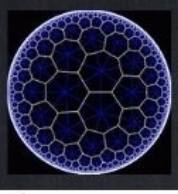
Blackboard

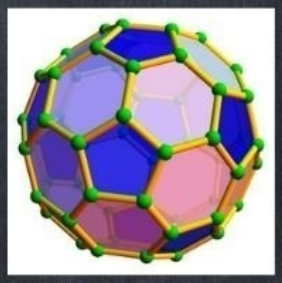


Issue 6

MTA (I)

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

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



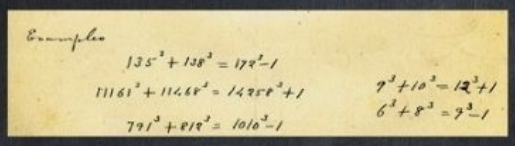
Here is the Ramanujan-Hardy formula for the calculation of the number of partitions:

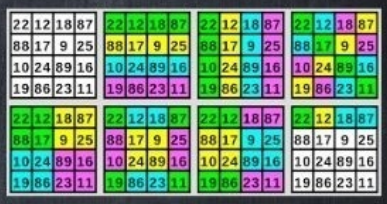
$$p(n) = \frac{1}{2\sqrt{2}} \sum_{k=1}^{v} \sqrt{k}\, A_k(n) \frac{d}{dn} \exp\left(\pi\sqrt{\frac{2}{3}}\frac{\sqrt{n-\frac{1}{24}}}{k}\right)$$

where

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

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





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

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

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

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

Registered Office

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

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

Blackboard

Bulletin of the Mathematics Teachers' Association (India)
Issue 6

January 2024

Contents

Editorial

Today is the National Mathematics Day 2023. If the great man after whom this day is so dedicated had been alive, he would have been 136.

We take this opportunity also to remember some of the celebrated mathematicians who have passed away recently. Foremost among them is C R Rao, who breathed his last 19 days before his 103rd birthday. Some of the other eminent mathematicians who left us in 2023 are Martin Davis who is known for his contributions to the solution of Hilbert's 10th problem, David Singmaster who was a pioneer of recreational mathematics and is also known for his book on the mathematical solution to the Rubik cube problem, and Robert Zimmer who greatly championed freedom of expression in higher education.

In this sixth issue of Blackboard, we have once again a representation of very diverse topics. A school teacher Pranashish Chandra Rinku talks about the importance of terminology in mathematics in order to have precision in communication. In particular, the popular usage of 'into' and 'by' may have exactly the opposite meanings, he argues.

We now have a regular contributor Jyotirmoy Sarkar from a US university who writes on a range of topics. In this issue, he has 3 articles on topics ranging from Probability to problems on slicing prisms and cylinders which motivate students to learn calculus and solve optimization problems. To exhibit the kind of problems discussed in the article on probability, consider the question of how many distinct values would be seen if a 12-sided die (a doedecahedron) is rolled 6 times. The author discusses the methods of simulation, complete enumeration, and mathematical derivation; apart from being interesting in its own right, this article should be of enormous help to instructors who teach probability. Incidentally, the rolling of a 12-sided die kind of problem can be formulated as a real-world problem as well. In the second article, the author employs a paper-folding problem to demonstrate how students may be motivated to study calculus and optimization problems; the phrasing of these problems are in terms of slicing a prism. The third article by Sarkar is written in collaboration with another author Colin Tully from another US university; it talks about slicing cylinders and, this too has applications to optimization problems.

Shailesh Shirali and Amber Habib have continued their series articles. Shirali writes on the history of calculus, and Habib on Tarski's formulation of the completeness axiom of real numbers. In addition, Shirali also writes about the beautiful nine-point circle

theorem, a classical result that does not seem to be as well known as it should be. Geetha Venkataraman descibes Euler's solution of the Konigsberg bridge problem in a captivating piece which has lovely pictures as well.

Jugal Verma and Lovy Singhal have written a completely self-contained narrative on Platonic solids. Precise mathematical formulations and complete proofs are hard to find in a single source, and this article is very valuable in that respect. It is ideal for discussion in a college classroom, where basic group theory has already been introduced.

In summary, many teachers are unaware of the existence of MTA (India), and the bulletin does not seem to have reached enough people. Though we would like to have many teachers contributing, this endeavor has not been successful so far.

On the occasion of bringing out this sixth issue of our bulletin, here is wishing all of the readers a very Happy New Year with beautiful and stimulating mathematical pursuits to dream of.

B. Sury
Indian Statistical Institute Bangalore
22 December, 2023

1 The Königsberg bridges problem and Euler's solution

Geetha Venkataraman

Dr. B. R. Ambedkar University Delhi
Email: geevenkat@gmail.com

1 The seven bridges of Königsberg

In the 1700s, the city of Königsberg was part of East Prussia. The river Pregel flowed through the city and created an island called Kniephof (meaning pubyard) as shown in the diagram below. There were seven bridges on the river.

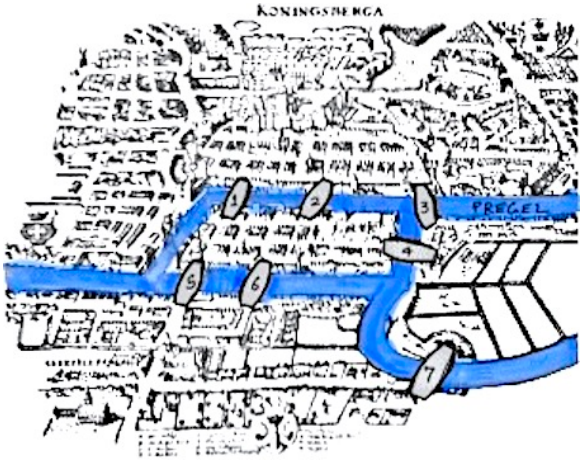


Figure 1: The seven bridges of Königsberg

The local names for the bridges were as follows: Bridge 1 was called Kraemer or Shopkeeper, Bridge 2 was called Schmiede or Blacksmith, Bridge 3 was known as Holz or Wooden, Bridge 4 was Honig or Honey, Bridge 5 was called Greune or Green, Bridge 6 was known as Koettel or Guts and the final Bridge 7 was Hohe or High. A puzzle that

the townsfolk wanted to solve was whether one could walk a trail in which each bridge is crossed but once and only once.

Königsberg is now the city of Kaliningrad in Russia. Based on the article [2] written a decade ago, of the seven original bridges, only two survive in their original form. Two were destroyed during World War II. Another pair was replaced by the Russians during road constructions. The seventh survives but is a version that was rebuilt in 1935 by the Germans. Below is a view of Königsberg as it was in the 18th century and a picture of the island in modern Kaliningrad with the bridge on the right being one of the original bridges.

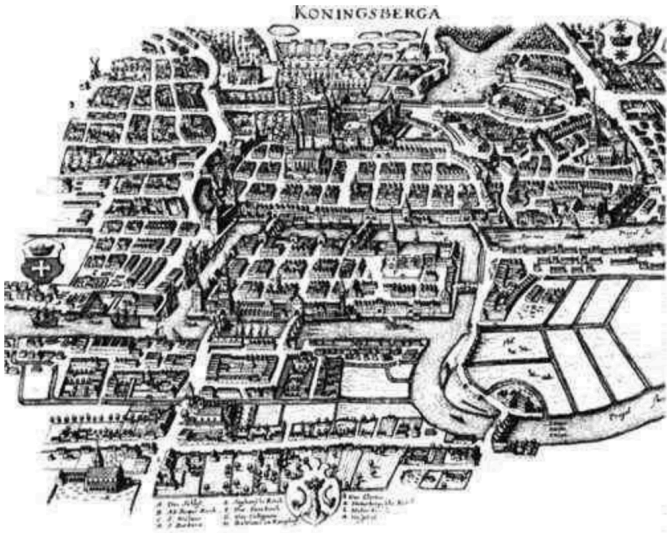


Figure 2: Königsberg in the 18th century

Figure 3: The island in 2007 with one of the original bridges on the right.

2 Euler's solution

The problem of walking a path in which the seven bridges of Königsberg are crossed once and exactly once was truly a puzzle. No one could find such a trail nor could they prove that such a trail could not exist. In 1736, Euler [1] not only proved that such a trail was not possible but in the process also nurtured the birth of Graph Theory. To learn more about Euler's life and work, see [3]. Below is an excerpt from the translation of Euler's 1736 paper.

2. The problem, which I understand is quite well known, is stated as follows: In the town of Königsberg in Prussia there is an island A, called "Kneiphof," with the two branches of the river (Pregel) flowing around it, as shown in Figure 1. There are seven bridges, a, b, c, d, e, f and g, crossing the two branches. The question is whether a person can plan a walk in such a way that he will cross each of these bridges once but not more than once.

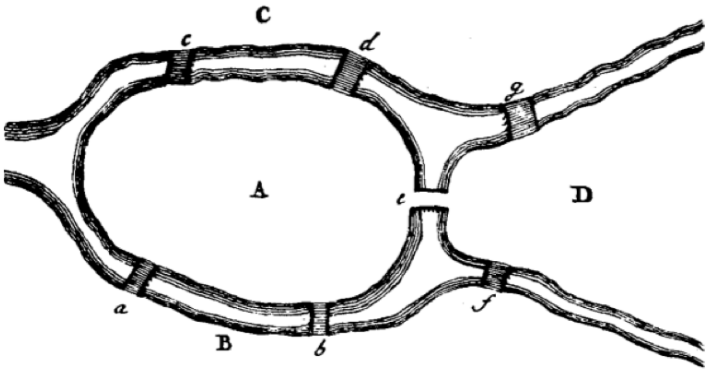


FIGURE 1

Figure 4: Excerpt from Euler's paper

Euler further writes in the paper that while all the possible paths could be carefully tabulated and inspected to find out whether the Königsberg Bridges Problem (KBP) had a positive solution or not, it would be very tedious because of the large number of possibilities. He also adds that if there were more than seven bridges in a similar type of problem then it would also be impossible to use this brute force method. He therefore decides that he should take a general case of the KBP problem and then try to see if there might be a solution or not. His generalised postulation was to consider a river with many branches (not just two as in the case of KBP) and any number of bridges and to determine whether there was a path or trail in which one could cross every bridge once but not more than once.

A key step in Euler's proof was the manner in which he labelled the land masses that are separated by the river. He then then used this to describe the trail or path made while traversing bridges from one land mass to the other. Consider the figure below.

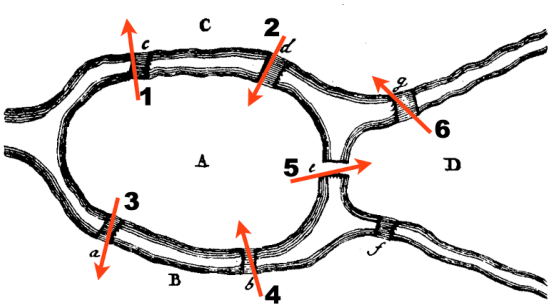


Figure 5: Labelling land masses separated by the river

We see from Figure 5 that the branches of the river Pregel have created four land masses. These have been labelled A, B, C, D. Euler uses the label $ACABADC$ to describe the path shown through the red arrows. Note that the label does not bother about which bridge has been used to traverse from one land mass to the other but tracks only that some bridge has been used to cross from one land area to the other. In the case described in the above figure, a person has started from land mass A, then moved to land mass C then back to A, then on to B, back to A, then to D and from there to C. Note that six bridges have been crossed and this has been represented by a sequence of seven letters, namely, $ACABADC$.

So if seven bridges have to be crossed, then this will be represented by a sequence of eight letters. Also Euler argued that if each bridge has to be crossed and crossed only once then we must have a sequence of eight letters in which AB (or BA in combination) will have to appear exactly twice as the bridges a and b both have to be crossed and they can each be crossed only once. Similarly for AC or CA. Thus AB, AC has to occur exactly twice whereas AD, BD, CD occurs exactly once in the sequence of eight letters representing the crossing of the 7 bridges, once and not more than once.

Now for Euler the problem boils down to whether we can create a sequence of eight letters using A, B, C, D such that AB, AC occurs twice and AD, BD, CD occurs only once. He then goes on to say that one should first check if such a creation is theoretical possible so that one does not waste time in constructing something impossible!

Euler then considers the configuration of two land masses and six bridges given in Figure 6.

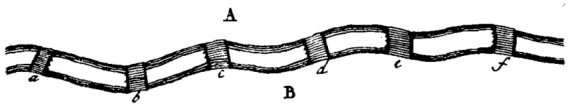


Figure 6: Two land masses and six bridges

He notes that if only one bridge is traversed A will occur once. If three bridges are traversed then A will occur exactly twice irrespective of whether you start from A or not. For example $AaBbAcB$ or $BdAeBfA$. he then concludes that if an odd number n of bridges are crossed between A and B, then A will occur exactly $(n+1)/2$ times in the sequence of letters representing the land masses traversed.

He then uses this knowledge to solve the KBP problem. Let us consider the four land masses A, B, C, D and the seven bridges a, b, c, d, e, f as in the figure below.

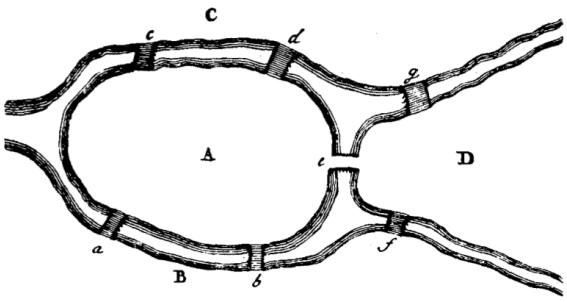


Figure 7: Seven bridges and four land masses

Also remember that crossing every bridge once and exactly once will mean a sequence of 8 letters involving A, B, C, D in which AB, AC occurs twice and AD, BD, CD occur exactly once. Now we apply the analysis which we got, namely that if an odd number n bridges emanating from a land mass A are crossed then A will occur in the sequence $(n+1)/2$ times. From Figure 7, we see that there are 5 bridges connecting land mass A to other land masses. Thus if each of these bridges is crossed exactly once, then A will have to occur $(5+1)/2 = 3$ times in the sequence. By a similar argument B, C and D will have to occur twice each. But then we have a sequence of 9 letters at least, which means that at least 8 bridges have been traversed. This means that at least one bridge has to be crossed twice. He then concludes that there is no path or trail by which one can cross the seven bridges of Königsberg once and exactly once. See the figure below for an excerpt from his paper.

For the Königsberg problem I would set up the tabulation as follows:

Number of bridges 7, giving 8 (= 7 + 1) bridges

A,	5	3
B,	3	2
C,	3	2
D,	3	2

The last column now adds up to more than 8, and hence the required journey cannot be made.

Figure 8: KBP has a negative answer

Euler also examines the case when an even number of bridges are traversed between two land masses. He concludes that if there are an even number n of bridges emanating from a land mass A, and the bridge traversing journey begins at A then the land mass A will occur will occur $(n/2)+1$ times and that it will be $n/2$ if the journey does not begin at A.

His final method is described as follows to decide whether a KBP type problem in general has a positive solution or not:

(1) Label the land masses separated by the water as A, B, C etc. This forms the first column as shown in Figure 8.

(2) Take the total number of bridges, say, m and write $m+1$ on top.

(3) In the second column, next to each capital letter, write the number of bridges that lead to or emanate from the region.

(4) Mark by an asterisk, those capital letters that have an even number of bridges leading into them.

(5) In the third column write $n/2$, if the number in the second column is the even number n. If the number in the second column is the odd number n then write $(n+1)/2$ corresponding to it in the third column.

(6) Add up the numbers in the third column. If the sum is m or equal to $m+1$ then Euler concludes that there is a positive solution. He however adds that when the total of the third column is equal to $m+1$ then the journey has to begin with a region not marked by an asterisk, whereas if the total of the third column is m then the journey must begin with a region marked by an asterisk.

(7) If the total of the third column is not m or $m+1$ then no such journey which crosses each bridge once and exactly once is possible.

He further analyses that the total sum of the numbers in the second column is twice the number of bridges. He then concludes that there have to be an even number of odd numbers occurring in the second column as the sum of the second column would be odd otherwise.

From this it is easy to see that if all the numbers in the second column are even, then the sum of the third column would be equal to m and so a positive solution to a KBP type problem will occur and any region can be a starting point. Similarly if exactly two entries in the second column are odd and the rest are all even, then the sum of the third column will be $m + 1$ and once again a positive solution for a KBP type problem will exist provided the journey is started at any of the two regions with an odd number of bridges leading to them. In all other situations the sum of the third column will exceed $m+1$ and so we cannot have a journey by which every bridge is crossed once and exactly once. The final conclusion in Euler's words is:

> **If there are more than two regions which are approached by an odd number of bridges, no route satisfying the required conditions can be found.**
>
> **If, however, there are only two regions with an odd number of approach bridges the required journey can be completed provided it originates in one of the regions.**
>
> **If, finally, there is no region with an odd number of approach bridges, the required journey can be effected, no matter where it begins. These rules solve completely the problem initially proposed.**

Figure 9: Solution to KBP type problem according to Euler

Now a days KBP type problems fall into the realm of graph theory. For an easy introduction to KBP type problems using graph theory and a graph theoretic solution to such problems see [4].

A note on the figures used

- Figures 4-9, have been taken from Euler's paper [1].
- Figure 1 has been taken from https://en.wikipedia.org/wiki/File:Konigsberg_bridges.png and modified under the CC ASA 3.0 license.
- Figure 2 is from https://mathshistory.st-andrews.ac.uk/Extras/Konigsberg/ and is likely in public domain.

- Figure 3 has been taken from https://en.wikipedia.org/wiki/File:Old_cathedral_of_Kaliningrad_in_Russia.jpg and is in public domain.

Bibliography

[1] L. Euler (1736), Solutio problematis ad geometriam situs pertinentis, Commentarii Academiae Scientiarum Imperialis Petropolitanae, Comm. Acad. Sci. Imp. Petropol, 8 (1736) 128–140 = Opera Omnia (I) 7 (191–156), 1–10, http://eulerarchive.maa.org//docs/originals/E053.pdf.

[2] Peter Taylor, What Ever Happened to Those Bridges?, Australian Mathematics Trust, University of Canberra, https://web.archive.org/web/20120319074335/http://www.amt.canberra.edu.au/koenigs.html, 19 March 2012.

[3] J. J. O'Connor and E. F. Robertson (1998), Leonhard Euler, https://mathshistory.st-andrews.ac.uk/Biographies/Euler/.

[4] Jonaki Ghosh, Amber Habib and Geetha Venkataraman, *Learning Mathematics through Modelling and Simulation: An Investigative Approach*, Universities Press (India) Pvt. Ltd., December 2022.

2 2 into 4 is not equal to 8

Pranashish Chandra Rinku

P.G.T. Mathematics
D.A.V. Public School Bariatu
Ranchi, Jharkhand - 834009

If I am not mistaken, the heading of this article has taken you by surprise. But it is so; 2 into 4 is not equal to 8. Then what is it equal to? It is 2.

I have more surprises for you. Here is another: 4 by 2 is not 2 but 8.

In our country we have been routinely using the term '**into**' to denote multiplication. But this usage is incorrect! In actuality, the term 'into' ought to mean division and not multiplication! Consider the following usages. "How many times does 3 go into 12?" The answer should obviously be "4 times" because $4 = 12 \div 3$. "How many times does 4 go into 20?" The answer should obviously be "5 times" because $5 = 20 \div 4$. And so on.

Incidentally, the word used for the sign of division (÷), obelus — the Greek word for '÷', is 'goes into' or 'into' for short, in addition to 'divides' or 'how many times'.

Thus when we say "2 into 4," what it really *ought* to mean is: "How many times does 2 go into 4?" In other words, what is the value of $4 \div 2$? Obviously, the answer should be 2.

Similarly, 6 goes 5 times into 30, so 6 into 30 = 5. The answer is **not** 180.

If we just say "2 into 4," the meaning is obviously ambiguous. To some it might mean, "What is the value if 2 is divided into 4 parts?" In this case the answer would naturally be 1/2 or 0.5.

In the case of multiplication, the correct word to use is '**times**'. Thus we should say: 3 times 4 is 12, or $3 \times 4 = 12$. Similarly, 5 times 6 is 30, or $5 \times 6 = 30$. Instead of 'times' we can also say 'multiplied by.' That would be just as correct.

When we specify the dimensions of matrices too we use the word 'by', e.g.: "This is a 3 by 2 matrix."

It may be noted that the term 'by' may be used for both multiplication and division — but with proper specification: 'multiplied by' for multiplication, and 'divided by' for division. It is best that we name the operation that goes with 'by' to avoid any ambiguity.

In the case of fractions, an expression such as 3/4 may be read as '3 over 4' or '3 upon 4' or '3 divided by 4' but **not** '3 by 4'.

When I explored these facts, I was surprised that since childhood we have been using the words 'into' for multiplication and 'by' for division, whereas things are exactly opposite in meaning! This wrong application has become normalised. I urge everyone, from Mathematics teachers to English teachers and other academicians and students to ponder this and not to use the wrong term.

It may be difficult to change a practice when we have got so used to it. But when we know the truth, it would be imprudent not to take the pains to change, would it not?

Note from the editors

One of the things for which mathematicians take pride in their subject is the precise and clear use of vocabulary. It is obviously necessary to be precise and unambiguous in the use of terminology, because we then become clear in our communication and reduce the possibility of gaps and misunderstandings. Collaborative work can happen successfully only if we are clear and consistent and precise in our use of words.

Not following this dictum would lead to a situation where the meaning of a mathematical statement is dependent on who has said it, and where it has originated from. This is surely not a desirable state of affairs.

So let us take this simple suggestion seriously!

3 The Nine-point Circle of a Triangle – Part 1

Shailesh Shirali

Sahyadri School KFI
Rajgurunagar, Khed
Pune – 410513

Email: shailesh.shirali@gmail.com

The nine-point circle theorem, a statement about nine significant points of a triangle, is one of the gems of elementary Euclidean geometry. Dan Pedoe writes, "This circle is the first really exciting one to appear in any course on elementary geometry" [1]. The result is not very well known at the school level — a lacuna which certainly needs correction.

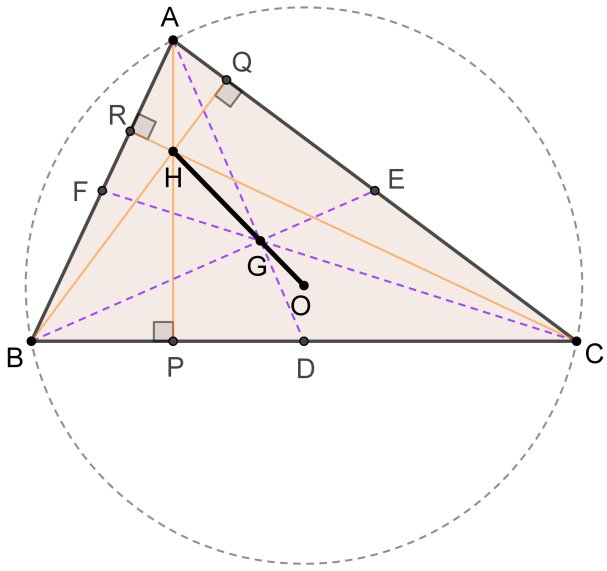


Figure 1: Line OGH is the Euler Line of $\triangle ABC$. The following is true: $\mathbf{OH} = 3\,\mathbf{OG}$.

Euler (1765; [2]) appears to have been the first to notice that for any non-equilateral

triangle $\triangle ABC$, the circumcenter O, the centroid G, and the orthocentre H lie in a straight line. Moreover, the placement of the points is such that G is the point of trisection of OH that is closer to O than to H. The line containing the three points is called the *Euler Line* of the triangle. See Figure 1.

We note in passing that the Euler line contains many other significant points of the triangle. But that is not the subject of this article, and we shall not say anything further about those points.

The nine-point circle

Theorem 1 (Nine-point circle theorem)**.** *For any triangle, the following nine points lie on a single circle: the midpoints of the three sides; the feet of the altitudes (i.e., the feet of the perpendiculars from the vertices of the triangle to the opposite sides); and the midpoints of the segments joining the vertices of the triangle to the orthocentre.*

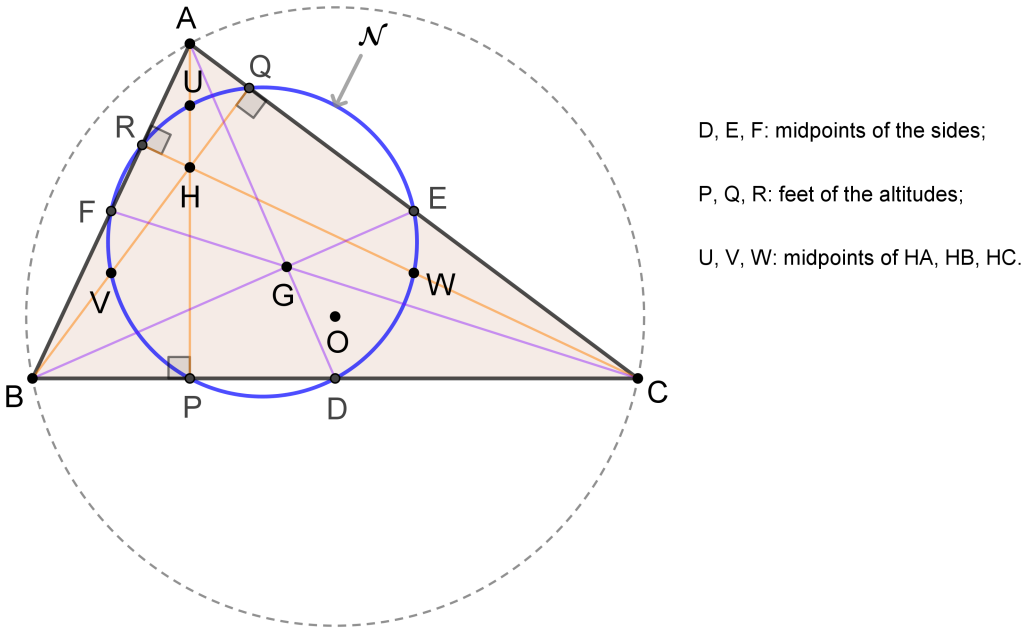


Figure 2: The Nine-Point Circle Theorem

Comments on the history of the theorem. There appear to be several mathematicians who independently noticed the existence of the nine-point circle: Leonard Euler; Karl Feuerbach; Benjamin Bevan; Jean-Victor Poncelet; Charles Brianchon; Joseph

Gergonne; Jakob Steiner; and others. It will not make much sense to try to establish priority. References [3] and [4] are highly recommended for interested readers.

Many proofs. There are many different proofs of the theorem. Perhaps all beautiful results in mathematics have this feature! In Part 1 of this article we describe a less known proof from pure geometry, followed by a proof using vector algebra. In Part 2 we establish a remarkable property of this circle first discovered by Feuerbach, and in Part 3 we describe a proof based on geometric transformations. Then we state and prove a remarkable affine generalisation of the theorem.

A proof using pure geometry

We first prove a simple result concerning quadrilaterals, first stated by Louis Brand ([7], [8]).

Theorem 2 (Brand; Eight-Point Circle Theorem)**.** *If the diagonals of a planar quadrilateral are perpendicular to each other, then the following eight points lie on a single circle: the midpoints of the four sides; and the feet of the perpendiculars dropped from the midpoints on the opposite sides of the quadrilateral. The centre of the circle is the mean centre of the vertices.*

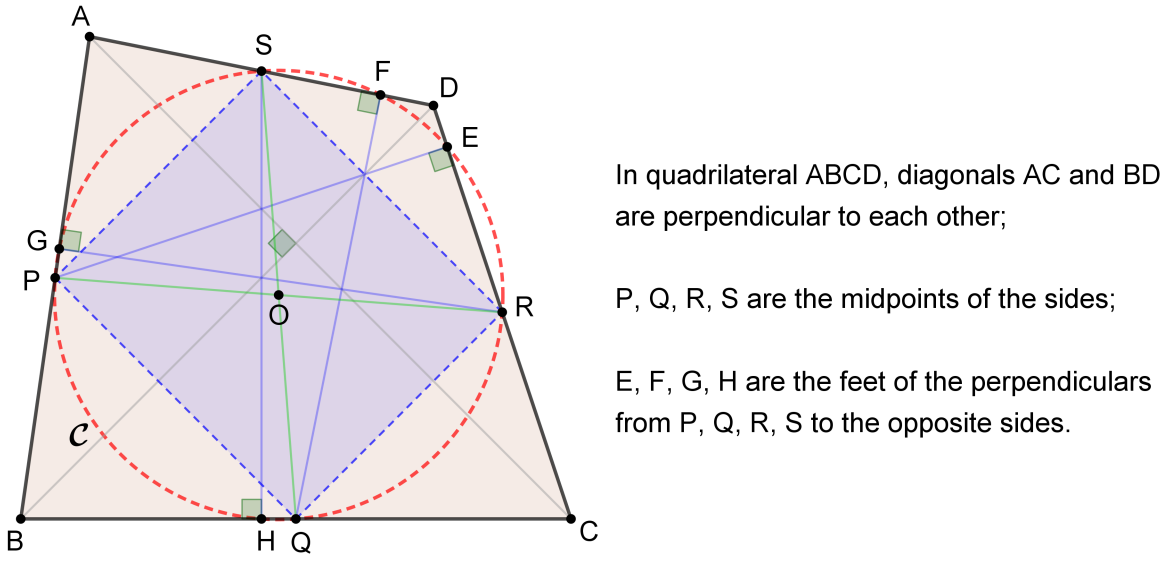


Figure 3: The Eight-Point Circle Theorem

Proof. Quadrilateral $PQRS$ is a rectangle, so there exists a circle $\mathcal{C}$ passing through its four vertices; its centre O is the common midpoint of PR and QS. As PR is a diameter of $\mathcal{C}$ and $\measuredangle PER = 90°$, it follows that E lies on $\mathcal{C}$. Similarly, points F, G, H lie on $\mathcal{C}$. The result follows. ■

Remark. We have an unusual situation here: the proof is shorter than the theorem statement!

Proof of the nine-point circle theorem. Now we show how to deduce the nine-point circle theorem from the eight-point circle theorem. We shall do so by examining three different quadrilaterals associated with the configuration of points in Figure 2.

We start by applying the theorem to quadrilateral $BCHAB$. Its diagonals are AC and BH; since $BH \perp AC$, the theorem can be applied. The midpoints of its four sides are U, W, D, F; they are the vertices of a rectangle. The feet of the perpendiculars from these points to the opposite sides are the points P, R, P, R (respectively). We infer that points U, W, D, F, P, R lie on a circle. See Figure 4.

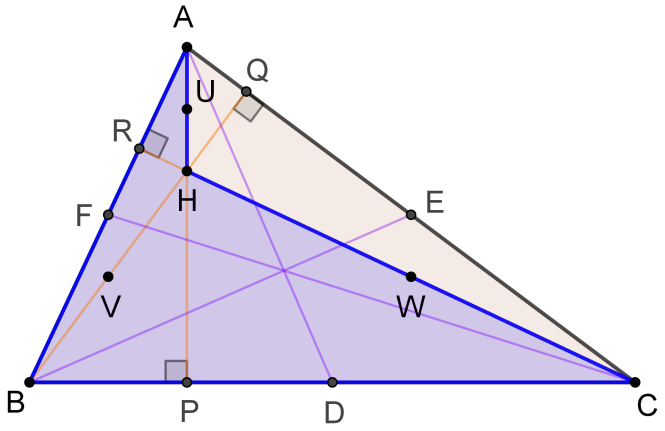


Figure 4: First application of the Eight-Point Circle Theorem

Next, we apply the theorem to quadrilateral $ABHCA$ (note that $AH \perp BC$); see Figure 5.

By reasoning in the same manner, we infer that points F, V, W, E, R, Q lie on a circle.

Lastly, we apply the theorem to quadrilateral $CAHBC$ (note that $CH \perp AB$); see Figure 6.

By reasoning yet again in the same manner, we infer that points E, U, V, D, Q, P lie on a circle.

