample nor large electromagnetic fields affected the movement. He was thus convinced of the view that some others had by then begun to express: the molecules of the liquid were moving around rapidly simply because the liquid had a temperature and the pollen grains were being kicked here and there by repeated collisions with these molecules.

Einstein had been wondering about ways to find direct evidence for the existence of

---

discrete molecules. Since there were already estimates of their sizes (of the order of $10^{-10}$ m; see section 2) which made it clear that they were too small to see through a microscope, he took a different approach. Unaware initially of the observations of Brown and the work that followed, he argued that a small particle suspended in a fluid should be agitated by collisions with molecules, and constructed a detailed quantitative theory of the motion of such a particle. His calculations showed that the resulting motion, for particles the size of pollen grains, should be seen easily through a microscope. He suggested cautiously that the movement he was predicting might be the same as the Brownian movement, which he had by then learnt about. As we shall see, the importance of his work was that it provided a decisive test for the kinetic theory and hence for the very existence of molecules.

The developments that followed Einstein’s theory of Brownian motion show us how the most profound physics and mathematics can emerge from simple, direct observation of natural phenomena. Ideas related to Brownian motion and random walks appear today in mathematics, all the natural sciences, engineering, linguistics, finance, economics, and even the social sciences. Accordingly, I think an acquaintance with the fundamentals of this phenomenon should be required of all scientists and engineers, at the very least. Brownian motion is visible evidence for the existence of molecules. Careful studies of this motion made it possible, as we shall see by and by, to count the number of atoms in a given mass of gas (i.e., to obtain Avogadro’s
number\(^3\)), therefore to infer the mass of a proton, and lastly to estimate the sizes of molecules, all using an ordinary light microscope!

This article discusses the physics and a bit of the mathematics of Brownian motion. I’ve tried to use only ‘plus-two’ and B Sc physics and mathematics (the ideal gas law, partial pressures, elementary kinetic theory, terminal velocity, combinations and permutations, vector calculus), so that science and engineering undergraduates should, in principle, be able to understand all of it if they stretch themselves a bit. The interested standard XI or XII student should be able to follow a fair fraction of the article, and readers with other backgrounds should get the general idea.

1.2 Einstein’s Key Result

How would you count the number of molecules in a mole (or any other quantity) of gas? Let’s start with high school physics: it is an experimental fact that the pressure \(p\), volume \(V\), and temperature \(T\) of \(\mu\) moles of a dilute gas obey

\[
pV = \mu RT, \tag{1}
\]

where \(R\) has the same value for all gases. The kinetic theory of gases, about which I will say more below, explains this observation and tells us that the constant

\[
R = N_A k_B \tag{2}
\]

where Avogadro’s number \(N_A\) is the number of molecules in one mole of any ideal gas, and \(k_B\) is known as Boltzmann’s constant. Now remember where \(k_B\) comes from: kinetic theory tells you that temperature is just molecular kinetic energy, i.e., that the molecules in a gas at temperature \(T\) have mean kinetic energy \((3/2)k_B T\), where \(k_B\) is simply an unknown constant of proportionality. This is because we measure temperature using a rather arbitrary scale: whereas the absolute zero of temperature is uniquely defined, the unit of temperature is not. From kinetic theory we know that if we double the absolute temperature we have doubled the kinetic energy of the molecules, but thermometers do not tell us the actual value of the kinetic energy in Joules. Thus the ideal gas law allows us to measure \(N_A k_B\), but not \(N_A\) by itself. If we could measure \(k_B\) independently, we would then have succeeded in counting the number of molecules in a mole.

What Einstein showed was that a particle of radius \(a\) executing Brownian motion in a fluid with viscosity \(\eta\) and temperature \(T\) would wander, in a time \(t\), a distance\(^4\)

\[R(t) \sim \sqrt{Dt},\]

with the ‘diffusion constant’ \(D\) determined entirely by \(k_B\) and the known quantities \(T\), \(\eta\), and \(a\). A measurement of \(R(t)\) through a microscope would thus yield a value for \(k_B\) and hence, as argued above, for Avogadro’s number \(N_A\). Einstein extracted \(N_A\) by applying his theory to the motion of sugar molecules in solution. Now you can’t see a sugar molecule through a microscope, so the properties of its Brownian motion had to be inferred by measuring the diffusion or spread of the concentration of sugar through un stirred water. Perrin got \(N_A\) by applying Einstein’s

\(^3\)I shall assume that the readers of this article are all familiar with the idea of a mole or gram molecular weight of a substance, since the concept is taught in the X or XI standard. Moles of hydrogen, oxygen and water, respectively weigh 2, 32 and 18 grams, and \(N_A\), the number of molecules in one mole, named in honour of Avogadro, is about \(6 \times 10^{23}\). In particular, note that measuring \(N_A\) tells us that one hydrogen atom (and hence one proton) weighs about \(1.7 \times 10^{-24}\) gram.

\(^4\)I will frequently make estimates of this sort where the \(\sim\) symbol means that I have left out purely numerical factors of order unity, but have included the dependence on all physical parameters.
theory to infer $k_B$ from the vertical distribution of heavier-than-water micron-sized resin particles, as well as from direct measurements of their Brownian motion. The resulting values for $N_A$ and the size of molecules agreed wonderfully with those due to Clausius and Maxwell (see section 2) – a triumph for the kinetic theory.

1.3 Structure of the Article

The rest of this article is organised as follows: In the next section, we recall why molecules were postulated to begin with, and how the kinetic theory extracts one set of estimates of $N_A$ and molecular sizes from macroscopic experiments. In section 3, we present and solve a mathematical model for the motion of a Brownian particle, and use it to show that the particle wanders a distance proportional to the square-root of the time. We then reconstruct the remarkable connection that Einstein established between diffusion and viscous friction, which allows us to infer Boltzmann’s constant from experiment. We show in addition how a direct consideration of the equilibrium between gravity and Brownian motion when the particles are less dense than the fluid in which they are suspended gave Einstein an independent and direct route to Boltzmann’s constant. We also discuss briefly the experiments of Perrin. The article concludes in section 4.

2. Molecules and Kinetic Theory

By the late 19th century it was generally agreed that matter was made of discrete bits called atoms or molecules. As you learnt in the 11th or 12th standard, the first hints in this direction came from chemistry: gases reacted in simple proportions by volume; the masses of oxygen which react with a given mass of nitrogen to form its various oxides are again in simple integer ratios; and so forth. Simple but striking chemical facts such as these were accounted for nicely by two crucial ideas: (i) that chemical reactions involved the combination of fixed numbers of atoms; and (ii) Avogadro’s hypothesis that different gases of the same volume, and at the same temperature and pressure contained the same number of molecules. As a result, it became possible to estimate the ratios of the masses of molecules, but not the values of the masses themselves. Of course, no one then knew why Avogadro’s hypothesis should be true. It was not until the success of the molecular-kinetic theory of heat, or simply kinetic theory, that the existence of molecules was firmly established. The kinetic theory accounted for the observed properties of gases which you’ve studied in school, such as the gas equation (1), in terms of a picture of an enormous number of molecules flying around at random and bouncing elastically off the walls of the container, provided one defined temperature $T$ by the statement that the average kinetic energy per molecule is equal to $(3/2)k_BT$. Using this theory, Clausius and Maxwell were able to express measurable macroscopic properties of gases in terms of molecular parameters like the size $d$ of a molecule, the number $n$ of molecules per unit volume of gas, and the average speed $\bar{v}$ of the molecules. They found what really mattered was the mean
free path ℓ, i.e., the mean distance a molecule travels before it collides with another molecule.

From Figure 2 it is easy to convince yourself that

\[ \ell \sim 1/(nd^2). \]  

(3)

They found that the kinematic viscosity was

\[ \nu \sim \bar{v}\ell \]  

(4)

which you could have guessed using dimensional analysis since the units of \( \nu \) are length² time⁻¹, and that the speed of sound in a gas was

\[ c_s \sim \bar{v}. \]  

(5)

Straightforward measurements of \( c_s \) and \( \nu \) thus yield the combination \( \ell \sim nd^2 \). The measured mass density of a gas gives us \( nm \), where \( m \) is the (unknown) mass of a molecule. Knowing the mass density of the liquid or solid phase of the same material gives us \( md^{-3} \) if we assume that in that phase the molecules are closely packed (one molecule per \( d^3 \)). Put together, these considerations yield values for \( n \), \( m \), and \( d \). For air at room pressure and temperature we find \( \nu \approx 2 \times 10^{-5} \text{ m}^2\text{s}^{-1} \) and \( c_s \approx 300 \text{ m/s} \). In its liquid phase, the mass density is \( \sim 1000 \text{ kg/m}^3 \). This leads to answers which we now know to be quite reasonable: \( d \sim 10^{-10} \text{ m} \) and \( n \sim 3 \times 10^{25} \) molecules per \( \text{m}^3 \), i.e., about \( 6 \times 10^{23} \) molecules in a mole.

2.1 The Need for an Independent Test

Clearly, a crucial test of the kinetic theory would be to see if the above estimates agreed with those from an entirely independent experiment. In this context, note that although kinetic theory was based on a molecular description, its successes at the time lay in its ability to account for macroscopic phenomena, averaged over a large number of molecules. How much more satisfying it would be to have direct evidence for molecules, in the form of events caused by the impact of a small number of them! Taking the kinetic theory seriously, if you looked at a tiny region in a gas or liquid for a tiny instant of time, then sometimes you’d see a molecule going one way, sometimes another, and sometimes maybe you’d see no molecule at all. Although on average the molecules you saw would be as likely to be going in any direction as in any other, there would be fluctuations about this average behaviour. It is an important principle of statistics that if the number of impacts on the particle is small their effect will not average out instantaneously. So if the particle were itself small or if it were floating around in a partially evacuated chamber so the number density of molecules was small, these fluctuations could become large and deliver an appreciable kick to the particle, in a direction which would vary randomly from one kick to the next. The resulting movement is what we see as Brownian motion. This was Einstein’s idea. In the next section we model the trajectory of the Brownian particle by a ‘random walk’ whose properties will lead us along one path to our final result.
3. Random Walks

3.1 Brownian Particles are Just Big Molecules

A notable feature of the theory of Brownian motion is that it requires no understanding of the detailed nature of the collisions between the molecules and the Brownian particles, of how many collisions take place per second, and at what angle of incidence, and so forth. All one does is to treat the Brownian particle as just another molecule, only larger than the rest. Kinetic theory then tells us, crucially, that a Brownian particle of mass $M$ has the same mean kinetic energy $(3/2) k_B T$ as any other molecule in the system, and hence a randomly directed mean speed

$$ \bar{V} = \sqrt{\frac{3k_B T}{M}}. \quad (6) $$

The collisions with the molecule enter in one other way: they damp out the instantaneous velocity of the Brownian particle in a time $\tau$ determined by a macroscopic property of the fluid, namely its viscosity. We can imagine that the Brownian particle then abruptly acquires a speed $\bar{V}$ again through collisions, once again in a random direction. Physically it is clear that this is not a very rapid way for the particle to get around, since the random direction of the velocity will take the particle as often in one direction as in the opposite direction. The product $\bar{V}^2 \tau$ has units of length$^2$ time$^{-1}$, as does the diffusion coefficient $D$ defined later in equation (14); if you therefore guessed that the ‘random walk’ defined above in fact leads to a particle ‘diffusing’ a typical distance $\sim (Dt)^{1/2}$ with $D \sim \bar{V}^2 \tau$, you’d be right. Let us make these ideas mathematically precise, as Einstein did. My presentation is equivalent to his, although not identical in detail.

3.2 How Far does a Brownian Particle Travel?

Assume that the trajectory of the Brownian particle is a sequence of discrete steps of length $b$ in a completely random direction. Each step takes a time $\tau$, which means that during the step the particle has a speed $\bar{V} = b/\tau$. Let the direction of the $j$th step be given by the unit vector $\mathbf{u}_j$, which can point in any direction with equal probability. The $\mathbf{u}_j$s for different $j$ are statistically independent. After a time $t \equiv N \tau$, a particle which started from the origin would be at a point (see Figure 3)

$$ \mathbf{R}(t) = b \sum_{j=1}^{N} \mathbf{u}_j. \quad (7) $$

Each $\mathbf{u}_j$ has mean zero, since it is distributed uniformly over all possible directions, so that the average displacement

$$ \langle \mathbf{R}(t) \rangle = b \sum_{j=1}^{N} \langle \mathbf{u}_j \rangle = 0. \quad (8) $$
This average tells you nothing about how far from the origin the particle is likely to be. The average vanishes only because there is no preferred direction in which the particle is likely to be found. We could look instead at the average of the absolute value \(|R(t)|\); it turns out to be easier to calculate the mean-squared displacement, i.e., the variance of \(R(t)\) which, from (7), is

\[
\Delta^2(t) \equiv \langle R(t) \cdot R(t) \rangle \\
= b^2 \sum_{i=1}^{N} \sum_{j=1}^{N} \langle \mathbf{u}_i \cdot \mathbf{u}_j \rangle. \tag{9}
\]

This sum is easy to do: the \(\mathbf{u}_i\) for distinct \(i\) are statistically independent, so that for \(i \neq j\), \(\langle \mathbf{u}_i \cdot \mathbf{u}_j \rangle = \langle \mathbf{u}_i \rangle \cdot \langle \mathbf{u}_j \rangle = 0\); and for \(i = j\), \(\langle \mathbf{u}_i \cdot \mathbf{u}_j \rangle = 1\) since they are unit vectors, so we see that the root-mean-square displacement is

\[
\Delta(t) = b \sqrt{N} = \sqrt{\bar{V}^2 \tau t} \tag{10}
\]

with \(\bar{V}\) and \(t\) as defined above.

The above result is probably familiar to you in its one-dimensional form: if you add \(N\) numbers each of which is equally likely to be +1 or −1, the resulting sum has mean zero and standard deviation \(\sqrt{N}\). In fact, for a one-dimensional random walk, you can quite easily get the entire probability distribution, and use this to solve immediately the problem in any number of dimensions. For instance: simple counting arguments, using the ideas of combinations and permutations you learnt in school, tell you that the probability that a random walk in one dimension (step length \(\delta\), step duration \(\tau\)) achieves a net displacement \(k\delta\) (where \(k\) can be positive or negative) after \(N\) steps is

\[
P_N(k) = \frac{N!}{[(N + k)/2]![(N - k)/2]!} 2^{-N}. \tag{11}
\]

Another way of seeing why this average is zero: \(\mathbf{u}_i \cdot \mathbf{u}_j = \cos \theta\) where \(\theta\) is the angle between the two vectors. \(\theta\ & 180^\circ - \theta\ are equally probable since the two vectors are independent. These make contributions to \(\langle \cos \theta \rangle\) which cancel.
For $N \gg k$

$$P_N(k) \approx \sqrt{\frac{2}{\pi N}} e^{-k^2/2N}$$  \hfill (12)

asymptotically (think about how you might prove this), as you can convince yourself by plotting $P_N(k)$ versus $k$ for a reasonably large value of $N$, say 20. For large $N$ it’s convenient to work in terms of the coordinate $x = k\delta$ and the time $t = N\tau$, treated as continuous rather than discrete variables. Then the probability $P(x, t)dx$ of finding the particle between $x$ and $x + dx$ at time $t$ is given by

$$P(x, t) = \frac{e^{-x^2/4Dt}}{\sqrt{4\pi Dt}}$$  \hfill (13)

where the ‘diffusion coefficient’

$$D = \frac{\delta^2}{2\tau}. \hfill (14)$$

For a walk in $q$ dimensions, since the displacements along the coordinate axes are statistically independent, we can make the random elementary step of length $b$ by adding vectorially $q$ elementary random steps of length $\delta$ in the coordinate directions. Thus $b = \sqrt{q}\delta$, and the distribution is just the product of $q$ distributions of the form (13):

$$P(R, t) = \frac{e^{-R^2/4Dt}}{(4\pi Dt)^{q/2}}, \hfill (15)$$

and

$$D = \frac{b^2}{2q\tau} = \frac{\bar{V}^2\tau}{2q}, \hfill (16)$$

so that the root-mean-square displacement of equation (10) is

$$\Delta(t) = \sqrt{2qDt} \hfill (17)$$

with $D$ as in (16). So measuring the diffusion coefficient $D$ will give us $\bar{V}^2\tau$. If we can relate $\tau$ to an independently measurable physical quantity, we can then find $k_B$ using (6).

Let’s do this qualitatively first. Macroscopically, our Brownian particle is just a particle moving through a viscous fluid. The effect of the fluid is put in simply as a force $f = -\Gamma v$ proportional to the particle velocity $v$, with a coefficient $\Gamma$ of viscous friction. The equation of motion for the particle velocity in one dimension is

$$M \frac{dv}{dt} = -\Gamma v. \hfill (18)$$

You all have probably solved an equation like (18), with an additional constant force on the right-hand side, when you studied ‘terminal velocity’. We see from (18) that our particle, starting from any initial speed (say $\bar{V}$), will effectively lose this speed in a time of order $M/\Gamma$. In other words,

$$\tau \sim \frac{M}{\Gamma} \hfill (19)$$
so that, from (6), (16), and (19),

\[ D \sim \frac{k_B T}{\Gamma}. \]  

(20)

This turns out to be exactly correct, i.e., \( D \) is precisely \( \frac{k_B T}{\Gamma} \), as we’ll see shortly. \( \Gamma \), moreover, can be obtained independently, as we shall see below. We’ll return to the complete solution of the problem after two short digressions. One of these is on ‘noisy’ differential equations, the other on numerical simulations of random walks.

### 3.2.1 Langevin Equations

One way to model Brownian motion is to include the kicks of the molecules in Newton’s second law of motion (18) for the Brownian particle in the form of a ‘noise’. This is done by adding to the right-hand side a function \( f(t) \) defined only by its statistical properties. The resulting object is called a ‘stochastic differential equation’ (often a ‘Langevin equation’ after the man who invented it in 1908) and can be solved using techniques [2] which are now standard. This approach is now the method of choice for solving problems in the dynamics of statistical systems. The noise, while random, is not arbitrary: just as (20) connects diffusion and friction, there is a strict relation between \( \Gamma \) and the statistical properties of the noise. This sort of approach is now widely used even for systems where the noise comes not from equilibrium thermal fluctuations but from other sources such as chaos. In such truly nonequilibrium situations there is no general relation between the noise statistics and the damping coefficients in the problem. I will say no more about this here: the interested reader can look at the ‘Suggested Reading’ at the end of this article.

### 3.2.2 Random Walks through Simulations and Coin-tosses

Try generating random walks on a PC and measuring their properties. It’s not too difficult and it is very instructive. Figure 4 shows the results of such a study. To test our predictions (10), (11) or (12), you’ll have to generate a large number of walks, each starting from the origin, and extract the probability of finding the particle at various locations. It’s fun to do the same thing by tossing a coin: a head means a step to the right, a tail to the left. A hundred tosses divided into groups of ten give you ten short random walks. You could even use it to check if the coin is biased. Try it!

### 3.3 Viscosity and Diffusion: the Stokes–Einstein Relation

Returning to the main narrative, recall that we’re trying to use Brownian motion to obtain a value for Boltzmann’s constant \( k_B \). I presented qualitative arguments for this which I will now make quantitative and complete, by expressing \( D \) in terms of the viscosity of the fluid.

To do this, note first, as you can and should check explicitly, that the distribution (15)
obey the ‘diffusion equation’

\[
\frac{\partial P(R, t)}{\partial t} = D \nabla^2 P.
\]  

(21)

This means that if we now take a collection of \( N \) independent random walkers (pollen grains), not interacting with each other, and start them all out at the origin, then the mean number of grains we’d find per unit volume at a point \( R \) at time \( t \) would be given (why?) by \( n(R, t) = NP(R, t) \). Thus \( n(R, t) \) would also obey the diffusion equation (21). A useful way of writing this assuming, for simplicity, variation only along the \( z \) direction, is

\[
\frac{\partial n}{\partial t} = -\frac{\partial j}{\partial z}
\]  

(22)

where the particle current, that is the number of particles crossing unit area normal to the \( z \) axis in unit time is

\[
j(z, t) = -D \frac{\partial n}{\partial z} \equiv j_{\text{diff}}.
\]  

(23)

Now imagine imposing, along the \(-z\) axis, a constant force \( F \) on the particles. In the typical experiment this force is gravity, in which case \( F = Mg - W \), where \( M \) is the mass of a Brownian particle, \( g \) the acceleration due to gravity, and \( W \) the weight of water displaced by the Brownian particle. In the absence of Brownian motion and walls this would cause each particle to drift downwards with a terminal speed \( F/\Gamma \) where \( \Gamma \) is the friction coefficient opposing the motion of the particles as a result of the
viscosity of the fluid. For spherical particles of radius $a$ much larger than molecular dimensions, moving slowly through a fluid with viscosity $\eta$, with the condition that the fluid in contact with the particle is at rest with respect to the particle, G G Stokes showed in 1851 that $\Gamma = 6\pi \eta a$. The resulting current due to the force $F$ would then be $j_F = -n(z)F/\Gamma = -n(z)(Mg - W)/\Gamma \equiv j_{grav}$ for gravity. At equilibrium, when Brownian motion is included, $j_{diff}$ from (23) and $j_F$ should balance, i.e.,

$$\frac{nF}{\Gamma} = -D \frac{dn}{dz}.$$  \hfill (24)

giving a steady, $z$-dependent profile for the concentration (see Figure 5).

We can derive an almost identical condition in a rather different way, without ever looking at the underlying dynamical processes; and this will finally give us the vital relation between $D$ and the friction $\Gamma$. Instead of balancing currents, let us balance the forces on the Brownian particles in the steady state arising from the interplay of gravity and Brownian motion. Treat the combined system of Brownian particles and molecules as a mixture of gases. Then the partial pressure $^6$ of the Brownian particles (known as the osmotic pressure), now a function of the vertical coordinate $z$, is simply $p_{osm}(z) = n(z)k_B T$. Since the osmotic pressure varies with depth, the difference between the osmotic pressures at the top and the bottom of a layer of particles at height $z$ and thickness $dz$ results in a force $-(dp_{osm}/dz)dz = -k_B T (dn/dz)dz$ on the layer. Thus the ‘osmotic force’ per unit volume is $f_{osm} = -k_B T dn/dz$. This is balanced by the gravitational force per unit volume, giving

$$n(z)F = n(z)(Mg - W) = -k_B T \frac{dn}{dz}.$$  \hfill (25)

---

$^6$ A concept that you have seen in B Sc physics, if not earlier.
Comparing (24) and (25) leads us to Einstein’s famous result

\[ D = \frac{k_B T}{\Gamma} = \frac{k_B T}{6\pi \eta a} \]  

(26)

where the second equality comes from using Stokes’ estimate for the friction. This is the result promised in section 1.2, which lets us extract \( k_B \) from a measurement of diffusion.

Comparing (24) to our earlier expression (16) for \( D \) with dimension \( q = 3 \), the step time \( \tau \) turns out to be \( M/3\pi \eta a \) for a Brownian particle with mass \( M \). For particles with \( a = 1\mu m \) in water, and with a density comparable to that of water, this yields \( \tau \approx 0.4\mu s \). This makes it amply clear that you cannot measure the instantaneous speed of the Brownian particle by observations through a microscope, a fact unclear to some early workers.

The Stokes–Einstein relation (26) connects \( D \), a property of fluctuations, to the viscosity, a measure of friction or dissipation. If you go on to study statistical physics you will find many such ‘fluctuation-dissipation’ relations. The physics underlying all these is the same: friction and thermal fluctuations are engendered by the same processes, at least in systems at thermal equilibrium.

### 3.4 A Test-tube Atmosphere

Einstein also used the relation (25) directly to get \( k_B \). Solving (25) gives a depth-dependent concentration

\[ n(z) = n(0)e^{-Fz/k_B T}. \]  

(27)

Thus the concentration of suspended particles in a test-tube (see Figure 5), provided the particles have a higher mass density than the fluid, obeys the same law as the density of air in the atmosphere (neglecting the variation of gravity with height).

### 3.5 Experiments

Einstein used his theory along with existing data on the diffusion (the spread of an initially concentrated region of sugar) in water to get a value for \( k_B \). Since he couldn’t see the sugar molecules, he needed one more piece of information from which he could infer their size. This he did by combining measurements of the diffusivity with those of the viscosity of the solution as a function of concentration. For this he had to calculate the change in viscosity of a fluid when small particles were added to it. He found that the fractional change, to first order in the concentration, was 2.5 times the volume fraction of particles added, a result still used by chemical engineers. The trouble with using this approach for the diffusion of molecules is that the Stokes result and Einstein’s calculation of the excess viscosity are for a particle large enough that the fluid can be treated as a continuum around it, an assumption not obviously valid for the flow of water around sugar. Although on dimensional grounds a law like (26)

\[ D = \frac{k_B T}{\Gamma} = \frac{k_B T}{6\pi \eta a} \]  

(26)

This prompted Perrin to remark in ch.IV of his book [1] that the paths of Brownian particles are like the continuous but nowhere differentiable functions that mathematicians discuss and that this means that such objects are far from being mere mathematical curiosities. In the preface to his book, he makes similar remarks about the surfaces of aggregates of colloidal particles. These ideas are clearly the precursors of the notion of a fractal [4].
might hold, there is no reason to trust the $6\pi$. In fact, if we assume the fluid slips a bit around the particles, smaller numbers such as $3\pi$ are possible, which might explain why Einstein’s estimate for Avogadro’s number from this approach is small by a factor of two.

Perrin observed the concentration as a function of height in a suspension of resin particles, by actually counting the number of particles seen in a small reference volume at various heights and verified the law (27). He also measured the diffusion coefficient and used it to give $k_B$ using the Stokes–Einstein relation. The values he got by the two methods were quite consistent with each other and with modern values. They agreed, moreover, with those obtained from the kinetic theory estimates I mentioned earlier. The existence of molecules was thus firmly established.

4. Conclusion

This article was meant not only to communicate a classic piece of physics but to make the point that physics is about the things we see around us. I hope I’ve convinced the reader of the importance of keeping one’s eyes open – especially when looking through a microscope.

Meanwhile, the subject that Einstein invented before he got so busy with relativity and quantum mechanics lives on, in colloid science – the world of particles much bigger than a molecule but much smaller than a mustard seed, statistical physics, mathematics, finance, the working of motor proteins in cells, the analysis of DNA sequences, and so forth. The random walk is the simplest model for a polymer chain, and the physics of rubber elasticity can be understood from such a model. The mathematics of random walks is at the base of one formulation of quantum mechanics.

The bibliography at the end of this article consists mainly of classic treatments of the subject. To learn more about what people are doing with Brownian motion today, try searching for it on the Web and then looking at some of the sites in detail. My attempt, through a popular search engine, turned up 4732 sites! Some of these were for physics courses, but many were from wildly diverse areas, which will give you an idea of the extraordinary reach of the subject.

Acknowledgements

I thank Rama Govindarajan, Rajaram Nityananda and Ajay Sood for useful suggestions and Ashwin Pande, Rahul Siddharthan and Chinmay Das for timely help with figures.

Suggested Reading


[6] A Kumar et al., *Physics: A textbook for senior secondary classes*, NCERT, New Delhi, 1988, Chapter 9 and the accompanying exercises are a lucid and entertaining introduction to atoms, molecules, and Brownian motion.


The Foucault pendulum is an elegant device that demonstrates the rotation of the Earth. After describing it, we will elaborate on an interesting geometrical relationship between the dynamics of the Foucault pendulum and Thomas precession discussed in the last installment\textsuperscript{1}. This will help us to understand both phenomena better.

The first titular president of the French republic, Louis-Napoleon Bonaparte, permitted Foucault to use the Pantheon in Paris to give a demonstration of his pendulum (with a 67 meter wire and a 28 kg pendulum bob) on 31 March 1851. In this impressive experiment, one could see the plane of oscillation of the pendulum rotating in a clockwise direction (when viewed from the top) with a frequency $\omega = \Omega \cos \theta$, where $\Omega$ is the angular frequency of Earth’s rotation and $\theta$ is the co-latitude of Paris. (That is, $\theta$ is the standard polar angle in spherical polar coordinates with the $z$-axis being the axis of rotation of Earth. So $\pi/2 - \theta$ is the geographical latitude). Foucault claimed, quite correctly, that this effect arises due to the rotation of the Earth and thus showed that one can demonstrate the rotation of the Earth by an \textit{in situ} experiment without looking at celestial objects.

This result is quite easy to understand if the experiment was performed at the poles or equator (instead of Paris!). The situation at the North Pole is as shown in Figure 1. Here we see the Earth as rotating (from west to east, in the counter-clockwise direction when viewed from the top) underneath the pendulum, making one full turn in 24 hours. It seems reasonable to deduce from this that, as viewed from Earth, the plane of oscillation of the pendulum will make one full rotation in 24 hours; so the angular frequency $\omega$ of the rotation of the plane of the Foucault pendulum is just $\omega = \Omega$. (Throughout the discussion we are concerned with the rotation of the plane of oscillation of the pendulum; not the period of the pendulum $2\pi/\nu$, which – of course – is given by the standard formula involving the length of the suspension wire, etc.). At the equator, on the other hand, the plane of oscillation does not rotate. So the formula, $\omega = \Omega \cos \theta$, captures both the results correctly.

It is trivial to write down the equations of motion for the pendulum bob in the rotating frame of the Earth and solve them to obtain this result [1, 2] at the linear order in $\Omega$. Essentially, the Foucault pendulum effect arises due to the Coriolis force in the rotating frame of the Earth which leads to an acceleration $2v \times \Omega$, where $v$, the velocity of the pendulum bob, is directed tangential to the Earth’s surface to a good approximation.

---

Jean Bernard Léon Foucault (1819–1868) was a French physicist, famous for the demonstration of Earth’s rotation with his pendulum. Although Earth’s rotation was not unknown then, this easy-to-see experiment caught everyone’s imagination.

If we choose a local coordinate system with the $Z$-axis pointing normal to the surface of the Earth and the $X, Y$ coordinates in the tangent plane at the location, then it is easy to show that the equations of motion for the pendulum bob are well approximated by

$$
\ddot{X} + \nu^2 X = 2\Omega_z \dot{Y}; \quad \ddot{Y} + \nu^2 Y = -2\Omega_z \dot{X},
$$

where $\nu$ is the period of oscillation of the pendulum and $\Omega_z = \Omega \cos \theta$ is the normal component of the Earth’s angular velocity. In arriving at these equations we have ignored the terms quadratic in $\Omega$ and the vertical displacement of the pendulum. The solution to this equation is obtained easily by introducing the variable $q(t) \equiv \dot{X}(t) + i\dot{Y}(t)$. This satisfies the equation

$$
\ddot{q} + 2i\Omega_z \dot{q} + \nu^2 q = 0.
$$

The solution, to the order of accuracy we are working with, is given by

$$
q = X(t) + iY(t) = (X_0(t) + iY_0(t)) \exp(-i\Omega_z t),
$$

where $X_0(t), Y_0(t)$ is the trajectory of the pendulum in the absence of Earth’s rotation.

It is clear that the net effect of rotation is to cause a shift in the plane of rotation at the rate $\Omega_z = \Omega \cos \theta$. Based on this knowledge and the results for the pole and the equator one can give a ‘plain English’ derivation of the result for intermediate latitudes by saying something like: “Obviously, it is the component of $\Omega$ normal to the Earth at the location of the pendulum which matters and hence $\omega = \Omega \cos \theta$.”

The first-principle approach, based on (1), of course has the advantage of being rigorous and algorithmic; for example, if you want to take into account the effects of ellipticity of Earth, you can do that if you work with the equations of motion. But it does not give you an intuitive understanding of what is going on, and much less a unified view of all related problems having the same structure. We shall now describe an approach to this problem which has the advantage of providing a clear geometrical picture and connecting it up – somewhat quite surprisingly – with Thomas precession discussed in the last installment.

One point which causes some confusion as regards the Foucault pendulum is the following. While analyzing the behavior of the pendulum at the pole, one assumes that the plane of rotation remains fixed while the Earth rotates underneath it. If we make the same claim for a pendulum experiment done at an intermediate latitude, – i.e., if we say that the plane of oscillation remains invariant with respect to, say, the “fixed stars” and the Earth rotates underneath it – it seems natural that the period of rotation of the pendulum plane should always be 24 hours irrespective of the location! This, of course, is not true and it is also intuitively obvious that nothing happens to the plane of rotation at the equator. In this way of approaching the problem, it is not very clear how exactly the Earth’s rotation influences the motion of the pendulum.

To provide a geometrical approach to this problem, we will rephrase it as follows [3, 4]. The plane of oscillation of the pendulum can be characterized by a vector
normal to it or equivalently by a vector which is lying in the plane and tangential to the Earth’s surface. Let us now introduce a cone which is coaxial with the axis of rotation of the Earth and with its surface tangential to the Earth at the latitude of the pendulum (see Figure 2). The base radius of such a cone will be $R \sin \theta$, where $R$ is the radius of the Earth and the slant height of the cone will be $R \tan \theta$. Such a cone can be built out of a sector of a circle (as shown in Figure 3) having the circumference $2\pi R \sin \theta$ and radius $R \tan \theta$ by identifying the lines OA and OB. The ‘deficit angles’ of the cone, $\alpha$ and $\beta \equiv 2\pi - \alpha$, satisfy the relations:

\[(2\pi - \alpha)R \tan \theta = 2\pi R \sin \theta \quad \text{(4)}\]

which gives

\[\alpha = 2\pi(1 - \cos \theta); \quad \beta = 2\pi \cos \theta. \quad \text{(5)}\]

The behavior of the plane of the Foucault pendulum can be understood very easily in terms of this cone. Initially, the Foucault pendulum is started out oscillating in some arbitrary direction at the point A, say. The direction of oscillation can be indicated by some straight line drawn along the surface of the cone (like AC in Figure 3). While the plane of oscillation of the pendulum will rotate with respect to a coordinate system fixed on the Earth, it will always coincide with the lines drawn on the cone which remain fixed relative to the fixed stars. When the Earth makes one rotation, we move from A to B in the flattened out cone in Figure 3. Physically, of course, we identify the two points A and B with the same location on the surface of the Earth. But when a vector is moved around a curve along the lines described above, on the curved surface of Earth, its orientation does not return to the original value. It is obvious from Figure 3 that the orientation of the plane of rotation (indicated by a vector in the plane of rotation and tangential to the Earth’s surface at B) will be different from the corresponding vector at A. (This process is called parallel transport and the fact that a vector changes on parallel transport around an arbitrary closed curve on a curved surface is a well-known result in differential geometry and general relativity.)

