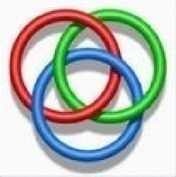
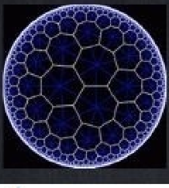


Blackboard

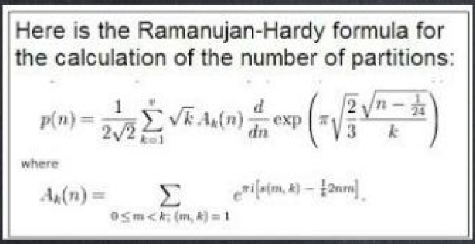
Issue 7

MTA (I)



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

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

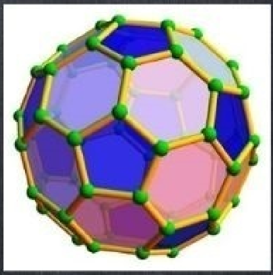


Here is the Ramanujan-Hardy formula for the calculation of the number of partitions:

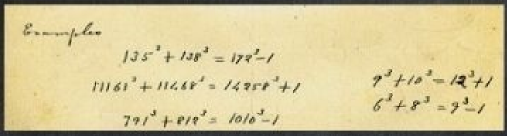
$$p(n) = \frac{1}{2\sqrt{2}} \sum_{k=1}^{v} \sqrt{k}\, A_k(n) \frac{d}{dn} \exp\left(\pi \sqrt{\frac{2}{3}} \frac{\sqrt{n - \frac{1}{24}}}{k}\right)$$

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+ ... = 1



Examples

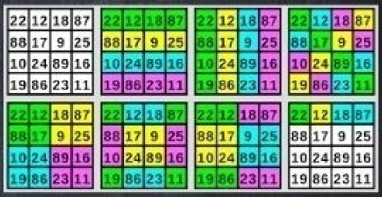
$135^3 + 138^3 = 172^3 - 1$

$11161^3 + 11468^3 = 14258^3 + 1$

$791^3 + 812^3 = 1010^3 - 1$

$9^3 + 10^3 = 12^3 + 1$

$6^3 + 8^3 = 9^3 - 1$



22	12	18	87
88	17	9	25
10	24	89	16
19	86	23	11

22	12	18	87
88	17	9	25
10	24	89	16
19	86	23	11

22	12	18	87
88	17	9	25
10	24	89	16
19	86	23	11

22	12	18	87
88	17	9	25
10	24	89	16
19	86	23	11

22	12	18	87
88	17	9	25
10	24	89	16
19	86	23	11

22	12	18	87
88	17	9	25
10	24	89	16
19	86	23	11

22	12	18	87
88	17	9	25
10	24	89	16
19	86	23	11

22	12	18	87
88	17	9	25
10	24	89	16
19	86	23	11

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

$$\begin{tikzcd}
 & 0 \arrow[r] \arrow[d] & \mathbb{Z}^2 \arrow[r, "\begin{pmatrix} 3 & 0 \\ 0 & 2 \end{pmatrix}"] \arrow[d, "\begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 2 \end{pmatrix}"'] & 3\mathbb{Z} \oplus 2\mathbb{Z} \arrow[r] \arrow[d] & {} \\
{} \arrow[r] & \mathbb{B}^a_d \arrow[r] & \mathbb{Z}^4 \arrow[r, "\begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}"] & \mathbb{Z}^2 \arrow[r] & {}
\end{tikzcd}$$

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

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 7

August 2024

Contents

Editorial

In this issue of Blackboard, the readers have a rich variety of articles to read, ranging from those written by a 11-year student to series articles penned by our editors.

Our youngest author Advay Misra describes combinatorial optimization principles which enable us to understand the mathematical underpinnings of optimization of sorting or grouping problems. Swastika Dey, an undergraduate student, considers a problem posed in the Regional Mathematical Olympiad competition in 2001. She solves it, and her method of proof makes it possible to obtain generalizations also. Prithwijit De, who is the National Co-ordinator in charge of the Mathematical Olympiad Program in our country, has written a delightful article on problems involving polynomials with integer coefficients.

Shailesh Shirali continues his series on the history of Calculus; the fifth instalment deals with the contributions of James Gregory (1638-1675) who accomplished a lot within his short life. Some of these aspects do not seem to be widely known even to those who keep interest in the history of mathematics. Shirali also has a second article following up on the nine-point circle of a triangle.

One other article which could be used to teach students even at the high school level, addresses the celebrated Bretschneider's formula (1842). This formula gives the area of a convex quadrilateral in terms of the side lengths and the sum of one pair of opposite angles. The original proof needed trigonometry, and later proofs using analytic geometry, vector algebra were given by others. In the article by Jyotirmoy Sarkar here, the proof remarkably uses only the extended Pythagorean theorem, a plane Euclidean geometric result that has been known for at least two and a half millennia. The same author has written another article on an extremization problem where the proof is discovered using a thought process that may be called a dialogue in the spirit of Socrates. In a third article by this author along with Bikas Sinha, the authors describe a fascinating account of a neighbor elimination problem involving random walks; they have dedicated this article to Krishna Athreya, a celebrated Probabilist, who passed away recently.

An extremely captivating topic is that of a mathematical study of egoism versus altruism. Kishan Suthar, a student from IIM Indore, has written on the scenario when Nash equilibrium and Berge equilibrium are compared.

Our fellow editor Anisa Chorwadwala has discussed dot products or more general inner products in an easy informal style; her article is entitled ‘We are what we think we are’.

It also gives me great pleasure in bringing to the notice of the mathematical community—especially the community comprised of mathematics teachers—certain laurels won by our fellow editor Neena Gupta.

In order to honor women who have made fundamental and sustained contributions to the mathematical sciences, the Association for Women in Mathematics (AWM) set up an annual Lecture Series in the name of Emmy Noether, one of the greatest mathematicians, who shaped modern algebra. From 2015, this series is being sponsored by AWM jointly with the American Mathematical Society (AMS). It is a matter of great pride for the Indian mathematical community that Neena Gupta, our fellow editor and a Professor at the Indian Statistical Institute Kolkata, has been selected to deliver the 44th AWM-AMS Emmy Noether Lecture during January 8-11, 2025. I mention in passing that Neena Gupta has also won several other accolades including the Shanti Swarup Bhatnagar Prize, the Ramanujan Prize, and the Ganit Ratna award. We are fortunate to have her in our editorial board.

In this context, a brief description of Emmy Noether’s enormous influence on mathematics is given by Neena Gupta and her doctoral advisor Amartya Kumar Dutta.

We draw attention to teachers that the MTA (India)’s Regional Conference on Mathematical Education is to be held in Cochin University of Science and Technology during September 21-22, 2024. Teachers are encouraged to apply for participation.

All in all, I believe that in this issue of Blackboard, there is something for anyone interested in some aspect of mathematics. I hope the issue brings pleasure in reading and sharing the thoughts communicated in these diverse articles.

B. Sury
Indian Statistical Institute Bangalore
31 August, 2024

1 A Geometric Proof of Bretschneider's Formula

Jyotirmoy Sarkar

Mathematical Sciences, Indiana University Indianapolis, USA
Email: jsarkar@iu.edu

Abstract
The celebrated Bretschneider's formula (1842) gives the area of a convex quadrilateral whose sides are a, b, c, d, and the sum of one pair of opposite angles is 2θ. Its original proof used trigonometry; later proofs used analytic geometry, vector algebra, or complex analysis. Our proof uses the extended Pythagorean theorem, a plane Euclidean geometric result known at least two and a half millennia ago. As such, our proof is accessible to high school students.

Key Words and Phrases: Pythagorean theorem, Extended Pythagorean theorem, Heron's formula, Brahmagupta's formula, Bretschneider's formula, Hobson's version, Coolidge's version, Ptolemy's inequality, Ptolemy's theorem

Mathematics Subject Classification: 51M04

Preamble

Come let our imaginations run wild for a while: Hop on my time machine and let us travel back to 90 CE to visit the Library of Alexandria, Egypt, where the master teacher Heron's mathematics students (and their academic descendants) have assembled for a conference in his honor. While the master taught the celebrated formula for the area of a triangle, given its three sides, people in his inner circle knew that until the day he breathed his last, the master was looking for just as elegant a formula for the area of a convex quadrilateral, given its four sides and the sum of one pair of opposite angles — alas in vain.

Our time machine must return to the present day within six hours, or else we will be propelled into antiquity with no hope of return. Therefore, we can attend only the Saturday morning session of the conference, and present within two hours the much desired formulas for the area of a convex quadrilateral. Moreover, since I will be exhausted from operating the time machine, would you do me a favor by giving the talk? I trust you will do a fantastic job if you carefully learn the proofs by reading this paper. A word of caution: We must take utmost care to use knowledge that the conference attendees already have or can pick up immediately when presented logically.

Thank you for agreeing to partner with me in this extraordinary trip. Let us begin by collecting the chronological development of the requisite formulas in Section 1. Then in Sections 2–4, we will document what we plan to present at the first century conference — taking liberty to give some references and remarks for our modern day readers. If those good folks of the first century can understand our talk, we hope our modern-day high school students will also. In Section 5, we conclude with some teaching suggestions for our time, including some exercises for the diligent student.

1 Arrange Some Results Chronologically

Given the three sides BC, CA, AB of lengths a, b, c, respectively, a triangle ABC is well defined and has an area given by Heron's (c. 60 CE) formula (see [1, 2])

$$\Delta(a,b,c) = \sqrt{t(t-a)(t-b)(t-c)}, \tag{1}$$

where $t = (a+b+c)/2$ is the semi-perimeter of the triangle.

However, given the four sides, a quadrilateral is not well defined: The four sides allow for a continuum of quadrilaterals as the sum of one pair of opposite angles increases while the sum of the other pair decreases (the sum of all four angles being 2π), or equivalently as one diagonal increases in length and the other decreases. One may ask: Among all such quadrilaterals, which one has the largest area? Indeed, the area is the largest when the quadrilateral is convex and cyclic (that is, the sum of each pair of opposite angles is π). There are many proofs of this claim. See [3] for an Euclidean plane geometric proof.

For a cyclic, convex quadrilateral, of lengths a, b, c, d and semi-perimeter $s = (a+b+c+d)/2$, the area is given by Brahmagupta's (c. 598–c. 668 CE) formula (see [1])

$$K_{cyclic} = \sqrt{(s-a)(s-b)(s-c)(s-d)}. \tag{2}$$

What about the area of a more general quadrilateral, not necessarily cyclic?

Bretschneider's formula (1842). See [2]. The area of a convex quadrilateral $ABCD$ whose sides AB, BC, CD, DA are of lengths a, b, c, d, semi-perimeter $s = (a+b+c+d)/2$, and the sum of one pair of opposite angles is 2θ, is given by

$$K_B = \sqrt{(s-a)(s-b)(s-c)(s-d) - abcd\,(r+1)/2}. \tag{3}$$

where, in our modern day language, $r = \cos 2\theta$, so that $(r+1)/2 = \cos^2\theta$. However, the first century mathematicians would prefer that we interpret r as the ratio of the directed length of the projection of one line segment onto another line to the length of the first line segment when the two lines meet at an angle 2θ. See Figure 1.

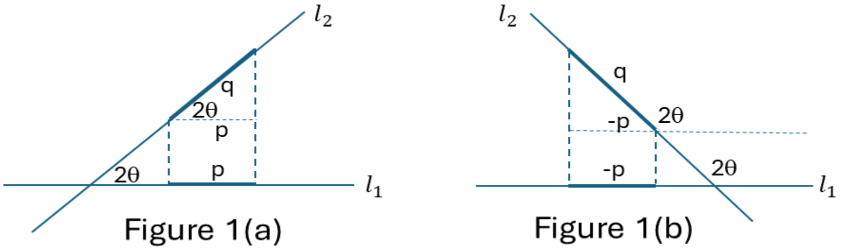


Figure 1(a)

Figure 1(b)

Figure 1: Lines l_1 and l_2 intersect at an angle 2θ. If segment q on l_2 is projected on to segment p on l_1, then r is defined as $p/q = \cos(2\theta)$. (a) When $2\theta < \pi/2$, we have $p > 0$, so $r > 0$; and (b) when $2\theta > \pi/2$, we have $p < 0$, or $r < 0$.

Formula (3) was proved in 1842 by Carl Anton Bretschneider and confirmed by Friedrich Strehlke in the same volume of *Archiv* (see [2, 9]). The formula was also independently discovered by Karl Georg Christian von Staudt in 1842. All these proofs use trigonometry. For a modern-day proof, still using trigonometry, see [4, 5]. For a vector algebra proof, without trigonometry, see [6]. We, on the other hand, provide an Euclidean geometric proof.

Hobson's version of Bretschneider's formula (1897). See [10]. The area of a convex quadrilateral $ABCD$, whose sides AB, BC, CD, DA are of lengths a, b, c, d, and diagonals AC, BD are of lengths p, q, respectively, is given by

$$K_H = \frac{1}{4}\sqrt{4p^2q^2 - (a^2 - b^2 + c^2 - d^2)^2}. \tag{4}$$

Hobson's book [10] on trigonometry naturally proves (4) using trigonometry. A complex analytic proof is given in [11]. Here, we give a Euclidean geometric proof.

Coolidge's version of Bretschneider's formula (1939). See [12]. Given the same information as in Hobson's version, and letting $s = (a + b + c + d)/2$ denote the semiperimeter of the quadrilateral, the area of quadrilateral $ABCD$ is given by

$$K_C = \sqrt{(s-a)(s-b)(s-c)(s-d) - [(ac+bd)^2 - p^2q^2]/4}. \tag{5}$$

We invite the interested reader to verify right away the algebraic equivalence between (5) and (4). Later, they can compare their verification with ours given in Section 3.

The equivalence of (5) and (3) would follow from $2abcd(r + 1) = (ac + bd)^2 - p^2q^2$, or equivalently from

$$p^2q^2 = a^2c^2 + b^2d^2 - 2abcd\, r, \tag{6}$$

where $r = \cos 2\theta$, or as explained in Figure 1, the ratio of the directed length of the projection and the length of the original line segment. The fascinating feature of (6) is that it relates the six distances within the six pairs of vertices to the sum of one pair of opposite angles of $ABCD$. See a trigonometric proof of (6) in [10] and a complex analytic proof in [13]. Here, we provide a Euclidean geometric proof of (6). This is the crux of our new discovery.

When specialized to a cyclic quadrilateral (for which $2\theta = \pi$; hence, $\cos 2\theta = -1$), Bretschneider's formula (3) yields Brahmagupta's formula (2). If $d = 0$ (so that $p = c$ and $q = a$), then clearly (6) holds, and each of (2), (3), (4), (5), yields Heron's formula (1).

For none of the versions of Bretschneider's formula (3), (4) or (5), did we find in the literature a Euclidean geometric proof. To fill the gap, in Section 3, we prove (4), which is algebraically equivalent to (5) as mentioned above. Then, in Section 4, we prove (6). Both proofs rely on the extended Pythagorean theorem, which has been known for over two and a half millennia.

2 The Pythagorean Theorem and its Extension

The well-known Pythagorean theorem (PT) was discovered in various places between 1800 BCE and 600 BCE. It appears in Euclid's *Elements* (see [14]) as Proposition I.47. It has over 370 independent proofs (see [15])! We encourage the readers to learn as many proofs as they wish and rank them according to their preferences. For their benefit, we reproduce here the proof based on similar triangles (for that is a key idea in our proof of (6)).

Theorem 1 (The Pythagorean Theorem). In a right-angled $\triangle ABC$ with right angle at C, if we let $a = BC, b = CA, c = AB$, then $c^2 = a^2 + b^2$.

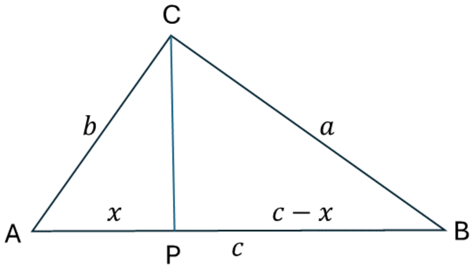


Figure 2: The altitude on the hypotenuse splits the given right triangle into two triangles each of which is similar to the given right triangle, and hence to each other.

Proof. See Figure 2. Drop a perpendicular CP onto AB. Let $x = AP$, then $BP = c - x$. Note that $\triangle ACP$ and $\triangle ABC$ are similar since both are right-angled and the angle at A is common to them. Hence, $x : b = b : c$, or $b^2 = cx$. Likewise, $\triangle BCP$ and $\triangle BAC$ are similar since both are right-angled and the angle at B is common to them. Hence, $(c - x) : a = a : c$, or $a^2 = c(c - x)$. Therefore, by adding, we get $a^2 + b^2 = c(c - x) + cx = c^2$. □

The PT holds only for a right-angled triangle. For a not-necessarily-right-angled triangle, what is the corresponding result? The extended Pythagorean theorem (EPT) answers this question. The EPT appears in Euclid's *Elements* (see [14]) as Propositions II.12 and II.13 for obtuse $\measuredangle C$ and acute $\measuredangle C$, respectively.

Theorem 2 (Extended Pythagorean Theorem (EPT)). In a $\triangle ABC$, let $a = BC, b = CA, c = AB$ and x be the projection of CA onto CB, then

$$c^2 = a^2 + b^2 - 2ax. \tag{7}$$

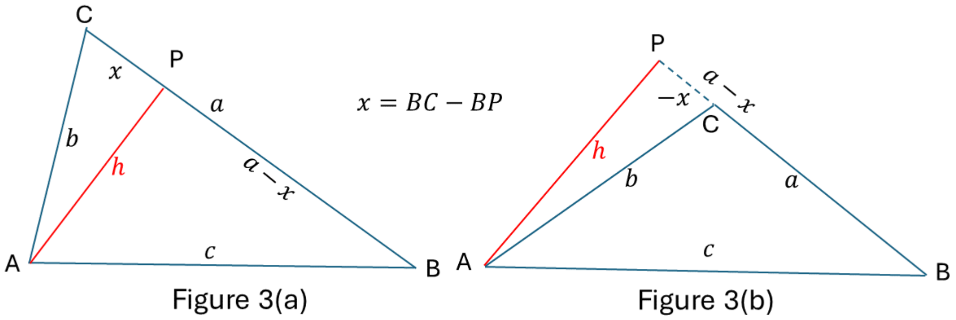


Figure 3: From A, drop a perpendicular AP to CB: (a) P is inside CB when $\measuredangle C$ is acute; (b) P is outside CB when $\measuredangle C$ is obtuse.

Proof. If $\measuredangle C$ is a right angle, then $x = 0$, and the result holds by the PT. Otherwise, let AP be perpendicular to BC. Then in both Figures 3a and 3b, $BP = a - x$. By the PT applied to right-$\triangle ABP$, we have $c^2 = h^2 + (a - x)^2$. Also, by the PT applied to right-$\triangle ACP$, we have $b^2 = h^2 + (\pm x)^2$. Therefore, by subtraction, we have $c^2 - b^2 = a^2 - 2ax$. □

Remark 1. The EPT was renamed by Jamshid al-Kashi, a 15th century Persian mathematician, as the *law of cosines*, for by then trigonometry had become a more convenient language: Writing $x = b \cos C$, (7) became

$$c^2 = a^2 + b^2 - 2ab \cos C. \tag{8}$$

However, being equivalent to (7), (8) is not truly a trigonometric result!

3 Proving Heron's Formula (1), Hobson's Version (4), and Coolidge's Version (5)

Both Heron's formula (1) and Hobson's version (4) are corollaries to Theorem 2. Whereas it is well known that Heron's formula follows from the EPT (7), stunningly, it is little known that Hobson's version also follows from the EPT almost identically! The small difference is that Heron's formula uses (7) once, but Hobson's version uses it *twice*.

Corollary 2.1 (Heron's formula). The area of a triangle with sides a, b, c is given by (1).

Proof. Refer to Figures 3a and 3b where $h^2 = b^2 - x^2$. The area of $\triangle ABC$ is $\Delta = ah/2$. Quadrupling, squaring, and using the EPT (7), we get

$$\begin{aligned} 16\Delta^2 &= (2a)^2(b^2 - x^2) = (2ab)^2 - (2ax)^2 \\ &= (2ab)^2 - (a^2 + b^2 - c^2)^2. \end{aligned}$$

By factoring the difference of two squares, and factoring each factor again, we have

$$\begin{aligned} 16\Delta^2 &= [2ab + a^2 + b^2 - c^2]\,[2ab - a^2 - b^2 + c^2] \\ &= [(a+b)^2 - c^2]\,[-(a-b)^2 + c^2] \\ &= (a+b+c)(a+b-c)(a-b+c)(-a+b+c) \\ &= 16t(t-c)(t-b)(t-a), \end{aligned}$$

completing the proof of (1).

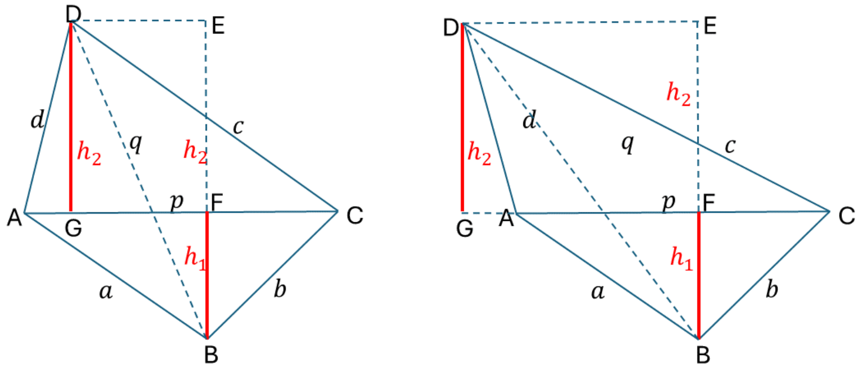


Figure 4(a)

Figure 4(b)

Figure 4: Drop perpendiculars BF and DG to AC. Complete the rectangle $DGFE$. Relative to vertex A, points F and G are (a) on the same side, or (b) on opposite sides.

Corollary 2.2 (Hobson's version). A convex quadrilateral $ABCD$ with four sides a, b, c, d and diagonals p, q has area given by (4).