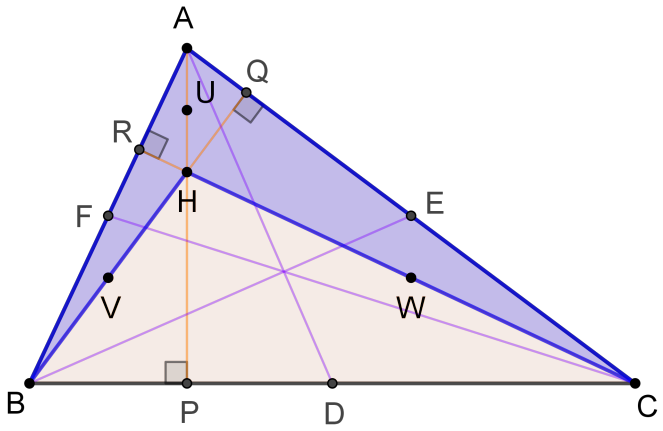


Figure 5: Second application of the Eight-Point Circle Theorem

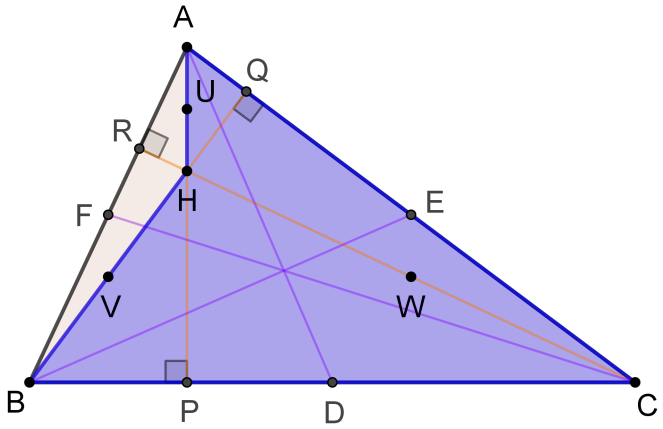


Figure 6: Third application of the Eight-Point Circle Theorem

Now consider these three circles, each passing through six of the nine points:

$$\text{Circle } (U, W, D, F, P, R), \quad \text{Circle } (F, V, W, E, R, Q), \quad \text{Circle } (E, U, V, D, Q, P).$$

These three collections of six points each make up a total of 18 points in which each point is counted twice. The union of any two collections is the full set of nine points, and each collection intersects the others in three points:

- $\{U, W, D, F, P, R\} \cap \{F, V, W, E, R, Q\} = \{F, R, W\}$;
- $\{F, V, W, E, R, Q\} \cap \{E, U, V, D, Q, P\} = \{E, Q, V\}$;
- $\{E, U, V, D, Q, P\} \cap \{U, W, D, F, P, R\} = \{D, P, U\}$.

Since there is just one circle that passes through three points not in a straight line, it

follows that the three circles are actually one circle, and so the nine points in question lie on a single circle $\mathcal{N}$. ■

Remark 1. Since points D, P, U lie on $\mathcal{N}$ and $\measuredangle DPU = 90°$, it follows that DU is a diameter of $\mathcal{N}$. Similarly, EV and FW are diameters of $\mathcal{N}$.

Remark 2. The three collections of six points display an interesting set of combinatorial properties! We are reminded of configurations in finite geometry.

A proof using vectors

We now describe a very compact proof using vector algebra.

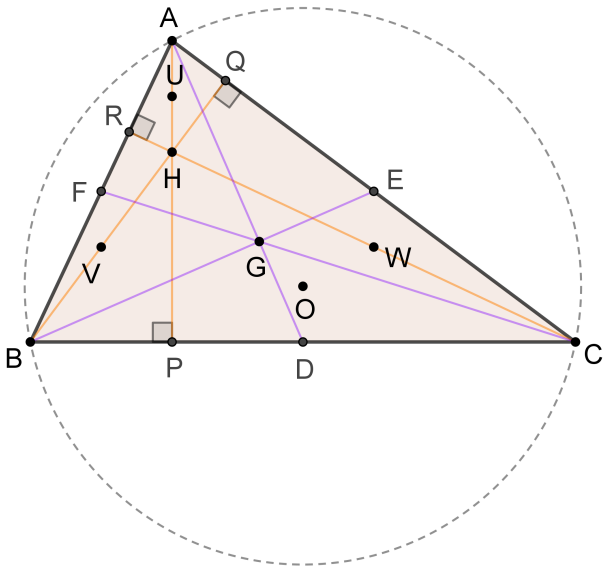


Figure 7: We must show that the nine points $D, E, F, P, Q, R, U, V, W$ lie on a circle

Let O be the point of reference of the system of vectors. We denote the position vectors of the various points in Figure 7 by the corresponding letter in bold font. For example, we denote vector $\overrightarrow{OA}$ by $\mathbf{a}$, vector $\overrightarrow{OH}$ by $\mathbf{h}$, and so on. As we have chosen the circumcenter O of $\triangle ABC$ to be the origin of the vector system, it follows that $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$ have equal length:

$$|\mathbf{a}| = |\mathbf{b}| = |\mathbf{c}|\,. \tag{1}$$

We may take this common length to be 1. Since D, E, F are midpoints of the sides, we have

$$\mathbf{d} = \frac{1}{2}(\mathbf{b}+\mathbf{c}), \qquad \mathbf{e} = \frac{1}{2}(\mathbf{c}+\mathbf{a}), \qquad \mathbf{f} = \frac{1}{2}(\mathbf{a}+\mathbf{b}). \tag{2}$$

As is well known, the position vector of the centroid G is given by

$$\mathbf{g} = \frac{1}{3}(\mathbf{a} + \mathbf{b} + \mathbf{c}). \tag{3}$$

We now show a striking result: the position vector of the orthocentre H is given by

$$\mathbf{h} = \mathbf{a} + \mathbf{b} + \mathbf{c}. \tag{4}$$

The simplest way of proving this is by showing that the point H' defined by $\mathbf{h}' = \mathbf{a}+\mathbf{b}+\mathbf{c}$ lies on all three altitudes and therefore must be the orthocentre.

We have, $\mathbf{h}' - \mathbf{a} = \mathbf{b} + \mathbf{c}$, hence

$$(\mathbf{h}' - \mathbf{a}) \cdot (\mathbf{b} - \mathbf{c}) = (\mathbf{b} + \mathbf{c}) \cdot (\mathbf{b} - \mathbf{c}) = |\mathbf{b}|^2 - |\mathbf{c}|^2 = 1 - 1 = 0, \tag{5}$$

so $AH' \perp BC$. It follows that H' lies on the altitude from vertex A to side BC. In the same way, we show that H' lies on the altitude from vertex B to side CA, and H' lies on the altitude from vertex C to side AB. Hence H' coincides with the orthocentre H, and $\mathbf{h} = \mathbf{a} + \mathbf{b} + \mathbf{c}$.

An immediate corollary of what we have just found is this:

$$\mathbf{h} = 3\,\mathbf{g}, \qquad \therefore \quad \overrightarrow{OH} = 3\,\overrightarrow{OG}. \tag{6}$$

This implies the property found by Euler: the circumcenter O, the centroid G, and the orthocentre H lie in a straight line, with $OH = 3\,OG$.

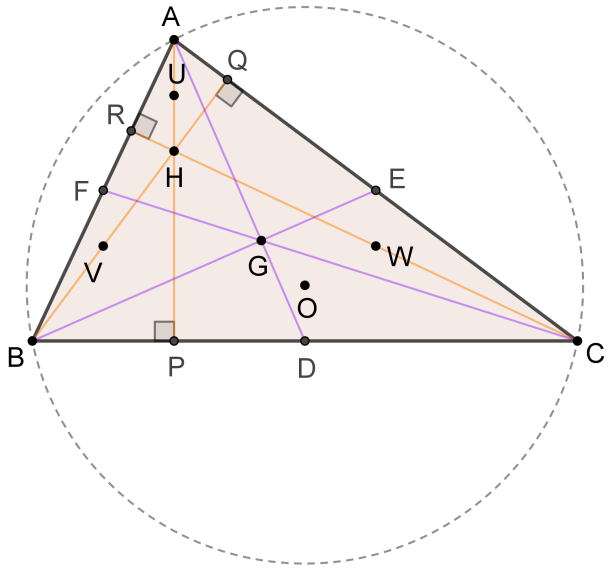


Figure 8: The figure repeated for convenience

Having obtained the position vector of H, we now have the position vectors of U, V, W as well:

$$\mathbf{u} = \mathbf{a} + \frac{1}{2}(\mathbf{b} + \mathbf{c}), \qquad \mathbf{v} = \mathbf{b} + \frac{1}{2}(\mathbf{c} + \mathbf{a}), \qquad \mathbf{w} = \mathbf{c} + \frac{1}{2}(\mathbf{a} + \mathbf{b}). \tag{7}$$

Observing these position vectors and the position vectors of the midpoints of the sides,

$$\mathbf{d} = \frac{1}{2}(\mathbf{b} + \mathbf{c}), \qquad \mathbf{e} = \frac{1}{2}(\mathbf{c} + \mathbf{a}), \qquad \mathbf{f} = \frac{1}{2}(\mathbf{a} + \mathbf{b}).$$

we are quickly led to the following observation. Let N be the midpoint of OH. The position vector of N is then

$$\mathbf{n} = \frac{1}{2}(\mathbf{a} + \mathbf{b} + \mathbf{c}), \tag{8}$$

hence we have

$$\mathbf{n} - \mathbf{d} = \frac{1}{2}\mathbf{a}, \qquad \therefore \ \ |\mathbf{n} - \mathbf{d}| = \frac{1}{2}, \tag{9}$$

i.e., ND has length $\frac{1}{2}$. We similarly find that NE and NF have length $\frac{1}{2}$. Again, we have

$$\mathbf{n} - \mathbf{u} = -\frac{1}{2}\mathbf{a}, \qquad \therefore \ \ |\mathbf{n} - \mathbf{u}| = \frac{1}{2}, \tag{10}$$

i.e., NU has length $\frac{1}{2}$. We similarly find that NV and NW have length $\frac{1}{2}$.

Finally, consider point P. Notice that N, being the midpoint of OH, is also the midpoint of UD:

$$\frac{1}{2}(\mathbf{u} + \mathbf{d}) = \frac{1}{2}\left(\mathbf{a} + \frac{1}{2}(\mathbf{b} + \mathbf{c}) + \frac{1}{2}(\mathbf{b} + \mathbf{c})\right) = \frac{1}{2}(\mathbf{a} + \mathbf{b} + \mathbf{c}) = \mathbf{n}. \tag{11}$$

Now we use elementary geometry rather than vector algebra. Consider $\triangle UPD$ which is right-angled at vertex P. The midpoint N of its hypotenuse UD is therefore equidistant from the vertices U, P, D. It follows that NP has length $\frac{1}{2}$. We similarly find that NQ and NR have length $\frac{1}{2}$.

Hence, the point N is equidistant from the nine points $D, E, F, U, V, W, P, Q, R$. These nine points therefore lie on a circle centred at N, the midpoint of OH. This establishes the existence of the nine-point circle. But we have found more: the centre N of the nine-point circle lies halfway between the circumcenter O and the orthocentre H, and its radius is half the radius of the circumcircle of $\triangle ABC$. ■

More to come ... In Parts 2 and 3, we show a remarkable tangency property of the nine-point circle and a beautiful generalisation of the result in an affine setting.

Bibliography

[1] Daniel Pedoe, *Circles: a Mathematical View*." The Mathematical Association of America.

[2] Wikipedia, "Euler line." From https://en.wikipedia.org/wiki/Euler_line

[3] Joe Wilson, "History of the Nine Point Circle." From http://jwilson.coe.uga.edu/EMT668/EMT668.Folders.F97/Anderson/geometry/geometry1project/historyofninepointcircle/history.html

[4] J. S. Mackay, "History of the Nine-point Circle," *Proceedings of the Edinburgh Mathematical Society*, (11). pages 19-61. From https://www.cambridge.org/core/journals/proceedings-of-the-edinburgh-mathematical-society/article/history-of-the-ninepoint-circle/F35C6273A2DC54CE806063838A3CB3A0

[5] Amy Edwards, Hannah Gottlieb, and Michael Kedroske, "The Nine Point Circle." From https://ninepointcircle.weebly.com/history.html

[6] Wikipedia, "Nine-point circle." From https://en.wikipedia.org/wiki/Nine-point_circle

[7] Louis Brand, "The Eight-Point Circle and the Nine-Point Circle." From https://www.jstor.org/stable/2304083

[8] Joe Wilson, "The Eight-Point Circle and the Nine-Point Circle: Mathematical History Overlooked." From http://jwilson.coe.uga.edu/EMT668/EMT668.Folders.F97/Anderson/geometry/geometry1project/eightpointcircle/eightpointcircle.html

[9] Weisstein, Eric W. "Nine-Point Circle." From MathWorld–A Wolfram Web Resource. https://mathworld.wolfram.com/Nine-PointCircle.html

[10] Michael de Villiers,"Nine Point Conic and Generalization of Euler Line." From http://dynamicmathematicslearning.com/ninepointconic.html

[11] Christopher Bradley, "The Nine-point Conic and a Pair of Parallel Lines." From https://people.bath.ac.uk/masgcs/Article119.pdf

4 Slicing a Prism Optimally: Motivating Students to Study Calculus

Jyotirmoy Sarkar

Indiana University Indianapolis
Department of Mathematical Sciences
402 N Blackford Street
Indianapolis, IN 46202-3216, USA

Email: jsarkar@iupui.edu

Abstract: A sufficiently tall right prism with a regular polygonal base will be sliced by a plane which is the perpendicular-bisector of KT, where K is either a midpoint of a side of the base or a vertex and T is on the diametrically opposite vertical line on the prism. How should you choose T: (1) to minimize the area of the cross-sectional cut; (2) to minimize the volume of the part of the prism under the cut and containing K; and (3) to maximize the volume of the part of the prism below the reflection of the bottom base about the first cutting plane?

Before solving these 3D problems, we solve the (relatively easier) corresponding problem of paper folding in 2D. Our aim is to motivate students to study calculus to solve optimization problems.

A Confession

Will you be my confidante if I confess to you a misdemeanor I committed? The fold mark on pages 87–88 of *Harry Potter and the Deathly Hallows*, owned by the Marion County Public Library, is my misdeed. I was so engrossed in the story that when my wife called me to dinner, I quickly folded the page instead of fetching a bookmark. Soon after dinner, I returned to the book. But the damage was already done and I could not undo it.

Today, twelve years later, as penance I am writing a mathematics paper on paper folding. I fondly hope the public library will not only forgive me, but also save a copy of this paper next to their Harry Potter collection.

1 Optimize While You Fold a Paper

A standard A4 sheet of paper $ABCD$ with $AB = CD = 210$ mm and $BC = AD = 297$ mm will be folded along a segment XY, where X is on AB and Y is on AD, so that the vertex A will fall on a point T on BC. See Figure 1. How will you choose T so that either (i) the length of XY is minimum; or (ii) the area of $\triangle AXY$ is minimum; or (iii) the area of $\triangle BXT$ is maximum?

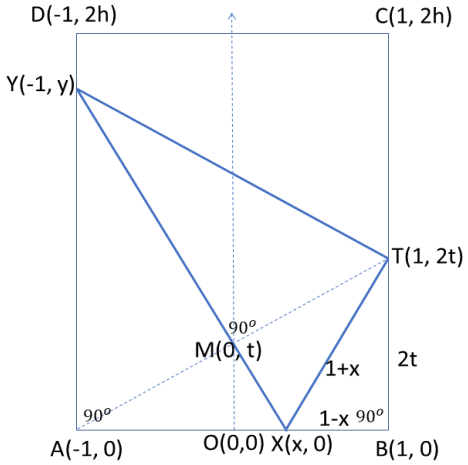


Figure 1: Fold the paper so that vertex A falls on a point T on edge BC, making a crease along XY with X on AB and Y on AD, and either (1) minimize the fold length XY, or (2) minimize the area of $\triangle AXY$, or (3) maximize the area of $\triangle BXT$.

We found Problem (i) in [5], as Exercise 3.7.73. The solution is worked out in [2]. For a dynamic demonstration, see [8]. Problem (ii) we found on the internet Q&A site [4]. Problem (iii) appears in [10]. For other aspects of paper folding see [9].

Astute readers will note that the actual measures of the paper are unimportant since the units can be changed. Also, while the given paper has aspect ratio $297/210 = 1.41428 \approx \sqrt{2}$, the problem can be posed for other rectangles with different aspect ratios. For example, in the United States a letter size paper has dimensions 11 in. by 8.5 in., with an aspect ratio 1.2941. Hence, without loss of generality, let the width of the paper be $AB = CD = 2$ and the height $AD = BC = 2h$, where h is sufficiently larger than 1. For

now, we assume $h > 1.25$; after finding the optimal solutions, we will know how small h can be.

Let $BT = 2t$. We shall express the lengths of various segments and areas of various regions in terms of t. Impose a co-ordinate system with the midpoint $O(0,0)$ of AB as origin and OB as the horizontal axis. Using the Pythagorean theorem in right $\triangle ABT$, the hypotenuse AT has length $2\sqrt{1+t^2}$, midpoint $M = (0,t)$ and slope t. The equation of XY, the perpendicular bisector of AT, is $y - t = (-1/t)\,x$, or

$$x + ty = t^2. \tag{1}$$

In Equation (1), if we put $y = 0$, we get $OX = x = t^2$, implying that $AX = 1+t^2$. If we put $x = 0$, we get $OM = t$, and if we put $x = -1$, we get $AY = y = (1+t^2)/t = t + 1/t$. To ensure that X is on AB and Y is on AD, we must have $1 + t^2 \leq 2$ and $t + 1/t \leq 2h$. Hence, we must choose t in the interval $\left[h - \sqrt{h^2-1}, 1\right]$. Lastly, the length of the fold is $XY = \sqrt{(1+x)^2 + y^2} = (1+t^2)\sqrt{1+1/t^2} = (1+t^2)^{3/2}/t$, the area of right $\triangle AXY$ is $(1+t^2)^2/(2t)$, and the area of right $\triangle BXT$ is $t\,(1-t^2)$.

To optimize (minimize or maximize) an objective function, it suffices to optimize a monotonic transformation: A monotonically increasing function should be *similarly* optimized; a monotonically decreasing function must be *oppositely* optimized.

Since $y = x^{2/3}$ is monotonically increasing, the length of XY is minimized at

$$\arg\min\ (1+t^2)/t^{2/3} = \arg\min\ (t^{-2/3} + t^{4/3}). \tag{2}$$

Equating the first derivative of the last objective function in (2) with respect to t to zero, we get $(-2/3)\,t^{-5/3} + (4/3)\,t^{1/3} = 0$, or $t^2 = 1/2$. Also, we verify that the second derivative is $(10/9)\,t^{-8/3} + (4/9)\,t^{-2/3} > 0$. Therefore, the objective function is convex from below. Hence, $t = \sqrt{1/2} \approx 0.7071$ minimizes XY, with minimum value $3\sqrt{3}/2 = 2.598076$. The reader may see [2] for an alternative derivation.

Likewise, since $y = x^{1/2}$ is monotonically increasing, area$(\triangle AXY)$ is minimized at

$$\arg\min\ (1+t^2)/t^{1/2} = \arg\min\ (t^{-1/2} + t^{3/2}). \tag{3}$$

Equating the first derivative of the last objective function in (3) to zero, we obtain $(-1/2)\,t^{-3/2} + (3/2)t^{1/2} = 0$, or $t^2 = 1/3$. Also, we verify that the second derivative is $(3/4)\,(t^{-5/2} + t^{-1/2}) > 0$. Hence, $t = \sqrt{1/3} \approx 0.57735$ minimizes area of $\triangle AXY$, with minimum value $8\sqrt{3}/9 = 1.539601$.

Finally, the area of $\triangle BXT$ is maximized at

$$\arg\max\ t\,(1-t^2) = \arg\max\ (t - t^3). \tag{4}$$

Equating the first derivative of the objective function in (4) to zero, we obtain $1 - 3t^2 = 0$, or $t^2 = 1/3$. Also, we verify that the second derivative $-6t < 0$. Hence, $t = \sqrt{1/3}$ maximizes area of $\triangle BXT$, with maximum value $2\sqrt{3}/9 = 0.384900$.

1.1 Commentaries on the Optimal Solutions

What minimum aspect ratio of the paper must we assume? The optimal solutions to Problems (i)–(iii) are attained at $t = \sqrt{1/2}, \sqrt{1/3}, \sqrt{1/3}$, respectively. Thus, $AX = 1 + t^2 < 2$ ensures that X falls on AB. The corresponding values of AY are $3/\sqrt{2}, 4/\sqrt{3}, 4/\sqrt{3}$, respectively. To ensure that Y will fall on AD for all three objectives, we must assume that $AY \leq 2h$, or $h \geq 2/\sqrt{3} = 1.154701$.

That the optimal solutions to Problems (ii) and (iii) are the same is a pleasant surprise: Different desirable objectives are attained at the same argument! More surprise is in store for you. Substituting the solution $t = \sqrt{1/3}$ into the various lengths, we see that $AT = 4/\sqrt{3} = AY = TY$. That is, $\triangle ATY$ is equilateral. Hence, we can find T geometrically: Simply draw $\measuredangle DAT = 60^o$.

Figure 2 shows the measures of all line segments and all angles when $t = \sqrt{1/3}$. Excluding $\triangle ATX$ and $\triangle ATY$, all other triangles have $(30^o, 60^o, 90^o)$ angles; hence, they are similar and each hypotenuse is twice as long as the shorter leg. Two such congruent triangles (for example, $\triangle AXY$ and $\triangle TXY$, or any two among $\triangle AXM$, $\triangle TXM$, $\triangle TXB$) can be joined along the longer legs to form an equilateral triangle.

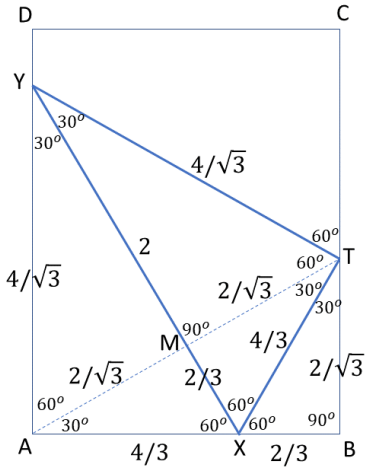


Figure 2: The optimal solution to problems (ii) and (iii) generates an equilateral $\triangle ATY$ and several $(30^o, 60^o, 90^o)$-triangles.

Moreover, $\triangle ATY$ is the largest connected equilateral triangle that can be cut out of the given rectangular sheet of paper $ABCD$, provided $AD = h \geq 2/\sqrt{3}$. But if $1 \leq h < 2/\sqrt{3}$, then the largest connected equilateral triangle inscribed in the rectangular sheet has side length AD.

2 Geometric Solutions

To emphasize the power of calculus in solving optimization problems, we also present the geometric solutions of Problems (i)–(iii) — but in reverse order (you will soon realize why). Calculus permits routine steps: first differentiate (a monotonic transformation of) the objective function, then solve the first-order condition and finally verify the second-order condition. In contrast, our geometric solutions to the three problems require specific ingenuity such as 3D visualization, trigonometric transformation, and both. Interested readers are invited to construct alternative proofs.

2.1 A geometric solution to Problem (iii)

The area of right $\triangle BXT$ is $t(1-t^2)$. Figure 3 shows the objective function $t(1-t^2)$ as the volume of a rectangular cuboid with sides $t,\ 1-t,\ 1+t$, respectively. The function achieves the value 0 at $t=0$ and at $t=1$, and it is positive on interval $(0,1)$. Therefore, it is maximized at some argument in the interval $(0,1)$. We will find the argmax by equating the areas of some combinations of boundary faces of the rectangular cuboid.

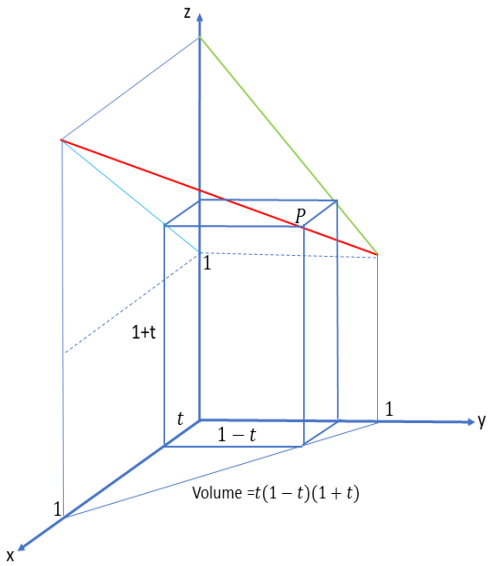


Figure 3: The objective function $t-t^3$ is the volume of a right rectangular cuboid with sides $t,\ 1-t,\ 1+t$, respectively.

As t increases to $t+\Delta$, for a tiny $\Delta>0$, the volume of the rectangular cuboid changes along three faces: Along the front face (parallel to the yz-face), it increases by $\Delta\,(1-t-\Delta)(1+t+\Delta)$; along the top face (parallel to the xy-face), it increases by $\Delta\,t(1-t)$; and along the right face (parallel to the xz-face), it decreases by $\Delta\,t(1+t+\Delta)$. (I

contemplated showing you a movie. But I decided it is more beneficial that you should imagine it.) By eliminating the common multiple Δ and then letting $\Delta \to 0$, the volume is maximized when the total area of the front face and the top face (faces that increase in area) equals the area of the right face (which decreases in area); or when

$$(1-t)(1+t) + t(1-t) = t(1+t). \tag{5}$$

Algebraic simplification of (5) yields $3t^2 = 1$, or $t = \sqrt{1/3}$. Hence, the maximum area of $\triangle BXT$ is $t(1-t^2) = 2\sqrt{3}/9$ with sides $BX = 2/3, BT = 2/\sqrt{3}, XT = 4/3$.

Remark 1. Here is an algebraic solution to Problem (i) using the AM-GM inequality (the arithmetic mean is at least as large as the geometric mean) to the set of three numbers $\{t^2, (1-t^2)/2, (1-t^2)/2\}$. Accordingly, $t^2(1-t^2)^2/4 \le (1/3)^3$; or taking square roots, we have $t(1-t^2) \le 2/(3\sqrt{3})$, with equality if and only if $t^2 = (1-t^2)/2$, or $t = \sqrt{1/3}$. Hence, the maximum area of $\triangle BXT$ is $t(1-t^2) = 2\sqrt{3}/9$.

For more examples of this sort, see [3].

2.2 A geometric solution to Problem (ii)

The area of $\triangle AXY$ is $(1+t^2)^2/(2t)$. Letting $t = \tan\theta$, we have $1+t^2 = 1/\cos^2\theta$. Taking reciprocal, which is a monotonically decreasing function, and using trigonometric double angle formulas, (one-fourth) the area of $\triangle AXY$ is minimized at

$$\begin{aligned} \arg\max_{0\le\theta\le\pi/4} 8\sin\theta\ \cos^3\theta &= \arg\max_{0\le\theta\le\pi/4} 2\sin 2\theta\ (1+\cos 2\theta) \\ &= \arg\max_{0\le\theta\le\pi/4} (2\sin 2\theta + \sin 4\theta). \end{aligned}$$

The last objective function is depicted in Figure 4 as the total area of three disjoint triangles OBP, POQ, QOA within the upper semicircle of a unit circle. We must choose $2\theta \in [0, \pi/2]$ to maximize this total area.

Given diameter AB of the unit circle, we must choose the point P on the semicircle (and the point Q on the semicircle will be immediately determined so that $\measuredangle BOP = \measuredangle POQ$) to maximize the total area of triangles OBP, POQ, QOA. Even if P and Q are allowed to be completely unrelated to each other, the total area is maximum when Q bisects arc PA and P bisects arc BQ; or when arcs BP, PQ, QA are equal; or when $\pi - 4\theta = 2\theta$; or $\theta = \pi/6$; or $t = \tan\theta = \sqrt{1/3}$. Hence, the minimum area of $\triangle AXY$ is $(1+t^2)^2/(2t) = 8\sqrt{3}/9$ with its sides $AX = 4/3, AY = 4/\sqrt{3}, XY = 8/3$.

Remark 2. Here is an algebraic solution to Problem (ii) using the AM-GM inequality. Substituting $u = 1/(1+t^2)$, minimizing the area $(1+t^2)^2/(2t)$ of $\triangle AXY$ is equivalent

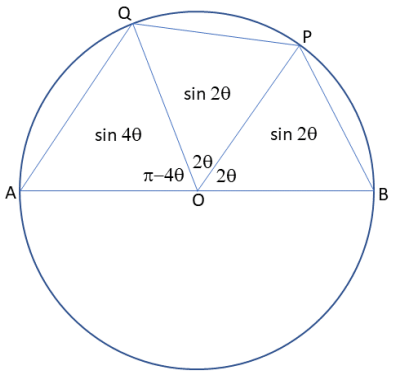


Figure 4: The objective function $2\sin 2\theta + \sin 4\theta$ is the total area of the three triangles OBP, POQ, QOA within the unit semi-circle with areas $\sin 2\theta, \sin 2\theta, \sin 4\theta$, respectively.

to maximizing the square of its inverse $4u^3(1-u)$. However, applying the AM-GM inequality to four numbers $\{u/3, u/3, u/3, 1-u\}$, we have $u^3(1-u)/27 \leq (2/4)^4$; or, $u^3(1-u) \leq 27/16$, with equality if and only if $u/3 = 1-u$, or $u = 3/4$, or $t = \sqrt{1/u - 1} = \sqrt{1/3}$. The minimum area of $\triangle AXY$ is $(1+t^2)^2/(2t) = 8\sqrt{3}/9$.

2.3 A geometric solution to Problem (i)

The length of XY is $(1+t^2)^{3/2}/t$. Substituting $t = \tan\theta$ and taking reciprocal, the length of XY is minimized at

$$\begin{aligned}\arg\max_{0\leq\theta\leq\pi/4} \sin\theta\ \cos^2\theta &= \arg\max_{0\leq\theta\leq\pi/4} \sin\theta\ (1-\sin^2\theta)\\ &= \arg\max_{0\leq s\leq\sqrt{1/2}} s\,(1-s^2),\ \text{where } s = \sin\theta.\end{aligned}$$

The last equality holds because the sine function is monotonically increasing on $[0, \pi/4]$. However, in subsection 2.1, we have seen that $s(1-s^2)$ is maximized at $s = \sqrt{1/3}$, which is in $(0, \sqrt{1/2})$. Hence, the optimal solution is at $\sin\theta = \sqrt{1/3}$, or at $t = \tan\theta = \sqrt{1/2}$. Hence, the minimum value of XY is $(1+t^2)^{3/2}/t = 3\sqrt{3}/2$.

3 Slicing a Right Prism

Our world being three dimensional, any problem we solve in two dimensions, naturally inspires us to pose related problems in three dimensions. Below we extend the problem of folding a rectangular paper to the problem of slicing a right prism. Preserving symmetry,

we shall restrict attention to regular polygonal bases with $n = 4, 3, 6$ sides (Why this strange order?) together with a bisector that joins a midpoint of one side or a vertex to the opposite vertex or the midpoint of the opposite side. Such a bisector splits the base into two symmetric halves. Since a right prism is generated when a base is translated orthogonally to itself, any line segment in the base generates a rectangle. Here, our chosen bisector generates a rectangle which splits the prism symmetrically and plays the role of the sheet of paper studied in the previous sections. In particular, without loss of generality, we assume that the base bisector is of length 2 and the prism has sufficiently large height $h \geq (t + 1/t)/2$, where t is the optimal solution yet to be discovered.

In this section, we discuss three special prisms: when $n = 4$, the base is a square (S); when $n = 3$, the base is an equilateral triangle (T); when $n = 6$, the base is a regular hexagon (H). We invite interested readers to extend the optimal slicing problems to other right prisms with different bases: for example, regular polygons with $n = 5, 7, 8$ sides, and arbitrary polygonal bases with arbitrary number of sides.

3.1 Slicing a Right Prism with a Square Base

A right prism with a square base $ABCD$ will be sliced by a plane which is (Sa) the perpendicular-bisector of MT, where M is the midpoint of AD and T is on the vertical line through the midpoint N of BC; or (Sb) the perpendicular-bisector of AT, where T is on the vertical edge through C. How should you choose the vertical height of T: (1) to minimize the area of the cross-sectional cut; (2) to minimize the volume of the part of the prism below the cutting plane and containing A; and (3) to maximize the volume of the part of the prism below a new plane which is the reflection of the bottom base about the first cutting plane and containing C?

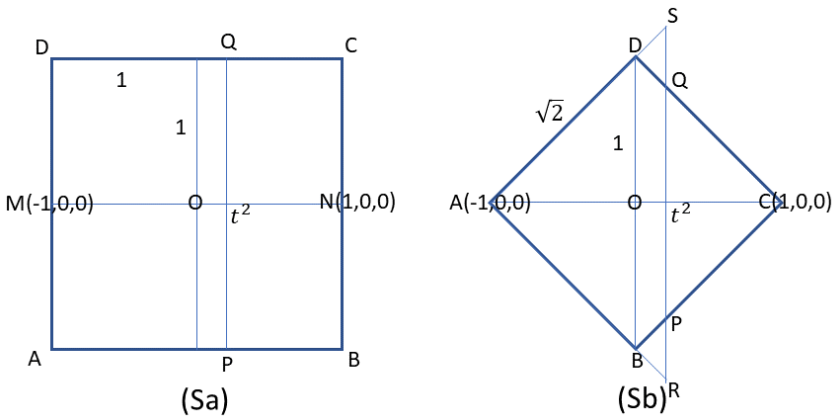


Figure 5: The square base of a right prism together with a bisector (Sa) MN, or (Sb) AC, which symmetrically bisects the base

Problem (Sa) of this subsection is equivalent to the paper folding problem of the previous section, except that the paper is endowed with a constant thickness of 2, the same as its width! We impose a coordinate system with the midpoint O of MN as origin, ON as x-axis, square base $ABCD$ of the prism in the xy-plane, and the height $2h$ of the prism in the z-direction sufficiently large. Then

$$M(-1,0,0), N(1,0,0), A(-1,-1,0), B(1,-1,0), C(1,1,0), D(-1,1,0).$$

The solutions to the three parts, with ingredients taken from Figure 1, are as follow: (Sa1) the cross-sectional cut attains a minimum area of $2 \cdot XY = 3\sqrt{3}$ when $t = \sqrt{1/2}$; (Sa2) the volume of the part of the square-based prism below the cut attains a minimum of $2 \cdot \triangle AXY = 16\sqrt{3}/9$ when $t = \sqrt{1/3}$; and (Sa3) the volume of the part of the square-based prism below the reflection of the base on the first cut attains a maximum of $2 \cdot \triangle BXT = 4\sqrt{3}/9$ when $t = \sqrt{1/3}$.