Clearly, the orientation of the vector changes by an angle $\beta = 2\pi \cos \theta$ during one rotation of Earth with period $T$. Since the rate of change is uniform throughout because of the steady state nature of the problem, the angular velocity of the rotation of the pendulum plane is given by

\[\omega = \frac{\beta}{T} = \frac{2\pi}{T} \cos \theta = \Omega \cos \theta. \quad \text{(6)}\]

This is precisely the result we were after. The key geometrical idea was to relate the rotation of the plane of the Foucault pendulum to the parallel transport of a vector characterizing the plane, around a closed curve on the surface of Earth. When this closed curve is not a geodesic – and we know that a curve of constant latitude is not a geodesic – the orientation of this vector changes when it completes one loop. There are more sophisticated ways of calculating how much the orientation changes for a given curve on a curved surface. But in the case of a sphere, the trick of an enveloping
The plane of oscillation of the pendulum will rotate with respect to a coordinate system fixed on the Earth, but it will always coincide with the lines drawn on the cone which remain fixed relative to the fixed stars. (Figures 2,3)

This derivation also allows one to understand the Thomas precession of the spin of a particle.

cone provides a simple procedure. (When the pendulum is located in the equator, the closed curve is the equator itself; this, being a great circle is a geodesic on the sphere and the vector does not get ‘disoriented’ on going around it. So the plane of the pendulum does not rotate in this case.)

This is good, but as I said, things get better. One can show that an almost identical approach allows one to determine the Thomas precession of the spin of a particle (say, an electron) moving in a circular orbit around a nucleus [5].

We saw in the last installment [6] that the rate of Thomas precession is given, in general, by an expression of the form

$$\omega dt = (\cosh \chi - 1) \left( d\hat{n} \times \hat{n} \right),$$

(7)

where \( \tanh \chi = v/c \) and \( v \) is the velocity of the particle. In the case of a particle moving on a circular trajectory, the magnitude of the velocity remains constant and we can integrate this expression to obtain the net angle of precession during one orbit. For a circular orbit, \( d\hat{n} \) is always perpendicular to \( \hat{n} \) so that \( \hat{n} \times d\hat{n} \) is essentially \( \hat{n} \) which integrates to give a factor \( 2\pi \). Hence the net angle of Thomas precession during one orbit is given by

$$\Phi = 2\pi (\cosh \chi - 1).$$

(8)

The similarity between the net angle of turn of the Foucault pendulum and the net Thomas precession angle is now obvious when we compare (8) with (5). We know that in the case of Lorentz transformations, one replaces real angles by imaginary angles which accounts for the difference between the cos and cosh factors. What we need to do is to make this analogy mathematically precise which will be our next task. It will turn out that the sphere and the cone we introduced in the real space, to study the Foucault pendulum, have to be introduced in the velocity space to analyze Thomas precession.

As a warm-up to exploring the relativistic velocity space, let us start by asking the following question: Consider two frames \( S_1 \) and \( S_2 \) which move with velocities \( v_1 \) and \( v_2 \) with respect to a third inertial frame \( S_0 \). What is the magnitude of the relative velocity between the two frames? This is most easily done using Lorentz invariance and four-vectors (and to simplify notation we will use units with \( c = 1 \)). We can associate with the 3-velocities \( v_1 \) and \( v_2 \), the corresponding four-velocities, given by

\[ u_1' = (\gamma_1, \gamma_1 v_1) \]

and

\[ u_2' = (\gamma_2, \gamma_2 v_2) \]

with all the components being measured in \( S_0 \).

On the other hand, with respect to \( S_1 \), this four-vector will have the components

\[ u_1'' = (1, 0) \]

and

\[ u_2'' = (\gamma, \gamma v), \]

where \( v \) (by definition) is the relative velocity between the frames. To determine the magnitude of this quantity, we note that in this frame \( S_1 \) we can write \( \gamma = -u_1' u_2'' \). But since this expression is Lorentz invariant, we can evaluate it in any inertial frame. In \( S_0 \), with \( u_1' = (\gamma_1, \gamma_1 v_1) \), \( u_2' = (\gamma_2, \gamma_2 v_2) \) this has the value

$$\gamma = (1 - v^2)^{-1/2} = \gamma_1 \gamma_2 - \gamma_1 \gamma_2 v_1 \cdot v_2.$$
Simplifying this expression we get

$$v^2 = \frac{(1 - v_1 \cdot v_2)^2 - (1 - v_1^2)(1 - v_2^2)}{(1 - v_1 \cdot v_2)^2}$$

$$= \frac{(v_1 - v_2)^2 - (v_1 \times v_2)^2}{(1 - v_1 \cdot v_2)^2}.$$  \hspace{1cm} (10)

Let us next consider a 3-dimensional abstract space in which each point represents a velocity of a Lorentz frame measured with respect to some fiducial frame. We are interested in defining the notion of ‘distance’ between two points in this velocity space. Consider two nearby points which correspond to velocities \(v\) and \(v + dv\) that differ by an infinitesimal quantity. By analogy with the usual 3-dimensional flat space, one would have assumed that the ‘distance’ between these two points is just

$$|dv|^2 = dv_x^2 + dv_y^2 + dv_z^2 = dv^2 + v^2(d\theta^2 + \sin^2 \theta d\phi^2),$$  \hspace{1cm} (11)

where \(v = |v|\) and \((\theta, \phi)\) denote the direction of \(v\). In non-relativistic physics, this distance also corresponds to the magnitude of the relative velocity between the two frames. However, we have just seen that the relative velocity between two frames in relativistic mechanics is different and given by (10). It is more natural to define the distance between the two points in the velocity space to be the relative velocity between the respective frames. In that case, the infinitesimal ‘distance’ between the two points in the velocity space will be given by (10) with \(v_1 = v\) and \(v_2 = v + dv\). So

$$dl_\nu^2 = \frac{(dv)^2 - (v \times dv)^2}{(1 - v^2)^2}.$$  \hspace{1cm} (12)

Using the relations

$$(v \times dv)^2 = v^2(dv)^2 - (v \cdot dv)^2; \quad (v \cdot dv)^2 = v^2(dv)^2$$  \hspace{1cm} (13)

and using (11) where \(\theta, \phi\) are the polar and azimuthal angles of the direction of \(v\), we get

$$dl_\nu^2 = \frac{dv^2}{(1 - v^2)^2} + \frac{v^2}{1 - v^2}(d\theta^2 + \sin^2 \theta d\phi^2).$$  \hspace{1cm} (14)

If we use the rapidity \(\chi\) in place of \(v\) through the equation \(v = \tanh \chi\), the line element in (14) becomes:

$$dl_\nu^2 = d\chi^2 + \sinh^2 \chi (d\theta^2 + \sin^2 \theta d\phi^2).$$  \hspace{1cm} (15)

This is an example of a curved space within the context of special relativity. This particular space is called (three-dimensional) Lobachevsky space.

If we now change from real angles to the imaginary ones, by writing \(\chi = i\eta\), the line element becomes

$$-dl_\nu^2 = d\eta^2 + \sin^2 \eta (d\theta^2 + \sin^2 \theta d\phi^2),$$  \hspace{1cm} (16)

which (except for an overall sign which is irrelevant) represents the distances on a 3-sphere having the three angles \(\eta, \theta\) and \(\phi\) as its coordinates.
Of these three angles, $\theta$ and $\phi$ denote the direction of velocity in the real space as well. When a particle moves in the $x-y$ plane in the real space, its velocity vector lies in the $\theta = \pi/2$ plane and the relevant part of the metric reduces to

$$dL^2 = d\eta^2 + \sin^2 \eta d\phi^2$$

which is just a metric on the 2-sphere. Further, if the particle is moving on a circular orbit with constant magnitude for the velocity, then it follows a curve of $\eta =$ constant on this 2-sphere. The analogy with the Foucault pendulum, which moves on a constant latitude curve, is now complete. If the particle carries a spin, the orbit will transport the spin vector along this circular orbit. As we have seen earlier, the orientation of the vector will not coincide with the original one when the orbit is completed and we expect a difference of $2\pi(1 - \cos \eta) = 2\pi(1 - \cosh \chi)$. So the magnitude of the Thomas precession, over one period is given precisely by (8). I will let you work out the details exactly in analogy with the Foucault pendulum and convince yourself that they have the same geometrical interpretation.

When one moves along a curve in the velocity space, one is sampling different (instantaneously) co-moving Lorentz frames obtained by Lorentz boosts along different directions. As we described in the last installment, Lorentz boosts along different directions do not, in general, commute. This leads to the result that if we move along a closed curve in the velocity space (treated as representing different Lorentz boosts) the orientation of the spatial axes would have changed when we complete the loop.

It turns out that the ideas described above are actually of far more general validity. Whenever a vector is transported around a closed curve on the surface of a sphere, the net change in its orientation can be related to the solid angle subtended by the area enclosed by the curve. In the case of the Foucault pendulum, the relevant vector describes the orientation of the plane of the pendulum and the transport is around a circle on the surface of the Earth. In the case of Thomas precession, the relevant vector is the spin of the particle and the transport occurs in the velocity space. Ultimately, both the effects – the Foucault pendulum and Thomas precession – arise because the space in which one is parallel transporting the vector (surface of Earth, relativistic velocity space) is curved.

Suggested Reading

When the small angle approximation is not made, the exact solution of the simple pendulum is a Jacobian elliptic function with one real and one imaginary period. Far from being a physically meaningless mathematical curiosity, the second period represents the imaginary time the pendulum takes to swing upwards and tunnel through the potential barrier in the semi-classical WKB approximation\(^1\) in quantum mechanics. The tunneling here provides a simple illustration of similar phenomena in Yang–Mills theories describing weak and strong interactions.

1. Introduction

Consider a point mass \(m\) at the end of a rigid massless stick of length \(l\). The acceleration due to gravity is \(g\) and the oscillatory motion is confined to a vertical plane. Denoting the angular displacement of the stick from the vertical by \(\phi\) and taking the gravitational potential to be zero at the bottom, the constant total energy \(E\) is given by

\[
E = \frac{1}{2} ml^2 \left( \frac{d\phi}{dt} \right)^2 + mgl(1 - \cos \phi)
= mgl(1 - \cos \phi_0),
\]

(1)

where \(\phi_0\) is the maximum angle. Isolating \((d\phi/dt)^2\) and taking its square root gives \(\phi\) as an implicit function of the time \(t\) through

\[
t = \sqrt{\frac{l^2}{2g} \int_0^{\phi} \frac{d\phi}{\cos \phi - \cos \phi_0}^{1/2}}
= \sqrt{\frac{l^2}{3g} \int_0^{\phi} \frac{d\phi}{\left(\sin^2(\phi_0/2) - \sin^2(\phi/2)\right)^{1/2}}}. \tag{2}
\]

Following the treatment in Landau–Lifshitz [1], we define \(\sin \xi = \sin(\phi/2)/\sin(\phi_0/2)\).

\(^{1}\) When the exact quantum mechanical calculation of the tunneling probability for an arbitrary potential \(U(x)\) is impracticable, the Wentzel–Kramers–Brillouin approach, invented independently by all three authors in the same year as the Schrödinger equation, provides an approximate answer.

**Keywords**

Pendulum, Jacobi elliptic functions, tunneling, WKB, instantons.
in terms of which the period $T$ becomes

$$T = 2\sqrt{\frac{l}{g}} \int_{\phi_0}^{\frac{\pi}{2}} d\phi \left( \frac{\sin^2(\phi_0/2) - \sin^2(\phi/2)}{\sin^2(\phi_0/2) - \sin^2(\phi/2)} \right)^{1/2}$$

$$= 4\sqrt{\frac{l}{g}} \int_{0}^{\frac{\pi}{2}} \frac{d\xi}{(1 - k^2 \sin^2 \xi)^{1/2}} .$$

This can be used to define $K(k)$, the complete elliptic integral of the first kind, via

$$T \equiv 4\sqrt{\frac{l}{g}} K(k) .$$

When $k^2 = (\sin(\phi_0/2))^2$ can be neglected, one recovers the period $2\pi \sqrt{l/g}$ of the simple small angle pendulum. Note that the standard elliptic function notation $K(k)$ is a bit misleading: $K$ is really a function NOT of $k$, but of $k^2$. Hence for $k^2 \ll 1$, we can consistently keep $k$ as a small non-zero quantity while neglecting $k^2$. The substitution $y = \sin \xi$ shows that (3) can also be written as

$$K(k) = \int_{0}^{\frac{\pi}{2}} \frac{d\xi}{(1 - k^2 \sin^2 \xi)^{1/2}}$$

$$= \int_{0}^{1} \frac{dy}{((1 - y^2)(1 - k^2 y^2))^{1/2}} .$$

2. Elliptic Functions and Integrals

Let us convert the definite integral (5) into an indefinite one by replacing the upper limit 1 by $y$. Also introducing the dimensionless time variable $x = t \sqrt{g/l}$, this yields

$$x = \int_{0}^{y} \frac{dy'}{((1 - y'^2)(1 - k^2 y'^2))^{1/2}} .$$

The right-hand side of (6) is known as the Legendre elliptic integral of the first kind and is denoted by $F(\arcsin y, k)$. The inverse of $F$, implicitly defined by (6), is the Jacobian elliptic function $y = sn x [2]$. Assembling all this, the exact time dependence of the angle is found to be

$$\phi(t) = 2 \arcsin \left( \sin \left( \sqrt{\frac{g}{l}} t \right) \sin \frac{\phi_0}{2} \right) ,$$

which becomes the familiar

$$\phi(t) = \phi_0 \sin \left( \sqrt{\frac{g}{l}} t \right)$$

in the limit $\sin \phi_0 \approx \phi_0$. 

---

320
Box 1. Elliptic Functions

A function \( f(z) \) is said to be doubly periodic in the complex \( z \)-plane with periods \( \omega_1 \) and \( \omega_2 \) if \( f(z + n \omega_1 + m \omega_2) = f(z) \) for any pair of integers \( n, m \). Double periodicity along two independent directions requires the ratio \( \omega_1/\omega_2 \) to have an imaginary part. Thus the function maps a single parallelogram cell \( \Pi \) defined by \( \omega_1 \) and \( \omega_2 \) to the entire \( z \)-plane. If the singularities of \( f(z) \) consist of poles, the resulting meromorphic doubly-periodic function is called an elliptic function. By Liouville’s theorem, \( f(z) \) must have poles in \( \Pi \) if it is not to be a constant. Since the parallel sides of the parallelogram \( \partial \Pi \) bounding \( \Pi \) give equal and opposite contributions, and \( f(z) \) is meromorphic, \( \oint_{\partial \Pi} f(z) \, dz = 0 \) by Cauchy’s theorem. This means the residues of the poles in \( \Pi \) must add up to zero. The simplest way to get vanishing total residue is either to choose two simple poles with opposite residues, or one double pole in \( \Pi \). These correspond to the Jacobi and the Weierstrass elliptic functions, respectively. The latter, denoted by \( \wp \), is easier to represent explicitly (in the form of an infinite double sum over all lattice points), but our present problem of a planar pendulum involves the Jacobi elliptic functions (interestingly, the Weierstrass ones appear in the treatment of the spherical pendulum!). The lattice is generated by \( \omega_1 = K \) and \( \omega_2 = iK(k’) \) that are defined in (5) and (9). The double periodicity properties follow by integrating (6) in the complex \( y \)-plane along different contours: \( \text{sn}(2K - y) = \text{sny} \) is proven by choosing a path that goes from 0 to \( y \) to 1 along the real axis, encircles 1 in a clockwise sense and comes back to \( y \) on a path going left just below the real axis. Using \( \text{sn}(-y) = -\text{sny} \) which follows from (6), we have \( \text{sn}(2K + y) = -\text{sny} \). Shifting the argument again by another \( 2K \), we get \( \text{sn}(4K + y) = +\text{sny} \). Extending the contour from 1 to \( 1/k \), clockwise encircling \( 1/k \) and coming back to \( y \) on a path just below the real axis yields \( \text{sn}(y + 2ik’') = \text{sny} \) on account of (9) and the fact that for \( 1 < x < 1/k \), \( \sqrt{1 - \omega^2} = \pm i\sqrt{\omega^2 - 1} \).

Historically, the study of elliptic functions originated around the end of the seventeenth century in the study of arc lengths of ellipses and a curve called Bernoulli’s lemniscate, leading to the integral \( \int_0^\alpha dx/\sqrt{1-x^4} \). Count Fagnano presented an addition theorem for this integral in 1750. Euler extended the theorem to integrals with \( \sqrt{1 + ax^2 + bx^4} \) in the denominator. Legendre’s studies of more general integrands such as \( E + F/\sqrt{Q(x)} \), where \( E \) and \( F \) are rational functions and \( Q \) is a cubic or quartic polynomial culminated in his 1832 Treatise on Elliptic functions and Euler integrals. In 1828 Abel and Jacobi almost simultaneously defined new functions by inverting such integrals and showed that they were doubly-periodic. The terminology was also changed: these functions are now called elliptic, while Legendre’s elliptic functions are referred to as elliptic integrals. Jacobi’s elliptic functions can also be expressed as the ratios of quadratic combinations of Jacobi \( \Theta \)-functions that are quasiperiodic rather than periodic in the complex plane. The so-called Ramanujan \( \Theta \)-function is a further generalization of Jacobi \( \Theta \)-functions, but it is not due to Ramanujan himself. Ramanujan introduced what he called “mock \( \Theta \)-functions”, which are nowadays regarded as a class of ‘modular forms’, defined by their transformation properties under modular transformations of the complex plane. Such a transformation takes \( z \) to \( (az + b)/(cz + d) \), where \( a, b, c, d \) are integers and \( ab - cd = 1 \). For more on Ramanujan’s remarkable work in this area we refer the reader to the book Ramanujan’s lost notebook, George E Andrews and Bruce C Berndt, Part I, Springer, New York, 2005.
This is an appropriate point to mention an essential difference between trigonometric and elliptic functions: Considered as functions of the complex variable \( z = x + iy \), the former are bounded and have a real period along the \( x \)-axis, but blow up exponentially as \( y \) goes to positive or negative infinity, while the latter have an additional imaginary period along the \( y \)-axis. We will have to state this and a few other properties of elliptic functions without proof, as the required complex analysis is a bit involved. We refer the interested reader to Mathews and Walker [2], whose treatment we mostly follow.

For \( \text{sn} \), we have already met the real period \( K(k) \) in (4); the imaginary period \( 2K' \) is defined via

\[
K' = \int_{1}^{1/k} \frac{dy}{((y^2 - 1)(1 - k^2 y^2))^{1/2}}. \tag{9}
\]

The double-periodicity property means that

\[
\text{sn}(z + 4K) = \text{sn}(z + 2iK') = \text{sn}z. \tag{10}
\]

It can be shown that \( K'(k) = K(k') \) if \( k'^2 = 1 - k^2 \) by going over to the variable \( z = \sqrt{(1 - k^2 y^2)/(1 - k^2)} \) in (9). This ‘duality relation’ will be useful in relating the pendulum’s behavior near the top and bottom points.

### 3. The Physical Meaning of the Imaginary Period

We will now show that the imaginary period \( K' \) coincides with the interval of imaginary time the pendulum spends while ‘swinging’ from \( +\phi_0 \) to \( -\phi_0 \) not via the bottom point \( \phi = 0 \) as in its normal motion, but by continuing beyond \( +\phi_0 \) to the top point \( \phi = \pi \) and all the way down in the same direction to \( \phi = 2\pi - \phi_0 \), which is of course the same position in space as \( -\phi_0 \). During this ‘motion’, the total energy \( mgl(1 - \cos \phi) \) is below the potential energy \( mgl(1 - \cos \phi_0) \), making the kinetic energy \( \frac{1}{2}ml^2(\frac{d\phi}{dt})^2 \) negative; hence the time \( t \) may be regarded as imaginary. Although we started with a familiar classical mechanics problem, this is clearly reminiscent of quantum tunneling, which raises the question of what an \textit{ab initio} WKB-like treatment of the problem without referring to elliptic functions would reveal.

Specifically, we wish to calculate the imaginary time \( T' \) that the mass takes to cover the above described path. To do this, all we have to do is to change the limits in the integral (2), giving

\[
T' = \int_{\phi_0}^{2\pi - \phi_0} \frac{d\phi}{(d\phi/dt)} = 2\sqrt{\frac{l}{g}} \int_{\phi_0}^{\pi} \frac{d\phi}{(\sin^2(\phi_0/2) - \sin^2(\phi/2))^{1/2}}, \tag{11}
\]

where in the last equation we have used the symmetry of the potential around \( \phi = \pi \). In the new integration range \( \pi \geq \phi \geq \phi_0 \), \( \sin \xi = \sin(\phi/2)/\sin(\phi_0/2) \geq 1 \). Setting
The real and imaginary 'paths' are most clearly understood by referring to the potential. The solid line represents the real time oscillations, while the dotted one corresponds to imaginary time tunneling.

\[ y = \sin \xi \text{ as before gives} \]

\[ T' = 2i \sqrt{\frac{l}{g}} \int_1^{1/k} \frac{dy}{\sqrt{(y^2 - 1)(1 - k^2 y^2)}} = 2i \sqrt{\frac{l}{g}} K'(k) . \]  

We note that \( i = \sqrt{-1} \) and the \( y \)-direction period \( K'(k) \) given in (9) naturally appear! The second, purely imaginary period of the Jacobi elliptic function indeed coincides with the WKB tunneling time.

4. Instantons, Small Oscillations, Duality

The duality relation \( K'(k) = K(k') = K(\sqrt{1 - k'^2}) \) mentioned earlier allows us to identify and interrelate some interesting limiting cases in the pendulum's behavior. Our problem is non-linear, and it happens to share some important aspects of the also non-linear Yang–Mills field theory. In the \( k^2 \approx 0 \) small-amplitude limit, both our pendulum and Yang–Mills fields can be treated perturbatively, meaning just a few lowest powers in the field amplitudes (\( \phi \) in our case) need be considered. An example in the pendulum problem is the first anharmonic term used in [3]. However, in both cases, the imaginary time behavior of the small or even zero amplitude limit contains rich non-perturbative phenomena related to the real time perturbative regime via duality. For the Yang–Mills case, we refer the reader to Huang [4]. In our problem, we start with the 'perturbative limit' \( k^2 = 0 \).

4.1 The \( k^2 = 0 \) Case

Since \( K(k^2 = 0) = \pi/2 \), we are in the small angle regime where \( T = 2\pi \sqrt{l/g} \).

Equation (9) then gives

\[ T' = 2i \sqrt{\frac{l}{g}} \int_1^{\infty} \frac{dy}{\sqrt{(y^2 - 1)}} = 2i \sqrt{\frac{l}{g}} K'(0) . \]  

In both cases, the imaginary time behavior of the small or even zero amplitude limit contains rich non-perturbative phenomena related to the real time perturbative regime via duality.
According to the ‘Standard Model’ of fundamental particle physics, strong, electromagnetic and weak interactions are mediated by gauge fields. The familiar electric and magnetic fields interact only with charges and currents but not directly with themselves, resulting in a linear theory. The Yang–Mills gauge fields used for describing strong and weak interactions, however, have self-interactions, and are fundamentally non-linear. The pendulum considered here is a surprisingly useful toy model for illustrating novel features (such as instantons) of Yang–Mills fields that arise from non-linearity. For our purposes, it will be sufficient to consider the simplest Yang–Mills gauge fields as electric and magnetic fields with an additional internal symmetry index $a$ that runs from 1 to 3; thus we now have $E_a$ and $B_a$ instead of the electric field $E$ and the magnetic field $B$. Just as the electromagnetic fields can be obtained from the potentials $A$ and $\Phi$ via $E = -\dot{A} - \nabla \Phi$ and $B = \nabla \times A$ (c, the speed of light is taken as unity here), the Yang–Mills fields are obtained from potentials $(A_a(r, t), \Phi_a(r, t))$, albeit in a more complicated and non-linear way. The potentials are the analogues of the ‘coordinate’ $\phi(t)$ and the ‘electric fields’ $E_a$ the analogues of the ‘velocity’ $\dot{\phi}$. The field energy density is proportional to $E_a \cdot E_a + B_a \cdot B_a$, where, pursuing the analogy, the first and second terms can be identified as ‘kinetic’ and ‘potential’ energies, respectively (the Einstein convention of summing over the repeated index $a$ is understood). When one switches to imaginary time to investigate tunneling, $E_a \rightarrow iE_a$, and the energy density becomes $-E_a \cdot E_a + B_a \cdot B_a$. This allows non-vanishing field configurations obeying $E_a = \pm B_a$ with zero energy. Such fields are said to be self-dual or anti-self-dual, and the equation essentially says (square root of the kinetic energy) = ± (square root of the potential energy), just as in (15). In particular, there are (anti) self-dual fields that vanish over all space at $t = -\infty$, become non-zero for later times and vanish again at $t = +\infty$. Although the fields are zero at $t = \pm\infty$, the potentials need not be, and the field that interpolates between two topologically different potentials for vanishing fields in the infinite past and the infinite future is called a Yang–Mills instanton. There are infinitely many topologically distinct ‘vacua’ with vanishing fields, just like the pendulum’s minima at $\phi = 2\pi n, n = 0, \pm 1, \pm 2, \ldots$. A similar situation obtains for an electron in a periodic lattice potential with infinitely many different minima; because the electron can tunnel from one minimum to another, its ground state is a superposition of all the minima. This is an example of a Bloch wave.

The integral is elementary, resulting in $K'(0) = 2\text{arccosh}(\infty) = \infty$. This is of course the degeneration of an elliptic function to a trigonometric one as the imaginary period goes to infinity. The duality relation then shows that $K(1)$, the real oscillation period of a pendulum starting at the top unstable equilibrium point $\phi = \pi$, is also infinite. The reason is simply that the motion cannot start without an initial perturbation, however small.

Let us next examine the actual function $\phi(t)$ in imaginary time, setting the energy $E$ (and thus also $k$) strictly to zero so that the pendulum starts at $\phi = 0$ in the infinite past. After canceling out overall constants, the zero-energy condition becomes

$$-(\phi)^2 + (2g/l)(1 - \cos \phi) = 0,$$  

(14)

The real oscillation period of a pendulum starting at the top unstable equilibrium point $\phi = \pi$, is also infinite.
or, using half angles and taking square roots,

\[ \phi = \pm \frac{2g}{l} \sin \frac{\phi}{2}. \]  

(15)

Let us choose the plus sign. The solution will be a function of \( t - t_0 \), with \( t_0 \) an arbitrary constant which we set to zero, resulting in

\[ \phi(t) = 4 \arctan(\exp \omega t). \]  

(16)

This is an example of a topological instanton [5], a semiclassical solution that tunnels in imaginary time. It starts from \( t = -\infty \) at the original ground state \( \phi = 0 \), and ends up at the topologically distinct one \( \phi = 2\pi \) at \( t = +\infty \). The tunneling could also go in the opposite direction. This corresponds to taking the minus sign in (15) and might be called an anti-instanton. Thus the infinite period \( T' = \frac{2i}{\sqrt{gK}} \) found in (13) can be viewed as the tunneling time of the instanton, which is defined as a classical solution interpolating in imaginary time between two ground states. The ground state of unquantized Yang–Mills theory is approximated by a Bloch wave-like superposition of precisely such topologically inequivalent vacua connected by instantons, although the topological inequivalence of Yang–Mills vacua is not as intuitively graspable as our elementary example.

According to the Feynman path integral recipe the tunneling probability amplitude for the instanton is proportional to \( \exp(iS/\hbar) \), where \( S \) is the action for the entire path. The latter can be calculated exactly in imaginary time \( i\tau \). The result is a purely imaginary action, turning the amplitude to \( \exp(-|S|/\hbar) \). We start with

\[ |S| = | \int_{-\infty}^{+\infty} L(t) dt | = \int_{-\infty}^{+\infty} \left\{ \frac{1}{2} ml^2 \left( \frac{d\phi}{dt} \right)^2 + mgl (1 - \cos \phi) \right\} dt. \]  

(17)

Using (14), one can eliminate the \( \phi^2 \) term in favor of the potential energy, giving

\[ |S| = 2mgl \int_{-\infty}^{+\infty} (1 - \cos \phi) \, dt. \]  

(18)

Replacing \( dt \) by \( d\phi/(d\phi/dt) \), using half-angle formulas and remembering that \( \phi(-\infty) = 0 \) and \( \phi(+\infty) = 2\pi \), the integral becomes

\[ |S| = 2m \sqrt{gl^3} \int_0^{2\pi} \sin \frac{\phi}{2} \, d\phi = 8m \sqrt{gl^3}. \]  

(19)

Thus the instanton’s contribution to the action is \( \exp(-8m \sqrt{gl^3}/\hbar) \). It can be shown that the value of \( |S| \) in (19) is a minimum (other than the trivial one \( S = 0 \)) and therefore the instanton solution dominates the imaginary time Feynman path integral for the problem.
Box 3. The Action, WKB and Feynman’s Path Integral

Let us limit ourselves to one coordinate $x$ and the velocity $\dot{x}$. This is sufficient for the essential concepts, and our problem happens to be one-dimensional in any case. The Lagrangian is $L = m \dot{x}^2/2 - U(x)$, where $m$ is the mass and $U$ is the potential energy of a non-relativistic particle. The action $S$ is given by $\int_0^t L dt'$ and fixing the initial and final positions $x(0)$ and $x(t)$, one calculates an $S(t)$ for each conceivable path connecting these points. The classical path is found by extremizing $S$, which yields the equations of motion. In Feynman’s path integral formulation of quantum mechanics, the quantity of interest is the probability amplitude, denoted by $< x(t), t|x(0), 0 >$, which, by Feynman’s recipe, equals $A \Sigma \exp(iS[x(t)]/\hbar)$. The sum is over all paths and $A$ is a normalization constant. Each path thus weighs in with its phasor, i.e., a complex number of unit modulus. Since $S$ is stationary around the classical path, the phasors for nearby paths are nearly parallel. They add up to give a huge contribution, while the randomly oriented phasors from non-classical paths cancel each other out. For macroscopic objects, the angle $\alpha(S) = (S_cl - S)/\hbar$ represents a big deviation in the direction of the phasor from the classical; it will thus lead to destructive interference with other non-classical-path phasors. An elementary particle, on the other hand, can follow non-classical paths for which $\alpha(S)$ remains within a small angular range.

Of particular interest for our problem are tunneling paths where the momentum $p(x) = \sqrt{(E - U(x))/2m}$ is imaginary because $U > E$. When an exact quantum mechanical calculation of the tunneling probability for an arbitrary $U(x)$ is impracticable, the Wentzel–Kramers–Brillouin approach, invented independently by all three authors in the same year as the Schrödinger equation, provides an approximate answer. The method is based on adding up the phase changes $\exp(ip(x)dx/\hbar)$ as the approximate plane wave moves in steps of length $dx$. For a tunneling event with imaginary $p(x)$ from $x_1$ to $x_2$, this gives a transmission amplitude proportional to $\exp(-\int_{x_1}^{x_2} \sqrt{(E - U(x))/2m})dx$. The relation of this expression to an also semi-classically evaluated Feynman recipe follows from putting $\int L dt = \int (pdx/dt - E) dt = \int p dx - E \tau$, where $\tau$ is the transit time. Note that the semi-classical approach involves taking $E$ as constant during the passage and the first term is just the WKB expression. A detailed calculation (R. Shankar, *Principles of quantum mechanics*, second edition, Kluwer Academic/Plenum publishers, New York, 1994.) shows that the factor $< x_2(t), t|x_1(0), 0 >$ actually includes another factor of $+E \tau$ in the exponential, leaving only the WKB term.

4.2 The $k = 1$ Case

For completeness, let us also note that the result $K(0) = \pi/2$ for the small-angle simple pendulum implies $K(0) = K'(1) = \pi/2$ by the duality relation. Physically, this means that while the real time oscillations from $\phi_0 = -\pi$ to $\phi_0 = +\pi$ take forever, the (imaginary) tunneling time around the top point from $-\pi + \varepsilon$ to $\pi - \varepsilon$ is the same as the usual period of a small-angle pendulum.

If given a sufficiently high initial velocity, our pendulum can of course also complete full rotations in vertical circles in a finite amount of time (instead of swinging from $\phi_0 = -\pi$ to $\phi_0 = +\pi$ with the infinite period $K'(0)$). The period of the full rotational motion can be made arbitrarily small by increasing the initial energy, and therefore
cannot be identified with \( K \) or \( K' \) for any \( k = \sin(\phi_0/2) \). Another way of achieving full rotations is of course to drive the pendulum with an external motor. Interestingly, if the familiar value \( \omega = \sqrt{g/l} \) is chosen, there is neither compression nor tension in the rod at the topmost point.

5. Discussion: Tunneling and Imaginary Time in Nature

Quantum tunneling as a microscopic phenomenon is of course ubiquitous: The nitrogen atom tunnels back and forth across the equilateral triangle of hydrogen atoms in the ammonia (NH\(_3\)) molecule about 2.4\( \times 10^{10} \) times per second; alpha particles tunnel through the repulsive Coulomb wall in nuclei and get out; a DC current flows across a thin insulating Josephson junction between two superconductors, to name a few familiar examples. It is now even the basis of some high technology devices such as tunnel diodes and scanning-tunneling microscopes. What we tried to show here is that a very familiar classical mechanical system also reveals connections with quantum mechanics and non-perturbative phenomena in Yang–Mills theories in an exact mathematical treatment, but the actual probability that a macroscopic pendulum will exhibit quantum tunneling is of course fantastically small. For example, for \( m = 1 \) kg and \( l = 1 \) m, \( \exp(-8m\sqrt{gl^3}/\hbar) \) is of order \( \exp(-10^{35}) \)! An appreciable probability is only possible if the argument of the exponential is of order unity. In SI units, this requires \( l^{3/2} \sim (0.263 \times 10^{-34})/m \), which leaves a very narrow window for simultaneously physically meaningful values of \( m \) and \( l \). A length of 1 m leads to masses of the order of a billionth of an average atomic mass, while a mass of 1 kg produces a length nearly at the Planck scale of \( 10^{-34} \) m. On the other hand, if \( l = 1 \) nm, one finds \( m \sim 10^{-22} \) kg. These last numbers are intriguingly within a few orders of magnitude of current nano-technology applications [6] involving simple harmonic oscillators, but considering the weakness of gravity relative to the competing forces at such small scales, it is very unlikely that any evidence for the ‘imaginary time behavior’ of a simple pendulum will ever be seen in a controlled experiment.

According to Hawking [7] and Vilenkin [8], however, the universe itself, the ultimate uncontrolled experiment and the ultimate macroscopic entity, may have started in imaginary time and switched to our usual time ‘later’ (whatever ‘later’ may mean in this setting)! In the Vilenkin version, the imaginary time arises in connection with the universe tunneling out of a state of zero energy just like the instanton here; indeed, the process is described by something called the ‘de Sitter–Hawking–Moss instanton’. In the Hawking version, on the other hand, imaginary time is principally used to smooth out the Big Bang singularity. For more details and the debate about whether the approaches are equivalent, we refer the readers to Vilenkin’s book.

Suggested Reading

Mirrors and Merry-Go-Rounds *

Joseph Samuel

This is an elementary introduction to rotations in three dimensions, using reflections to naturally introduce spinors. It provides a stepping stone to higher mathematics and some new perspectives.