Proof. Refer to Figure 4. Drop perpendiculars $h_1 = BF$ and $h_2 = DG$ on to AC. Draw rectangle $DGFE$. By the PT applied to right-$\triangle DBE$, we have $(h_1+h_2)^2 = q^2 - DE^2 = q^2 - GF^2$. The area of quadrilateral $ABCD$ is $K = p(h_1+h_2)/2$. Quadrupling, squaring, and using the EPT (7) *twice* — once in $\triangle ABC$ and again in $\triangle ADC$ — we get

$$\begin{aligned}
16K_H^2 &= (2p)^2(q^2 - GF^2) = (2p)^2(q^2 - \{AF - AG\}^2) \\
&= (2pq)^2 - \{2p \cdot AF - 2p \cdot AG\}^2 \\
&= 4p^2q^2 - \{(a^2 + p^2 - b^2) - (d^2 + p^2 - c^2)\}^2 \\
&= 4p^2q^2 - (a^2 - b^2 + c^2 - d^2)^2,
\end{aligned}$$

completing the proof of Hobson's version. □

Corollary 2.3 (Coolidge's version). A convex quadrilateral $ABCD$ with four sides a, b, c, d and diagonals p, q has area given by (5).

Proof. Suffices to prove that (5) follows from (4), which involves algebraic simplification, very much like that in Heron's formula (1) shown in Corollary 2.1. Adding and subtracting $4(ac + bd)^2$ to the expression for $16K_H^2$ in the proof of Corollary 2.2, and pairing terms appropriately, we have

$$16K_H^2 = [4(ac + bd)^2 - (a^2 - b^2 + c^2 - d^2)^2] - 4[(ac + bd)^2 - (pq)^2].$$

Thereafter, in the first pair by factoring the difference of two squares, and factoring each factor again, we have

$$\begin{aligned} & 4(ac+bd)^2-(a^2-b^2+c^2-d^2)^2 \\ = & \left[2(ac+bd)+a^2-b^2+c^2-d^2\right]\left[2(ac+bd)-a^2+b^2-c^2+d^2\right] \\ = & \left[(a+c)^2-(b-d)^2\right]\left[-(a-c)^2+(b+d)^2\right] \\ = & \left[(a+b+c-d)(a-b+c+d)\right]\left[(a+b-c+d)(-a+b+c+d)\right] \\ = & 16(s-d)(s-b)(s-c)(s-a). \end{aligned}$$

Hence,

$$16K_H^2=16(s-a)(s-b)(s-c)(s-d)-4[(ac+bd)^2-(pq)^2]=16K_C^2,$$

showing that (5) is equivalent to (4). Hence, (5) holds true. □

4 The EPT Proves Bretschneider's Formula (3)

The next theorem uses the EPT to prove (6), and hence (3).

Theorem 3 (Segments-Angle Sum Identity) If a convex quadrilateral $ABCD$ has sides $a=AB, b=BC, c=CD, d=DA$, diagonals $p=AC, q=BD$, and sum of one pair of opposite angles $2\theta=\measuredangle ABC+\measuredangle ADC$, then identity (6) holds. That is,

$$p^2q^2=a^2c^2+b^2d^2-2abcd\,r$$

where r is the ratio of the projection of a line segment of length bd onto a line that makes an angle 2θ with the segment.

Proof. In Figure 5(a), without loss of generality, assume $b=1$. Holding $\triangle ABC$ fixed, rotate $\triangle ACD$ about the point C until CD aligns with the ray CB and $\triangle ACD$ is relocated to $\triangle A'CD'$. See Figure 5(b). Choose point T on ray CA so that $D'T$ is parallel to BA. Then $\triangle D'CT$ and $\triangle BCA$ are similar with the former being a c-multiple of the latter, since $b=1$. Hence, $TD'=c\cdot AB=ac$ and $TC=c\cdot AC=pc$.

Next, note that $\triangle A'CT$ is similar to $\triangle BCD$ because

$$\measuredangle A'CT=\measuredangle A'CD'+\measuredangle D'CT=\measuredangle DCA+\measuredangle BCA=\measuredangle BCD,$$

and $TC:CA'=pc:p=c:1=DC:CB$. Hence, $TA':DB=A'C:BC=p:1$. Or $TA'=p\cdot DB=pq$.

Finally, note that $\measuredangle TD'A'=\measuredangle TD'C+\measuredangle CD'A'=\measuredangle ABC+\measuredangle CDA=2\theta$. Hence, by applying the EPT to $\triangle TD'A'$, we have

$$(pq)^2=(ac)^2+(bd)^2-2(ac)(bd)\,r$$

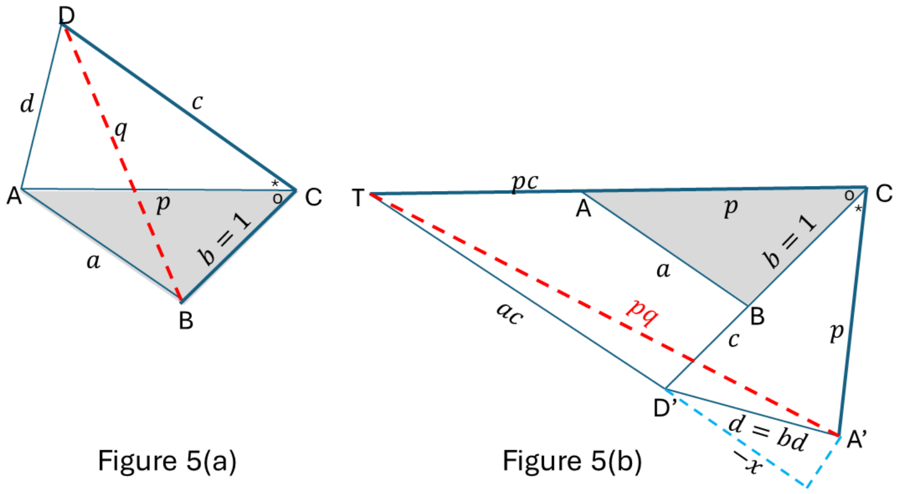


Figure 5(a)

Figure 5(b)

Figure 5: Rotate $\triangle ACD$ about C to relocate it at $\triangle A'CD'$ such that CD' aligns with ray CB. Then dilate $\triangle BCA$ to form $\triangle D'CT$.

where $r = -x/(bd)$ with x being the directed length of the projection of $D'A'$ onto TD'. Thus, we have proved (6). □

Remark 2. In our modern-day language, we write $r = \cos 2\theta$.

The following corollary is immediate.

Corollary 3.1 (Ptolemy's theorem). In the proof of Theorem 3, by using the triangle inequality in $\triangle TD'A'$, we have Ptolemy's inequality

$$pq \leq ac + bd. \tag{9}$$

Furthermore, $pq = ac + bd$ if and only if $\measuredangle TD'A' = 2\theta = \pi$; that is, $ABCD$ is cyclic. This is called Ptolemy's theorem.

5 Summary and Teaching Recommendation

In Section 2, we proved PT and the EPT. In Section 3, we proved Hobson's version (4) using the EPT, and recalled its equivalence to Coolidge's version (5) by successive factoring. In Section 4, in Theorem 3, we proved (6) using the EPT, thereby establishing that Coolidge's version (5) is equivalent to Bretschneider's formula (3). Thus, we have proved Bretschneider's formula (3) using Euclidean plane geometry and some simple algebra.

We did not find in the literature any Euclidean geometric proof of any of the three versions (3)–(5) of Bretschneider's formula. Therefore, we produced Euclidean geometric proofs based on the Pythagorean theorem and its extension, results that were known at least five centuries before Heron discovered his formula (1) for the area of a triangle in c. 60 CE!

Here is our recommendation: Henceforth, to learn Bretschneider's formula, students need not wait until they learn trigonometry, analytical geometry, vector algebra or complex analysis. High school students should learn both Heron's formula (1) and Hobson's version (4) of Bretschneider's formula soon after learning the PT and the EPT (7), which today is known as the law of cosines (8). Next, they will learn Coolidge's version (5) of Bretschneider's formula by successive factoring. Not too long after that, they will be ready to learn the Bretschneider's formula (3) by using Theorem 3, which proves (6). They will immediately know Brahmagupta's formula (2), Ptolemy's inequality (9) and Ptolemy's theorem. They will also know that the area of a quadrilateral, given its four sides, is maximized when opposite angles are supplementary, or equivalently, when the quadrilateral is cyclic.

To the diligent reader, we assign the following exercises involving convex rectilinear figures.

1. Find the area of $\triangle ABC$ in which (i) $a = BC, b = CA, \gamma = \measuredangle C$; and (ii) $\alpha = \measuredangle A, \beta = \measuredangle B, c = AB$. Specialize to the case when $\alpha = \beta$. In (i) and (ii), find the missing side(s) and angle(s) of $\triangle ABC$.

2. Let P be a point inside a convex quadrilateral $ABCD$ such that $a = PA, b = PB, c = PC, d = PD$ and $\alpha = \measuredangle APB, \beta = \measuredangle BPC, \gamma = \measuredangle CPD, \delta = \measuredangle DPA$. Find the area of quadrilateral $ABCD$. Specialize to the case when $\alpha = \gamma$ and $\beta = \delta$.

3. In a convex quadrilateral $ABCD$, let the midpoints of AB, BC, CD, DA be denoted by E, F, G, H, respectively. Show that the area of $EFGH$ is half that of $ABCD$.

4. On the board is drawn a convex quadrilateral and its diagonals. What is the fewest number of segments or angles you must measure to find its area?

5. Find the area of a convex, cyclic pentagon $ABCDE$ given $a_1 = AB, a_2 = BC, a_3 = CD, a_4 = DE, a_5 = EA$.

6. Find the area of a convex pentagon $ABCDE$ given $a = AB, b = BC, c = CD, d = DE, e = EA, u = AC, v = CE, w = EB, x = BD, y = DA$.

7. Find the area of a convex pentagon $ABCDE$ given $a = AB, b = BC, c = CD, d = DE, e = EA, \alpha = \measuredangle EAB$ and either (i) $\beta = \measuredangle ABC$, or (ii) $\gamma = \measuredangle BCD$.

Acknowledgement

I thank my colleagues Patrick Morton and Michal Misiurewicz for some discussions, literature searches and for giving feedback on an earlier draft.

Bibliography

[1] Nelsen, Roger B.: Heron's formula via proof without a word. *The College Mathematics Journal* **32**(4), 290 (2001)

[2] Wikipedia, the Free Encyclopedia: Heron's formula.
https://en.wikipedia.org/wiki/Heron'sformula

[3] Sarkar, Jyotirmoy: The largest quadrilateral Is cyclic: A new geometric proof. *The College Mathematics Journal* **54**(4), 378–384 (2023)

[4] Wikipedia, the Free Encyclopedia: Brahmagupta's formula.
https://en.wikipedia.org/wiki/Brahmagupta'sformula

[5] Bretschneider, Carl Anton: Untersuchung der trigonometrischen Relationen des geradlinigen Viereckes. (Investigation of the trigonometric relations of the rectilinear quadrilateral.) *Archiv der Mathematik und Physik* **2**, 225–261 (1842)

[6] Strehlke, Friedrich: Zwei neue Satze vom ebenen und spharischen Viereck und Umkehrung des Ptolemaischen Lehrsatzes. (Two new theorems about the plane and spherical quadrilateral and reversal of the Ptolemaic theorem.) *Archiv der Mathematik und Physik* **2**, 323-326 (1842)

[7] Wikipedia, the Free Encyclopedia: Bretschneider's formula.
https://en.wikipedia.org/wiki/Bretschneider'sformula

[8] Pritsker, Boris: *Geometrical Kaleidoscope.* Dover, New York 49 (2017)

[9] Weisstein, Eric W.: Bretschneider's Formula. From MathWorld–A Wolfram Web Resource. https://mathworld.wolfram.com/BretschneidersFormula.html

[10] Hobson, Ernest William: *A Treatise on Plane Trigonometry.* Cambridge University Press, Cambridge, England 204–205 (1897). Reprinted as *A Treatise on Plane and Advanced Trigonometry.* Dover, New York 204–205 (1957)

[11] Lopes, Andre van Borries, dos Santos, Rogerio Cesar and dos Reis, Paulo Rodrigo Alves: Bretschneider's Formula Via Complex Numbers. Unpublished manuscript (2024)

[12] Coolidge, Julian Lowell: A historically interesting formula for the area of a quadrilateral. *The American Mathematical Monthly* **46**(6), 345–347 (1939)

[13] Andreescu, Titu and Andrica, Dorin: *Complex Numbers from A to ... Z*, 2 ed. Birkhauser, Boston, MA (2014)

[14] Euclid; Heath, Thomas Little, ed. and tr.; Heiberg, Johan Ludvig: *The thirteen books of Euclid's Elements*. University Press, Cambridge (1908)

[15] Loomis, Elisha Scott: *The Pythagorean Proposition*, 2 ed. Edwards Brothers, Ann Arbor, Michigan (1940). Reissued by the National Council of Teachers of Mathematics (1968).

Appendix: Answers

Do not read the answers listed below until after you have tried the exercises given in Section 5 on your own or in collaboration with a like-minded math enthusiast:

1.(i) $\Delta = (1/2)ab\sin\gamma$; first find c by using the EPT, then $\alpha = \arcsin((a/c)\sin\gamma)$ and $\beta = \arcsin((b/c)\sin\gamma)$.

1.(ii) $\Delta = (1/2)c^2 \sin\alpha\, \sin\beta/\sin(\alpha+\beta)$, $a = c\sin\alpha/\sin(\alpha+\beta)$, $b = c\sin\beta/\sin(\alpha+\beta)$, and $\gamma = \pi-\alpha-\beta$. When $\alpha = \beta$, $\Delta = (1/4)c^2\tan\alpha$, $a = b = c\sec\alpha/2$, and $\gamma = \pi - 2\alpha$.

2. $K = (1/2)[ab\sin\alpha + bc\sin\beta + cd\sin\gamma + da\sin\delta]$. When $\alpha = \gamma$ and $\beta = \delta$, A, P, C are colinear, and B, P, D are colinear, $K = (1/2)(a+c)(b+d)\sin\alpha = (1/2)pq\sin\alpha$, where $p = a+c, q = b+d$ are the lengths of the diagonals.

3. Indeed, $EFGH$ is a parallelogram with $EF = HG = AC/2$. Hence, $\triangle BEF = \triangle DGH = K/4$. Likewise, $\triangle AEH = \triangle CFG = K/4$. Hence, the area of $EFGH$ is $K/2$.

4. Measure only three items: the two diagonals p, q and the angle α between them. Then as in (2), $K = (1/2)pq\sin\alpha$.

5. First find the circum-diameter by solving $\sum_{i=1}^{5} \arcsin(a_i/D) = \pi$. Then the area of the pentagon is $(1/2)\sum_{i=1}^{5} a_i\sqrt{D^2 - a_i^2}$.

6. Add the areas of $\triangle ABC, \triangle CDE, \triangle ECA$, using Heron's formula (1).

7.(i) Let AE and BC intersect at T. Find AT, BT. Using $CT, ET, \measuredangle CTE$ find CE using the EPT. Then the area of the pentagon is $|\triangle CTE - \triangle BTA| + \triangle CDE$.

7.(ii) Find BE and BD using the EPT. Then add the areas of $\triangle ABE, \triangle BCD, \triangle BDE$.

2 When Nash meets Berge: A Mathematical Understanding of Egoism and Altruism

Kishan Suthar

Indian Institute of Management Indore
Prabandh Shikhar, Rau - Pithampur Rd
Indore, Madhya Pradesh 453556.
Email: f21kishans@iimidr.ac.in

Abstract
Sharing economy is becoming more and more relevant nowadays, where individuals' choices are crucial in shaping social outcomes. When a set of individuals are rational, their interaction can be mathematically modeled as a game. Non-cooperative (or strategic) games are usually solved using the concept of Nash equilibrium, when the individuals select self-optimizing options from available alternatives. In this article, we first outline a known proof of the existence of Nash equilibrium using Sperner's lemma, and use the same to provide a simple proof of the existence of Nash equilibrium for zero-sum games. This study also reviews an unconventional solution concept known as the Berge equilibrium, based on society's moral golden rule. A comparative analysis between Nash and Berge equilibria reveals that the latter presents a more suitable solution concept for modeling scenarios with altruistic individuals, as long as individual rationality is maintained. In conclusion, this paper highlights the limitations of the Berge equilibrium and suggests potential future research opportunities that could fortify and extend the theoretical underpinnings of Berge's altruism framework.

1 Introduction

Studies on social dilemma are primarily concerned with resolving the inherent tradeoff that exists between an individual's pursuit of self-optimizing outcomes, commonly referred to as self-centered behaviour, versus the individual considering others' optimal outcomes, often termed altruistic behaviour [Wade-Benzoni and Tost, 2009]. A fundamental aspect of the mathematical analysis applied to address social dilemmas entails the conversion of individual preferences into real-valued utilities, a process that allows for a quantifiable evaluation of these preferences. Expanding on the pioneering contributions of John von Neumann and Oskar Morgenstern, as presented in "Theory of Games and Economic behaviour"[Von Neumann and Morgenstern, 1947], mathematicians and economists have made substantial contributions to the field by investigating and resolving the challenges posed by games with different characteristics. In this context, a "game" denotes the dynamic interaction among individuals (or players) engaged in the decision-making process.

In a seminal study ([v. Neumann, 1928]), Von Neumann demonstrated that a "two-player zero-sum game" always possesses an optimal solution. John Nash was the first to explore "n-player nonzero-sum games" and proved that there always exists an equilibrium point in games with a finite set of players, each player having a finite set of strategies [Nash, 1951]. The Nash equilibrium is a widely used concept in economics, social science, management, engineering, biology, and many more fields. Nash's work assumed that each player holds complete knowledge about other players' strategies. Extending this research, [Harsanyi, 1967] introduced the concept of an incomplete information game where each player has some belief(s) about other players' strategies (or payoffs) and proposed the concept of Bayesian Nash equilibrium in this context.

The mathematician Claude Berge proposed the concept of equilibrium in games with altruistic players [Berge, 1957], unlike von Neumann's and Nash's works where players aim to optimize their respective payoff functions. The equilibrium concept that Berge initially introduced is the concept of the $P/K-$equilibrium where P and K are coalitions of players, coalition K wanting to optimize the payoffs of coalition P's players. V. I. Zhukovskiy [Zhukovskiy, 1985] extended Berge's work to individual players and defined the Berge equilibrium.

The Berge equilibrium is based on the society's moral golden rule of "doing good for others as you would want them to do to you", a notable departure from classical game theory. The moral golden rule is also called "reciprocal altruism" and is applicable in numerous social scenarios. For many decades, Berge's work encountered significant limitations due to language barriers, rendering its application relatively confined to the Russian context. Nonetheless, recent research by [Larbani and Zhukovskii, 2017] has played a pivotal role in disseminating and promoting Berge's work with broader applications.

This article undertakes a comparative examination of the works of Nash and Berge, seeking to unearth additional opportunities for their practical implementation. To facilitate this comparison, we begin by elucidating Nash's groundbreaking theorem regarding the existence of an equilibrium in non-cooperative games, subsequently contrasting this concept with the Berge equilibrium. We offer concise yet comprehensive explanations of both solution frameworks and provides illustrative examples to clarify their implications. For the sake of simplicity and brevity, we limit our examples to two-player games.

2 Non-cooperative Games and Nash Equilibrium

A game theoretic approach to solving a decision-making problem consists of a number of players in a game, the information each player has regarding the game, the set of strategies available to each player, and the payoff function or utility linked to each player's choice of strategy (or strategies).

Generally, game-theoretic models are broadly classified into non-cooperative and cooperative games based on the problem's context and interactions between the players. In this article, we confine our discussion to non-cooperative games. For an exposition to cooperative games, we refer the readers to the book by von Neumann and Morgenstern [Von Neumann and Morgenstern, 1947] as well as to the foundational work by Shapley [Shapley et al., 1953].

In this section, we first define some terminology necessary to understand John Nash's work [Nash, 1951] on n-player non-cooperative games. Then, we present a simple proof of existence of the Nash equilibrium for two-player zero-sum games. In Section 3, we define the Berge equilibrium and compare it with the Nash equilibrium.

Definition 1. *A non-cooperative game, G, in strategic (or normal) form consists of*

- *a finite set of players, $N = \{1, 2, ..., n\}$,*
- *for each player, a finite set of action-choices; S_i being the set of action-choices of Player i, and*
- *for each player, a payoff function; $p_i : S \to \mathbb{R}$ being the payoff function of Player i, which maps every $n-$tuple of action-choices from the set $S = S_1 \times S_2 \times ... \times S_n$ to a real number.*
- *the objective of each player. That is, for each player, is that player a maximizer (who wants to obtain maximum possible payoff) or a minimizer (wanting to minimize her payoff function).*

Terminology and Notation:

- *An action choice is called a pure strategy.*
- *An $n-$tuple of action choices is called a pure strategy profile.*
- *A mixed strategy, X_i is a probability distribution over the strategy set S_i of Player i.*
- *We may extend strategy profiles and payoff functions to include mixed strategies too; the set $X = X_1 \times X_2 \times ... \times X_n$ being the set of $n-$tuples of (mixed) strategies (or (mixed) strategy profiles) $x = (x_1, x_2, ...x_n)$, and the function $p_i : X \to \mathbb{R}$ being the expected payoff function of Player i.*
- *We use $p_i(x_i, x_{N\backslash i})$ to denote the payoff of Player i when she plays strategy $x_i \in X_i$ against $(N\backslash i)$[1] players' strategies $x_{N\backslash i} \in X_{N\backslash i}$.*
- *G is called a complete information game (or a game with complete information) if all players know the sets of strategies, payoff functions and objectives of each other.*

We now state Nash's theorem (Theorem 1) which posits the existence of an equilibrium point in any finite n-player game wherein each player has a strategy that optimizes their respective (expected) payoff, taking into account the equilibrium strategies of the other players. Ever since, this equilibrium point has been called the "Nash equilibrium". We formally define the Nash equilibrium below and, though Nash just called it an "equilibrium point", we start using the term "Nash equilibrium" from Theorem 1 itself.

Definition 2. *Suppose G is a non-cooperative, complete information game in strategic form where all players are maximizers. A strategy profile $x^* = (x_1^*, x_2^*, ...x_n^*)$ of the players is a Nash equilibrium if*

$$p_i(x_i, x^*_{N\backslash i}) \leq p_i(x_i^*, x^*_{N\backslash i}), for\ all\ x_i \in X_i$$

In other words, $for\ all\ i \in N$, if Player i plays strategy x_i^* and $N\backslash i$ players play $x^*_{N\backslash i}$, then $x^* = (x_1^*, x_2^*, ...x_n^*)$ is said to be a Nash equilibrium (N.E., in short) if no individual player $i \in N$ has any incentive to deviate from their respective N.E. strategy x_i^*.

Theorem 1. *[Nash, 1951] Every non-cooperative, complete information game with a finite number of players, where each player has a finite number of pure strategies, has at least one Nash equilibrium point.*

[1] For ease of notation, we denote singleton $\{i\}$ as i and the set N without player i as $N\backslash i$.