For Problem (Sb), where the bisector of the base joins two opposite vertices A and C, we impose a coordinate system with the midpoint O of AC as origin, OC as x-axis, OD as y-axis and the vertical height of the prism as z-axis. Then

$$A(-1,0,0), B(0,-1,0), C(1,0,0), D(0,1,0).$$

(Sb1) The area of the part of the base $\{|x| + |y| \leq 1, z = 0\}$ to the left of the line $PQ = \{x = t^2, z = 0\}$ is $2 - (1-t^2)^2 = 1 + 2t^2 - t^4$. Hence, the cross-sectional area of the first cut is $(1 + 2t^2 - t^4)\sqrt{1 + t^{-2}}$. To minimize the area, it suffices to minimize its square

$$(1 + 2t^2 - t^4)^2(1 + t^{-2}) = t^{-2} + 5 + 6t^2 - 2t^4 - 3t^6 + t^8.$$

The derivative of the squared area with respect to t is

$$\begin{aligned} -2t^{-3} + 12t - 8t^3 - 18t^5 + 8t^7 &= -2t^{-3}[1 - 6t^4 + 4t^6 + 9t^8 - 4t^{10}] \\ &= -2t^{-3}[(1 - 3t^4)^2 + 4t^6(1 - t^4)], \end{aligned}$$

which is negative for all $0 < t \leq 1$. Hence, the minimum area of cross-section is attained at $t = 1$. We call this a corner solution since the minimum occurs at an endpoint of the range of permissible values of the argument. The minimum area of the cutting plane is $2\sqrt{2}$ when the cross-section is a rhombus with side length $\sqrt{3}$ and two diagonals of lengths 2 and $2\sqrt{2}$, respectively.

(Sb2) If the first cutting plane intersects AC at $(t^2, 0, 0)$, then the part of the prism below the cutting plane and containing A is the remaining solid when from a pyramid (on a right triangular base ARS and apex Y) are cut off two smaller pyramids (on two congruent triangular bases BRP and DSQ). This part of the prism has volume

$[(1+t^2)^3-2(t^2)^3]/(3t)$. To minimize this volume, it suffices to minimize $t^{-1}+3t+3t^3-t^5$. Equating the derivative to zero, we obtain

$$-t^{-2}+3+9t^2-5t^4=0, \text{ or}$$
$$-t^{-2}[1-3t^2-9t^4+5t^6]=0.$$

Do not panic: We do not have to solve a sixth-degree polynomial; we are only dealing with a cubic equation in t^2. See [1] and [6] for how to solve cubic equations. Admittedly, the cubic formula is much less known than the quadratic formula. If you have not already seen it before, now is a good time to learn it. Applying the cubic formula, the solution to $1-3t^2-9t^4+5t^6=0$ is $t^2=0.2131$, or $t=0.461692$. Therefore, the minimum volume under the first cut is 1.2751.

(Sb3) The part of the prism below the reflection of the base about the first cutting plane and containing C is a pyramid (on a right triangular base QCR and height $2t$). Hence, this part of the prism has volume $(1-t^2)^2\cdot 2t/3$. To maximize this volume, it suffices to maximize $(1-t^2)\sqrt{t}=t^{1/2}-t^{5/2}$. Equating the derivative to zero and solving, we get $t^2=1/5=0.2$, or $t=0.4472$. The maximum volume of pyramid $QCRT$ is 0.1908.

3.2 Slicing a Right Prism with an Equilateral Triangular Base

A right prism with an equilateral triangular base ABC will be sliced by a plane which is (Ta) the perpendicular-bisector of MT, where M is the midpoint of AC and T is on the opposite vertical edge through B; or (Tb) the perpendicular-bisector of AT, where T is on the opposite vertical line through the midpoint of N of BC. How should you choose T: (1) to minimize the area of the cross-sectional cut; (2) to minimize the volume of the part of the prism below the cutting plane and containing A; and (3) to maximize the volume of the part of the prism below the reflection of the bottom base about the first cutting plane and containing B?

For Problem (Ta), we impose a coordinate system with the midpoint O of MB as origin, OB as x-axis, etc. Then

$$M(-1,0,0), A(-1,-2/\sqrt{3},0), B(1,0,0), C(-1,2/\sqrt{3},0).$$

(Ta1) The area of the part of the base to the left of the line $PQ=\{x=t^2, z=0\}$ is the difference between the areas of two equilateral triangles ABC and PBQ, or $[2^2-(1-t^2)^2]/\sqrt{3}=(3-t^2)(1+t^2)/\sqrt{3}$. Hence, the cross-sectional area of the first cutting plane is $(3-t^2)(1+t^2)^{3/2}/(\sqrt{3}t)$. To minimize this area, it suffices to minimize three times its square, or $(3t^{-1}-t)^2(1+t^2)^3$, whose derivative is

$$\begin{aligned}&2(3t^{-1}-t)[(-3t^{-2}-1)(1+t^2)+3t(3t^{-1}-t)](1+t^2)^2\\&=-2(3t^{-1}-t)(1+t^2)^2t^{-2}[3-5t^2+4t^4]\\&=-2(3t^{-1}-t)(1+t^2)^2t^{-2}[(5/4-2t^2)^2+3-25/16],\end{aligned}$$

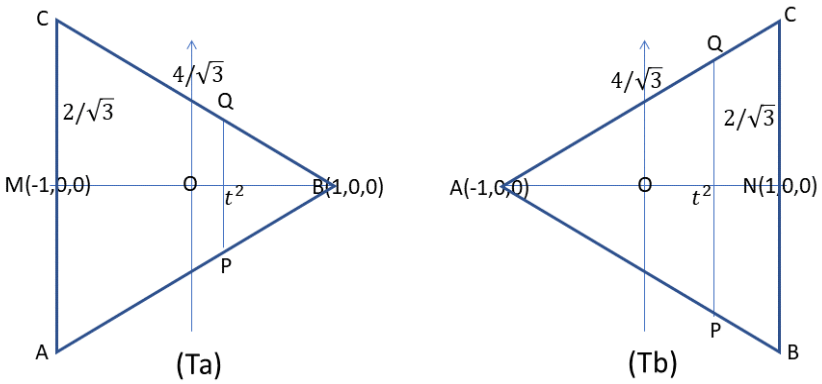


Figure 6: The equilateral triangular base of a right prism with a bisector (Ta) MB, or (Tb) AN, which symmetrically bisects the base

which is negative for all $0 < t \leq 1$. Hence, the minimum area of cross-section is attained at $t = 1$, a corner solution. The minimum area is $4\sqrt{2/3}$ when the cross-section is an isosceles triangle with one side of length $4/\sqrt{3}$, the altitude on that side of length $2\sqrt{2}$, and the other two equal sides of length $2\sqrt{7/3}$.

(Ta2) If the first cutting plane intersects MB at $(t^2, 0, 0)$, then the part of the prism below the cut and containing A is a solid that can be sliced into thin rectangles with cuts parallel to the yz-plane at x-coordinates in $(-1, t^2)$. Integrating the areas of these slices, this part of the prism has volume

$$\begin{aligned}\int_{-1}^{t^2} \left(\frac{t^2 - x}{t}\right) \frac{2(1-x)}{\sqrt{3}}\, dx &= \frac{2}{\sqrt{3}t} \int_{-1}^{t^2} [t^2 - (1+t^2)x + x^2]\, dx \\ &= \frac{2}{\sqrt{3}t}\left[t^2(t^2+1) - (1+t^2)\frac{t^4-1}{2} + \frac{t^6+1}{3}\right] \\ &= [5t^{-1} + 9t + 3t^3 - t^5]/(3\sqrt{3})\end{aligned}$$

To minimize this volume, we equate the derivative to zero, or

$$\begin{aligned}-5t^{-2} + 9 + 9t^2 - 5t^4 = 0, \text{ or} \\ -(1+t^{-2})[5 - 14t^2 + 5t^4] = 0.\end{aligned}$$

Since $(1 + t^{-2}) > 0$ for all t, equating the other quadratic factor to zero and solving, see [7], we get $t^2 = (14 - \sqrt{96})/10 = 0.4202$, or $t = 0.648232$. Also, the second derivative is $10t^{-3}(1 - 2t^6) + 18t > 0$. Hence, the minimum volume is 2.74243.

(Ta3) The part of the prism below the reflection of the base about the first cutting plane is a pyramid (on base equilateral $\triangle PBQ$ and height $2t$) with volume

$$\frac{1}{3} \cdot \frac{(1-t^2)^2}{\sqrt{3}} \cdot 2t = \frac{2}{3\sqrt{3}}(1-t^2)^2\, t.$$

To maximize this volume, it suffices to maximize $(1-t^2)\sqrt{t} = t^{1/2} - t^{5/2}$. Equating the derivative to zero, we get $0 = (1/2)t^{-1/2} - (5/2)t^{3/2}$, solving which we get $t^2 = 1/5 = 0.2$, or $t = 0.4472$. Also, the second derivative is $(-t^{-3/2} - 15t^{1/2})/4 < 0$. Hence, the maximum volume of pyramid $PBQT$ is 0.110165.

For Problem (Tb), choose the origin at the midpoint of AN, and let

$$A(-1,0,0), B(1,-2/\sqrt{3},0), C(1,2/\sqrt{3},0), N(1,0,0).$$

(Tb1) The part of the base to the left of the line $PQ = \{x = t^2, z = 0\}$ is an equilateral triangle APQ with side length $2(1+t^2)/\sqrt{3}$. Hence, the cross-sectional area of the first cut is

$$\frac{(1+t^2)^2}{\sqrt{3}} \cdot \frac{\sqrt{1+t^2}}{t} = \frac{(1+t^2)^{5/2}}{\sqrt{3}t}.$$

To minimize the area, it suffices to minimize (any multiple of) its 2/5-th power, or $(1+t^2)t^{-2/5} = t^{-2/5} + t^{8/5}$. Equating the derivative to 0, we have $0 = -(2/5)t^{-7/5} + (8/5)t^{3/5}$, whence $t^2 = 1/4$, or $t = 1/2$. Also, one can check that the second derivative is $(14/25)t^{-12/5} + (24/25)t^{-2/5} > 0$. Hence, the minimum area of the cross-sectional cut is $(25/16)\sqrt{5/3} = 2.017179$.

(Tb2) If the first cutting plane intersects AN at $(t^2, 0, 0)$, then the part of the prism below the first cutting plane and containing A is a pyramid (on an equilateral triangular base APQ of side length $2(1+t^2)/\sqrt{3}$; hence area $(1+t^2)^2/\sqrt{3}$) and height $(1+t^2)/t$. Hence, this part of the prism has volume $(1+t^2)^3/(3\sqrt{3}t)$. To minimize this volume, it suffices to minimize $(1+t^2)t^{-1/3} = t^{-1/3} + t^{5/3}$. Equating the derivative to zero, we have $-(1/3)t^{-4/3} + (5/3)t^{2/3} = 0$, whence $t^2 = 1/5 = 0.2$, or $t = 0.4472$. Also check that the second derivative is $(4/9)t^{-7/3} + (10/9)t^{-1/3} > 0$. Hence, the minimum volume of pyramid $APQY$ is $(72/125)\sqrt{5/3} = 0.743613$.

(Tb3) The part of the prism below the reflection of the base about the first cutting plane and containing N is a solid that can be sliced into thin rectangles with cuts parallel to the yz-plane at x-coordinate in $(t^2, 1)$. Hence, it has volume

$$\begin{aligned}\int_{t^2}^1 \frac{(x-t^2)2t}{1-t^2}\,\frac{2(1+x)}{\sqrt{3}}\,dx &= \frac{4t}{\sqrt{3}(1-t^2)}\int_{t^2}^1 [-t^2 + (1-t^2)x + x^2]\,dx \\ &= \frac{4t}{\sqrt{3}}\left[-t^2 + \frac{1-t^4}{2} + \frac{1+t^2+t^4}{3}\right] \\ &= 2t[5-4t^2-t^4]/(3\sqrt{3}).\end{aligned}$$

To maximize this volume, it suffices to maximize $5t - 4t^3 - t^5$. So, we equate its derivative to zero to obtain $5 - 12t^2 - 5t^4 = 0$, solving which, we get $t^2 = (-12 + \sqrt{244})/10 = 0.36205$, or $t = 0.601706$. Also check that the second derivative is $-24t - 20t^3 < 0$. Hence, the maximum volume is 0.7922275.

3.3 Slicing a Right Prism with a Regular Hexagonal Base

A right prism with a regular hexagonal base $ABCDEF$ will be sliced by a plane which is either (Ha) the perpendicular-bisector of MT, where M is the midpoint of AF and T is on the vertical line through the midpoint N of CD; or (Hb) the perpendicular-bisector of AT, where T is on the vertical edge through D. How should you choose T so that: (1) you minimize the area of the cross-sectional cut; (2) you minimize the volume of the part of the prism below the cutting plane and containing A; and (3) you maximize the volume of the part of the prism below a new plane which is the reflection of the bottom base about the first cutting plane and containing D?

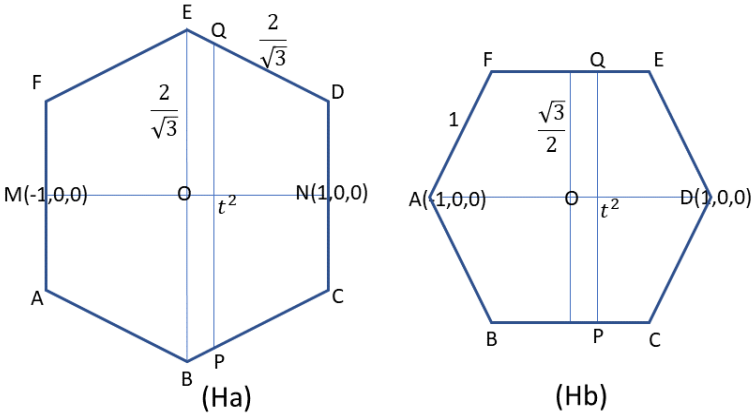


Figure 7: The regular hexagonal base of a right prism with a bisector (Ha) MN, or (Hb) AD, which symmetrically bisects the base

Avoiding monotonous repetitions, we only document the objective functions, optimal values and optimal solutions to problems (1)–(3) of versions (Ha) and (Hb):

(Ha1) The first cutting plane has area $(3+4t^2-t^4)\sqrt{1+t^{-2}}/\sqrt{3}$, which attains a minimum value of 4.790292 at $t^2 = 0.7363$, or $t = 0.858038$.

(Ha2) The part of the prism under the first cutting plane and containing A has volume $(4/t + 9t + 18t^3 - 5t^5)/(3\sqrt{3})$, which attains a minimum value of 2.787458 at $t^2 = 0.2085932$, or $t = 0.45672$.

(Ha3) The part of the prism under the reflection of the base about the first cutting plane has volume $2(4t - 5t^3 + t^5)/(3\sqrt{3})$, which attains a maximum value of 0.5354766 at $t^2 = 0.3698 = (15 - \sqrt{145})/8$, or $t = 0.608112$.

(Hb1) The first cutting plane has area $\sqrt{3}(3/4 + t^2)\sqrt{1+t^2}/t$, which attains a minimum value of 3.725936 at $t^2 = 0.4114 = (\sqrt{112} - 4)/16$, or $t = 0.641434$.

(Hb2) The part of the prism under the first cutting plane and containing A has volume $\sqrt{3}(7/t + 18t + 12t^3)/24$, which attains a minimum value of 1.767889 at $t^2 = (\sqrt{1332} - 18)/72 = 0.2569$, or $t = 0.506850$.

(Hb3) The part of the prism under the reflection of the base about the first cutting plane has volume $2t(1-t^2)^2/\sqrt{3}$, which attains a maximum value of 0.304119 at $t = 1/3$.

3.4 Optimal solutions to various prisms

We list in a table the optimal values of t derived in the previous subsections for various prisms and different objectives.

Table 4.1: Optimal values of t for various prisms and different objectives

Regular polygonal base shape	Left end of base bisector at	Minimum area of the first cut	Minimum volume under the first cut	Maximum volume under the reflection
triangle	(Ta) mid-side	1	0.6482	$\sqrt{1/5}$
	(Tb) vertex	1/2	$\sqrt{1/5}$	0.6017
square	(Sa) mid-side	$\sqrt{1/2}$	$\sqrt{1/3}$	$\sqrt{1/3}$
	(Sb) vertex	1	0.4616	$\sqrt{1/5}$
hexagon	(Ha) mid-side	0.8580	0.4567	0.6081
	(Hb) vertex	0.6414	0.5068	1/3

Curious readers should note that the optimal solutions to Problems (Ta1) and (Sb1) are both $t = 1$, and the optimal solutions to Problems (Ta3), (Tb2) and (Sb3) are all $t = \sqrt{1/5}$. Interested readers may extend Table 4.1 by including other right prisms with different bases: for example, regular polygons with $n = 5, 7, 8$ sides.

4 Summary and an Open Problem

After solving a 2D problem of optimally folding a rectangular sheet of paper, we posed and solved the corresponding 3D problem of optimally slicing a prism with a regular polygonal base with $n = 3, 4, 6$ sides. We left to the reader to solve similar problems for $n = 5, 7, 8$, etc. sides. To the aspiring researcher we pose the following open problem.

Slicing a Cylinder Optimally: A right circular cylinder with a specified diameter AB will be sliced by a plane $\mathbb{F}$ which is the perpendicular-bisector of AT, where T is on the vertical line through B. How should you choose T to accomplish the following tasks:

(C1) minimize the area of $\mathbb{F}$; (C2) minimize the volume of the part of the cylinder below $\mathbb{F}$ and containing A; and (C3) maximize the volume of the part of the cylinder below a new plane which is the reflection of the bottom base about $\mathbb{F}$ and containing B?

We urge readers to identify other 2D problems that may be extended to 3D problems.

Acknowledgement

I concede that solutions to the paper folding problems exist on many internet sites other than the ones cited. I am grateful to Dr. Randall Stoetz for making wooden models of sliced prisms. I thank my colleague, Dr. Li, for some helpful comments. I am indebted to a referee's generous suggestion that led to Remarks 1 and 2.

Bibliography

[1] Lee Johnson, How to Solve Cubic Equations, Sciencing. https://sciencing.com/solve-cubic-equations-8136094.html

[2] Anil Kumar, Calculus: Application of Derivatives. https://www.youtube.com/watch?v=WL7o7p09R0o

[3] Ivan Niven (1981), *Maxima and Minima Without Calculus*, Vol. 6 of Dolciani Mathematical Expositions, 1981, Washington, DC: American Mathematical Society.

[4] Stack Exchange: Mathematics, Folding a rectangular paper and finding the area of the triangle so formed. https://math.stackexchange.com/questions/1092941

[5] James Stewart, *Calculus*, 8 ed., 2016, Boston, MA: Cengage Learning.

[6] Eric W. Weisstein, Cubic Formula, MathWorld—A Wolfram Web Resource. https://mathworld.wolfram.com/CubicFormula.html

[7] Eric W. Weisstein, Quadratic Equation, MathWorld—A Wolfram Web Resource. https://mathworld.wolfram.com/QuadraticEquation.html

[8] Eric W. Weisstein, Optimize the length of the crease of a folded piece of paper, MathWorld—A Wolfram Web Resource. https://demonstrations.wolfram.com/

[9] Wikipedia, the Free Encyclopedia, Mathematics of paper folding, https://en.wikipedia.org/wiki/Mathematics_of_paper_folding

[10] Paper folding: Maximizing the area of a triangle, Leading to Success in Algebra 2 Workshop 2003. https://studylib.net/doc/5888914/

Jyotirmoy Sarkar's research areas include enumeration, probability, statistics, and reliability theory. He enjoys reading, 'riting, 'rithmetic and R-coding.

5 Slicing a Cylinder Optimally

Jyotirmoy Sarkar

Indiana University Indianapolis
Department of Mathematical Sciences
402 N Blackford Street
Indianapolis, IN 46202-3216, USA

Email: jsarkar@iupui.edu

Collin Tully

Purdue University
Department of Computer Science
305 N University Street
West Lafayette, IN 47907, USA

Email: collin.tully@protonmail.com

Abstract: A sufficiently tall right-circular cylinder will be sliced by a plane which is the perpendicular-bisector of AT, where A is any point on the circular base and T is on the vertical line on the cylinder through a point B diametrically opposite to A. How should you choose T: (1) to minimize the area of the cut; (2) to minimize the volume of the part of the cylinder under the cut and containing A; and (3) to maximize the volume of the part of the prism below the reflection of the bottom circular base about the first cutting plane?

This paper is in response to an open problem posed in [3], from which the reader can review the calculus technique to solve optimization problems. Here we also document some new results and solve transcendental equations.

1 Picking a Point on a Cylinder

A right circular cylinder is at least as tall as it is wide; that is, the height is no shorter than the diameter. The goal is to pick a point T on the vertical curved surface to optimize each one of three desired objective functions, to describe which we need to define some other quantities.

Let the vertical line through T meet the circular base $\mathbb{B}$ at B. Let BA be a diameter of the base with center O. Join AT and consider its perpendicular bisector plane $\mathbb{F}$. Reflect the base $\mathbb{B}$ about $\mathbb{F}$ to obtain a second plane $\mathbb{G}$. See Figure 1. Because a cylinder is a solid of revolution about its axis, without loos of generality, we impose a coordinate system with $O = (0, 0, 0), A = (-1, 0, 0), B = (1, 0, 0)$ and $T = (1, 0, 2t)$. Thus, choosing T is equivalent to choosing t in the interval $(0, 1)$.

We shall separately solve the following three optimization problems: (1) Choose t to minimize the area of $\mathbb{F}$; (2) Choose t to minimize the volume of the portion of the cylinder below $\mathbb{F}$ and containing A; and (3) Choose t to maximize the volume of the portion of the cylinder below $\mathbb{G}$ and containing B.

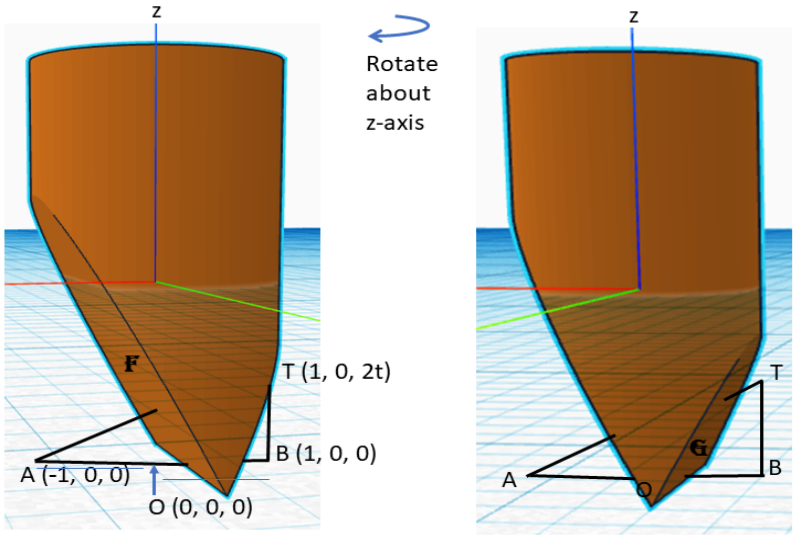


Figure 1: Let AB be a diameter of the bottom circular plane of a right circular cylinder twice as tall as the diameter. Choose a point T vertically above B so that when the cylinder is sliced with a plane orthogonal to AT: (1) the area of the cutting surface $\mathbb{F}$ is minimum, (2) the volume of the missing portion under $\mathbb{F}$ is minimum, and (3) the volume of the missing portion under $\mathbb{G}$ is maximum.

2 Folding a Paper Optimally

Let us begin with a simpler problem, whose sources and solutions are given in [3].

A standard A4 sheet of paper $ABCD$ with $AB = CD = 210$ mm and $BC = AD = 297$ mm will be folded along a segment XY, where X is on AB and Y is on AD, so that the vertex A will fall on a point T on BC. See Figure 1. How will you choose T so that either (i) the length of XY is minimum; or (ii) the area of $\triangle AXY$ is minimum; or (iii) the area of $\triangle BXT$ is maximum?

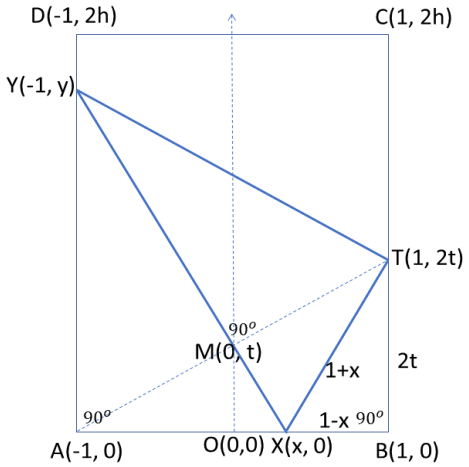


Figure 2: Fold the paper so that vertex A falls on a point T on edge BC, making a crease along XY with X on AB and Y on AD, and either (1) minimize the fold length XY, or (2) minimize the area of $\triangle AXY$, or (3) maximize the area of $\triangle BXT$.

Readers may see [4], [6], [8] for demonstrations and derivations of solutions to these problems. Here we simply jot down the answers.

Writing $BT = 2t$, the length of the fold is $XY = (1+t^2)^{3/2}/t$, the area of right $\triangle AXY$ is $(1+t^2)^2/(2t)$, and the area of right $\triangle BXT$ is $t\,(1-t^2)$. Using standard single-variable calculus method, one finds that (1) $t = \sqrt{1/2} \approx 0.7071$ minimizes XY, with minimum value $3\sqrt{3}/2 = 2.598076$; (2) $t = \sqrt{1/3} \approx 0.57735$ minimizes area of $\triangle AXY$, with minimum value $8\sqrt{3}/9 = 1.539601$; and (3) $t = \sqrt{1/3}$ maximizes area of $\triangle BXT$, with maximum value $2\sqrt{3}/9 = 0.384900$.

3 A Cylinder is a Special Prism

Moving from a 2D rectangular piece of paper to a 3D right prism with a regular polygonal base on n sides, [3] works out the optimal choices of T for $n = 4, 3, 6$ in two cases: (a) A is a vertex of the regular base; and (b) A is the midpoint of any one side of the base. In both cases, B is on the base diametrically opposite to A. As the number of sides of a regular polygonal base increases, the base approaches a circle. When the base is circular, the prism becomes a cylinder, and cases (a) and (b) become identical. A right circular cylinder with a specified diameter AB will be sliced by a plane which is the perpendicular-bisector $\mathbb{F}$ of AT, where T is on the vertical line through B. How should you choose T to accomplish the following tasks: (C1) minimize the area of the cut $\mathbb{F}$; (C2) minimize the volume of the part of the cylinder below $\mathbb{F}$ and containing A; and (C3) maximize the volume of the part of the cylinder below a new plane $\mathbb{G}$ which is the reflection of the bottom base $\mathbb{B}$ about $\mathbb{F}$ and containing B?

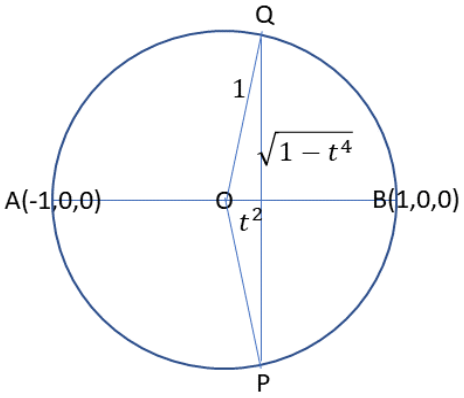


Figure 3: If a circle has a diameter AB and an orthogonal line PQ through a point on AB other than the center, what are the areas to the two sides of PQ?

We must find the mathematical expressions for the objective functions in Problems (C1)–(C3). As a first step, let us study how the area of the circular base of the cylinder, $\{x^2 + y^2 \leq 1, z = 0\}$, shown in Figure 3, is split by a chord PQ orthogonal to AB. Writing $PQ = \{x = t^2, z = 0\}$, the area of the region $\mathbb{B}_R$ to the right of PQ equals the area of the circle sector $OPBQ$ *minus* the area of triangle OPQ, or

$$\cos^{-1}(t^2) - t^2\sqrt{1-t^4}; \tag{1}$$

and hence the area of the region $\mathbb{B}_L$ to the left of the line PQ equals

$$\pi - \cos^{-1}(t^2) + t^2\sqrt{1-t^4}. \tag{2}$$

Remark 1. Note that when $t^2 = 0$, the areas to the right and the left of PQ, evaluated from (1) and (2), are both $\pi/2$, which agree with what we know from ele-

mentary geometry. Also, the areas of $\mathbb{B}_R$ and $\mathbb{B}_L$ to the right and the left sides of $P'Q' = \{x = -t^2, z = 0\}$ are obtained either by evaluating (1) and (2) at $-t^2$, or by evaluating (2) and (1) at t^2, because $\cos^{-1}(-t^2) = \pi - \cos^{-1}(t^2)$.

In solving optimization Problems C1–C3, involving transcendental expressions (1), (2) or their derivatives, eventually we resort to numerical computations.

3.1 Solution to Problem (C1)

Recall from Figure 2 that $AY = AX/t$. Hence, the cutting plane $\mathbb{F}$ is a magnification of $\mathbb{B}_L$, the part of the base circle to the left of PQ, by a factor $XY/XA = \sqrt{1+t^{-2}}$. Therefore, $\mathbb{F}$ has area

$$A(t) = \left[\pi - \cos^{-1}(t^2) + t^2\sqrt{1-t^4}\right]\sqrt{1+t^{-2}}. \tag{3}$$

Proposition 3 (Minimizing the area of $\mathbb{F}$)**.** *The minimum area of* $\mathbb{F}$ *is* 4.3767562326*, attained at* $t = 0.692758$*, or* $t^2 = 0.479914$*.*

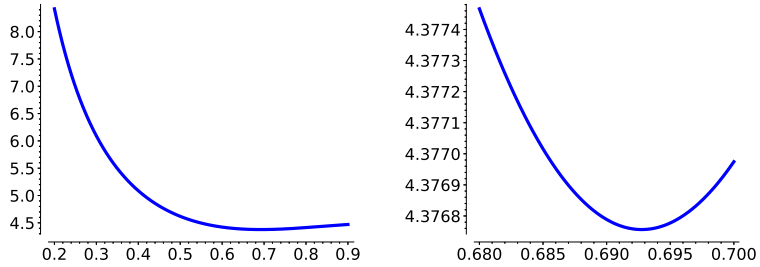


Figure 4: The objective function $A(t)$, the area of the cut $\mathbb{F}$

Proof of Proposition 3: Let us substitute $u = t^2$. Then expression (3) for the area of the first cutting plane $\mathbb{F}$ becomes $f(u) = \sqrt{1+1/u}\left(\pi + u\sqrt{1-u^2} - \cos^{-1}(u)\right)$, to minimize which we must solve the first-order condition

$$f'(u) = \frac{\sqrt{1-u^2}(\cos^{-1}(u) - \pi) - 4u^4 - 3u^3 + 4u^2 + 3u}{2u^2\sqrt{1-u^2}\sqrt{1+1/u}} = 0.$$

Equivalently, we must solve

$$\sqrt{1-u^2}(\cos^{-1}(u) - \pi) = 4u^4 + 3u^3 - 4u^2 - 3u. \tag{4}$$

The left-hand-side of (4), say $g(u) = \sqrt{1-u^2}(\cos^{-1}(u) - \pi)$, has derivative

$$g'(u) = \frac{(\pi - \cos^{-1}(u))u}{\sqrt{1-u^2}} - 1,$$

which is strictly increasing on $[0,1]$ because factors $\pi - \cos^{-1}(u)$, u and $(1-u^2)^{-1/2}$ are strictly increasing on $[0,1]$. Therefore, by Taylor's theorem, in any interval $(a,b) \subset (0,1)$, the function $g(u)$ is bounded by linear functions as follows:

$$g(a) + g'(a)(u-a) < g(u) < g(a) + g'(b)(u-a).$$

Using the above two bounds with the choice of interval (a,b) as $(0, 0.47)$, $(0.48, 0.88)$ or $(0.89, 1)$, we note that (4) does not have a solution in these intervals. Also, since f'' is strictly negative over $[0.88, 0.89]$, there exists no minimizer in $[0.88, 0.89]$. Finally, since f'' is strictly positive over $[0.47, 0.48]$, there exists a unique solution to (4) in $[0.47, 0.48]$. This unique solution can be found using a numerical method. For instance, using Newton's method, we found the solution $u \approx 0.4799$. Because the second-order condition holds (f'' is positive), $u \approx 0.4799$ is the argmin with the corresponding minimum value $f(u) \approx 4.3768$. Since $\lim_{u\to 0+} f(u) = \infty$ and $\lim_{u\to 1-} f(1) = \sqrt{2}\pi$, we conclude that the global minimum of f is attained at $u \approx 0.4799$ or $t = \sqrt{u} \approx .6928$. The minimum area of $\mathbb{F}$ is 4.3768. ■

3.2 Volume of a Partial Slant-Half Cylinder

We shall use the following lemma to find the volume of the portion of the cylinder under the first cut $\mathbb{F}$ or under the reflection $\mathbb{G}$ of the base about the first cut.