1. Introduction

If you look around you, you will see rotations everywhere. Wheels on cars, fans, spin on cricket balls, giant wheels, merry-go-rounds... Your head rotates when you look around. The Earth rotates and it’s a very good thing that it does! But for the rotation of the Earth, our days would be long indeed and our lives correspondingly short! Some of us would get roasted and the others frozen. The Sun rotates, as do black holes, the solar system, the galaxy and our local cluster of galaxies. Looking down in scale instead of up at the sky, some bacteria have rotary engines to propel themselves. Molecules rotate. Electrons and many other elementary particles have spin. Nuclei have spin and this leads to a life-saving medical probe: Magnetic Resonance Imaging (MRI).

Scientists encounter rotations in many areas of research. In physics we deal with the mechanics of rotations (of galaxies, stars, black holes, nuclei or elementary particles) and with the mathematics of rotations. Mechanics is a wide term including classical mechanics, quantum mechanics and statistical mechanics and rotation is studied in all three branches of mechanics. This article will focus on some mathematical aspects of rotations, trying to understand the ‘space of rotations’. As you will see, reflecting on rotations can make your head spin! Prerequisites for following this article are a familiarity with complex analysis, trigonometry, vector calculus, google to look up unfamiliar terms and most importantly some paper and pencils to supply missing steps.

2. Topology

Here is a simple experiment which you can do: take a glass of water in the palm of your hand and, keeping it upright (so the water doesn’t spill) rotate it by 360° anticlockwise (2\pi in radians). You will find that your arm is twisted uncomfortably. Continue the anticlockwise rotation and you find that your arm untwists after 720° (4\pi radians)! There is a sense in which 4\pi rotation is trivial but 2\pi is not. Understanding this simple experiment leads us naturally to topology and higher mathematics.

1 Another example is the Dirac Belt trick, which has the same intellectual content. See http://www.gregegan.net/APPLETS/21/21.html

Keywords
Rotations, spinors, reflections.

Rotations can be broken up into reflections.

Euler studied rigid body motion and in the process anticipated spinors.

**Box 1. Leonhard Euler (1707–1783)**

Euler had many theorems, one has to be specific about which theorem of Euler one is referring to. Apart from the theorem on rotations used in the text, Euler has theorems in several branches of mathematics and physics, from number theory to fluid mechanics. He had phenomenal powers of concentration and could work while dandling a child on his knee! He left behind a stack of manuscripts which kept mathematicians busy till recently – sorting, annotating and publishing them. They have now given up on this task and are just scanning the rest of Euler’s manuscripts and uploading them on the web.

Let’s now set ourselves to understand rotations in three dimensions. Consider ordinary three-dimensional space and fix an origin so that the coordinates of points in space are described by three real numbers \((x, y, z)\). These form a vector and we will sometimes denote it by \(\vec{r}\). Rotations are linear transformations of space that preserve the lengths of vectors: \(\vec{r} \cdot \vec{r} = x^2 + y^2 + z^2\) is invariant under the transformations. But not all linear transformations that preserve length are rotations. Reflections also preserve length, as you can see in a plane mirror. Reflections are relatively clean operations – they reverse some components of a vector and preserve others. The main point of this article is that two reflections lead to a rotation and so rotations can be broken up into reflections. Reflections are in some ways simpler than rotations. (Mirrors are “silver and exact” to borrow words from the poet Sylvia Plath!) We will use reflections to understand rotations. The turning of a giant wheel or a merry-go-round at a fair can be understood from quite abstract points of view. One would not have expected higher mathematics to lurk in such a commonplace phenomenon as rotation.

Consider linear transformations of \(\mathbb{R}^3 (\vec{r} \in \mathbb{R}^3)\)

\[
\vec{r}' = R\vec{r},
\]

which preserve length:

\[
\vec{r}' \cdot \vec{r}' = \vec{r} \cdot \vec{r}.
\]

From (2) it follows that \(R\) satisfies

\[
R^TR = 1,
\]

where \(R^T\) is the matrix transpose of \(R\). We say that \(R\) is an orthogonal matrix. It also follows by considering \(\vec{r}(\lambda) = \vec{r}_1 + \lambda \vec{r}_2\), (where \(\lambda\) is an arbitrary real number), and its transformation, that \(R\) preserves inner products (and therefore angles) between vectors

\[
\vec{r}_1' \cdot \vec{r}_2' = \vec{r}_1 \cdot \vec{r}_2.
\]

From (3) it follows that \((\det R)^2 = 1\) or \(\det R = \pm 1\). Transformations with \(\det R = +1\) are called rotations. These preserve the handedness of frames. Those transformations with \(\det R = -1\) are called reflections. These reverse the handedness of frames and are called improper. The transformations which preserve length form a group called \(O(3)\), which includes rotations as well as reflections. \((O\) means orthogonal (see equation (3))). The 3 in \(O(3)\) tells us the dimension we are in, which is three dimensions. Rotations form a subgroup called \(SO(3)\). \((S\) stands for special, which means \(\det R = 1\)).

Euler’s theorem on rotations (see Box 1) states that every rotation leaves some direction invariant. Euler’s theorem can be proved by noting that over the complex numbers the characteristic equation

\[
\det(R - \lambda I) = 0
\]
of $R$ is a cubic polynomial with real coefficients. From (3), (re-written as $R^*R = 1$, since $R$ is real), we see that $R$ is unitary, which implies $|\lambda| = 1$ and the eigenvalues lie on the unit circle. Since the coefficients of the characteristic equation are real, if $\lambda$ is an eigenvalue, so is its complex conjugate $\bar{\lambda}$. The three roots must be of the form $(e^{i\theta}, e^{-i\theta}, 1)$ since their product has to be unity from $\det R = 1$. The eigenvector corresponding to $\lambda = 1$ is the direction that is left invariant by the rotation. This proves Euler’s theorem: Every rotation in three dimensions has an axis $\hat{\nu}$ which is unchanged by the rotation. (Note that Euler’s theorem is true in all odd dimensions and not true in all even dimensions).

What are all the rotations? In other words what is the space of rotations? We want to find a space so that each point of the space corresponds to a rotation and each rotation is represented by one point. We know that rotations are characterised by an axis $\hat{\nu}$ (which needs two numbers to specify it) and an angle $\theta$ which needs one number. The space of rotations is three-dimensional. So, we will not be able to represent the space of rotations on a sheet of paper, but we can use points of three-dimensional space. We use the tip of the vector

$$\vec{V}_R = \theta \hat{\nu}$$

to represent a rotation. If $\theta = 0$, $\vec{V}_R = 0$, and the rotation by zero about any axis is represented at the origin. This is the identity of $SO(3)$. Rotations by non-zero angles are represented by non-zero vectors in the direction of the axis. What is the range of $\theta$? We remember that rotation by $\pi$ about $\hat{\nu}$ is the same as rotation by $\pi$ about $-\hat{\nu}$. Thus, $\theta$ goes from 0 to $\pi$. The picture of the rotation group $SO(3)$ looks like a solid ball. The centre of the ball is the identity. (We must remember though that adding two $\vec{V}$ vectors doesn’t make sense. The solid ball is just a representation of the rotations. To compose rotations we have to do more work. Simply adding the corresponding vectors is definitely wrong, because rotations don’t commute!) Although the solid ball has a boundary, we must remember that opposite points on its surface represent the same rotation.

This last remark leads to an interesting observation. There are closed curves in $SO(3)$ that start from the origin, go to a point $V_R = \pi \hat{\nu}$ at $\theta = \pi$ and return from its antipode $-\pi \hat{\nu}$ to the origin (Figure 1). The total angle of rotation as one traverses this curve is $2\pi$. Such curves cannot be shrunk to a point. We say that $SO(3)$ is ‘multiply connected’.

However, transversing this loop twice (rotating by $4\pi$) leads to a curve which can be shrunk to a point! $2\pi$ rotations are topologically non-trivial. $4\pi$ rotations are topologically trivial. This is the mathematics behind the simple experiment with a glass of water that we started with.

$SO(3)$ can be represented as a ball with opposite points of the boundary identified, i.e., to be regarded as the same point. Equivalently, consider a sphere in four dimensions $(x_1, x_2, x_3, x_4)$ all real with $x_1^2 + x_2^2 + x_3^2 + x_4^2 = 1$. If we suppose $x_4$ positive, the

---

**Figure 1.** The space of rotations is a sphere with opposite points identified (declared to be the same or glued together). The center of the sphere is marked O and represents the zero rotation. The vector shows a particular rotation whose direction is the axis of rotation and whose length represents the angle. The radius of the sphere is $\pi$. Also shown is a curve that starts from O, goes to A at the edge of the sphere and returns to O from the antipodal point, which is also labelled A. This is a closed curve in the space of rotations which cannot be continuously shrunk to a point.

The space of rotations is the boundary of a four-sphere with antipodal points identified.
set of points on the four-sphere can be identified with the ball in three dimensions $x_1^2 + x_2^2 + x_3^2 < 1$. Similarly, supposing $x_4$ negative we get another copy of this ball. For $x_4$ zero, we must remember the original identification of opposite points. Thus $SO(3)$ has the same structure as a sphere in four dimensions with opposite points identified. This space is called $\mathbb{R}P^3$ or real projective space. Such a space is easier to visualise in lower dimensions: $\mathbb{R}P^2$ is a sphere with opposite points identified. Such spaces come up in physics quite naturally. In liquid crystals, (look at your wristwatch or mobile to see them) there are long rod-like molecules that align to produce interesting optical properties. These molecules are long but do not point in any direction since both their ends are the same. Unlike vectors, these molecules do not have an ‘arrow’. The parameter that describes the ordering alignment in liquid crystals is a headless vector, (a line segment) or a point in $\mathbb{R}P^2$.

Using the four-sphere to represent rotations is a natural trick in view of the topological ideas we went through. Such methods were used to understand tops by Cayley and Klein (the ‘Cayley–Klein’ parameters) long before the quantum mechanical ‘spin’ of elementary particles was discovered. Notice that there are two antipodal points in the four sphere which represent the same rotation. This number two is closely related to spinors. We will see that reflections also bring in this number two naturally and lead us to spinors.

3. Reflections

Euler used improper elements of $O(3)$ to understand rotations. We will write $\tilde{R}$ for reflections. Two reflections give a rotation since $\det \tilde{R}_1 \tilde{R}_2 = (-1)^2 = 1$. This is easy to visualise in two dimensions. Reflecting the plane first in the line $P_1$ and then in the line $P_2$ leads to a rotation by $2\phi$. This is easy to see for vectors lying in either of the two lines $P_1$ or $P_2$. With some work you can convince yourself that it is in fact true for all vectors. You can also place two mirrors at an angle $\phi$ and see that the net effect of two reflections is a rotation by $\theta$, which is twice the angle $\phi$ between the mirrors. Equivalently,

$$\phi = \theta/2.$$  

This appearance of half angles is a characteristic of spinors! As we will see, spinors are objects that return to themselves only after a $4\pi$ rotation, unlike vectors, that return after a $2\pi$ rotation.

In three dimensions, we reflect in a plane $P$ whose unit normal is $\hat{p}$. The vector $\hat{p}$ is perpendicular to all vectors in $P$. The reflection operation is

$$\vec{r}' = \vec{r} - 2(\vec{r} \cdot \hat{p})\hat{p}$$  \hspace{1cm} (5)

and it reverses the component of $\vec{r}$ perpendicular to $P$, while preserving its parallel component. Note that replacing $\hat{p}$ by $-\hat{p}$ does not change matters since reflection is
bilinear in $\hat{p}$. Reflecting twice in the same plane gives back the original vector

$$\vec{r}'' = \vec{r}' - 2(\vec{r}' \cdot \hat{p})\hat{p} = \vec{r}.$$  

But reflection first in plane $P_1$ and then in plane $P_2$ gives a rotation whose axis is $\hat{n}$ and angle is $\theta$. Evidently, the angle of rotation is twice the angle between the planes.

$$\cos (\theta/2) = \hat{p}_1 \cdot \hat{p}_2.$$  

The axis must lie in the intersection of planes $P_1$ and $P_2$, since this direction is unchanged in both reflections. The axis must be perpendicular to both $\hat{p}_1$ and $\hat{p}_2$ and therefore lies along $\hat{p}_1 \times \hat{p}_2$ (which is non-zero unless $\hat{p}_1 \propto \hat{p}_2$).

For example, a reflection in the $xy$ plane takes

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} \rightarrow \begin{pmatrix} x \\ y \\ -z \end{pmatrix}.$$  

It simply reverses one of the coordinates of the vector $(x, y, z)$. Instead of arranging $(x, y, z)$ as a vector we could arrange it as a matrix $(X = \vec{\sigma} \cdot \vec{r} = \sigma_1 x + \sigma_2 y + \sigma_3 z)$, i.e.,

$$X = \begin{pmatrix} z & x - iy & x + iy \\ x - iy & x & -z \\ x + iy & -z & x \end{pmatrix},$$

where $\sigma_1, \sigma_2, \sigma_3$ are the Pauli matrices. You can find explicit forms for these in any quantum mechanics book, but we will not need these forms. They satisfy $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 1$, $\sigma_1 \sigma_2 = i \sigma_3$, $\sigma_2 \sigma_3 = i \sigma_1$ and $\sigma_3 \sigma_1 = i \sigma_2$. The different $\sigma$ anticommute with each other ($\sigma_1 \sigma_2 = -\sigma_2 \sigma_1$ and so on). Reflection in the $xy$ plane is just

$$X' = -\sigma_3 X \sigma_3.$$  

This follows because $\sigma_3$ anticommutes with $\sigma_1, \sigma_2$ and when the second $\sigma_3$ is moved through $X$ to cancel the first we end up reversing only $z$. More generally, reflection in the plane perpendicular to $\hat{p}$ is simply

$$X' = -p X p$$  

with $p = \hat{p} \cdot \vec{\sigma}$. Since $p^2 = 1$, $X'' = X$ as expected for reflections. If you reflect about $P_1$ followed by a distinct plane $P_2$, the result is a rotation. $X \rightarrow +p_2 p_1 X p_1 p_2$, where $p_2 p_1$ is

$$p_2 p_1 = (\hat{p}_2 \cdot \vec{\sigma})(\hat{p}_1 \cdot \vec{\sigma}) = \hat{p}_2 \cdot \hat{p}_1 + i(\hat{p}_2 \times \hat{p}_1) \cdot \vec{\sigma}.$$  

The angle of rotation is given by (we can suppose $0 \leq \theta \leq \pi$)

$$\cos (\theta/2) = \hat{p}_2 \cdot \hat{p}_1.$$  

The axis is given by $(\hat{p}_2 \times \hat{p}_1)$. The size of the vector $\hat{p}_2 \times \hat{p}_1$ is the square root of

$$(\hat{p}_2 \times \hat{p}_1) \cdot (\hat{p}_2 \times \hat{p}_1) = (\hat{p}_1 \cdot \hat{p}_1)(\hat{p}_2 \cdot \hat{p}_2) - (\hat{p}_1 \cdot \hat{p}_2)^2 = 1 - \cos^2(\theta/2) = \sin^2(\theta/2).$$
Writing \( \hat{n} = (\hat{p}_2 \times \hat{p}_1)/\sin(\theta/2) \) (assuming that \( \theta \neq 0, \pi \)) we have

\[
p_2 p_1 = \cos(\theta/2) + i \sin(\theta/2) \hat{n} \cdot \vec{\sigma}.
\]

This useful formula can also be written as

\[
p_2 p_1 = e^{i(\theta/2)\hat{n} \cdot \vec{\sigma}}.
\]

To see this simply expand the exponential as a power series and collect terms remembering that \((\vec{\sigma} \cdot \hat{n})^{2m} = 1\). Composing rotations is easy

\[
\begin{align*}
(\cos(\theta_2/2) + i \sin(\theta_2/2) \hat{n}_2 \cdot \vec{\sigma}) \times (\cos(\theta_1/2) + i \sin(\theta_1/2) \hat{n}_1 \cdot \vec{\sigma}) \\
= & \cos(\theta_2/2) \cos(\theta_1/2) - \sin(\theta_2/2) \sin(\theta_1/2) \hat{n}_1 \cdot \hat{n}_2 \\
& \quad + i \cos(\theta_1/2) \sin(\theta_2/2) \hat{n}_2 \cdot \vec{\sigma} + i \cos(\theta_2/2) \sin(\theta_1/2) \hat{n}_1 \cdot \vec{\sigma} \\
& \quad + i \sin(\theta_2/2) \sin(\theta_1/2) (\hat{n}_1 \times \hat{n}_2) \cdot \vec{\sigma}.
\end{align*}
\]

From this we can read off the axis and angle of the composite rotation \((\hat{n}, \theta)\)

\[
\begin{align*}
\cos(\theta/2) &= \cos(\theta_1/2) \cos(\theta_2/2) - \sin(\theta_1/2) \sin(\theta_2/2) \cos(\psi), \\
\sin(\theta/2) \hat{n} &= \cos(\theta_1/2) \sin(\theta_2/2) \hat{n}_2 + \cos(\theta_2/2) \sin(\theta_1/2) \hat{n}_1 \\
& \quad + \sin(\theta_2/2) \sin(\theta_1/2) \hat{n}_1 \times \hat{n}_2,
\end{align*}
\]

where we write \(\psi\) for the angle between \(\hat{n}_1\) and \(\hat{n}_2\).

4. Conclusion

The Pauli matrices \(\sigma^j\) are normally introduced as generators of rotations, satisfying commutation relations

\[
[\sigma^i, \sigma^j] = \sigma^i \sigma^j - \sigma^j \sigma^i = 2i \epsilon^{ijk} \sigma^k,
\]

where \(\epsilon\) is a completely antisymmetric tensor with \(\epsilon_{123} = 1\). In the present case what we used to represent reflections was the anticommutation relation

\[
\{\sigma^i, \sigma^j\} = \delta^{ij} \sigma^i + \sigma^i \sigma^j = 2 \delta^{ij}.
\]

In mathematics, objects that satisfy such anticommutation relations are called Clifford algebras. Starting from very elementary ideas, we step towards Clifford algebras and spinors.

In three dimensions we found that a \(4\pi\) rotation can be deformed to the identity. This is an example of a general topological characterisation of spaces. In any space one considers different ways in which a rubber band can be placed. Let us mark a point on the rubber band and keep its location fixed. We regard two configurations of the rubber band equivalent, if one can be deformed to the other continuously. We consider two configurations and multiply them by traversing them in succession. This forms a
group in the mathematical sense, called the fundamental group of the space. What we saw from the experiment is that the fundamental group of $SO(3)$ has two elements, $\{1, -1\}$. Examples of the $-1$ element are $2\pi$ rotations. Elements of the first are $4\pi$, $6\pi$ or $0\pi$.

Earlier $\vec{r}$ which was a vector in three-dimensional real space was replaced by $X$, a matrix in a two-dimensional complex space. Vectors in this two-dimensional complex space are called spinors. The study of elementary particles demands the use of spinors to describe particles like electrons, which have half-integral spin. Spinors form a representation of the group $SU(2)$, which can be identified with the sphere in four dimensions. This sphere ‘wraps around’ the space of rotations twice. We say it gives a double cover.

We learn in quantum mechanics that the wave function of a spin $s$ object acquires a phase of $\exp(is\theta)$ when it is rotated by $\theta$. For a spin-half particle it takes a $720^\circ$ rotation (or a $4\pi$ rotation) for the wave function to return to itself. This is closely related to the experiment with the glass of water we started with: $4\pi$ rotations are trivial in a topological sense.

In 2 dimensions it turns out that $4\pi$ rotations cannot be shrunk to a point. Note that we essentially needed the third dimension to shrink a $4\pi$ rotation to nothing. The space of 2-d rotations is a circle and its fundamental group consists of integers, the number of times the rubber band winds around the circle. In two dimensions spin does not have to be half integral and can take any value. This has implications in two-dimensional physics, which is currently an exciting area of research.

In relativity we have four-vectors $\begin{pmatrix} x \\ y \\ z \\ t \end{pmatrix}$ instead of three-vectors $\begin{pmatrix} x \\ y \\ z \end{pmatrix}$.

These four-vectors are easily accommodated.

Simply extend $X$ as

$$X = \begin{pmatrix} t + z & x - iy \\ x + iy & t - z \end{pmatrix}.$$ 

The earlier discussion goes through. We now have spinors of the Lorentz group $SO(3, 1)$ instead of the rotation group $SO(3)$. We can also write

$$X = t + i\mathbf{x} + j\mathbf{y} + k\mathbf{z},$$

where $i$, $j$, $k$ satisfy $i^2 = -1$, $ij = k$ and cyclic. These define the quaternions introduced by Hamilton (Box 2).

In closing we remark that mathematics is not a spectator sport. While it is fun to read, mathematics enters your system only if you struggle with it. To help you do this here are some exercises.

Two dimensions are special: spin does not have to be half integral.

The idea of a spinor carries over to relativity theory.
Box 2. William Rowan Hamilton (1805–1865)

Hamilton invented quaternions while trying to generalise complex numbers. In Dublin, you can find a bridge (now Broom bridge, earlier Brougham bridge) on which William Rowan Hamilton carved their defining relations. Hamilton’s quaternions are closely related to spinors and now very much a part of a good theoretical physics education. However at the time they were invented, their relevance to the physical world was seriously doubted. The great physicist Lord Kelvin (earlier known as William Thompson) took a dim view of quaternions: “Quaternions came from Hamilton after his really good work had been done; and though beautifully ingenious, have been an unmixed evil to those who have touched them in any way... Vector is a useless survival, or offshoot from quaternions, and has never been of the slightest use to any creature.”

Lord Kelvin may have been a great physicist, but he was a poor prophet! He did not foresee how important Hamilton’s ideas would turn out to be in the microscopic world.

Exercises

1. Let $R$ be a rotation. It can be expressed as a product of two reflections in more than one way. How many ways can this be done?

   $$ R = \tilde{R}_1 \tilde{R}_2. $$

2. Given rotations $R$ and $S$, we can write

   $$ R = \tilde{R}_1 \tilde{R}_2; \quad S = \tilde{S}_1 \tilde{S}_2. $$

   Use the freedom in 1. above to choose $\tilde{R}_2 = \tilde{S}_1$ so that these reflections cancel out. We find then

   $$ RS = \tilde{R}_1 \tilde{S}_2. $$

   Use this trick to derive the formula for composing rotations. Look up Hamilton’s theory of turns on the web to see the connection.

Acknowledgements

It is a pleasure to thank C S Hegde for reading through the manuscript and suggesting improvements.

Suggested Reading

Electrons in Condensed Matter *

TV Ramakrishnan

1. Introduction

A hundred years ago, J J Thomson tore light negatively charged particles out of many different kinds of atoms by subjecting them to strong electric fields in discharge tubes. He showed that the same particle (with identical charge $e$ and charge to mass ratio $(e/m)$) came out of different atoms. This established the electron as a basic ingredient of atoms and molecules, which constitute all common forms of matter.

In dilute gases of atoms and molecules, electrons lead stable, private lives; an electron is confined to its atom, or is shared between the atoms of a molecule. The electron is attracted and bound to the positively charged nucleus, and the resulting atom or molecule is stable and is electrically neutral. This is a reason why it took so long to free and know electrons.

As atoms or molecules are brought close to each other, i.e as matter condenses to a liquid or a solid, atoms exert strong forces on each other. This often frees outer electrons from bondage to individual atoms or molecules; they are free to move around the entire liquid or solid. The variety of electronic behaviour which then results is astonishingly diverse and stretches our imagination to its limits. It is influenced by the spatial arrangement of atoms, by the atomic heritage of the electrons, by the fact of their being identical and together and by the interactions between them. In this and two other articles, one by D N Bose on Transistors – From Point Contact to Single Electron and another by A K Raychaudhuri on Seeing with Electrons, we briefly celebrate a century of this continuing journey of exploration and discovery. What electrons do in condensed matter is not only surprising and interesting; it is also a large part of physics and materials based technology which is continuing to transform the world. This will also be brought out wherever possible.

I will start with a brief account of early discoveries and ideas pointing to some roles electrons play in condensed matter. It was however only with the development of quantum mechanics (that is, the correct laws of motion of electrons and other material objects) in the 1920’s, that some basic patterns in the electronic properties of condensed matter could be rationalized. Finally, some examples illustrating the unusual diversity of electronic behaviour in condensed matter, eg. ferromagnetism, superconductivity, metal to insulator transitions and quantized Hall effect will be described.

Perhaps the most important for us here is the free electron theory of metals, developed by P Drude in 1900, within three years of the discovery of the electron. The theory is simple and has many consequences some of which agree very well with experiment, and others which conflict equally spectacularly!

1 In the calculation so far there is no absorption of electromagnetic energy, hence it is all reflected.

2. Beginnings: The Drude Free Electron Theory of Metals

As the idea that the electron is a constituent of atoms gained ground, a number of implications became clear. Perhaps the most important for us here is the free electron theory of metals, developed by P Drude in 1900, within three years of the discovery of the electron. The theory is simple and has many consequences some of which agree very well with experiment, and others which conflict equally spectacularly! However, with relatively minor changes (which take into account properties of the electron discovered later) this picture is still an accurate approximation to the properties of many metals.

Suppose atoms in metals are so close that electrons from different atoms ‘run into’ each other, lose their atomic identity, and move about freely in the solid. These free electrons help conduct electrical current, as described below, and the high electrical conductivity of metals such as aluminium is ascribed to their having a high density of free, mobile electrons. One problem with this explanation is that some other equally dense solids, eg diamond, are insulators! We return to this later.

Is there any experimental way one can see that there are free electrons in metals? How many are there? One answer is the following: consider a free electron in an oscillatory electric field $\vec{E}(t) = \vec{E}_0 e^{-i\omega t}$ such as could be due to an electromagnetic wave of angular frequency $\omega$. The electron experiences a force $e\vec{E}(t)$ and is forced to oscillate in tune. If the electron position is $\vec{r}(t)$, we have

$$e\vec{E}(t) = m\ddot{\vec{r}}(t) = -m\omega^2 \vec{r}(t) \quad (1)$$

since $\vec{r}(t)$ oscillates with frequency $\omega$. Thus the electron has an induced electric dipole moment $e\vec{r}(t)$, and if there are $n$ free electrons per unit volume, the induced dipole moment per unit volume is $ne\vec{r}(t) = -(ne^2/m\omega^2) \vec{E}(t)$. This induces an electric field $\vec{E}_i = -(4\pi ne^2/m\omega^2)\vec{E}$ so that the dielectric constant $\kappa$ of the free electron gas is

$$\kappa(\omega) = 1 - (4\pi ne^2/m\omega^2). \quad (2)$$

From this, we infer a frequency dependent refractive index

$$n(\omega)\sqrt{\kappa(\omega)} = \sqrt{(1 - (4\pi ne^2/m\omega^2))} \quad (3)$$

We notice that for $\omega < \omega_p = (4\pi ne^2/m)^{1/2}$, the refractive index is pure imaginary; light does not go through the metal, but decays exponentially into it starting from the surface. The incident energy is reflected back.\footnote{In the calculation so far there is no absorption of electromagnetic energy, hence it is all reflected.}

These facts are close to what we know from experience. More interestingly, for $\omega > \omega_p$, $n(\omega)$ is real, and the metal is transparent! This transparency of metals for high frequency light was observed by Wood in 1930 or so. In addition to confirming the free electron picture, it allows us to estimate $n$ from a measurement of the transparency frequency $\omega_p$ (known as the plasma frequency). For example for Na, 5.8 eV. This leads to an $n$ of $2.6 \times 10^{22}/\text{cm}^3$, which is about one electron per sodium atom. This
is reassuring, since the Na atom does have one loosely bound electron outside the closed shell.

Suppose the electron collides with random obstacles (vibrating ions, ions out of place, other electrons) so that the momentum gained by it (in an electric field say) is lost at a rate $\tau^{-1}$ ($\tau$ is called collision or relaxation time). This damping force is included in the force equation (1) as a term $-(m\vec{v})/\tau$ on the left hand side. (This is very much like the Stokes viscous force on a ball falling through a fluid, which goes as $6\pi\eta\vec{v}$).

One can again find the dielectric function and the electrical conductivity $\sigma$, which is seen to be of the form

$$\sigma(\omega) = (ne^2\tau/m)(1 - i\omega\tau)^{-1} = \sigma_0(1 - i\omega\tau)^{-1}. \quad (4)$$

This is the famous Drude formula for complex conductivity which can be checked by measuring $\sigma$ at different frequencies. It relates the dc conductivity $\sigma_0$ to electron density and the relaxation time $\tau$. It does describe the ac conductivity of metals well.

It can also be used to show that ac electric field does not penetrate the metal uniformly, but is exponentially damped out inside, starting from the surface, with a characteristic distance $\lambda = \sqrt{c/\sigma_0\omega}$, called the skin depth. This phenomenon is called the skin effect, and is well known in ac circuits, and is one of the reasons why connecting wires for high frequency circuits especially, have many strands (much more skin!)

The earlier dielectric function formula (2) is obtained from (4) in the limit $\omega\tau \gg 1$. So, here we have a simple, detailed and successful description of electrical (dc and ac) and optical properties of metals in a classical free electron picture.
$k_B$ is Boltzmann’s constant $\approx 1.38 \times 10^{-23}$ Joules per degree kelvin. According to simple kinetic theory, a free particle has energy $3k_B T/2$ and a bound particle (like the aluminium ion) $3k_B T$. The specific heat is just the coefficient of proportionality between energy and $T$ in this simple model.

Is anything wrong? Well, yes, as Drude and others realized very early. As mentioned above, this picture does not explain why some solids are metals with free electrons, while others, nearly equally dense, are not. Secondly, if there are so many free electrons, they ought to contribute to the specific heat. For example, aluminium has three free outer electrons per atom. Its specific heat should be $(3k_B + 3 \times 3k_B/2)$ per atom, $3k_B$ for the vibrating atom, and $3 \times (3k_B/2)$ for the three freely moving electrons. The value at room temperature is $3.1k_B$; the electronic specific heat is missing! The next stage in the electronic theory of solids clears up these two puzzles, as we see now.

3. Electron Waves in Solids

In the 1920’s, it became clear that electrons are waves (this is discussed in other articles in this issue). Electron waves of different wavelengths inhabit the solid, much like a string fixed at both ends supports different modes. In a large system, these modes are very closely spaced; the energy spacing between two successive modes goes inversely as the volume of the system. Now a given mode or wavelength state can be occupied by only two electrons, these differing in their intrinsic spin or magnetic moment orientation; this is the Pauli exclusion principle. So, at zero temperature, the lowest energy states are occupied one by one, up to a maximum depending on the number of electrons. This maximum, called Fermi energy $\varepsilon_F$, is about 30,000 to 60,000 K in temperature units. (If we write $\varepsilon_F = k_B T_F$, where $k_B$ is the Boltzmann constant, then $T_F$ is between $3 \times 10^4$ to $6 \times 10^4$ K. Suppose we heat this solid to a

Figure 2. Thermal energy can excite only a small fraction of the total electrons.
temperature $T$, i.e., give each electron an energy $k_B T$. The only ones which can actually increase their energy are the small fraction $(k_B T / \varepsilon_F)$ situated near the Fermi energy. For the others, the states to which they could go on gaining an energy $k_B T$ are fully occupied! Thus the energy gained by a gas of $N$ electrons is approximately $N k_B T (k_B T / \varepsilon_F)$. The specific heat is the temperature derivative of this energy, and is $2Nk_B (k_B T / \varepsilon_F)$. The factor in brackets is the reduction with respect to a classical gas value due to the wave nature of the electron and the Pauli exclusion principle; it is only 1% or so at room temperature. So electrons cannot contribute much to the specific heat of metals at room temperature. Both this, and the linear temperature dependence (easily measured at low temperatures) are indeed in agreement with experiment.

Now consider the wavelength of the highest occupied electronic state. If this is the same as the lattice spacing, there is resonance; the electron is repeatedly scattered back and forth by the periodically arranged lattice of ions, the scattered and original waves have a definite phase relationship and there is interference. This resonant interference lowers or increases the energy of the state by a discrete amount, called the energy gap. Thus when certain numbers of electrons (even numbers per unit cell) fill the solid, the next possible propagating state is separated from the highest occupied state by an energy gap. (The closely spaced occupied states form a ‘band’, called the valence band; the unoccupied or partially occupied states are part of the conduction band.)

Thus whether a crystalline solid is a metal or not depends on the number of outer electrons per unit cell of the crystal. If the number is odd, as in Cu where it is one, the solid is a metal. If the number is even, as in diamond or silicon where it is four, the solid can be an insulator. We say ‘can be’, because in a solid, the wavelength for resonance depends on the direction of propagation of the wave. Thus the energy after reduction (or increase) by the gap depends on direction; the highest occupied states may not then be separated by a gap from the lowest unoccupied states! So, for example, lead with four electrons per unit cell is a metal.

(Whether a solid is called a semiconductor or an insulator depends on the size of the band gap. If the band gap is about an eV or less, sufficient electrons can be thermally excited across the energy gap so that the conductivity is 0.1 to 1 (ohm cm)$^{-1}$ at room temperatures. Such substances are called semiconductors; those with a larger gap or lower room temperature conductivity are called insulators).

The above basic classification of solids, depending on the periodicity of the solid, and on the wave nature of the electron, holds up generally, and is substantiated by detailed calculations of the energies of various modes, called band structure. There are two spectacular ways in which this picture also fails; both are subjects of continuing attention, needing new ideas. One is the insulating nature of nonperiodic or glassy solids; some of the best insulators are glassy, even liquid! The notion of band gaps mentioned above is tied to waves in a periodic structure, interference, electron count etc. A new general idea, called localization of states in a random medium,
is needed and has been developed to rationalize some properties of nonperiodic electronic systems. The other is the insulating nature of some periodic solids with odd numbers of electrons per unit cell; these ought to be metallic. For example, \( \text{La}_2\text{CuO}_4 \) has, in its square planar unit cell, one \( \text{Cu}^{++}\text{O}^{--} \) ‘molecule’. \( \text{O}^{--} \) has a filled shell configuration, while \( \text{Cu}^{++} \) has nine d electrons in the outermost shell. It ought to be a metal, but is an insulator. What is the reason for the existence of such ‘Mott’ insulators? On replacing some La with Sr in \( \text{La}_2\text{CuO}_4 \) a high temperature superconductor \( \text{La}_{2-x}\text{Sr}_x\text{CuO}_4 \) \((0.1 < x < 0.2)\) is obtained, with unusual normal state properties. Does this have something to do with the parent compound being a Mott insulator? We do not know yet; questions of this sort are very much at the heart of modern condensed matter physics.

4. Unusual Lives of Interacting Electrons

We have pretended so far that electrons move freely in solids, or at best that they are aware only of the periodic arrangement of atoms which diffract them. This is surprisingly useful fiction. But there is an amazing variety of phenomena, possible only because electrons exert forces on each other; a small selection of these is presented here, with very sketchy explanations. The idea is to convey a flavour of the possibilities.