To understand a simple proof of Nash's theorem (Theorem 1), we need two additional concepts, namely Brouwer's fixed point theorem and Sperner's lemma which are explained in the following subsection, after defining a few other mathematical terms and notations that are required.

2.1 Brouwer's Fixed Point Theorem and Sperner's Lemma

[Brouwer, 1911] showed that any continuous mapping from a compact (that is, closed and bounded), convex set to itself has a fixed point (Definition 3). [Kakutani, 1941] generalized Brouwer's fixed point theorem from point-to-point functions to set-valued functions. [Nash, 1951] proved the existence of an equilibrium point using Kakutani's fixed point theorem [Kakutani, 1941] and used Brouwer's fixed point theorem [Brouwer, 1911] in his Ph.D. thesis. Here, we focus only on Brouwer's fixed point theorem and state it using motivation from [Katz, 2017]. Interested readers may read more about the applications of fixed point theorems to game theory in [Border, 1985].

Definition 3. *For a given function $f : X \to X$, $x \in X$ is a fixed point of f if $f(x) = x$.*

In Theorem 2, we state Brouwer's fixed point theorem applied to the r-dimensional unit ball, $B^r = \{x \in \mathbb{R}^r | \sum_{i=1}^{r} x_i^2 \leq 1\}$, which is compact and convex.

Theorem 2. *[Brouwer, 1911] Suppose $f : B^r \to B^r$ is a continuous function. Then, f has a fixed point. In other words, there exists $x \in B^r$ such that $f(x) = x$.*

This theorem can be proved in different ways and one easy way to understand it is using the Sperner's lemma [Park and Jeong, 2003]. Before stating the Sperner's lemma, we define the following terms. Interested readers may refer to [Rudin, 1953] for detailed explanations and proofs.

Definition 4. *An r-simplex is the convex hull formed by r+1 affinely-independent vertices. The convex combination of r+1 such vertices $x_0, x_1, ..., x_r$ forms an r-simplex Δ^r where,*

$$\Delta^r = \Big\{\sum_{i=0}^{r} \lambda_i x_i \Big| \sum_{i=0}^{r} \lambda_i = 1, \lambda_i \geq 0 \, for \; i = 0, 1, ..., r\Big\}$$

From Definition 4, we observe that a 0-simplex gives a point. A 1-simplex is a line segment formed by joining two affinely-independent points in one-dimensional space.

Similarly, a 2-simplex is a triangle in two-dimensional space and a 3-simplex is a tetrahedron in three-dimensional space. The convex hull of any non-empty subset of vertices of a simplex is called a face of the simplex. For example, figure 1 shows a Δ^3 simplex and its faces. The concept of simplices (or simplexes) serves as a basis for understanding different combinations of pure strategies (represented as vertices) and the existence of fixed points in a strategy space.

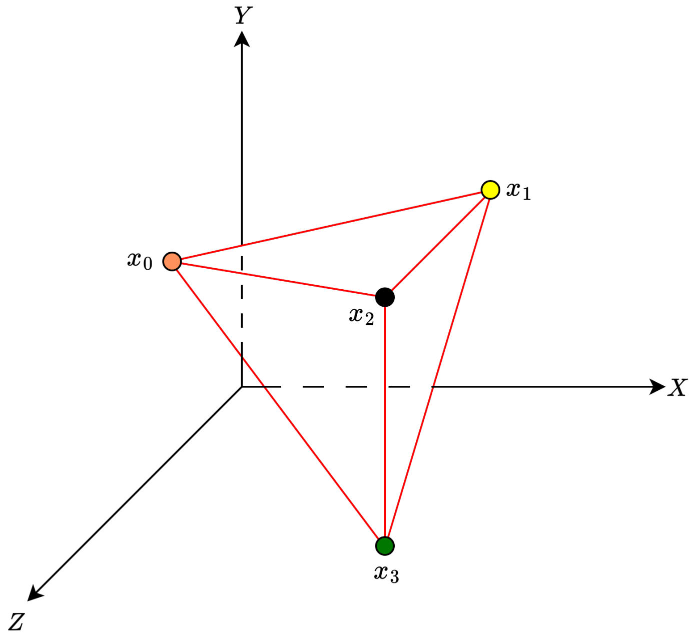


Figure 1: A Δ^3 *Simplex*

Definition 5. *A simplicial subdivision of a given simplex Δ^r is an operation that divides Δ^r into smaller sub-simplices such that*

(i) any two sub-simplices are disjoint and share a common face, and

(ii) the subdivision is homeomorphic to the original simplicial complex.

Definition 5 suggests that for a given simplex Δ^r, we can divide it into sub-simplices such that the properties of the original simplex are retained in these subdivisions. One example of this subdivision process is barycentric subdivision, where triangulation is used for such subdivision. Suppose we have k sub-simplices within the original simplex, then the intersection of any two sub-simplices share a common face. The union of all such k sub-simplices is the original simplex Δ^r. This process of defining simplices through a set of common faces is known as triangulation of the simplex.

Definition 6. *In an r-simplex Δ^r, a proper labeling (or proper coloring) of a simplicial subdivision involves assigning different colors to the $r+1$ vertices of Δ^r, and the points on each face of Δ^r only contains colors of the vertices defining that face.*

Lemma 1. *[Sperner, 1928] Every properly colored simplicial subdivision of r-simplex contains a closed cell (or part or subdivision) which is completely colored (that is, with all vertices having different colors).*

Geometric explanation of Sperner's lemma

First, we discuss a geometric explanation of Sperner's lemma and restrict our explanation to the two-dimensional case. The readers are encouraged to refer to the formal proof of Sperner's Lemma in the work by [McLennan and Tourky, 2008]. Let us consider a polygon whose vertices are labeled with different colors and properly triangulated. There exists a triangle inside the polygon that has all three different colors, and there is always an odd number of such triangles.

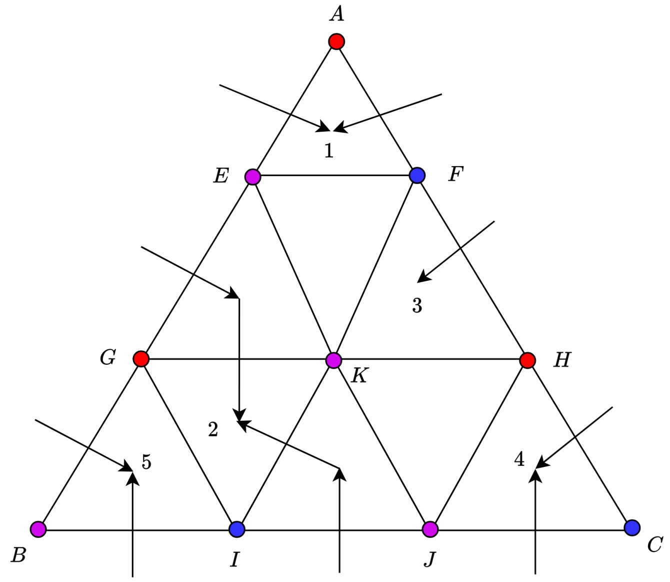


Figure 2: Sperners Lemma for a triangle.

For example, let us consider a triangle as shown in Figure 2 whose vertices A, B and C are colored red, pink, and blue, respectively. The triangle is further divided into sub-triangles, and each sub-triangle is color labeled. There is only one rule: the opposite side of any vertex does not contain that vertex's color. The inside coloring is done arbitrarily. Once the labeling is done, let us draw a path from the edge that contains two different colors. Once we reach inside the triangle, let us move further with the same rule. We stop moving once we do not find an edge with those two different colors. In the figure (Figure 2), let us move to the top left edge that contains red and pink vertices, there is no other red-pink edge once we reach triangle 1. Following the same procedure, we number all such triangles, namely, 2, 3, 4 and 5. We observe that all these triangles contain different colored vertices and these are an odd number of these triangles.

Example of proof of Brouwer's fixed point theorem using Sperner's lemma

In this subsection, we outline the proof of Brouwer's fixed point theorem using Sperner's lemma. We use a simple example of a 2-simplex $\Delta^2 = \{(x, y, z) \in \mathbb{R}_+ | x + y + z = 1\}$ as shown in Figure 3.

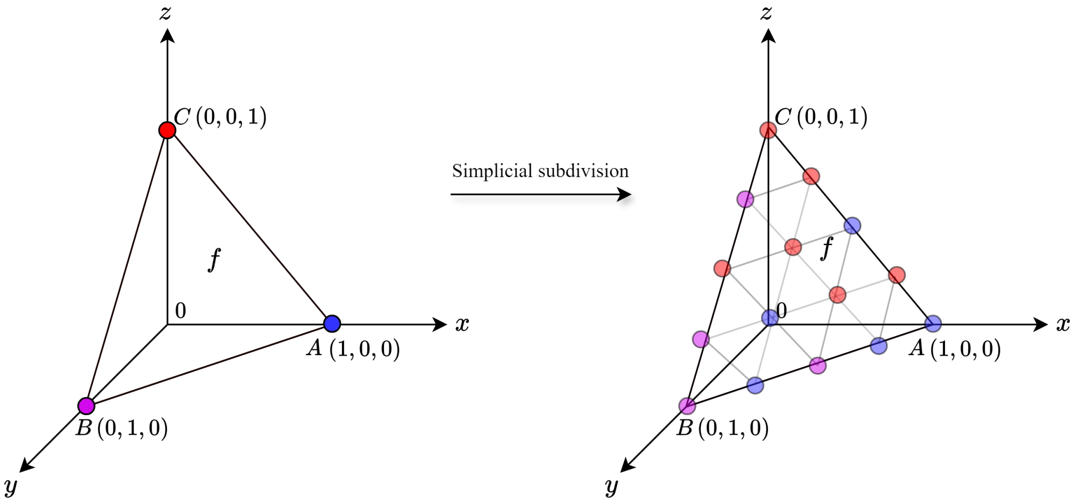


Figure 3: 2-simplex representation

The coordinates of the vertices of this simplex are $A = (1, 0, 0)$, $B = (0, 1, 0)$ and $C = (0, 0, 1)$. Let f be a continuous function from Δ^2 to itself which is decreasing in x, y and z. This means that, coordinate-wise, $A > f(A)$, $B > f(B)$ and $C > f(C)$. We assign colors blue, pink, and red to A, B, and C, and divide the simplex Δ^2 into l sub-simplicial parts (or triangles). As f decreases in x, y, and z, it cannot be increased in x moving towards edge BC; therefore, we cannot color it blue. Similarly, the function cannot be increased in y and z moving towards edges AC and AB, respectively. This gives us a properly labeled simplicial subdivision of Δ^2 and, hence, we can use Sperner's lemma (Lemma 1). From Sperner's lemma, there exists at least one completely colored sub-triangle in Δ^2. Repeating the above process, we may divide it into further simplicial subdivisions, inside which there exists another completely colored sub-traingle. Proceeding this way, when the simplicial sub-parts become infinitesimally small, we find that the function f does not decrease or increase in any direction, for points in the completely colored sub-simplex. In other words, in this recursive process, f approaches a fixed point in Δ^2, and eventually reaches that fixed point, thereby proving Brouwer's fixed point theorem.

2.2 A Proof of Existence of Nash Equilibrium for Matrix Games

As discussed earlier, [Nash, 1951] proved the existence of an equilibrium point for "n-player nonzero-sum games". In this subsection, we will first define bimatrix games (or "two-player nonzero-sum game"), matrix games (or "two-player zero-sum game"), and state von Neumann's minmax theorem for matrix games. We will, then, provide a simple proof of the existence of Nash equilibria in matrix games.

Definition 7. *A two-player non-cooperative game, G, in strategic (or normal) form is called a bimatrix game. As there are only two players, their payoff functions can be represented as payoff matrices, $A = (a_{ij})_{k\times m}$ and $B = (b_{ij})_{k\times m}$, A being Player 1's payoff matrix and B being Player 2's payoff matrix. Player 1 chooses a row (or a probability distribution over the rows), and Player 2 chooses a column (or a probability distribution over the columns).*

With reference to Definition 1, $N = \{1, 2\}$ is the set of players, $S_1 = \{1, 2, \ldots, k\}$ is the set of pure strategies of Player 1 or the set of rows, $S_2 = \{1, 2, \ldots, m\}$ is the set of pure strategies of Player 2 or the set of columns, $a_{ij} = p_1(i, j)$ and $b_{ij} = p_2(i, j)$ are the payoffs of Players 1 and 2 respectively, when Player 1 chooses $i \in S_1$ and Player 2 chooses $j \in S_2$. We call Player 1 the row player and Player 2 the column player.

The payoff matrices are

$$A = \begin{bmatrix} a_{11} & a_{12} & . & . & a_{1m} \\ a_{21} & a_{22} & . & . & a_{2m} \\ . & . & . & . & . \\ . & . & . & . & . \\ a_{k1} & a_{k2} & . & . & a_{km} \end{bmatrix}, \quad B = \begin{bmatrix} b_{11} & b_{12} & . & . & b_{1m} \\ b_{21} & b_{22} & . & . & b_{2m} \\ . & . & . & . & . \\ . & . & . & . & . \\ b_{k1} & b_{k2} & . & . & b_{km} \end{bmatrix}$$

and the bimatrix game is denoted as (A, B).

Proof. In this section, for simplicity and convenience, we use a slightly different notation from that used in Definition 1; x (instead of x_1) for Player 1's mixed strategy and y (instead of x_2) for Player 2's mixed strategy. The expected payoff of Player 1 and Player 2 are, hence, $x^T Ay$ and $x^T By$, respectively. Suppose both players want to maximize their respective expected payoffs. The strategy pair (x^*, y^*) is a Nash equilibrium of (A, B) if,

$$x^T Ay^* \leq x^{*T} Ay^* \;\; for\; all\; x \in X$$
$$x^{*T} By \leq x^{*T} By^* \;\; for\; all\; y \in Y$$

In other words, x^* and y^* are mutual best response strategies.

If the game is a zero-sum game, we have $A = -B$. It is enough to consider the payoff matrix, A, of Player 1 where Player 1 maximizes her (expected) payoff, and Player 2 minimizes his (expected) payoff. In zero-sum games, one player's gain is the other player's loss. When analyzing a zero-sum game to find a pair of mutual best response strategies, it may occur that it is not enough to consider only pure strategies. John Von Neumann proved that if we permit mixed strategies, then every zero-sum game always possess an optimal value (or optimal expected payoff to Player 1) and a pair of optimal strategies of the players. This led to the renowned solution concept for zero-sum games known as von Neumann's "minimax theorem" (or the "minmax theorem") [Von Neumann and Morgenstern, 1947], stated in Theorem 3.

Theorem 3. *[v. Neumann, 1928] Suppose $p : X \times Y \to \mathbb{R}$ is a continuous function, where $X \subseteq \mathbb{R}^k$ and $Y \subseteq \mathbb{R}^m$ are compact, convex sets. When $p(\cdot, y_i)$ is concave in x_i, for a given value of $y_i \in Y$, and $p(x_i, \cdot)$ is convex in y_i, for given value of $x_i \in X$, then*

$$\min_{y_i \in Y} \max_{x_i \in X} p(x_i, y_i) = \max_{x_i \in X} \min_{y_i \in Y} p(x_i, y_i).$$

Theorem 3 says that when one player maximizes the payoff, and the other player minimizes the same, then, an optimal solution always exists, whether in pure or mixed form. This solution is termed as the saddle point of a game. Another approach to demonstrate the existence of a solution in a zero-sum game is by formulating the given game as a linear program ([Dantzig, 1951]). Here, we provide a proof based on Brouwer's fixed-point theorem as done in Nash's work for nonzero-sum games [Nash, 1951]. We consider a simple example of a two-player zero-sum game for clarity and ease of understanding.

Let us consider a matrix game G in which the row player (Player 1) maximizes the payoff while the player, who plays columns (Player 2), minimizes the payoff. The row player's payoff matrix is $A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$ and the column player's payoff matrix is $-A$.

Let x, where $x^T = (x_1 \quad x_2)$, be a strategy vector of Player 1 and y, where $y^T = (y_1 \quad y_2)$, be a strategy vector of Player 2, where $x_i \geq 0$, $y_j \geq 0$, $\sum_i x_i = 1$ and $\sum_j y_j = 1$, for $i, j = \{1, 2\}$. The strategy vectors represent probability distributions over the players' respective sets of strategies. The expected payoff of Player 1 playing game G is $E(x, y) = x^T Ay$.

We define a function $f_i(x, y)$ for each pure strategy i of Player 1 as follows:

$$f_i(x, y) = max\{_i(Ay) - x^T Ay, 0\} \tag{1}$$

where $_i(Ay)$ is the i^{th} entry of (Ay).

We use this function to improve the probability vector x, $x^T = (x_1 \quad x_2)$, as

$$\bar{x}^T = \begin{pmatrix} \dfrac{x_1 + f_1(x, y)}{x_1 + x_2 + f_1(x, y) + f_2(x, y)} & \dfrac{x_2 + f_2(x, y)}{x_1 + x_2 + f_1(x, y) + f_2(x, y)} \end{pmatrix}.$$

As $\sum\limits_i x_i = 1$, we have $\bar{x}^T = \begin{pmatrix} \dfrac{x_1 + f_1(x, y)}{1 + f_1(x, y) + f_2(x, y)} & \dfrac{x_2 + f_2(x, y)}{1 + f_1(x, y) + f_2(x, y)} \end{pmatrix}$.

Similarly, we define a function $g_j(x, y)$ for Player 2 as

$$g_j(x, y) = max\{_j(x^T(-A)) - x^T(-A)y, 0\} \tag{2}$$

The revised probability vector, y, for Player 2 is

$$\bar{y}^T = \begin{pmatrix} \dfrac{y_1 + g_1(x, y)}{1 + g_1(x, y) + g_2(x, y)} & \dfrac{y_2 + g_2(x, y)}{1 + g_1(x, y) + g_2(x, y)} \end{pmatrix}.$$

We define a function, ϕ, which updates the pair of probability distributions (x, y) to $(\bar{x}^T, \bar{y}^T)$. That is,

$$\phi(x, y) = (\bar{x}^T, \bar{y}^T) \tag{3}$$

where the domain of ϕ is a compact, convex set (as it is a pair of mixed strategy sets, each mixed strategy being a convex combination of pure strategies) and $\phi(x, y)$ is a continuous function. Using Brouwer's fixed point theorem (Theorem 2), it can be shown that there exists a point (x^*, y^*) such that $\phi(x^*, y^*) = (x^*, y^*)$. We now show that this fixed point is a Nash equilibrium of the game G.

If possible, let $f_1(x^*, y^*) + f_2(x^*, y^*) > 0$. Then, $_i(Ay^*) > x^{*T} Ay^*$ for all i.

However, $x^{*T} Ay^* = \sum\limits_i x_i^* (_i(Ay^*))$, which contradicts the above inequality.

Therefore, $f_1(x^*, y^*) + f_2(x^*, y^*) = 0$.

Similarly, $g_1(x^*, y^*) + g_2(x^*, y^*) = 0$.

The probability vectors $\bar{x}^*$ and $\bar{y}^*$ that update x^* and y^*, respectively, are

$$\bar{x}^{*T} = \begin{pmatrix} \dfrac{x_1^* + f_1(x^*, y^*)}{1 + f_1(x^*, y^*) + f_2(x^*, y^*)} & \dfrac{x_2^* + f_2(x^*, y^*)}{1 + f_1(x^*, y^*) + f_2(x^*, y^*)} \end{pmatrix},$$

$$\bar{y}^{*T} = \begin{pmatrix} \dfrac{y_1^* + g_1(x^*, y^*)}{1 + g_1(x^*, y^*) + g_2(x^*, y^*)} & \dfrac{y_2^* + g_2(x^*, y^*)}{1 + g_1(x^*, y^*) + g_2(x^*, y^*)} \end{pmatrix}.$$

As $f_1(x^*, y^*) + f_2(x^*, y^*) = 0$ and $g_1(x^*, y^*) + g_2(x^*, y^*) = 0$, it follows that

$$\begin{aligned} {}_i(Ay^*) &\leq x^{*T}Ay^* \text{ for all } i, \text{ and} \\ {}_j(x^{*T}(-A)) &\leq x^{*T}(-A)y^* \text{ for all } j, \end{aligned}$$

which is, in fact, the definition of Nash equilibrium (Definition 7). □

In the above proof, our goal is to find a function ϕ that has fixed points, and these fixed points are Nash equilibrium points. We also need to find a strategy that gives maximum payoff to Player i. The function $f_i(x, y)$ represents the possible improvement in expected payoff by switching to pure strategy i. For example, if Player 1 plays $x_1 = 1$ and $x_2 = 0$ pure strategy then expected payoff is $a_{11}y_1 + a_{12}y_2$. Similarly, playing pure strategy $x_2 = 1$ and $x_1 = 0$ results expected payoff of $a_{21}y_1 + a_{22}y_2$. The function $f_i(x, y)$ gives the value that much can be improved by changing the strategies. The vectors $\bar{x}^T$ and $\bar{y}^T$ are normalized by dividing $\bar{x}^T$ with $x_1 + x_2 + f_1(x, y) + f_2(x, y)$ and $\bar{y}^T$ with $y_1 + y_2 + g_1(x, y) + g_2(x, y)$.

The above approach can be extended for equilibrium analysis for n-player nonzero-sum games too as shown by [Nash, 1951].

2.3 Limitations of Nash Equilibrium

Despite its strong mathematical foundation and popularity, Nash equilibrium has limitations. One of the primary issues is that Nash equilibrium does not always lead to a Pareto-efficient outcome.[2] This inefficiency arises due to the absence of cooperation among the players, as Nash equilibrium focuses on individual optimization rather than collective welfare. Secondly, multiple Nash equilibria can exist in many strategic interactions in the same game. The presence of multiple equilibria poses challenges in deducing or predicting which equilibrium will be selected and leads to questions about the stability of the outcomes. Thirdly, Nash equilibrium assumes a lack of cooperation among the players, and guarantees an equilibrium point only against unilateral deviation and not when players may collectively deviate. Lastly, Nash equilibrium assumes that each player solely focuses on maximizing her own utility or payoff. This assumption may not always reflect the real-world dynamics of various strategic interactions, where cooperation and communication can significantly impact the outcomes.

[2] A Pareto-efficient outcome is one where no player can benefit without another player being worse off.