Lemma 1 (Volume of the upper portion of a slant-half cylinder). Suppose that the right circular cylinder $\{x^2 + y^2 \le 1, |z| \le 1\}$ is sliced in half by the slant plane $\{x = z\}$ making 45^o angles with the top and the bottom circular cross-sections. If the slant-half that contains the bottom circular base $\{x^2 + y^2 \le 1, z = -1\}$ is further sliced by a plane $\{z = h\}$ parallel to the base at height $h \in [-1, 1]$ (see Figure 5), then the upper portion of the slant-half cylinder has volume

$$V(h) = -h\cos^{-1} h + \sqrt{1-h^2}\,(2+h^2)/3. \tag{5}$$

Proof of Lemma 1: Both cases (Ca) $h \in [0,1]$ and (Cb) $h \in [-1,0]$, shown in Figure 5, are proved using the same technique. Let us thinly slice the upper portion of the slant-half cylinder with planes parallel to the base. By integrating the areas of the

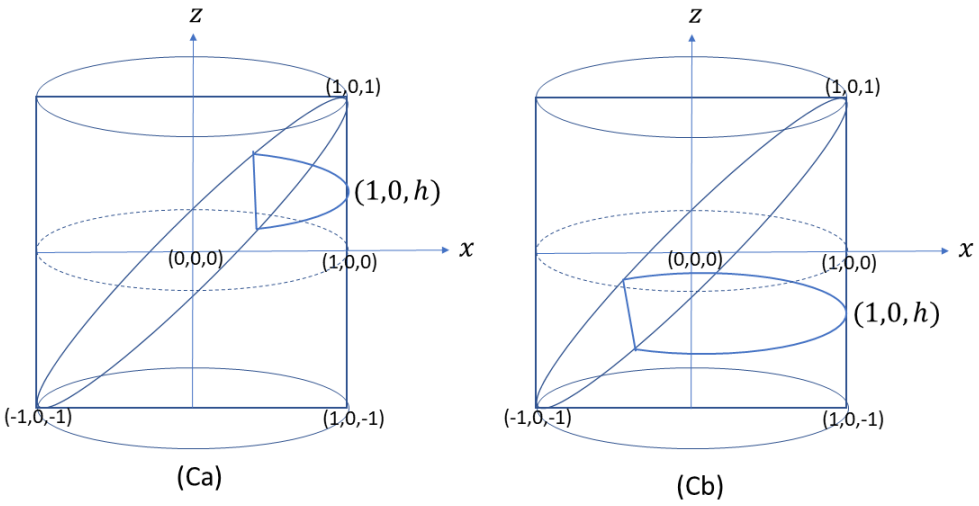


Figure 5: A cylinder, as high as the base diameter, is sliced into equal halves by a 45^o slant cut. Then the slant-half cylinder with a circular base at the bottom is sliced again by a plane cut parallel to the base at height (Ca) $h \in [0, 1]$, or (Cb) $h \in [-1, 0]$. What is the volume of the upper part of the slant-half cylinder?

cross-sections of these thin slices given by (1), and making a change of variable $z = \cos\theta$, the volume of the upper portion of the slant-half cylinder is seen to be

$$\begin{aligned} V(h) &= \int_h^1 \left\{\cos^{-1} z - z\sqrt{1-z^2}\right\} dz \\ &= \int_0^{cos^{-1}h} \{\theta - \sin\theta\ \cos\theta\} \sin\theta\, d\theta \\ &= \int_0^{cos^{-1}h} \theta\ \sin\theta\, d\theta - \int_0^{cos^{-1}h} \sin^2\theta\ \cos\theta\, d\theta = V_1(h) - V_2(h),\ \text{ say,} \end{aligned}$$

where, applying integration by parts, we have

$$\begin{aligned} V_1(h) &= -\theta\ \cos\theta\,|_0^{\cos^{-1}h} + \int_0^{cos^{-1}h} \cos\theta\, d\theta = -h\ \cos^{-1}h + \sqrt{1-h^2} \\ V_2(h) &= \sin^3\theta\,|_0^{\cos^{-1}h} - 2\int_0^{cos^{-1}h} \sin^2\theta\ \cos\theta\, d\theta = (1-h^2)^{3/2} - 2V_2(h), \end{aligned}$$

whence $3V_2(h) = (1-h^2)^{3/2}$; or $V_2(h) = (1-h^2)^{3/2}/3$. Hence, the expression for the volume of the upper portion of the slant-half cylinder becomes

$$V(h) = V_1(h) - V_2(h) = -h\ \cos^{-1}h + \sqrt{1-h^2} - (1-h^2)^{3/2}/3,$$

which simplifies to (5), completing the proof of the lemma. ■

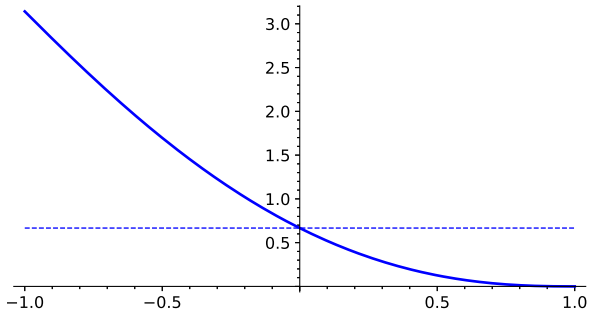


Figure 6: The volume above a horizontal slice at height h of a slant-half cylinder

Remark 2. By evaluating (5) at $h = 1, -1$, we obtain $V(1) = 0, V(-1) = \pi$, which agree with our prior knowledge. However, $V(0) = 2/3$ is a pleasant surprise: How did π vanish from the expression? Furthermore, from (5), we note that $V(-h) - V(h) = h\,[\cos^{-1}(-h) + \cos^{-1} h] = h\,\pi$, for all $h > 0$, which agrees with the value we anticipate from geometric symmetry, $V(-h) - V(h) = (1/2) \cdot \pi \cdot 2h = \pi\, h$.

3.3 Solution to Problem (C2)

If the cutting plane $\mathbb{F}$ intersects AB at $(t^2, 0, 0)$, then the part of the cylinder below $\mathbb{F}$ and containing A has volume $W_1(t)$, which equals $1/t$ times the volume of the upper slant half cylinder with $h = -t^2$ in (5); that is, the desired volume is

$$W_1(t) = V(-t^2)/t = t\,\cos^{-1}(-t^2) + \sqrt{1 - t^4}\,(2 + t^4)/(3t). \tag{6}$$

Proposition 4 (Minimizing the volume under $\mathbb{F}$)**.** *The minimum volume under the cutting plane* $\mathbb{F}$ *is* 2.2382434*, attained at* $t \approx 0.52775$*, or* $t^2 = 0.278520$*.*

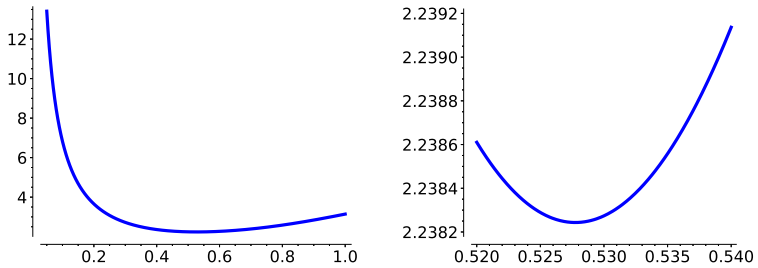


Figure 7: The objective function $W_1(t)$, the volume under the cut $\mathbb{F}$

Proof of Proposition 4: We wish to minimize the function

$$W_1(t) = t\,\cos^{-1}(-t^2) + \sqrt{1-t^4}\,(2+t^4)/(3t).$$

Note that the first derivative

$$W_1'(t) = \left(\frac{5}{3}t^2 - \frac{2}{3}t^{-2}\right)\sqrt{1-t^4} + \cos^{-1}(t^2)$$

is monotonically increasing on $(0,1)$ because the second derivative

$$W_1''(t) = \frac{10t^5+2t}{3\sqrt{1-t^4}} + \frac{2}{3}\left(5t+2t^{-3}\right)\sqrt{1-t^4}$$

is strictly positive, being the composition of strictly positive functions. Therefore $W_1(t)$ is strictly convex, and hence has a unique minimum in the interval $(0,1]$. The argmin is determined numerically, using Newton's method as in Proposition 3, to be $t \approx 0.52775$. ■

3.4 Solution to Problem (C3)

If the cutting plane $\mathbb{F}$ intersects AB at $(t^2,0,0)$, then the part of the cylinder under the reflection of the base about $\mathbb{F}$ and containing B has volume $W_2(t)$, which equals $2t/(1-t^2)$ times the volume of the upper slant-half cylinder with $h = t^2$ in (5); that is, the required volume is

$$W_2(t) = V(t^2)2t/(1-t^2) = 2t\left[-t^2\,\cos^{-1}(t^2) + \sqrt{1-t^4}\,(2+t^4)/3\right]/(1-t^2). \qquad (7)$$

Proposition 5 (Maximizing the volume under the reflection)**.** *The maximum volume under the reflection $\mathbb{G}$ of the base $\mathbb{B}_L$ about $\mathbb{F}$ is* 0.4485601341 *attained at* $t^2 = 0.2627454$*, or* $t = 0.512587$*.*

Proof of Proposition 5: Substituting $u = t^2$ on the right-hand-side of (7), and setting the derivative with respect to u to zero, we must solve

$$5u^5 - 7u^4 - 7u^3 + 5u^2 + 2u + 2 = 3u\sqrt{1-u^2}\,\cos^{-1}(u)(3-u). \qquad (8)$$

Factoring the polynomial on the left-hand-side of (8), we have

$$-(5u^3 - 7u^2 - 2u - 2)(1-u^2) = 3u\sqrt{1-u^2}\,\cos^{-1}(u)(3-u).$$

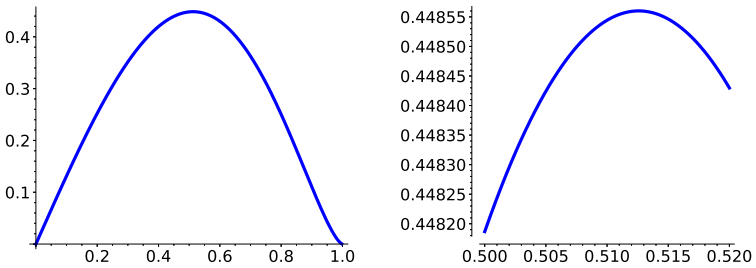


Figure 8: The objective function $W_2(t)$, the volume under the plane $\mathbb{G}$

Next, canceling out the common factors from the two sides, it suffices to solve

$$(5u^3 - 7u^2 - 2u - 2)\sqrt{1-u^2} = 3u\cos^{-1}(u)(u-3). \tag{9}$$

Because functions $\cos^{-1}(u)$ and $\sqrt{1-u^2}$ are strictly concave and decreasing on $[0,1]$, for any $u \in [a,b] \subset [0,1]$, the following two inequalities hold:

$$\cos^{-1}(b) \le \cos^{-1}(u) \quad \text{and} \quad \sqrt{1-u^2} \le \sqrt{1-a^2} - \frac{a}{\sqrt{1-a^2}}(u-a).$$

Moreover, because $(5u^3-7u^2-2u-2) < 0$ and $(u-3) < 0$ on $(0,1)$, for any $u \in [a,b] \subset [0,1]$, we have

$$\begin{aligned}&(5u^3 - 7u^2 - 2u - 2)\left(\sqrt{1-a^2} - \frac{a}{\sqrt{1-a^2}}(u-a)\right)\\ &\le (5u^3 - 7u^2 - 2u - 2)\sqrt{1-u^2} = 3u\cos^{-1}(u)(u-3) \le 3u\cos^{-1}(b)(u-3),\end{aligned}$$

whence we find the following necessary (though not sufficient) condition to guarantee that a proposed interval (a,b) contains a local optimum:

$$(5u^3 - 7u^2 - 2u - 2)\left(\sqrt{1-a^2} - \frac{a}{\sqrt{1-a^2}}(u-a)\right) \le 3u\cos^{-1}(b)(u-3).$$

Using this quartic polynomial, we determine that there exists no local optimum in the intervals $(0, 0.26)$ and $(0.27, 0.98)$. Thereafter, we check that no solution exists in $[0.98, 1]$, but a unique solution exists in $[0.26, 0.27]$. As in Propositions 3 and 4, using Newton's method, we find the unique optimum at $u \approx .2627$ or $t \approx 0.5126$. The maximum volume under the reflection of the base about $\mathbb{F}$ is 0.44856. ■

4 Optimizing Some Linear Combinations

While we have solved three optimization problems separately, need may arise to optimize some linear combinations of the three objective functions. For example, one may

wish to minimize $L_1(t) = \lambda\, A(t) + W_1(t) - W_2(t)$ or $L_2(t) = \lambda\, A(t) + W_1(t)$, where λ is a constant multiplier of a two-dimensional measure of area that makes the product compatible to a three-dimensional measure of volume making the addition (and subtraction) meaningful.

For illustration, choosing $\lambda = 0.05$, we found the optimal solutions are $t = 0.5304$ and $t = 0.5370$ (with minimum values 2.0173 and 2.4649, respectively). Moreover, the linear combinations L_1 and L_2 are not sensitive to the choice of λ, and so the optimal t remains similar for all $\lambda \in [0, 1]$.

For a quick reference, let us summarize the solutions to all optimization problems solved in this paper in Table 5.1.

Table 5.1: Target objective functions, optimal solutions and optimal functional values (when $\lambda = 0.05$)

objective	optimal t	optimal t^2	optimal value
maximize W_2	0.5126	0.2627	0.4486
minimize $W_1 - W_2$	0.5235	0.2740	1.7901
minimize W_1	0.5278	0.2785	2.2382
minimize $L_1 = \lambda\, A + W_1 - W_2$	0.5304	0.2813	2.0173
minimize $L_2 = \lambda\, A + W_1$	0.5370	0.2883	2.4649
minimize A	0.6928	0.4799	4.3768

Thus, the optimal value of objective function L_1, compared to that of L_2, is about 18.15% lower when $\lambda = .05$, and about 20% lower when $\lambda = 0$. The argmin for L_2 is slightly larger than that for L_1.

5 Summary

Responding to the open problem in [3], we have sliced a right circular cylinder optimally to achieve three desirable objectives: minimize the area $A(t)$ of the cutting surface $\mathbb{F}$, minimize the volume $W_1(t)$ of the portion of the cylinder below the cutting surface $\mathbb{F}$ and maximize the volume $W_2(t)$ of the portion of the cylinder below the reflection $\mathbb{G}$ of the base circle about the cutting surface $\mathbb{F}$.

Similar problems remain open for optimally slicing a prism with a regular polygonal base having $n > 6$ sides. We urge readers to identify other 2D problems that may be extended to 3D problems. For instance, given a convex quadrilateral in 2-D, the point whose total distance from the four vertices is minimum is the point of intersection of the

two diagonals. Given a tetrahedron in 3-D, find the point whose total distance from the four vertices is minimum.

Acknowledgement

We are grateful to Dr. Randall Stoetz for making wooden models of sliced cylinders. We also thank a referee for suggesting some improvements.

Bibliography

[1] Anil Kumar, Calculus: Application of Derivatives. https://www.youtube.com/watch?v=WL7o7p09R0o

[2] Ivan Niven (1981), *Maxima and Minima Without Calculus*, Vol. 6 of Dolciani Mathematical Expositions, 1981, Washington, DC: American Mathematical Society.

[3] Jyotirmoy Sarkar, Slicing a Prism Optimally: Motivating Students to Learn Calculus. *Blackboard: Bulletin of the Mathematics Teachers' Association (India)*. Issue 6, 2024, pp 25–40.

[4] Stack Exchange: Mathematics, Folding a rectangular paper and finding the area of the triangle so formed. https://math.stackexchange.com/questions/1092941

[5] James Stewart, *Calculus*, 8 ed., 2016, Boston, MA: Cengage Learning.

[6] Eric W. Weisstein, Optimize the length of the crease of a folded piece of paper, MathWorld—A Wolfram Web Resource. https://demonstrations.wolfram.com/

[7] Wikipedia, the Free Encyclopedia, Mathematics of paper folding, https://en.wikipedia.org/wiki/Mathematics_of_paper_folding

[8] Paper folding: Maximizing the area of a triangle, Leading to Success in Algebra 2 Workshop 2003. https://studylib.net/doc/5888914/

6 Platonic Solids

Lovy Singhal

Email: lovysinghal@gmail.com

Jugal Verma

Department of Mathematics
Indian Institute of Technology Bombay
400076

Email: jkv@math.iitb.ac.in

Abstract: In this expository article, we discuss in detail all the five regular polyhedra better known as Platonic solids and examine their various properties *viz.* uniqueness, groups of symmetries and relationships with the group of rotations in $\mathbb{R}^3$ etc.

1 History & Omnipresence

Regular polyhedra refer to the special category of three-dimensional solids all of whose faces are congruent regular convex polygons. They have been known to human beings at least since the time of the ancient Greeks. Most of the Great Pyramids of Egypt correspond to the upper half of an octahedron. Indeed, the Pythagoreans knew about the cube, the tetrahedron and the octahedron (see § 2 for more). They had also developed knowledge of various geometric properties of these solids such as lengths of their diagonals, radii of the smallest spheres in which they can be inscribed etc. Hermann Weyl put it best as [13]:

> *One might say that the existence of the cube, the tetrahedron and the octahedron is a fairly trivial geometric fact. But the discovery of the dodecahedron and the icosahedron is certainly one of the most beautiful and singular discoveries made in the whole history of mathematics.*

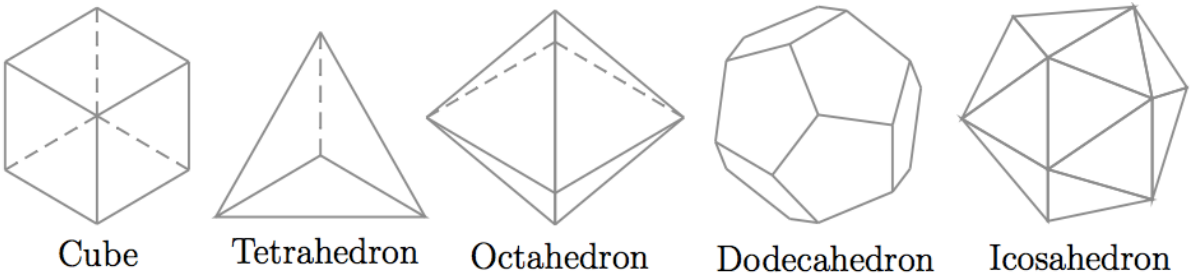


Figure 1: Regular polyhedra.

Cromwell [5] proved that the following statements about a regular polyhedron are equivalent:

1. All of its vertices lie on a sphere.

2. All of its dihedral angles are equal in magnitude.

3. All faces meeting at any of its vertices are regular polygons having the same number of sides.

4. All of its solid angles are equal.

5. All of its vertices are surrounded by the same number of faces.

There are exactly five such solids, as we will prove in § 4. They are: the cube, the tetrahedron, the octahedron, the dodecahedron, and the icosahedron. Plato in his *Timaeus* associated tetrahedron with fire, octahedron with air, cube with the Earth and icosahedron with water while the whole universe itself was linked to dodecahedron [14]. Long before the celebrated German astronomer Johannes Kepler established the three laws of planetary motion, he tried to explain sizes of planets and interplanetary distances with the help of Platonic solids. The six spheres in his arrangement of regular solids correspond to six planets, Saturn, Jupiter, Mars, Earth, Venus and Mercury. These spheres are separated from each other by a cube, tetrahedron, dodecahedron, octahedron and icosahedron [7]. We remind the reader that the outer planets were yet to be discovered by Kepler's time! As Cromwell says, "Fascination with these solids led both Kepler and Plato to use them in their theories of the cosmos."

Circa 1887, the redoubtable German scientist Ernst Haeckel drew pictures of various single-celled creatures called radiolaria. He named three of them as *Circoporus octahedrus*, *Circogonia icosahedra* and *Circorrhegma dodecahedra* because they look very akin to the eponymous Platonic solids [8]. As any jeweller would swear, polyhedra manifesting themselves as cut gemstones also occupy an important place in bullion markets.

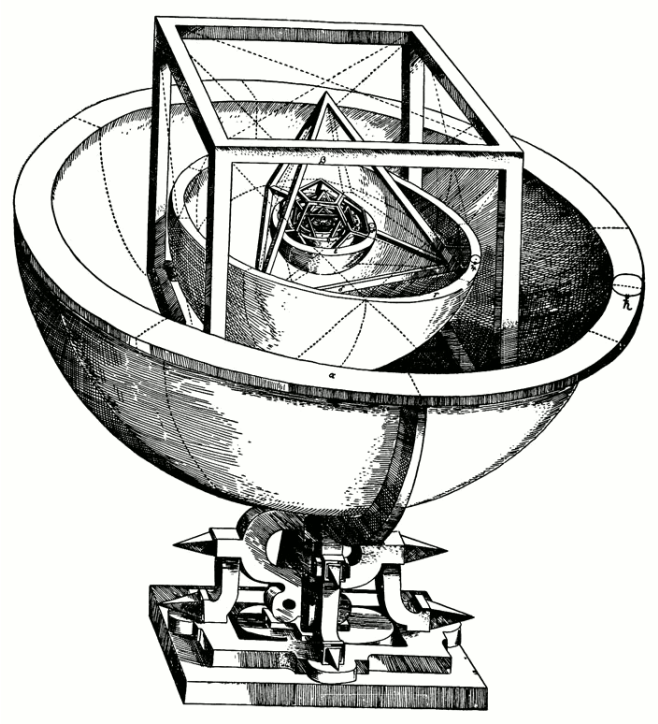


Figure 2: An illustration of Kepler's Platonic solid model of the Solar system.

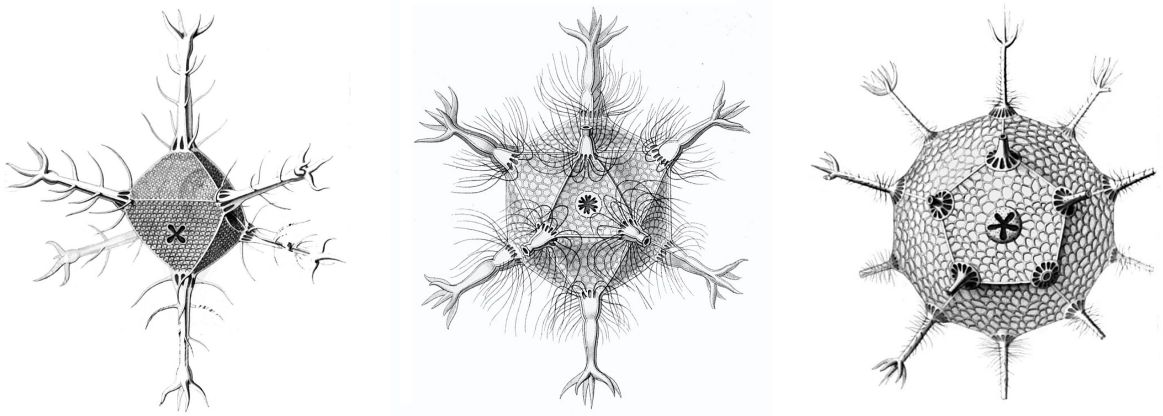


Figure 3: (from left to right) *Circoporus octahedrus*, *Circogonia icosahedra* and *Circorrhegma dodecahedra*.

Euclid around 300 BC is understood to have compiled all the results in geometry and number theory known till his time. The end-product of this massive exercise was *Elements* – a colossal work neatly divided into 13 books, last of which deals with solid geometry and ends with a discussion on Platonic solids in particular. Much of this book is due to another mathematician Theaetetus of Athens (ca. 417–369 BC). Some historians have gone to the extent of claiming that the chief goal of Euclid's *Elements* was construction of the Platonic solids [1, p. 213].

> *The textbook that shall really replace Euclid has not yet been written and probably never will be.*
>
> – Encyclopaedia Britannica

Though the study of polyhedra had largely ceased to be a research mathematician's call during the Middle Ages, it has witnessed renewed interest in the last century or so with links discovered to deep topics like group theory, geometry of singularities etc. (see [9] for example). It is, therefore, our sincere hope that this introduction has convinced the reader to continue with the upcoming sections.

2 Geometry of Platonic Solids

In this section, we introduce all of the five regular polyhedra along with only some of their interesting geometric properties. The relevance and ubiquity of polyhedra in life outside of mathematics is further cemented.

Cube

A cube is perhaps the next solid, second only to a sphere, which comes to mind when one thinks of the concept of a solid itself. It has 8 vertices, 12 edges and 6 faces. The diameter of the sphere circumscribing a cube is $\sqrt{3}$ times the length of its side being equal to the length of its (body) diagonal. Some of you might have heard of the famous *Delian problem* which asks the following:

> Given a cube, construct another cube whose volume is double the first.

If we take the original cube to have edge length one, its volume then becomes $1^3 = 1$. The problem above is thereby reduced to constructing the edge of a new cube of length x satisfying $x^3 = 2$. With the help of ideas from *Field Theory,* one can show that

such a construction is impossible using only a compass and an unmarked ruler (see [3, § 13.4]).

Tetrahedron

A tetrahedron has 4 vertices, 6 edges and 4 faces. The diameter of its circumsphere is $\sqrt{3}/2$ times the length of any of its six equally long edges. The geometry of a tetrahedron takes central place in the area of organic chemistry in the sense that many properties of organic compounds are explained on its basis. As a matter of fact, one of the most-reputed journals on this branch of knowledge is called the *Tetrahedron* [11]. Our polyhedron is no less important in the chemistry of other inorganic compounds involving *tetravalent* ions. We must point out here that many molecules of viable commercial importance are known to possess a tetrahedral structure, e.g., methane (see Fig. 4).

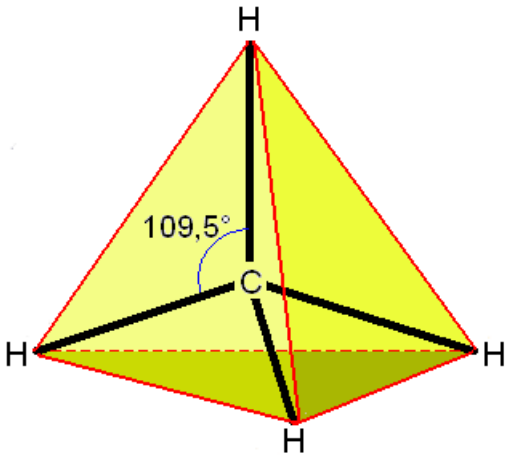


Figure 4: A molecule of methane (CH_4).

Octahedron

An octahedron has 6 vertices, 12 edges and 8 faces. The diameter of the sphere circumscribing an octahedron is $\sqrt{2}$ times the length of its side. The octahedron has received its large share of fame by lending its shape to most of the Pyramids built in Egypt. Besides, its geometry plays an indispensable role in areas such as organic chemistry, molecular orbital theory, crystallography etc. [12]

Dodecahedron

Some people have suggested that the dodecahedron was abstracted and 'perfected' by the Pythagoreans after being inspired by the iron pyrite crystals found in Sicily, Italy [4]. A dodecahedron has 20 vertices, 30 edges and 12 faces. All of its faces are regular pentagons, three of which meet at each of its vertices. This polyhedron has five different cubes inscribed in it. Conversely, these cubes provide one of the easiest method to construct a dodecahedron. On each face of a cube, we put a tent-like solid having two

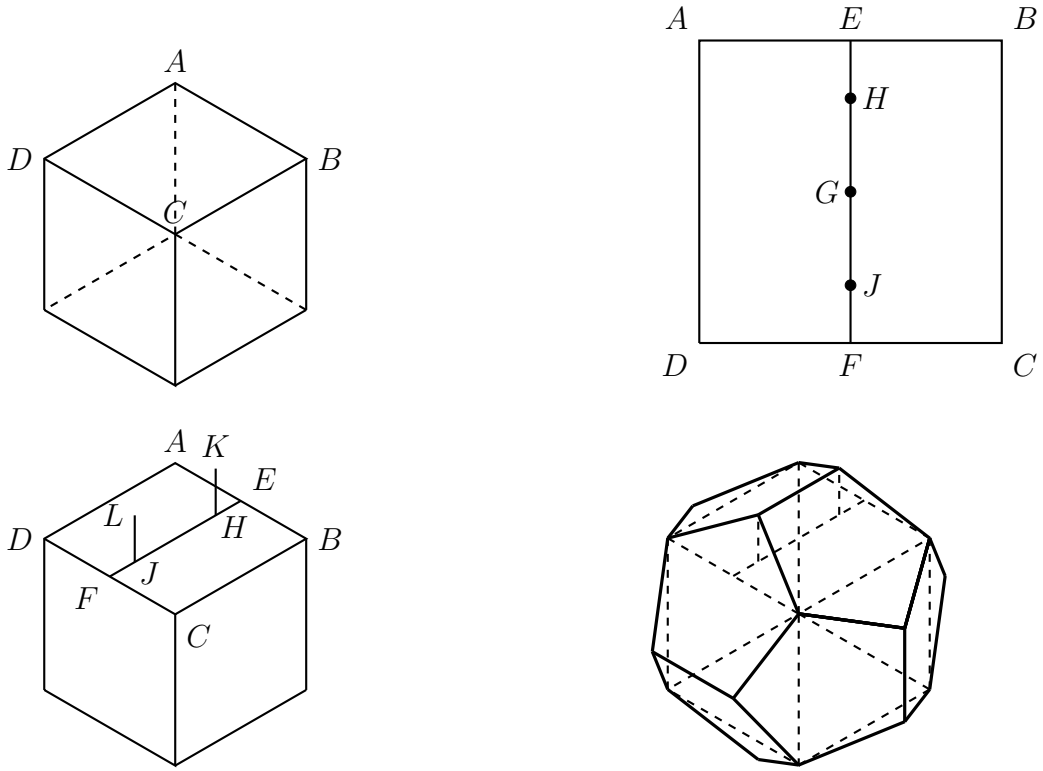


Figure 5: Constructing a dodecahedron.

vertices. On the face $ABCD$ of the cube, mark the mid-points E, F and G of the sides AB, CD and EF respectively. Divide EG and GF to get

$$|GH| : |EH| \;=\; |GJ| : |FJ| \;=\; \varphi$$

where $\varphi = (\sqrt{5}+1)/2 \approx 1.618\ldots$ is the golden ratio. Erect normals to the surface HK and JL so that $HK = JL = HG$. The roofs fit together and we get a dodecahedron. This construction is the one used by Euclid in his book to construct the dodecahedron on a paper. Interestingly, we found this same construction to be the easiest while drawing the dodecahedron on our computer for the purpose of this article and being completely unaware of the fact that Euclid had already prescribed so. This cannot be brushed aside as a mere coincidence and is yet another evidence for the broader compatibility and timeless relevance of Euclid's ideas in general.

Icosahedron

An icosahedron has 12 vertices, 30 edges and 20 faces. Five equilateral triangles meet at each of the vertices of an icosahedron. The bases of these five triangles form a regular pentagon. Pairs of opposite sides of an icosahedron help to constitute *golden rectangles* whose longer sides are diagonals of these pentagons (see Figure 6). Every two of these rectangles intersect at right angles. The 12 vertices of these rectangles are the exact same 12 vertices of our icosahedron. And once again, this awesome property of the icosahedron has been put to use by us while drawing for this article.

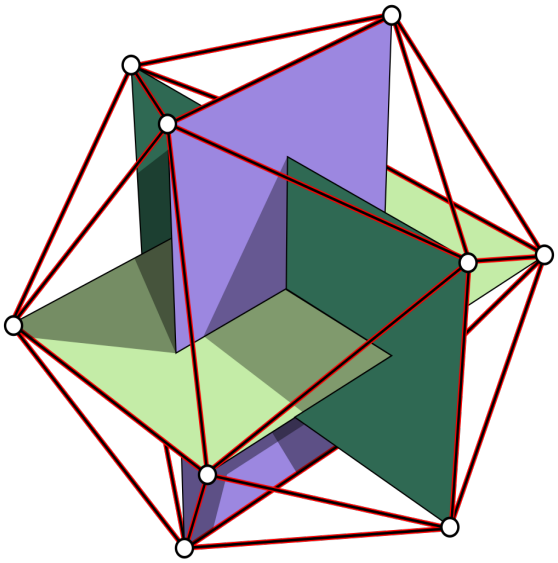


Figure 6: Golden rectangles and icosahedron.