**Ferromagnetism:** Iron and many other metals, intermetallic compounds, alloys, and insulators, are magnetized below a certain temperature. The magnetic moments of the outer electrons spontaneously align themselves parallel to each other, at temperatures...
as high as a thousand degrees (1043 K for Fe). A little thought shows that this has to be due to interactions between electrons. First consider an atom. If the outer electronic shell is incompletely filled, the atom will have a magnetic moment. In solids this is mainly due to the intrinsic or spin magnetic moment of the electrons; the orbital moment is ‘quenched’ by the low local symmetry of the solid. So, each atom has a spin magnetic moment. Or does it? If electrons can move in and out of atoms in a solid, what is to preserve the magnetic integrity of each atom, its magnetic moment? Can there be magnetic moments in metals? It turns out there can be, but these are objects mixed with their surroundings, nonintegral in general and often disappearing at low temperatures. It is clear that the question of their ordering involves electrons being exchanged from one moment to the other. One can imagine two limits. The atomic limit is one where the intermoment electron mixing is weak. The energy of purely magnetic interaction between near neighbour moments (dipole–dipole interaction) is of the order of a few degrees Kelvin: ordering temperatures are more than a hundred times higher. So magnetic order has to involve exchange of electrons. The opposite is the itinerant limit, where the ‘up’ spin electrons and the ‘down’ spin electrons freely and separately (independently) move in the solid, forming bonds. If they are truly independent, they will occupy the same energy levels one by one (Pauli exclusion principle) and there will be no net magnetization. There has to be repulsive interaction between up and down spin electrons; this gives maximum repulsive energy when their numbers are equal, and thus favours ferromagnetism (in which the numbers become unequal). So we see that ferromagnetism in metals requires exchange and interaction, and raises questions about the nature of moments in metals. These questions are not fully answered yet, even for iron.

Can there be magnetic moments in metals? It turns out there can be, but these are objects mixed with their surroundings, nonintegral in general and often disappearing at low temperatures.

Ferromagnetism in metals requires exchange and interaction, and raises questions about the nature of moments in metals. These questions are not fully answered yet, even for iron.
Superconductivity: In 1908, Kamerlingh Onnes succeeded in liquefying helium at 4 K or so. In the succeeding years, in his laboratory at Leiden, many solids were cooled in liquid helium, and their properties were studied. One aim was to check if as predicted by Drude, resistivity of pure metals continues to decrease as atomic vibrations are chilled out. This indeed happens; but mercury suddenly seemed to lose \textit{all} electrical resistance below 4.2 K or so. The resistivity of a mercury sample was observed to decrease from 40 $\Omega$ at room temperature to about 0.08W at 4.3 K, and to then plummet suddenly to unmeasurably low values, below $3 \times 10^{-6}$ $\Omega$! This new state of superconductivity (see R Srinivasan in Suggested Reading), discovered in 1911, took about fifty years to be fully understood. It occurs not just in mercury, but in many metals and alloys at low temperatures.

The superconductor is not just a perfect conductor with no electric fields $\vec{E}$ inside it; it is also a perfect diamagnet with $\vec{B} = 0$! No magnetic field is admitted, as discovered by Meissner and Ochsenfeld in 1933. The perfect conductivity raises the following question: How is it that electrons that were being randomly scattered by various objects, causing electrical resistance, are no longer affected by the same objects at an infinitesimally lower temperature? The crude answer is that the entire collection of electrons in the superconductor goes into a single coherent state, which pervades the solid and is rigid against local deformations. What is the nature of this ‘macroscopic quantum coherent’ state? Bardeen, Cooper and Schrieffer, in their successful microscopic theory proposed in 1957, argued that electrons condense in pairs, and that these pairs are in phase with each other. Now, pairing of electrons

![Figure 5. Pairing of electrons via lattice vibrations.](image-url)
seems unlikely; they repel each other electrically, and because of the Pauli exclusion principle, cannot be in the same state. Bardeen, Cooper and Schrieffer proposed an unusual mechanism of attraction. An electron whizzes past the lattice of ions, deforming it in some place. The ionic deformation being heavy, moves slowly. Another electron travelling in the opposite direction comes by later, and experiences the attraction due to this still present lattice deformation (caused by an electron no longer there!). The retarded (i.e. delayed) attraction mediated by lattice deformation (phonons) can overcome the instantaneous repulsion, and causes pairing of electrons with zero total momentum (and total spin). This mechanism seems to be the right one for many superconductors.

The recently discovered high $T_c$ or cuprate superconductors, eg. La$_{2-x}$Sr$_x$CuO$_4$, Ba$_2$Cu$_3$O$_7$ almost certainly become superconducting due to some other causes. Their metallic state above the superconducting transition temperature $T_c$ is in detail unlike that of any metal we know. For example, the cuprates are metallic in the plane containing the Cu$^{++}$ ions, but behave like an unusual kind of insulator perpendicular to the plane, and become bulk three dimensional superconductors below $T_c$. As mentioned earlier, the parent compound La$_2$CuO$_4$ is a Mott insulator (which orders antiferromagnetically at about 200 K). The Sr doped compound La$_{2-x}$Sr$_x$CuO$_4$ is an unusual metal that becomes a high $T_c$ superconductor. How does this transformation take place? We do not understand these mysteries yet; hopefully it will take less than fifty years.

One very bizarre consequence of macroscopic quantum coherence is the effect theoretically predicted by Josephson in 1961 (he was a Ph D student at Cambridge then). If two superconductors are separated by a very thin insulating layer, a steady dc voltage difference, say $V$, can be maintained between them. When this is done, an ac signal, of angular frequency $\omega = (2 eV/\hbar)$ is generated! The two superconductors are in single quantum states with precisely defined phases. The applied voltage changes the phase difference at a steady rate: $\phi = (2 eV/\hbar)$. The factor two is a giveaway; the phase is of charged electron pairs. Indeed the ac Josephson effect is observed, and is used among other things as a quantum voltage standard.

**Metal - Insulator Transitions:** One idea basic to our thinking of electrons in condensed matter, is the distinction between metals and insulators. In the former, occupied and unoccupied states near the Fermi energy are extended in space, and can carry current. In the latter, there is either a gap between extended occupied states and unoccupied states, or these states are localized and unable to carry current. These seem fundamentally distinct phases electronically. Surprisingly, many systems undergo a transition from one to the other as some parameter is varied. One well known example is V$_2$O$_3$, which on cooling suddenly becomes an insulator, with a millionfold increase in resistivity. How is it that electrons which were there (moving about, spread around) are no longer there?

In the last few years, another family of oxides eg La$_{2-x}$Ca$_x$MnO$_3$ has been extensively
explored all over the world, including a group in Bangalore (C N R Rao and A K Raychaudhuri). These manganites show a transition from a high temperature paramagnetic insulating phase to a low temperature ferromagnetic metallic phase. The transition occurs at a temperature of about 230 K. Near the transition, a relatively small magnetic field of order a tesla or so, can change the electrical resistivity by factors of two or more! This colossal magnetoresistance is astonishing because the energy scale associated with the magnetic field is $\mu_B H = 2 K$ for 1 T. This is a hundred times smaller than the energy scale $k_B T_c$ of the transition, and ten thousand times smaller than electronic energies which are of order $10^4 K$ in temperature units. How can such a small change in magnetic energy make such a large change in physical properties? Can a small tail wag a big dog? We do not know the reasons yet. As it turns out for many fundamentally interesting phenomena, colossal magnetoresistance may also find applications, this time in magnetic recording.

**Quantized Hall Effect:** We conclude our selection of unusual doings of electrons in solids with a beautiful quantum effect discovered by von Klitzing in 1980. He was investigating anomalies in resistance and magnetoresistance of a gas of electrons confined to move in a plane. (This is achieved by a semiconductor arrangement called an inversion layer). A large magnetic field, of ten Tesla or more, is applied perpendicular to the plane, and the inversion layer is at a low temperature, of order a degree or so. Under these conditions von Klitzing found that the Hall resistance showed well defined plateaus as a function of magnetic field. The Hall resistance is the ratio of the in plane voltage developed across the sample, and the current flowing perpendicular to it in the plane. (The magnetic field is perpendicular to both). The surprising thing was the value of the resistance; it is $(h/e^2)$ where $h$ is Planck’s constant, $e$ the electron charge, and $i$ is an integer. Why is the Hall resistance quantized in terms of fundamental constants? How accurately is it quantized? We cannot here discuss the explanation, which finally has to do with the effect of a magnetic field on the phase of the electron wave, but the second question is easier to answer. So far as we can tell, the relation is exact. It is accurate and reproducible to a few parts in $10^8$; there is a quantum resistance standard $(h/e^2) = 25812.80 \Omega$.

A few years after this discovery, in 1982, Tsui, Störmer and Gossard discovered the fractional quantum Hall effect, in which the conductance is not $i(e^2/h)$, but $\nu(e^2/h)$ where $\nu$ is an odd denominator fraction, e.g. $\nu = 1/3, 2/5 \ldots$ An explanation of this required completely new ideas, due to Laughlin (who also provided the general framework for understanding the integral quantum Hall effect). Strong magnetic fields, low temperature electron interactions and some disorder are essential to the observation of these unique quantum phenomena. Laughlin predicted that the lowest stable excited particle state for a $\nu = 1/3$ fractional quantum Hall state has a charge $(e/3)$; its statistics is also fractional (it is neither a fermion nor a boson). Indeed such $(e/3)$ charge excitations have been directly observed in electrical noise experiments at very low temperatures. So, as a consequence of interactions and magnetic field, new quasi particles are generated with fractional quantum numbers; they actually exist.
As the first century of our acquaintance with the electron nears its end, this most fundamental and stable of all particles seems to be coming apart, assuming unexpected forms. In systems containing many electrons that interact strongly with each other, the collection settles to some ‘ground’ state as disorderly thermal motions cease (i.e. as temperature $T \to 0$). Particle excitations out of this ground state, or excited ‘quasi’ particles can be quite unlike electrons in their basic properties. This is clear in several one and two dimensional systems. For example, in one dimension, for an interacting many electron system, there are excitations carrying spin but no charge, moving with a different speed from excitations carrying charge but no spin! (This is called spin-charge separation). In polyacetylene, namely $(\text{CH})_x$ (where $x$ is very large, of order a thousand or more) there are alternating single and double carbon-carbon bonds; the molecule forms a long chain. One can have defects such as two double bonds or two single bonds adjacent to each other. It turns out that such defects are mobile, and have a charge $(e/2)$ for the former and $(-e/2)$ for the latter. We have already mentioned above the fact that if a collection of electrons moving in a plane is subjected to very high magnetic fields, for some density of electrons, the quasiparticle excitations have charge exactly $(e/3)$. These examples illustrate strikingly the fact that the effective low excitation energy degrees of freedom of a background interacting electron fluid can be qualitatively different from electrons. (This of course raises again the question: what is an electron? Is this charged magnetic clump of matter, the lightest member of a family, an excitation out of some more basic ground stuff?).

Considering the slow and sedate beginnings of our century old journey into the realm of electrons and their unexpected doings in condensed matter that we have come to know of in the last few decades, the next century promises to unfold wonderful new worlds.
Postscript (added by December 2021)

In the nearly quarter of a century since the article was originally written, there has been both evolution and revolution in this field. This supplementary note selectively touches on a few of these.

The quantum Hall effect (QHE) was mentioned then as a relatively new finding. This has evolved into a many splendored thing, as we mention below. The clear recognition that the integral Quantum Hall system is a piece of topological quantum matter has given rise to the biggest revolution in the field. Novel kinds of topological materials are being predicted and made; this is an explosively growing area with huge societal consequences, and is a large part of the large subject of quantum materials (including quantum computing materials).

The electron ‘falling apart’ was mentioned there as a recent realization. It has become quite ‘old hat’ by now; while originally the ‘falling apart’ of the electron was in the sense of its spin and charge degrees of freedom having qualitatively different dynamics in materials with strong electronic correlations, there are increasingly many realizations and speculations of there being quite novel kind of emergent quasiparticles in electronic many body systems.

A completely new electronic material, graphene, an atomic film of carbon atoms bonded as in a sheet of graphite, was discovered and isolated in 2004.

Details of some of these developments are mentioned below very sketchily. Applications which range from quantum computing to hydrogen energy storage materials are not mentioned.

We start with the quantum Hall effect, which continues to be a cornucopia of completely new phenomena and ideas. A glimpse of this is given by a viewpoint article in Nature Reviews (Physics) in 2020, marking forty years of the discovery. Characteristically, the article has more than a dozen co-authors, including von Klitzing, the discoverer of the effect. They emphasize new directions. For example, there are sections on the quantum spin Hall effect, quantum anomalous Hall effect, quantum thermal Hall effect, QHE in graphene, QHE in 3 dimensional systems, quantum entanglement, synthetic fields for atomic QHE (in cold atom systems) and topological states of light (realized in photonic crystals). This is in addition to amazing developments in what may be termed ‘conventional’ QHE. For example, more than seventy distinct quantum Hall states below a filling appropriate to the first Landau level have been observed; each is topologically different. A large number of both odd denominator fractions and even denominator fractions have been observed.

The quantization of Hall conductance is extremely precise (to more than a part per billion), unaffected by disorder in the MOSFET system etc. (the quantization is now used to define the values of the fundamental constants e and h; this is quantum...
metrology). The early realization that the Hall conductance in units of \((e^2/h)\) is a topologically invariant integer (a Chern number) ‘protected’ by an intrinsic energy gap inspired the revolutionary realization that there is a direct connection between topology and condensed matter physics.

The existence and importance of topological numbers robustly characterizing certain geometrical features of a system has been known for a long time. For example, with a sphere, we associate a number zero, and with a doughnut with a single hole the number one etc. The number is unchanged on deforming the sphere (say into an ellipsoid or even into a disk; think of a spherical piece of dough being pressed into different shapes) or deforming the doughnut (e.g. into a cup with one handle). That one can associate such topologically nontrivial quantum numbers with electronic states in solids is perhaps the greatest new development in condensed matter science and has obvious implications, now being realized in a large number of materials. Completely unexpected topological properties have been predicted and found. For example, \(\text{Bi}_2\text{Se}_3\) is a semiconductor characterized by band inversion; at the interface between it and free space, there are necessarily zero energy (chiral) electron states; the surface is inevitably metallic. The consequences of this are known in depth. There is another family of materials called Weyl semimetals in which electrons and holes have ‘Weyl’ dispersion, namely the energy \(E\) and momentum \(\mathbf{p}\) are related as \(E \propto \sigma_i \mathbf{p}\) where \(\sigma_i\) are the Pauli spin matrices. This also has bizarre experimental consequences. It has been shown recently that nearly a quarter of crystalline semiconductors are topologically ‘interesting’ so that this is not an arcane phenomenon. Further, the properties of topological semiconductors are protected by the semiconducting band gap, a large energy compared to room temperatures, so that these are not low temperature properties but are accessible at room temperatures.

Graphite, a common material (lead pencils) can be thought of as a number of weakly coupled atomic layers of carbon, mutually bonded and stacked one on top of another. The single layer (a planar honeycomb lattice) is graphene, isolated and studied in 2004 (Geim and Konstantilov, Nobel Prize 2010) by a primitive method in which they peeled off a flake from graphite using scotch tape, and deposited this on a silica substrate, occasionally obtaining a single layer. Electronically, it is a semiconductor with zero band gap, a semimetal. The highest occupied states (valence band) are made up of sigma bonded electrons and are separated from the unoccupied states (valence band) by zero gap. Because of the nature of the honeycomb lattice, the energy of these electron states is linearly dependent on their momentum (not on their square as is generally the case for free electrons or for electrons in solids at band extrema!). This (Dirac) dispersion combined with the gaplessness, both observed here for the first time, is the electronic signature of graphene and is what is expected for relativistic massless fermions. The mobility of an electron in the graphene conduction band is phenomenally large; its mechanical strength is about a hundred times larger than that of steel with the same thickness. Because of such electrical and mechanical properties, it is called a wonder material. Single layer graphene, bilayer graphene
with a small (∼1.1°) twist angle, rolled graphene sheets (carbon nanotubes) etc. are home to a large number of exotic phenomena.

The other major departure in quantum condensed matter physics dating to the nineteen eighties is the discovery of superconductivity in cuprates at unprecedentedly high temperatures. Here, the theoretical description both of the superconductivity and of the ‘normal’ phase from which it arises below $T_c$ continue to be a matter of controversy. Many new phenomena characteristic of these materials have been discovered and explored; in their totality, they add to the mystery of the phenomenon. One very simple fact, first noted in cuprates, is that their electrical resistivity depends linearly on temperature, from above $T_c$ to the highest temperatures for which measurements can be made. As seen in twisted bilayer graphene, the latter can be as high as the effective Fermi temperature of the collection of charge carriers. Interestingly, when superconductivity in cuprates is destroyed by a high magnetic field, the resulting material also has resistivity linear in temperature. The ubiquity of the effect (it is not confined to cuprates, and is believed to be present in all strongly correlated systems) and the apparent constancy of the slope (in renormalized units) are major questions calling out for understanding.

**Suggested Reading**


The author wishes to thank Ayan Guha, Physics Department, IISc for the illustrations.
Emerging Trends in Topological Insulators and Topological Superconductors *

Arijit Saha and Arun M Jayannavar †

Topological insulators are a new class of materials which are characterized by a bulk band gap like ordinary band insulators but have protected conducting states on their edges or surfaces. These states emerge due to the combination of spin-orbit coupling and time reversal symmetry. Also, these states are insensitive to scattering by non-magnetic impurities. A two-dimensional topological insulator has one dimensional edge states in which the spin-momentum locking of the electrons give rise to quantum spin Hall effect. A three-dimensional topological insulator supports novel spin-polarized 2D Dirac fermions on its surface. These topological insulator materials have been theoretically predicted and experimentally observed in a variety of 2D and 3D systems, including HgTe quantum wells, BiSb alloys, and Bi₂Te₃, Bi₂Se₃ crystals. Moreover, proximity induced superconductivity in these systems can lead to a state that supports zero energy Majorana fermions, and the phase is known as topological superconductors. In this article, the basic idea of topological insulators and topological superconductors are presented along with their experimental development.

1. Introduction

In condensed matter systems, atoms with their electrons can form many different states of matter, such as crystalline solids, magnets, and superconductors. These states can be classified by the concept of symmetry breaking. For the above mentioned examples, translational, rotational, and gauge symmetries respectively, are spontaneously broken. Before 1980, the principle of broken symmetry was the key concept for the classification of states of matter. The discovery of quantum Hall effect (QHE) in 1980 [1] provided the first example of a quantum state where no spontaneous symmetry was broken. In QHE, the electrons confined in a 2D electron gas, subjected to a strong magnetic field of the order of few tesla [1] manifests an entirely different type of topological order. Its behavior is independent of its specific geometry. Hence, the quantum Hall (QH) state was topologically distinct from all previously known states of matter 1.

In recent times, a new class of topological state has emerged called the quantum spin hall (QSH) phase or the topological insulators (TI) [2–7]. Such states are topologically distinct from all other known states of matter, including the QH states. QSH systems

---

† Deceased


**Keywords**

Superconductor, quantum Hall effect, topological insulator, Majorana fermions.
Arijit Saha and Arun M Jayannavar

When an edge-state electron encounters an impurity, it still propagates along the same direction as backscattering is prohibited along the same edge. This is the key reason why the QH effect is topologically robust.

Figure 1. (a) Schematic of a QH bar in which the upper edge contains a forward mover and the lower edge contains only a backward mover. Here, both the 1D edges are spinless. (b) Cartoon of a QSH bar in which both the 1D edges are spinful and spin-momentum locked i.e. of helical nature. The upper edge contains a forward mover with up spin and a backward mover with down spin. The spin and momentum directions are reversed for the lower edge. are insulating in the bulk which means that they have an energy gap separating the conduction and valence bands. In spite of having a bulk band gap, they contain metallic edges or surface bands on the boundary. These boundary edges or surface states are topologically protected and immune to scalar (non-magnetic) impurities. This means backscattering is prohibited by such impurities along the edge or surface. Moreover, these boundary states are protected by time reversal (TR) symmetry. Here lies the important difference between a QSH state and a QH state. The latter requires an external magnetic field which explicitly breaks the TR symmetry. In contrast, QSH states are TR invariant and do not require an external magnetic field. The signatures of QSH states have been experimentally observed in HgTe quantum wells [8], in BiSb alloys [9], and in Bi$_2$Se$_3$, Bi$_2$Te$_3$ bulk crystals [10, 11].

2. Quantum Spin Hall Effect

In QH effect, a strong magnetic field is applied perpendicular to a 2D electron gas in a semiconductor. Here, magnetic field breaks the TR symmetry. At low temperatures and high magnetic fields, the electrons flow along the edges of the 2D sample. In Figure 1(a), we present the schematic of a QH bar geometry where the upper and bottom 1D edges are separated by the bulk. At these two spinless 1D edges, electrons propagate in a chiral fashion i.e. only in forward (right moving) or backward (left moving) direction. This is in contrast to normal 1D systems where electrons can flow in both the directions. Hence, the top or bottom edge of a QH bar contains only half the degrees of freedom compared to a normal 1D system. When an edge-state electron encounters an impurity, it still propagates along the same direction as backscattering is prohibited along the same edge. This is the key reason why the QH effect is topologically robust. Such dissipationless transport mechanism can be very useful for semiconductor devices.

Now, one asks the question, can we still realize a QH effect without a magnetic field i.e. without breaking TR symmetry? In recent times, it has been observed that certain materials with strong spin-orbit coupling (SOC) can exhibit such intriguing phenomena. SOC arises in a material due to inversion asymmetry as well as crystal asymmetry. It is a relativistic effect and acts like an internal magnetic field without violating the TR symmetry. Within such materials, we can leave the spin-up forward mover and the spin-down backward mover on the upper edge. In the bottom edge,
the spin and the associated momentum directions are reversed. This is illustrated in Figure 1(b). A system with such edge states is said to be in a QSH state, because it has a net transport of spin forward along the top edge and backward along the bottom edge, just like the separated transport of charge in the QH state. This phenomenon is known as QSH effect which was independently predicted by Kane–Mele [12] and Bernevig–Huges–Zhang [13] in certain theoretical models with SOC.

Although, QSH edges consist of both forward and backward movers, backscattering by non-magnetic impurity is still forbidden. The reason behind this can be attributed to the fact that to have backscattering, spin of the carriers also has to be flipped. Such spin-flip scattering process is forbidden for a non-magnetic/scalar impurity. If the impurity carries a magnetic moment, then the TR symmetry is broken, and backscattering is possible due to the spin-flip process caused by the magnetic impurity. In that sense, the robustness or topology of the QSH edge state is protected by the TR symmetry. The possibility of obtaining symmetry protected, dissipationless spin current through QSH systems can be very useful for future generation spintronic devices [14].

3. 2D Topological Insulators

We already mentioned that SOC i.e. coupling between spin and orbital motion is a relativistic effect most pronounced in heavy elements (elements with large Landé g-factor). Although all materials have SOC, only few of them turns out to be topological insulators. Here, we discuss a general mechanism for finding a TI [13, 4]. It was predicted particularly for mercury telluride (HgTe) quantum wells which is believed to be a 2D TI.

The typical band dispersion of a 2D TI is shown in Figure 2. Here, the bulk conduction band and the bulk valence band is separated by an insulating gap like an ordinary band insulator. The 1D helical edge states appear within the gap with a linear dispersion. The general mechanism behind the appearance of such edge states is band inversion in which the usual ordering of conduction band and valence band is inverted by SOC. This mechanism, we discuss next in detail, for the case of HgTe.

In most common semiconductors, the conduction band is formed by $s$ orbital electrons, and the valence band is formed from electrons in the $p$ orbital. However, in some heavy elements like Hg and Te, the SOC is so large that the $p$ orbital band is pushed above the $s$ orbital i.e. the bands are inverted. HgTe quantum wells can be fabricated by sandwiching the material between cadmium telluride (CdTe) (see Figure 3), which owns similar lattice spacing as HgTe but has weaker SOC. Therefore, as one increases the thickness $d$ of the HgTe layer, the SOC strength is enhanced for the entire quantum well. For a thin quantum well, as shown in Figure 3(a), CdTe has a dominating effect and the bands follow normal ordering. The $s$-like conduction sub-band $E$ is located above the $p$-like valence sub-band $H$, and the system behaves like a trivial insulator.
The signature of 2D TI was observed in a recent experiment in which HgTe quantum wells were grown by molecular beam epitaxy method.

With the enhancement of $d$ above a critical thickness $d_c$, the H sub-band is pushed above the E sub-band by SOC as illustrated in Figure 3(b). Due to band inversion, a pair of gapless 1D edge states carrying opposite spins appear, and they disperse linearly all the way from valence band to conduction band (see Figure 2). This pair of edge states is also known as ‘Kramer’s pair’ (TR partner) and cannot be removed by external perturbations. This is one of the topological signatures of a 2D TI.

The signature of 2D TI was observed in a recent experiment [8] in which HgTe quantum wells were grown by molecular beam epitaxy method. The thickness of HgTe layer was tuned by a gate voltage. It was observed that when $d > d_c$ ($d_c \sim 6.5$ nm) i.e. the system is in the topological phase, conductance appears to be quantized ($2e^2/h$) as the two edge states of TI act as two conducting 1D channels contributing $e^2/h$ each. In contrast, when $d < d_c$, conductance comes out to be vanishingly small akin to trivial band insulator.
4. 3D Topological Insulators

Here, we briefly discuss the phenomenology of 3D TI [5]. Note that, the pair of 1D edge states for our previous 2D TI crosses at $k = 0$ which is already depicted in Figure 2. Near the crossing point, the dispersion of these states follow a linear relation. This is exactly the dispersion relation one obtains in relativistic quantum mechanics from the Dirac equation for a massless free fermion in 1D. Thus, the same equation can be used to describe QSH edge states. Similar picture can be generalized to a 3D TI which owns 2D surface states at the boundary. These surface states consist of 2D massless Dirac fermions and the corresponding dispersion forms a single Dirac cone as depicted in Figure 4. Similar to the 2D case, the crossing point of the surface states is located at the tip of the cone i.e. $k_x = k_y = 0$. The latter is also a TR invariant point at which Kramer’s degeneracy is protected by the TR symmetry. Also note that, each momentum at the surface has only a single spin state at the Fermi level $E_F$ (spin-momentum locking), and the spin direction rotates as the momentum moves around the Fermi surface (see Figure 4). Thus, these surface states exhibit non-trivial spin textures and carry a geometrical Berry’s phase\(^3\) \([15]\) of $\pi$, which makes them topologically distinct from ordinary surface states. When disorder or scalar impurities are incorporated on the surface, backscattering is prohibited, and the metallic surface states remain robust against disorder i.e. they do not become localized or gapped.

From materials point of view, bismuth telluride (Bi\(_2\)Te\(_3\)) and bismuth selenide (Bi\(_2\)Se\(_3\)) are examples of 3D TI. These materials have been investigated experimentally using angle resolved photo emission spectroscopy (ARPES) method \([10, 11]\) and the single Dirac cone nature of the surface states was experimentally observed.

\(^3\) Berry’s phase is a geometrical phase acquired by the wave-function of a quantum particle, subjected to a Hamiltonian depending on slowly varying (adiabatic) time dependent parameters.

---

Figure 4. Schematic of the surface dispersion relation of a typical 3D TI. The 2D surface states reveal a single Dirac cone. Rotation of the spin degrees of freedom around the Fermi surface exhibits spin texture.
Furthermore, spin-resolved measurements probe the spin textures of the surface and confirm that the electron’s spin indeed lies in the plane of the surface and is always perpendicular to the momentum, which is in agreement with the theory.

5. Topological Superconductors

Topological insulators, discussed above, are not superconductors by themselves. However, superconductivity can be induced in them via a process called ‘proximity effect’. In this process, if a non-superconducting material is kept in close contact to a bulk superconductor, then superconducting correlation can tunnel through the non-superconducting material up to a certain length scale, depending on the dimensions of the system and the interface of the two kinds of materials.

Fu and Kane in their seminal work [16], adopted this idea and proposed that if one places a 3D TI material (Bi₂Te₃ or Bi₂Se₃) in close proximity to an ordinary superconductor, then superconductivity can be induced in it via the proximity effect (see Figure 5). Moreover, the interface between a topological insulator and a superconductor may allow the creation of an ‘emergent’ particle that neither material supports by itself. As discussed earlier, the 2D surface electrons of 3D TI are massless Dirac fermions. Hence, a superconductor deposited on the surface opens an excitation gap, which can be closed locally by a magnetic field. The magnetic field penetrates as an Abrikosov vortex shown in Figure 5. If a vortex line runs from the superconductor into the topological insulator, then a zero-energy Majorana fermion (MF) is trapped in the vicinity of the vortex core as illustrated in Figure 5. Therefore, a proximity induced TI with zero-energy MF is commonly known as topological superconductor.

![Figure 5. Schematic of a topological superconductor where a 3D TI is placed in close proximity to a conventional s-wave superconductor. Majorana fermion is formed in the vicinity of the vortex core.](image_url)
Arijit Saha and Arun M Jayannavar

6. Majorana Fermions

The underlying mechanism behind the emergence of MF in 3D TSC is rather a complex subject, and we refer to [7, 17, 18, 19] for further details. On the contrary, we discuss here a more physical picture of MF based on 1D systems [20] which has been experimentally investigated very recently. To start with, one can ask the following question: What is MF and what is its significance from the application point of view? MF was originally proposed by Ettore Majorana in 1937 while finding the real solution of Dirac equation. Majorana fermions (MFs), occurring at exactly zero energy (also known as Majorana zero modes), have the remarkable property of being their own antiparticles. In nanoscience and condensed-matter physics, being its own antiparticle means that a MF must be an equal superposition of an electron and a hole state. Mathematically, this property can be expressed as an equality between the particle’s creation and annihilation operators \( \gamma^\dagger = \gamma \). Also, MFs are massless, spinless, and electrically neutral. Furthermore, Majorana zero modes are believed to exhibit a special kind of quantum statistics – so called non-Abelian exchange statistics – which is neither Fermi–Dirac, nor Bose–Einstein like. This special property endows MFs to be used as a building block for the next generation topological quantum computers, which would be exceptionally well protected from errors or decoherence [21].

The first toy model, for the realization of MF in 1D, was put forth by Kitaev [22]. In this model, one starts from a 1D tight-binding chain with unconventional spinless \( p \)-wave superconducting pairing. The superconducting gap \( \Delta \), and hopping \( t \), are assumed to be the same for all lattice sites. \( \mu \) is the chemical potential set for the system. The Majorana physics can be understood for a special parameter regime when \( t = \Delta \) and \( \mu = 0 \). In this regime, the chain becomes TSC, and two unpaired zero energy MFs \( \gamma_0 \) and \( \gamma_N \), are located non-locally at the two ends of the chain. This feature is illustrated in Figure 6. In the general case, however, the two MFs are not completely localized only at the two end sites of the chain, but decay exponentially away from the end. The MFs remain at zero energy only if the chain is long enough so that they do not overlap. For a finite chain, the two unpaired Majorana wave-functions can also overlap and become a normal Dirac fermion.

Figure 6. Cartoon of two unpaired MFs \( \gamma_0 \) and \( \gamma_N \), located at the two ends of the 1D chain. The intermediate ones have paired to become Dirac fermions.
7. Realization of MF in 1D nanowire

Kitaev’s chain can be realized in a 1D nanowire (NW) made of a semiconductor with strong SOC. The basic idea came from two independent seminal works of Oreg–Refael–von Oppen [23] and Lutchyn–Sau–Das Sarma [24]. If a 1D NW is placed in close contact to a conventional bulk s-wave superconductor, then superconductivity can be induced in the NW via the proximity effect as shown in Figure 7(a). A constant magnetic field $B$ is applied parallel to the NW. Here, $B_{so}$ denotes the direction of the spin-orbit field in the NW. Note that the direction of $B$ and $B_{so}$ should be perpendicular to each other to realize the desired phenomena. Under suitable circumstances, this NW becomes a TSC, and a pair of MFs appears at the two ends of the NW. These two non-local Majorana bound states are denoted by the two green dots in Figure 7(a).

After setting up the basic architecture required for our purpose, now we discuss how the MFs emerge in the NW. The red and blue curves in Figure 7(b) illustrate the band structure of the NW in the limit $B = 0$. The strong SOC, present in the NW, shifts the two parabolic bands depending on their spin polarization along the axis of the spin-orbit field. Switching on a magnetic field $B$, TR symmetry is broken, and a Zeeman gap opens up at $k = 0$, which is the crossing point of the two parabolas. The new band energies are sketched by the black solid curves of Figure 7(b). When the Fermi level $\mu$ resides within the gap, the wire appears ‘spinless’ (see Figure 7(b)).

Next, we introduce the proximity induced pairing potential $\Delta$. Hence, the gap at zero momentum \( i.e. k = 0 \) decreases with the enhancement of $\Delta$ and closes completely when $B \geq \sqrt{\Delta^2 + \mu^2}$. In this situation, the NW enters into the topological superconducting phase. Now, if we focus on the ‘spinless’ regime and project away the upper unoccupied band, then an effective intraband $p$-wave pairing mediated by $\Delta$ appears, which connects smoothly to the phenomenon demonstrated before by the Kitaev’s toy model. Since the NW is in the topological superconducting phase, a pair of MFs emerge localized at the wire’s ends (see Figure 7(a)). For larger values of $\Delta$, the gap reopens, but now in a non-topological superconducting state where the NW no longer appears ‘spinless’ resulting in a trivial phase. Therefore, the phase transition between...
the topological and non-topological superconducting states can only take place at the point where the gap at \( k = 0 \) closes. This can be achieved by satisfying the criterion \( B = \sqrt{\Delta^2 + \mu^2} \).

Very recently, the signature of MFs has been experimentally observed by Mourik et al., [25] and Das et al., [26] in 1D NW systems. The experiment has been performed in a set-up similar to Figure 7(a). The NW is made of indium antimonide (InSb) or indium arsenide (InAs) which has a large ‘\( g \’-factor (\( g \approx 50 \)) i.e. strong SOC. Niobium nitride (NbN) is used as a bulk superconductor to induce superconductivity in the NW. The signature of MFs is revealed via the transport measurements. When the external magnetic field, applied parallel to the NW, satisfies the criterion \( B = \sqrt{\Delta^2 + \mu^2} \), zero energy Majorana bound states appear at the wire’s ends. The tunneling conductance shows a large peak (quantized to \( 2e^2/h \) in ideal conditions) at zero-bias when the Majorana mode is present, and no peak when it is absent. Such zero-bias peaks can be interpreted as an experimental evidence for the Majorana zero mode.