3 Berge Equilibrium

Berge equilibrium deals with the moral golden rule “Do good for others as you would want them to do to you”, contrary to Nash’s theorem where players are self-optimizers. Reciprocity is an inherent part of many applications, where players are not self-optimizers but help others and get help in return. For example, the government plans to design subsidies to improve the social conditions of poor people, a family business with multiple units, the supply chain of food banks and humanitarian operations, hospital emergency operations, climate change, and many more. Therefore, Nash’s assumption that the sole objective of every player is to optimize their respective payoffs is not applicable in many social problems. In such social applications, the Berge equilibrium helps to find acceptable solutions. The research by V.I. Zhukovskiy ([Zhukovskiy, 1985]) accelerated the use of this concept in social science and economics. For quite some time, most of the research publications on Berge equilibrium were in Russian, and the concept was not well-known among a wider audience. However, we are seeing increasing work on Berge equilibrium, including [Larbani and Zhukovskii, 2017], [Colman et al., 2011], [Larbani and Nessah, 2008], [Nessah et al., 2007], and [Abalo and Kostreva, 2005]. The main challenge faced by researchers is “individual rationality”, which was directly satisfied in the Nash equilibrium and was hard to establish in Berge’s theory. We explain this challenge in the subsequent subsection (Section 3.1).

Definition 8. *[Zhukovskiy, 1985] Suppose N is the set of players of a non-cooperative, complete information game, where all players are maximizers. A strategy profile $x^B = (x_1^B, x_2^B, ..., x_n^B)$ is said to be a Berge equilibrium if*

$$p_i(x_i^B, x_{N\setminus i}) \leq p_i(x^B)$$

where $i \in N$, $p_i(\cdot)$ is the payoff function of Player i, and x_i^B is Player i’s Berge strategy.

Recall Nash’s theorem where $(x_i^*, x_{N\setminus i}^*)$ is an equilibrium strategy tuple if $p_i(x_i, x_{N\setminus i}^*) \leq p_i(x_i^*, x_{N\setminus i}^*), for\ all\ x_i \in X_i$. The distinction between the definitions of Nash and Berge equilibria lies in the selection of strategy. In Nash equilibrium, players choose strategies that maximize their own respective payoffs, whereas in Berge equilibrium, players select strategies that maximize the payoffs of all other players’ (excluding their own). We can find Nash and Berge equilibria as follows:

$$\left.\begin{aligned} x_i^* &= \arg\max_{x_i \in X_i} p_i(x_i, x_{N\setminus i}^*) \\ x_i^B &= \arg\max_{x_{N\setminus i} \in X_{N\setminus i}} p_i(x_i^B, x_{N\setminus i}) \end{aligned}\right\} \text{for all} i \in N.$$

Definition 8 says that Player i’s payoff on playing strategy x_i^B is higher when other players play $x_{N\setminus i}^B$, than when other players play $x_{N\setminus i}$. Alternatively, Player i’s payoff function is maximized when other players choose their Berge strategies, and in return, Player i plays a Berge strategy that maximizes others' payoffs. [Larbani and Nessah, 2008] showed that the convexity and compactness of strategy sets, and continuity and concavity of payoff functions are not sufficient for the existence of Berge equilibrium. [Nessah et al., 2007] provided necessary and sufficient conditions for the existence of Berge equilibrium.

3.1 Limitations of Berge Equilibrium

In this subsection, we state the limitations of Berge equilibrium and illustrate the same using two-player games. If Player 1's strategy set is X_1 and Player 2's strategy set is X_2, then we say individual rationality is achieved when the following inequalities are satisfied, given that the players play strategies $x^* = (x_1^*, x_2^*)$ [Von Neumann and Morgenstern, 1947],

$$p_1(x^*) \geq \max_{x_1 \in X_1} \min_{x_2 \in X_2} p_1(x_1, x_2)$$

$$p_2(x^*) \geq \max_{x_2 \in X_2} \min_{x_1 \in X_1} p_2(x_1, x_2)$$

[Salukvadze et al., 2020] provided an example of a Berge equilibrium that fails to meet the individual rationality condition. Expanding on this, we present another simple illustrative example and identify additional conditions under which individual rationality is compromised.

Example 1. *Consider the following game with two players. Suppose the strategy set for Player 1 is $X_1 = [0, 1]$ and strategy set for Player 2 is $X_2 = (-\infty, \infty)$. Suppose the payoff function for Player 1 is $p_1(x_1, x_2) = x_2^2 - x_1 x_2$ and payoff function for Player 2 is $p_2(x_1, x_2) = x_1^2 - x_1$ where $x_1 \in X_1$ and $x_2 \in X_2$. The strategy pair (x_1^B, x_2^B) is a Berge equilibrium if*

$$\max_{x_2 \in X_2} p_1(x_1^B, x_2) = p_1(x_1^B, x_2^B)$$

$$\max_{x_1 \in X_1} p_2(x_1, x_2^B) = p_2(x_1^B, x_2^B)$$

Using the first order conditions $\frac{\partial p_1}{\partial x_2} = 0$ and $\frac{\partial p_2}{\partial x_1} = 0$, we get $x_1^B = \frac{1}{2}$ and $x_2^B = \frac{1}{4}$ which gives us payoff $p_1(x_1^B, x_2^B) = \frac{-1}{16}$ and $p_2(x_1^B, x_2^B) = \frac{-1}{4}$. Using von-Neumann's maximin theorem, $p_1^* = \max\limits_{x_1 \in X_1} \min\limits_{x_2 \in X_2} p_1(x_1, x_2)$, we find $x_1^* = 0$ and $x_2^* = 0$ that gives payoff $p_1^*(x_1^*, x_2^*) = 0$.

Here, we note that $p_2^*(x_1^*, x_2^*) > p_2(x_1^B, x_2^B)$, which shows that Berge equilibrium fails to satisfy the individual rationality condition. To alleviate this difficulty, K.S. Vaisman proposed a solution which is called the Berge-Vaisman equilibrium [Vaisman, 1995].

Remark: It is interesting to note that the optimal solution of either inner minimization problem $\min_{x_2 \in X_2} p_1(x_1, x_2)$ of minimax theorem *or* maximization of a function $p_1(x_1, x_2)$ as $\max_{x_2 \in X_2} p_1(x_1^B, x_2)$ in Berge equilibrium, gives a boundary point based on the function characteristics. Refer figure 4.

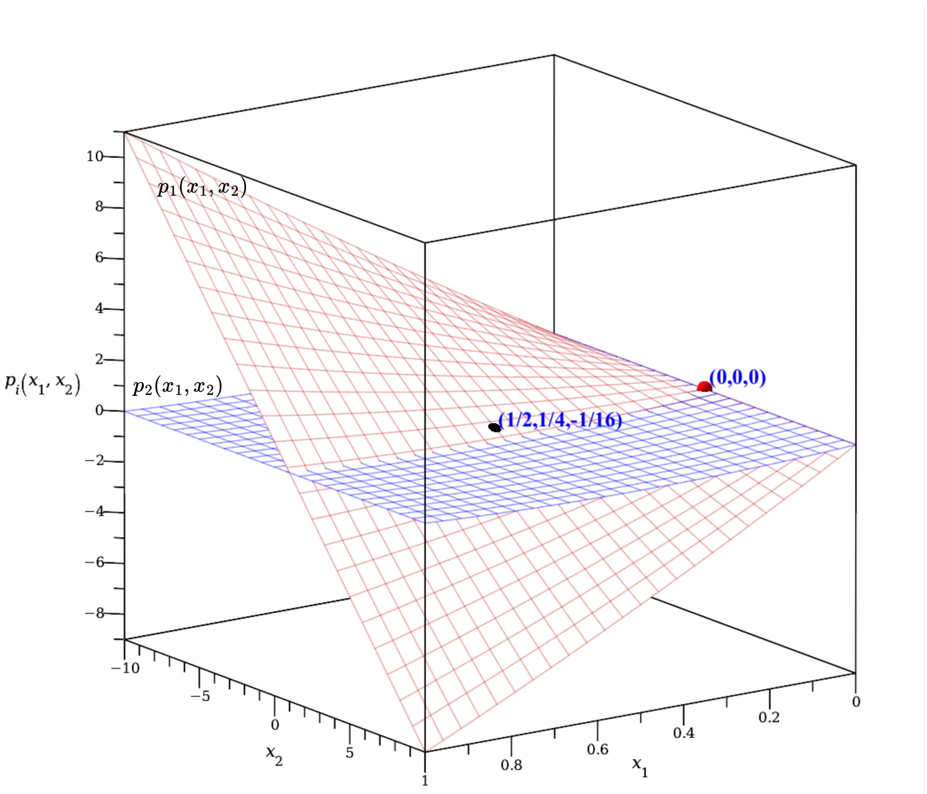


Figure 4: B.E. example

Definition 9. *[Zhukovskii and Chikrii, 1994] The strategy profile* $x^B = (x_1^B, x_2^B, ..., x_n^B)$ *is a Berge-Vaisman equilibrium if*

$$p_i(x^B) \geq p_i(x_i^B, x_{N\backslash i})$$
$$p_i(x^B) \geq \alpha_i$$

where $\alpha_i = \max_{x_i \in X_i} \min_{x_{N\backslash i} \in X_{N\backslash i}} p_i(x_i, x_{N\backslash i})$ *is the minimum threshold set by Player* i *(assuming all players are maximizers).*

Definition 9 is a revision of Definition 8, incorporating the concept of individual rationality. That is, the Berge-Vaisman equilibrium is the Berge equilibrium under conditions of a minimum guaranteed payoff to Player i, even if the remaining players, $N\backslash i$, minimize

Player i's payoff. As demonstrated in Example 1, achieving individual rationality may not always be possible when finding a Berge equilibrium. Therefore, the Berge-Vaisman equilibrium offers a more appropriate solution approach.

4 Comparison of Nash and Berge Equilibria

The main difference between Nash equilibrium (or N.E.; Definition 2) and Berge equilibrium (or B.E.; Definition 8) is related to the change in payoffs because of deviation from equilibrium strategies. In a Nash equilibrium, Player i's payoff decreases when she deviates from her best response strategy while, in a Berge equilibrium, Player i's payoff decreases when other players deviate from their best response strategies. Alternatively, we can say that Player i gains by focusing on her own strategies if she decides to play according to the Nash equilibrium, and she gains by maximizing other players' payoffs if all players adopt their respective Berge strategies. [Colman et al., 2011] offered examples illustrating two-player pure strategy Nash and Berge equilibria. In this context, we present additional instances that allow for a comparative analysis of both equilibrium solutions.

4.1 Two-player Bimatrix Game Examples

In this section, we provide a few examples of bimatrix games (two-player nonzero-sum games) in which both players are maximizers and each player has two strategies. We find out pure strategy Nash and Berge equilibria of these games, if they exist, and compare the results. Following the notation from Definition 7, we can represent a bimatrix game as a pair of payoff matrices for two players, denoted as (A, B). Each row and column entry in this pair is represented by a pair of payoff values (a_{ij}, b_{ij}).

Example 2. *Non-existence of pure strategy Nash and pure strategy Berge equilibria*

Consider the following two-player nonzero-sum game.
The payoff matrix of a game is given as $(A, B) = \begin{bmatrix} (6,10) & (10,3) \\ (8,6) & (3,8) \end{bmatrix}$ where A is the Player 1's (row player) payoff matrix and B is the Player 2's (column player) payoff matrix.

Pure strategy N.E.: In this game, if the column player selects the first column to maximize his payoff, the row player's best response strategy is to choose the second row to maximize her own payoff. Conversely, if the row player opts for the second row, the

column player's best response strategy is to choose the second column. However, the row player does not play the second row if the column player plays the second column. This analysis confirms that the game does not have a pure strategy Nash equilibrium.
Pure strategy B.E.: If the players maximize each other's payoffs (instead of their own), the column player's best response in the first row is the second column, and the row player's best response in the second column is the second row. In the second row, the column player's best response is the first column, and in the first column, the row player's best response is the first row. Therefore, this game also does not have a pure strategy Berge equilibrium.

Example 3. *Existence of pure strategy Nash equilibrium and non-existence of pure strategy Berge equilibrium*

Consider the following two-player non-zero sum game.

$(A, B) = \begin{bmatrix} (1,5) & (10,0) \\ (0.5,0) & (0,5) \end{bmatrix}$ where A and B are Player 1's and Player 2's payoff matrices, respectively.

Pure strategy N.E.: In this scenario, if the row player selects the first row, the column player's best response strategy is to choose the first column. If the column player opts for the first column, the row player's best response is to select the first row. This indicates a pure strategy Nash equilibrium with payoff pair (1,5).
Pure strategy B.E.: If the row player plays the first row, the column player will play the second column to maximize the row player's payoff. However, to improve the column player's payoff, the row player chooses the second row if the column player plays the second column. If the row player plays the second row, then the column player chooses the first column where the row player's payoff is higher than the second column. Therefore, this game does not possess pure strategy B.E.

Example 4. *Non-existence of pure strategy Nash equilibrium and existence of pure strategy Berge equilibrium*

Consider the following bimatrix game with payoff matrix pair $(A, B) = \begin{bmatrix} (5,5) & (7,4) \\ (6,3) & (6,4) \end{bmatrix}$.

Pure strategy N.E.: If the row player plays the first row, then the column player's best interest is to play the first column. Given that the column player plays the first column, the row player's best response strategy is to choose the second row. Now, if the row

player plays the second row, then the column player selects the second column, which indicates that this game has no pure strategy N.E.
Pure strategy B.E.: If the row player plays the first row, then the column player chooses the second column. Given that the column player selects the second column, the row player is indifferent between playing the first or second row. If the row player plays the second row, then the column player is indifferent between choosing the first or second column. Therefore, there are two pure strategy B.E. with pairs of payoffs (7,4) and (6,4). Note that the Berge equilibrium (7,4) offers a better payoff to Player 1 compared to the Berge equilibrium (6,4).

Example 5. *Existence of pure strategy Nash and pure strategy Berge equilibrium*

Consider the following bimatrix game with payoff matrix pair $(A, B) = \begin{bmatrix} (10,1) & (5,10) \\ (11,8) & (6,12) \end{bmatrix}$.

Pure strategy N.E.: In this game, if the row player plays the first row, then the column player's best response strategy is to choose the second column. If the column player plays the second column, then the row player chooses the second row. Given that the row player plays the second column, the column player selects the second column which gives the pure strategy N.E. with payoff pair (6,12).
Pure strategy B.E.: If the column player plays the first column then, the row player chooses the second row. Given that the row player chooses the second row, to maximize the the row player's payoff, the column player chooses the first column, and we get the pure strategy B.E. with payoff pair (11,8). The N.E. gives better payoff to the column player compared to the B.E. while the B.E. returns higher payoff to the row player compared to the N.E.

4.2 Application of Berge Equilibrium in Social Environment

The Berge equilibrium has many social applications, and we give a few examples here in which we find a pure strategy Berge equilibrium. We, first, look at a well-known example, the prisoners' dilemma.

Prisoner's dilemma
In this game, each player has to decide whether to confess or lie. If only one of the players confesses then he gets free and the other player gets five years of jail, and vice versa. If both players confess, then both get three years of jail; if both lie, then both get one year of jail. The bimatrix representation of this game is $(A, B) = \begin{bmatrix} (-3,-3) & (0,-5) \\ (-5,0) & (-1,-1) \end{bmatrix}$, where both players want to minimize the number of years in jail (or, equivalently, maximize

their respective payoff entries in (A, B). In this bimatrix, the first row and first column represent the respective confessing strategies, and the second row and second column show the respective lying strategies. We find that (confess, confess) is a pure strategy N.E. with corresponding pair of payoffs (-3,-3). This is clearly not Pareto efficient. According to Berge's strategy, if the prisoners are altruistic then the column player chooses to lie against the row player's lie strategy and vice versa, which gives us the pure strategy B.E. with payoffs (-1,-1). This is better than the N.E. for both the players and is Pareto efficient. Therefore, both players get the benefits of choosing the strategy which makes the other player's payoff better.

[Axelrod and Hamilton, 1981] empirically showed that when the prisoner's dilemma game is played repetitively for a finite number of rounds assuming both players are unaware of the total number of rounds, the "Tit for Tat" strategy outperforms all other strategies. Therefore, we can say that if both players only choose the lie strategy in repetitive prisoner's dilemma game, then the result of this game is similar to the result of a single round B.E. game. However, this does not hold true if one of the players deviates from the lie strategy and chooses to confess.

Trust game

Let us consider a second example called the "trust game" where there are two firms in the market run by CEOs sharing a common family. We assume that both firms produce a homogeneous product. Now, during the selling period, each firm shares information with each other. If both firms trust each other, then both will gain 100 units of payoff. However, if one of the firms chooses not to trust, that firm gains 150 units of payoff whereas the other firm gains only 50. If both firms decide not to trust each other, then both will achieve 80 units. We represent this game in bimatrix form as $(A, B) = \begin{bmatrix} (100, 100) & (50, 150) \\ (150, 50) & (80, 80) \end{bmatrix}$ where the first row and the first column represent the "trust" strategies of Players 1 and 2 respectively. The second row and second column represent the "do not trust" strategies. If the players choose their Nash strategies, then both get 80 units. (The pair of N.E. payoffs is (80,80)). Now, as both CEOs share a common family, if both choose an altruistic strategy for each other, they can achieve 100 units each by playing their respective Berge strategies (The pair of B.E. payoffs is (100,100)).

Hawk and Dove game

Now, let us consider a game with a total value of $M > 0$ available, and two players need to divide this value among themselves. We call a strategy "Hawk" when players become aggressive and call a strategy "Dove" when players choose peace. If both players are aggressive, then they both get $\frac{M}{2} - C$ value of payoff where $C > 0$ is the cost of being aggressive. If both choose peace then they divide the value equally $\frac{M}{2}$. Generally, the Hawk and Dove game (also known as the "chicken game") represents a situation where one player must "swerve" to avoid a collision, or both players are penalized for being "aggressive". However, for the purpose of understanding, we consider the altruistic behaviour of the players and find B.E. to observe the difference in payoff compared to

N.E.

We represent this game in bimatrix form as $(A, B) = \begin{bmatrix} (\frac{M}{2} - C, \frac{M}{2} - C) & (M, 0) \\ (0, M) & (\frac{M}{2}, \frac{M}{2}) \end{bmatrix}$ where the first row-first column represents the "Hawk-Hawk" strategy pair and the second row-second column represents the "Dove-Dove" strategy pair. The value of N.E. and B.E. depends on the value of $\frac{M}{2} - C$.

Case 1: $\frac{M}{2} - C > 0$
In this case, if Player 2 chooses column 1 then Player 1's best response is to choose row 1 and that gives us the N.E. payoff pair $(\frac{M}{2} - C, \frac{M}{2} - C)$. However, if we compute the B.E. of this game then the pair of payoffs is $(\frac{M}{2}, \frac{M}{2})$. It is evident that achieving peace is the better option for both players in this scenario.

Case 2: $\frac{M}{2} - C < 0$
In this case, there are two pure Nash equilibria, $(Hawk, Dove)$ and $(Dove, Hawk)$, with corresponding payoff pairs $(M, 0)$ and $(0, M)$. This suggests that players should choose Hawk when the other player chooses Dove, and vice versa. Additionally, there exists one mixed strategy Nash equilibrium. Now, if we consider Berge equilibrium, then playing $(Dove, Dove)$ remains an equilibrium strategy and is better for both the players. Therefore, when the cost of being aggressive is higher, choosing peace is the better option. This rationale suggests that irrespective of payoff gain or loss, choosing peace and dividing the value equally is better for both the players.

5 Concluding Remarks and Future Research Directions

In this article, we initially outline Nash's seminal contributions to game theory. Nash showed that every game with a finite number of players, where each player has a finite set of strategies has at least one equilibrium point. We outline the proof of Nash's theorem using Brouwer's fixed point theorem and sperner's lemma. We, then, present a simple proof of Nash's theorem for "two-player zero-sum games". According to Nash's theory, each player is primarily motivated by self-optimization and does not take into account other players' benefits while making strategic decisions. In contrast to Nash's individualistic approach, we delve into the concept of Berge equilibrium, which is grounded in the principle of "reciprocal altruism" or the "moral golden rule". In Berge's theory, each player makes decisions that benefits others (and themselves due to reciprocity), resulting in mutual gains. However, due to its limitations in terms of individual rationality and its relatively uncommon theoretical foundation, the Berge equilibrium has not gained widespread adoption in practical applications. Nevertheless, we make an effort to explore the potential of Berge equilibrium in social contexts.

In numerous real-world scenarios, not all players base their decisions solely on personal gain. For instance, examining Berge's equilibrium is of interest in the design of government subsidy programs, where the government aims to maximize social welfare rather than pursuing self profit maximization or self cost minimization. Another significant application arises in firms' sustainable and climate change policy decisions. Firms can collaborate to establish policies that contribute to goal attainment rather than competing on carbon footprint minimization. While some literature has explored the application of Berge equilibrium in Cournot and Bertrand competition models, there is a notable gap in its study across various market models in economics. Healthcare represents another important arena for the application of Berge equilibrium. In the context of organ transplant operations conducted by multiple hospitals, Nash's non-cooperative model may not be suitable for guiding hospitals in their decision-making processes. It is worth noting that Nash's theory has been extended to a wide array of applications, encompassing different game types while in contrast, Berge's theory remains largely unexplored in these variations. Hence, researchers interested in the Berge equilibrium can undertake further investigations into its applicability in diverse scenarios that can be modelled as a game.

Bibliography

[Abalo and Kostreva, 2005] Abalo, K. and Kostreva, M. (2005). Berge equilibrium: some recent results from fixed-point theorems. *Applied Mathematics and Computation*, 169(1):624–638.

[Axelrod and Hamilton, 1981] Axelrod, R. and Hamilton, W. D. (1981). The evolution of cooperation. *Science*, 211(4489):1390–1396.

[Berge, 1957] Berge, C. (1957). *Théorie générale des jeux à n personnes*, volume 138. Gauthier-Villars Paris.

[Border, 1985] Border, K. C. (1985). *Fixed point theorems with applications to economics and game theory*. Cambridge University Press.

[Brouwer, 1911] Brouwer, L. E. J. (1911). Über abbildung von mannigfaltigkeiten. *Mathematische Annalen*, 71(1):97–115.

[Colman et al., 2011] Colman, A. M., Körner, T. W., Musy, O., and Tazdaït, T. (2011). Mutual support in games: Some properties of Berge equilibria. *Journal of Mathematical Psychology*, 55(2):166–175.

[Dantzig, 1951] Dantzig, G. B. (1951). A proof of the equivalence of the programming problem and the game problem. *Activity Analysis of Production and Allocation*, 13.

[Harsanyi, 1967] Harsanyi, J. C. (1967). Games with incomplete information played by "bayesian" players, i–iii, Part i. The basic model. *Management Science*, 14(3):159–182.

[Kakutani, 1941] Kakutani, S. (1941). A generalization of Brouwer's fixed point theorem.

[Katz, 2017] Katz, J. (2017). Brouwer's fixed point theorem. *The University of Chicago*. Available at: https://web.stanford.edu/~saberi/lecture1.pdf.