Buckminsterfullerene, a carbon allotrope discovered in 1985, belongs to the celebrated family of fullerenes and is made up of 60 carbon atoms alone [10]. In structural terms, it is a *regular truncated* icosahedron in which no two pentagons share an edge. Three of the discoverers Harold Kroto, Robert Curl and Richard Smalley were awarded the 1996 Nobel Prize in Chemistry. The shape of this molecule resembles that of a football, one which has been filled using pentagons and hexagons (see Figure 7). It is worthwhile noting that the smallest member of the family of fullerenes C_{20} is actually a dodecahedron which goes by the same name.

Many of the geometric facts stated above were first compiled by Euclid in his *Elements* – Book 13, in which he also gave detailed proofs for them. We have refrained ourselves from writing the same statements as given in various translations of the *Elements* (see [6] for example). With due regards to Euclid, it is our humble opinion that the statements are too clumsy to be easily understood by an average 21st-century reader.

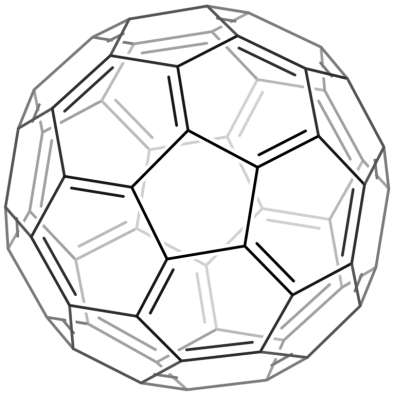
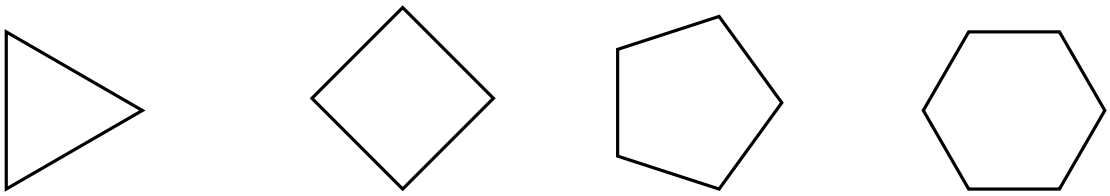


Figure 7: Molecular structure of Buckminsterfullerene.

3 Groups of rotational symmetries of Platonic solids

In this section, we intend to develop the concept of symmetry and rigid motions starting with a geometrical description and moving on to abstract group-theoretic concepts. We will then be describing the groups of symmetries of our five Platonic solids. Along the way, we will learn about the very important concept of group action. These shall be a pre-requisite for understanding the next two sections.

Symmetry of an object can be thought of as an operation under which the object remains unchanged. Consider a regular k-gon in the Cartesian plane $\mathbf{R}^2$ centered at the origin. We are interested in its rotation about its own axis (normal to our plane and passing



(a) $k = 3$ (b) $k = 4$ (c) $k = 5$ (d) $k = 6$

Figure 8: The first four regular k-gons.

through the origin) by angles which are integer multiples of $2\pi/k$. One can easily work out that any such rotation will bring the regular k-gon to itself in the two–dimensional plane. What we see here is an example of a more general concept, namely the *action* of a group on a set. In this case, the group consists of rotations by $n\theta$ where $n = 0, 1, 2, \ldots, k-1$ and $\theta = 2\pi/k$. The set being acted upon consists of all the vertices of our k-gon, or equivalently, all the k sides of a regular k-gon.

More generally, let G be a group and S be some non-empty set. By an *action* of G on S, we refer to a rule assigning some element $gs \in S$ for each pair $g \in G$ and $s \in S$.

Otherwise said, it is a map from $G \times S \to S$ sending

$$(g, s) \;\mapsto\; gs.$$

Our mapping must satisfy the following conditions:

1. for the identity element $e \in G$, we have $es = s$ for all $s \in S$, and

2. $(g_1 g_2)s = g_1(g_2 s)$ for all triples $g_1, g_2 \in G$ and $s \in S$ (*associativity*).

The set S endowed with such an action of G is also called a *G-set*.

Fix $g \in G$ and define a map m_g from S to itself given by

$$m_g(s) = gs \quad \forall\ s \in S.$$

We claim that m_g is bijective. This is because the second condition stated above implies that m_g has an inverse map, namely $m_{g^{-1}}$ which stands for action of the group element g^{-1} on S. Let us check

$$m_{g^{-1}}\big(m_g(s)\big) \;=\; g^{-1}\big(g(s)\big) \;=\; \big(g^{-1}g\big)(s) \;=\; es \;=\; s \text{ for all } s \in S.$$

We may similarly conclude that $m_g \circ m_{g^{-1}} = \mathrm{id}_S$ as well.

The next logical step is to partition the set S into orbits.

Definition 3.1. The *orbit* of an element s in a G-set S is the subset

$$Gs \;:=\; \{\, a \in S \mid a = gs \text{ for some } g \in G \,\}.$$

For example, the orbit of a vertex of a regular k-gon under the operation of the cyclic group of rotations C_k is the set of all vertices of the k-gon. As another example, let us consider the symmetric group S_n acting on the set $\{1, 2, \ldots, n+1\}$. The two orbits for this action are

$$\{1, 2, \ldots, n\} \text{ and } \{n+1\}.$$

With the help of these two examples, one should be able to see that a set S is indeed partitioned by the action of the group G into different orbits. In other words, the orbits for a group operation are equivalence classes given by the relation

$$a \sim b \quad \text{iff} \quad b = ga \text{ for some } g \in G. \tag{1}$$

The group G acts on each orbit independently of its action on other orbits. In particular, any element $g \in G$ does not carry elements from one orbit to another orbit for if it does so, then the two orbits are essentially the same as each orbit is an equivalence class for the relation given by (1).

Definition 3.2. The *stabilizer* of any element $s \in S$ is the subset G_s of G consisting of all elements which fix s, i. e.,

$$G_s \;:=\; \{\, g \in G \mid gs = s \,\}.$$

It is easy to check that G_s is a subgroup of G. The former is a cousin to the kernel of a group homomorphism from the world of group actions. As an example, consider the action of the group D_3 on the set of sides of an equilateral triangle via rotations and reflections. The stabilizer of any side consists of the identity transformation and the reflection which flips the triangle about the altitude bisecting the given side.

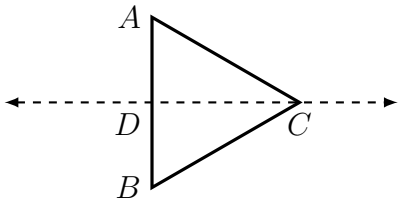


Figure 9: Reflection about CD sends the side AB to itself.

We can write precisely in terms of the stabilizer G_s when two elements $g, g' \in G$ act in an identical fashion on the element $s \in S$,

$$gs \;=\; g's \quad \text{iff} \quad g^{-1}g' \;\in\; G_s. \tag{2}$$

Now, let P be any subgroup of a group G. The left cosets

$$kP \;=\; \{\, kp \mid p \in P \,\}$$

partition the group G as

$$x \sim y \quad \text{iff} \quad y^{-1}x \in P$$

is clearly an equivalence relation on the set G.

Definition 3.3. The set of left cosets of a subgroup P of a group G is called the *(left) coset space* of G by P and is denoted as G/P.

We must emphasize that G/P will not be a group unless P is a normal subgroup of G. It will be of fundamental importance for us that G acts on the coset space G/P in a canonical manner. Given $g \in G$ and coset $C \in G/P$, the coset gC is defined as

$$gC \;=\; \{\, g\alpha \mid \alpha \in C \,\}.$$

If for example $C = g_1P$, then $gC = gg_1P$. This means that our group G acts *transitively* on the coset space G/P. For every pair of cosets

$$C_1 = g_1P \text{ and } C_2 = g_2P$$

in G/P where $g_1, g_2 \in G$, the group element $g_2g_1^{-1}$ takes coset C_1 to C_2.

Proposition 3.1. Let S be a G-set and $s \in S$. Also, let H be the stabilizer subgroup of s. There exists a natural bijection φ between the coset space G/H and the group orbit Gs given by

$$gH \longleftrightarrow gs. \tag{3}$$

The mapping φ respects the actions of G on G/H and S in the sense that

$$\varphi(gC) = g\varphi(C) \quad \forall g \in G,\ C \in G/H.$$

Proof. Clearly, the compatibility of φ with the actions of G is beyond any reasonable doubt subject only to questions over its existence. What we are left with is to prove that (3) is well-defined.

Suppose the symbols g_1H and g_2H represent the same coset for some group elements g_1 and g_2. We claim that $g_1s = g_2s$ too. This is so because $g_1H = g_2H$ iff $g_1^{-1}g_2 \in H$ or equivalently, $g_2 = g_1h$ for some $h \in H$. The last identity in turn implies that $g_2s = g_1hs = g_1s$ since $h \in H$, the stabilizer of s. We can even reverse the direction of our logic to conclude that $\varphi : G/H \to Gs$ is in fact injective.

The surjectivity of φ can be seen from the fact that all elements of Gs look like gs for some $g \in G$ and φ carries the coset gH to gs. ■

We have now earned the wherewithal to discuss transformations of the three-dimensional space. Any rotation of $\mathbb{R}^3$ about the origin can be described completely by two quantities, a nonzero vector $\mathbf{v}$ which lies along the axis of rotation, and a nonzero angle θ, the angle by which the rotation takes place about the vector $\mathbf{v}$. To think of the simplest example, a rotation by an angle θ about the unit vector $\mathbf{e_1}$ can be jotted down as

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{pmatrix}. \tag{4}$$

Indeed, every rotation of $\mathbb{R}^3$ is a linear map from $\mathbb{R}^3 \to \mathbb{R}^3$ but its matrix with respect to the standard basis can be increasingly involved.

Definition 3.4. The set of all $n \times n$ orthogonal matrices with real entries form the *orthogonal group*

$$\mathrm{O}_n := \{A \in GL_n(\mathbb{R}) \mid AA^t = I\}.$$

By its very definition, O_n is a subgroup of the *general linear group* $\mathrm{GL}_n(\mathbb{R})$. Next, the determinant of any orthogonal matrix is ± 1 since $AA^t = I$ implies

$$(\det A)^2 = \det A \cdot \det A^t = 1.$$

Definition 3.5. The subset of orthogonal matrices with determinant $+1$ forms a subgroup of O_n, namely the *special orthogonal group*

$$\mathrm{SO}_n \;=\; \{\, A \in GL_n(\mathbb{R}) \mid AA^t = I,\ \det A = 1 \,\}.$$

The orthogonal matrices with determinant $+1$ are called *orientation preserving* while those with determinant -1 are said to be *orientation reversing.*

Definition 3.6. A *rigid motion* or *isometry* of $\mathbf{R}^n$ is a map $f : \mathbf{R}^n \to \mathbf{R}^n$ preserving distances between every pair of its points. For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,

$$\| f(\mathbf{x}) - f(\mathbf{y}) \| \;=\; \| \mathbf{x} - \mathbf{y} \|.$$

Proposition 3.2. Let $f : \mathbb{R}^n \to \mathbb{R}^n$. Then, the following are equivalent:

i) It is a rigid motion which fixes the origin.

ii) It preserves dot products, i. e., $f(\mathbf{x}) \cdot f(\mathbf{y}) = \mathbf{x} \cdot \mathbf{y}$ for all $\mathbf{x}, \mathbf{y} \in \mathbf{R}^n$.

iii) It is left multiplication by an orthogonal matrix.

Theorem 1. A matrix B is a rotation of $\mathbf{R}^2$ ($\mathbf{R}^3$) iff $A \in \mathrm{SO}_2$ (SO_3).

Proof. We first observe that every rotation ρ is a rigid motion as it keeps the length of every vector unchanged. Thereafter, Proposition 3.2 tells us that ρ is left multiplication by some orthogonal matrix B. Moreover, $\det B = 1$ since the determinant function varies continuously with the rotation angle θ whereas $\theta = 0$ corresponds to the identity transformation I_n with $\det I_n = 1$. Thus, the matrix associated with any rotation is special orthogonal.

Conversely, let $B \in \mathrm{SO}_2$. Then, it is a 2×2 orthogonal matrix with determinant $+1$. Let $\mathbf{v_1}$ be the first column $B\mathbf{e_1}$ of B which is a unit vector owing to the orthogonality of B. There exists a rotation R such that $R\mathbf{e_1} = \mathbf{v_1}$. Thus, $\widetilde{B} = R^{-1}B$ fixes $\mathbf{e_1}$ where both B and R (by the first half of the proof above) are special orthogonal matrices. Hence, $\widetilde{B} \in \mathrm{SO}_2$ as well since the latter is a group under matrix multiplication. This will mean that columns of $\widetilde{B}$ form an orthonormal basis of $\mathbf{R}^2$ of which the first column is $\mathbf{e_1}$. The second column of $\widetilde{B}$ being orthogonal to $\mathbf{e_1}$ has no option but to be $\mathbf{e_2}$. Thus, we see that $\widetilde{B} = I$ and $B = R$ is very much a rotation. ■

In the three-dimensional space, Proposition 3.2 implies that any rigid motion fixing the origin is given by some orthogonal martix. Every element in SO_3 satisfies this criterion. Since $\rho \in \mathrm{SO}_3$ must preserve dot products, it sends the plane P orthogonal to $\mathbf{v}$ onto itself and P becomes an invariant subspace for $\rho : \mathbf{R}^3 \to \mathbf{R}^3$. After a suitable change of basis, we can see that the restriction of ρ to P is a rotation. The requirement that a rotation of $\mathbf{R}^3$ must also fix some non-zero vector $\mathbf{v}$ is the same as saying that $\mathbf{v}$ is an eigenvector for our map ρ with the corresponding eigenvalue being 1.

Lemma 3.1. Every matrix $A \in \mathrm{SO}_3$ has an eigenvalue 1.

We leave the proof of this lemma as an easy exercise for the reader. Its statement tells us that left multiplication by any matrix $A \in \mathrm{SO}_3$ fixes a nonzero vector $\mathbf{u_1}$ whose length can be normalized to 1. This is because $A\mathbf{u} = \lambda\mathbf{u}$ forces

$$A(c\mathbf{u}) \;=\; c\,(A\mathbf{u}) \;=\; c\,(\lambda\mathbf{u}) \;=\; \lambda(c\mathbf{u}).$$

We then extend $\{\mathbf{u_1}\}$ to obtain an orthonormal basis $\mathcal{S} = \{\mathbf{u_1}, \mathbf{u_2}, \mathbf{u_3}\}$ of $\mathbf{R}^3$. Observe that

$$T = P^{-1}AP, \text{ where } P = [\mathcal{S}] = [\,\mathbf{u_1}\;\mathbf{u_2}\;\mathbf{u_3}\,]$$

represents the same transformation with respect to the basis $\mathcal{S}$ as A does with respect to the standard basis $\{\mathbf{e_1}, \mathbf{e_2}, \mathbf{e_3}\}$. The bottomline is that $T \in \mathrm{SO}_3$ as T is orthogonal with both A and P being orthogonal while $\det T = \det A = 1$. The first column of T is $\mathbf{e_1}$ owing to $\mathbf{u_1}$ being the eigenvector of A with eigenvalue 1. Now, the other columns must be orthogonal to $\mathbf{e_1}$ whence T has the following shape

$$\left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & T_2 \end{array}\right)$$

where T_2 is a 2×2 real matrix. Since $T \in \mathrm{SO}_3$, we have $T_2 \in \mathrm{SO}_2$ and is therefore a rotation. This implies that T has the form given in (4) and, thereby, represents a rotation. In turn, A represents a rotation too.

Before we begin describing groups of rotational symmetries of each of the five Platonic solids, let H be a subgroup of any group G. We know that for any coset aH of H in G, $|\,H\,| = |\,aH\,|$. The number of such disjoint cosets of H in G is called the *index of* H, written as $[\,G : H\,]$ or $|\,G\,/\,H\,|$. Also, G is a union of all of these pairwise disjoint cosets of H whence

$$|\,G\,| \;=\; |\,H\,| \cdot |\,G\,/\,H\,|. \tag{5}$$

More generally, let S be a G-set. Combining Proposition 3.1 with (5),

Proposition 3.3 (Counting formula). Let G act on S and $s \in S$. Then,

$$|\,G\,| \;=\; |\,G_s\,| \cdot |\,Gs\,|$$

where Gs and G_s are respectively the orbit and stabilizer of s. In other words, the order of the orbit of s is equal to the index of the stabilizer of s.

An immediate consequence of Proposition 3.3 is that the order of a group orbit divides the order of the group. Consider, for example, the group G of orientation-preserving symmetries of a regular polyhedron P. It follows from Proposition 3.2 and Theorem 1 that these symmetries are all rotations of $\mathbf{R}^3$, or to say in SO_3. We are interested in the action of G on the set V of all vertices of P. The stabilizer of a vertex s is the subgroup of rotations about the line ℓ joining the origin (which we assume to be the centre of mass of our solid) and s by integer multiples of $2\pi/q$, where q is the number of regular p-gons which meet at any vertex of P. Let the polyhedron P have v many vertices. Now, the orbit of s is the whole of V as any vertex can be taken to any of its adjacent vertices. Because of Proposition 3.3, $|\,G\,| = qv$. Arguing similarly for edges and faces, we obtain

$$|\,G\,| \;=\; 2e \;=\; pf \;=\; qv$$

where e and f are respectively the number of edges and faces of P.

3.1 Group of rotational symmetries of a cube

As already mentioned in § 2, a cube has 8 vertices, 12 edges and 6 faces. The lines called the *axes of rotation,* about whom certain rotations are the symmetries of our solid, are of three kinds:

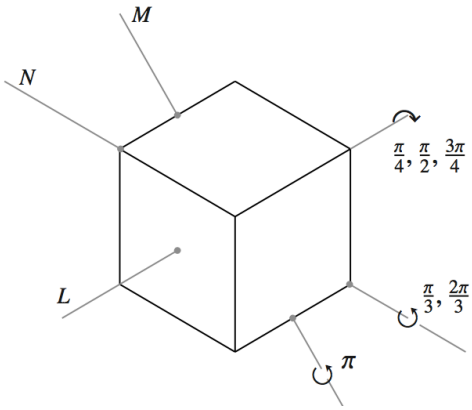


Figure 10: Rotational symmetries of a cube.

i) those normal to any pair of its opposite faces and passing via their centres of mass (e. g. axis L in Figure 10),

ii) those which join opposite edges and pass via their mid-points (e. g. axis M), and

iii) those which pass via diagonally opposite vertices of the cube (e. g. axis N in the figure above).

For any axis of the same type as L (see Figure 10), rotations by angles $\pi/2$, π and $3\pi/2$ radians take the cube onto itself in the three-dimensional space aside from the identity transformation. About axes of the second kind given above, rotation by an angle of π is the only non-trivial transformation which is a symmetry of the cube. For axes of the third kind, rotations by angles $2\pi/3$ and $4\pi/3$ are also symmetries of our solid. In a cube, there are 3 such different axes of the first kind, 6 of the second and 4 of the third kind. Hence, there are $(3 \times 3 + 6 \times 1 + 4 \times 2 =)$ 23 total such transformations apart from the identity transformation itself which together constitute all of its rotational symmetries. Recall that by symmetry, we not only imply here that all the vertices, edges and faces of our cube are permuted amongst themselves but also that the solid occupies the exact same three-dimensional space as before.

The reader must convince herself that any transformation of the cube that is a composition of some of these 24 rotations is once again a rotation belonging to this set. Next, rotations about axes of type L by either $\pi/2$ or $3\pi/2$, when repeated four times gives us the identity transformation of the cube. Indeed, these rotations are of order 4. Similarly, rotations about axes of types L and M by an angle of π are of order 2 while rotations about any axis of type N by angles $2\pi/3$ and $4\pi/3$ are of order 3 each. This set of 24 rotational symmetries of a cube constitutes its *group of rotational symmetries.* It has $(3 \times 1 + 6 \times 1 =)$ 9 elements of order two, 8 elements of order three and 6 elements of order four besides the trivial transformation.

3.2 Group of rotational symmetries of a tetrahedron

As has been explained already, a tetrahedron has 4 vertices, 6 edges and 4 faces. Its axes of rotation are of two kinds:

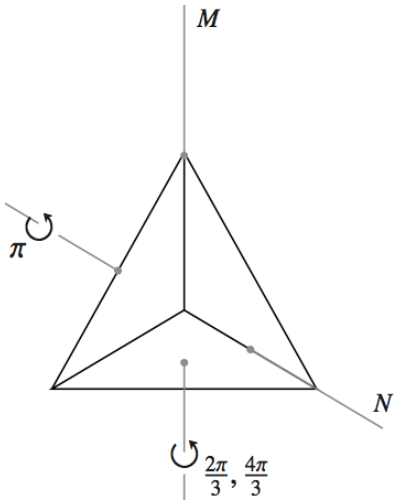


Figure 11: Rotational symmetries of a tetrahedron.

i) those passing via a vertex and the centroid of its opposite face (e. g. axis M in Figure 11), and

ii) those passing via mid-points of pairs of opposite edges (e. g. axis N given above).

About a line passing through a vertex and center of opposite face, rotations by angles $2\pi/3$ and $4\pi/3$ map the tetrahedron onto itself as does the identity transformation. For an axis of the second kind, rotation by π radians is the only non-trivial symmetry of our solid. There are four such axes of the first type corresponding to the four vertices of a tetrahedron, and three of the second kind. Thus, there are $(4 \times 2 + 3 \times 1 =)$ 11 rotations, exclusive of the identity transformation, which are symmetries of the solid.

As in the case of the cube, we can assure ourselves here too that any transformation of the tetrahedron obtained by composition of some of these 12 rotations is again a rotation. Further, rotations about axes of type M by angles $2\pi/3$ and $4\pi/3$ are both of order three whereas rotations about any axis of type N by π are of order two. Hence, there are 12 elements in the group of rotational symmetries of the tetrahedron including $(3 \times 1 =)$ 3 elements of order two and $(4 \times 2 =)$ 8 elements of order three. This group is also known as the *tetrahedral group* T of rotations.

3.3 Group of rotational symmetries of an octahedron

We earlier saw in § 2 that an octahedron has 6 vertices, 12 edges and 8 faces. Its axes of rotation, about whom specific rotations take the solid to itself, are again of three kinds: those passing via

i) a pair of opposite vertices (e. g. axis L in Figure 12),

ii) a pair of mid-points of opposite edges (e. g. axis M above), and

iii) a pair of centres of opposite faces (e. g. axis N in the figure).

For any axis of the first type, rotations by angles $\pi/2$, π and $3\pi/2$ are symmetries of the octahedron alongside the identity transformation. Insofar as axis of the second kind are concerned, rotation by π radians is the only non-trivial transformation which is a symmetry of our solid. About axes of the third kind, rotations by angles $2\pi/3$ and $4\pi/3$ are symmetries of an octahedron apart from the identity transformation of course. Now, we have 3 different axes of the first, 6 of the second and 4 axes of the third type. All in all, there are $(3\times3+6\times1+4\times2 =)$ 23 such non-trivial rotations which are symmetries of an octahedron and together with the identity transformation, they constitute the group of rotational symmetries of an octahedron.

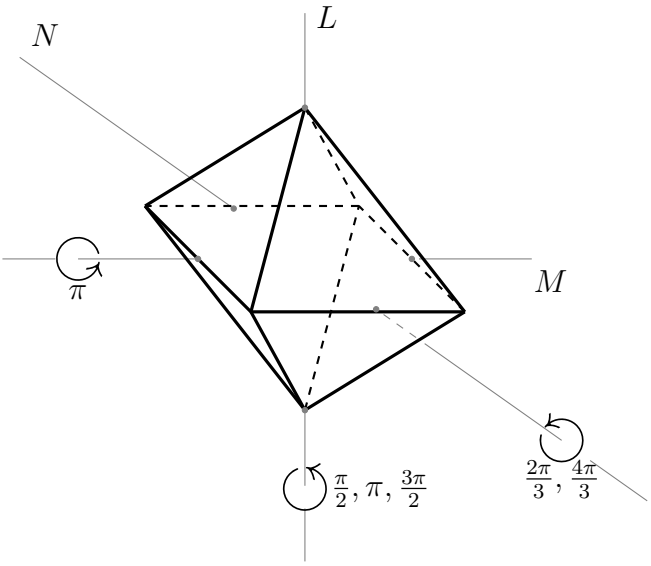


Figure 12: Rotational symmetries of an octahedron.

Of these, rotations about axes of the first and second type by an angle of π are of order two each whereas rotations about axes of the third type by angles $2\pi/3,\ 4\pi/3$ are of order three. Similarly, rotations about axes of the first kind by angles $\pi/2$ and $3\pi/2$ are of order four each. Hence, there are a total of 24 elements in the group of rotational symmetries of an octahedron of which $(6 \times 1 + 3 \times 1 =)$ 9 elements are of order two, $(4 \times 2 =)$ 8 elements of order three, and $(3 \times 2 =)$ 6 elements of order four besides the trivial element – having a striking similarity with the structural organization of the group of rotational symmetries of the cube. As a matter of fact, it can be shown that these groups are actually the same. We call this to be the *octahedral group* O of rotations.

3.4 Group of rotational symmetries of a dodecahedron

As seen before, a dodecahedron has 20 vertices, 30 edges and 12 faces. It has all the three types of axes of rotation as does an octahedron (see §§ 3.3) about whom certain rotations are symmetries of the dodecahedron. We now have $10, 15$ and 6 axes of the first, second and third type, respectively. Here, rotations about axes of the first type by angles $2\pi/3$ and $4\pi/3$ are non-trivial symmetries whereas for axes of the second type, rotation by π is the only transformation, other than the identity transformation, which is a symmetry of our solid. About axes of the third type, rotations by angles $2\pi/5,\ 4\pi/5,\ 6\pi/5$ and $8\pi/5$ are non-trivial symmetries of a dodecahedron. Hence there are $(10 \times 2 + 15 \times 1 + 6 \times 4 =)$ 59 such rotations of $\mathbb{R}^3$, which are the symmetries of our solid besides the identity transformation. Arguing similarly as in earlier cases, this set of 60 rotational symmetries of a dodecahedron form a group. Among these rotations, non-trivial rotations about axes of the first, second and third type are of order three, two

and five, respectively. Thus, there are 60 elements in the group of rotational symmetries of a dodecahedron among which $(15 \times 1 =)$ 15 elements are of order two, $(10 \times 2 =)$ 20 elements of order three and $(6 \times 4 =)$ 24 elements of order five.

3.5 Group of rotational symmetries of an icosahedron

It has been mentioned in § 2 that an icosahedron has 12 vertices, 30 edges and 20 faces. We also have all the three types of axes of rotation as for an octahedron or a dodecahedron about whom specific rotations are symmetries of an icosahedron. As can be seen in Figure 6, there are $6, 15$ and 10 axes of the first, second and third type, respectively. Moreover, rotations about axes of the first type by angles $2\pi/5$, $4\pi/5$, $6\pi/5$, and $8\pi/5$ are symmetries of our solid besides the identity transformation. About an axis of the second type, rotation by an angle of π is the only non-trivial transformation which is a symmetry of an icosahedron. Rotations about axes of the third type by $2\pi/3$ and $4\pi/3$ are symmetries of an icosahedron of order three each.

4 Uniqueness of Platonic Solids

The last Proposition of Book 13 of *Elements* classifies all regular polyhedra.

Theorem 2. There are exactly five regular polyhedra.

In order to circumvent Euclid's involved proof given in the *Elements*, we present a group-theoretic argument. It employs a rather famous formula due to Descartes (1639) which was rediscovered and popularized by Euler (1751). Later on, Cauchy gave a very clever proof which we sketch below. A *convex polyhedron* in $\mathbb{R}^n$ is the convex hull of a finite set of points in $\mathbb{R}^n$.

Lemma 4.1 (Euler's formula). Let P be a convex polyhedron in $\mathbb{R}^3$ with v vertices, e edges and f faces. Then, we have

$$v - e + f \;=\; 2.$$

Sketch of Proof. Remove a face (say the top one) of the polyhedron. It suffices to prove that $v-e+f = 1$ for our polyhedron with one face removed, because we can see that the removed face shares all its vertices and edges with the remaining entity. Flatten out this face-deficient polyhedron onto the plane (see Part (iii) of Figure 13). Next, triangulate this newly obtained planar image by drawing sufficiently many diagonals dividing each

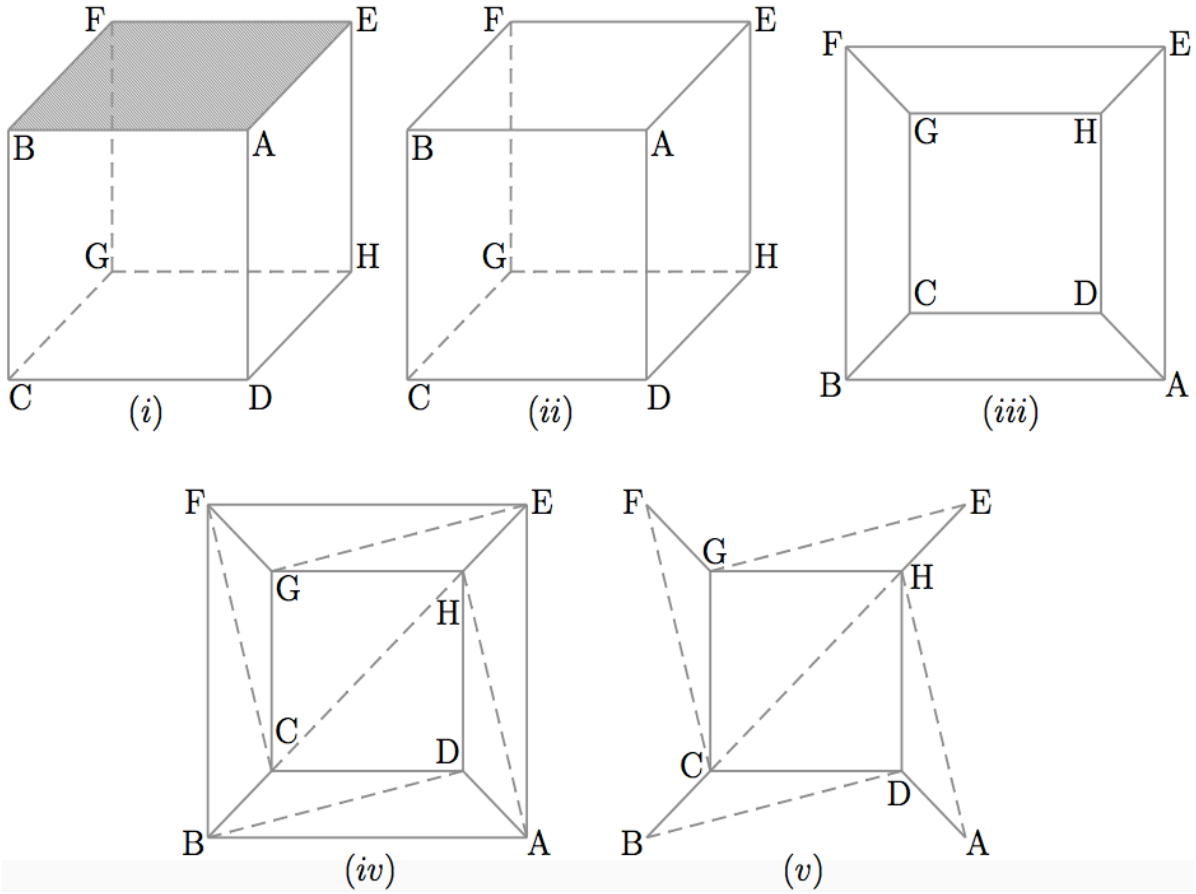


Figure 13: Method of triangulation.

polygon into triangles (see Part (iv) of Figure 13). Each diagonal adds one edge and one face, thus keeping the value $v - e + f$ unchanged.

We now start removing the triangles from the boundary (Part (v) of Figure 13). The triangles on the boundary have either one or two edges on the boundary. If the triangle being removed has only one edge on the boundary, then its removal decreases one edge and one face, thus leaving the value $v - e + f$ unchanged. If the triangle being removed has two edges common with the boundary, then its removal decreases e by two, f by one and v by one, once again keeping the value $v - e + f$ invariant. On continuing in this manner iteratively, we will be ultimately left with just one triangle which clearly satisfies $v - e + f = 1$. ■

Proof of Theorem 2. If a regular polygon has q regular p-gons meeting at each of its vertices, we will say that it is a (p, q)-regular polyhedron. For such a polyhedron, we have agreed before in § 3 that

$$qv = 2e = pf. \tag{6}$$

This implies

$$\frac{v}{1/q} = \frac{e}{1/2} = \frac{f}{1/p} = \frac{v-e+f}{1/q-1/2+1/p} = \frac{2}{1/q-1/2+1/p} = \frac{4pq}{4-(p-2)(q-2)}.$$

In particular, $(p-2)(q-2) < 4$. Thus, the only possible (p,q)-regular polyhedra are for values of the pair (p,q) corresponding to

$$(3,3),\ (3,4),\ (3,5),\ (4,3) \text{ or } (5,3).$$

Above five regular polyhedra correspond to the tetrahedron, the octahedron, the icosahedron, the cube and the dodecahedron, respectively. ■

5 Finite subgroups of SO_3 and Platonic solids

Our aim in this last section is to prove that finite subgroups of the special orthogonal group SO_3 are precisely the groups of symmetries of Platonic solids and those of regular polygons.