8. Conclusions and Outlook

In this article, we have provided a pedagogical introduction to the exciting field of topological insulators, topological superconductors, and Majorana fermions in condensed matter systems. We emphasize that a 2D TI has 1D helical edge states exhibiting QSH effect. A 3D TI supports 2D surface states which form a single Dirac cone. We also discuss that these systems can enter into a topological superconducting phase supporting Majorana fermions at the vortex core. Finally, we illustrate that topological superconducting phase can also be realized in a 1D NW with strong SOC and proximity coupled to ordinary \( s \)-wave superconductor. In the topological superconducting phase, two non-local Majorana zero modes appear at the two ends of the NW. The zero bias peak, appearing in the tunneling conductance signal, reveals the experimental signature of MFs.

In TI systems, one of the recent interest is to understand the effects of electron-electron interaction in them. A TI with strong Coulomb interaction is called a ‘fractional topological insulator’ which is one of the current topics of research in this direction. Interface between TI and other non-topological materials is also a subject of modern interest. In the context of MF, the smoking gun signal of Majorana zero mode is still lacking. Further experimental investigations is needed to explore their physical properties in detail. Moreover, there is a need for additional theoretical research to understand the experimental findings. The final goal is of course to be able to control and manipulate quantum information stored in Majorana-based qubit systems that can be implemented for topological quantum computation.

Acknowledgment

Arun M Jayannavar thanks DST, India, for financial support (through J. C. Bose National Fellowship).
Suggested Reading


---

*RESONANCE-75 | Promoting Science Education*
In this article it is pointed out how different layers of substructure of matter were revealed to us by experiments which were essentially very similar to the famous $\alpha$-particle scattering experiment performed by Rutherford. This experiment, which revealed the nuclear structure of an atom, paved the way towards our current understanding of the fundamental constituents of matter and shaped the course of physics for the 20th century.

1. ‘Looking’ Inside Matter

We are all celebrating the centenary of the remarkable discovery by Rutherford that all the positive charge and almost all the mass of an atom is concentrated in a tiny region, christened ‘nucleus of an atom’ by him. This discovery, in fact, shaped the course of physics for the entire century\(^1\). After the discovery of atomic structure of matter, this was the next step into our journey towards an understanding of what lies at the ‘heart of matter’. Rutherford in fact, ‘split’ the atom! The importance of this step is underscored when we note that a physicist like Feynman had hailed our knowledge of atomic structure of matter as the one piece of understanding worthy of passing on to the future, should all but one piece of the entire scientific knowledge be destroyed. The series of experimental and theoretical investigations that began with Rutherford’s experiment have now helped us understand that the basic building blocks of nature are quarks and leptons, the quarks making up protons/neutrons which in turn make up the nuclei that form then atoms along with electrons. The atoms in turn make the molecules and so on. In fact, the simple drawing in Figure 1 depicts how different layers of structure of matter have been revealed at different distance/energy scales. This picture helps us appreciate the magnitude of importance of Rutherford’s discovery.

In general, there have been two basic ways in which physicists have arrived at this current understanding of the substructure of matter. One is by noting similarities and patterns in the properties of the composites – like atoms, nuclei and various particles such as proton, neutron, pions, etc., – and the second is to scatter off beams of particles from a target. Rutherford’s experiment has pioneered the second way. In some sense the experiments being carried out today at the LHC (Large Hadron Collider) – where relativistic beams of protons or positive heavy ions collide with each other, hunting possibly for answers about the laws of physics that function at the heart of matter and at the beginning of the Universe – are but a logical conclusion of the kind of experiment performed by Rutherford.

\(^1\) It is interesting that Rutherford’s Nobel Prize was for the ‘Chemistry of Radioactive substances’! This makes Rutherford one of the two scientists, along with Albert Einstein, who did not get the Nobel Prize for their most well-known work!

Keywords
Rutherford, structure of photon, quark, deep inelastic scattering.
The experiments being carried out today at the LHC are but a logical conclusion of the kind of experiment performed by Rutherford.

## 2. Large-Angle Scattering Reveals the Atomic Nucleus

Let me recall here the basic arrangement of Rutherford’s experiment (Figure 2). He studied scattering of energetic $\alpha$ particles (the beam) emitted by radioactive nuclei from a thin gold foil (the target) and the scattered $\alpha$ particles were counted with a microscope through the scintillations these produced on hitting the zinc sulphide screen (the detector). Here the experimental observation was that the fraction of $\alpha$ particles scattered at large angles was much larger than expected if the positive charge in the atom was spread out all over the atom. In Rutherford’s words, “It was about as credible as if you had red a fifteen inch shell at a piece of tissue paper and it came back and hit you.” A simple understanding as to why this indicates that the charge and mass was concentrated at a ‘point’ can be obtained by looking at Figure 3. As one sees from Figure 3a, the smaller the impact parameter (perpendicular distance from the central line) of the approach of the $\alpha$ particles, the higher will be the angle through which it will get deflected. If the electrostatic charge felt by the incoming $\alpha$
The substructure of an atom was revealed through the observed angular distribution of the scattered \( \alpha \) particles.

Figure 3. A schematic depiction why substructure will lead to more large angle scattering. Here \( b_2, \ldots, b_5 \) etc., the perpendicular distances from the central line, are the impact parameters of particles 2, \ldots, 5.

Now, the watchful among the readers may point out an obvious flaw in what I have said so far. If Rutherford’s experiment showed that the atom had a point-like nucleus, how is it that as per our current wisdom summarised in Figure 1, this ‘point-like’ nucleus is further made up of nucleons: neutrons and protons. This has simply to do with our ability to ‘resolve’. Recall that even when we decipher the structure of an object visually, it involves scattering of light from the object which we see with our eyes and/or microscopes. In this case, smaller the wavelength of the light used, higher is the resolving power. The resolving power possible in the scattering experiments like that of Rutherford is decided by the ‘wavelength’ of the probing beam particles is concentrated in a small region, one would expect more scattering through large angles than otherwise would be the case (Figure 3b). The \( x \)-axis here is the angle of scattering. This shows how the substructure of an atom was revealed through the observed angular distribution of the scattered \( \alpha \) particles.
Rutherford’s experiment showed that the atom had a point-like nucleus; our current wisdom is that this ‘point-like’ nucleus is further made up of nucleons: neutrons and protons. This has simply to do with our ability to ‘resolve’.

“It has long been my ambition to have available a copious supply of atoms and electrons which will have energies transcending those of the $\alpha$, $\beta$ particles.”

– Rutherford

Rutherford’s wish was fulfilled by Cockroft and Walton who built, in 1932, the first accelerator.

You can get the information on the actual 2-mile long accelerator which was built at Stanford and the detector which was used to detect and measure properties of the scattered electron from the website of the SLAC laboratory:
http://www2.slac.stanford.edu/vvc/nobel/1990nobel.html
or Interactions.org or http://www.physics.ox.ac.uk/documents/PUS/dis/SLAC.htm

of particles. De Broglie’s hypothesis of wave–particle duality, put forward in 1924 following Einstein’s idea that light shows both wave like and particle like behaviour, implies that a beam of particles with momentum $|\vec{p}| = p$, is given by $\lambda = h/p$. Here $h$ is the Planck’s constant. While for a particle moving with a non-relativistic velocity $v = |\vec{v}|$, the momentum is given by $p = m_0 v$ for a particle of rest mass $m_0$, in general the magnitude of the momentum $p$ is related to the energy $E$ of the particle via $E^2 = m_0^2 c^4 + p^2 c^2$. Since the energies of the $\alpha$ particles used by Rutherford were of the order of a few million electron volts (1 eV being the energy gained by an electron when it falls through a potential difference of 1 volt), his probe could resolve a size greater than 1/100 th of an Angstrom. As we know now the size of a nucleus is a few Fermi, which is 100 thousandth of an Angstrom. Thus as far as Rutherford’s experiment was concerned, the nucleus was a ‘point’ with an extension smaller than the least count of his probe, viz., the wavelength of the $\alpha$ particles. Thus this probe could resolve the atom with the size of few Angstroms into a point-like nucleus and electrons, but not further.

3. Nuclear Analog of Rutherford’s Experiment

The study of properties of nuclei and observed systematics in their masses, magnetic moments, etc., had already indicated to the physicists that the nuclei too may be composites of nucleons. To determine the spatial distribution of the mass and charge of a nucleus and/or a nucleon one needs to do a nuclear analog of Rutherford’s experiment, but now with beams of particles accelerated to high energies such that the corresponding wavelengths, given by de Broglie’s formula above, are smaller than those of the $\alpha$ particles used by Rutherford. Again to quote Rutherford, “It has long been my ambition to have available a copious supply of atoms and electrons which will have energies transcending those of the $\alpha$, $\beta$ particles.” Rutherford’s wish was fulfilled by Cockroft and Walton who built, in 1932, the first accelerator (and won a Nobel Prize for that in 1951!). Beginning from there, the development in accelerator physics and nuclear/particle physics went hand in hand.

In 1951, the first electron–nuclei scattering experiments were done with electron beams of about 15.7 MeV. In 1953, Hofstadter performed the nuclear analog of Rutherford’s experiment with still higher energy electron beams, accelerated to energies of a few hundred MeV. These electrons had wavelengths substantially smaller than those of the 7 MeV $\alpha$ particles used by Rutherford.

Note the similarity between the beam–target–detector arrangement in Figure 2 and that in Figure 4. The process studied by Hofstadter was

$$e^- (E_e) + A \rightarrow e^- (E'_e) + A.$$  

In Rutherford’s experiment the detector could be moved easily to measure $\alpha$ particles scattered at different angles. In case of Hofstadter’s experiment the huge detector built for the purpose, could catch the scattered electrons at a few fixed angles. Once the
angle of the scattered electron was given, the energy $E'_e$ was uniquely fixed. It is not too difficult to calculate the fractional number of electrons which would be scattered through a solid angle $d\Omega$. $d\sigma/d\Omega$ is proportional to this fraction. It is possible to show that, for a spherically symmetric charge distribution,

$$
\frac{d\sigma}{d\Omega}_{\text{charge distn.}} = |F(Q^2)|^2 \left( \frac{d\sigma}{d\Omega} \right)_{\text{point}},
$$

where $F(Q^2)$ is nothing but the Fourier transform of the ‘normalised’ charge distribution $\rho(\vec{R})$, with $Q^2 = |\vec{q}|^2$. $\vec{q}$ is the momentum transfer from the incoming electron to the scattered electron, given by $\vec{p}_e - \vec{p}'_e$. $Q^2$ is a more convenient variable than the scattering angle and is clearly related to it. The distribution of the scattered electron at different angles, which in turn gave $F(Q^2)$, indicated that the charge of the nucleus was concentrated in a region of size of the order of a few Fermi’s ($1$ Fermi = $100000$th of an Angstrom). Just like before, the information on electrons scattered at different angles indicated the presence of a mass/charge distribution and gave an idea of its extension. In principle, similar measurements can also yield information on the ‘shapes’ of nuclei as well.

Figure 4. The schematic of the nuclear analog of the Rutherford experiment performed at the Stanford Linear Accelerator (SLAC).

Figure 5. The panel on the left shows the Stanford Linear Accelerator and the one at right shows the detector. Courtesy: SLAC
Incidentally, this way of obtaining information about mass/charge distributions of the target is very similar to how one obtains the crystal structure by looking at the X-ray diffraction patterns for solids.

To summarise, the spatial distribution of the target modifies the $Q^2$ dependence compared to the expectation for a point scatterer. For a point scatterer, by definition $F(Q^2)$ will be a constant. In fact it can be shown that at $Q^2 \ll 1/\langle R^2 \rangle$, $F(Q^2) = 1 - \langle R^2 \rangle Q^2 / 6$. This formula then very clearly explains why Rutherford ‘saw’ the nucleus to be a ‘point’ even though now we know it to be as big as a few Fermi’s. Our ability to infer and study the structure of an object from scattering experiments is possible only when $\langle R^2 \rangle Q^2 \approx 1$. That is, smaller the spatial extension, higher the energy required to see the evidence for this structure. Hence, for energies such that $\lambda_e \sim R_{\text{target}}$, one sees a deviation of $F(Q^2)$ from $\sim 1$. This is exactly how one got information on the spatial extension of the nucleus.

4. Proton Too has Finite Size as well as Constituents!

Since the nucleus was expected to be a composite of nucleons the finite size of the nucleus did not come as a surprise to anybody. The next step was to perform the analog of Rutherford experiment but with a proton target; the nucleus was replaced by a proton, and electron beam energies were raised to 2000 MeV and above (Figure 4).

It is interesting to note that physicists had reasons to expect that the proton too may be a composite. The value of 5.58 found experimentally for its gyromagnetic ratio was very different from the value 2 of an electron. It was known that for any spin-half charged point particle the expected value for the gyromagnetic ratio is 2 according to the Dirac equation, which is nothing but a relativistic wave equation for a spin-half particle. Worse still, neutron which is neutral should have no magnetic moment at all, but in reality has magnetic moment $= -\frac{|e|}{2M_p}$. This indicated that the proton and neutron must be at least charge distributions.

Hofstadter performed the proton/neutron analog of the Rutherford experiment in 1954. The exact process studied by him was:

$$e(E_e) + p \rightarrow e(E'_e) + p.$$  \hspace{1cm} (1)

Energy and momentum conservation suggests that for a given value of incident electron energy $E_e$ and scattering angle $\theta$, the scattered electron energy $E'_e$ should in fact have a fixed value $E_0$, given by

$$E_0 = \frac{E_e}{1 + 2E_e/M_p \sin^2(\theta/2)}.$$  \hspace{1cm} (2)

The finite size of the proton was confirmed by these scattering experiments (just like that for nuclei) from measurements of the form factors $F(Q^2)$ for protons/neutrons and studying their $Q^2$ dependence. The experimental results indicated that the size of the proton is $\sim 100,000$ times smaller than an atom: a Fermi. Hofstadter was awarded the Nobel Prize in Physics for the year 1961 for this research into sizes of nuclei and nucleons.
A natural question to ask then was: Is this proton just a charge distribution OR is there something inside? The real surprise came when $E_e$ was increased even further! The process studied now was slightly different from the good old Rutherford scattering:

$$e^- + p \rightarrow e^- + X,$$

where $X = \pi, K, p, \bar{p}, \ldots$; and we sum over all $X$. Such a measurement is called inclusive measurement. Since the final state now contains many more particles than the proton, this process is called Deep Inelastic Scattering (DIS). The energies of the electron $E_e$ were now increased yet again by a factor of 5–10 more, to 10,000–20,000 million electron volts (10–20 Giga electron Volts: GeV). The resolution possible with this electron beam was now 1/100 compared to the size of the $p/n$. So this beam should be able to ‘see’ inside the proton and find out whether it too was made up of something more fundamental.

The surprising fact was that unlike the elastic scattering process of equation (1), the energy of the scattered electron $E'_e$, for a given angle of scattering, was found to have many different values and not just one single value $E_0$. This led the physicists to suspect that, maybe, the $p$ had something inside it. At still higher values of $E_e$, the scattered electron again began to have a unique value $E'_0$, different from that for a proton $E_0$, given by equation (2). This observation in the experiment performed by Friedman, Kendall and Taylor in 1968 could be interpreted, through the insight offered by Feynman and Bjorken, to mean that at these high energies the $\lambda_e$ was now small enough to feel the individual scatterers inside the proton\(^4\). The experimentalists were awarded the Nobel Prize in the year 1990. The main point here was again that the experiments indicated the presence of scattering at larger angles than expected if the proton did not have a structure, just like Rutherford’s experiment.

The above-mentioned experiments on electron–proton scattering at the Stanford Linear Accelerator (SLAC) were followed, between 1970-1990, with those using $\mu$ (muon) and $\nu$ (neutrino) beams of energies upto 800 GeV incident on nuclei and protons. The last word in probing the structure of the proton by colliding it with an electron beam was given by the experiments at the ep collider HERA at the German high energy physics laboratory DESY in Hamburg, which collided $e^-/e^+$ beams of 30 GeV on proton beams of 800 GeV. This corresponds to doing an experiment with electron beams of energy ~ $10^5$ GeV incident on a stationary proton target\(^5\). The corresponding wavelengths of the electron beam are then of the order of $10^{-18} \text{ m} = 10^{-3} \text{ fm}$. These experiments have not revealed any evidence for further substructure. Not just that, the predictions for a very wide class of processes from theories which assume the quarks to be absolutely point-like up to very high energies seem to agree with experimental measurements to a very high degree of accuracy. Thus we have an ‘indirect’ indication that the quarks do not seem to have a substructure. It is a matter of great interest that even now, one of the first experimental results at the LHC, was about putting limits on the substructure of quarks \cite{1,2}, by once again looking whether we have an excess of scattering at higher angles than expected from our theories. So the methodology of the Rutherford experiment continues to guide us even today!

\(^4\) It is another very interesting story that similar experiments, at still higher energies, could be used to show that these scatterers with an electromagnetic charge inside the proton, seemed to have exactly the same properties as the quarks postulated by Murray Gell-Mann and the neutral scatterers could be identified with the gluons required in a theory of strong interactions put forward by Gell-Mann, Fritsch and Leutwyler. But that is a story for another day.

\(^5\) Compare this with the energy of 10–20 GeV of the first electron beams at SLAC, which discovered that proton has quarks inside it.

The methodology of the Rutherford experiment continues to guide us even today!
The discussion in the article indicated how one track of scientists' journey in the quest of elementary constituents of matter, essentially used extensions of the pioneering experiment by Rutherford in 1911. These were three different types of experiments, one using $e^-$, $\mu^-$ and $\nu$ beams of energies $O(1)$ TeV with stationary nuclei as targets, second one using colliding electron-proton beams (this was equivalent to a Hofstadter experiment with $e^\pm$ beams of energies $\sim 50$ TeV) and the third one was experiments with $pp$ collisions, with each $p$ beam of energy $\sim 6$ TeV such as the LHC.

A parallel track on which this journey proceeded was of the development of a mathematical framework to describe the interactions among these fundamental building blocks of nature. This development of understanding of four fundamental interactions as being mediated by force carriers which themselves are elementary particles was also made possible because of the same experiments mentioned above. These scattering experiments where fundamental constituents scatter against each other, were able to provide evidence for the existence of all the particles predicted from theoretical considerations. The discovery of the heaviest boson, the Higgs boson $h$, in 2012 at CERN, is the latest contribution of the 'Rutherford-type' experiments (a full century after the original experiment) in shaping our understanding of the elementary particles and fundamental forces among them.

Figure in the left panel below shows one of the processes for production of the $h$ at the LHC, in the scattering of gluons inside each of the colliding protons. The $h$ thus produced decays instantaneously. An event in which a Higgs is produced and decays into two photons at the LHC, as seen in the CMS detector, is shown in the right hand panel. The figure on the right hand side is taken from https://home.cern/news/series/lhc-physics-ten/higgs-boson-revealing-natures-secrets

Suggested Reading


We humans are perennially searching for new sources of energy. This is because we are puny creatures but our dreams are big. Our physical prowess is limited, but we want to move mountains, change the course of rivers, and fly to the stars. We need a genie to come out of the proverbial bottle at the flick of a finger and do our bidding! Energy is such a genie and we never seem to have enough of it. This article is devoted to a short history of our search for new sources of energy including the latest quest – to light and sustain starfire (the fire of nuclear fusion) in the laboratory. This search has led to several encounters with Agni, the Lord of Fire in one of his manifestations, in fashioning fuels for storage and future use or in burning up fuels for energy release. As we trace this history, we stand in awe of Agni and his various manifestations and cannot help but follow our ancestors in singing a hymn in his praise – hence the title of this article.

The modern human being needs a minimum amount of energy to live a life of comfort and dignity. In fact, the per capita electricity consumption in a given country is a good indicator of the human development index of that country. India has been moving up this graph steadily after independence and is still a long way away compared to developed countries. Developing countries like India are still at 1/5th of the average world consumption, 1/17th of that of the average developed world and 1/50th of that of North America. As more and more of the developing countries climb up the curve of the human development index, the overall requirements of energy in the world will be enormous. Thus the search for new sources of energy continues, nay, gets intensified, year after year.

As mentioned above, the average electricity consumption per person for living a comfortable life (i.e., for heating, cooling, cooking, lighting, transporting, manufacture, etc.) in the developed world is 6000 Kilo Watt Hours (KWH) / year. This is about 17 KWH or 17 units of electricity per day. We in India have available only 1 KWH per person per day. One KWH is the energy used up by a 100 watt bulb in ten hours of continuous operation or the energy used by a typical air conditioner in half an hour’s operation. It is the equivalent of energy used up in lifting about two thousand 100 kilogram bags of rice through a height of 2 meters. Thus we see that modern creature comforts do not come easy.

Based on the Twelfth Kumari L. A Meera Memorial Lecture, delivered on December 11, 2003, Bangalore.
†Deceased

Keywords
Energy security, fossil fuel, thermo-nuclear fusion, Tokamaks
The use of muscle power to fulfill our energy needs has very limited utility.

1. Muscle Power

This was the first source of energy used by man, first his own and then that of his slaves and animals. The great pyramids in Egypt were built with muscle power and the galley ships of Rome which were used to conquer a good part of Europe were powered by the muscle powers of slaves. Closer to home, even today we see horses, bullocks and camels pulling carts, ploughing fields and grinding oil seeds and men and women engaged in various forms of manual labor. However, it is easy to show that the use of muscle power is of limited utility in a world as hungry for energy as it is today.

Take the case of human labor. The energy source is food. The average intake of food per day per person is about 2500 Calories which is equivalent to about 3 KWH of energy. Most of this energy is used up for the business of staying alive, for example keeping warm (did you know that each of us is radiating about 100 watts of infrared radiation during all our waking hours?), doing repair and maintenance of muscle and tissue, and for growth. At the most we may be able to divert about 10 percent of input energy during waking hours to do useful work of ~0.15 KWH per person per day. This also shows that the use of muscle power to fulfill our energy needs has very limited utility.

Let us now discuss how animals (including humans) manage to do work and generate heat, so essential for life. Basically, animals are controlled furnaces “slowly burning” the fuel of food to generate energy. For example, sugars and fats from the food chemically combine with the oxygen that we take in during respiration and form carbon dioxide and water and release energy. This energy is partly dissipated as heat in our bodies and is partly stored in our muscles for future use. This storage also takes the form of a chemical reaction. Energy is soaked up by the Adenosine Diphosphate (ADP) molecule which combines with inorganic phosphate in the body in an endothermic reaction to form an Adenosine Triphosphate (ATP) molecule. The basic reaction rates are made decent even at the relatively low body temperatures by the strategy of employing special catalysts (enzymes) to accelerate the chemical reactions. ATP is the basic storehouse of energy in our bodies. Whenever we need to do some work, our brain sends signals to our muscles which contract and reconvert ATP to ADP releasing the stored energy which is then used for doing useful work and producing heat. Thus we face our first encounter with Agni. Living matter is a slow fire undergoing combustion and converting energy trapped in food into a usable form.

2. Fossil Fuels

We now turn to the next most basic source of energy, namely, the combustibles – things that burn. This is our direct encounter with Agni. Fire was discovered about 25000 years ago and has dominated human civilization since then. Our early ancestors
discovered natural fire and only learnt to keep it alive by adding combustibles like wood, dry leaves, and fat from dead animals. Later they learnt how to ignite a fire at will. Fire revolutionized human life by contributing to heat, light (freedom from cold and fear), cooking (hence health), weaponry (smelting of ores), and so on. It was really put to work for the first time in the steam engines of the 19th century and led to the Industrial Revolution and the explosive growth of Western civilization. Thus began the age of fossil fuels. The internal combustion engine and the electric generator revolutionized the way we use energy. Coal, oil and natural gas became key commodities. The 20th century has been the century of fossil fuels. Availability and the price of fuels have determined the pace of world economic development. It has also colored the political map of the world. The question we must wrestle with is – can we depend on fossil fuels as the main source of energy in the 21st century and beyond? Let us look at this question in a fundamental quantitative manner.

Fossil fuels like coal, oil and gas are hydrocarbons produced from remains of plants and animals which are trapped for millions of years under the earth due to geological upheavals. They have energy stored in them just like food and they also release it by a process of combustion.

\[ \text{hydrocarbons} + \text{oxygen} = \text{water} + \text{carbon dioxide} + \text{energy} \]

(FIRE)

In either case (food or fossil fuels) energy is liberated because the electronic linkages binding atoms in a fuel are broken and rearranged by the chemical combustion process with the release of energy. Typical binding energies stored in electronic linkages are of the order of a fraction of an electron volt (eV) per atom and a molecule has several such linkages which get rearranged during combustion. The typical amount of energy released per molecule is of the order of 1 eV which is about $10^{-19}$ watt sec (joule).

It is known that one gram molecule (mole) of any substance has $6 \times 10^{23}$ molecules. Thus we can calculate the amount of energy liberated by burning one mole of the fuel. This leads to the conclusion that 1 ton of hydrocarbon fuel typically liberates a few thousand KWH of energy. The total known reserves of fossil fuel on the earth is approximately $10^{12}$ tons, which is equivalent to about $10^{16}$ KWH of energy. On the other hand, the energy required by the 5 billion inhabitants of this earth for a reasonable lifestyle is of order $6000 \times 5 \times 10^9$ KWH per year, i.e. $3 \times 10^{13}$ KWH per year. Thus our fossil reserves will last us only for about 300 years. Detailed estimates show that gas will finish in 50 years, oil in 100 years and coal in about 300 years.

We thus reach the following fundamental conclusions regarding fossil fuels. It has taken almost a billion years for the earth to build up the $10^{12}$ tons of fossil fuel reserves and humanity can squander them away in less than 500 years. It is worth noting that fossil fuels have many other applications (e.g., manufacture of petrochemicals) and form a component of wealth which truly belongs to all our future generations – our children, grandchildren and their progeny. The hard question is, should we unilaterally...
It seems clear that fossil fuels are not a long term solution to the energy problem of the world. The second major issue with continued and enlarged use of fossil fuels for all of our energy needs is the environmental degradation that it causes, and the possibility that it may be leading to the greenhouse effect and the warming of the earth’s atmospheric temperature which could have far reaching and detrimental effects on life on this planet. This may mean that we have to clean-up the environment and that the cost of the clean up has to be added to the cost of fossil fuels. If this is done, the fossil fuels will no longer remain cheap. Another major issue with fossil fuels is their uneven distribution on the globe. Countries with 6 percent of the world’s population are endowed with 66 percent of the oil wealth of the world. This has led to serious political conflicts, wars and a great deal of energy insecurity among nations who have no direct access to the oil wealth. From these and many other similar considerations, it seems clear that fossil fuels are not a long term solution to the energy problem of the world.

We saw that food and fossil fuels had ‘energy’ stored in them in the form of electronic linkages of atoms. How did the energy get stored in these molecules and where did it really come from? We know that it came from the Sun, that fierce manifestation of Agni up in our skies. Let us see how this works.

Solar radiation is directly utilized by the green pigment in plants (chlorophyll) to carry out photosynthesis.

\[
\text{carbon dioxide} + \text{water} + \text{light} = \text{carbohydrates} + \text{oxygen}
\]

Our Sun aims about \(10^{17}\) watts of power in the form of solar radiation in the direction of the earth. About a hundredth of a percent of this power is trapped by chlorophyll in plants and is employed in photosynthesis and the resulting growth of plant matter. About one millionth of the total plant matter on the earth is trapped by the geological upheavals and converted to fossil fuels. Thus \(10^7\) watts of the incoming solar energy is trapped into fossil fuels. This gives us an upper limit to the possible fossil fuel reserves as energy equivalent of \(10^7\) watts for a billion years. This translates into \(10^{17}\) KWH of fuels which is close to \(10^{13}\) tons of fossil fuels. This number is at the same time awesome and frightening. It is gratifying and awe-inspiring because we see that, Agni at work in the form of the Sun, has provided us all this fossil fuel wealth so that we have remained warm and comfortable for all these thousands of years. It is frightening and sobering because we note that the upper limit is close to our known reserves, that is, things are really running out and there is not a lot more there which we haven’t tapped yet.

3. **Renewable Energy Sources**

We now turn to renewable sources of energy. Out of a total of \(10^{17}\) watts of solar energy that is directed at the earth, about two-thirds reaches the earth’s surface and is directly converted to heat. This energy, if trapped, can be used and various modes of its utilization come under the banner of solar energy. About one third is used
up in evaporation, precipitation and storage in the water cycle. This is the energy which we see in mighty rivers rolling down mountainsides and which can be utilized in various forms of water mills, hydroelectric projects, etc. Another half a percent is stored in kinetic energy of winds that can be used by windmills and wind-farms for producing electricity. A really small amount, namely a hundredth of a percent is trapped by photosynthesis for storage in vegetation and plants. Of this, as already stated, a miniscule fraction (one millionth) gets stored by the geological upheavals in the bowels of the earth in the various forms of fossil fuels. Let us see what the prospects of using solar energy in its various manifestations are.

We start with hydropower, the energy of the moving water in the rivers has been used since antiquity. It was used in water mills to grind corn, run irrigation systems, saw wood, power textile looms, and so on. The modern version of usage is in the form of running dynamos in hydroelectric power stations. This form of electricity generation has many good features. It is non-polluting, renewable and relatively less expensive. Its demerits are that its capacity is limited by geography and that it typically takes up a large land area, thus displacing a large number of families. In India, the total potential capacity is about 85 Gigawatts (nearly ten percent of our total requirement eventually). However, the actual installed capacity is less than a third of the total potential and there is substantial room for growth.

If we look at solar power we see that it is too dilute and diffuse to be directly of much use for centralized urban industrial complexes. It may be used mostly for residential applications such as solar cookers, water heaters, solar refrigeration, etc. Problems which have not been yet solved satisfactorily include the problem of storage of the energy during night or cloudy periods. Solar photovoltaics technology, i.e. direct conversion of sunlight into electricity by the use of semiconductors, is relatively inefficient and expensive, and has therefore found application only in niche areas like space vehicle power systems. Another method by which solar energy may be used is through the generation of biomass. Schemes for exploitation include fast-growing cash crops used for direct burning or for conversion into alcohols for use as liquid fuels, or even for conversion into biogas which can be conveniently transported for use in homes and small industries. There were great hopes that these sources could be quickly assimilated and integrated with the rural economies in India. But the process of adoption has been slow and the integration has not really taken off.

Wind power has limited applicability in coastal areas where the wind velocities tend to be high. The installed capacity in India is about 1.7 Gigawatts but its capacity utilization is quite low (about ten percent).

We finally turn towards Agni in the Sun. What is the fundamental source of energy in the Sun? We know that our Sun has been burning brightly for about 5 billion years. It has thus spewed out $4 \times 10^{26} \text{ (watts of radiation)} \times 5 \times 10^9 \text{ (no of yrs)} \times 3 \times 10^7 \text{ (seconds /yr)} \sim 6 \times 10^{43} \text{ joules of energy!}$ The Sun has a mass of about $10^{33} \text{ grams.}$ Even if we assume that every atom of Sun ‘burnt’ in a chemical fire, rearranging electronic
Nuclear energy is perhaps the biggest discovery of the twentieth century. In nuclear fission reactions, heavy elements like uranium are made to fragment into lighter nuclei which have more binding energy per nucleon; and energy is released in the process.

In another type of nuclear reaction, the so-called nuclear fusion process, energy is released when light nuclei like hydrogen combine to form heavier nuclei like helium.

Nuclear energy is perhaps the biggest discovery of the twentieth century. It all started with Einstein, who while working out the consequences of his special theory of relativity, came to the startling conclusion that mass and energy are interconvertible and that mass is actually a very concentrated storehouse of energy. His famous equation $E = mc^2$ says it all. Here $c$ is the velocity of light in vacuum, which is a very large number. Thus each gram of matter, if destroyed, can generate about $10^{14}$ joules of energy and the entire energy liberated by the Sun till now could have been generated by the destruction of only $10^{29}$ gms (i.e., .01 percent) of the solar mass, which would not have had any other measurable or significant consequences.

Let us understand this source of energy a little better. The Bohr-Rutherford model of the atom had already shown that the mass of the atom is concentrated in a tiny nucleus (with a size of order $1/\text{ten million millionth}$ of a centimetre) which is positively charged and made up of protons and neutrons, whereas all the electrons (which are negatively charged and make the atom neutral) form a swarm about $100000$ times larger in size. Hence nuclear reactions, as opposed to chemical reactions (which only rearrange the electronic linkages), could change mass and release energy if some mass is lost in the rearrangement of nuclear matter. There are two significant types of nuclear reactions which may be used for energy release. In the nuclear fission reactions, heavy elements like uranium are made to fragment into lighter nuclei which have more binding energy per nucleon, i.e., have nucleons sitting in deeper energy wells. The result is that energy is released in the process. This is the basis of the atomic bomb and also the power-producing nuclear reactors which are widely used around the world. In another type of nuclear reaction, the nuclear fusion process, energy is released when light nuclei like hydrogen combine to form heavier nuclei like helium. Again the nucleons go into deeper energy wells when they form helium and the energy difference is released as the kinetic energy of the products of the nuclear reaction. Typically, the binding energy inside the nucleus arises because of the strong nuclear forces. Per nucleon is a million times stronger than the binding energy per atom which is due to electromagnetic forces (through the various electronic linkages). Hence, weight for weight, nuclear reactions release a million times more energy than chemical reactions. In other words energy packing in nuclei is a million times denser than in atoms.
The Sun is powered by the energy that is released by the thermonuclear fusion of protons. The reaction

\[ \text{H} + \text{H} + \text{H} + \text{H} = \text{Helium} + \text{electrons} + \text{energy} \]

in which four protons fuse to produce a helium nucleus is catalyzed by the presence of carbon, nitrogen and oxygen nuclei. Thermonuclear fusion works because of the high temperatures inside the Sun. High temperatures are necessary because the positively charged protons would normally repel each other because of Coulomb (electrostatic) forces. At high temperatures, the kinetic energy of motion can give the nuclei ability to approach each other and sufficient proximity assists in quantum mechanical tunneling through the Coulomb barrier and their ultimate fusion. Thermonuclear fusion has been successfully demonstrated on the earth in the so-called hydrogen bombs. However, the peaceful utilization of nuclear fusion energy is more difficult. This is because one needs to find a good containment device for matter at very high temperatures. In the Sun and the stars, matter is contained by intense gravitational fields. However, on the earth that is not practical. Hence magnetic bottles are used. We shall discuss these in more detail a little later.

Fission reactors have been exploited commercially for more than 50 years. Thus today France produces more than 75% of its power in fission reactors, USA 25%, Japan 33%, Korea 40% and so on. India is producing only 3% of its power requirements by the nuclear fission process. Currently, there is an ambitious program of expansion with a desire to produce about 20 Gigawatts of nuclear power by the year 2020.