[Larbani and Nessah, 2008] Larbani, M. and Nessah, R. (2008). A note on the existence of Berge and Berge–Nash equilibria. *Mathematical Social Sciences*, 55(2):258–271.

[Larbani and Zhukovskii, 2017] Larbani, M. and Zhukovskii, V. (2017). Berge equilibrium in normal form static games: a literature review. *Izv. IMI UdGU*, pages 80–110.

[McLennan and Tourky, 2008] McLennan, A. and Tourky, R. (2008). Using volume to prove Sperner's lemma. *Economic Theory*, 35(3):593–597.

[Nash, 1951] Nash, J. (1951). Non-cooperative games. *Annals of Mathematics*, pages 286–295.

[Nessah et al., 2007] Nessah, R., Larbani, M., and Tazdait, T. (2007). A note on Berge equilibrium. *Applied Mathematics Letters*, 20(8):926–932.

[Park and Jeong, 2003] Park, S. and Jeong, K. S. (2003). A proof of the Sperner lemma from the Brouwer fixed point theorem. In *Nonlinear Analysis Forum*, volume 8, pages 65–68.

[Rudin, 1953] Rudin, W. (1953). *Principles of Mathematical Analysis*.

[Salukvadze et al., 2020] Salukvadze, M. E., Zhukovskiy, V. I., et al. (2020). *The Berge Equilibrium: A Game-Theoretic Framework for the Golden Rule of Ethics*. Springer.

[Shapley et al., 1953] Shapley, L. S. et al. (1953). A value for n-person games.

[Sperner, 1928] Sperner, E. (1928). Ein satz über untermengen einer endlichen menge. *Mathematische Zeitschrift*, 27(1):544–548.

[v. Neumann, 1928] v. Neumann, J. (1928). Zur theorie der gesellschaftsspiele. *Mathematische Annalen*, 100(1):295–320.

[Vaisman, 1995] Vaisman, K. S. (1995). The Berge equilibrium. *Cand. Sci. (Phys.-Math.) Dissertation, St. Petersburg, 110 pp. (In Russian)*.

[Von Neumann and Morgenstern, 1947] Von Neumann, J. and Morgenstern, O. (1947). *Theory of games and economic behavior, 2nd rev.* Princeton University Press. (Newer (2007) edition: 60th Anniversary Commemorative Edition).

[Wade-Benzoni and Tost, 2009] Wade-Benzoni, K. A. and Tost, L. P. (2009). The egoism and altruism of intergenerational behavior. *Personality and Social Psychology Review*, 13(3):165–193.

[Zhukovskii and Chikrii, 1994] Zhukovskii, V. I. and Chikrii, A. A. (1994). Linear quadratic differential games. *Naoukova Doumka, Kiev.*

[Zhukovskiy, 1985] Zhukovskiy, V. I. (1985). Some problems of non-antagonistic differential games. Mathematical methods in operations research. *Bulgarian Academy of Science Sofia (In Russian).*

3 A Brief Encounter with Integer Polynomials

Prithwijit De

Homi Bhabha Centre for Science Education,
TIFR, Mumbai.
Email: prithwijit@hbcse.tifr.res.in

A polynomial with integer coefficients is called an *integer polynomial* and the general form of such a polynomial $P(x)$ of degree $n(\geq 1)$ is

$$P(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$$

where $a_0, a_1, \ldots, a_n$ are integers and $a_n \neq 0$. If $P(x)$ is an integer polynomial then for any two distinct integers u and v, $P(u) - P(v)$ is divisible by $u - v$. If $P(x) = ax + b$ then

$$P(u) - P(v) = a(u - v)$$

and the result is obvious. If degree of $P(x)$ is n then

$$P(u) - P(v) = \sum_{k=1}^{n} a_k (u^k - v^k) \tag{1}$$

and by virtue of the identity

$$u^k - v^k = (u - v)(u^{k-1} + u^{k-2} v + \cdots + u v^{k-2} + v^{k-1})$$

it follows that every term in the sum on the right hand side of (1) is divisible by $u - v$, and hence the same holds for $P(u) - P(v)$. This result has some nice applications as we shall now see.

Example 1

Let a, b, c be three distinct integers. Does there exist an integer polynomial $P(x)$ with $P(a) = b$, $P(b) = c$ and $P(c) = a$?

Suppose there exists such a polynomial $P(x)$. Then $a-b$ divides $P(a)-P(b)=b-c$, $b-c$ divides $P(b)-P(c)=c-a$, and $c-a$ divides $P(c)-P(a)=a-b$. Thus

$$|a-b| \leq |b-c| \leq |c-a| \leq |a-b|$$

which leads to

$$|a-b| = |b-c| = |c-a|$$

and this does not hold unless $a=b=c$.

Ponder This:

Let $a_1, a_2, \ldots, a_n$ be $n(>3)$ distinct integers. Does there exist an integer polynomial $P(x)$ with $P(a_k)=a_{k+1}$ for $k=1,2,\ldots,n-1$ and $P(a_n)=a_1$?

Example 2

Let f be a monic polynomial with integer coefficients and suppose that there are four distinct integers a, b, c, d for which

$$f(a)=f(b)=f(c)=f(d)=12.$$

Show that there is no integer k for which $f(k)=25$.

Let $g(x)=f(x)-12$. Then $g(x)$ is a monic polynomial and $g(a)=g(b)=g(c)=g(d)=0$. Therefore

$$g(x)=(x-a)(x-b)(x-c)(x-d)h(x) \tag{2}$$

where $h(x)$ is a monic integer polynomial. If there exists an integer k such that $f(k)=25$ then $g(k)=13$. That is

$$(k-a)(k-b)(k-c)(k-d)h(k)=13. \tag{3}$$

Since a, b, c, d are distinct, we may assume that $a<b<c<d$. Thus $k-a>k-b>k-c>k-d$. As 13 is a prime, it cannot be written as a product of at least four distinct integers and we arrive at a contradiction. Therefore such an integer k does not exist.

Ponder This:

In the proof above we claimed that $h(x)$ is a monic integer polynomial. Is the claim correct?

Example 3

Let $P(x) = x^n + a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \cdots + a_1x + a_0$ be a polynomial with integer coefficients, such that, $P(0)$ and $P(1)$ are odd integers. Does $P(x)$ have an integer root?

Suppose u is an integer root of $P(x)$. Observe that since $P(0)$ is odd, $P(0) \neq 0$. Therefore $u \neq 0$. Now, u divides $P(u)-P(0) = -P(0)$ and $u-1$ divides $P(u)-P(1) = -P(1)$. Since both $P(0)$ and $P(1)$ are odd, both u and $u-1$ must be odd, which is absurd.

Ponder This:

If $P(m)$ and $P(m+1)$ are odd for some integer m, does $P(x)$ have an integer root?

Example 4

Does there exist an integer polynomial with $P(x)$ such that $|P(n)|$ is a prime number for every positive integer n?

Suppose there exists such a polynomial. Let $P(1) = p$, p a prime number. Then,

$$P(1+tp) - P(1)$$

is divisible by p for $t = 1, 2, \ldots$. As $P(1)$ is divisible by p we have $P(1+tp)$ divisible by p for each t. Since $|P(1+tp)|$ approaches infinity as t approaches infinity, $|P(1+tp)| > p$ for infinitely many values of t. This shows that there are infinitely many positive integers n such that $|P(n)|$ is composite. Therefore such a polynomial does not exist.

There are many more examples based on the "$u - v$ divides $P(u) - P(v)$" theme. We leave one such example for the reader to enjoy before moving on to another type of problems involving integer polynomials.

For the reader:

At some integer points an integer polynomial $P(x)$ assumes the values 1, 2, and 3. Prove that there exists no more than one integer at which $P(x)$ assumes the value 5.

We now shift our focus on another result concerning and integer polynomial and its rational roots.

Let $x = u/v$, u,v integers with $v \neq 0$, be a root of the integer polynomial $P(x)$ defined earlier. Assume that the greatest common divisor of u and v is 1 (i.e u and v are relatively prime). Then u divides a_0 and v divides a_n. To see why it is true, substitute

$x = u/v$ in $P(x)$ and clear the denominators by multiplying both sides by $(uv)^n$ to obtain

$$a_n u^n + a_{n-1}u^{n-1}v + \cdots + a_1 uv^{n-1} + a_0 v^n = 0, \tag{4}$$

whence u divides $a_0 v^n$ and v divides $a_n u^n$. As u and v are relatively prime u does not divide v^n and v does not divide u^n. Therefore u divides a_0 and v divides a_n. If $a_n = 1$ then $v = \pm 1$. Thus for a monic integer polynomial any rational root is an integer root. Let us see some applications.

Example 5

Prove that a cubic integer polynomial $ax^3 + bx^2 + cx + d$ where ad is odd and bc is even must have an irrational root.

Since ad is odd, $d \neq 0$ and hence 0 is not a root of $p(x) = ax^3 + bx^2 + cx + d$. Suppose all roots of $p(x)$ are rational and let them be r_i/s_i, where the integers $r_i \neq 0$, $s_i \neq 0$ and $(r_i, s_i) = 1$ for $i = 1, 2, 3$. Since $p(r_i/s_i) = 0$ we have

$$ar_i^3 + br_i^2 s_i + cr_i s_i^2 + ds_i^3 = 0 \tag{5}$$

whence $r_i|d$ and $s_i|a$. Since both a and d are odd, r_i and s_i are odd for each $i \in \{1, 2, 3\}$. Observe that $ar_i^3 + ds_i^3$ is even (being the sum of two odd numbers). Therefore, from (5) we see that $br_i^2 s_i + cr_i s_i^2$ must be even, implying that the two summands must be of the same parity. Since bc is even, both b and c must be even, otherwise the parities of $br_i^2 s_i$ and $cr_i s_i^2$ will be different. Now,

$$a(r_1/s_1 + r_2/s_2 + r_3/s_3) = -b, \tag{6}$$

which leads to

$$a(r_1 s_2 s_3 + r_2 s_3 s_1 + r_3 s_1 s_2) = -bs_1 s_2 s_3. \tag{7}$$

Observe that the left hand side of (7) is odd but the right hand side is even. A contradiction. Therefore $p(x)$ must have an irrational root.

Ponder This:

Let a, b, c be odd integers. Can the polynomial $ax^2 + bx + c$ have rational roots?

We observed earlier that a rational root of a monic integer polynomial is an integer root. The following example offers an exception to this fact.

Example 6

Let p be a prime number and $n \geq 3$. Let

$$Q(x) = px^n + a_{n-2}x^{n-2} + a_{n-3}x^{n-3} + \cdots + a_1x + a_0$$

be an integer polynomial with $a_0 \neq 0$ and g.c.d $(p, a_{n-2}, a_{n-3}, \ldots, a_1, a_0) = 1$. Then any rational root of $Q(x)$ is an integer root.

If $x = u/v$ is a rational root where u, v are non-zero integers and g.c.d$(u, v) = 1$ then

$$pu^n + a_{n-2}u^{n-2}v^2 + \cdots + a_1uv^{n-1} + a_0v^n = 0. \tag{8}$$

Observe that v^2 divides p. As p is a prime its only positive divisors are 1 and p. But $\sqrt{p}$ is not an integer. Hence $v^2 = 1$ and we are done.

Ponder This:

Let p be a prime, $n \geq 3$ and $k \in \{2, 3, \ldots, n\}$ be a positive integer. Is every rational root of the integer polynomial

$$p^l x^n + a_{n-k}x^{n-k} + a_{n-k-1}x^{n-k-1} + \cdots + a_1x + a_0$$

an integer root, where $a_0 \neq 0$, g.c.d$(p, a_{n-k}, \ldots, a_1, a_0) = 1$ and $1 \leq l \leq k - 1$?

The next two examples have a number theoretic flavour but surprisingly these can be solved using integer polynomials in a clever manner.

Example 7

If the non-zero integers a, b, c are such that $\dfrac{ab}{c} + \dfrac{bc}{a} + \dfrac{ca}{b}$ is an integer, then each of ab/c, bc/a, and ca/b is an integer.

Apparently there is no connection with an integer polynomial as far as we can see. But observe that

$$\left(x - \frac{ab}{c}\right)\left(x - \frac{bc}{a}\right)\left(x - \frac{ca}{b}\right) = x^3 - \left(\frac{ab}{c} + \frac{bc}{a} + \frac{ca}{b}\right)x^2 + (a^2 + b^2 + c^2)x - abc$$

is a monic integer polynomial with rational roots. So its roots must be integers.

Example 8

The positive integers a, b, c are such that $\frac{a}{b}+\frac{b}{c}+\frac{c}{a}$ and $\frac{a}{c}+\frac{b}{a}+\frac{c}{b}$ are positive integers. Prove that $a=b=c$.

The crucial observations are

- $\frac{a}{b}.\frac{b}{c}+\frac{b}{c}.\frac{c}{a}+\frac{c}{a}.\frac{a}{b}=\frac{a}{c}+\frac{b}{a}+\frac{c}{b}$;
- $\frac{a}{b}.\frac{b}{c}.\frac{c}{a}=1$.

These suggest looking at

$$\left(x-\frac{a}{b}\right)\left(x-\frac{b}{c}\right)\left(x-\frac{c}{a}\right)=x^3-\left(\frac{a}{b}+\frac{b}{c}+\frac{c}{a}\right)x^2+\left(\frac{a}{c}+\frac{b}{a}+\frac{c}{b}\right)x-1,$$

which is a monic integer polynomial with rational roots. Thus the roots are positive integers. As their product is 1, each of them must be 1, which readily shows $a=b=c$.

It is evident that any integer polynomial assumes an integer value when evaluated at an integer. Thus it is *integer-valued* at the integers. But it is not necessary for an integer-valued polynomial to have integer coefficients. For instance, the polynomial

$$P_n(x)=\binom{x}{n}:=\frac{x(x-1)(x-2)\dots(x-n+1)}{n!}$$

($n\geq 2$) is integer-valued but it has rational coefficients which are not integers. Interestingly any integer-valued polynomial $g(x)$ of a given degree can be written uniquely as

$$g(x)=\sum_{k=1}^{\text{deg(g)}}c_k\binom{x}{k}$$

where the coefficients c_k are integers and which depend on the polynomial $g(x)$.

This was just a small offering on integer polynomials to kindle the readers' interest in the topic. We hope that it propels them and encourages them to embark on an expedition to the world of integer polynomials.

Suggested reading

1. *Polynomials*, Edward J. Barbeau, Springer-Verlag, 1989.
2. *Polynomials*, Victor V. Prasolov, Springer-Verlag, 2010.

4 Optimizing Group Formation: A Mathematical examination of maximizing fun!

Advay Misra

Grade VIII
Sanskriti School
New Delhi, India.

Abstract
Students are sorted into groups throughout their academic lives. At times, sorting and grouping is necessitated by the need to work in small teams to accomplish certain tasks such as class projects, research work, or for better supervision. At other times sorting and grouping is required for logistical purposes such as accessing certain facilities, field trips or just pure fun!

Assignment of students to groups can be a challenge on multiple counts – for administrative reasons such as mandating/disallowing certain groupings; a need to ensure diversity; imperatives of group cohesiveness/productivity. This paper however primarily seeks to examine the issue from the prism of trying to democratize the process and to produce an optimal result for the group as a whole.

I describe the use of combinatorial optimization principles, specifically the applications of the Stable Roommate problem to understand the mathematical underpinnings of the issue and to arrive at a Pareto-optimal solution. The solution is implemented in Python language.

1 Introduction

Into the picnic bravely marched the 24!

Room 103 was abuzz as everyone chatted animatedly about the upcoming school trip. Post the exams and the school annual day, the prospect of being away from textbooks and classrooms had ignited a contagious excitement among the seventh-graders. The school trip to Jaipur was the talk of the town, or at least the talk of the school canteen and hallways.

The enthusiasm in the room, however, was tempered by the impending task of forming groups. The question of who would be bunking with whom, lingered. There was a subtle tension. The challenge was clear – creating groups that balanced personalities, interests, and the unspoken middle school *omerta*.

"I call dibs on being with you guys!" shouted X, waving at his two best friends, Y and Z [I will not even attempt to use real names!]. "That's not fair! We should mix things up," retorted M, eyeing the group of friends uneasily. Friends spoke with each other, alliances were forged and dissolved, some lobbied, others weighed their options. The class eventually muddled through to a draft list.

As the list was read out, a murmur of approval as well as of silent dissent rippled through the room. The wisdom of the crowds had not necessarily maximized utility in this case. This got me thinking as to what could be best ways to organize ourselves into groups.

2 Standing on the shoulders of giants

The branch of Mathematics that deals with the counting, arranging, and combining objects is referred to as *combinatorics*. It is closely related to many other areas of mathematics and has many applications ranging from logic to statistical physics and from evolutionary biology to computer science [1]. Combinatorics involves the study of discrete structures and is concerned with questions like "How many ways can a set of objects be arranged?" or "How many subsets of a certain size can be formed from a given set?"

Key concepts and topics in combinatorics include: Permutations (arrangement of objects in a specific order); Combinations (selection of objects without regard to the order); Binomial Coefficients (ways to choose k elements from a set of n elements); Pigeonhole Principle (distribution of n items into m containers; with $n > m$).

Now consider, if you will, the following problem. Imagine you have a small backpack, and you're off on an adventure. In your backpack, you can only carry a certain weight of items because it's not very big (neither are you!). You would like to take along a bunch of toys, each with different weights and values. Some are heavy, and some are light. The *knapsack problem* is to figure out the best way to fill your backpack with toys, so that you carry the most valuable toys without going over the weight limit. You want to make sure you choose the toys that give you the most fun (value) without making your backpack too heavy.

Putting it mathematically, the challenge is to:

Maximize the function $\sum_{i=1}^{n} x_i v_i$ subject to conditions $\sum_{i=1}^{n} w_i x_i \leq W$ and $x_i \in \{0, 1\}$.

Here, we are choosing from n items numbered from 1 up to n, each with a weight w_i and a value v_i, along with a maximum weight capacity W. This particular variation is known as the *0-1 variation*, since we restrict the number x_i of each kind of item to either zero or one [2].

The field of study that deals with such problems is *Combinatorial optimization.* In combinatorial optimization problems, our goal is to optimize an objective function, subject to a set of constraints, while exploring a finite though often very large solution space. The solutions to these problems are often discrete and combinatorial in nature, meaning they involve selecting a combination of elements or making discrete decisions.

Combinatorial optimization problems can be challenging because the solution space is often exponentially large, making it impractical to evaluate all possible solutions. A famous collection of especially hard problems are the *NP-hard* (Nondeterministic Polynomial-Time hard) ones – these are like searching from hundreds of keys to a lock. There are many potentially wrong choices and only one correct choice, which once known is easy to replicate and verify. Various algorithms and heuristics, such as genetic algorithms, simulated annealing, and integer programming, are commonly employed to find near-optimal solutions within a reasonable amount of time.

The sorting/grouping of a set of individuals into smaller groups is a fundamental problem in combinatorial optimization. In particular, an example of such a problem is the *Stable Marriage Problem.*

3 Marriages are made in heaven; Stable Marriages are made in algorithms!

The *stable marriage problem* (also called the stable matching problem or SMP) deals with the problem of finding a stable matching between two equally sized sets of elements given an ordering of preferences for each element (See [6]). A matching is a bijection (one-to-one correspondence) from the elements of one set to the elements of the other set.

The Wikipedia article on the stable marriage problem describes it as follows [5]:

> A matching is not stable if:
>
> There is an element A of the first matched set which prefers some given element B of the second matched set over the element to which A is already matched, and B also prefers A over the element to which B is already matched.
>
> In other words, a matching is stable when there does not exist any pair (A, B) which both prefer each other to their current partner under the matching.
>
> The stable marriage problem has been stated as follows:
>
> Given n men and n women, where each person has ranked all members of the opposite sex in order of preference, marry the men and women together such that there are no two people of opposite sex who would both rather have each other than their current partners. When there are no such pairs of people, the set of marriages is deemed stable.

So, how does all this help us to form groups in our school trip? Enter Stable Roommates!

4 Stable Roommates: “You can count on me like one, two, three!”

The *stable-roommate problem* (SRP) is the problem of finding a stable matching for an even-sized set. A matching is a separation of the set into disjoint pairs (”roommates”). The matching is stable if there are no two elements that are not roommates and that both prefer each other to their current roommate under the matching. This is distinct from the stable-marriage problem in that the stable-roommates problem allows matches between any two elements, not just between classes of ‘men’ and ‘women’.

Over the years a number of algorithms have been propounded to solve such problems. Popular among them are the Hungarian Algorithm [3] and the Irving Algorithm [4]. However, unlike the stable marriage problem, a stable matching may fail to exist for certain sets of participants and their preferences. For a minimal example of a stable pairing not existing, consider 4 people A, B, C, and D, whose rankings are:

$$\begin{aligned}A &: (B, C, D),\\ B &: (C, A, D),\\ C &: (A, B, D),\\ D &: (A, B, C).\end{aligned}$$

In this ranking, each of A, B, and C is the most preferable person for someone. In any solution, one of A, B, or C must be paired with D and the other two with each other (for example AD and BC). Yet, for anyone who is partnered with D, another member will have rated them highest, and D's partner will in turn prefer this other member over D. In this example, AC is a more favorable pairing than AD, but the necessary remaining pairing of BD then raises the same issue, illustrating the absence of a stable matching for these participants and their preferences.

An efficient algorithm was given by Irving in 1985 [4]. The algorithm will determine, for any instance of the problem, whether a stable matching exists, and if so, will find such a matching. Irving's algorithm has $O(n^2)$ complexity, provided suitable data structures are used to implement the manipulation of the preference lists and identification of rotations. The algorithm consists of two phases.

In Phase 1, participants propose to each other, in a manner similar to that of the *Gale-Shapley algorithm* for the stable marriage problem. Each participant orders the other members by preference, resulting in a preference list—an ordered set of the other participants. Participants then propose to each person on their list, in order, continuing to the next person if and when their current proposal is rejected.

A participant will reject a proposal if they already hold a proposal from someone they prefer. A participant will also reject a previously-accepted proposal if they later receive a proposal that they prefer. In this case, the rejected participant will then propose to the next person on their list, continuing until a proposal is again accepted. If any participant is eventually rejected by all other participants, this indicates that no stable matching is possible. Otherwise, Phase 1 will end with each person holding a proposal from one of the others.

Consider two participants, p and q. If q holds a proposal from p, then we remove from q's list all participants x after p, and symmetrically, for each removed participant x, we remove q from x's list, so that q is first in p's list; and p is last in q's, since q and any x cannot be partners in any stable matching. The resulting reduced set of preference lists together is called the Phase 1 table. In this table, if any reduced list is empty, then there is no stable matching. Otherwise, the Phase 1 table is a stable table.