Theorem 3. Every finite subgroup G of SO_3 is isomorphic to one amongst

C_k : cyclic group of rotations by integer multiples of $2\pi/k$ about a line ℓ,

D_k : the dihedral group of order $2k$ of symmetries of a regular $k-$gon,

T : the tetrahedral group of 12 rotational symmetries of a tetrahedron,

O : the octahedral group of 24 rotations of a cube/octahedron, or

I : the icosahedral group of 60 rotations of an icosahedron/dodecahedron.

The proof of this theorem will require some more groundwork from us. Let G be a subgroup of SO_3 of order N. Every non-trivial element $g \in G$ is a rotation about some unique line ℓ. Hence, g fixes the two points of intersection of the line ℓ with the unit sphere $S^2 \subset \mathbf{R}^3$. We call these two points to be the *poles* of g. As an example, consider the group G of rotational symmetries of any regular polyhedron Δ. Then, the points on S^2 which are radial projections of the vertices, the mid-points of edges and the centers of faces of Δ will be poles for the corresponding elements of G.

Lemma 5.1. The subgroup G acts on the sphere so as to carry the set P of all poles to itself, i. e., G acts on P.

Proof. Let p be a pole of some $g\ (\neq \mathrm{id}_3\,) \in G$. We need to show that hp is a pole for every $h \in G$. In other words, hp is fixed by some $g'\ (\neq \mathrm{id}_3\,) \in G$. The group element hgh^{-1} fits the bill as

$$hgh^{-1}(hp) = hgp = hp,$$

where $hgh^{-1} \neq \mathrm{id}_3$ since $g \neq \mathrm{id}_3$. ■

We will now be making inferences about the group G by counting the number of poles. One might rush to claim that the total number of poles is $2N-2$, two each for every element of G except for the identity. This is, however, incorrect because it might be the case that two or more non-trivial elements of G share a point as one of their poles. The stabilizer of any pole p is the cyclic subgroup of G of all rotations by integer multiples of some angle θ about the line ℓ joining the origin to the pole p. If $|\,G_p\,| = r_p$, then $\theta = 2\pi/r_p$. Counting formula (Proposition 3.3) then gives us

$$|\,G_p\,| \cdot |\,Gp\,| \;=\; |\,G\,|.$$

For the sake of convenience, we write this equation as

$$r_p \cdot n_p = N,$$

where n_p is the number of poles in the orbit Gp of p and there are $(r_p - 1)$ elements with p as one of their poles. On the other hand, every $g \in G$ excluding the identity element has 2 poles. We, therefore, have

$$\sum_{p\,\in\,P} (r_p - 1) \;=\; 2N - 2. \tag{7}$$

If p and q are in same orbit, then $|\,G_p\,| = |\,G_q\,|$ because $Gp = Gq$ and

$$|\,G_p\,| \cdot |\,Gp\,| \;=\; |\,G_q\,| \cdot |\,Gq\,| \;=\; |\,G\,|.$$

On clubbing the terms on the left side of (7) corresponding to poles in the same orbit, and there are n_p such terms for the orbit of p, we get

$$\sum_{j} n_j(r_j - 1) \;=\; 2N - 2,$$

where the index j runs over the different orbits $Gp_1, Gp_2, Gp_3, \ldots,$ n_j equals $|\,Gp_j\,|$ and $r_j = |\,G_p\,|$ for any $p \in Gp_j$. Dividing both sides by $N = n_j r_j$,

$$\sum_{j} \left(1 - \frac{1}{r_j}\right) \;=\; 2 - \frac{2}{N}. \tag{8}$$

The above formula turns out to be the dark horse in the proof of Theorem 3. Each summand on the left side of (8) satisfies

$$1 - \frac{1}{r_j} \;\geq\; \frac{1}{2}$$

as $r_j = |\, G_p \,|$ is the order of the stabilizer G_p of some pole p and the stabilizer subgroup of any pole must have at least one non-trivial element. On the other hand, the right side of (8) is at most 2 with $N = |\, G \,|$ being positive. It follows that there can be a maximum of three pairwise disjoint orbits for the action of the roational symmetry group G of a regular polyhedron Δ on the set of poles P. We will now head to enlist all the three possible cases:

Only one orbit This implies $2 - \frac{2}{N} = 1 - \frac{1}{r}$ for some r. This is impossible since $2 - 2/N \geq 1$ for any non-trivial group G while $1 - \frac{1}{r} < 1$ for all $r > 0$. Hence, no solution exists in this case.

Two orbits We have here that

$$\big(1 - \frac{1}{r_1}\big) + \big(1 - \frac{1}{r_2}\big) \;=\; 2 - \frac{2}{N}$$

which happens if and only if

$$\frac{1}{r_1} + \frac{1}{r_2} \;=\; \frac{2}{N}. \tag{9}$$

Next, note that $r_j \leq N$ since G_{p_j} is a subgroup of G. This will mean $1/r_j \geq 1/N$ for $j = 1, 2$ and as a consequence, the identity (9) can hold only when $r_1 = r_2 = N$. It also implies $n_1 = n_2 = 1$ leading to two orbits having a single element each. Thus, there are two poles p and q both fixed by every element of the group G. Otherwise said, each element of G is a rotation about the same common line ℓ joining antipodal poles p and q. The conclusion is that G is the cyclic group C_N of rotations by integer multiples of $2\pi/N$ about the axis passing through p and q.

Three orbits This is the most important of all the three cases and the one which gives us all the remaining finite subgroups in Theorem 3. For us, formula (8) now reduces to

$$\frac{1}{r_1} + \frac{1}{r_2} + \frac{1}{r_3} - 1 \;=\; \frac{2}{N}. \tag{10}$$

Without loss of generality, we make take the r_j's to be arranged in ascending order $r_1 \leq r_2 \leq r_3$. Then, $r_1 = 2$. Else if $r_j \geq 3$ for all j, the right side of (10) becomes ≤ 0 and this cannot give a solution to our equation. Next,

1) The first two orders r_1 and r_2 are both 2. Let $r_3 = k$ be any positive integer ≥ 2 so that $N = 2k$. Then, $n_3 = 2$ so that there exist exactly two poles in

the orbit $Gp_3 = \{p_3, q_3\}$ (say). Every element $g \in G$ fixes both p_3 and q_3 pointwise or swaps them. Hence, we have that every element of G is either a rotation about the line ℓ joining p_3 and q_3 or a rotation by an angle of π about some line ℓ' orthogonal to ℓ. It is clear that our group G is the dihedral group D_k of symmetries of a regular k-gon.

2) Suppose $r_2 \geq 3$. We must first state that $r_2 \geq 4$ is not possible as that would make $r_3 \geq 4$ and the right side of (8) to be ≤ 0. On similar grounds, the possibility of $r_1 = 2, r_2 = 3$ and $r_3 \geq 6$ is ruled out. Thus, only three cases remain be analyzed further:

   i) $(r_1, r_2, r_3) = (2, 3, 3),\ N = 12;$

   ii) $(r_1, r_2, r_3) = (2, 3, 4),\ N = 24;$ and

   iii) $(r_1, r_2, r_3) = (2, 3, 5),\ N = 60.$

   What follows is a brief analysis of the three cases written above.

For case (i), we get the corresponding triple $(n_1, n_2, n_3) = (6, 4, 4)$. The poles in the second orbit correspond to the vertices of a tetrahedron Δ. Clearly, n_1 is the number of edges brought into the picture by radial projections of their respective mid-points on the unit sphere whereas n_3 is the number of faces of Δ. The group G is the tetrahedral group of rotations T.

For case (ii), we have $(n_1, n_2, n_3) = (12, 8, 6)$. The poles in the orbit Gp_2 are obtained by taking radial projections of the vertices of a cube on the unit sphere while those in the third orbit correspond to the mid-points of its faces. Hence, $G = O$.

For case (iii), the sizes of the orbits are given by

$$(n_1, n_2, n_3) = (30, 20, 12).$$

The poles in the second orbit correspond to the vertices of an icosahedron while those in Gp_3 to that of a dodecahedron. Thus, we have $G = I$ being the icosahedral group of rotations.

One might still wonder at this stage that since all the poles in an orbit are evenly spaced on the unit sphere S^2, there can be many more regular polyhedra. However, this is not quite the case, because one can actually make the solids and see that except for the orbits stated here, poles in any other configuration do not arise from a regular polyhedron. Our claim is also supported by the uniqueness of Platonic solids proved before in § 4.

Acknowledgments

The first author thanks the DST, Govt. of India, for financial support under the KVPY fellowship scheme during the period 2005 – 11. He is also grateful to IIT Bombay for its hospitality during the summers of 2006 and 2007. This article is an outgrowth of a talk delivered by the second author at a workshop organized by HBCSE, Bombay in the Fall of 2003.

Bibliography

[1] G J Allman, *Greek Geometry from Thales to Euclid*, Hodges, Figgis, & Company, 1889.

[2] M A Armstrong, *Groups and Symmetry*, Springer-Verlag, New York, 1988.

[3] M Artin, *Algebra*, Prentice-Hall of India, New Delhi, 1991.

[4] J Baez, *Fool's Gold*, 2011. https://math.ucr.edu/home/baez/golden.html

[5] P R Cromwell, *Polyhedra*, Cambridge University Press, Cambridge, 1997.

[6] Euclid, T L Heith, *The Thirteen Books of Euclid's Elements, Vol. 3 (Books X-XIII)*, Dover Publications; Second Edition Unabridged, Cambridge, 1956.

[7] J V Field, *Kepler's Geometrical Cosmology*, Bloomsbury Publishing, 2013.

[8] E Haeckel, *Report on the Radiolaria collected by H. M. S. Challenger during the years 1873-76*, Her Majesty's Stationery Office, Edinburgh, 1887.

[9] J v Hoboken, *Platonic solids, binary polyhedral groups, Kleinian singularities and Lie algebras of type A, D, E*, Masters' Thesis, Univ. of Amsterdam, 2002.

[10] H W Kroto, J R Heath, S C O'Brien, R F Curl and R E Smalley, C_{60}: Buckminsterfullerene, *Nature* **318**, 162–163 (1985).

[11] "Tetrahedron", *2020 Journal Citation Reports*, Web of Science (Science ed.), Thomson Reuters, 2021.

[12] A F Wells, *Structural Inorganic Chemistry*, Oxford: Clarendon Press, 1984.

[13] H Weyl, *Symmetry*, Princeton University Press, New Jersey, pp.73-77, 1982.

[14] D Zeyl and B Sattler, Plato's *Timaeus*, *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University, 2023.

7 Historical Roots of Calculus – Part 3

Gerardus Mercator, Edward Wright, James Gregory, Isaac Barrow

Shailesh Shirali

Sahyadri School KFI
Rajgurunagar, Khed
Pune – 410513

Email: shailesh.shirali@gmail.com

In Part 1 of this article we considered some of the early approaches to finding tangents to curves; in particular, the work of Roberval, Descartes and Fermat. Then, in Part 2, we looked at some early approaches to finding area; in particular, the work of Archimedes and his "method of exhaustion." Now in Part 3 we study ideas that emerge from the work of individuals such as the cartographer Gerardus Mercator and the mathematicians Edward Wright, John Napier, Henry Bond, James Gregory, and Isaac Barrow.

Remarkably, the story has its origins in cartography. Let us start by explaining what is an *equal area cylindrical map projection*. Imagine a cylinder into which the Earth fits snug, touching it along the equator, so the axis of the cylinder coincides with the line through the North Pole and South Pole. Now let radial lines be drawn perpendicular to the central axis, drawn outward till they intersect the circumscribing cylinder. The image of each point on the Earth is the point where the radial line intersects the cylinder. Thus the map of the continents and the oceans is projected onto the cylindrical surface. (See Figure 1.) The cylinder may now be cut open along a line parallel to the axis, thus yielding a rectangular map of the Earth. This projection has the following features.

1. Circles of constant latitude are projected to straight lines parallel to the midline of the rectangle (i.e., the line corresponding to the equator). Moreover, they all have the same length in the projected map. In actuality, parallels of latitude naturally get smaller as we approach the polar regions. It follows that there is a steadily increasing distortion in the East-West direction as we approach the polar regions.

2. Meridian lines are projected to straight lines perpendicular to the line corresponding to the equator.

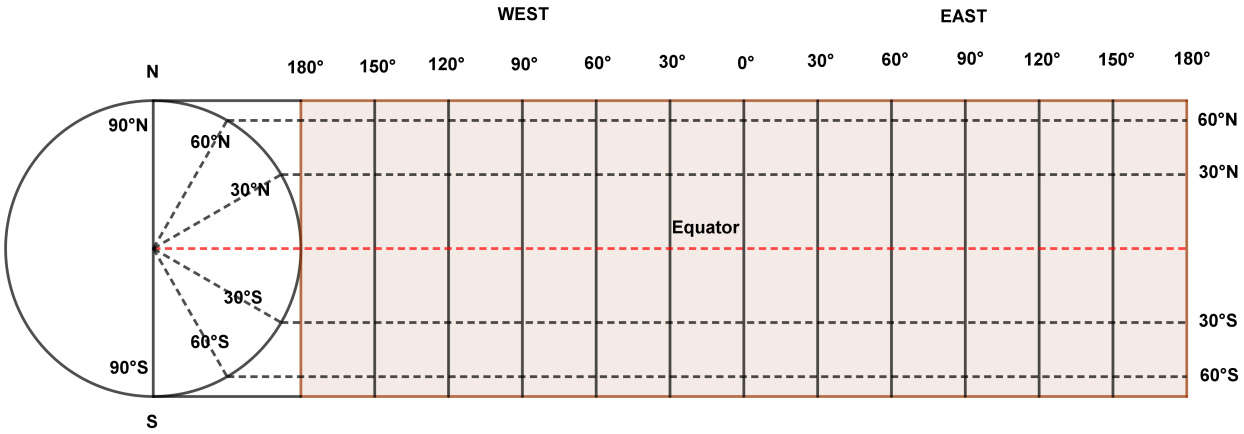


Figure 1: Side view of the equal area cylindrical projection

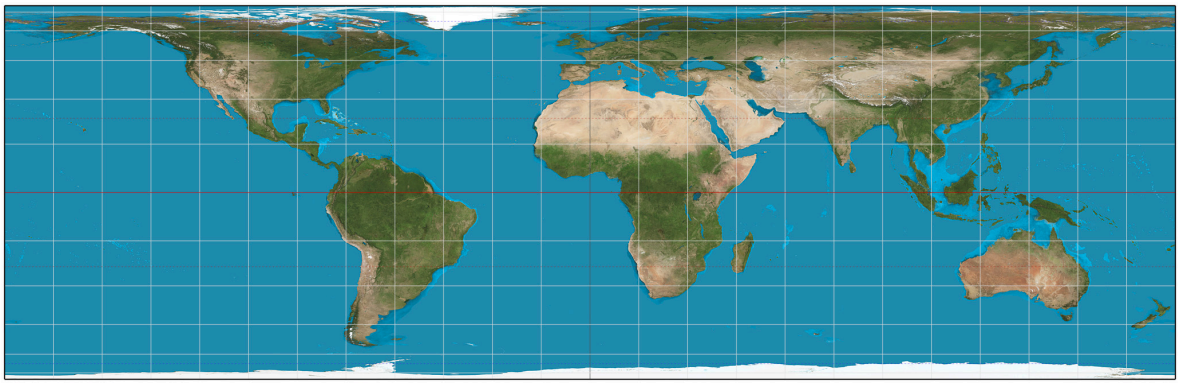


Figure 2: A world map that uses the cylindrical equal area projection. Source: https://en.wikipedia.org/wiki/Cylindrical_equal-area_projection

3. As noted above, there is a steadily increasing East-West dilation as we approach the polar regions. At the same time there is a North-South contraction. It can be shown that *these two effects are exact reciprocals of one another* (this is quite easy to show), which means that areas remain invariant. Therefore, this is an *equal area projection*: the image of each region on the rectangular sheet is the same as the area of the corresponding region on the Earth's surface. (Historically, this is the way that Archimedes proved the surface area of a sphere of radius R is $4\pi R^2$. From the above observation, this follows immediately, because the surface area of the enclosing cylinder is $2\pi \cdot R \cdot 2R = 4\pi R^2$.)

A typical world map that uses the cylindrical equal area projection is shown in Figure 2. Note the distortion (the EW dilation and the NS contraction) in the extreme northern and extreme southern regions. This is an inevitable feature of the cylindrical equal area projection.

In 1569 Mercator hit upon a cylindrical map projection that became known as the *Mercator projection*. It had the following features:

1. It too is a rectangular map projection: the parallels of latitude are straight lines parallel to the equator, while the meridians are lines perpendicular to the equator.

2. However, the spacing of the parallels of latitude (i.e., along the meridians) is fixed in such a way that *the North-South distortion is exactly the same as the East-West distortion*. This means, obviously, that locally, directions are not distorted. An important consequence of this is that a navigator using such a map would be able to set a constant direction of motion simply by drawing a straight line on the map. (Navigators refer to such lines as *rhumb lines*.) Such a map would naturally have great utility for navigators. This manner of drawing a map is known as the Mercator projection.

3. Like the cylindrical equilateral projection, the Mercator projection too results in distortion at the polar regions — but in a very different manner. Shapes are not distorted, *but areas are grossly distorted*. As a result, polar regions look much larger on such a map than they actually are.

A typical world map that uses the Mercator projection is shown in Figure 3. Note the distortion in area in the extreme northern regions and extreme southern regions; Greenland looks larger than Africa, and very much larger than Australia! This is an inevitable feature of the Mercator projection.

Now comes the really interesting mathematical component of this map projection. How do we work out the spacing along the meridians so that the parallels of latitude are placed correctly? Since the spacing is not uniform, this constitutes a non-trivial problem.

In the discussion below, we assume throughout that the Earth is a perfect sphere of radius R.

Consider the parallel at latitude θ (see Figure 4). It is a circle, and simple trigonometry tells us that the radius of the circle is $R\cos\theta$. Therefore the circumference of this parallel of latitude is $2\pi R\cos\theta$. But on the Mercator map this is shown as a horizontal line on the rectangle having the same length as the equator (length $2R$). Therefore at latitude θ, there is an East-West stretch by a factor of $1/\cos\theta$, i.e., $\sec\theta$. Therefore, by the nature of the Mercator projection, at this latitude we must have a North-South stretch by $\sec\theta$. This means that the gap between two parallels of latitude separated by an infinitesimal amount $\mathrm{d}\theta$ must be given by $\sec\theta\ \mathrm{d}\theta$. And this in turn implies that the gap between the parallel of latitude α and the equator is equal to

$$\int_0^\alpha \sec\theta\ \mathrm{d}\theta. \tag{1}$$

Figure 3: A world map that uses the Mercator projection. Source: https://en.wikipedia.org/wiki/Mercator_projection#

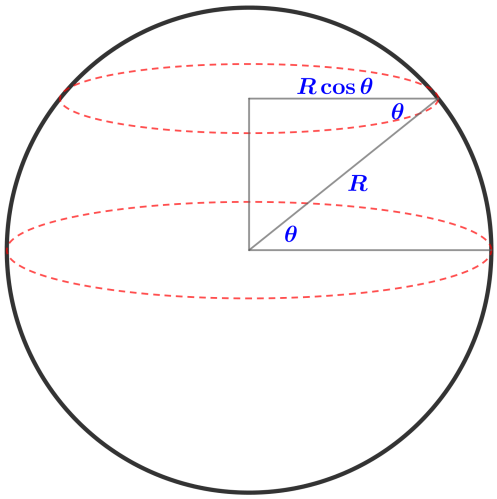


Figure 4: Geometry of the Earth at the parallel of latitude θ.

Therefore to accurately draw a map using the Mercator projection, we need to be able to evaluate the integral (1).

Now recall the year in which all these developments are taking place — 1569. The integral calculus developed by Newton and Leibniz lies a full century in the future! So how does Mercator evaluate the value of (1)? (It goes without saying that Mercator could not have used the notation shown in (1). We have used modern notation purely for convenience and simplicity.) The short answer is that we have no idea at all on what basis he drew the Mercator projection! He simply presented his table, leaving no record of his thinking in the matter. We do not even know whether he argued in the way that we have described. (Most likely he did *not* think on the above lines. His interest was mainly in preserving the straightness of directions. Attempts have been made in recent years to reconstruct his thinking, using only geometry and geometrical instruments; e.g., see [6].)

In the late 1580s a serious attempt was made by Thomas Harriot to explain Mercator's projection. But he did not publish his work, and his results did not enter into public circulation.

Then in 1599 a mathematician named Edward Wright managed to give an explanation of the table by presenting (in essence) the derivation we have given above. (As already noted, the formal machinery and symbolism of calculus did not exist at the time. Though methods to find the slope of the tangent to a curve were available, they were ad hoc and applicable only to certain kinds of curves. Similarly, methods to find the area under a curve were available, but they too were ad hoc and applicable only to certain kinds of curves.. Most importantly, the connection between these two problems had

not been discovered. That all-important development — the fundamental theorem of calculus — had to wait for Newton and Leibniz.) To evaluate the integral, he used a *finite summation method* (equivalent to what we now call 'Riemann summation'). He had to use a published table of values of the secant function. Accurate tables of all the trigonometric functions had been computed by that time and were available for acute angles at intervals of $1'$. It needs to be mentioned here that all this work lies prior to the development of logarithms by John Napier (in 1614). One can only imagine the amount of manual labour that Edward Wright had to put in computing the table!

In the years following the invention of logarithms, tables were published of logarithmic sines and logarithmic tangents; these were regarded as indispensable for navigation, surveying and astronomy. The indefatigable Edward Wright was involved in this endeavour as well (see [7])!

Then in 1640 there occurred a truly remarkable episode of 'pattern spotting' of a kind that would make any data scientist proud. A mathematics teacher, Henry Bond, happened to compare Wright's tables with a table of logarithmic tangents and noticed a remarkable agreement between the published values of (1) and the values of $\ln\tan\left(\frac{\theta}{2}+\frac{\pi}{4}\right)$. This led him to *conjecture* — purely on the basis of numerical experimentation! — that

$$\int_0^{\alpha} \sec\theta \, \mathrm{d}\theta = \ln\tan\left(\frac{\alpha}{2}+\frac{\pi}{4}\right). \tag{2}$$

This is an astonishing development. No explanation for the equality came forth, and the conjecture remained an unsolved problem. It led to wonderment and puzzlement in the mid-17th century, in much the same way as the Basel problem did in the 1720s and 1730s till it was resolved by Euler. Interestingly, the problem was noticed by Isaac Newton in the 1660s.

The resolution came in 1666-1670 in the work of Nicholas Mercator (no relation, despite the name, to Gerardus Mercator), John Wallis, James Gregory, and Isaac Barrow. See [8] and [9] for details on this wonderful episode in the early history of calculus. The story goes roughly as follows.

Mathematicians of the time were aware that the area under a rectangular hyperbola is given by the logarithmic function; specifically that

$$\int_0^{t} \frac{\mathrm{d}x}{1+x} = \ln(1+t). \tag{3}$$

(Once again, we have expressed this relation using modern symbolism.) If we now expand $\frac{1}{1+x}$ as a geometric series,

$$\frac{1}{1+x} = 1 - x + x^2 - x^3 + \cdots \tag{4}$$

(of course we must have $|x| < 1$ but such considerations lie far in that future!) and do term-by-term integration, then we arrive at the result

$$\ln(1+t) = t - \frac{t^2}{2} + \frac{t^3}{3} + \frac{t^4}{4} - \cdots . \tag{5}$$

The substitution $t \mapsto -t$ leads to

$$\ln(1-t) = -t - \frac{t^2}{2} - \frac{t^3}{3} - \frac{t^4}{4} - \cdots , \tag{6}$$

and now by subtraction:

$$\frac{1}{2}\ln\left(\frac{1+t}{1-t}\right) = t + \frac{t^3}{3} + \frac{t^5}{5} + \cdots . \tag{7}$$

This result was known. Now consider the integral $\int \sec x \, \mathrm{d}x$. We write:

$$\begin{aligned}
\int \sec x \, \mathrm{d}x &= \int \sec^2 x \cdot \cos x \, \mathrm{d}x = \int \sec^2 x \, \mathrm{d}(\sin x) \\
&= \int \frac{1}{1-\sin^2 x} \, \mathrm{d}(\sin x) \\
&= \int \left(1 + s^2 + s^4 + s^6 + \cdots\right) \, \mathrm{d}s \qquad\qquad (\text{where } s = \sin x) \\
&= s + \frac{s^3}{3} + \frac{s^5}{5} + \frac{s^7}{7} + \cdots \\
&= \frac{1}{2}\ln\left|\frac{1+s}{1-s}\right| .
\end{aligned} \tag{8}$$

This yields:

$$\int_0^\theta \sec x \, \mathrm{d}x = \frac{1}{2}\ln\left|\frac{1+\sin\theta}{1-\sin\theta}\right| . \tag{9}$$

The reader will surely agree that this is quite a remarkable derivation!

Isaac Barrow came up with another method that may well be the first-ever use of the partial fractions technique for integration:

$$\begin{aligned}
\int \sec x \, \mathrm{d}x &= \int \frac{1}{\cos x} \, \mathrm{d}x = \int \frac{\cos x}{\cos^2 x} \, \mathrm{d}x = \int \frac{\cos x}{1-\sin^2 x} \, \mathrm{d}x \\
&= \int \frac{\cos x}{(1+\sin x)(1-\sin x)} \, \mathrm{d}x \\
&= \frac{1}{2}\left(\int \frac{\cos x}{1+\sin x}\mathrm{d}x + \int \frac{\cos x}{1-\sin x}\mathrm{d}x\right) \\
&= \frac{1}{2}\big(\ln|1+\sin x| - \ln|1-\sin x|\big) \\
&= \frac{1}{2}\ln\left|\frac{1+\sin x}{1-\sin x}\right| .
\end{aligned} \tag{10}$$

To round off the above discussion, we note that

$$\begin{aligned}\frac{1}{2}\ln\left|\frac{1+\sin x}{1-\sin x}\right| &= \frac{1}{2}\ln\left|\frac{1+\sin x}{1-\sin x}\cdot\frac{1+\sin x}{1+\sin x}\right| \\ &= \frac{1}{2}\ln\left|\frac{(1+\sin x)^2}{\cos^2 x}\right| \\ &= \ln\left|\frac{1+\sin x}{\cos x}\right| = \ln|\sec x+\tan x|, \end{aligned} \tag{11}$$

and also:

$$\begin{aligned}\frac{1}{2}\ln\left|\frac{1+\sin x}{1-\sin x}\right| &= \frac{1}{2}\ln\left|\frac{\sin^2\frac{x}{2}+\cos^2\frac{x}{2}+2\sin\frac{x}{2}\cos\frac{x}{2}}{\sin^2\frac{x}{2}+\cos^2\frac{x}{2}-2\sin\frac{x}{2}\cos\frac{x}{2}}\right| \\ &= \frac{1}{2}\ln\left|\frac{(\cos\frac{x}{2}+\sin\frac{x}{2})^2}{(\cos\frac{x}{2}-\sin\frac{x}{2})^2}\right| \\ &= \ln\left|\frac{1+\tan\frac{x}{2}}{1-\tan\frac{x}{2}}\right| = \ln\left|\tan\left(\frac{x}{2}+\frac{\pi}{4}\right)\right|, \end{aligned} \tag{12}$$

and we have obtained the desired results.

In forthcoming articles of this series we will study some of the work of James Gregory, Gottfried Leibniz, Colin Maclaurin and Isaac Newton.

As should be clear from the material presented above, the history of calculus is more like a tangled thicket than an upright coconut tree. So we have a lot more material to cover!

Bibliography

[1] Wikipedia, the free encyclopedia, "Cylindrical equal-area projection." https://en.wikipedia.org/wiki/Cylindrical_equal-area_projection

[2] Weisstein, Eric W. "Cylindrical Equal-Area Projection." From MathWorld–A Wolfram Web Resource. https://mathworld.wolfram.com/CylindricalEqual-AreaProjection.html

[3] Wikipedia, the free encyclopedia, "Mercator projection." https://en.wikipedia.org/wiki/Mercator_projection#

[4] Weisstein, Eric W. "Mercator Projection." From MathWorld–A Wolfram Web Resource. https://mathworld.wolfram.com/MercatorProjection.html

[5] Robert Israel, "Mercator's Projection." From https://personal.math.ubc.ca/~israel/m103/mercator/mercator.html

[6] Gyula PÃ¡pay, “Mercator’s Geometric Method in the Construction of His Projection from 1569”, KN - Journal of Cartography and Geographic Information (2022) 72:261â€“267, from https://doi.org/10.1007/s42489-022-00115-5

[7] John Russell Napier and Edward Wright, “A Description of the Admirable Table of Logarithmes, With a declaration of the most plentiful, easy and speedy use thereof in both kindes of Trigonometrie, as also in all Mathematicall calculations.” From https://openlibrary.org/works/OL15676445W/A_Description_of_the_Admirable_Table_of_Logarithmes

[8] H. S. Carslaw, “The Story of Mercator’s Map.: A Chapter in the History of Mathematics”, *The Mathematical Gazette*, Jan., 1924, Vol. 12, No. 168 (Jan., 1924), pp. 1-7, from https://www.jstor.org/stable/3603395

[9] V. Frederick Rickey and Philip M. Tuchinsky, “An Application of Geography to Mathematics: History of the Integral of the Secant”, *Mathematics Magazine*, May, 1980, Vol. 53, No. 3 (May, 1980), pp. 162-166, from https://www.jstor.org/stable/2690106

8 Historical Roots of Calculus – Part 4

Gottfried Leibniz

Shailesh Shirali

Sahyadri School KFI
Rajgurunagar, Khed
Pune – 410513

Email: shailesh.shirali@gmail.com

In Part 1 of this article we considered some early approaches to finding tangents to curves: the work of Roberval, Descartes and Fermat. In Part 2, we looked at some early approaches to finding area; in particular, the work of Archimedes and his "method of exhaustion." Then, in Part 3, we studied an episode associated with the cartographer Gerardus Mercator; namely, the quadrature of the secant function. Now, in Part 4, we look at how Gottfried Leibniz discovered the series for π that bears his name (and that of James Gregory, who discovered the series independently and in another manner). (The same series was discovered earlier by the great Indian mathematician Madhava. We shall describe that work later in this series.)

Leibniz's contribution

To anyone interested in the history of mathematics, and the history of calculus in particular, we warmly recommend watching the documentary [1]. In this short video, the presenter Jeremy Gray takes the viewer through two historically important phases in the history of calculus — the work done independently by Newton and Leibniz — by displaying the actual notebooks maintained by them. Interestingly (and luckily for us), both individuals were in the practice of writing detailed notes in their notebooks, thereby preserving a wonderful record for history. As Gray remarks,

> *What's really fascinating is that the original writings recording the discoveries of both of these men are preserved. In the university library in Cambridge we have the notebooks that Newton kept between 1665 and 1667, and in*

> *Hannover, Leibniz's notes from 1676 are preserved as well. They provide a fascinating glimpse into the process of mathematical discovery that both of these men used.*

In Leibniz's case we see the ideas developing in his mind on a real-time basis: the way he develops the symbols for the derivative and the integral; one almost hears him talking to himself, working out the reasons for the choice of symbols. He finalised all the notation in a two-week period of intense mental activity from late October to early November 1675, and this is the notation still in use today, many centuries later! Gray remarks, in the same documentary:

> *Our story of Leibniz begins in London in 1673. In January of that year he presented to the Royal Society a calculating machine he had invented incorporating several novel features. He was elected a fellow of the Royal Society on the strength of this invention. All his life Leibniz worked to mechanize all reasoning processes. He wanted to formalize the rules of logic so that any logical argument or mathematical proof could be produced by machine. Leibniz saw the calculating machine he took to the Royal Society as just the first stage in the development of such a logical machine, and all his life he worked to improve his calculating machines. …These ideas of Leibniz are important since they do much to explain his way of working and the particular care he went to to invent a powerful and flexible notation for his calculus.*

The last line in this quote may explain why we are still using Leibniz's notation today.

Though their final results are equivalent, Newton and Leibniz approached the problem in very different ways, and this explains why the notation they developed was not the same. Consider a curve $\mathcal{C}$ in the (x, y)-coordinate plane. To work out a way for finding the tangent to $\mathcal{C}$ (this was the standard terminology of the time in referring to the slope of a curve), Newton thought in terms of *motion*. He visualised a pair of points moving along the two axes. The x- and y-coordinates of these two points define some point P on the coordinate plane, and this point traces out a curve $\mathcal{C}$. Each coordinate has an instantaneous velocity which varies with time. Newton referred to these as *fluxions* and denoted them by $\dot{x}$, $\dot{y}$, respectively. The desired slope is of $\mathcal{C}$ at P then $\dot{y}/\dot{x}$.

Leibniz, on the other hand, thought exclusively in terms of *infinitesimals*; as this is very close to the modern approach, we do not need to describe it here. We should add here that Newton was not unfamiliar with the notion of an infinitesimal but for various reasons preferred to use the 'fluxions' approach. Leibniz, in contrast, used the approach of infinitesimals at every stage of his work.