5. Limitations of Fission Technology

The exploitation of fission technology faces certain important challenges. Firstly, there is the problem of radioactive waste disposal. Fission processes lead to some long-lived radioactive products which have to be properly disposed of. One technology which has been developed is that of reducing the volume of the waste and then immobilizing it in a solid matrix. This solid waste residue is then stored deep inside some unused mines at sufficient depth. This technology has been mastered and appears to be quite safe, but is still facing opposition by environmental groups. Other methods which are under development include actinide burning in fission and fusion reactors wherein the long-lived radioactive waste products are themselves subjected to neutron irradiation to transmute them into short-lived isotopes and also to extract some further energy from the nuclei. These methods are still in an early stage of development. A second problem faced in the case of fission reactors involves radiation hazards and safety. Fear in this regard has accentuated because of well-known accidents like those at the Three Mile Island or Chernobyl reactor accident. Whereas one cannot ignore the concern in the minds of the public, it is to be remarked that with sufficient care and precautions such incidents have been few and far between, and the record of the fission power industry so far has been quite good. There is an irrational fear of
With the fast realization that absence of fission power means either energy starvation and/or facing severe environmental degradation problems, there is an increased realization that fission power has to play its legitimate role in the energy scene of a rapidly developing world.

Fusion looks like the ultimate solution to the energy problems of mankind. Major programs for its development are going on around the world. invisible radiation in the minds of the public which can only be removed by education and actual statistical data on accidents/casualties in fission power industries versus the hazards of competitive energy industries. It is also necessary that this data be collected and disseminated by some independent watchdog agency rather than by the promoters of this or that technology. A third problem faced by the fission reactor technology is the uneven distribution of the fuels like uranium and thorium which makes the energy security problem severe for some nations. Lastly, with fission there is the possibility of proliferation, owing to the fissile material falling in the hands of rogue nations and/or of terrorist groups, leading to serious problems. All these challenges, each one of which can be confronted and solved, have given fission a bad name with the result that the growth in fission power stations has been impeded and there has been a relatively stagnant period in their development for the past decade or so. However, with the fast-growing realization that absence of fission power means either energy starvation and/or facing severe environmental degradation problems, there is an increased awareness that fission power has to play its legitimate role in the energy scene of a rapidly developing world.

6. Relative Merits of Fusion Technology

We now turn to a discussion of the nuclear fusion process. This process has not yet been converted into a commercially viable option. Why should we invest in its development? There are several outstanding merits of this process because of which the stakes in its rapid development and deployment are high. What are these merits?

Firstly, the fuel for this process is essentially limitless. Deuterium is a heavier isotope of hydrogen, and is a component of heavy water, which is naturally distributed in the ratio 1 part in 6000 in sea water. Thus if we can use pure deuterium fusion reactions we have literally oceans of fuel, which can last mankind for millions of years. If we want to use tritium, it can be bred from lithium which is widely distributed in the earth's crust and the oceans. Both deuterium and tritium will be readily accessible to most nations, giving widespread energy security. Secondly, the energy is very clean. There is no atmospheric pollution, no greenhouse gas emission. Thirdly, the radioactivity from waste products in fusion is negligible, and in principle, can be eliminated totally. This is because, unlike fission, the reaction products themselves are non-radioactive. Radioactivity is induced in the surrounding materials by the fast neutrons coming from the fusion reaction. One can reduce its hazards by utilizing low activation materials, which are currently a major area of development. One can also totally eliminate it by the use of advanced fusion reactions like proton–boron reactions and deuterium-helium 3 reactions. However, this will only be possible when the containment problems of high-temperature plasmas by magnetic fields are properly solved. Fourthly, fusion reactions are inherently safe; the reaction is difficult to ignite and there is no possibility of a chain reaction or a melt-down. Lastly, there are no dangers of proliferation, no worries that some rogue nation or terrorist group
will steal strategic material for a nuclear weapon. Thus fusion looks like the ultimate solution to the energy problems of mankind. Major programs for its development are going on around the world.

7. Methods for Achieving Fusion in the Laboratory

How does one achieve fusion in the laboratory? As discussed above, matter has to be kept at a very high temperature so that a sufficient number of fusion reactions might occur. Thus for a deuterium-tritium mixture one needs to create a temperature of about 100 million degrees, and also have the product of the density and confinement time exceed a critical value. Only then can a fusion fire be ignited – only then can the great thermonuclear Agni be invoked. In magnetic fusion one starts with matter at sub-atmospheric densities (say about 100000 times rarer than air), heats it to about 100 million degrees and holds the hot matter away from material walls for several seconds in cages made of non-material magnetic field lines. If this is done, a sufficient number of fusion reactions take place in the fire to throw out energetic neutrons which may be trapped in outer blankets to create heat, generate steam and run steam turbines to generate electricity as in a fission reactor. The energy of the helium atoms produced is re-absorbed in the hot matter to keep it hot even as we keep adding more and more fuel to it. In this manner, a fusion fire once ignited will stay ignited.

Figure 1. Components of a fusion reactor. Tokamak Schematic, is reproduced from the US National Fusion Energy Science Website (http://www.fusionscience.org/technical/ASslides/Asslids.html)
Matter at hundred million degrees is in the plasma state, a state in which each atom is broken up by collisions into nucleus and electrons. Thus a plasma is a collection of hot nuclei and electrons. This swarm of charged particles has many special properties which distinguish it from a normal gas. From the fusion engineer’s point of view it has useful properties: for instance it can be heated by electric currents and microwaves, which allow us to take it to ignition temperatures. Indeed, such temperatures have been achieved in the magnetic containers in various laboratories around the world. But it also has bad properties such as loss of heat by electric and magnetic storms created by collective effects (‘mob effect’ by the swarms of charged particles), which cause the fire to cool down when it should stay hot and ignited. It has taken the science of plasmas several decades to develop enough sophistication to understand why these mob effects arise, how they may be avoided or (sometimes) how they may be used cleverly to assist in lighting and maintaining a fusion fire. Starting from the early days of fusion research, the key parameter has been improved by a factor of about 100000, and now one is within a factor of 5 of what is needed for a successful commercial fusion reactor. In some large tokamaks in USA and Europe, the production of about 20 Megawatts of fusion power steadily for several seconds has actually been demonstrated. The next step is the demonstration that such plasmas can be maintained steadily for hours together so that fusion fire once ignited may continue to give energy by addition of more and more fuel. This is the aim of the next generation of these experiments.
The largest of these experiments, which will have plasmas of sizes very similar to that in a commercial fusion reactor, is being done internationally in the so-called ITER (International Thermonuclear Experimental Reactor) experiment. ITER will be built at a cost of about 5 billion dollars (Rs 25000 crores), which will be shared by the partners USA, Europe, Russia, Japan, Canada, China and South Korea. Each partner has to contribute a minimum of ten percent towards the cost of the project. Furthermore they are also typically supporting a national program of comparable magnitude to stay abreast of the latest developments. The engineering design of ITER is over and construction is likely to start in 2006. The experiment will start yielding key results by the year 2015. Most people believe that the earliest that fusion reactors may be commercially deployed is by the year 2035 and the latest is by the year 2050.

8. Fusion Research in India

Our national program in fusion started in the early 1980s at the Institute for Plasma Research in Gandhinagar. We have indigenously built a tokamak ADITYA which was commissioned in 1989, and on which experiments have been conducted since then. ADITYA is like a miniature artificial Sun which is contained by magnetic cages for fractions of a second and in which temperatures of several million degrees have been achieved. We are now in the process of assembling our second generation experiment, a steady state superconducting tokamak SST1, in which the key feature will be keeping the fusion fire alive for 1000 seconds. This experiment should be operational within a year and is likely to be the first experiment...
Figure 4. SST1, the superconducting steady state tokamak, which will hold a fusion plasma for a 1000 seconds, being fabricated and assembled at the Institute for Plasma Research, Gandhinagar

keeping the hot fusion plasma alive for such a long time. The experiment will therefore yield key information of interest to ITER. The cost of participation in ITER is about ten times larger than our present resources and so it appears that we shall have to keep abreast of the international developments only by conducting a clever national program.

9. Epilogue

We now begin the process of summing up what we learnt by following the trail of Agni in search of sources of energy. In the process we shall also touch upon some new ideas which may form the basis of future options and also take a deeper look at Agni and at ourselves. We learnt that muscle power is out because of its inefficiency and our large needs. Fossil fuels have been extensively used for about a hundred years or so. In the long run the workhorse will be nuclear energy, especially its release using the thermonuclear fusion fire. Renewable energy sources like hydro, wind, and solar power are definitely in, but may play a subsidiary role.
New ideas on improved utilization of solar energy abound, and are active areas of research. Thus it is believed that accelerated photosynthesis and/or dissociation may be achieved by using modern tools of genetic engineering. Work is currently being done on the development of algae/bacteria which would use sunlight and water as inputs and generate hydrogen. Hydrogen is an excellent fuel since it is a gas and can be readily transported by pipes, and is totally non-polluting since it produces only water as a combustion product. Trapping of the energy in sunlight into hydrogen would thus be an excellent development. Another attractive idea for the utilization of solar energy is the solar satellite concept. In this concept solar power stations are put in earth-stationary orbits where they would see the Sun all the time (no problem of night and day), would not occupy valuable real estate on the ground, would not be affected by dust, atmosphere and clouds, etc., and would beam concentrated energy to ground power stations through microwave beams. This concept mitigates many of the shortcomings of solar energy utilization. Detailed design and development work at the component level is in progress. There are also dreamers like Freeman Dyson, who feel that we should not be satisfied with the small fraction (hundred millionth) of its energy which the Sun chooses to beam at us. Dyson would like to trap all of the Sun’s energy by using human engineering skills on a grand scale. Thus he would take a planet like Saturn, break it up and use its material to surround the Sun completely with a spherical shell near the orbit of Jupiter. One would then line the inside of this spherical shell with equipment which converts solar energy into a usable form and transport it to earth for utilization by us, the energy hungry hordes of humans! Dyson believes that technological civilizations more advanced than us might already be carrying out engineering projects on a planetary scale like this.

We now ask a final question in our journey tracing the sources of energy. We found that most of the sources were ultimately traced to nuclear fuels either burning in the nuclear fire of our Sun or some day burning in our reactors. We may ask: from where does the energy trapped in nuclear fuels originate? For this we have to trace the genesis of elements. This problem has been studied in detail by cosmologists and astrophysicists. It is believed that in the beginning there was the Big Bang, and the Primordial Fire or the earliest form of Agni was born. This fire started as a ball of radiation. As it expanded and cooled, it created material particles like quarks and gluons forming the so-called quark-gluon plasma. As the ball cooled further, the quarks condensed into protons, neutrons and some of the lightest elements like helium, lithium, and boron. Soon gravitation rolled dense matter in local regions into fusion furnaces known as stars. Heavier elements like carbon, nitrogen and oxygen were formed in the fusion furnaces of young stars. Iron is the most stable nucleus and would be the end product of all nuclear reactions. Elements beyond iron are formed by neutron capture processes. We see the rich variety of elements because cooking in the primitive universe and stars is imperfect. If the cooking had been perfect there would be iron everywhere and there would be no nuclear fuels.
To put this all into a perspective, we may take a deeper look at ourselves. What are we?
We are highly organized collections of molecules made out of the same constituent
atoms, the same partially cooked elements - hydrogen, carbon, nitrogen, and oxygen.
Each atom of our body is billions of years old. Thus each hydrogen nucleus of my
body was fashioned from quarks in that primeval Agni 14 to 15 billion years ago!
Each carbon, nitrogen and oxygen nucleus here on the earth has been fashioned by
imperfect cooking in the fusion furnace of some star, spewed out into space during its
death throes and picked up by the gas cloud from which our Sun and its planets have
emerged. With this perspective we understand that this whole game is being played by
Agni with its partially cooked range of nuclei. Thus, in spite of our modern scientific
background, in expressing our deepest yearnings, our sentiments are no different from
those of our Vedic ancestors who sang hymns to Agni, as in the Rg Veda.

Thou art the Source
and fashioned every limb of mine...

Thou art the Sustainer
and provide Energy to keep me together
And when I am ready to Sleep
into thy bosom only will I disappear!
Quadrupole Ion Traps*
Localising Charged Particles by Electric and Magnetic Fields

Pushpa M Rao, Richard D’Souza and S A Ahmad

During the last two decades there has been tremendous progress in the technique of trapping and cooling ions using quadrupole ion traps. Using these trapping techniques one can have charged particles of a single species confined near the trap centre, which enables to carry out studies of these ions in a well-controlled environment. The long storage times of the ions, possible in these traps, results in the elimination of transit-time broadening making it possible to do precision spectroscopic measurements on these ions. Several important experiments with single electron or ion have been undertaken to address problems related to basic physics, such as the measurement of the electron radius, precision measurements of fundamental parameters and tests of the predictions of quantum mechanics.

1. Introduction

Wolfgang Paul was awarded the 1989 Nobel Prize in Physics for his immense contributions to the physics and techniques of ion trappings, and in the Nobel Lecture delivered on 8th December, 1989 he said “Experimental physics is the art of observing the structure of matter and of detecting the dynamic processes within it”. However, in order to understand the complicated behaviour of some natural processes, one has to measure the relevant parameters involving the matter-light interaction as precisely as possible. Measurements of atomic characteristics like energy levels of an atom or probability of transition between these levels is usually performed on a collection of atoms. Furthermore, effects such as collisions complicate and modify the values of the parameters relevant for understanding the physics of the atom. However in certain studies it is necessary to carry out investigations on an isolated single atom/ion. Such studies have been carried out recently, resulting in some exciting observations in the field of the interaction of radiation with atoms.

A single atom at rest in free space, free of uncontrolled perturbations, would be the ideal dream of any atomic physicist. Ion traps have almost fulfilled this dream: Observation of a single atomic ion, almost at rest in a nearly perturbation-free environment, over a long period of time has become a practical reality.

The idea of trapping charged particles developed from studies on electrical discharges and has continuously evolved from extensive research in the fields of mass filters and particle accelerators. Investigation of the properties of an electrical discharge between a very thin filament cathode and a cylindrical anode led to the earliest trap

A single atom at rest in free space, free of uncontrolled perturbations, would be the ideal dream of any atomic physicist. called the Kingdon trap in honour of K H Kingdon. Study of the properties of the electrical discharges between coaxial cylinders in the presence of an axial magnetic field by F M Penning in 1936 led to the development of the Penning trap. The important result of this work was that the electron path between the two electrodes could be very long due to the tendency of the magnetic field to force the electrons into circular orbits around the axis. In 1953 Wolfgang Paul in Bonn, investigated the non-magnetic quadrupole mass filter, which revolutionised mass spectrometry. His studies led to the development of the Paul or radiofrequency ion trap for the atomic ions. Since then various kinds of ion traps have been built which cater to specific investigations on the ions. Development of the Penning trap, using a dc electric field along with a magnetic field, was primarily by Hans Dehmelt and his associates in Seattle. Dehmelt trapped a single electron in a Penning trap by an electromagnetic potential. Dehmelt called this single electron bound to the gravitational field, the ‘geonium atom’. The first atomic hyperfine structure experiment on trapped ions was performed by Dehmelt’s group using the stored-ion exchange-collision technique in a Paul trap which paved the way for some of the subsequent experiment for atomic frequency standards. Confinement of charged particles made it possible to investigate some of the most fundamental properties of matter-radiation interaction. The immense impact of these studies was acknowledged in 1989 by the award of the Nobel Prize in Physics to Hans Dehmelt and Wolfgang Paul (half of the prize was awarded to Norman F Ramsay for introduction of oscillatory field in atomic-beam magnetic resonance method).

2. Confinement of Ions

The purpose of an atom or ion trap is to confine the motion of the atomic or ionic particles to a small region of space. In this article we discuss the trapping of charged particles, which is understandably far easier than the confinement of neutral atoms, as the forces which can be exerted by electromagnetic fields on the latter are far smaller.

How does one go about trapping an ion? The simplest solution is to have the ion elastically bound to an origin by a restoring force \( F = -kr \) that increases linearly with the distance \( r \). In other words, if the ion moves back and forth about an equilibrium position through a parabolic potential \( \Phi \) that varies as

\[
\Phi = \frac{1}{2} k r^2 = \frac{1}{2} k (x^2 + y^2 + z^2),
\]

then the particles are elastically bound to the origin. However, any potential in free space has to satisfy Laplace’s equation \( \nabla^2 \Phi = 0 \); it can be shown that the above potential (1) does not. This is the essence of Earnshaw’s theorem, which states that “A charged particle cannot be held in stable equilibrium by electrostatic forces alone” (see Box 1).
Box 1. Earnshaw’s Theorem

Earnshaw’s theorem states that it is not possible to arrange, in free space, charged particles in stable equilibrium. Let us consider the following arrangement in one dimension (Figure A), of a positive charge particle placed in between two negative charge particles. For this arrangement of charges to be in equilibrium, it can be worked out that the magnitude of the negatively charged particles has to be four times that of the positive charge placed at equi-distance ‘a’ from the two negative charges. This arrangement cannot be in a stable equilibrium, for even a slight displacement of any of the charges will render the entire system unstable. It is essentially a consequence of the fact that the potential \( \phi(x) \) satisfying Laplace’s equation \( \nabla^2 \phi = 0 \) can have no maxima or minima.

The potential energies of each charge due to the other two can be easily worked out. For example the potential energy of the charge \(-4q\) in the field of the other two charges is given by

\[
W_{-4q} = -4q \left( \frac{q}{y} - \frac{4q}{a+y} \right) = 4q^2 \left( \frac{3y-a}{y(a+y)} \right),
\]

where \(y\) is the distance between the charge \(+q\) and one of the displaced charges \(-4q\). For \(0 \leq y < \infty\), the potential energy between the charge \(+q\) and the charge \(-4q\) on the left side is shown by the curve ABC (Figure A), whereas for the other \(-4q\) charge on the right for where \(0 \geq y > -\infty\) the potential energy curve is shown by DEF. When the charges \(-4q\) are stationary, the energy of the charge \(+q\) is

\[
W_q = q \left( \frac{-4q}{a-z} + \frac{4q}{a+z} \right) = \frac{-8q^2z}{a^2 - z^2},
\]

where \(z\) is the displacement of the charge \(+q\) from the equilibrium position. The potential energy curve of charge \(q\) for \(0 \leq z \leq a\), is given by the curve MNP. It is interesting to note that the maxima of all three potential energy curves correspond to charges in equilibrium. It is for this reason that there is no stable equilibrium.

Figure A.  
Figure B.  
Box 1 continued . . .
Earnshaw’s theorem can be visualised in two dimensions (Figure B) by an arrangement consisting of a system of four hyperbolic electrodes, with adjacent electrodes oppositely charged as shown in Figure 1. The minimum distance between the opposite electrodes is $2\tau_0$. The solid lines in Figure B represent equipotential surfaces and the electric field is zero at the centre. A test charge placed at the centre would be at equilibrium, but not a stable one. While the zero potential of this point is a minimum with respect to the quadrants 1 and 3 it is a maximum with respect to quadrants 2 and 4. Such a point is called a saddle point. Thus charges cannot be in stable equilibrium in electrostatic fields alone.

This necessitates a more involved arrangement if one wishes to trap the ions in three dimensions. Let us consider the electric quadrupole potential of the form

$$\Phi(x, y, z) = A(x^2 + \beta y^2 + \gamma z^2). \quad (2)$$

The Laplace condition $\nabla^2 \Phi = 0$ imposes that the constants $\alpha$, $\beta$ and $\gamma$ satisfy the condition $\alpha + \beta + \gamma = 0$. A simple way to satisfy this condition is by setting $\alpha = -\beta = 1, \gamma = 0$, which results in the two dimensional field

$$\Phi(x, y) = A\alpha(x^2 - y^2). \quad (3)$$

This potential can be generated by a set of four hyperbolically shaped electrodes separated by a distance $2\tau_0$ and linearly extended in $z$-direction, which is essentially the principle of the two-dimensional mass filter shown in Figure 1. The other possibility is by setting $\alpha = \beta, \gamma = -2\alpha$, which generates the three dimensional field

$$\Phi(x, y, z) = A\alpha(x^2 + y^2 - 2z^2), \quad (4)$$

the details of which will be discussed in the subsequent sections.

Figure 1. (left) Electrode structure (of the two-dimensional mass filter) required to generate the potential given in (3).

Figure 2. (right) Electrode structure required to produce the three-dimensional rotationally symmetric quadrupole fields used in the ion trap.
3. Three Dimensional Confinement

The three-dimensional quadrupole trap field given by (4) can be generated by a three-electrode structure as shown in Figure 2. There are two end-cap electrodes separated by a distance $2z_0$ and a ring electrode of radius $r_0$, ($r_0^2 = 2z_0^2$) whose surfaces are hyperboloids of revolution about the $z$-axis.

For trapping the positive ions the two end-caps are held at a static positive potential with the ring electrode held at negative potential. Applying a potential $U_0$ between the end caps and the ring electrodes, (4) becomes

$$\Phi(x, y, z) = \frac{U_0}{2r_0^2} (2z^2 - x^2 - y^2),$$

(5)

where the field strength of this potential is given by

$$E_x = \left( \frac{U_0}{r_0^2} \right) x; \quad E_y = \left( \frac{U_0}{r_0^2} \right) y; \quad E_z = -\left( \frac{2U_0}{r_0^2} \right) z,$$

and for an ion of mass ‘$m$’, and charge ‘$q$’, we have

$$\frac{d^2x}{dt^2} = \left( \frac{2qU_0}{mr_0^2} \right) x = 0$$

(6a)

$$\frac{d^2y}{dt^2} = \left( \frac{2qU_0}{mr_0^2} \right) y = 0$$

(6b)

$$\frac{d^2z}{dt^2} + \left( \frac{qU_0}{mr_0^2} \right) z = 0$$

(6c)

Equation (6c) which represents a simple harmonic motion shows that the ion is trapped in a harmonic potential along the $z$-axis. However, the potential in the $xy$-plane defocuses the ions, which is of course consistent with Earnshaw’s theorem as mentioned above. At the trap centre, the potential forms a saddle and the charged particles will be confined either in the radial plane or in the axial direction, but will escape in the other direction (Figure 3). In the Penning trap this is overcome by the addition of a magnetic field (Figure 4a), while in the Paul trap a RF potential is applied to the electrodes (Figure 4b).

4. The Penning Trap

It is now clear that the dc field provides confinement of the ions along the axial direction while there is a repulsive force in the radial $xy$-plane. Confinement of the ions in the Penning trap is achieved by adding a constant magnetic field $B$ along the $z$-axis to the electrostatic field (Figure 4a). The three original translational degrees of
Figure 3. Equipotential surface (5) in an ion trap as a function of the axial coordinate $z$ and the radial distance $r$. A positive ion starting from rest on the $z$-axis would oscillate along it, but any radial displacement would lead to instability. This can be prevented by alternating the potential in an RF (Paul) trap or adding a magnetic field $B$ directed along the $z$-axis in the dc (Penning) trap.

freedom of the ion now become three modes of motion in the trap. The ion moving along the $z$-axis sees a parabolic potential,

$$\Phi(x, y, z) = \frac{U_0}{2r_0^2} (2z^2 - x^2 - y^2) = \frac{U_0}{2r_0^2} (2z^2 - r^2)$$

(7)

and will undergo harmonic oscillations (6c) with a frequency $\omega_z$ (axial frequency),

$$\omega_z^2 = \frac{2qU_0}{mr_0^2}$$

(8)

As the ion starts to escape along the radial plane, the magnetic field turns back the ion into a circular orbit (see Box 2). Thus in the radial plane the ion undergoes a circular motion with the centre at the origin. For this orbit, due to the radial component $E_r$,

Figure 4. The cross-section of ion trap used in the (a) Penning and (b) Paul configurations.
Box 2. Cyclotron and Magnetron Motion

To understand the dynamics of the motion of the ion in the radial plane, let us first assume the presence of only the magnetic field $B$ along the $z$-axis and the ion velocity in the radial plane. The ion in this case will undergo a cyclotron motion of frequency $\omega_c = qB/m$.

Next consider a constant electric field $E$ in the $x$-direction. The ion will follow a cycloid motion in the radial plane, which can be described qualitatively as follows: Suppose initially the ion is at rest near the origin; the magnetic force is zero and the electric field accelerates it away from the centre in the $x$-direction. As the speed of the ion increases, the magnetic force increases which pulls the ion back towards the direction of $y$-axis as shown in Figure A. The faster it moves, the stronger the force due to the magnetic field which eventually turns the ion back to the $y$-axis, and as the ion is moving against the electric field, its speed decreases and eventually comes to zero on the $y$-axis.

The entire process is periodically repeated and the ion slowly drifts away from the centre in the radial plane. This can be quantitatively worked out using the Lorentz force $\vec{F} = q(\vec{E} + \vec{v} \times \vec{B})$.

In the ion trap the electric field in the $x$-$y$ plane is radial. The slow drift motion, described in the above paragraph, is modified into a circular motion. This slow circular drift in the radial plane is called the magnetron motion.

The electric force

$$qE_r = -q \frac{\partial \Phi}{\partial r} = \frac{qU_0 r}{r_0^2}$$

almost balances the magnetic force $qvB$, that is

$$\frac{qU_0 r}{r_0^2} = qvB.$$  

Hence, the frequency of this circular motion $\omega_m$, which is a constant of the trap, is given by

$$\omega_m = \frac{v}{r} = \frac{U_0}{r_0^2 B}.$$  

the cyclotron frequency $\omega_c$ given by,

$$\omega_c = \frac{qB}{m}.$$  

\[10\]
Figure 5. Ion trajectory (left) in an ion trap, with all three fundamental modes (cyclotron $\omega_c = 2\pi v_c$, magnetron $\omega_m = 2\pi v_m$ and axial $\omega_z = 2\pi v_z$) excited, and the projections (right) of that trajectory onto the $xy$ plane.

A more accurate analysis, taking into account the coupling of electric and magnetic fields, shows that the frequencies of cyclotron ($\omega_c$) and magnetron motion ($\omega_m$) (Figure 5) are slightly modified

\[
\omega'_c = \frac{1}{2} \omega_c + \left( \frac{\omega_c^2}{4} + \frac{\omega_z^2}{2} \right)^{1/2} \tag{11}
\]

\[
\omega_m = \frac{1}{2} \omega_c - \left( \frac{\omega_c^2}{4} + \frac{\omega_z^2}{2} \right)^{1/2} \tag{12}
\]

Here $\omega'_c$ is the modified cyclotron frequency.

Confinement is assured if the action of the magnetic field exceeds the defocusing force of the electric field. In order to obtain stable orbits, we must have real values for $\omega_c$ and $\omega_m$, that is,

\[
\frac{\omega_c^2}{2} > \omega_z^2 \tag{13}
\]

is the condition for stability in the Penning trap.

5. The Paul Trap

In a Paul trap, an oscillating electric potential is applied between the ring and the two end-cap electrodes (Figure 4b) in conjunction with the static electric potential $U_0$. The trap potential then has the form

\[
\Phi(x, y, z) = \left( \frac{U_0 - V_0 \cos(\Omega t)}{2r_0^2} \right) (x^2 + y^2 - 2z^2), \tag{14}
\]

where $V_0$ and $\Omega$ are the amplitude and frequency, respectively of the oscillating electric potential.
The trap is stable in the axial direction and unstable in the radial plane for half the cycle and vice versa for the next half of the cycle. However, owing to the field inhomogeneity, the force averaged over a period of the oscillating field does not average to zero but is directed towards the regions of weak field, that is, towards the trap centre. At the trap centre, there is no field variation; therefore there is no motion due to the RF field and ideally a particle set at the trap centre would remain at rest.

The equation of motion for a particle of mass $m$ and charge $q$ is

$$\frac{d^2}{dt^2}\begin{pmatrix} x \\ y \\ z \end{pmatrix} + \frac{e}{mr_0^2} [U_0 - V_0 \cos(\Omega t)] \begin{pmatrix} x \\ y \\ -2z \end{pmatrix} = 0.$$  \hspace{1cm} (15)

If the following substitutions are made,

$$a = \frac{8eU}{mr_0^2\Omega^2}; \quad q = \frac{4eV_0}{mr_0^2\Omega^2}; \quad \tau = \frac{1}{2} \Omega t,$$

then we obtain the following set of Mathieu equations

$$\frac{d^2}{dt^2}\begin{pmatrix} x \\ y \\ z \end{pmatrix} + [a - 2q \cos(2\Omega t)] \begin{pmatrix} x \\ y \\ -2z \end{pmatrix} = 0.$$  \hspace{1cm} (17)

It has been shown that (17) possesses bounded (stable) and unbounded (unstable) solutions depending on the values of ‘$a$’ and ‘$q$’. Figure 6 shows a section of the

![Figure 6](image.png)

**Figure 6.** A section of the Mathieu stability diagram for the Paul trap. The motion is simultaneously stable both in the axial and radial directions only within the shaded regions.
Extreme reduction of Doppler broadening can be achieved, due to the possibility of very effective cooling of the trapped ions.
As mentioned above ion traps find numerous applications in the field of high-resolution optical double resonance spectroscopy, single-ion spectroscopy, precision mass measurements of stable and radioactive isotopes, atomic frequency standards, and some of these are briefly discussed below.

(a) Optical-RF Double Resonance Spectroscopy: The conventional techniques of high resolution spectroscopy are limited by large Doppler widths (see Box 3) of spectral lines, typically of the order of 0.05 cm$^{-1}$ $\approx$ 1500 MHz. These optical techniques are incapable of achieving the required resolution and precision for measuring small splitting/shifts in the energy levels. This is mainly due to the fact that a small energy separation is being measured indirectly by taking the difference between two very large optical frequencies.

In double resonance spectroscopy two electromagnetic fields of different frequencies are simultaneously in resonance with two atomic or molecular transitions sharing a common energy level. The first electromagnetic field takes the atoms/molecules to an excited state and the second field, which could be in the RF or microwave region, induces transitions between energy levels which are very closely spaced. As the transition is taking place in the RF region, the Doppler line width is reduced. Furthermore in a trap the ion motion is restricted to amplitudes which are small compared to the wavelength of the microwave radiation. Thus there is no first-order Doppler effect and in addition there is elimination of the time-of-flight broadening, allowing high-resolution optical double resonance spectroscopy to be performed. Using this technique very high resolution and accuracy has been achieved, e.g. a line width of 2 mHz (millihertz) has been obtained.

(b) Single Ion Spectroscopy: Ion traps make it possible to study a single cold ion, and the use of laser cooling techniques further removes all orders of Doppler effects on the atomic transitions. Thus a series of experiments in basic physics have been performed, such as observation of quantum jumps, experimental test of quantum Zeno effect, a precision test of the linearity of quantum mechanics and so on. E Peik and others (1994) have performed laser cooling and quantum jumps on a single Indium ion. For a single ion they observed quantum jumps and out of the $^3P_0$ level. Quantum Zeno effect was experimentally tested by Itano and others (1990) with atomic level measurements realised by means of a short laser pulse.

(c) Precision Mass Measurement: Very precise values of $g$-factors and mass ratios have been obtained by measuring ion oscillation frequency in a Penning trap. As shown in (7-11), the motional frequencies in Penning traps are related to cyclotron frequency $\omega_c = eB/m$ and the basic idea for $g$-factor and mass measurements is a precise determination of the cyclotron frequency of the trapped particles.

The apparatus set up at the ISOLDE facility at European Centre for Nuclear Research, Geneva for the direct measurement of masses of radioactive isotopes uses ions delivered by the online separator and involves two Penning traps arranged in tandem. Kluge and others (1993) have measured the masses of unstable isotopes.
Box 3. Line Broadening Mechanism

Spectral lines in discrete absorption or emission spectra are never strictly monochromatic. Even with very high resolution one observes a spectral distribution $I(\omega)$ of the observed or emitted intensity around the central frequency $\omega_0 = (E_f - E_i)/\hbar$ corresponding to an atomic transition between upper level $E_f$ and lower level $E_i$. Some of the broadening mechanisms of spectral lines are discussed below.

**Natural Linewidth:** If $\tau$ is the mean lifetime of an energy level $E$ then according to the uncertainty principle its energy $E$ can be determined only with an uncertainty $\Delta E = \hbar/\tau$. This results in the total uncertainty

$$\Delta E = \Delta E_i + \Delta E_f = \hbar \left( \frac{1}{\tau_i} + \frac{1}{\tau_f} \right),$$

which leads to the natural line broadening of a spectral transition

$$\Delta \omega = \frac{\Delta E}{\hbar} = \left( \frac{1}{\tau_i} + \frac{1}{\tau_f} \right).$$

**Doppler Linewidth:** The Doppler width arises due to the thermal motion of atoms absorbing or emitting radiation. Suppose that an atom in the excited energy levels $E_f$ moving with a velocity $\vec{v}_f$ emits a photon of energy $\hbar \omega$ and momentum $\hbar \vec{k}$, resulting in the transition $E_f \rightarrow E_i$ where $E_i$ is the lower energy level. The emission of radiation causes the atom to recoil to a new velocity $\vec{v}_i$. Conservation of momentum demands, that

$$M \vec{v}_f = M \vec{v}_i + \hbar \vec{k}, \quad \text{(a)}$$

where $M$ is the mass of the atom. Energy conservation under non-relativistic conditions is

$$E_f + \frac{1}{2} M \vec{v}_f^2 = E_i + \frac{1}{2} M \vec{v}_i^2 + \hbar \omega. \quad \text{(b)}$$

If $\omega_0$ is the frequency of light, which would be emitted if the atom had zero velocity before and after the emission, then

$$\hbar \omega_0 = E_f - E_i. \quad \text{(c)}$$

Elimination of $\vec{v}_f, E_i, E_f$ from (b) with use of (a) and (c) gives,

$$\omega = \omega_0 + \frac{\hbar k^2}{2M} = \omega_0 + \frac{\omega v_i}{c} + \frac{\hbar \omega^2}{2M c^2}. \quad \text{(d)}$$

The second term is the linear (first-order) Doppler effect describing the well-known $\Delta \omega = \vec{k} \cdot \vec{v}$ Doppler shift. Thus the width can be reduced by making observations perpendicular to the emitting atoms, as in experiments with collimated atomic beams. The third term in (d) is the photon recoil effect. Typical orders of magnitude for some quantities on the right of (d) are

$$\frac{v}{c} \approx 10^{-5}, \quad \frac{\hbar \omega}{2Mc^2} \approx 10^{-9}.$$
Although the first order Doppler width can be completely eliminated by using various techniques, the second order Doppler effect, which is purely relativistic, persists.

Consider an atom (modelled as an oscillator) travelling along the $z$-axis. We observe the emission say along the $x$-direction. If the period of the stationary oscillator (atom) is $\tau_0$ then the period in the laboratory frame is $\tau = \tau_0 / \left(1 - \frac{v^2}{c^2}\right)^{1/2}$, due to time dilation. Hence the frequency of emission $\omega = 2\pi / \tau$ will be,

$$\omega = \omega_0 \left(1 - \frac{v^2}{c^2}\right)^{1/2} = \omega_0 - \frac{\omega_0 v^2}{2c^2}.$$  

The second term here is the second order Doppler effect. Note that this term is independent of the direction of $\vec{v}$ and can be eliminated only by making $\vec{v} = 0$.