A stable table, by definition, is the set of preference lists from the original table after members have been removed from one or more of the lists, and the following three conditions are satisfied (where reduced list means a list in the stable table):

(i). p is first on q's reduced list if and only if q is last on p's,

(ii). p is not on q's reduced list if and only if q is not on p's if and only if q prefers the last person on their list to p; or p, the last person on their list to q,

(iii). no reduced list is empty.

Stable tables have several important properties, which are used to justify the remainder of the procedure:

Any stable table must be a subtable of the Phase 1 table, where subtable is a table where the preference lists of the subtable are those of the supertable with some individuals removed from each other's lists. In any stable table, if every reduced list contains exactly one individual, then pairing each individual with the single person on their list gives a stable matching. If the stable roommates problem instance has a stable matching, then there is a stable matching contained in any one of the stable tables.

Any stable subtable of a stable table, and in particular any stable subtable that specifies a stable matching as in 2, can be obtained by a sequence of rotation eliminations on the stable table. These rotation eliminations comprise Phase 2 of Irving's algorithm. The pseudo code for the algorithm is summarized as under:

```
T = Phase 1 table;
while (true) {
    identify a rotation r in T;
    eliminate r from T;
    if some list in T becomes empty,
        return null; (no stable matching can exist)
    else if (each reduced list in T has size 1)
        return the matching M = {{x, y} | x and y are on each other's
                                    lists in T}; (this is a stable matching)
}
```

5 Stable Roommates: "Are we there yet?"

Not quite. While we have got a mechanism for solving the Stable Roommates problem, the above solutions essentially work in cases where the "room size" is two. i.e., each person is to be matched with one other person. The formation of bigger groups (such as groups for a certain field trip for Grade VII!) presents an additional layer of challenge.

In order to generalize the solution, we would need a mixture of heuristics and a recursive algorithm. The algorithm in [7] "uses cardinal method (rating each member on a scale) instead of the usual preference ordering, because it carries more information, is easier to collect, and makes aggregating preferences easier."

The group size is defined at the outset. Subsequently, each group is randomly initialized with a single member. The left over members then propose to each group in order of their preference (This is similar to the Gale-Shapley algorithm) .

Each group accepts one new member in each iteration. At the end of each iteration, we swap members between groups if swapping improves overall score and continue the process till the last member is grouped. Finally, we iterate one last time to check if swapping members improves overall score. Using this approach, we implemented the algorithm using Python. The code is included as an appendix to this article.

6 Conclusion: Understanding the Magic!

The above is an attempt to understand the theories underlying the formation of groups. There exists an opportunity to introduce a number of further refinements such as embargoes/mandates on certain groupings; need to rotate and encourage gregariousness and democratic conflict resolution.

The algorithm discusses above has implications in matching service providers with recipients according to their preference/dis-preference (like service aggregators); matching students to classes; scheduling etc.

Friendships don't fall within the neat mandates of Mathematics. Perhaps that's what makes them magical. However, being able to understand the mechanics behind the process helps understand the world a bit better.

Acknowledgements

This paper has benefited from discussions with Dr. Robert Irving, Honorary Research Fellow (School of Computing Science), University of Glasgow. Dr. Irving has done pioneering work in the field of combinatorics. A link to his publications is at https://glasgow.academia.edu/RobertIrving.

Bibliography

[1] Garrett Birkhoff, *Lattice Theory* (3rd ed., reprinted with corrections). American Mathematical Society, 1984.

[2] Maya Hristakeva and Dipti Shrestha, *Different approaches to solve the 0/1 knapsack problem*, https://micsymposium.org/mics_2005/papers/paper102.pdf

[3] Derek Bruff, *The Assignment Problem and the Hungarian Method.* https://web.archive.org/web/20120105112913/http://www.math.harvard.edu/archive/20_spring_05/handouts/assignment_overheads.pdf

[4] Robert W. Irving, An Efficient Algorithm for the “Stable Roommates” Problem, *Journal of Algorithms* **6**, pp 577–595 (1985). https://uvacs2102.github.io/docs/roomates.pdf

[5] Wikipedia, *Stable marriage problem.* https://en.wikipedia.org/wiki/Stable_marriage_problem

[6] Wikipedia, *Stable roommates problem.* https://en.wikipedia.org/wiki/Stable_roommates_problem

[7] Anjay Goel, *Stable Roommate Generalised.* https://github.com/AnjayGoel/Stable-Roommate-Generalised/blob/main/README.md

Appendix: Python code

```
## Code written by Advay Misra;
## Companion code for
## 'Optimizing Group Formation: A Mathematical Examination of Having Fun';

import sys
import random

# a variety of ways to get input

k = int(sys.argv[1])
n = int(sys.argv[2])
filename = sys.argv[3]
if filename == "ranking":
    preferences = [random.sample(list(range(n)), n) for m in range(n)]
elif filename == "scores":
    min_score = int(sys.argv[4])
    max_score = int(sys.argv[5])
    preferences = [[random.randint(min_score, max_score)
                    for l in range(n)] for m in range(n)]
else:
    preferences = [a.split(',') for a in open(filename).read().splitlines()]

for m in range(n):
    preferences[m][m] = 0

# print 2-d list nicely

def formatted_print(lst):
    print('\n'.join('\t'.join(str(cell) for cell in row) for row in lst), end='\n\n')

# important functions for scoring

def group_score(group, preferences):
    r = 0
    for member1 in group:
        for member2 in group:
            r += preferences[member1][member2]
    return r

def score(groups, preferences):
    r = 0
    for group in groups:
        r += group_score(group, preferences)
    return r

# flip two people's assignment

def flip(groups, groups_dict, flop):
    groups[groups_dict[flop[0]]][groups[groups_dict[flop[0]]].index(flop[0])] = flop[1]
    groups[groups_dict[flop[1]]][groups[groups_dict[flop[1]]].index(flop[1])] = flop[0]
    groups_dict[flop[0]], groups_dict[flop[1]] = groups_dict[flop[1]],
                                                                 groups_dict[flop[0]]


# announce preferences

print('\n\n')
print('Preferences:')
formatted_print(preferences)
print('\n\n')
```

```
# the actual algorithm

# random initial assignment
groups = []
free = set(range(n))
groups_dict = {}
while free:
    new = random.sample(list(free), min(k, len(free)))
    free -= set(new)
    for member in new:
        groups_dict[member] = len(groups)
    groups.append(new)
random_score = score(groups, preferences)
random_assgn = tuple(tuple(group) for group in groups)

print('\n\n')
print("Initial, random assignment:")
formatted_print(random_assgn)
print(f'Score: {random_score}')
print('\n\n')

best_flip = 1
current_score = random_score

while best_flip is not None:
    best_flip = None
    best_flip_score = 0
    for person1 in range(n):
        for person2 in range(n):
            if groups_dict[person1] == groups_dict[person2] or person1
                                                    >= person2: continue
            flip(groups, groups_dict, (person1, person2))
            flip_score = score(groups, preferences)
            print(f"Swapping {person1} with {person2} gives a score of
                                                   {flip_score}.")
            flip(groups, groups_dict, (person1, person2))
            if flip_score > best_flip_score:
                best_flip_score = flip_score
                best_flip = person1, person2
    if best_flip_score >= current_score:
        flip(groups, groups_dict, best_flip)
        print(f"The best of those, swapping {best_flip[0]} with {best_flip[1]},
                                       gives a score of {best_flip_score}")
    else:
        print("Local maximum attained.")
        break
    current_score = best_flip_score


print('\n\n')
print('Preferences:')
formatted_print(preferences)
print('\n\n\n\n')
print("Initial, random assignment:")
formatted_print(random_assgn)
print(f'Score: {random_score}')
print('\n\n\n\n')
print("Final assignment:")
formatted_print(groups)
print(f"Score: {score(groups, preferences)}")
print('\n\n')

## Finis!
```

5 Generalization of an RMO problem

Swastika Dey

B. Stat. Hons.
Indian Statistical Institute
Kolkata, India.
Email: swastikadey256@gmail.com

1 Introduction

In this article, we consider a generalization of a problem that appeared in the Regional Mathematical Olympiad (RMO) 2001. Let us first solve the RMO problem. The method of proof gives us an idea to obtain a generalisation.

2 The RMO Problem and its Solution

Problem 1. *Find the number of natural numbers n such that*

$$\left[\frac{n}{99}\right] = \left[\frac{n}{101}\right]. \tag{1}$$

(where $[x]$ denotes greatest integer function of x).

Firstly, we note that for a solution of $[n/99] = [n/101]$, the common value is less than 50. For, if $[n/99] = k = [n/101]$, then

$$99(k+s) = n = 101(k+t)$$

for some $0 \leq s, t < 1$. Hence $2k = 99s - 101t < 99$, giving $k < 50$. So, we only need to count the solutions of $[n/99] = k = [n/101]$ for $k = 0, 1, \ldots, 49$. This is easy to do, and we get:

$$\left[\frac{n}{99}\right] = \left[\frac{n}{101}\right] = 0 \Leftrightarrow n = 0, 1, 2, 3, \ldots, 98$$
$$\left[\frac{n}{99}\right] = \left[\frac{n}{101}\right] = 1 \Leftrightarrow n = 101, 102, 103, \ldots, 197$$
$$\left[\frac{n}{99}\right] = \left[\frac{n}{101}\right] = 2 \Leftrightarrow n = 202, 203, 204, \ldots, 296,$$

etc., until

$$\left[\frac{n}{99}\right] = \left[\frac{n}{101}\right] = 49 \Leftrightarrow n = 4949.$$

Hence the total number of solutions in non-negative integers is the sum of the first 50 odd numbers, $99+97+95+\cdots+3+1 = 2500$. Thus, the number of solutions in positive integers n is 2499.

From the above we get the first generalization:

Proposition 1. *For any $a \in \mathbb{N}$, the number of solutions $n \in \mathbb{N}$ to the equation*

$$\left[\frac{n}{2a-1}\right] = \left[\frac{n}{2a+1}\right]$$

is equal to $\big((2a-1) + (2a-3) + (2a-5) + \cdots + 1\big) - 1 = a^2 - 1$.

We have a more general result below.

Notation

We shall write, for the rest of this article, $\boxed{f(n,x) = [n/x]}$.

Proposition 2. *For any $a, b \in \mathbb{N}$, $b > a$, the number of solutions $n \in \mathbb{N}$ to the equation*

$$f(n,a) = f(n,b)$$

is equal to

$$\frac{(k+1)\big((b-a)k + 2i\big)}{2} - 1,$$

where k and i are defined as follows:

$$k = \left[\frac{a}{b-a}\right], \quad i = a - (b-a)k.$$

3 Examples

We consider a few examples to explain the above formula.

Example 1

The case $a = 4$, $b = 6$

The equation is

$$f(n,4) = f(n,6), \quad n \in \mathbf{N}.$$

We find that $f(n,4) = 0 = f(n,6)$ for $n = 0,1,2,3$; and $f(n,4) = 1 = f(n,6)$ for $n = 6,7$. There are no other values of n for which the two sides are equal. Therefore the total number of solutions in natural numbers is $4 + 2 - 1 = 5$.

Example 2

The case $a = 4$, $b = 5$

The equation is

$$f(n,4) = f(n,5), \quad n \in \mathbf{N}.$$

We find that:

- $f(n,4) = 0 = f(n,5)$ for $n = 0,1,2,3$;
- $f(n,4) = 1 = f(n,5)$ for $n = 5,6,7$;
- $f(n,4) = 2 = f(n,5)$ for $n = 10,11$; and
- $f(n,4) = 3 = f(n,5)$ for $n = 15$.

There are no other values of n for which the two sides are equal. Therefore the total number of solutions in natural numbers is $4 + 3 + 2 + 1 - 1 = 9$.

Example 3

The case $a = 4$, $b = 7$

The equation is

$$f(n,4) = f(n,7), \quad n \in \mathbf{N}.$$

We find that:

- $f(n,4) = 0 = f(n,7)$ for $n = 0, 1, 2, 3$;
- $f(n,4) = 1 = f(n,7)$ for $n = 7$.

There are no other values of n for which the two sides are equal. Therefore the total number of solutions in natural numbers is $4 + 1 - 1 = 4$.

Example 4

The case $a = 5$, $b = 8$

The equation is

$$f(n,5) = f(n,8), \quad n \in \mathbf{N}.$$

We find that:

- $f(n,5) = 0 = f(n,8)$ for $n = 0, 1, 2, 3, 4$;
- $f(n,5) = 1 = f(n,8)$ for $n = 8, 9$.

There are no other values of n for which the two sides are equal. Therefore the total number of solutions in natural numbers is $5 + 2 - 1 = 6$.

Example 5

The case $a = 6$, $b = 9$

The equation is

$$f(n,6) = f(n,9), \quad n \in \mathbf{N}.$$

We find that:

- $f(n,6) = 0 = f(n,9)$ for $n = 0, 1, 2, 3, 4, 5$;
- $f(n,6) = 1 = f(n,9)$ for $n = 9, 10, 11$.

There are no other values of n for which the two sides are equal. Therefore the total number of solutions in natural numbers is $6+3-1=8$.

4 Some general results

We now study a few families of problems of this type: "Find the number of $n \in \mathbb{N}$ such that $f(n,a)=f(n,b)$, for given $a,b \in \mathbb{N}$, $a<b$." By studying the results, we will arrive at the general statement described earlier.

We first note that for any solution of $f(n,a)=f(n,b)$, where $a<b$, the common value of the two quantities must be less than $a/(b-a)$. For, if $f(n,a)=k=f(n,b)$, then

$$a(k+s)=n=b(k+t)$$

for some $0 \leq s,t < 1$. Hence $(b-a)k = as-bt < a$, giving $k < a/(b-a)$.

Problem 2. *For a given $a \in \mathbb{N}$, to find the number of solutions $n \in \mathbb{N}$ to the equation*

$$f(n,a)=f(n,a+1).$$

Solution. *In this case the common value of the two quantities cannot exceed a.*

We observe the following:

$$\begin{aligned}
f(n,a)=f(n,a+1)=0 &\quad \textit{when } n=0,1,2,\ldots,a-1,\\
f(n,a)=f(n,a+1)=1 &\quad \textit{when } n=a+1,a+2,\ldots,2a-1,\\
f(n,a)=f(n,a+1)=2 &\quad \textit{when } n=2a+1,2a+2,\ldots,3a-1,\\
\cdots=\cdots\ ,&\\
f(n,a)=f(n,a+1)=a &\quad \textit{when } n=a^2-1.
\end{aligned}$$

Hence the required number of solutions is

$$a+(a-1)+(a-2)+\cdots+1-1=\frac{a(a+1)}{2}-1.$$

This may also be written as

$$\frac{(k+1)(k+2i)}{2}-1.$$

Problem 3. *For a given $a \in \mathbb{N}$, to find the number of solutions $n \in \mathbb{N}$ to the equation*

$$f(n,a)=f(n,a+2).$$

Solution. *Here the common value of $f(n,a)$ and $f(n,a+2)$ cannot exceed $a/2$. Let $k=[a/2]$ and $a=2k+i$ where $i\in\{0,1\}$; then the common value cannot exceed $(a-i)/2$. We observe the following:*

$$\begin{aligned}
f(n,a)=f(n,a+2)&=0 \quad \textit{when } n=0,1,2,\ldots,a-1,\\
f(n,a)=f(n,a+2)&=1 \quad \textit{when } n=a+2,a+3,\ldots,2a-1,\\
f(n,a)=f(n,a+2)&=2 \quad \textit{when } n=2(a+2),2(a+2)+1,\ldots,3a-1,\\
\cdots&=\cdots.
\end{aligned}$$

If a is even ($i=0$), there is no value of n for which the common value $(a-i)/2$ is attained. If a is odd ($i=1$), there is just one value of n (namely, $n=2a^2+3a$) for which the common value $(a-1)/2$ is attained. Hence there are i solutions of $f(n,a)=f(n,a+2)=(a-i)/2$.

Therefore the required number of solutions is

$$\begin{aligned}
\big(a+(a-2)+(a-4)+\cdots+i\big)-1&=(2k+i)+(2k-2+i)+\cdots+i-1\\
&=\frac{(k+1)(2k+2i)}{2}-1.
\end{aligned}$$

Problem 4. *For a given $a\in\mathbb{N}$, to find the number of solutions $n\in\mathbb{N}$ to the equation*

$$f(n,a)=f(n,a+3).$$

Solution. *Here the common value of $f(n,a)$ and $f(n,a+3)$ cannot exceed $a/3$. Let $k=[a/3]$ and $a=3k+i$ where $i\in\{0,1,2\}$; then the common value cannot exceed $(a-i)/3$. Proceeding as earlier, we find that:*

$$\begin{aligned}
f(n,a)=f(n,a+3)&=0 \quad \textit{when } n=0,1,2,\ldots,a-1,\\
f(n,a)=f(n,a+3)&=1 \quad \textit{when } n=a+3,a+4,\ldots,2a-1,\\
f(n,a)=f(n,a+3)&=2 \quad \textit{when } n=2(a+3),2(a+3+1,\ldots,3a-1,\\
\cdots&=\cdots.
\end{aligned}$$

Like earlier we find that there are i solutions of $f(n,a)=f(n,a+3)=(a-i)/3$.

Hence the required number of solutions is

$$\begin{aligned}
\big(a+(a-3)+(a-6)+\cdots+i\big)-1&=(3k+i)+(3k-3+i)+\cdots+i-1\\
&=\frac{(k+1)(3k+2i)}{2}-1.
\end{aligned}$$

We observe that the final formula has the same form as earlier. It is easy to work out the general form and therefore the generalized result.

5 Generalization

Problem 5. *Let $a, m \in \mathbb{N}$ be given. Let $[a/m] = k$ and $i = a - mk$, so $i \in \{0, 1, 2, \ldots, m-1\}$. Find the number of solutions n to the equation*

$$f(n, a) = f(n, a + m),$$

in terms of m, k, i.

Solution. *By enumerating the possibilities as earlier, we find that the number of solutions is*

$$\begin{aligned}\big(a + (a-m) + (a-2m) + \cdots + i\big) - 1 &= \big((mk+i) + (mk - m + i) + \cdots + i\big) - 1 \\ &= \frac{(k+1)(mk+2i)}{2} - 1. \qquad \blacksquare\end{aligned}$$

The above analysis also yields a solution to the problem when it is stated in this form:

Problem 6. *Let $a, b \in \mathbb{N}$ be given, $b > a$. Find the number of solutions n to the equation*

$$f(n, a) = f(n, b),$$

in terms of a, b.

Solution. *Put $m = b - a$ in the solution of Problem 6.*

6 A Short R Code for The Problem

```
#taking two random integers

c<– sample(1 : 100000,1) d<– sample(1 : 100000,1)

#assigning a to the min and b to the max

a<–0 b<–0 if(c!=d){
      a <– min(c,d)
      b <– max(c,d)
      print(a)
      print(b)
} a b

#first observe no. of solutions when [n/a]=[n/b]=0

count_0 = 0 for(n in 1 : (a-1)){
     if(floor(n/a)==floor(n/b)){
          count_0 = count_0 + 1
          count_0
     }
}

#observe that for all non-(-ve) integer i, no. of solutions for
[n/a]=[n/b]=i > no. of solutions for [n/a]=[n/b]=(i + 1), so
[n/a]=[n/b]=0 has greater no. of solutions than any natural no. i in
[n/a]=[n/b]=i and the largest i can't be more than no. of solutions
for [n/a]=[n/b]=0. So, any solution n of [n/a]=[n/b] can't exceed
(no. of solutions for [n/a]=[n/b]=0)^2

max_count = (count_0 +1)*(count_0 +1) count0 = 0

#finding total no. of solutions

for(n in 1 : max_count){
     if(floor(n/a)==floor(n/b)){
          count0 = count0 + 1
          count0
     }
} count0

#checking if formula gives same answer

m<- (b-a) k<-floor(a/m)
r<- a %% m
count_1 <- ((k+1)*(m*k + 2*r))/2 -1 if(count0 == count_1){
     print(”yes”)
}
```

7 Solution to RMO Problem

In our generalized formula, put $a = 99$, $b = 101$. Then $m = 2$, $k = 49$, $i = 1$. The required number is then

$$\frac{(49+1)\times(2\times 49+2\times 1)}{2}-1=\frac{50\times 100}{2}-1=2500-1=2499.$$

Bibliography

[1] M. R. Modak, S. A. Katre, V. V. Acharya, V. M. Sholapurkar, *An Excursion In Mathematics*, Bhaskaracharya Pratishthana, 2018.

[2] David M. Burton, *Elementary Number Theory*, 7th ed., McGraw Hill, 2023.

6 Socratic Dialogues to Discover a Proof

Jyotirmoy Sarkar

Indiana University Indianapolis
Department of Mathematical Sciences
402 N Blackford Street
Indianapolis, IN 46202-3216, USA.
Email: jsarkar@iu.edu

Abstract
Given a point A away from a line l and a point B belonging to line l, find a point T on l such that if an ant walks straight from A to T, turns at a right angle (without passing through l) and again walks straight, then when it meets the line AB at S, it will be as close to A as possible.

We document a thought process (or a mental dialogue) a diligent inquirer might follow to discover a solution to this problem. Our objective is to give a young scholar a taste of mathematical research. We conclude with a formal statement and proof of a Euclidean geometric theorem that solves this optimization problem.

Key Word and Phrases: minimization problem, coordinate geometry, intermediate value theorem, Euclidean geometry, angle bisector theorem, similar triangles.

To Reveal the Discovery Process

In mathematics, it is customary to first write the statement of a result and then present its deductive proof. The proof (unless by contradiction) typically proceeds from hypotheses to conclusion following intermediate logical steps involving axioms, definitions, assumptions, and already-proved results. This streamlined method is praised for its beauty and elegance. However, often, it fails to illuminate the inspiration, imagination, and creativity in discovering the proof.

Here, we reveal the discovery process to inspire young scholars to conduct mathematical research. To do so, we adopt a Socratic method of conversation between an inquirer (Kajal Thakur) and a guide (Professor Mrinal Chand). Kajal is a typical college student studying biology who lives in the on-campus dormitory and actively participates in cultural programs. Professor Chand lives in the faculty quarters at the end of the campus across a pond from the student dorm.

Having revealed the discovery process, we revert to the customary "theorem-proof" method to illustrate proper mathematical writing. We strongly urge diligent readers to drop the paper after reading the problem statement in Section 1 and instead solve the problem. Later, they may return to read the rest of the paper to compare their solutions with ours and to pick up some tips.

Section 1 poses the problem. Section 2 presents Socratic conversations between Kajal and Dr. Chand in five subsections. Section 3 presents both an informal and a formal Euclidean geometric proof. Section 4 resolves some previously raised issues.