A result that Leibniz uses repeatedly is his *transmutation theorem* which we now describe. Let a curve $\mathcal{C}$ have equation $y = f(x)$. For an arbitrary point P on $\mathcal{C}$, we show how to

locate another point U. Then, as P moves on $\mathcal{C}$, U describes a curve $\mathcal{K}$ with its own equation. (See Figure 1.)

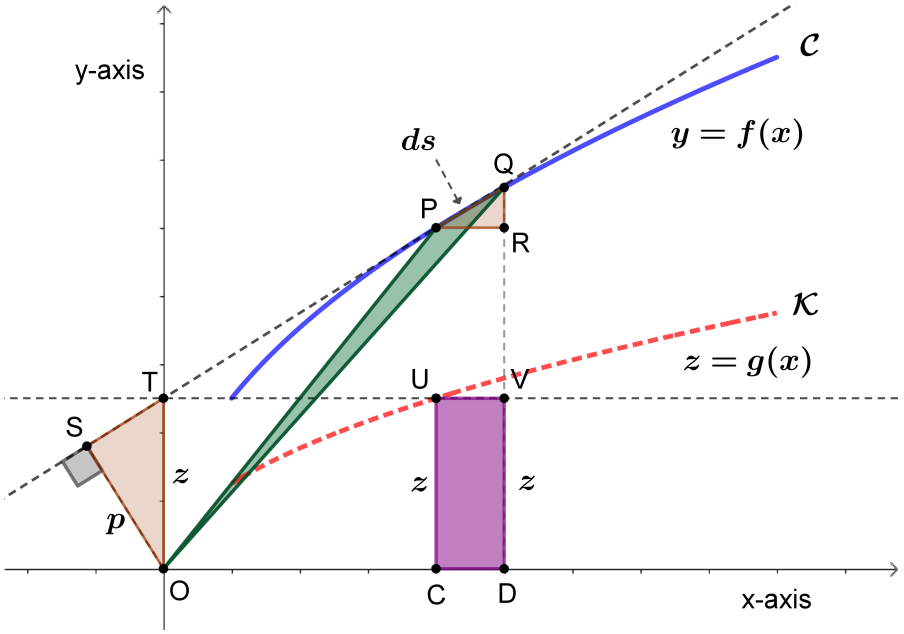


Figure 1

In Figure 1, P, Q are a pair of points on $\mathcal{C}$, very close together. Let the tangent at P to $\mathcal{C}$ cut the y-axis at T. (Note that the tangent may be regarded as coincident with line PQ.) Let the perpendicular from the origin O to the tangent cut the tangent at S. Let U be the point where the horizontal line through T cuts the vertical line through P. In the same way, point Q on the curve gives rise to point V. Let R be the point where the horizontal line through P cuts the vertical line through Q. It will be readily seen that as P moves on $\mathcal{C}$, U describes another curve $\mathcal{K}$ with its own equation. Leibniz refers to this operation as a 'transmutation.' (In Figure 1, V is seen as not quite lying on the 'transmuted' curve, but this is only because P and Q are not that close to one another.)

Leibniz now proves the following:

$$\text{Area of sector } OQP = \frac{1}{2} \cdot \text{Area of rectangle } UCDV. \tag{1}$$

The result is easy to prove. As P and Q are taken to be very close together, the arc PQ of the curve may be regarded as coincident with segment PQ. As indicated in Figure 1, let the lengths of PQ, OS and OT be ds, p and z, respectively. Note that triangle PQR is similar to triangle OTS. Hence

$$\frac{ds}{PR} = \frac{z}{p}, \qquad \therefore \; p \cdot ds = z \cdot PR. \tag{2}$$

Therefore we have:

$$\text{Area of sector } OQP = \frac{1}{2}\, p{\cdot}ds = \frac{1}{2}\, z{\cdot}PR = \frac{1}{2}\, z{\cdot}UV = \frac{1}{2}\,\text{Area of rectangle } UCDV. \quad (3)$$

Summing this result over a large number of infinitesimally small, adjacent arcs of the curve $\mathcal{C}$, we are led to the result depicted in Figure 2.

$$\text{Area of sector } OBAO = \frac{1}{2} \cdot \text{Area of the region } UU'V'VU. \quad (4)$$

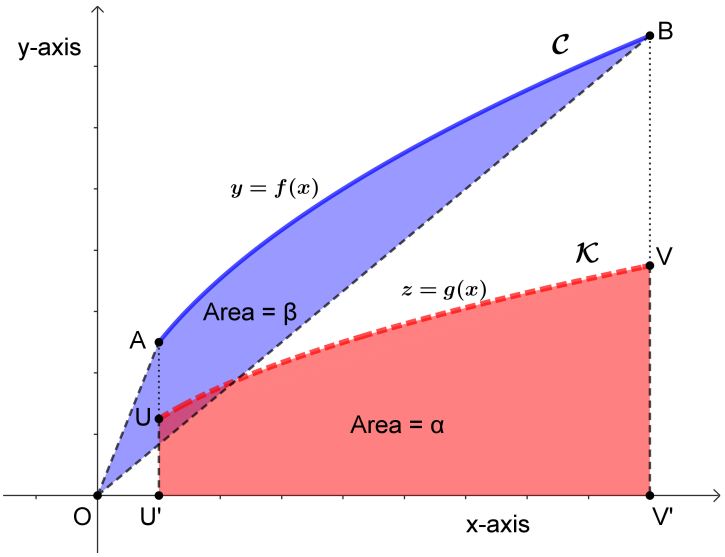


Figure 2: The areas of the two shaded regions are related as follows: $\alpha = 2\beta$

In the figure, $\mathcal{K}$ is the transmuted form of $\mathcal{C}$, and the points corresponding to A and B (on $\mathcal{C}$) are U and V (on $\mathcal{K}$), respectively.

Now let us see how Leibniz applies this to the quadrature of a circle.

Figure 3 shows a quarter circle with unit radius, centred at the point $O = (1, 0)$. It passes through the points $A(0, 0)$ and $B(1, 1)$. Consider the shaded region which is a segment of the circle; its area is

$$\frac{\pi}{4} - \frac{1}{2}. \quad (5)$$

Let us find the equation of the transmuted curve. Take a typical point P on the quarter circle. Let the tangent to the circle at P cut the y-axis at T, and let U be the point where the horizontal line through T cuts the vertical line through P. Let $P = (x, y)$ and $U = (x, z)$. Let m be the slope of the circle at P. Then we have:

$$m = \frac{PU}{TU} = \frac{y - z}{x}, \quad \text{and also} \quad m = \frac{-1}{y/(x-1)} = \frac{1 - x}{y}. \quad (6)$$

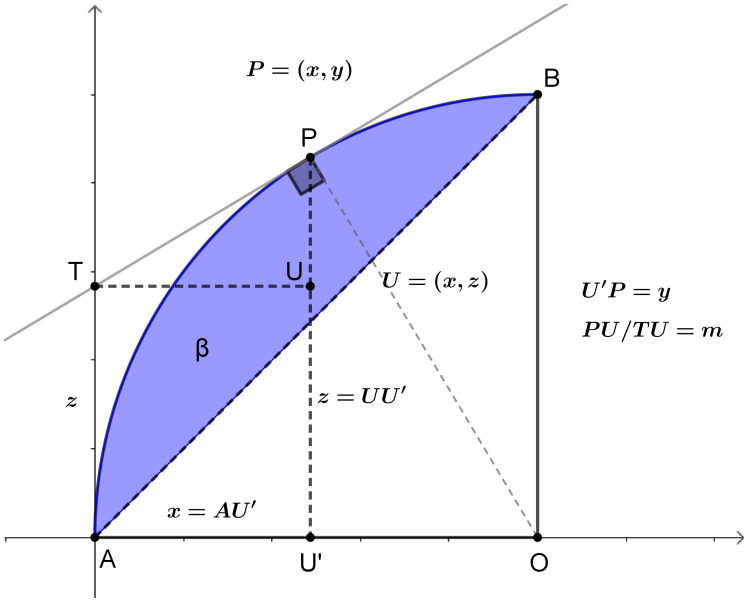


Figure 3

(The latter relation is true because the tangent is perpendicular to OP.) Hence:

$$\frac{y-z}{x} = \frac{1-x}{y}, \qquad \therefore \quad z = \frac{x}{y}. \tag{7}$$

Since the equation of the circle is $(x-1)^2 + y^2 = 1$, it follows that

$$y = \sqrt{2x - x^2}, \qquad \therefore \quad z = \frac{x}{\sqrt{2x-x^2}} = \sqrt{\frac{x}{2-x}},$$

and so,

$$z^2 = \frac{x}{2-x}, \qquad x = \frac{2z^2}{1+z^2}. \tag{8}$$

With the equation of the transmuted curve at hand, we may now sketch it accurately. Figure 4 gives a sketch of the circle and the transmuted curve $\mathcal{K}$ on the same set of axis.

Note the two shaded regions in Figure 4, with areas α and β, respectively. From Leibniz's transmutation result we know that $\beta = \alpha/2$. We have just shown that $\beta = \pi/4 - 1/2$. It follows that

$$\frac{\alpha}{2} = \frac{\pi}{4} - \frac{1}{2}. \tag{9}$$

It now remains to get an expression for α by some other line of reasoning. This will then yield an expression for π.

To find α in some other way, Leibniz 'tips' the problem over and looks at the situation 'sideways.' Figure 5 shows what he has in mind: rather than directly compute α, why

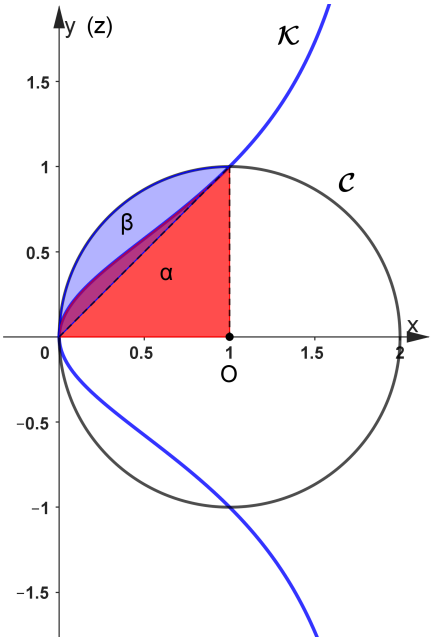


Figure 4: The equation of $\mathcal{K}$ is $z^2 = \frac{x}{2-x}$

not compute the area γ of the region that is the complement of α with respect to the unit square with vertices at $(0,0)$, $(1,0)$, $(1,1)$ and $(0,1)$? Since $\gamma + \alpha = 1$, once we get γ we immediately obtain α as well.

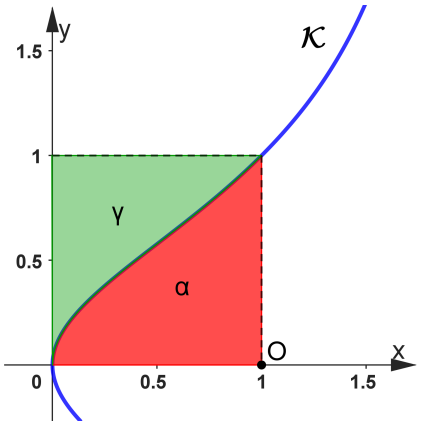


Figure 5

Conveniently for us, it is easier to obtain an expression for γ than for α, because the function defining α involves a square root. We have already obtained x in terms of z:

$$x = \frac{2z^2}{1+z^2}.$$

Hence, expressed in modern notation (the limits for z too are 0 to 1),

$$\gamma = \int_0^1 \frac{2z^2}{1+z^2}\, dz. \tag{10}$$

To simplify this, Leibniz expresses the fraction $1/(1+z^2)$ as an infinite series (this is permissible since $0 < z < 1$):

$$\frac{1}{1+z^2} = 1 - z^2 + z^4 - z^6 + \cdots . \tag{11}$$

Hence:

$$\frac{2z^2}{1+z^2} = 2\left(z^2 - z^4 + z^6 - z^8 + \cdots\right), \tag{12}$$

and

$$\begin{aligned}
\gamma = \int_0^1 \frac{2z^2}{1+z^2}\, dz &= 2\int_0^1 \left(z^2 - z^4 + z^6 - z^8 + \cdots\right)\, dz \\
&= 2\left(\frac{z^3}{3} - \frac{z^5}{5} + \frac{z^7}{7} - \frac{z^9}{9} + \cdots\right)\Bigg|_0^1 \\
&= 2\left(\frac{1}{3} - \frac{1}{5} + \frac{1}{7} - \frac{1}{9} + \cdots\right), \\
\therefore\ \frac{\gamma}{2} &= \frac{1}{3} - \frac{1}{5} + \frac{1}{7} - \frac{1}{9} + \cdots .
\end{aligned} \tag{13}$$

Since $\alpha = 1 - \gamma$ and $\pi/4 = \alpha/2 + 1/2$, it follows that

$$\begin{aligned}
\frac{\alpha}{2} &= \frac{1}{2} - \left(\frac{1}{3} - \frac{1}{5} + \frac{1}{7} - \frac{1}{9} + \cdots\right) \\
\therefore\ \frac{\pi}{4} &= 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \cdots .
\end{aligned} \tag{14}$$

We have arrived at the well known series which was discovered independently by Leibniz, Gregory, and Madhava.

Closing remarks

Several important remarks need to be made concerning the above derivation.

Leibniz's approach to notation. We remarked earlier in the article on how Leibniz was always on the lookout for better notation to make his methods easier to use. In his notebooks one can see the steady progression in his thinking. He realised early on that the problem of finding tangents (i.e., finding slope) and the problem of quadrature

(i.e., finding area) are inverse problems, in the sense that the tangency problem involves computing differences of ordinates, whereas the area problem involves computing sums of ordinates.

Leibniz associated the letter d (for difference) with the procedure for finding slope. He noticed that the action of taking differences brings down the dimension by 1, whereas the action of taking sums seems to increase the dimension by 1 (the slope function for x^2 is $2x$; the slope function for x^3 is $3x^2$; etc). So he started by placing the d in the denominator to reflect this behaviour (i.e., $\frac{y}{d}$ to denote the slope of y), but later found this to be inconvenient and so juxtaposed it with the symbol for the function itself (i.e., dy and not $\frac{y}{d}$). Mathematicians have stuck to this notation ever since!

In the case of area, he started by using the extremely awkward notation $Omn\,l$; here, 'Omn' is a short form for 'Omnia' which is Latin for 'all' and 'l' stands for length. The reason behind this symbolism is that area was regarded as stacking infinitely many lengths together and summing them. So Leibniz was writing statements like $Omn\,l\,x = \frac{1}{2}x^2$, $Omn\,l\,x^2 = \frac{1}{3}x^3$, and so on. Notationally, this is clearly very inconvenient. A little while later he discarded this in favour of S (for 'sum'), which he wrote in an ornate, elongated form: $\int$. Just a few days later he adopted the practise of including 'dx' with the symbol for integration. (All this action took place in the span of two weeks in early November 1675!) Thus was born the symbol for integration which we continue to use today.

Leibniz's use of known results. In the computation of γ (above), Leibniz had made use of results such as

$$\int z^2\,dz = \frac{z^3}{3}, \qquad \int z^4\,dz = \frac{z^5}{5}, \qquad \ldots \tag{15}$$

Leibniz did not have to derive these results from first principles; they were already widely known, having been found by earlier mathematicians such as Fermat.

Effect on Leibniz's life of the above discovery. Leibniz's derivation of the above series was one of his first major mathematical achievements, and it did much to propel him to fame among mathematicians and scientists of the time. It parallels (to some extent) what happened to Euler when he solved the Basel problem half a century later.

Leibniz went on to apply his transmutation theorem to find the areas corresponding to a large number of curves: the cycloid, the hyperbola, and so on. Probably he himself was surprised at the ease with which he was able to compute these areas using his new approach. (We may describe some of these results in a future article of this series.)

Connection between the transmutation theorem and integration by parts. We close this article by looking more closely at the transmutation theorem of Leibniz. We discover — much to our pleasure! — that this result is simply another form of the

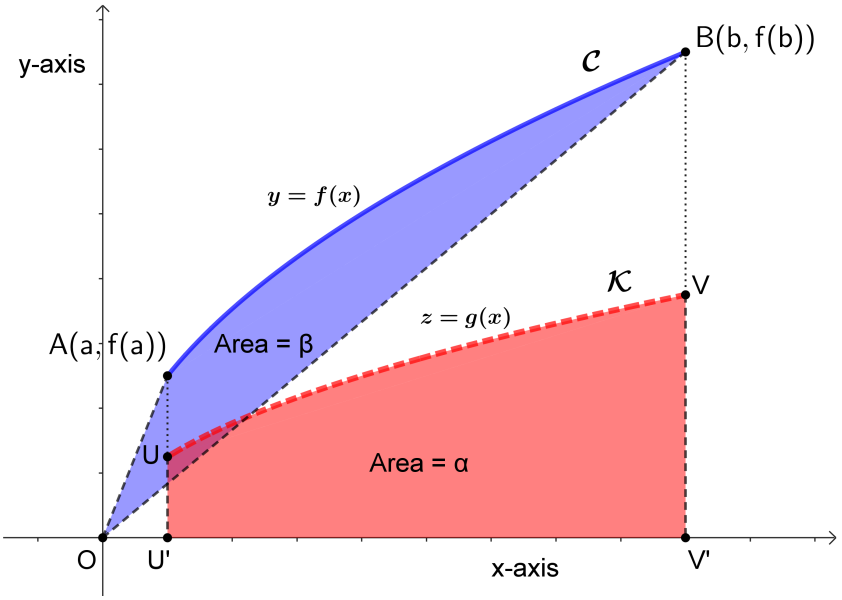


Figure 6: Implications of the transmutation theorem

rule for integration by parts that we use so frequently. Let us show how. We consider again the result proved in Figure 2. We have drawn the figure again here for convenience (Figure 6).

The following should be clear:

$$\beta + \text{Area of } \triangle BOV' = \text{Area of } AU'V'B + \text{Area of } \triangle AOU'. \tag{16}$$

Hence:

$$\frac{1}{2}\alpha + \frac{1}{2}bf(b) = \text{Area of } AU'V'B + \frac{1}{2}af(a). \tag{17}$$

That is,

$$\int_a^b g(x)\ dx = af(a) - bf(b) + 2\int_a^b f(x)\ dx. \tag{18}$$

Now consider again the procedure by which we constructed the point U from the point P. Figure 7 shows a simplified form of Figure 1 (redrawn for convenience).

The slope of $\mathcal{C}$ at P is $y' = \frac{dy}{dx}$. Hence, with $y = f(x)$ and $z = g(x)$:

$$y' = \frac{PU}{TU} = \frac{y-z}{x}, \qquad \therefore\ z = y - x\frac{dy}{dx}, \qquad \therefore\ g(x) = f(x) - x\frac{dy}{dx}. \tag{19}$$

Substituting this relation into the previous equality we get:

$$\int_a^b \left(f(x) - x\frac{dy}{dx}\right) dx = af(a) - bf(b) + 2\int_a^b f(x)\ dx. \tag{20}$$

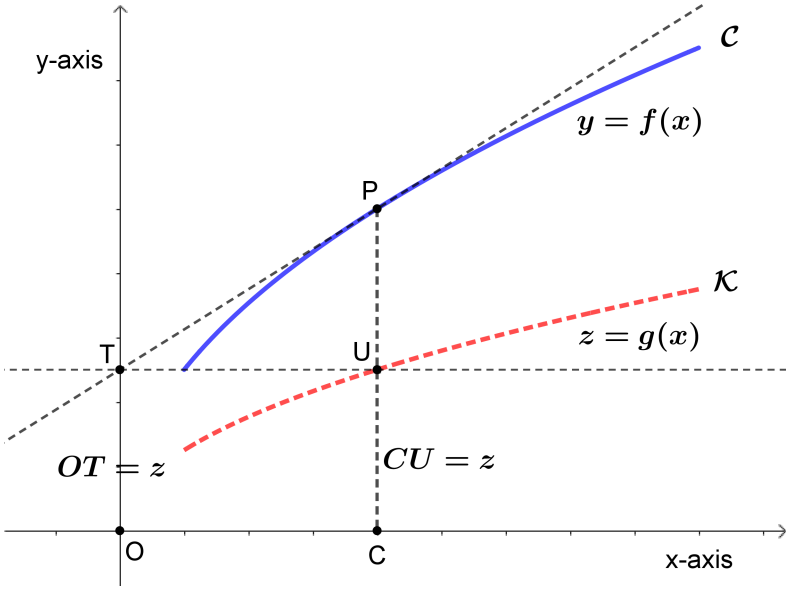


Figure 7

This simplifies to:

$$\int_a^b f(x)\ dx - \int_{x=a}^{x=b} x\ dy = af(a) - bf(b) + 2\int_a^b f(x)\ dx,$$

or:

$$\int_a^b f(x)\ dx = |xy|_a^b - \int_{y=f(a)}^{y=f(b)} x\ dy. \tag{21}$$

This is exactly the formula for integration by parts!

Further reading. We warmly recommend watching the video [1] and reading the following: [2], [3], [4] and [6].

Bibliography

[1] Open University (UK) & BBC Two. "The Birth Of Calculus (1986)." https://www.youtube.com/watch?v=0bPg3ki9GOI

[2] David Dennis and Susan Addington, "Pascal and Leibniz: Sines, Circles, and Transmutations." http://www.quadrivium.info/MathInt/Notes/Transmutation.pdf

[3] Ranjan Roy, "The Discovery of the Series Formula for π by Leibniz, Gregory and Nilakantha." *Mathematics Magazine*, Dec., 1990, Vol. 63, No. 5 (Dec., 1990), pp.

291-306. Available at https://maa.org/sites/default/files/images/upload_library/22/Allendoerfer/1991/0025570x.di021167.02p0073q.pdf

[4] Wikipedia, the free encyclopedia, "History of calculus." https://en.wikipedia.org/wiki/History_of_calculus

[5] Wikipedia, the free encyclopedia, "Fluxion." https://en.wikipedia.org/wiki/Fluxion

[6] Mactutor, "Gottfried Wilhelm von Leibniz." https://mathshistory.st-andrews.ac.uk/Biographies/Leibniz/

[7] Jimmy Iskandar, "Leibniz - Short Biography and Work." https://math.berkeley.edu/~robin/Leibniz/work.html

In forthcoming articles of this series we will study some of the work of James Gregory, Madhava, Colin Maclaurin and Isaac Newton. We will also study the quadrature of the cycloid and hyperbola by Leibniz.

As should be clear from the material presented above, the history of calculus is more like a tangled thicket than an upright coconut tree. So we have a lot more material to cover!

9 History in the Classroom: Revisiting the Completeness Property

Amber Habib

Shiv Nadar IoE
Gautam Buddha Nagar – 201314

Email: amber.habib@snu.edu.in

Abstract: The creation of the completeness axiom for real numbers is one of the markers of the modern conception of mathematical rigor. In today's classroom, it serves to demarcate school mathematics from university mathematics. Many facts that are taken as 'visually obvious' in school, are given formal support in university courses with the help of the completeness axiom. This step forward is not an easy one for students, with one reason being the usual presentation of the completeness axiom via the least upper bound property. In this article, we present an alternative formulation due to Alfred Tarski, which seems to lead to simpler proofs in the initial stages of analysis. As samples of its use, we give two applications. First, we combine it with Heron's method to prove the existence of square roots of real numbers. Second, we show that the natural logarithm is a bijection.

1 Versions of the Completeness Axiom

It was only in the late nineteenth century that mathematicians realised that they needed a better understanding of real numbers. In particular, the German mathematician Richard Dedekind – who was then teaching in a polytechnic – realised that the instruction of calculus lacked grounding in suitable axioms. What was needed was a formalisation of the idea that real numbers form a continuum, that they completely fill a line, something rational numbers fail to do.

Dedekind [1] provided the following **completeness axiom** for this purpose: Suppose A

and B are non-empty subsets of the set $\mathbb{R}$ of real numbers, such that (a) $\mathbb{R} = A \cup B$, and (b) $a \in A$ and $b \in B$ implies $a \leq b$. Then there is a unique $\alpha \in \mathbb{R}$ such that $a \leq \alpha \leq b$ for every $a \in A$ and $b \in B$.

With the help of this completeness axiom one can prove all the results that intuition suggests, such as the Cantor intersection property, the Archimedean property, the intermediate value theorem, the mean value theorem, and so on.

The typical use of the axiom is to show that a number with a given property exists by creating a collection A of underestimates and another collection B of overestimates. If these cut up $\mathbb{R}$ as described above, then the completeness axiom gives us a candidate α. These proofs tend to be complicated by the requirement that A and B should cover all of $\mathbb{R}$.

Over time, teachers and textbooks have moved to using an equivalent formulation called the **least upper bound property**.[1] This requires some preliminary definitions. First, a real number M is called an **upper bound** of a subset A of $\mathbb{R}$ if $M \geq a$ for every $a \in A$. If A has any upper bound, we say that A is **bounded above**. In this case, if there is an upper bound that is smaller than all other upper bounds, we call it the **least upper bound** of A.

The least upper bound property (usually abbreviated to 'LUB property') states that if a non-empty subset of the real numbers is bounded above then it has a least upper bound.

The LUB property simplifies proofs since we are only required to work with one set, and that set doesn't need any special property beyond being bounded above. On the other hand, students struggle with the concept of least upper bound, and especially the distinction between least upper bound and maximum. This confusion is an effective roadblock to utilising the desired benefits of the simpler proofs.

There have been other reformulations of the completeness axiom. The article [5] collects several of them. There is one that Alfred Tarski gave in 1937, which offers the benefits of the original completeness axiom without its complications. Tarski's axiom goes as follows: Suppose A and B are non-empty subsets of the set $\mathbb{R}$ of real numbers, such that $a \in A$ and $b \in B$ implies $a \leq b$. Then there is $\alpha \in \mathbb{R}$ such that $a \leq \alpha \leq b$ for every $a \in A$ and $b \in B$.

Theorem. The axioms of Dedekind and Tarski are equivalent.

[1] As I started to write this article, I checked a dozen books on calculus and real analysis. All but one gave the LUB property and called it the completeness axiom. The exception was [4], a fascinating book which succeeds in teaching calculus formally without using the ϵ-δ formulations.

Proof. Assume Dedekind's axiom. Now, suppose $A, B \subseteq \mathbb{R}$ are non-empty and that $a \leq b$ whenever $a \in A$ and $b \in B$. Take any $b \in B$. It is an upper bound of A, hence the set B' of all upper bounds of A is non-empty. Similarly, the set A' of all lower bounds of B' is non-empty. By definition, we have $a \leq b$ whenever $a \in A'$ and $b \in B'$. Further, take any $x \in \mathbb{R}$. If $x \notin A'$ then it is not a lower bound of B'. So $\exists b \in B'$ such that $b < x$. But then x is an upper bound of A, hence in B'. So $\mathbb{R} = A' \cup B'$. Dedekind's axiom can now be applied to get $\alpha \in \mathbb{R}$ such that $a \leq \alpha \leq b$ for every $a \in A'$ and $b \in B'$. This α also satisfies $a \leq \alpha \leq b$ for every $a \in A$ and $b \in B$.

Now assume Tarski's axiom. Suppose A and B are non-empty subsets of the set $\mathbb{R}$ of real numbers, such that (a) $\mathbb{R} = A \cup B$, and (b) $a \in A$ and $b \in B$ implies $a \leq b$. Tarski's axiom gives an $\alpha \in \mathbb{R}$ such that $a \leq \alpha \leq b$ for every $a \in A$ and $b \in B$. We need to show this α is unique. We see that α has to be either the maximum element M of A or the minimum element m of B. Non-uniqueness is only possible if both of these exist and are distinct. But then their midpoint would not belong to either A or B. ■

An important application of the completeness axiom is the **Archimedean property**. We present its derivation from Tarski's axiom:

Theorem. The set $\mathbb{N}$ of natural numbers is not bounded above in $\mathbb{R}$.

Proof. Suppose $\mathbb{N}$ is bounded above. Then the set B of all upper bounds of $\mathbb{N}$ is non-empty. Further, if $a \in \mathbb{N}$ and $b \in B$ then $a \leq b$. Hence, by Tarski's axiom, there is a real number α such that $a \leq \alpha \leq b$ for every $a \in \mathbb{N}$, $b \in B$.

Now, $\alpha - 1 < \alpha$ implies $\alpha - 1 \notin B$. Hence there is an $N \in \mathbb{N}$ such that $N > \alpha - 1$. But then $N + 1 \in \mathbb{N}$ and $N + 1 > \alpha$, a contradiction. ■

2 The Existence of Square Roots

The previous article in this series [2] had given a geometric introduction to Heron's method for estimating square roots. The method is as follows: Let x be any positive real number and let m be an overestimate of its positive square root, that is, $m^2 > x$. Define $m' = \frac{1}{2}\left(m + \frac{x}{m}\right)$. The number m' is an improved overestimate of $\sqrt{x}$. For example, if we take $x = 2$ and $m = 2$, we get $m' = \frac{3}{2} = 1.5$. Repeated application of Heron's method gives $m'' = \frac{17}{12} = 1.417$ and $m''' = \frac{577}{408} = 1.414215$. The last is accurate to 6 significant places.

We shall combine Heron's method with Tarski's axiom to prove the existence of square roots of positive real numbers. We denote the set of positive real numbers by $\mathbb{R}^+$. This proof was published earlier in [3].

Theorem. Let $x \in \mathbb{R}^+$. Then there is $y \in \mathbb{R}^+$ such that $y^2 = x$.

Proof. Define $A = \{a \in \mathbb{R}^+ : a^2 < x\}$ and $B = \{b \in \mathbb{R}^+ : b^2 > x\}$. Now observe that A and B are non-empty:

- If $x > 1$ then $1 \in A$, while if $x \leq 1$ then $x/2 \in A$.
- In all cases, $x + 1 \in B$.

Now $a \in A$ and $b \in B$ implies that $a^2 < x < b^2$, and hence $a < b$. Tarski's axiom gives $y \in \mathbb{R}^+$ such that $a \leq y \leq b$ for every $a \in A$, $b \in B$.

We note that if $y \in A$ then y is the greatest element of A, while if $y \in B$ then y is the least element of B. Therefore, if we show that A has no greatest member and B has no least member, we will have ruled out both $y^2 < x$ and $y^2 > x$, leaving $y^2 = x$ as the only possibility.

Let us first show that B has no least member. Take any $m \in B$. Apply Heron's method to m:

$$m' = \frac{1}{2}\big(m + \frac{x}{m}\big) = \frac{x}{m} + \frac{1}{2}\big(m - \frac{x}{m}\big).$$

Then $m^2 > x$ gives $m > \frac{x}{m}$, hence $0 < m' < m$. Further, $m'^2 > \dfrac{x^2}{m^2} + \dfrac{x}{m}\big(m - \dfrac{x}{m}\big) = x$ gives $m' \in B$. Hence B has no least element.

Next, observe that $t \in A$ if and only if $\frac{1}{t} \in B' = \{m \in \mathbb{R}^+ : m^2 > \frac{1}{x}\}$. By the argument above, B' has no least element. It follows that A has no greatest element.

This completes the proof of $y^2 = x$. ■

3 The Logarithm

The natural logarithm can be defined by $\log x = \int_1^x \frac{1}{t}\,dt$ for $x > 0$. Its algebraic properties and monotonicity follow easily from this definition. The fact that it is a bijection between $\mathbb{R}^+$ and $\mathbb{R}$ is harder to establish. We shall do so in two stages, using Tarski's axiom for the first one.

Theorem. The natural logarithm has the intermediate value property: If $b > a > 0$ and $\log b > L > \log a$ then there is $\alpha \in (a, b)$ with $\log \alpha = L$.

Proof. Define $A = \{x \in \mathbb{R}^+ : \log x < L\}$ and $B = \{x \in \mathbb{R}^+ : \log x > L\}$. Then $a \in A$ and $b \in B$.

Suppose $x \in A$ and $y \in B$. Then $\log x < L < \log y$. Since log is an increasing function, we must have $x < y$. Hence, by Tarski's axiom, there is a real number α such that $x \leq \alpha \leq y$ for every $x \in A$, $y \in B$.

Suppose $\log \alpha < L$. Then $\log \alpha = L - \epsilon$ with $\epsilon > 0$. For any $\delta > 0$,

$$\log(\alpha + \delta) = \log \alpha + \int_{\alpha}^{\alpha+\delta} \frac{1}{t}\, dt \leq (L - \epsilon) + \int_{\alpha}^{\alpha+\delta} \frac{1}{\alpha}\, dt = (L - \epsilon) + \frac{\delta}{\alpha}.$$

If we choose $\delta = \frac{1}{2}\epsilon\alpha$, we get $\log(\alpha + \delta) \leq L - \frac{\epsilon}{2} < L$. Hence $\alpha + \delta \in A$, a contradiction. This eliminates $\log \alpha < L$.

If $\log \alpha = L + \epsilon$, we can show in a similar fashion and using $\delta = \frac{1}{2}\epsilon a$, that $\alpha - \delta \in B$. This eliminates $\log \alpha > L$.

Therefore $\log \alpha = L$. ■

Theorem. The function $\log\colon \mathbb{R}^+ \to \mathbb{R}$ is a bijection.

Proof. We have $\log 1 = 0$. So we have to show every non-zero y is in the range of log.