**Transit-Time Broadening:** If the interaction time of atoms with the radiation field is small compared with the spontaneous lifetime of the excited level, then it leads to transit-time or time-of-flight broadening. Consider an atom interacting with monochromatic light of frequency $\omega$ during the interval $T$. The atom experiences the following field

$$E(t) = \begin{cases} A_0 \cos(\omega_0 t) & 0 \leq t \leq T \\ 0 & t > T \end{cases}$$

which in the frequency domain is obtained from the Fourier transform

$$E(\omega) \equiv \frac{1}{\sqrt{2\pi}} \int_0^T A_0 \cos(\omega_0 t) e^{-i\omega t} dt.$$  

The intensity profile of the spectral line, $I(\omega) = E^*(\omega) E(\omega)$ is, for $(\omega - \omega_0) \ll \omega$,

$$I(\omega) = C \frac{\sin^2\left[\frac{(\omega - \omega_0)T}{2}\right]}{(\omega - \omega_0)^2},$$  

where $C$ is a constant. This is a function with a full half-width $\delta\omega_T = 5.6/T$ (see Figure A). The intensity profile of the spectral line is shown for, (i) rectangular and (ii) Gaussian laser pulses. For a beam of fast ions with typical velocity $3 \times 10^8$ cm/sec, the time required to traverse a laser beam of width 0.1 cm is less than $10^{-9}$ sec which is shorter than the spontaneous lifetimes of most atomic levels. With $T \approx 10^{-9}$ sec, the full-width at half maximum of the transit time broadening $\omega_T \approx 10^9$ Hz.

![Figure A.](image-url)
\[(^{142}\text{Cs} \ t_{1/2} \sim 1.8 \text{ sec})\] with a measurement accuracy of \(\delta m/m \approx 10^{-7}\) and resolving power \(m/\Delta m \approx 106\).

\textbf{(d) Frequency Standards:} The most important possible application of cooled-trapped ions is the development of trapped ion frequency standards for both optical and microwaves regions. Ion traps offer the potential for achieving cooled ion species with very narrow resonances (with line widths a fraction of a Hertz, both in optical and microwave regions) with high accuracy and stability. A variant of the quadrupole trap capable of localising a trapped ion within much less than an optical wavelength has been realised by S R Jefferts, C Monroe and D J Wineland (1995).

\textbf{Suggested Reading}


The Challenge of Fluid Flow
What One Can and Cannot Do

Roddam Narasimha

1. The Navier–Stokes Equations

In Part 1\(^1\) we saw how diverse flow phenomena can be, even when we are looking at such common fluids as air and water (more complex fluids – from paint to slurry – are another big story by themselves). We saw how in many of those flows both order and disorder are simultaneously present. We have no theories that can handle these mixed order-disorder phenomena well.

But we do believe that the basic laws governing fluid motion are known. They are consequences of Newton’s laws of motion and his concept of viscosity. The laws are best written down in the formalism of a continuous field, along the lines that Leonhard Euler introduced in 1755 – exactly 250 years ago. For an incompressible viscous fluid of density \(\rho\) and viscosity \(\mu\), the flow is governed by the equations of conservation of mass and momentum,

\[
\text{div } \mathbf{u} = 0
\]

and

\[
\rho \frac{\text{d}\mathbf{u}}{\text{d}t} \equiv \rho \left( \frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \text{grad}) \mathbf{u} \right) = -\text{grad } p + \mu \nabla^2 \mathbf{u} + \mathbf{F},
\]

where \(\mathbf{u} = \mathbf{u}(\mathbf{x}, t)\) is the (vector) velocity field, dependent on the position vector \(\mathbf{x}\) and time \(t\), \(p = p(\mathbf{x}, t)\) is the pressure field, and \(\mathbf{F} = \mathbf{F}(\mathbf{x}, t)\) is a body force (per unit volume of fluid).

Note the presence in the equations of the quadratically nonlinear term \(\mathbf{u} \cdot \text{grad } \mathbf{u}\). This nonlinear term is the crux of the problems of fluid dynamics; if it had not been there the subject would legitimately have become ‘classical’ by now. It is only in the 20th century that fluid dynamicists began slowly learning how to handle the non-linearity, but the learning process is far from over.

2. Prandtl’s Idea

The first shot here was fired by Ludwig Prandtl almost exactly a hundred years ago, so I should say a few words about it in this centenary year. In 1904 this remarkable engineer read a paper at the 3rd International Congress of Mathematics held at Heidelberg, and showed how, in the limit of large Reynolds numbers, the equations could be made


Keywords
Fluid dynamics, turbulence, chaos, nonlinearity, mathematics of flow.
Box 1. The Navier–Stokes Equations

We have written down these equations for an incompressible viscous fluid. They assume that the fluid is continuous (no voids, for example), and that the state of internal stress in the fluid depends linearly on the local rate of strain that the fluid experiences during its motion. The principles of conservation of mass and momentum are then sufficient to obtain the partial differential equations that govern the motion: one expressing continuity or conservation of mass, and three scalar equations (constituting one vector equation) each of which expresses conservation of momentum in an independent direction.

Because fluid velocity and pressure depend in general on both space and time, the governing equations are partial differential equations. The term that is responsible for the difficulty of solving these equations is what may be called the advective acceleration, written as \((u \cdot \text{grad}) u\) in the text. This term represents the acceleration experienced by a fluid particle, originally at a given position and time, as it moves a short interval of time later to a neighbouring position where the fluid velocity is slightly different. Because the increment in velocity and the distance travelled by the particle both depend on the velocity field, this advective acceleration is nonlinear in the fluid velocity. This nonlinearity has profound consequences, and results in all the complexities that we see in the motion of even such common fluids as air and water.

The nonlinear term in the Navier–Stokes equations is the crux of the problems of fluid dynamics; if it had not been there the subject would legitimately have become ‘classical’ by now.

By example Prandtl brought to an end the war between ancient hydraulics and 19th century hydrodynamics, which till then had scorned each other.

manageable without sacrificing nonlinearity. His approach can be summarized in the recipe, ‘Divide, conquer and unify’. The first big step was to realize that the flow could be divided into different **regions**: let us call them, in the simplest cases, inner and outer (assuming there are only two such regions). The corresponding inner and outer **equations**, each the limiting form of the governing equation in its respective region, are separately conquered, i.e. solved, yielding inner and outer **solutions**. By cleverly specifying their boundary conditions these separate solutions are then ‘unified’ into a composite solution, which was later shown to be asymptotic to the exact solution everywhere in the large Reynolds number limit.

Prandtl posed the problem for the flow past an aligned flat plate, which was trivial in 19th century Eulerian (**inviscid**) hydrodynamics \((\mu = 0)\), but was fundamental to the new fluid dynamics he was creating. By an ingenious mixture of extensive visual observation (in a $40 water channel), mathematical approximation and numerical calculation, he produced an answer for the drag or resistance of the plate that in principle solved the classical d’Alembert paradox (of no resistance in a non-viscous fluid). In the process of solving this special problem, however, Prandtl did many other things. He showed how, at high Reynolds number Re, fluid flows tend to fold or squash into layers; he invented what later became the more formal method of matched asymptotic expansions for handling singular perturbation problems; by example he brought to an end the war between ancient hydraulics and 19th century hydrodynamics, which till then had scorned each other (hydraulics was dismissed as a science of variable constants, hydrodynamics as the mathematics of dry wa-
ter); and, in the process, he founded modern fluid dynamics, giving it the tools by which many of the earlier ‘paradoxes’ that had plagued the subject could be resolved one by one.

3. Some Things We Know

Thanks to that modern fluid dynamics we now know a lot about many fluid flows – enough to build aircraft traveling at Mach 3 or carrying nearly a thousand passengers (the A380 can take up to 873 people including crew), to make rockets that can shoot us to the moon and bring us back, to predict weather a few days in advance, to manage flows so as to enhance or diminish heat transfer between a solid surface and a fluid or to promote or suppress mixing between two fluids. But we still do not understand the central problem of turbulent flows, which remains fundamentally unsolved. (As the word appears repeatedly in this article, I should explain what I mean by understanding a phenomenon. In the first place, it implies that there is a quantitative explanation that is derived from accepted first principles – e.g. the Navier-Stokes equations – if necessary by exploiting reasonable approximations using well-tested methods, but not appealing to additional experimental data. Related to this is a second kind of explanation, consistent with the first but possibly qualitative, where the phenomenon in question may be shown to emerge from others which are already ‘understood’.) If the physics underlying the phenomenon is utterly new, it is natural that understanding in the above sense will emerge only slowly. To illustrate the bizarre situation in fluid dynamics, consider the common plumbing problem of estimating the pressure loss suffered by water flowing through a pipe. Man has been pushing water through pipes and channels for thousands of years; our ancestors did that already very well in the Indus valley civilization of some 4000 years ago. Thousands of engineers make confident and successful designs using data codified into diagrams of the type shown in Figure 1. But it is only the green line in the diagram, representing laminar flow, that is ‘understood’. The rest is known, by a mix of testing and ingenious heuristic argument about turbulent flow, but not really understood. To highlight this extraordinary situation, let me note that there is a new analysis of turbulent pipe flow that claims that some of the results of the kind shown in Figure 1 can be wrong by as much as 60% at extremely high Reynolds numbers; this new analysis is not yet either confirmed or refuted.

4. Deterministic Chaos

This has been the enduring mystery of fluid flows: its governing laws are known, nevertheless even everyday phenomena, seen by our eyes all the time, cannot always be explained solely from those laws. Richard Feynman called turbulent flow “the greatest puzzle of classical physics”. But the adjective ‘classical’ there should be carefully interpreted, for ‘classical’ is often equated with ‘understood’, i.e. intellectually dead.
This of course is far from being the case with regard to turbulent flows. I think it would be more accurate to say that Newtonian mechanics has turned out to be full of deep, unresolved, sometimes even unsuspected mysteries (in spite of having been dubbed as ‘classical’ with the advent of relativity and quantum mechanics). One such phenomenon, unsuspected till some forty years ago, is deterministic chaos. This forces together two concepts – of necessity and chance, of law and accident, or (in Upanishadic terms) of niyati and yadrecha – concepts that had earlier been thought of as two competing, mutually incompatible views of the nature of the universe. The discovery that paradigmatically deterministic Newtonian systems can behave in ways that appear random has had such a profound effect on our thinking that Sir James Lighthill, occupant of the same prestigious Lucasian Chair that Newton had held some 300 years earlier in Cambridge, felt compelled to say in 1986:

We [i.e. the community of scientists pursuing ‘classical’ mechanics] collectively wish to apologize for having misled the general educated public by spreading ideas about the determinism of systems satisfying Newton’s laws of motion that, after 1960, were to be proved incorrect.

And chaos has now been detected in such exemplars of the alleged ‘clockwork’ of the universe as the planetary system, the pendulum and the elastic string. Einstein famously said that he did not believe in a God who played dice, but would he have believed in a Newtonian God who played deterministic nonlinear games whose outcomes would be effectively indistinguishable from the results of playing dice?
Incidentally Newtonian chaos is not unrelated to the turbulence of fluid flow – chaos is generic turbulence, so to speak, that may be encountered in non-fluid dynamical systems as well. Indeed, a key advance in the emergence of the concept of chaos was the study of a highly idealized form of convective weather – going one or two nonlinear steps beyond the sree-border pattern of Figure 3, Part 1. That study, undertaken in the early 1960s by the American meteorologist E N Lorenz, showed how convection can become erratic, explained why weather is unpredictable beyond a certain time horizon, showed the relation of these properties to those of some simple nonlinear maps, and visually displayed for the first time the mathematical object that later came to be called a strange attractor. This theory of chaos has certainly solved the philosophical problem of how turbulence can emerge out of the Navier-Stokes equations. But it has unfortunately not otherwise been of great help in understanding or predicting turbulent flows for the simple reason that the number of degrees of freedom in a flow diverges as the Reynolds number increases.

5. The Fundamental Problem

I believe it was John von Neumann who came closest to seeing the true nature of the fundamental problem of fluid dynamics. He first of all realized that:

From the point of view of theoretical physics turbulence is the first clear-cut instance calling for a new form of statistical mechanics.

He then went on to say:

The impact of an adequate theory of turbulence on certain very important parts of pure mathematics may be even greater [than on fluid dynamics].

Von Neumann thought that there was some hope to “break the deadlock by extensive but well-planned computational efforts”; and this was one of the reasons that he got so deeply involved in the development of computer technology. But computing solutions of the Navier-Stokes equations is more like doing experiments (only they are numerical instead of physical), and will not automatically provide understanding. In any case, even on the most powerful computers available in the world today, we cannot reach Reynolds numbers higher than of order \(10^4\). To see that we still have a long way to go, look at Figure 2, which is a collection of data compiled by Anatol Roshko on the pressure at the back (more precisely rear stagnation point) of the same kind of cylinder that sheds Kármán vortices at lower Reynolds numbers. First of all note that the data (collected from many different sources) show surprisingly small scatter. Now how can the pressure on a body whose cross-section is a platonically perfect circle vary with Reynolds number in such a non-simple way? (Could that crazy variation be a set of signatures of the many transitions that keep occurring as the flow folds into complex layers?) Who would dare to guess (based on that data)
Two problems on the Navier–Stokes equations are among the seven million-dollar prize problems posed by the Clay Foundation (Box 2) – along with the Riemann hypothesis and the Poincaré conjecture. That I feel is the right company for turbulence, and shows why fluid dynamics continues to be such an enduring challenge.

6. What Can be Done

While waiting for mathematical paradise, there is a great deal that fluid dynamicists can do and have done about these problems. Engineers of course cannot always wait for understanding (as the great electrical engineer Oliver Heaviside pointed out, we do not stop eating just because we do not understand digestion). But engineers would love a good theory – it would cut development costs dramatically (I can hear some of you saying ugh!). In the absence of a theory, one engineering way is to treat each flow on its merits, carry out extensive tests (see Box 3) and make handbooks or catalogues with diagrams like Figure 1 (using physical or computational experiments) – or their more modern computerised equivalents. But I think a great deal more can be done.
Box 2. The Clay Institute’s Million-dollar Prize Problems

The Clay Mathematics Institute (Cambridge, MA) is a private foundation financed by the businessman Landon T Clay. The Institute has offered a million-dollar prize for solutions of each of 7 ‘Millennium Prize’ problems. These include such celebrated and long-standing open questions as the Riemann hypothesis and the Poincaré conjecture. One of the seven problems is concerned with ‘Navier–Stokes existence and smoothness’.

The basic question underlying the Navier–Stokes problem is to show whether smooth, physically reasonable solutions of the three-dimensional Navier–Stokes equations exist or not. (The two-dimensional problem was solved by the Soviet mathematician Ladyzhenskaya (1922–2004) in the 1960s.)

As solution of the problem, proof is demanded of one of four statements. The flavour of these statements is conveyed by the following question. If the initial velocity field is sufficiently smooth everywhere, and the forcing function \( F(x,t) \) is also similarly sufficiently smooth everywhere and at all times, does the Navier–Stokes solution for the velocity field remain smooth with finite energy, or can it blow up? (For a statement of the problem with more precise mathematical qualifications, see the Clay Mathematical Institute website: [http://en.wikipedia.org/wiki/Millennium_Prize_Problems](http://en.wikipedia.org/wiki/Millennium_Prize_Problems).) Even more simply stated, do smooth initial conditions always yield smooth solutions for all times or not? For the Euler equations (valid for an inviscid fluid, \( \nu = 0 \)), the evidence seems to suggest that blow-up is possible. Answers to such questions can help us to understand the way that viscous fluid flows are drawn out into thin sheets or filaments of vorticity, as in Figure c in Box 4.

7. Some Basic Ideas

First of all, I believe there are two keys to appreciating fluid flows – i.e. to getting an intuitive feel for their structure (which is still far short of predicting them from first principles). The two keys, I propose, are

**INSTABILITY, NONLINEARITY**

As we have already seen, fluid flows tend naturally to instability under most conditions barring the mildest (very low Reynolds or Rayleigh numbers, for example). The nature of the instability depends on the particular flow type (jet, boundary layer, convection, rotation etc.), and can vary widely from flow to flow. That is, instability is general, but its character is flow-specific. This tendency to crumple into instability, enabling small disturbances to grow, can be the first step leading to chaos and turbulence. It is no wonder therefore that instability has been seen as a central issue by some of the biggest names in physics and fluid dynamics: Sommerfeld, Rayleigh, Heisenberg, Chandrasekhar – and most relentlessly and successfully by Prandtl and his pupils and by G I Taylor. But the final stages of transition, ending up in breakdown, are essentially nonlinear, and not yet understood.

Fluid flows tend naturally to instability under most conditions barring the mildest.
Box 3. What We Can Measure

Here are some typical examples of what one can measure by recent techniques. The current trend in fluid flow measurements is towards development of tools that measure whole pressure, velocity and temperature fields, rather than making point measurements.

Figure a. (left). A recently developed method of measuring pressure distribution on models in wind tunnels is the use of pressure-sensitive luminescent coatings (‘paints’). This picture shows pressure measurements made on an aircraft model using a pressure-sensitive paint developed at NAL.

Figure b. (right). Whole-field velocity measurements using particle image velocimetry near the trailing edge of a flapped aircraft wing.

That leads to nonlinearity, which has several effects. The first, strangely, is to fix order. By this I mean that a linear instability mode which has started growing can have its amplitude limited by nonlinear saturation, producing the stable order of which we saw several illustrations earlier (e.g. convective rolls, Kármán vortices), without necessarily changing the mode much.

The second effect of nonlinearity is to fold the flow into layers, as first analysed by Prandtl in the simple case of flow past a flat plate. But as thin layers fold into even thinner ones (see e.g. Figure 3), there is a cascade in scales. And there can be layers within layers! Of course the folding and squashing can result not only in sheets, but scrolls, filaments, shocks and various other types of singularities.

Thirdly nonlinearity can generate chaos – as we have already described.

And, by a combination of all these mechanisms nonlinearity can produce strange mixtures of chaos and order over a wide range of scales – the vicitra-vibhava of the Vāsiṣṭha.
Box 4. What We can Compute

Here are some typical examples of the kind of computations that can be made today. Each of them is thought to be an exact numerical solution of the governing equations and initial-boundary conditions – that is, without the aid of any additional empirical or semi-empirical assumptions, hypotheses, models etc.

**Figure a.** Solution of Navier–Stokes equations for an aerofoil at high angle of attack (~ 29°). Instantaneous contours of vorticity. Reynolds number = 45,000 based on chord.

**Figure b.** Diametral cross-section of a turbulent jet, like that shown in *Figure 9, Part 1*. Solutions of the full Navier–Stokes equations, Re = 1600 based on nozzle velocity and diameter. Colour coding indicates values of the vorticity component along the direction of jet flow, i.e. perpendicular to the plane of the paper. Arrows show velocity vectors in the plane of the paper just outside the core of the jet. (Left) Conventional jet. (Right) Jet with volumetric heating, simulating the release of latent heat of condensation of water vapour in a cloud. Notice the dramatic changes in both velocity and vorticity fields, explaining in part why flight through clouds in an aircraft can be so bumpy.

**Figure c.** (left). A solution of the Navier–Stokes equations for decaying turbulence stirred up in a box of size 2048³. Reynolds number = 270 based on r.m.s. fluctuating velocity and the Taylor microscale. Picture shows contours of constant vorticity. The solution (with others like it) takes several weeks on one of the most powerful computers in the world, the Advanced Simulation and Computing Q machine at Los Alamos.

**Figure d.** (right). A solution of the Euler equations, colour-coded to indicate pressure distribution on the Light Combat Aircraft.
**Figure 3.** A vortex pair descends on a circular cylinder, and folds itself into layers. Smoke flow visualization at 40 ms intervals, basic cylinder Reynolds number 1500.

To these two pillars of intelligent thinking about fluid flows, we should perhaps add a third,

**SCALING AND MATCHING**

which is a way of reasoning about length and time *scales* in the flow. This is important for several reasons. First of all fluid flows have a wide range of scales (because of the layer-making property discussed above). Secondly scaling arguments help us to separate the more nearly universal features of a flow from those that are less so. Thirdly a successful scaling argument helps us to condense vast quantities of experimental data into manageable *patterns*. For example, if the scales characterizing wake flows are known, then all wakes become *instances* of a universal wake, i.e. all of them collapse into universal functions and numbers in appropriately scaled variables. (Of course these functions and numbers cannot still be *predicted*, and usually have to be found from physical or numerical experiments.)
When the flow velocity in turbulent flow fluctuates in an apparently random way (as seen for example in Figure 7, Part 1), one useful method of description is in terms of its (frequency) spectrum, or more precisely the power spectral density. In the case of the turbulent velocity field \((u')\) this is commonly done in terms of the wave-number, which is a ‘spatial’ frequency measured say in cycles per unit length, related to the more familiar temporal frequency (cycles per second) through a translational velocity. If \(k\) is the magnitude of the wave number, one can express the mean kinetic energy in the fluctuating motion per unit mass of fluid as \(\frac{1}{2} (u')^2\), and write it as the integral

\[
\frac{1}{2} (u')^2 = \int_0^\infty E(k) \, dk.
\]

The spectral density \(E(k)\) can be inferred by measurement of the mean square value of outputs from appropriate filters through which components of the velocity signal \(u'\) are made to pass. Today it is more convenient to evaluate the spectrum by digitizing measured \(u'\) components and doing a fast Fourier transform on them in a computer.

One of the most celebrated of such scaling arguments is due to Kolmogorov (although the fundamental core of the argument had been used earlier in turbulent channel flow by Clark Millikan – at least that is the way I see it). Kolmogorov analysed the spectrum of turbulence, which he considered as determined by an energy cascade from large scales in the motion to small scales (rather as the crushing of coal leaves us with pieces of various sizes – from a few big lumps to a lot of fine dust). He proposed that the ‘small eddies’ (high wave numbers) are universal and depend only on the viscosity and the energy dissipation, and the large eddies are flow-specific and inviscid. Crucially, he further postulated that there is a range of intermediate scales over which both scaling arguments are valid, i.e. they are matchable. This led to the prediction that, over that range of intermediate scales (often called the inertial subrange), the spectrum should be proportional to \(k^{-5/3}\) (where \(k\) is the wave number). This should be true in any flow – whether it is a jet stream in the atmosphere, tidal flow in the oceans, or the boundary layer on an aircraft wing. Figure 4 shows how successful the argument is.

I have vastly oversimplified the reasoning here, and must hasten to caution you that universality may not be as common as it is sometimes thought to be. Kolmogorov himself felt compelled to revise his argument nearly twenty years after he first put it forward. Nevertheless, the organizing power of a successful scaling argument is enormous, its chief attraction being that it is minimalist in the hypotheses it makes (unlike in the specialized industry that churns out basically \textit{ad hoc} turbulence models, for example). But then we cannot prove when (or even whether) such scaling arguments follow from the Navier-Stokes equations, and, even if the arguments are valid,
What engineers often do today is use any tool or method that will help: theory of course, whenever it is available, testing, computing, simulation, scaling arguments and, increasingly, mathematical modelling.

the corresponding universal numbers and functions have of course to be determined from experiment of one kind or other.

Such scaling arguments have been used very widely – some times with success, at other times in controversy. They can be seen as simple applications of a kind of group theory, the centre-piece of the argument being a postulate on what the relevant group is.

There is, of course, THEORY – a lot of it in fact, although the central problem of the turbulent solutions of the Navier–Stokes equations in the limit of high Reynolds numbers remains unsolved. So the theories we possess have mostly to do either with linear problems such as inviscid flows without vorticity, small disturbance flows or low amplitude waves, highly viscous flows (i.e. at low Re) etc., or with limiting flow situations where the nonlinearity can be simplified because it is localized or approximated in some way.

What engineers often do today is use any tool or method that will help: theory of course, whenever it is available, testing, computing, simulation, scaling arguments and, increasingly, mathematical modelling. This modelling has to be distinguished from making direct appeal solely to first principles (e.g. Navier–Stokes). Instead, essentially new equations are devised, taking inspiration from but abandoning equivalence to the Navier–Stokes equations. Ingenuity lies in inventing equations that give the engineer the reduced information he wants, like the pipe data of Figure 1. Such models can be at many different levels – from codification into charts or tables like

Figure 4. The high-wave number end of turbulent spectra, scaled according to Kolmogorov. The agreement between atmospheric and oceanographic measurements, and the presence of an extended $k^{5/3}$ region, demonstrate the effectiveness of scaling and matching arguments.
Figure 1, through ordinary differential equations for each flow type (boundary layers, jets etc.), all the way to systems of nonlinear partial differential equations (for mean quantities) expected to be useful for wide classes of flows – inspired by but not deducible from the Navier-Stokes equations. Although none of them is spectacularly successful, many of these models are useful – sufficiently so that there is a minor industry across the world generating, testing and applying such models to a vast variety of fluid flow problems – from making better aircraft wings to estimating how pollutants disperse to forecasting weather or ocean state.

So I hope you may see in some small way why some of us love exploring fluid flows: they can be beautiful, crazy, fearsome, important; a challenge whether you want to stare at nature’s free displays, or visualize flows in the laboratory, or measure them with great precision, or control them so that they do your bidding; a happy hunting ground if you want to match your skills – of any kind – to try and unravel their secrets and to add to the great deal that is known; but, in essence, still beyond any mathematics or computers invented by man; still out there, so to speak, taunting us to see if we can understand, as we claim to know the basic laws.

If you also feel taunted by those flows, welcome to the fluid dynamicists’ club!

Acknowledgements

The contents of this article are taken from the Foundation Day Lecture given at the Inter-University Centre for Astronomy and Astrophysics, Pune, on 29 December 2004. I thank Dr Naresh Dadhich, Director of IUCAA, for gently insisting that I speak on why fluid flows are interesting rather than on the solution of some specific fluid-dynamical problems, and for permitting the publication of this article. I also thank Prof N Mukunda for the invitation to publish the text of the lecture in Resonance, and both him and Prof J H Arakeri for their comments and suggestions on an earlier version.

Ms K Nagarathna has as always been of immense help in putting the material together from the bits and pieces handed over to her at various random times.

Suggested Reading


If you are also taunted by the beautiful, crazy fearsome, important flows, welcome to the fluid dynamicists’ club!
Picture credits


Box 3. Figure a: Channaraju and P R Viswanath: PSP measurements on an aircraft model in NAL 1.2m blowdown wind tunnel, *NAL PD EA* 0313, 2003. Courtesy: P R Viswanath.


Box 4. Figure d: ADA/HAL Brochure.
In this article we first consider the importance of prediction of the monsoon, and events such as the intense rainfall event over Mumbai in July 2005. We then discuss how meteorologists make short-, medium-, and long-range forecasts and the concept of the limit of predictability in a chaotic system such as the atmosphere. Problems and prospects of prediction on different time-scales are discussed by using one example of short-range forecasts and the prediction of the monsoon by dynamical and statistical methods. Finally we consider measures of the skill of a forecast and how high the skill has to be for it to be useful for applications.

1. Introduction

From the beginning of this year there has been a steady increase in the inflation rate due to the dramatic rise in the price of rice, other food grains and oil. By July the inflation rate had risen above 11%, its highest in 13 years. As efforts were made to arrest it, the question arose as to when it would actually start decreasing. The response of one of the experts (Deputy Chairman of the Planning Commission) was very interesting. He suggested that the situation would improve if the monsoon turned out to be normal, as predicted\(^1\). Thus the impact of the vagaries of the monsoon on critical facets of our economy is perceived to be very significant. Not surprisingly, then, we never take the monsoon for granted. Every year, as the heat scorches the countryside in May, we start worrying about the upcoming monsoon and the media becomes obsessed with predictions about the monsoon rainfall.

Yet the Indian summer monsoon rainfall is one of the most reliable events in the tropical calendar. The typical year to year variation (i.e., the standard deviation) of the all-India Summer Monsoon Rainfall (ISMR) is only about 10% of the average rainfall of about 85 cm. The frequency distribution of ISMR over the period 1876–2007 is shown in Figure 1. Droughts are characterized by the ISMR anomaly (difference between the ISMR of a specific year and the long term average) being negative and of magnitude larger than 10% of the long term average. On the other hand, excess

\(^{1}\) In fact, the monsoon did turn out to be normal and the inflation rate did decrease in October. However, it is likely that the sharp decrease in oil prices associated with the economic meltdown was a far more important factor than the monsoon.
If the monsoon turns out to be a ‘normal’ monsoon, as in this year, the nation heaves a sigh of relief and carries on with business as usual.

rainfall years are characterized by positive ISMR anomalies of magnitude larger than 10% of the average. Normal monsoon years are characterized by the magnitude of the ISMR anomaly being less than 10% of the average. The distribution is not symmetric and is characterized by a longer tail with negative anomalies than that with positive anomalies. Over the 132-year period there have been 23 droughts and 19 excess rainfall years. Thus historical records show that the chance of the so-called normal monsoon is a little over 68%, of droughts around 17%, and of excess rainfall about 14%. While for the worst drought (1877) the ISMR deficit was 25 cm, for the season with maximum rainfall (1961) the ISMR anomaly was 17 cm. It is seen that the most likely value of ISMR (the mode) is around 90 cm, i.e., higher than the average and the rainfall is in the range of 83.75–91.25 cm in 44% of the years.

Why, then, are we so anxious about the monsoon? It turns out that although the amplitude of the variation of ISMR from year-to-year is not large, it has a substantial impact on the agricultural production in the country [1]. Before independence, this also implied a large impact on the economy of the country since the economy was primarily dependent on agriculture. With planned development since independence, the contribution of agriculture to the Gross Domestic Product (GDP) decreased substantially and led to the expectation that the impact of the monsoon on the economy would have also decreased. However, a recent analysis of the variation of the GDP and the monsoon has revealed that the impact of severe droughts on GDP has remained between 2 to 5% of GDP throughout [2]. The large impact of droughts on GDP (despite the substantial decrease in the contribution of agriculture to GDP) can be attributed to the indirect impact on the purchasing power of the large fraction of the population dependent on agriculture. Hence if the monsoon turns out to be a ‘normal’ monsoon, as in this year, the nation heaves a sigh of relief and carries on with business as usual. If it turns out to be a drought, there is a significant impact on agriculture and the economy, and major drought relief programmes are launched.
It has been shown [2] that while the magnitude of the adverse impact on food-grain production (IFGP) and the GDP (IGDP) of deficit rainfall is large, the positive impact of surplus rainfall is not large (Figure 2). In other words, there is an asymmetry in the response of food-grain production and GDP to the variation of the monsoon. It has been suggested that a possible reason for the relatively large asymmetry in the response of the foodgrain production after 1980, is that the strategies that would allow farmers to reap benefits of the good rainfall years (such as adequate investments in fertilizers and pesticides for rain-fed areas) are not economically viable in the current milieu. Such strategies would become economically viable if reliable predictions...
Meteorological forecasts are generated for different time-scales.

for ‘no droughts’ could be generated. Thus prediction of the interannual variation of ISMR and particularly for the occurrence or non-occurrence of the extremes continues to be extremely important.

In addition to prediction of the monsoon rainfall over the country as a whole, there is demand for prediction of some events such as the intense rainfall event on 26 July 2005 when Mumbai received 94.4 cm of rainfall on a single day, or of the severe cyclone that devastated Orissa in 1999, because of the enormous impact they have on a large number of people. There is also a need for several user-specific predictions such as prediction of low-level wind for sailors and for paragliding enthusiasts, quantitative precipitation forecasts for reservoir and flood management. The time-scales of the events for which prediction is required also varies with the application. Thus while some farmers need prediction for occurrence of a dry spell of duration of a week or more, for managers of reservoirs, prediction of the total rainfall in a month or a season is often adequate.

Meteorological forecasts are generated for different time-scales. Forecasts of daily weather with a lead time of 1–3 days are short-range forecasts and with a lead time of 3–10 days are called medium-range forecasts. Forecasts for monthly or seasonal rainfall come under the category of long-range forecasts. The official government agency which has the responsibility of disseminating short-, medium- and long-range forecasts in our country is the India Meteorological Department (IMD). At present, efforts are being made the world over to generate predictions over an intermediate time-scale, the so-called extended-range prediction with a lead time of 10 days to a month for rainfall, temperature, etc., averaged over about 5 days. Since spatial and temporal scales are inexorably linked, short-range forecasts are generated for the meteorological subdivisions of India (shown in Figure 3 in which the rainfall anomalies for the drought of 2002 are depicted for each subdivision) and for smaller spatial scales such as district level; whereas long-range forecasts are made for larger regions such as the all-India scale or for 3–4 sub-regions of the country.

In this article we first consider how the forecasts over short, medium and long range are generated (Section 2), mention an example of short range forecast (Section 3), and then focus on the problems and prospects of predicting the monsoon rainfall over India (Section 4). Finally we discuss measures of the skill of a forecast and the minimum skill which has to be attained by a forecast before it can be useful for decision making (Section 5).

2. How Do Meteorologists Generate Forecasts?

Scientists and laymen often find it difficult to understand the reasons for the painfully slow progress in forecasting the weather and climate in the modern-day milieu of satellites and computers. When solar eclipses can be predicted to fractions of a second and the position of a satellite millions of miles out in space can be pinpointed,
why can’t reliable weather predictions be made for a day, week, month, season or years in advance? In fact, the problem of predicting meteorological events (such as heavy rainfall over a region) is more complex because the atmosphere is unstable and the systems responsible for the events that we are trying to predict, such as clouds or a monsoon depression (in which thousands of clouds are embedded) are the culmination of the instabilities of the atmosphere\(^2\). They involve nonlinear interaction between different spatial scales from 3–4 kilometres (as in a single cloud) to hundreds of kilometres (as in a monsoon depression or a hurricane).

Let us first try to understand how predictions are generated. The state of the atmosphere at any point of time (in terms of temperature, wind, rainfall, etc. as a function of space) evolves according to Newton’s laws as applied to a compressible fluid in a rotating system. Hence the logical way of predicting the future state of the atmosphere (say 24 or 48 hours ahead) is to integrate the governing equations, starting with the observed state of the atmosphere at the initial instant as the initial condition, and the observed conditions at the surface of land or ocean as the boundary condition, for

---

24/48 hours. The errors in the short-range forecasts occur because (i) the models are not perfect (involving many assumptions like how sub-grid scale processes such as clouds affect the heating), and (ii) there are errors and gaps in the observations of the initial state.

An important question is: Even with a perfect model and high resolution observations, can we predict a week, month or a season ahead, the weather at a particular place at a specific instant viz. state of the atmosphere at that instant, at that point in space? In fact, even with a perfect model, it will never be possible to predict ‘weather’ more than about seven days ahead. This is because there is an inherent limit to predictability of weather. In a pioneering study Lorenz [3] showed that if we start integrating the governing equations from two very similar initial conditions (i.e., two similar states of the atmosphere), as they evolve, because of the instabilities in the atmosphere, the two solutions start diverging with time, i.e., the difference in the predicted states increases with time. By about seven days, the initial condition appears to be forgotten. The difference between the two states then becomes comparable to the difference between two states evolving from two randomly chosen initial conditions (not arbitrarily close ones as assumed earlier). Lorenz’s study introduced the concept of chaos and the atmosphere became the first known example of a chaotic system.