1 An Ant Needs Your Advice

On the Mathematics department's bulletin board was posted the following problem.

Problem of the Month: Point A is away from a line l, and point B belongs to line l. An ant will start its journey from A and walk straight to a point T on l, then it will turn at right angle (without passing through l) and again walk straight until it will meet line AB at S. The ant wants to reach this point S as close to A as possible. Advice the ant which point T on l it must initially go toward.

2 The Would-be Advisor Seeks Advice

A critical reader should understand the problem, read up on some background material, seek advice from reliable resource persons, converse with one's alter self, pursue different choices, struggle with unfamiliar concepts, overcome a few intermediate hurdles, discover new ideas and results until she sees the light at the end of the proverbial tunnel.

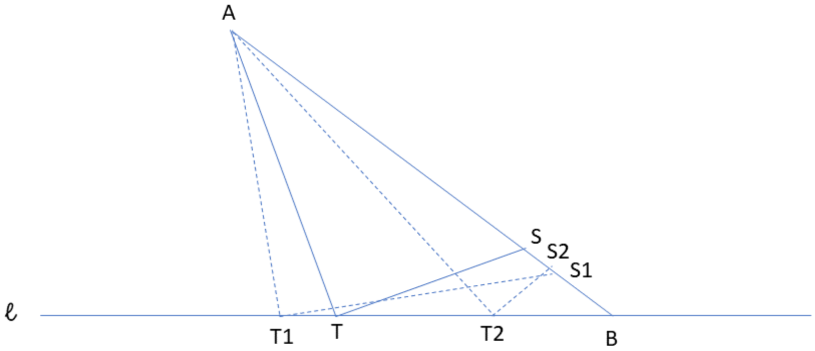


Figure 1: Starting at A, an ant heads to T on l, turns at right angle, and ends at S on AB. Which choice of T makes AS the shortest?

Afterwards, eliminating all extraneous details, she must find the smoothest path from hypothesis to conclusion.

In the forthcoming five subsections, our would-be ant advisor, Kajal Thakur, seeks advice from an encouraging guide, Professor Mrinal Chand.

2.1 Reformulating the Problem

KT: Greetings, Dr. Chand. I read on the bulletin board that you want someone to advise an ant find a point on a line. Can you please tell me more about this extra-ordinary request?

MC: Kajal, you read it right. Would you take up the challenge?

KT: I am not qualified to advise an ant. How do I communicate with ants? I do not know their language, nor do they know mine.

MC: No worries: Leave it to an entomologist. You focus on solving the mathematical problem at hand. I know, advising an ant does not sound like an attractive proposition. I wanted a line to join A and T. So, I said to myself: “A-n-T”. That’s the genesis of this ant.

KT: You are funny; I mean “punny”.

MC: Pun aside, if it helps, think of advising an elephant (**aleph ant**). If you give a wrong advise to an elephant, it will trample you. But if you give it a good advise, it will provide you a lifetime of free, royal mode of transportation.

KT: I get it: The stakes are high when dealing with an elephant. I will do my best to give it the perfect answer — with your help. Let me restate the mathematical problem the way I understand it.

MC: That's the spirit! It is a good idea to describe the problem in your own words. Explain what information is given to you and what you must show or find or prove.

KT: I will try. Let us refer to Figure 1. There is a line l containing a point B. Another point A is away from the line l. Of course, l and A together define a unique plane on which all actions take place. No information is known about the angle at B between BA and l.

MC: So far, so good. You have described the given information quite well. Now tell me what is being sought.

KT: To find a point T on l such that if the line through T orthogonal to AT meets AB at S, then the length of segment AS is minimized.

MC: Very good. Can you classify which branch of mathematics this problem belongs to?

KT: It belongs to calculus since we must *minimize* AS.

MC: That surely is possible. How would you proceed to minimize AS?

KT: I will express AS as a function of T, differentiate the function with respect to T, equate the derivative to 0, and solve for T.

MC: How do you differentiate with respect to a point T?

KT: Pardon me, Professor. I mean to differentiate the length of AS with respect to the length t of BT as T varies over l.

MC: Much better now. To do so, you need a coordinate system to express the lengths of different segments. Do you recall how to impose a coordinate system on a plane?

KT: Yes, Dr. Chand. I should choose the two mutually orthogonal axes. Wherever they intersect will be called the origin $(0, 0)$. Since I have two lines, l and AB, perhaps I can choose one of them as one axis, and the other axis should be orthogonal to the first. But where should they intersect? Again, I have two natural choices — A and B.

MC: You are making good choices. Keep searching for all options — natural or contrived. For each choice, express the length of AS, as a function of a suitably chosen variable.

KT: Thank you for the suggestion. May I go back to the dormitory and work on these choices. When is the best time to see you tomorrow?

MC: How about at the same time as you came in today?

KT: Goodbye, Professor.

The first thing Kajal did after leaving Professor Chand's office was to visit the library and borrow *Stewart* [9] and *Swokowski* [10].

2.2 Calculus is a Powerful Tool

Kajal spent half the night (eating chips and) sketching and scribbling on paper to express the length of AS under various choices for the coordinate axes. Four different choices (shown in Figure 2) led to expressions for the corresponding objective functions, which, when optimized, would produce the target point T the ant must initially proceed toward.

Equipped with the figures, the expressions of AS and the optimal solutions for T that minimizes AS, Kajal visited Professor Chand the next day.

[Dear readers, please STOP reading further. Instead, reconstruct Kajal's solutions based on each diagram in Figure 2. Can you draw a diagram different from these? You may return to the paper after doing these tasks.]

MC: Welcome back, Kajal. What's new and exciting?

KT: Dr. Chand, I followed your advice and pursued every choice. They all solve the optimization problem, even though the final expressions appear to differ!

MC: Indeed exciting! Do you have a favorite among these solutions?

KT: What do you mean by "a favorite"? Aren't they all equally good?

MC: Sure, they may be equally good, provided they are all correct. But doesn't one appeal to you more than the others? Perhaps because of the simplicity

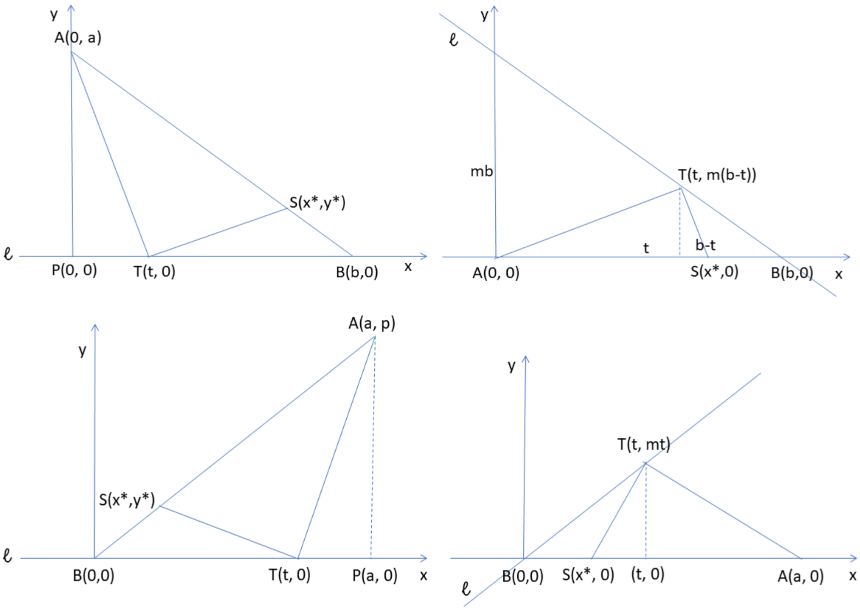


Figure 2: Four choices of coordinate axes: (1) (l, A), (2) (AB, A), (3) (l, B), (4) (BA, B). The coordinates of S and the length of AS are functions of t. The goal is to minimize AS as a function of t.

of the choice, the ease of calculations, the compactness of the final answer, and suggestiveness to other interesting results?

KT: Oh, I see. I have not compared the methods according to those criteria. Can we discuss them all? Please guide me in determining which one meets which criteria to what extent.

MC: Be my guest. Go to the board. Erase what's already on it.

KT: Thank you.

Kajal went to the board, drew the diagrams, wrote the objective functions and the optimal solutions, showing Professor Chand the derivations he had scribbled.

[Dear readers, I suppose you also have scribbled the derivations. If so, well done. I salute you. If not, do it promptly before reading further.]

MC: Can you walk me through your reasoning to derive the objective function? Begin with the coordinate axes you chose.

KT: Sure. Refer to the top left diagram in Figure 2. I'm calling this diagram the (l, A) case, because the horizontal axis is l and the vertical axis passes through A. The origin is neither A nor B, even though I had thought those were the natural choices.

MC: So, you allowed a slightly contrived choice. You are well within your right to do so. Keep talking.

KT: Since I projected A onto l, I labeled the foot of the perpendicular as $P(0, 0)$, and labeled $A(0, a)$ and $B(b, 0)$. If the target point is $T(t, 0)$, then the slope of AT is $-a/t$; hence, the slope of TS is the negative reciprocal, or t/a. Also, since TS passes through $T(t, 0)$, the equation of TS is $y = t(x - t)/a$.

MC: Sounds good to me. However, if your readers are young scholars, please provide them a reference — perhaps a textbook. What's next?

KT: I will refer them to *Swokowski* [10]. Furthermore, the equation of AB is $y/a + x/b = 1$. Therefore, by solving the system of equations for TS and AS, we find the intersection point $S(x^*, y^*)$, where

$$x^* = b\,\frac{t^2 + a^2}{bt + a^2} \quad \text{and} \quad y^* = a\left(1 - \frac{x^*}{b}\right) = a\,\frac{bt - t^2}{bt + a^2}.$$

Hence, the distance between A and S, by the Pythagorean theorem, becomes $AS = \sqrt{x^{*2} + (y^* - a)^2} = \sqrt{a^2 + b^2}\;(t^2 + a^2)/(bt + a^2)$.

MC: Bravo! You were very attentive in *Analytical Geometry* class.

KT: You are so kind. Thank you. Actually, I was sick that semester, and flunked the final exam. The Dean gave me another chance to retest. That's when I poured myself heart and soul into the subject.

MC: Way to go. It seems to have paid off well. What's your next step?

KT: I differentiated the length of AS with respect to t; I equated the derivative to 0; and I solved to get the minimizer $t^* = a(-a \pm h)/b = ab/(a \pm h)$, where $h = \sqrt{a^2 + b^2}$ is the hypotenuse of the right $\triangle APB$. In finding the minimizer, I used the quadratic formula. Anyone can read up on quadratic equation at the internet site [11].

MC: Nice. You have mentioned the reference even before I asked for it. But aren't you missing an important step? How do you know you have minimized and not maximized?

KT: Yeah, yeah, I remember now. I should check the second-order condition: the second derivative must be positive.

MC: Now you are demanding too much. The entire second derivative need not be positive. Suffices to ensure that the second derivative is positive at the solution to the first-order condition or at t^*.

KT: Ah, I stand corrected. Let me jot it down in my notebook to fill in the details later.

MC: While you take notes, may I suggest you read *Kumar* [5] to learn about some applications of derivatives. At this time, you are still missing one final task! You found the minimizer t^*, but what's the minimum value of AS?

KT: Okay, that's easy. I can evaluate AS at t^* to see that the minimum value is $2a/(1 \pm a/h)$.

MC: What's easy for you may be challenging for others. And vice versa. Easy or not, you must do it for the sake of completion.

KT: Noted; thank you. Do you agree that I have solved the problem?

MC: Not so fast. Don't you see the surprise glistening for your attention?

KT: How do you mean?

MC: You began your adventure with the objective to find an optimal T. Do you think you have found it?

KT: Sure, I have found the desired $T^*(t^*, 0)$, where $t^* = ab/(a \pm h)$. ... O my God, O my God! I have found two optimal solutions!! If I choose $T_1 = (ab/(a+h), 0)$, then $AS_1 = 2ha/(h+a)$. But if I choose $T_1 = (ab/(a-h), 0)$, then $AS_2 = 2ha/(h-a)$. That's out of the world crazy! Oh my God. Where did the second solution come from?

MC: Calm down, calm down. You can invoke God when you earn your free, royal transportation. There is yet much work to be done. Can you go back to the diagram and draw the two optimal solutions?

KT: I suppose I must return to my dorm and finish all the tasks you have assigned me today. I will bring back the drawings in a day or two.

MC: That is fine with me. But tomorrow I am off to a conference in Hyderabad. So meet me in five days. Meanwhile, go through all the other diagrams and raise the same questions and issues in the — what did you say? — the (l, A) case.

KT: Will do the same in the other cases: $(AB, A), (l, B)$ and (BA, B). Have a good trip to Hyderabad and back. Good bye.

Knowing that he can finish the assigned tasks in a couple of days at most, on his way back, Kajal stopped at the library, and borrowed *Kumar* [5] and *Polya* [7].

[Dear readers, follow this "modified golden rule": Do for Kajal as you would like Kajal to do for you. (Paraphrased from Luke 6:31 of the *Bible.* "Do to others as you would like them to do to you." — *New Living Translation.*)]

2.3 All Roads Lead to Home

KT: Hello, Dr. Chand. How was your conference?

MC: Fine. I ran into a Professor Murthy, who wants a summer undergraduate research student. I mentioned your name. How would you like to work on Biomathematics during the summer break?

KT: It will be an honor. Thank you for your recommendation.

MC: Now update me on your progress in advising the elephant.

KT: I worked out the details for all four cases. They are all fruitful.

MC: Do you have a favorite now?

KT: I sure do. My favorite is the (BA, B) case. Here, $B(0, 0)$ is the origin, $A(a, 0)$ is the ant's starting location on the horizontal axis BA, and since l passes through B, I assume its equation is $y = mx$. If the target point is $T(t, mt)$, then the destination is $S(t - m^2t^2/(a - t), 0)$.

MC: What makes it your favorite?

KT: The ease of computing and an immediate implication: Here, minimizing AS is equivalent to maximizing $BS = t - m^2t^2/(a - t)$. The solution is $t^* = a \pm p$, where $p = am/\sqrt{1 + m^2}$ is the perpendicular distance from A to l. When we choose $t^* = a - p$, the minimum length of AS is $p + m^2(a - p)^2/p = 2am(\sqrt{1 + m^2} - m)$, which is directly proportional to a, with a proportionality constant less than 1, as anticipated. I suppose I can leave to my readers to verify these claims, and to find out what happens if we choose $t^* = a + p$.

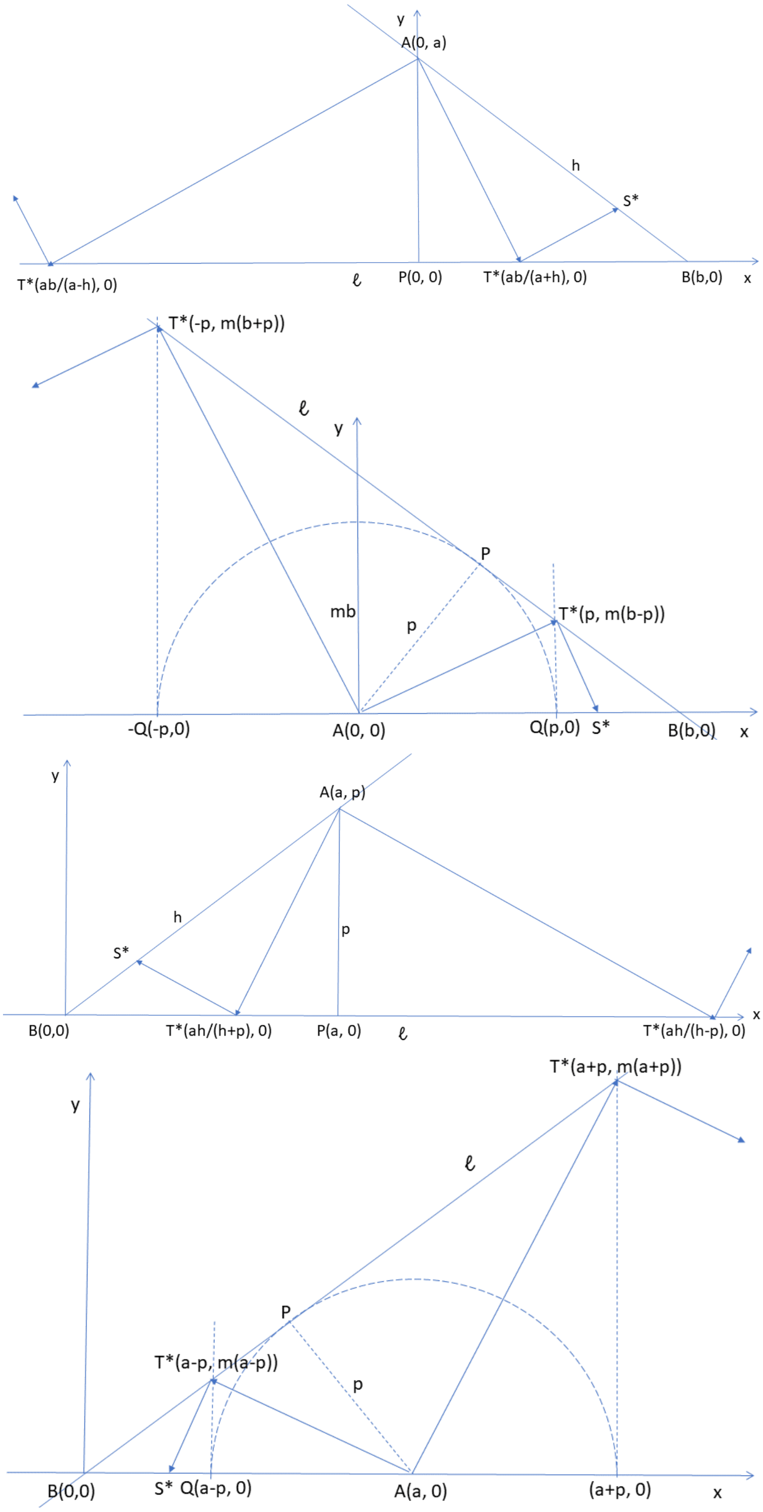


Figure 3: For all four choices of coordinate axes — (1) (l, A), (2) (AB, A), (3) (l, B), (4) (BA, B) — the optimal solutions to the target point T that minimizes the length of AS. We can show only one S^{*}; the other one falls beyond the page.

MC: Fantastic! I like that you leave some tasks for your readers. How do the other two cases compare with the two cases you told me about?

KT: In fact, case (l, B) is similar to case (BA, B) and case (AB, A) is similar to case (l, A). They are mirror images of each other!

MC: Then you should also favor case (l, B), shouldn't you?

KT: I suppose so. By the way, here is Figure 3 showing the diagrams of the *pairs* of optimal solutions in all four cases.

MC: Great. Can you find some interesting properties in these diagrams?

KT: What do you mean by "interesting properties"?

MC: I mean special features that might reveal the solutions more straightforwardly.

KT: I have to sleep over that and talk to you tomorrow.

MC: While you are at it, do you think there can be any other interesting coordinate system? Or have you exhausted them all?

KT: How can anything be simpler than the four cases I have already considered? Likely not.

MC: Think again. What about polar coordinates? That is, a ray being rotated about its starting point?

KT: That is interesting! Why didn't I think of that? I suppose I must excuse myself and go back to discovery mode.

MC: I like your spirit. All the best. Meet me tomorrow.

Kajal went to the library and borrowed *Abramson* [1].

[Dear readers, why don't you beat Kajal to this race to discover some features in Figure 3 and a polar coordinate method of minimizing AS?]

2.4 Trigonometry Precedes Cartesian Coordinate Geometry

KT: (excitedly) Dr. Chand, Dr. Chand, you were right! Absolutely right!!

MC: Right about what? What did you find?

KT: Here is Figure 4 showing a new diagram to minimize AS using polar coordinate. It works like a charm! No, like a magic!!

MC: Go to the board and speak with less hyperbole.

KT: I cannot contain myself. It is a miracle!

MC: Enough already. Explain your new method.

KT: As in the (l, A) case before, from A, drop a perpendicular to l of length p and meeting l at P. However, now let $A(0, 0)$ be the origin of the polar coordinate, or A is the point of rotation starting in the direction AP. Let $\measuredangle PAB = \alpha$ and $\measuredangle PAT = \theta$. Writing in polar coordinates (that is, the distance from A and the angle of rotation), we have $P(p, 0)$, $B(p/\cos\alpha, \alpha)$ and $T(p/\cos\theta, \theta)$. Then the destination is $S(p/[\cos\theta\ \cos(\alpha-\theta)], \alpha)$ on AB. To minimize AS, it suffices to maximize $f(\theta) = \cos\theta\ \cos(\alpha-\theta)$ for $0 < \theta \leq \alpha$.

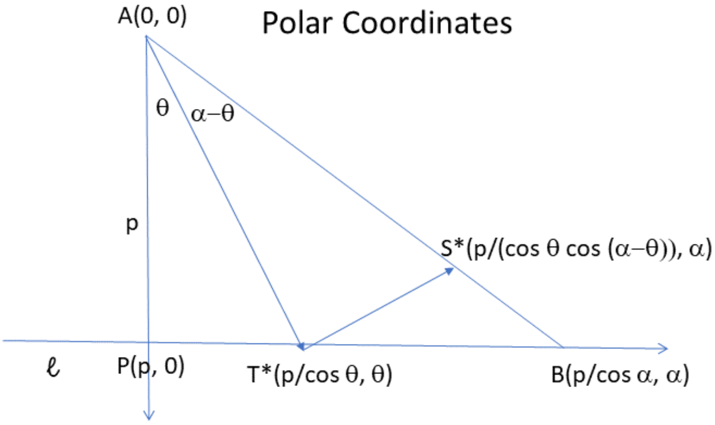


Figure 4: Writing in polar coordinate, let $A(0, 0)$ be the point of rotation starting in the direction AP where $P(p, 0)$. Let $B(p/\cos\alpha, \alpha)$. Which choice of $T(p/\cos\theta, \theta)$ for $0 < \theta \leq \alpha$ minimizes AS?

MC: And how do you propose to do that?

KT: I will solve the first-order condition $f'(\theta) = 0$ to obtain the critical value θ^*, and then check that $f''(\theta^*) < 0$.

MC: Good. You remembered the second-order condition. Way to go!

KT: I will leave to the reader to check the details: Simply put, the first-order condition simplifies to $\tan\theta = \tan(\alpha - \theta)$, whence the solution is $\theta^* = \alpha/2$. Moreover, $f''(\theta^*) = -2\cos(\alpha - 2\theta^*) = -2 < 0$.

MC: Very good. But aren't you missing another solution?