Consider any $y > 0$. The Archimedean property gives $N \in \mathbb{N}$ such that $\frac{y}{\log 2} < N$. Hence $y < N \log 2 = \log(2^N)$. So,

$$\log 1 < y < \log 2^N.$$

By the intermediate value property of log, there is $x \in (1, 2^N)$ such that $\log x = y$.

Now consider $y < 0$. There is $x > 0$ such that $\log x = -y$. Then $\log(x^{-1}) = -\log x = y$. So log is onto. ■

It is worth noting that we have managed to introduce the key features of the log function without waiting (as is usually done) for a general development of limits and continuity. In fact, Tarski's axiom is particularly well-suited to defining and obtaining the properties of the Riemann integral (via Darboux's approach). It reduces the proofs to simple checks of inequalities. The book [3] demonstrates this.

Bibliography

[1] R. Dedekind (Translated by W. W. Beman), *Essays on the Theory of Numbers*. Open Court Publishing Company, 1901. http://www.gutenberg.org/2/1/0/1/21016/

[2] A. Habib, History in the Classroom: Area and Algebra. *Blackboard*, Issue 3, 2021.

[3] A. Habib, *Calculus*, Indian Ed. Cambridge University Press, 2022.

[4] J. E. Marsden and A. Weinstein, *Calculus Unlimited*. Benjamin-Cummings Publishing Company, 1985. https://authors.library.caltech.edu/25054/.

[5] H. Teismann, Toward a More Complete List of Completeness Axioms. *The American Mathematical Monthly*, Vol. 130, No. 2, 2013, 99–114.

10 How Many Dots? How Many Distinct Values?

Jyotirmoy Sarkar

Indiana University Indianapolis
Department of Mathematical Sciences
402 N Blackford Street
Indianapolis, IN 46202-3216, USA

Email: jsarkar@iupui.edu

Abstract: If a fair six-sided die is rolled six times, how many dots in total will you see on the top face? If a fair dodecahedron (12-sided die) is rolled six times, how many distinct values will you see? Think over these questions, and compute the answers using a software.

At our university — read "unity in diversity" — undergraduate students pursuing different degree programs enroll in the same course *Introduction to Probability/Statistics* with different levels of mathematical preparation. Faced with a problem, they seek different types of solutions: (1) the numerical answer; (2) the formula that must be evaluated; (3) a computational technique; and (4) mathematical logic. Here we present two such problems — one we teach (mostly) in the classroom as an example, and the other we assign as a take-home exercise. All computations are done using the freeware R [1]. To the interested reader we pose an open problem.

1 A DRV and its PMF

Before presenting the problems, let us describe what students have already learned. Students have been taught that the answer to a question "How many?" need not be a number; it may be a discrete random variable (DRV) which takes on different values with associated probabilities that must add to one. Such a listing, or its graphical

presentation, or a formula that will produce this list, is called a probability mass function (PMF).

For instance, suppose that we draw 3 balls without replacement (WOR) (or with replacement, WR) from a basket containing 3 red and 7 blue balls. How many red balls will be drawn?

The number of red balls drawn, X, has a PMF given in Table 1, or in Figure 1, or in formulas (1)-(2). If sampled WOR, the distribution is hypergeometric$(3, 3, 7)$; if WR, it is binomial$(3, .3)$. See any introductory statistics book such as [2].

Table 10.1: PMF of the number of red balls drawn.

	x	0	1	2	3	sum
WOR	$P\{X = x\}$	.2917	.5250	.1750	.0083	1
WR	$P\{X = x\}$	.3430	.4410	.1890	.0270	1

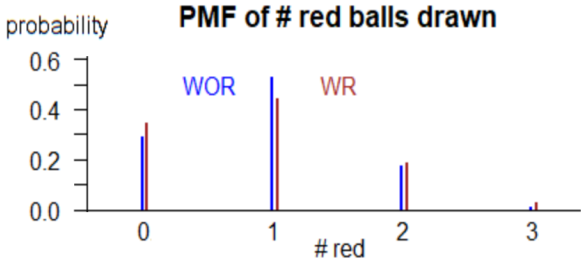


Figure 1: PMF of the number of red balls drawn.

$$\text{WOR: } P\{X = x\} = \frac{\binom{3}{x}\binom{7}{3-x}}{\binom{10}{3}}, \text{ for } x = 0, 1, 2, 3. \tag{1}$$

$$\text{WR: } P\{X = x\} = \binom{10}{x}\left(\frac{3}{10}\right)^{x}\left(\frac{7}{10}\right)^{3-x}, \text{ for } x = 0, 1, 2, 3. \tag{2}$$

Having defined and illustrated the notions of a DRV and its PMF, we dive into a more advanced example.

2 In-class Example

Example: How many dots will you see (1) on the top face when you roll a fair die once? (2) twice? (3) three times? (4) four times? (5) five times? (6) six times?

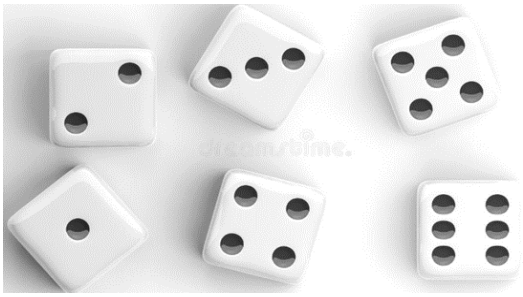


Figure 2: What is the PMF of the number of dots on the top face when a fair die is rolled 6 times?

When rolling the fair die once or twice, students can easily make a list of all possible equally likely outcomes and make a table of possible values of the total number of dots on top face(s) together with associated probabilities. For $k = 1, 2$, we have

$$(p_1(x) : x = 1, 2, \ldots, 6) = (1, 1, 1, 1, 1, 1)/6$$
$$(p_2(x) : x = 2, 3, \ldots, 12) = (1, 2, 3, 4, 5, 6, 5, 4, 3, 2, 1)/6^2$$

But for $k \geq 3$ rolls, most students are inclined to use a computer programming language to compute the PMF requiring different degrees of sophistication: Simulation requires a one line code; complete enumeration of all 6^k outcomes requires k nested "for" loops; recursive enumeration involves k steps; and mathematical derivation requires the celebrated generalized binomial theorem (see [5]), credited to Sir Isaac Newton. After describing the thought process needed in each method of solution, we compute all quantities using the freeware R [1]. To learn more about the use of R simulation in probability and statistics, see [4].

Solution 1: (Simulation) In a simulation we replicate a process many times and thereafter we summarize the results. A simulation of 6 rolls of a fair die, replicated 10^6 times, yields the following frequencies $f_6(x)$ for the total number of dots, whence we estimate the probabilities $P\{X = x\} \approx f_6(x)/10^6$ for $x = 6, 7, \ldots, 36$.

```
# simulation for total number of dots
table(replicate(10^6, sum(sample(1:6,6, replace=T))))
    6     7     8     9    10    11    12    13    14    15
   20   121   471  1228  2717  5357  9706 16131 24764 35851
   16    17    18    19    20    21    22    23    24    25    26
48021 61514 73561 83222 90694 92867 90727 83861 73373 61279 48198
   27    28    29    30    31    32    33    34    35    36
35529 24776 16169  9863  5455  2743  1186   444   133    19
```

Better still, symmetrize the estimates; that is, take

$$P\{X = x\} = P\{X = 42 - x\} \approx \frac{f_6(x) + f_6(42 - x)}{2 \times 10^6}.$$

Solution 2: (Complete enumeration) A complete enumeration accounts for every possible outcome and computes a desired quantity. Here we compute the sum for each of the 6^k equally likely outcomes of k rolls, and tabulate the total number of dots. For k rolls, the total ranges over k to $6k$ with exact frequencies $a_k(x)$. For $k = 6$, we compute $a_6(x)$ as follow:

```
# complete enumeration
total=c()
for (i1 in 1:6){
 for (i2 in 1:6){
  for (i3 in 1:6){
   for (i4 in 1:6){
    for (i5 in 1:6){
     for (i6 in 1:6){
total=c(total, i1+i2+i3+i4+i5+i6)
}}}}}}
table(total)
total    6    7    8    9   10   11   12   13   14   15
freq     1    6   21   56  126  252  456  756 1161 1666
total   16   17   18   19   20   21   22   23   24   25   26
freq  2247 2856 3431 3906 4221 4332 4221 3906 3431 2856 2247
total   27   28   29   30   31   32   33   34   35   36
freq  1666 1161  756  456  252  126   56   21    6    1
```

The frequencies are symmetric and unimodal (with a mode at $(6 + 36)/2 = 21$). These 6^k outcomes being equally likely, the PMF $p_k(x) = P_k\{X = x\}$ is the ratio

$$\frac{a_k(x)}{6^k}, \text{ for } x = k, k + 1, \ldots, 6k \tag{3}$$

Solution 3: (Recursive enumeration) In a recursive computation, we describe the boundary values and explain how the neighboring values are computed — in succession. The numerator $a_k(x)$ in (3) can be computed recursively. First, $a_1(x) = 1$ for $x = 1, 2, \ldots, 6$; and $a_1(x) = 0$ for all other x. Thereafter, for $k \geq 2$, conditioning on the outcome i of the first roll (or the sum j after $(k - 1)$ rolls), we have

$$a_k(x) = \sum_{i=1}^{6} a_{k-1}(x - i) = \sum_{j=x-6}^{x-1} a_{k-1}(j), \text{ for } x = k, \ldots, 6k \tag{4}$$

Equation (4) generalizes the construction of Pascal's triangle (see [6] for a dynamic demonstration) which lists the coefficients of the binomial expansion $(a+b)^n$ for positive integers $n \geq 1$. In Pascal's triangle the coefficients in a new line are obtained by taking the sums of successive pairs of coefficients in the previous line. Here, the coefficients in a new line are constructed by taking the sums of successive strings of six coefficients in the previous line (augmented by leading and trailing zeros).

Below we compute a_k for $2 \leq k \leq 6$.

```
# recursive computation
cumsum6=function(a){
   b=cumsum(c(a, rep(0,5)))
   l=length(a)-1
   b-c(rep(0,6), b[1:l])
} # end function
a1=rep(1,6)
a2=cumsum6(a1); c(a2, sum(a2), 6^2)
 [1]  (1  2  3  4  5  6  5  4  3  2  1) 36=36
a3=cumsum6(a2); c(a3, sum(a3), 6^3)
 [1]  (1   3   6  10  15  21  25  27  27  25  21  15  10   6   3   1) 216=216
a4=cumsum6(a3); c(a4, sum(a4), 6^4)
 [1]  (1    4   10   20   35   56   80  104  125  140  146
[12] 140  125  104   80   56   35   20   10    4    1) 1296=1296
a5=cumsum6(a4); c(a5, sum(a5), 6^5)
 [1]  (1    5   15   35   70  126  205  305  420  540  651  735  780
[14] 780  735  651  540  420  305  205  126   70   35   15    5    1) 7776=7776
a6=cumsum6(a5); c(a6, sum(a6), 6^6)
 [1]   (1     6    21    56   126   252   456   756  1161  1666
[11] 2247  2856  3431  3906  4221  4332  4221  3906  3431  2856  2247
[22] 1666  1161   756   456   252   126    56    21     6     1) 46656=46656
```

The vector $a_k = (a_k(x) : x = k, \ldots, 6k)$ being symmetric, that is, $a_k(x) = a_k(7k - x)$, it suffices to compute the first "half" of a_k and then fill-in the second half by writing the first half in reverse. Thus, symmetry is not just pretty; it simplifies computation.

A more efficient computation of a_k is given in the Appendix.

Solution 4: (Mathematical derivation) The probability generating function (PGF) of the number of dots in one single roll is $\sum_{x=1}^{6} t^x \, P\{X = x\} = (t + t^2 + \ldots + t^6)/6$. See [7]. Hence, the PGF of the sum of k *independent* rolls is $(t + t^2 + \ldots + t^6)^k/6^k$. Therefore, $a_k(x)$ is the coefficient of t^x in the expansion of $(t + t^2 + \ldots + t^6)^k = t^k(1 - t^6)^k(1 - t)^{-k}$. Using the generalized binomial theorem (see [5]), which allows the index n in $(a + b)^n$ to be any real number, $a_k(x)$ is the coefficient of t^x in

$$t^k \, (1 - t^6)^k \, (1 - t)^{-k} = t^k \left(\sum_{i=0}^{k} \binom{k}{i} (-t^6)^i \right) \left(\sum_{j=0}^{\infty} \binom{-k}{j} (-t)^j \right) \tag{5}$$

where $\binom{-k}{j} = (-k)(-k-1)\cdots(-k-j+1)/j! = \binom{k+j-1}{j}(-1)^j$.

We evaluate (5), for $3 \le k \le 6$ (in any order we want)

$$\begin{aligned}
k=3:\ & t^3\ \left(1-3t^6+3t^{12}-t^{18}\right)\times \\
& \left(1+3t+6t^2+10t^3+15t^4+21t^5+28t^6+\cdots\right) \\
& = t^3+3t^4+6t^5+10t^6+15t^7+21t^8+27t^9+\ldots+t^{18} \\
k=4:\ & t^4\ \left(1-4t^6+6t^{12}-4t^{18}+t^{24}\right)\times \\
& \left(1+4t+10t^2+20t^3+35t^4+56t^5+84t^6+\cdots\right) \\
k=5:\ & t^5\ \left(1-5t^6+10t^{12}-10t^{18}+5t^{24}-t^{30}\right)\times \\
& \left(1+5t+15t^2+35t^3+70t^4+126t^5+210t^6+\cdots\right) \\
k=6:\ & t^6\ \left(1-6t^6+15t^{12}-20t^{18}+15t^{24}-6t^{30}+t^{36}\right)\times \\
& \left(1+6t+21t^2+56t^3+126t^4+252t^5+462t^6+\cdots\right)
\end{aligned}$$

By the usual binomial theorem, where the index is a positive integer, the coefficient $\binom{k}{x}$ of $(-t^6)^x$ in $(1-t^6)^k$ equals the sum $\binom{k-1}{x}+\binom{k-1}{x-1}$ of coefficients in $(1-t^6)^{k-1}$. However, the coefficient $\binom{-k}{j}$ of $(-t)^j$ in $(1-t)^{-k}$ equals the cumulative sum of coefficients of $(-t)^i$ $(i=0,1,\ldots j)$ in $(1-t)^{-(k-1)}$; that is,

$$\binom{-k}{j} = \sum_{i=0}^{j}\binom{-(k-1)}{i}.$$

One advantage of a mathematical formula is that we can compute the coefficients for a particular k without first computing them for all values smaller than k. The other advantage is that, in its simplest form, a mathematical formula usually requires the fewest number of arithmetic operations. The interested reader should check that the coefficient of t^x is 0 for any $x < k$ and any $x > 6k$, and the coefficients of t^x and t^{7k-x} are the same for all $k \le x \le 6k$. For $k = 6$, we compute the frequencies below.

```
# mathematical calculations
k=6; k0=seq(0,k,1); q=choose(k,k0)*(-1)^k0
index=seq(0,5*k,1); e=length(index)
p=choose(-k, index)*(-1)^index
p + q[2]*c(rep(0,1*6),p[1:(e-1*6)]) +  q[3]*c(rep(0,2*6),p[1:(e-2*6)]) +
    q[4]*c(rep(0,3*6),p[1:(e-3*6)]) +  q[5]*c(rep(0,4*6),p[1:(e-4*6)]) +
    q[6]*c(rep(0,5*6),p[1:(e-5*6)])

 [6]    1    6   21   56  126  252  456  756 1161 1666
[16] 2247 2856 3431 3906 4221 4332 4221 3906 3431 2856 2247
[27] 1666 1161  756  456  252  126   56   21    6    1
```

Having taught the above example (mostly) in class, we assign a take-home exercise stated below, fully anticipating different types of solutions exhibiting different levels of sophistication.

3 Take-Home Exercise

Exercise: Six people were seated at a restaurant table. The waiter passed out to each dinner guest a small square card containing the names of all 12 months (see Figure 3) and asked them to mark any one month *at random*. After all guests made their choices, the waiter sorted the cards according to the month chosen. How many distinct months were chosen by the six dinner guests?

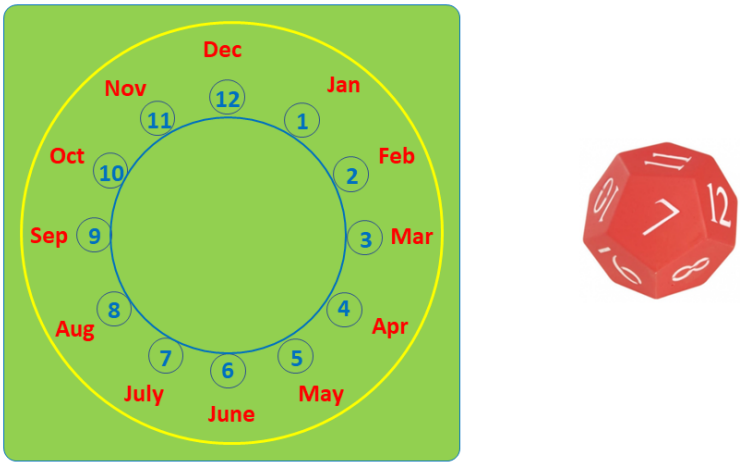


Figure 3: Choose any one month at random.

The exercise allows 6 guests to choose months *independently* of one another. As such, the problem is equivalent to rolling a fair dodecahedron (a 12-sided die) 6 times. See Figure 3 again. The number of distinct values chosen is a DRV on $\{1, 2, \ldots, 6\}$; but with what associated probabilities? We present several solutions.

Solution 1: (Simulation) Some students (mostly in Biology and Chemistry) looking for an approximate numerical answer conduct a simulation of the data collection conducted by the waiter.

```
# simulation for distinct values
dv=replicate(10^6, length(unique(sample(1:12,6,replace=T))))
```

```
table(dv)
      1        2        3        4        5        6      sum
      4     1429    39634   258959   477109   222865   10^6
.000004  .001429  .039634  .258959  .477109  .222865   1.00
```

According to the square root law of the standard error of estimating a proportion, see [3] for example, the simulation probabilities are correct to three decimal places. This claim is verified from the exact probabilities computed in the solutions below.

Solution 2: (Complete enumeration) Other students (in Physics and Engineering) prefer complete enumeration. Since every guest is allowed to choose in 12 possible ways at random, the total number of equally likely ways all 6 guests can choose is 12^6. For every such choice, compute the number of distinct months chosen, and summarize these $12^6 = 2,985,984$ values in a table.

```
# complete enumeration
distinct=rep(0,12^6);
i=1
for (i1 in 1:12){
 for (i2 in 1:12){
  for (i3 in 1:12){
   for (i4 in 1:12){
    for (i5 in 1:12){
     for (i6 in 1:12){
distinct[i]=length(unique(c(i1,i2,i3,i4,i5,i6)))
i=i+1;
}}}}}}
table(distinct); table(distinct)/12^6
```

```
# distinct         1       2       3       4       5       6
frequency         12    4092  118800  772200 1425600  665280
probability  .000004 .001370 .039786 .258608 ,477431 .222801
```

These probabilities support the degree of accuracy of the simulation claimed above in Solution 1.

Solution 3: (Recursive enumeration) Some students (especially in Applied Mathematics and Computer Science) prefer to compute the frequencies recursively. Irrespective of what the first guest chooses, exactly one month is selected. Thereafter, the second guest may select the same month as the first guest with probability (w.p.) 1/12, or a different month w.p. 11/12. If the first two guests have chosen the same month, then the third

guest may select the same month w.p. 1/12, or a different month w.p. 11/12. But if the first two guests have chosen two different months, then the third guest may select one of those two already chosen months w.p. 2/12, or a third month w.p. 10/12. Etc.

For $k \geq 1$, let b_k denote the k-dimensional vector

$$\left(12^{k-1}\, P_k\{X = x\} : x = 1, 2, \ldots, k\right)$$

after k guests have made their choices. Clearly, $b_1 = (1)$, and $b_2 = (1, 11)$. Thereafter, for $k \geq 3$, $b_k(1) = 1$, and for $x = 2, \ldots, k$,

$$b_k(x) = x * b_{k-1}(x) + (13 - x) * b_{k-1}(x - 1). \tag{6}$$

These calculations are done below for $k = 6$.

```
# recursive enumeration for k=6
one=c(1, 0, 0, 0, 0, 0);    two=1*one+11*c(0,one[1:5])
thrE=c(1,2,0,0,0,0)*two +c(0,c(11,10,0,0,0)* two[1:5])
four=c(1,2,3,0,0,0)*thrE+c(0,c(11,10,9,0,0)*thrE[1:5])
five=c(1,2,3,4,0,0)*four+c(0,c(11,10,9,8,0)*four[1:5])
six6=c(1,2,3,4,5,0)*five+c(0,c(11,10,9,8,7)*five[1:5])
six6; six6/12^5
[1]        1     341    9900   64350  118800   55440
[1]  .000004 .001370 .039786 .258608 .477431 .222801
```

Solution 4: (Mathematical derivation) Only a few students (in Pure Mathematics, Statistics and Actuarial Science) compute each probability using combinatorics. For $\{X = x\}$, with $1 \leq x \leq 6$, the 6 guests must break out into x groups, and then the groups must choose distinct months 1 through 12.

If all guests form one group only, it can be done in only 1 way. Thereafter, only one month can be chosen in 12 ways. It does not matter which month is chosen. Therefore, the number of ways is divided by 12 to get $b_6(1) = 1$. If all guests form two groups of sizes (1,5), (2,4), (3,3), these groups can be formed in $\binom{6}{1} = 6$, $\binom{6}{2} = 15$, $\binom{6}{3}/2 = 10$ ways, respectively. The last count is divided by 2 because the two groups of sizes (3,3) are interchangeable. Thereafter, two months are chosen in $12 * 11$ ways. Hence, dividing by 12, $b_6(2) = (6 + 15 + 10) * 11 = 341$.

If all guests form three groups of sizes (1,1,4), (1,2,3), (2,2,2), these groups can be made in $\binom{6}{4} = 15$, $\binom{6}{3}\binom{3}{2} = 60$, $\binom{6}{2}\binom{4}{2}/3! = 15$ ways, respectively. Again, the last count is divided by 3! because the three groups of sizes (2,2,2) can be rearranged in 3! ways. Thereafter, three months are chosen in ${}^{12}P_3$ ways. Hence, dividing by 12, $b_6(3) = (15{+}60{+}15) * {}^{11}P_2 =$ 9900. If all guests form four groups of sizes (1,1,1,3), (1,1,2,2), these groups can be made

in $\binom{6}{3} = 20$, $\binom{6}{2}\binom{4}{2}/2 = 45$ ways. Once again, the last count is divided by 2 because the two groups of sizes (2,2) can be interchanged in 2 ways. Thereafter, four months are chosen in ${}^{12}P_4$ ways. Hence, dividing by 12, $b_6(4) = (20 + 45) * {}^{11}P_3 = 64350$.

Suppose that all guests form five groups of sizes (1,1,1,1,2) in $\binom{6}{2} = 15$ ways. Thereafter, five months are chosen in ${}^{12}P_5$ ways. Hence, dividing by 12, $b_6(5) = 15 * {}^{11}P_4 = 118800$. If all guests form six groups, it can be done in only one way. Thereafter, six months are chosen in ${}^{12}P_6$ ways. Hence, dividing by 12, $b_6(6) = {}^{11}P_5 = 55440$.

Having already computed the frequencies $b_6(x)$ for $x = 1, \ldots, 6$, what more do we need a computer software for? We use it to verify our computations. Moreover, we might be interested in computing b_k for $k > 6$ rolls of a fair dodeccahedron, or perhaps for any number of rolls of a fair icosahedron (a 20-sided die).

```
# combinatorial calculations
fr6=factorial(12)/factorial(6)
fr5=choose(6,2)*factorial(12)/factorial(7)
   gr4=(choose(6,3)+choose(6,2)*choose(4,2)/2)
fr4=gr4*factorial(12)/factorial(8)
   gr3=(choose(6,4)+choose(6,3)*choose(3,2)+choose(6,2))
fr3=gr3*factorial(12)/factorial(9)
   gr2=(choose(6,5)+choose(6,4)+choose(6,3)/2)
fr2=gr2*factorial(12)/factorial(10)
fr1=12
freq=c(fr1, fr2, fr3, fr4, fr5, fr6)
freq;   freq/12;    freq/12^6
[1]      12    4092  118800 772200  1425600  665280
[1]       1     341    9900  64350   118800   55440
[1] .000004 .001370 .039786 .258608 .477431 .222801
```

4 An Open Problem

While my students could read my solutions only to the in-class example and had to work on the take-home exercise on their own, you, my dear and diligent readers, have had access to my solutions to both problems. To make it fair, let me assign you a new problem. Have fun solving it.

Open Problem: Fresh out of Medical School, Dr. Monica opened a clinic and instructed her secretary, "Schedule no more than five patients per day. Finalize the schedule at least a week in advance." At the end of the first year, the number of patients Dr. Monica saw on each clinic day exhibited the following PMF $p(x)$:

x	0	1	2	3	4	5
p(x)	.10	.10	.15	.35	.20	.10

Of course, Dr. Monica did not come to clinic on days with no patient scheduled. She also thought it was sub-optimal to come to clinic on days with only one or two patients scheduled. Thus spake she to her secretary: "From now on, please vacate the days with only one or two patients scheduled, reschedule these patients on the closest future day with fewer than 5 patients, but do not exceed the five-patients-per-day limit."

Assuming that the patient inflow distribution remains unchanged and every patient accepts the newly scheduled date, under the revised scheduling policy, what is the new PMF $q(x)$ of the number of patients Dr. Monica will see on each clinic day?

Clearly, $q(0) = .35, q(1) = 0, q(2) = 0$. Show that $q(3) = .177, q(4) = .147, q(5) = .326$. Best wishes on deriving this solution at any level of sophistication you choose.

My Perspective on Teaching Changed

We solved two probability problems using different methods that require different depths of thinking. In my younger days, I would insist that every student must learn every solution. Accordingly, I would diligently endeavour to realize that noble objective. Predictably, it inflicted much pain on the students and overwhelmed their fateful instructor with inevitable frustration.

Today I admit that students have different backgrounds, needs and ambitions. The instructor's responsibility is to inspire them *all* to ask the appropriate questions and chase after the solutions they are capable of pursuing. Instead of pushing wholesome food down everybody's throat, it is healthier to let each person choose their meal according to their own taste, appetite, ability to digest, and pleasure. *Bon appetit!*

Bibliography

[1] R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Free download at https://www.R-project.org/.

[2] Ronald E. Walpole, Raymond H. Myers, Sharon L. Myers and Keying E. Ye, *Probability and Statistics for Engineers and Scientists*, 9 ed, 2016.

[3] Vijay K. Rohatgi and A. K. Md. E. Saleh, *An Introduction to Probability and Statistics*, 2ed, Wiley Series in Probability and Statistics, 2008.

[4] Robert Parker, *R Lecture Notes*, University of Florida, 2020. https://users.phhp.ufl.edu/rlp176/Courses/PHC6089/R_notes/simulations.html

[5] Wikipedia, the Free Encyclopedia. Binomial Theorem. https://en.wikipedia.org/wiki/Binomial_theorem.

[6] Wikipedia, the Free Encyclopedia. Pascal's triangle. https://en.wikipedia.org/wiki/Pascal's_triangle.

[7] Wikipedia, the Free Encyclopedia. Probability-generating function. https://en.wikipedia.org/wiki/Probability-generating_function.

Appendix: Efficient Computation

In Section 2, Solution 3 of Problem 1 recursively computes the exact frequencies $a_k(x)$ of obtaining a total of $k \le x \le 6k$ dots in k rolls of a fair die. Having already computed a_m and a_n for $1 < m \le n$, there is a more efficient way to construct a_{m+n} without having to compute $a_{n+1}, a_{n+2}, \ldots, a_{n+m-1}$. How?

Compute the running inner product (RIP) $a_m \# a_n$ as follows. [Here, RIP stands for running inner product, not the more familiar tombstone inscription. Alert readers may adapt this procedure to multiply two large numbers having m and n digits, respectively, in only one step.] Write down the row vectors a_m and a_n on two horizontal strips of paper with entries equally spaced. Then holding vector a_m fixed on the left side (and above), bring vector a_n from the right side (and below) closer and closer creating a partial overlap that keep on increasing; for every possible overlap, compute the inner product and jot it down in a new vector a_k (where $k = m + n$) term by term. Continue until the sliding vector a_n passes out from the left side of the fixed vector a_m. In class, we demonstrate this RIP operation under the document camera by computing a_4, without first computing a_3, as $a_4 = a_2 \# a_2$: Clearly, $a_4(4) = 1(1) = 1$,

$$a_4(5) = (1,2,3,4,5,6,5,4,3,2,1)$$
$$(1,2,3,4,5,6,5,4,3,2,1) = 2(1) + 1(2) = 4$$
$$a_4(6) = (1,2,3,4,5,6,5,4,3,2,1)$$
$$(1,2,3,4,5,6,5,4,3,2,1) = 3(1) + 2(2) + 1(3) = 10$$
$$a_4(7) = (1,2,3,4,5,6,5,4,3,2,1)$$
$$(1,2,3,4,5,6,5,4,3,2,1) = 4(1) + 3(2) + 2(3) + 1(4) = 20$$

Etc.

```
# recursive enumeration via running inner product
```

```
rip=function(a,b){ m=length(a); n=length(b); p=c()
if(m<=n){ for (i in 1:m){p[i]=rev(a[1:i])%*%b[1:i]}
 for (j in 1:(n-m)){p[m+j]=rev(a)%*%b[(1+j):(m+j)]}
 for (k in 1:(m-1)){p[n+k]=rev(a)[1:(m-k)]%*%b[(n-m+k+1):n]}
p} # end if
else{print ("interchange a and b")}
} # end function
a1=rep(1,6);  a2=rip(a1,a1);  a4=rip(a2,a2); a6=rip(a2,a4);
a4; c(sum(a4),6^4);  a6; c(sum(a6),6^6)
a4=(  1   4  10  20  35  56  80 104 125 140 146
    140 125 104  80  56  35  20  10   4   1 )           [1]  1296     1296
a6=(  1    6   21   56  126  252  456  756 1161 1666
   2247 2856 3431 3906 4221 4332 4221 3906 3431 2856 2247
   1666 1161  756  456  252  126   56   21    6    1 )  [1] 46656    46656
```

Curious readers may wonder: Starting from $a_1 = (1, 1, 1, 1, 1, 1)$, what is the least number of RIP operations needed to compute a_k, for any given $k \geq 2$? Here is the binary strategy: If $k = 2^m$, compute $\{a_2, a_4, \ldots, a_{2^m}\}$ using m RIP operations. Otherwise, if $k = (2^m + b_{m-1}2^{m-1} + \ldots + b_1 2 + 1)2^n$, write k in binary notation $k = (1, b_{m-1}, \ldots b_2, b_1, 1, 0, \ldots, 0)_2$, where each of $b_{m-1}, \ldots, b_2, b_1$ is either 0 or 1, and there are n trailing zeros. Then compute and save $\{a_2, a_4, \ldots, a_{2^{m+n}}\}$ using $(m+n)$ RIP operations; thereafter, to $a_{2^{m+n}}$ successively enjoin a_{2^i} (which has been already computed and saved) for each $b_i = 1$ $(0 \leq i < m)$, requiring $(b_{m-1} + \ldots + b_1 + 1)$ more RIP operations. For example, to compute a_{50}, since $50 = (1, 1, 0, 0, 1, 0)_2$, the binary strategy needs 5+2=7 RIP operations to compute a_{50} as

$$\{a_2, a_4, a_8, a_{16}, a_{32}, a_{48} = a_{32} \# a_{16}, a_{50} = a_{48} \# a_2\}.$$

Sometimes, other strategies are as good as the binary strategy. For example, you may compute a_{50} as $\{a_2, a_4, a_8, a_{16}, a_{24}, a_{25}, a_{50}\}$, or $\{a_2, a_4, a_8, a_{16}, a_{17}, a_{34}, a_{50}\}$, or $\{a_2, a_4, a_8, a_{12}, a_{24}, a_{25}, a_{50}\}$, or $\{a_2, a_3, a_5, a_{10}, a_{20}, a_{40}, a_{50}\}$, each requiring 7 RIP operations. Can you compute a_{50} using fewer than 7 RIP operations?

Other times, a strategy better than the binary exists! For example, the binary strategy obtains a_{31} as $\{a_2, a_4, a_8, a_{16}, a_{24}, a_{28}, a_{30}, a_{31}\}$ requiring 8 RIP operations. But it is quicker to compute a_{31} as $\{a_2, a_3, a_5, a_{10}, a_{20}, a_{30}, a_{31}\}$ requiring 7 RIP operations.

Of course, if one permits an inverse running inner product (IRIP) operation (defined naturally by allowing subtraction), one can reduce the total number of operations further. For example, if an IRIP operation is permitted then we can compute a_{31} as $\{a_2, a_4, a_8, a_{16}, a_{32}, a_{31}\}$, requiring 5 RIP operations and one IRIP operation for a total of 6 operations.

We leave to the interested reader to discover the least number of RIP operations or mixed RIP and IRIP operations necessary to construct a_k for any given $k \geq 2$.