However, fortunately, every facet of the atmosphere is not chaotic on all time-scales. In fact, the variation of climatic elements averaged over different spatial and temporal scales (e.g., the interannual variation of the seasonal rainfall over the Indian region) arises partly from the variation of the conditions at the lower boundary of the atmosphere such as the sea surface temperature (SST) or snow cover over Eurasia. Hence such variables can be used as predictors for this time-scale. Thus seasonal forecasting is primarily a boundary value problem, while short- or medium-range weather forecasting is primarily an initial value problem. Extended range prediction will depend on the initial as well as boundary conditions.

The first short-range weather forecasts were made by meteorologists with empirical knowledge of how weather maps evolved from day-to-day. By the 1950s, development of physical models of the atmosphere on the one hand and detailed observations of the system on the other, led to insights into the physics of the variation on the scale of a few days. With the advent of satellites the density of observations increased enormously and with the phenomenal increase in computing power, complex models of the atmosphere that could simulate the short- and medium-range variation realistically, were developed by the 1980s. Now, the integration of such models with initial conditions obtained from the worldwide observation network, is a major input for weather prediction on these time-scales. Atmospheric models are run regularly for this purpose at IMD and the National Centre for Medium Range Weather Forecasting (NCMRWF).

Predictions on the seasonal to interannual scale can be generated by using ensemble runs of atmospheric models with specified boundary conditions and varying initial
conditions. Since oceans evolve more slowly than the atmosphere, the conditions at the surface of the ocean could be specified for these runs. Operationally at many centres in the world, long-range predictions are generated by running atmospheric models with specified boundary conditions or by running coupled models in which the oceans also evolve. For long-range predictions, an alternative approach is the traditional one, in which statistical models are used for prediction. These models are generally based on the links of the predictand (in our case rainfall) with prior values of that variable and/or other variables (such as pressure, temperature of the atmosphere or ocean, over the same or different regions of the atmosphere/ocean) discovered by analysis of large number of data sets.

3. Short Range Forecasts

We consider one example of a short-range forecast here. On 26/27 July 2005, Mumbai received unprecedented heavy rainfall, with its suburb Santa Cruz recording 94.4 cm of rainfall in 24 hours. There were reports of even heavier rainfall of 104.5 cm near Vihar lake. It disrupted life in the metropolis and led to a large number of deaths. The intensity of this event was not predicted either by IMD or by other operational forecasts generated by major weather prediction groups like UK Met office and US weather service. IMD’s prediction made 24 hours ahead, suggested a high probability of heavy rainfall (rainfall exceeding 12.5 cm) over the region. However, while rainfall at Mumbai exceeding 12.5 cm in a day is a very common event in the rainy season, rainfall over 90 cm in a day had never been experienced before. Had the forecast been more specific in terms of the probable intensity, the damage could have been reduced to some extent and a number of lives could have been saved.

A post facto analysis of the prediction of the Mumbai event [4] suggests that it would have been possible to predict the intensity of this event with reasonable accuracy, with high resolution atmospheric models, provided high resolution data (particularly on clouds organized over meso-scale and higher scales) available from satellites and quality-controlled local meteorological data were used in specifying the initial condition. Once the system which can assimilate relevant data from Doppler radars, from satellites, the high density meteorological observations in the metropolis as well as high resolution data on the terrain and land surface conditions is in place, it should be possible to generate reliable predictions of such events using the high resolution models available in the country.

4. Predicting the Indian Summer Monsoon Rainfall

4.1 Predictions with Dynamical Models

Consider first the prediction of the monsoon rainfall by integration of complex models of the atmosphere or the coupled atmosphere-ocean system based on equations
It is important to note that the breakthroughs in seasonal forecasting over the tropics have come from the phenomenal progress since the 80s in the understanding of the physics of El Nino–Southern Oscillation (ENSO) (see Parts 1 and 4), the dominant signal of the interannual variation of the coupled atmosphere-ocean system over the Pacific. The elucidation of the nature of ENSO, and unravelling of the underlying mechanisms led to development of models to a level at which they could realistically simulate the phenomenon and its impact on the climate of different regions.

Given the links between the Indian monsoon and ENSO, it was expected that it would be possible to simulate the interannual variation of the ISMR with atmospheric general circulation models when the observed SSTs are specified as a boundary condition. However, the results of several such studies suggest that the problem remains a challenging one. This is because, in addition to ENSO, another mode, the equatorial Indian Ocean Oscillation (EQUINOO), plays an important role in determining the interannual variation of the monsoon [5]. EQUINOO involves an oscillation between one phase in which the atmospheric convection over the eastern equatorial Indian Ocean is suppressed and that over the western equatorial Indian Ocean is enhanced and another phase with opposite signs of convection anomalies over the eastern and western equatorial Indian Ocean. Analysis of the simulations for the years 1979–95 by 20 state-of-the-art atmospheric general circulation models showed that while almost all models simulated the correct sign of the ISMR anomaly in 1988, a vast majority of the models failed to capture the anomaly for the excess monsoon season of 1994. Thus, the skill of the models in simulating the sign of the anomalies is not the same for all the droughts or excess rainfall years. During the excess monsoon season of 1988, ENSO was favourable while during 1994, ENSO was unfavorable but EQUINOO was favourable and excess rainfall occurred. On the whole, the skill of the models in simulating the sign of the anomaly for extreme ISMR seasons is higher for the extreme seasons which are associated with ENSO. Clearly more research and development effort is required to develop models which are capable of a realistic simulation of the links of the monsoon to EQUINOO.

The success of the atmospheric models in simulating the extremes of ISMR when they are linked to ENSO was achieved by concerted efforts under an international programme MONEG in the 90s under which the cases of 1987 and 1988 were studied with a slew of models. We now need research and development of atmospheric and coupled models to a level at which they can simulate realistically the response of the Indian monsoon to EQUINOO as well as ENSO. Once this is achieved it may be possible to generate reasonable predictions of the ISMR with dynamical models.

Predicting ISMR is one of the important mandates of the India Meteorological Department. Hence until the skill of the atmospheric and coupled models in simulating and predicting the interannual variation of the Indian monsoon improves, the empirical approach has to be adopted for operational forecasts.
4.2 Statistical Models

Forecasting of monsoon rainfall has been attempted for over a hundred years in India. In 1871 the Madras famine commission recommended that, “so far as it may be possible, with the advance of knowledge to form a forecast of the future, such aids should be made use of, though with due caution”. A major drought and famine occurred in India in 1877 soon after the IMD was established. The first long-range prediction in the world was made in 1886 by Blanford, who was the Chief Reporter of IMD, at the request of the colonial government in the wake of this drought. The prediction was based on the relationship between Himalayan snow cover and monsoon rainfall, discovered by Blanford in 1884. IMD has always been the responsible agency for the operational long-range forecasts of monsoon rainfall, which until recently have been based only on empirical models such as Blanford’s. Forecasts during the initial years were subjective and qualitative. In the early part of the last century, Sir Gilbert Walker initiated extensive studies of the worldwide variation of weather elements (e.g., pressure, temperature, etc.) to develop models for monsoon prediction. In 1909 Walker introduced an objective technique based on correlation and regression analysis. The first model used by Walker in 1909 for prediction of ISMR was a linear regression model based on four predictors (Himalayan snow accumulation at the end of May, South American pressure during March–May, Mauritius pressure in May and Zanzibar rain in April and May). However, assessment of the predictions by this model up to 1936 showed that, in spite of its early encouraging performance, the formula had broken down completely in the 15 years from 1921. While investigating the links of the Indian monsoon with atmospheric conditions over the rest of the globe, Walker discovered the Southern Oscillation, which is a see-saw of pressure between Darwin, Australia and Tahiti in the Pacific Ocean. This discovery was to play a major role in the phenomenal advances in the understanding and prediction of the interannual variability of the tropical ocean–atmosphere system witnessed over the last decade.

After the discovery of strong links between the El Nino and the Indian monsoon, the empirical models for monsoon prediction have developed rapidly. In the tradition of Walker, a large number of potential predictors have been identified by analysis of the ever-increasing data from conventional and satellite observations on many atmospheric and oceanic variables, and their lag correlation with the ISMR. Some of these parameters are related to El Nino and Southern Oscillation, others to snow over the Himalayas and Eurasia, and some to global and regional conditions on spatial scales ranging from one station (e.g., surface temperature at De Bilt in Holland) to hemispheric (e.g., northern hemispheric surface air temperature in January and February). In fact, as the sample of years increased with time, the correlation coefficient with several parameters became poor and for some of them even changed sign; hence many revisions were made on the model by changing the predictors. In 1988, IMD introduced the 16-parameter power regression and parametric models which were used operationally during the period 1988–2002. However there are very large errors
Figure 4. Performance of Operational Forecast (1988–2008).

in 1994, 1997, 2002 (Figure 4). In fact our analysis of the predictions generated by the empirical models used operationally by IMD during 1932–2002, suggests that the performance of these models, based on the relationship of the monsoon rainfall to atmospheric/oceanic conditions over different parts of the globe, has not been satisfactory [6].

After the failure of forecast in 2002, IMD introduced a new two-stage forecast strategy in 2003, according to which the first-stage forecast for the summer monsoon rainfall over the country as a whole is issued in April and the update is issued in June. Along with the update forecast, separate forecast for seasonal rainfall over broad homogeneous rainfall regions of India and July rainfall over country as a whole are also issued. In 2007, IMD introduced a new statistical forecasting system based on ensemble technique for the summer monsoon season using 8 predictors. In the ensemble method, instead of relying on a single model, all possible models based on all the combination of predictors are considered. Out of all the possible models, the best few models are selected based on their skill in predicting monsoon rainfall during a common period. The forecast is then generated from the weighted average of the forecast from the selected models. In fact the forecast for 2008 turned out to be rather accurate (Figure 4). But 2008 was a normal monsoon. Whether the model can forecast a drought or an excess rainfall season will be tested only when an extreme monsoon season does occur.

Our experience of the monsoon of 2006 suggests that incorporation of predictors associated with EQUINOO along with those associated with ENSO may improve predictions. Based on the analysis of predictors, IMD issued a long-range forecast for the 2006 monsoon season rainfall as 93% of long-period average. This inference of below normal was drawn based primarily on the warming tendency of SST
anomalies over the equatorial Pacific which suggested the development of an El Nino. The monsoon rainfall performance, in fact, was alarming till the third week of July with all-India cumulative rainfall departure being 13% below normal. However, rainfall activity revived by the third week of July and good rainfall activity extended almost unabated till the middle of September, thus improving the rainfall situation in the country. At the end of the monsoon season, seasonal rainfall was 100% of its long-period average. During August–September a positive phase of EQUINOO had developed with enhanced convection over the western equatorial Indian Ocean and suppressed convection over the eastern part. The enhanced rainfall during the second half of the monsoon season could be attributed to this. Had the development of positive EQUINOO phase by August been predicted, it might have been possible to predict that the deficit would certainly not be as large as expected from ENSO alone.

5. Assessment of Skill and Minimum Level Required for Applications

Obviously forecasts will be useful only if they are reliable. Hence it is important to assess the skill of forecasts generated by the different models used. For assessment of the skill of prediction of an event (such as rainfall at Mumbai greater than 70 cm in a day), a large number of predictions generated for an event have to be compared with the observations. Let the number of occasions on which the event was predicted and observed to occur be \(a\); in which the event was predicted but did not occur be \(b\); the number of occasions on which it was predicted not to occur, but occurred, be \(c\), and those for which it was predicted not to occur and did not occur, \(d\). The skill of the prediction is assessed by how large the hit rate (i.e., probability of correct forecasts \(= (a + d) / (a + b + c + d)\)) is vis a vis the false alarm rate (i.e., fraction of the times it was predicted but did not occur \(= b / (a + b)\)). The forecast is said to have a reasonable skill only when the hit rate is larger than the false alarm rate. For rare events, the threat score \(a / (a + b + c)\) is a more appropriate measure than the hit rate since \(d\) is much larger than \(a\) in this case.

In fact, the level of skill (i.e., the probability of correct forecast) has to be sufficiently high for a prediction to be useful for an application. How good is good enough, depends on the cost of adopting a strategy which is appropriate for the prediction and the expected benefit from such a change in strategy. For example, suppose that after an attack by insect pests, a farmer has to choose between two management strategies (e.g., to spray pesticide or not). Let the cost incurred in spraying the pesticide be \(C\). The benefit \(B\) in terms of enhancement of the yield due to the spraying of pesticides (which is necessarily greater than the cost \(C\) for spraying to be considered at all) will be realized only if it is not immediately followed by a wet spell. Thus, if it is predicted that a wet spell will not occur, the appropriate strategy would be to spray. However, it has been shown [7] that such a strategy will in fact be beneficial only if the probability that this prediction is correct is greater than \(C/B\).
At present in addition to the operational forecasts by IMD, predictions are also available from different meteorological centres of the world and some predictions are also generated by different groups in the country. However, a quantitative assessment of all the available predictions to determine their skill for different events of importance is yet to be made. Such an exercise has to be done on a continuing basis as and when new models are developed here or abroad, to ensure that predictions generated with models or a combination of models with the maximum possible skill are disseminated to users.

**Postscript (added by December 2021)**

Since 2010, there has been a significant progress in monsoon prediction capability in India with the sustained efforts made by the Ministry of Earth Sciences (MoES). In 2012, they launched the ambitious Monsoon Mission with a funding of Rs 400 crores and with an ultimate aim to improve monsoon prediction in all time scales. Till 2012, India Meteorological Department (IMD) was using a coarse resolution weather prediction model for short to medium range forecasts and indigenously developed statistical methods for seasonal forecasts. There was no methodology for predicting weather parameters and active-break phases of monsoon in extended range time scale of 10-30 days. Weather forecasts and warnings in extended range are absolutely required for proper planning in different sectors like agriculture and water management.

The first phase of the Monsoon Mission was completed in 2017 and the second phase was organized during 2017–2021 [8]. Ministry is now planning for the third phase of the said mission. Under the Monsoon Mission, the ministry could develop world class facilities for weather and climate forecasts in all time scales from short range (2–3 days) to seasonal (up to one season). State-of-the art dynamical models with higher resolution were adapted and introduced for operational forecasts with the technical support from the US Weather Services (NCEP) and UK Met office. Academic institutions in India and abroad have participated in this effort to develop the world class capability in monsoon prediction. IMD now uses two versions of global weather prediction models with a resolution of 12 km, which is one of the highest resolutions being used by weather prediction agencies around the world. For accounting uncertainties involved in initial conditions and model skill, an Ensemble Forecasting System (EFS) also was introduced for the first time in India. With the EFS, we are now able to estimate probabilities associated with severe/extreme weather events with a substantial lead time.

About 10 years ago, dynamical models had limited capability in prediction of seasonal mean monsoon rainfall and its variability. Due to systematic efforts made by monsoon researchers around the world, monsoon prediction skill of dynamical models has substantially improved during the recent years. This improvement in seasonal forecasts inspired the Ministry to adapt a high resolution dynamical climate model
for making seasonal forecasts. A Dynamical Seasonal Forecasting System was introduced in 2017 for the first time, which is being used by IMD. For understanding global warming and associated changes in extreme events, MoES also took an initiative for developing an Earth System Model (ESM) for making climate change projections. An ESM treats the whole Earth System as an interactive physical system. The ESM developed by the Centre for Climate Change Research (CCCR), Pune is being used to make climate change projections. For the first time, the climate change projections made from this model were included in the sixth climate change assessment report, prepared by the Inter-governmental Panel on Climate Change (IPCC).

Systematic and dedicated efforts made by MoES and other academic institutions are responsible for this quantum jump in monsoon prediction capability. However, there is still scope for further improvement in monsoon prediction in all time scale. Seasonal forecast is still below its potential limit. Short range forecasts can be further improved to get additional lead time of couple of days. Many grey areas in modeling are still remaining. We do not know fully how physics and dynamics of clouds and oceans work. Systematic atmospheric and oceanic observations including satellite data, dedicated efforts in modeling (higher resolution Ensemble systems), adoption of Artificial Intelligence/Machine Learning methods, better data assimilation techniques, investment in high performance computers and more importantly training of young students are required for further improvement in skill of monsoon prediction [7].

Suggested Reading

Surface topography (whether subaerial or submarine) is a product of plate tectonic, geological, climate and surface processes. Plate tectonic processes, predominantly operated through earthquake cycle, dominate, at least in the seismically active regions, and are responsible for a significant amount of topography build up. On a shorter time scale, they appear to cause devastation, but on a longer time scale, they lead to the development of diverse landscapes which not only makes them look beautiful but also inhabitable. The topography build-up, due to deformation caused by earthquake processes, depends upon the mechanism of deformation and earthquake occurrence. The earthquake occurrence mechanism is different in the plate boundary regions as compared to that in the plate interior regions. Here in this article, I discuss the nuances of crustal deformation in the interplate and intraplate regions in the context of India and India plate.

1. Introduction

We enjoy holidaying in the Himalaya, or in the Andaman island belt. The diverse landscape and diversity in the flora and fauna attract us. However, many a time we do not realise what actually causes such diversity. We just enjoy the diverse climate and scenic beauty of these regions and forget that the scenic landscape and climate are the product of tectonic, geological, climate and surface processes which occur through phases of active and quiescence periods. The active period could be short but could be very dramatic (which can sometimes be disastrous, e.g., the occurrence of earthquakes, landslides, floods, etc.), while the quiescence periods are generally longer and calm but these are the periods when preparations for those catastrophic events take place. It is the active period which changes the landscape most dramatically. If we happen to witness any of the active phase (e.g., the occurrence of a large magnitude earthquake), leading to loss of lives and property, then we are reminded of the power of these processes and forces shaping the landscape. So the processes which shape these beautiful and scenic landscapes, cause losses in a shorter time scale while on a longer time scale, they make these places attractive, distinct and inhabitable. Imagine life without plate tectonic forces; we wouldn’t be seeing the beautiful Himalayan ranges, no topography in the north would have led to less rainfall in the region, less rainfall and lack of eroded soil from the mountains would have made the plains less

DOI: https://doi.org/10.1007/s12045-021-1123-2

Keywords
India plate, interplate, intraplate, crustal deformation, strain, earthquake, slip.
Imagine life without plate tectonic forces; we wouldn’t be seeing the beautiful Himalayan ranges, no topography in the north would have led to less rainfall in the region, less rainfall and lack of eroded soil from the mountains would have made the plains less fertile. Similarly, no plate tectonic forces would have deprived us of the Andaman island belt or the vast ocean basins. Absence of climate and surface processes would have deprived us of the beautiful sand dunes of the Thar desert or the flat and fertile Indo-Gangetic plains.

It is now known that the landscape evolves in response to various processes which can broadly be classified into three categories—tectonic, geologic, climate and surface processes (Figure 1). Tectonic and geologic processes operate over a long time and involve deeper levels of earth, while the climate and surface processes (fluvial, mass wasting, glaciation, erosion, deposition, etc.) are generally shallow, and sometimes may be confined in spatial extent. Also, sometimes they are influenced by anthropogenic activities. However, both, tectonic (and geologic) and climate (and surface) processes, influence each other and often it is difficult to distinguish their effects. They may work in tandem or one may influence or lead to the other. For example, it was earlier believed that the uplift of Himalaya and Tibet led to the cooling and ice ages. But then this idea was challenged, and it was suggested that as the climate changed to more glacial conditions in the late Cenozoic times, enhanced rates of erosion within mountain belts caused increased rates of valley incision, which in turn incited isostatic uplift of the residual peaks [1]. Amongst the two processes, plate tectonic processes contribute the maximum and operate largely through earthquake cycles. In this article, I attempt to explore the diverse nature of deformation in the India plate boundary and interior regions. Broadly speaking, I attempt to address as to how the continents deform in response to the earthquake cycle. I acknowledge here that geologic processes have also shaped the Indian landscape, e.g., the large Deccan volcanic region (occupying parts of Maharashtra, Madhya Pradesh and Gujarat), which developed some 65 million years ago (Mya) due to plume activity. Similarly, the climate and surface processes (which includes the erosional and depositional processes and also the processes caused by human activities) also shape the landscape, e.g., Indo-Ganga flood plains or Thar desert. But, I will confine the discussion to plate tectonics and earthquake deformation.

Figure 1. A schematic depth section across the Indo-Gangetic plains, Himalaya and Tibetan plateau showing the climate, surface and tectonic processes in the context of the India plate and Himalayan region.
Figure 2. India plate along with the surrounding Africa, Antarctica, Sunda and Eurasia plates. India plate moves towards the northeast at a rate of $\sim 5$ cm/year. The right panel shows the zoomed version along with the earthquakes of magnitude more than five from 1970 to 2020 from United States Geological Survey (USGS). Stars mark the major and great earthquakes of the past 200 years. Koyna, Bhuj, Jabalpur earthquakes are also shown.

2. Earthquakes and Crustal Deformation Associated with India Plate

The theory of plate tectonics developed in the 1960s suggests that the seven major plates of the Earth, which consists of the crust and upper part of the mantle (extending up to a depth of $\sim 100$–$120$ km), are moving continuously with respect to each other in response to various plate driving forces; convection current in the mantle is one of the most prominent among them. In the process of relative motion between them, they interact with neighbouring plates either through convergence (at subduction/collision zones) or through divergence (at mid-oceanic ridges). At some places, the two plates, slide past each other (e.g., the San Andreas fault between the North American and Pacific plates). Majority of the relative motion between the two plates, particularly at the subduction/collision boundaries, is accommodated through earthquakes, as some of it is also accommodated in an aseismic manner (i.e., without causing earthquakes). The India plate is surrounded by the Africa and Arabia plates to the west and Eurasia and Sunda plates to the north and east (Figure 2). It rotates along a point (known as the Euler pole) somewhere in Germany, due to which the India plate moves at a rate of $\sim 5$ cm/year in the northeast direction and collides with the Eurasian plate in the north.
and subducts under the Sunda plate in the east. This convergence has led to the development of the Himalaya, the Indo-Burmese mountain arc, and the Sumatra-Andaman island arc. Thus, the resulting topography is largely the outcome of the convergence process which is achieved by the episodic movement through earthquakes. Hence topography, earthquakes and deformation are all linked together and are all outcome of plate tectonic movement and other climate and surface processes. In the past two centuries, 1905 Kangra, 1934 Nepal-Bihar, 1950 Assam (now Arunachal), 2004 Sumatra Andaman, 2005 Kashmir and 2015 Gorkha earthquakes are the examples of major and great earthquakes of India plate boundary region.

Larger the frequency of earthquakes, more the deformation and hence earthquakes and that surprises all of us because as per plate tectonics theory, the continent interiors should not deform and the deformation should occur only at the plate boundary. But now we understand that the continental interiors are not rigid and have weak zones, in terms of faults. Besides this, there are other factors (e.g., anthropogenic, meteorological and climatic processes) which influence stress state in the plate interiors which can cause earthquakes and deformation. The 1819 Allah Bund, 1967 Koyna, 1970 Bharuch, 1993 Killari, 1997 Jabalpur, and 2001 Bhuj earthquakes are the examples of plate interior earthquakes. Thanks to seismological and geodetic observations, now we understand that there is a marked difference between the mechanism of earthquake occurrence and accompanying deformation in the plate boundary and plate interior regions. Because of plate interaction at plate boundaries, earthquake frequency is large and so is the size of earthquakes. Larger the frequency of earthquakes, more the deformation and hence more pronounced the topography or the elevated landscape. Thus, topography is the proxy of the plate tectonics and the frequency of earthquakes. This is the reason we observe high topography in the plate boundary regions, e.g., the Himalayan region (Figure 2). In the plate interior regions, the topography is not that prominent.

However, even the subdued and smooth topography in the plate interior region implies that either the region was tectonically active earlier or/and could be active even now. Other than the difference in topography between the plate interior and the plate boundary regions, which is directly linked to the earthquake activity (both in terms of number and magnitude), the mechanism of earthquake occurrence or deformation is quite different in the two regions. The basic difference is evident from the definition itself. In the plate boundary regions, it is the interaction of two plates which causes earthquakes and resulting deformation (hence these earthquakes are referred to as the interplate earthquakes), whereas in the plate interior regions, it is the weak zones and mass heterogeneities due to its geological evolution within the plate which cause earthquakes and resulting deformation (hence these earthquakes are referred to as the intraplate earthquakes). Majority of earthquakes in the India plate interiors occur in regions with clear geomorphic surface expressions in terms of topography and geological diversity. These regions are referred to as the paleo/aborted/failed rift regions.

Majority of earthquakes in the India plate interiors occur in regions with clear geomorphic surface expressions in terms of topography and geological diversity. These regions are referred to as the paleo/aborted/failed rift regions.
failed rift, and Mahanadi failed rift. They all developed as nascent rifts sometime in the late Triassic period (∼ 220 Mya) when India plate was breaking away from the Africa and Antarctica plates during the Gondwana land breakup. However, due to some change in the direction of motion of the India plate, sometime in late Cretaceous period (∼ 100 My ago), the rifting could not sustain along these regions and the India plate continued its journey northward as a single plate. Had the rifting along those regions continued, the extent and shape of India plate would have been different, and possibly we would not have been referring it as the India plate, but by some other name. Nevertheless, these regions developed some weak planes through which high-density material from beneath came out, thus causing mass heterogeneities. Because of the continuous ongoing Indian plate movement, these regions accumulate strain and produce deformation and earthquakes along the failed rift zones. However, a large part of the topography in these regions (e.g., Vindhyan, Aravalli and Satpura mountain regions) is the relic of past geological and surface processes and currently, earthquakes contribute very little. We also see mountain chains close to the east and west coast of India. It is considered that this topography represents the retreated rifted margins of India which India plate once shared with other continents during the Gondwana breakup. Thus the built-up topography in these regions is mainly due to previous episodes of plate tectonic processes, and currently, erosion and crustal upwarping/flexure modify it. The seismicity in some such regions is largely due to processes other than plate tectonics.

Then, there are regions which have experienced strong and damaging earthquakes (e.g., the 1993 Latur earthquake) but do not exhibit any pronounced topography or do not have any geomorphic expression on the surface. These earthquakes do not appear to be linked with any known paleo structure and are truly stable continental region earthquakes.

3. Deformation Mechanism

The basic tenet of earthquake mechanism and deformation is the elastic rebound theory [2]. After the 1906 San Francisco earthquake, it was observed that the earthquake caused an offset in a fence that was built across the San Andreas fault. Henry Fielding Reid, Professor of Geology at Johns Hopkins University, concluded that the earthquake must have involved an ‘elastic rebound’ of previously stored elastic strain (Figure 3). It was as early as 1819, during the Allah Bund earthquake in the Rann of Kutch, when a fault scarp was noticed and was ascribed to the earthquake, but it took another ∼100 years to understand the mechanism and to formulate the elastic rebound theory in 1910. It may be noted here that at that time, the theory of plate tectonics was not in place and we had no idea how earthquakes occur. The concept of elastic rebound theory is applicable both to intraplate (plate interiors) and interplate...
Figure 4. Geodynamic models described by the relationship among earthquake characteristics, and tectonic structures. (a) Plate boundary: Earthquakes are associated with large-scale lithospheric structures. (b) Random: Large earthquakes can occur anywhere, but their magnitudes and recurrence intervals are limited by the low strain rate (e.g., Killari earthquake). (c) Localized weak zones: Earthquakes occur along small, local structures associated with a weak zone in the otherwise strong lithosphere (e.g., Koyna-Warna region). (d) Large scale weak zones: Earthquakes focus along long paleotectonic structures (e.g., the Narmada Son or Kutch failed rift) associated with a weak layer in the lower crust and/or upper mantle. Lateral variations in lithospheric strength concentrate strain along the structures [3].

In the plate boundary region, it is the relative motion between the two plates which accumulates as strain at plate boundary and gets released during the earthquake. This is referred to as the ‘stick’ and ‘slip’ motion and causes convergence between the two plates (Figure 4a). I cite an example from the Garhwal-Kumaun Himalaya here. Geological investigations in the region suggest that a long term convergence rate between the Indian plate and southern Tibet is \( \sim 18 \text{ mm/year} \). As per the elastic rebound theory, this convergence is achieved through strain accumulation (stick) and release (slip) mechanism. Thus in the interseismic period of strain accumulation, the underneath fault should be locked and should accumulate elastic strain (i.e., recoverable deformation) at the slip deficit rate of \( \sim 18 \text{ mm/year} \). This deficit will accumulate over a long period (say a few hundred years) until the strain reaches the strength of the rocks and then suddenly the accumulated strain is released during a great or major earthquake (i.e., coseismically). Thus the released strain balances the slip deficit accumulated in the interseismic period. Since the accumulated strain is elastic, the displacement rate across a region should exhibit a gradient in the rate, increasing from south to north (Figure 5). At the extreme north, where the underneath fault is not locked, it should correspond to the convergence rate of \( \sim 18 \text{ mm/year} \). In Figure 5, the variation in GPS measurement rates along an arc normal profile exactly exhibits that type of behaviour. The deformation rate is close to zero at sites which are either on the India plate or on the southern extreme of the Himalayan wedge. The velocity at sites increases gradually and reaches to a value of \( \sim 14 \text{ mm/year} \), and does not increase any further, which actually corresponds to the convergence rate of \( \sim 18 \text{ mm/year} \). The basic feature of the deformation profile exhibiting strain accumulation is that the far off sites from the strain accumulating fault show convergence, even though the fault is locked.
In contrast to the interplate regions, there may not be a long-distance convergence across a strain accumulating region in the intraplate regions. Thus, earthquakes in these regions may occur due to tectonic (e.g., internal plate deformation across weak zones) along with geologic (e.g., mass heterogeneity), climate and surface (e.g., erosion leading to flexure/upwarping) processes. In the context of the India plate, the deformation in the intraplate region can be classified using three models [3].
(1) The random model (Figure 4b), defined by the lack of significant lithospheric structure with no apparent geomorphic expression. The seismicity in such cases could be spatially and temporally random. The 1993 Killari (Latur) earthquake comes under this category. Sizes of such earthquakes are generally moderate to strong and their occurrence is random and rare. Nevertheless, such earthquakes pose the maximum hazard as they occur at unexpected places and take people by surprise. Our understanding of such earthquakes is minimal. The deformation caused by these events is highly localized, and the region may not witness another earthquake in the next few hundred or thousand years and may not even show the evidence of renewed strain accumulation. These earthquakes and their corresponding deformation migrate in space and time.

(2) The localized weak zone model (Figure 4c) assumes that earthquakes are limited to small areas of crustal weakness (e.g., the Koyna Warna region in Maharashtra). The deformation is very low (less than 1 mm/year) and localized. Even these regions may or may not have clear geomorphic expression and may experience repeated earthquakes in one’s lifetime. The Koyna Warna region where a strong earthquake occurred in 1967 is a good example of this category. The region is about 20 km long, and the GPS measurements across it show a relative motion of less than 1 mm/year [4].

(3) The large-scale weak zone model (Figure 4d) is characterized by moderate strain concentration (a few mm/year) across the major paleo-rift regions (e.g., Kutch, Narmada Son, Godavari) with clear geomorphic expression. These regions may experience major earthquakes that can cause huge damage, e.g., the 1881 Allah Bund and 2001 Bhuj earthquakes. Other examples of moderate magnitude earthquakes in such regions are the 1970 Bharuch, 1956 Anjar, 1997 Jabalpur, 1936 Satpura and 1969 Bhadrachalam earthquakes. It is possible that due to the compression exerted by the northeastward moving India plate, the entire failed rift may not be active at all times. The earthquakes and corresponding deformation may migrate along the failed rift, thus exhibiting spatial and temporal variations in strain rate. I cite here an example of the Kutch region [5]. The GPS measurements of crustal deformation here suggest that the Kutch paleo rift is under regional compression from both sides at a rate of \(\sim 4-5\) mm/year (Figure 6).

4. Concluding Remarks

Elastic rebound theory and plate tectonics explain the earthquake process and deformation associated with it. The two theories and the available geodetic data also help in differentiating the deformation mechanism between the interplate and intraplate earthquakes (Figure 7). However, even now the mechanism of intraplate earthquakes is not fully understood, particularly in case of randomly occurring earthquakes. Probably the resolution of geodetic data is not precise enough even now to capture the sub-mm/year level deformation associated with these earthquakes. Other questions pertaining to crustal deformation due to earthquake processes are: Do the rates of
Figure 6. (a) GPS measurements of crustal deformation (2006/09–2016) from the Kutch and adjoining region. Estimates of site velocity are shown in the Indian reference frame. Focal mechanism solutions of the 2001 Bhuj and 1956 Anjar earthquakes are also shown. Red filled circles are the aftershocks of the 2001 Bhuj earthquake (2001–2016). Stars denote historical earthquakes of past 200 years. Yellow lines are the major faults in the region. (b) Top panel shows the variation of northward site velocity along a north-south profile. The three curves show the modelled response of slip on faults on either sides of the paleo-rift. Bottom panel shows a schematic depth section across the Kutch paleo-rift and location of faults on the paleo-rift flanks. Locations of 1845, 1956, 2001 and 1819 earthquakes are only representative. Topography is exaggerated. KHF–Katrol Hill Fault, KMF–Kutch Mainland Fault, ABF–Allah Bund fault, NPF–Nagar Parkar Fault [5].
Figure 7. Difference in the interplate and intraplate deformation during interseismic period. The interplate deformation is generally widespread, large, and show long distance offset whereas, the intraplate deformation is localised, small and does not show any offset. Lower part of the figure in each case shows the locked faults in the two scenarios in a depth section.

deformation remain steady over time? Do all faults in a region rupture regularly and produce similar magnitude earthquakes? Can we really define the maximum magnitude earthquake for each fault and does it remain steady over time? Does the behavior of earthquakes and deformation remain similar on a fault over several earthquake cycles? And the biggest question is: Can we predict earthquakes by monitoring deformation in a region? We are still exploring the answers to these questions.

Acknowledgment

I am thankful to Dr Nandini Nagrajan for encouraging me to write this paper. I thank the anonymous reviewer who saved me from my overenthusiasm and my strong bias towards earthquake processes leading to topography build up. This is CSIR-NGRI contribution number NGRI/Lib/2020/Pub-43.

Suggested Reading

Resonance – Journal of Science Education started in 1996 was the culmination of intensive efforts by the Academy through a panel, which studied the state of science education in the country. The journal is targeted primarily at science education for undergraduate students and teachers and focuses on enriching the processes of teaching and learning science.

Resonance carries scientific articles written in a lucid style, on topics within the undergraduate curriculum, especially those that students often consider difficult to understand, new classroom experiments, emerging techniques and ideas, and innovative procedures for teaching specific concepts. Each issue highlights the contributions of an important scientist.