KT: Where do I get the other solution from?

MC: You want the ant to go from A to T on l. Why must it go to T between P and B? Why can't it go to l on the opposite side of B? That is, why restrict $\theta \in [0, \alpha]$? Why not let it be in $(-\pi/2, \pi/2)$?

KT: If I let $\theta < 0$, then the ant will return to line BA much farther away from A, and on the opposite side of B.

MC: Why don't you check it out by picking T on l on either side of P. Then, choose between the two optimal choices of T.

KT: I suppose I can do that. Why didn't I think of that?

MC: It's a rookie mistake: When solving trigonometric functions, you must be more careful. First, draw the graph of $\tan\theta$, then superimpose the graph of $\tan(\alpha - \theta)$. Remember that both functions are periodic with period π. Now find the points of intersection in $(-\pi/2, \pi/2)$.

KT: (draws the graphs on the board) Oh, I see. Indeed, there are two solutions: $\theta^* = \alpha/2$ and $\theta^* = (\alpha - \pi)/2$; they differ by a right angle!

MC: Excellent! What's the final step in the solution?

KT: I know: What's the minimum length of AS? If S is on AB, then AS is $p/\cos^2(\alpha/2)$. On the other hand, if S is on extended BA, then it is $p/\cos^2((\pi - \alpha)/2) = p/\sin^2(\alpha/2)$, which is much larger, I think. I admit I have just learned about this second solution. May I bring a revised diagram of the optimal solutions next Monday?

MC: You may. I know you are tied up the next few days with rehearsal for Tagore's dance drama *Tasher Desh* that you will stage on Founder's Day this weekend. Break a leg.

KT: Thanks. Are you coming to attend the cultural program?

MC: I wouldn't miss it. Whenever in town, I bring my whole family to this event. You will likely meet my twin daughters there.

KT: I look forward to meeting them. Goodbye.

2.5 Finding Solutions without Calculus

On Monday, Kajal visited Dr. Chand, carrying Figure 5 showing the *two* optimal solutions based on polar coordinates.

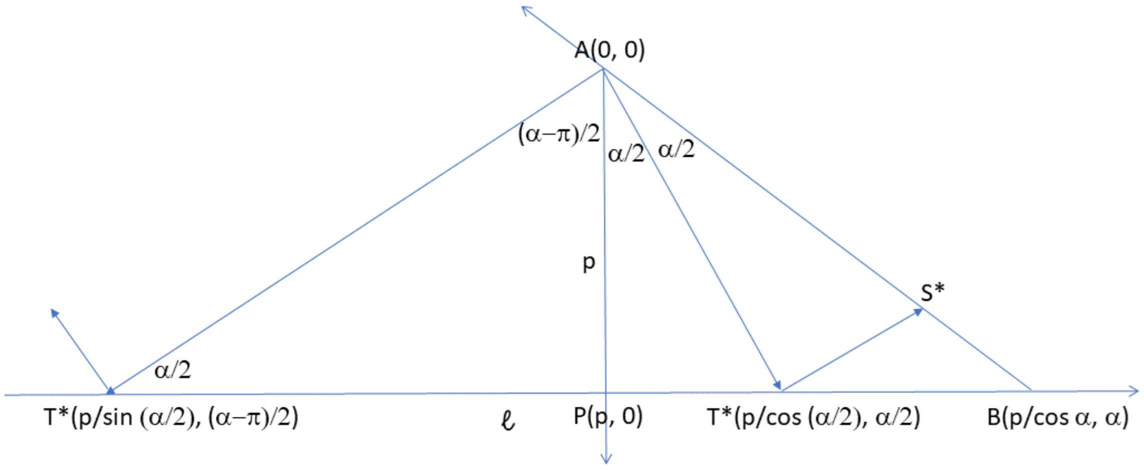


Figure 5: The optimal choices to minimize AS are when the ant heads from A to T making an angle $\alpha/2$ or $(\alpha - \pi)/2$ with AP.

KT: Dr. Chand, did you enjoy the cultural program?

MC: Very much. I am proud of your awesome production: The costumes were gorgeous, the performance spectacular, the dances lively, and the songs were catchy. They are still ringing in my ears. Let us see if your math research is equally impressive.

KT: I am trying. But for your help, I would have given up much earlier.

MC: What new thoughts do you have today?

KT: Please look at Figure 5. In essence, the ant is supposed to travel along the bisectors of the interior and the exterior angles BAP where AP is orthogonal to l. These two optimal directions are themselves orthogonal to each other. Only one of the target points S^* is shown in Figure 5, while the other is too far off and could not be fitted in the diagram. An imaginative reader can identify this missing point.

MC: Imagine that! Tell me, which is more elegant? Discovery based on Cartesian coordinates or polar coordinates?

KT: I must vote in favor of polar coordinates — at least in this instance.

MC: There is an even better reason to use polar coordinates: If you will use a bit of trigonometric identities, you can render the optimization problem almost trivial, requiring no calculus.

KT: No kidding! You are making me curious. Will you tell me how? Or do I have to discover it myself?

MC: You know me better than to ask. Of course, you should discover it yourself. Here's one option: Since you already know the final answers for θ^* are $\alpha/2$ and $(\alpha - \pi)/2$, respectively, you can change the variable into $\xi = \theta - \alpha/2$ and $\eta = \theta - (\alpha - \pi)/2$, and rewrite the objective function in terms of ξ and η, respectively, in these two cases.

KT: So, in stead of maximizing $\cos\theta \, \cos(\alpha - \theta)$, in the first case, we maximize $\cos(\alpha/2 + \xi) \, \cos(\alpha/2 - \xi)$, which after simplification becomes

$$\cos^2(\alpha/2)\cos^2\xi - \sin^2(\alpha/2)\sin^2\xi = \cos^2\xi - \sin^2(\alpha/2).$$

To maximize it, with respect to the new variable ξ, choose $\xi^* = 0$. The second case is similar: Replace α by $(\alpha - \pi)$, and ξ by η.

MC: Two comments and a question: First, if you did not know the answer, you might not know how to change the variable. Still, using the addition formula, you can see that

$$2\cos\theta \, \cos(\alpha - \theta) = \cos(2\theta - \alpha) + \cos(\alpha).$$

Then you could maximize $\cos(2\theta - \alpha)$ by choosing $\theta^* = \alpha/2$. Second, you have justifiably left the second case for the reader to fill in. But don't you need to know the graph of the cosine function?

KT: That would help. However, it is enough to know the definition of $\cos\xi$ as the length of the projection on the horizontal axis of a rotating line segment of unit length starting from the east (or right) direction and rotating counterclockwise by an angle ξ.

MC: Well said. Clearly, then, the projection is maximized at the start, or $\xi = 0$. Moreover, cosine is an even function; that is, $\cos(-\xi) = \cos(\xi)$. What implication does it have on AS?

KT: The objective function AS is also even. So, if you pick T_1, T_2 on PB on opposite sides of T^*, such that $\measuredangle T^*AT_1 = \measuredangle T^*AT_2$, then the ant's destination is the same point S on AB, with $AS > AS^*$.

MC: Also, the circle with diameter AS intersects l at T_1 and T_2.

KT: Thus, points on l come in pairs to determine the same S. Fascinating!

MC: Fascinating indeed. And you learned all these without using calculus!

KT: Are you saying calculus can be forgotten while solving this problem?

MC: Not quite. I am saying someone without the knowledge of calculus can still solve this problem using trigonometry.

KT: Then the problem belongs to trigonometry, known around 1000 CE, long before calculus was invented in the 1680's.

MC: You are right about the chronology. On chronology, I have more to say: Long before trigonometry flourished, another well-founded branch of mathematics existed. Do you know which one?

KT: Geometry. Euclidean, to be precise, already well-known in 300 BCE.

MC: Right you are. Now that you have discovered the solutions to the elephant's optimization problem, can you state a Euclidean geometric theorem to summarize the result? I recommend reading *Niven* [6] for examples of optimization problems solved without calculus.

KT: Can I get a 24 hour recess? I mean, I would like to work in solitude.

MC: The court is adjourned until tomorrow.

Kajal rushed to the library to borrow *Niven* [6], *Heath* [2] and *Katz* [3] to brush up on Euclid's *Elements* and to check the chronology of mathematical concepts.

3 Euclidean Geometry

After another all-nighter (while consuming a box of *BIKAJI Rimjim* — a spicy Indian snack), Kajal regained confidence and was ready to meet Professor Chand.

3.1 Collecting All Ingredients

KT: Good day to you, Dr. Chand. Geometry rules.

MC: Indeed, it epitomizes logical thinking. Although some errors were later corrected, written in 300 BCE, *Elements* remains the ancient tome with the second largest number of editions, after the *Holy Bible*. It is the most successful and most influential textbook ever written.

KT: I did not know that. I should check Wikipedia.

MC: So you must. Most assuredly, I am not making it up; I am only quoting from that same source.

KT: What's the secret to Euclid's success?

MC: It is due primarily to its logical presentation of most of the mathematical knowledge available to Euclid. Euclid's style has not been improved since: Begin with a few definitions, some truisms and some axioms; prove some elementary results; build on them to prove more advanced results; keep building more spectacular results that could not be anticipated. Use pure logic as the glue to errect the edifice.

KT: Is that why you asked me to find a Euclidean geometric proof of the optimal path the ant — nay, the elephant — must follow?

MC: If any result can be proved using Euclidean geometry, why should we settle for any other method? Moreover, geometry is so visual.

KT: I made some progress. I hope you can verify them and help me improve them whereever needed.

MC: I am all ears. Carry on.

KT: Given a point A away from a line l and a point B on l, find a point T on l so that if an ant travels along AT, turns at right angle (withoout passing through l) and continues to travel straight until it meets AB at S, then AS is minimized.

MC: Well stated. How do you find the target point T?

KT: Drop AP perpendicular to l. Let the bisector of $\measuredangle BAP$ meet l at the target T. Since there are two bisectors of $\measuredangle BAP$ — the interior and the exterior

— we have two solutions to T, consequently to S. We will measure AS in these two cases and admit the shorter one.

MC: Good. You narrated the answer perfectly. Tell me, what do you need to fill the chasm between the given and the desired?

KT: I must supply logical steps from the given conditions to the desired conclusion. I think I can do it with your help. Start me off, please.

MC: Suppose that you have found the optimal S. Consider the circle $\mathcal{C}$ with diameter AS. Then, $T \in l$ is on this circle. Do you know why?

KT: (thinking...) Because all angles in a semi-circle are right angles, and $\measuredangle ATS$ is one such right angle.

MC: Right you are. What other point of l is on this circle?

KT: (engrossed in thought for a long time) I can't find any.

MC: Don't give up so soon. Let me check my emails while you search.

KT: (thinking aloud) Where can the other point be? To the left of T or right? It makes no difference, for I can surely relabel T. Oh, I see.

MC: What do you see? Can you show me?

KT: Actually, I mean, I don't see; I can't see; no one can see. A second point $R \in l \cap \mathcal{C}$ is impossible! For otherwise, we can move S to $\tilde{S}$ slightly closer to A so that the circle with diameter $A\tilde{S}$ still intersects l, contradicting the minimality of AS. Hence, T is the only point common to l and $\mathcal{C}$. That is, l is tangent to $\mathcal{C}$ at T.

MC: (applauding) Right, again. What does this tangent property imply?

KT: (more thinking...) The radius through the point of tangency must be orthogonal to the tangent.

MC: When you refer to a tangent, you should also know about the corresponding normal — the line orthogonal to the tangent through the point of tangency. What can you say about the normal through T?

KT: The normal through T passes through the center M of circle $\mathcal{C}$.

MC: Nice. What can you say about $\triangle PAB$ and $\triangle TMB$?

KT: They are both right triangles and they share a common angle at B. They are similar. Hence,

$$PT : TB = AM : MB = TM : MB = PA : AB,$$

implying that TA bisects $\measuredangle BAP$, by the angle bisector theorem.

MC: I think you have all the right ingredients. You must arrange them in a proper sequence to streamline the logical proof. And remember not everyone will know the angle bisector theorem.

KT: I know my mission now. Give me, I plead, just one more day to straighten everything up.

MC: Petition granted. The court will be again in session tomorrow.

[Thus encouraged, Kajal feverishly wrote up the Euclidean geometric results; took a long cold shower; and slept long hours. The next day, Kajal showed the results to Professor Chand, who took a few minutes to read them silently. Then the two math enthusiasts chatted briefly.]

3.2 A Formal Euclidean Proof

First, we state and prove the angle bisector theorem. Then, we present the new result that solves the ant's problem.

Lemma 1. *(Angle Bisector Theorem) If the bisector of $\measuredangle BAC$ meets BC at T, then $BT : TC = BA : AC$. And conversely.*

Proof. Refer to Figure 6. Extend CA to D such that $AD = AB$. Join BD. Then the base angles of $\triangle ABD$ are equal; that is, $\measuredangle ABD = \measuredangle ADB$. But their sum equals the exterior angle of $\triangle ABD$; that is, $\measuredangle BAC = \measuredangle ABD + \measuredangle ADB$. Since AT bisects $\measuredangle BAC$, we have $\measuredangle CAT = \measuredangle TAB$. Hence, $\measuredangle CAT = \measuredangle TAB = \measuredangle ABD = \measuredangle ADB$. Since $\measuredangle CAT = \measuredangle CDB$ are corresponding angles, BD is parallel to TA. Hence, $\triangle CTA$ and $\triangle CBD$ are similar, whence $CT : TB = CA : AD = CA : AB$. This proves the "if" part of the theorem.

The converse is established using proof by contradiction. Suppose, if possible, that $CT : TB = CA : AB$, but AT does not bisect $\measuredangle BAC$. Instead, let the bisector of $\measuredangle BAC$ meet BC at U. Then by the "if" part of this theorem, $CU : UB = CA : AB$, implying that $CT : TB = CU : UB$, or $T = U$. This is a contradiction. □

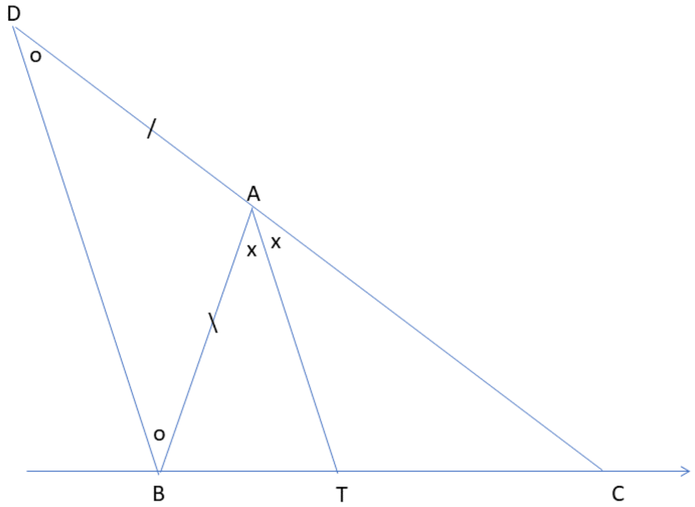


Figure 6: If the bisector of $\measuredangle BAC$ meets BC at T, then $BT : TC = BA : AC$. And conversely.

Theorem 1. *Given a point A outside line l and a point B on l, a point T on l, such that the perpendicular to AT through T intersects AB at S and AS is minimized, must be chosen so that AT bisects $\measuredangle BAP$, where AP is perpendicular to l.*

Construction. Let P be the foot of the perpendicular from A to l. Let the bisectors of interior and exterior $\measuredangle BAP$ meet l at T_1 and T_2, respectively. Let the perpendiculars to AT_1 and AT_2 meet AB at S_1 and extended BA at S_2 (not shown).

Proof. We prove the result for T_1, leaving the proof for T_2 to the reader.

Imagine a family of circles $\mathcal{C}_{AR}$ with diameter AR as R varies over the segment AB. Because A is away from l, if AR is too small (very close to 0), then $\mathcal{C}_{AR}$ does not intersect l. On the other hand, if AR is too large (say, $AR = AB$), then $\mathcal{C}_{AR}$ intersects l at two points. The larger AR is, the farther the two intersection points are. By the intermediate value theorem, there exists a unique point R_1 on AB such that $\mathcal{C}_{AR_1}$ intersects l exactly at one point T_1. That is, l is tangent to $\mathcal{C}_{AR_1}$ at T_1.

Next, we prove that AT_1 bisects $\measuredangle BAP$. Let the center of $\mathcal{C}_{AR_1}$ be denoted by M_1. The three radii M_1A, M_1T_1, M_1R_1 of $\mathcal{C}_{AR_1}$ are equal. Moreover, M_1T_1 and AP, both being perpendicular to l, are parallel. Hence, the similarity of $\triangle PAB$ and $\triangle T_1M_1B$ implies that $PT_1 : T_1B = AM_1 : M_1B$ as well as $T_1M_1 : M_1B = PA : AB$. However, $AM_1 = T_1M_1$ implies that all four ratios mentioned above are equal. In particular, $PT_1 : T_1B = PA : AB$. Hence, by Lemma 1, AT_1 bisects $\measuredangle BAP$. □

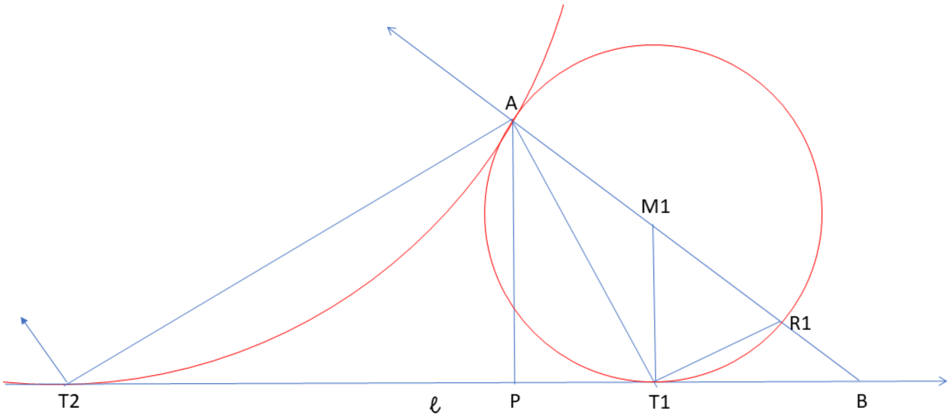


Figure 7: If the circle with diameter AR_1, with R_1 on AB, intersects l at one and only one point T_1, then AT_1 bisects interior $\measuredangle BAP$, where AP is perpendicular to l.

After reading the document, thus spoke Professor Chand, with eyes closed:

MC: Is this all? Or do you have any other evidence?

KT: This is all, your honor. I rest my case and await your judgement.

MC: (after a long pause) Have you read John Keats' *Endymion: A Poetic Romance*? There Book I begins as follows.

```
A thing of beauty is a joy for ever:
Its loveliness increases; it will never
Pass into nothingness; ...
```

Thank you for showing me a thing of beauty. Congratulations!

KT: No, thank *YOU*. It would not have been possible without your help.

MC: Nor without your persistence. I say you have won your royal ride.

KT: So, are you declaring that the elephant won't trample me? Thank God! And thank you for letting me experience this joy of discovery.

MC: You are welcome. If you wish to experience similar joy, read *Stanley* [8]. It explores many problems accessible to undergraduate students. More than solutions, it asks for perspectives, especially ones that expose the heart of what the problem aims to illuminate.

KT: I will surely check out the book. My journey on mathematical problem solving has just begun. Thank you for being my first tour guide.

MC: And I thank you for being more than a student. I am proud to call you my disciple.

KT: Pronaam, Guruji.

Figure 8: Kajal earns a royal mode of transportation.

4 Tying up Some Loose Ends

We concluded our exposition of the discovery of a new Euclidean geometric theorem. However, some issues raised during the mental journey must be addressed.

What are some features in Figure 3 that Kajal (and we) could utilize to discover the optimal T without requiring polar coordinates?

In Figure 3, in the first panel (l, A) and the third panel (l, B), we could note that $PT^* : T^*B = PA : AB$. Then, in view of Lemma 1, we could conclude AT^* bisects $\measuredangle BAP$. In the second panel (BA, B) and the fourth panel (AB, A), congruency of $\triangle APT^*$ and $\triangle AQT^*$ (by the right angle, hypotenuse, side criterion), implies $\measuredangle PAT^* = \measuredangle QAT^*$, or AT^* bisects $\measuredangle QAP = \measuredangle BAP$.

We urge diligent readers to fill in a few things Kajal left undisclosed:

1. In case (BA, B), prove that AS is proportional to BA with proportionality constant less than 1.

2. Derive the expression of the length of AS in cases (l, B) and (AB, A).

3. Derive the first-order condition to minimize the length of AS when using polar coordinates.

4. Draw the graph of $\tan\theta$, and superimpose the graph of $\tan(\alpha - \theta)$. Then find the points of intersection for $\theta \in (-\pi/2, \pi/2)$.

5. Maximize $\cos((\alpha - \pi)/2 - \eta)\cos((\alpha - \pi)/2 + \eta)$ with respect to η.

6. Prove Theorem 3.2 using T_2 in place of T_1.

As a parting gift to the diligent reader we pose this extended problem:

Generalized Problem: Point A is away from a line l, and point B belongs to l. An ant will start its journey from A and walk straight to a point T on l, then it will turn **by an angle $\boldsymbol{\beta \in (\pi/2, \pi)}$** (without passing through l) and again walk straight until it will meet line AB at S. The ant wants to reach this point S as close to A as possible. Advice the ant which point T on l it must initially go toward.

[Answer: Drop AP perpendicular to l. Let $\alpha = \measuredangle PAB$ and $\theta^* = (\alpha - \beta + \pi/2)/2$. Obtain T^* on l such that $\measuredangle PAT^* = \theta^*$ or $\theta^* - \pi/2$.]

Acknowledgment

This paper I offer as a guru-dakshina to Professor Bikas Kumar Sinha, who inspired me to solve puzzles and problems at many different levels. The idea for writing a mathematics paper in a dialogue format came from my involvement in a puppet show for children at First Mennonite Church of Indianapolis. I thank an anonymous referee for suggesting an alternative trigonometric simplification to liberate the optimization problem from calculus, and for suggesting a reference. I gratefully acknowledge the Editor's guidance. I hope this paper will inspire many young scholars to take up mathematical problem-solving as a hobby and/or profession.

Bibliography

[1] Jay Abramson, *Algebra and Trigonometry*, 2e. Arizona State University, 2021.

[2] Thomas L. Heath, *The Thirteen Books of Euclid's Elements* with Introduction and Commentary. Cambridge University Press, 1908.

[3] Victor J. Katz, *A History of Mathematics: An Introduction*, 2e. Reading, MA: Addison Wesley Longman, 1998.

[4] John Keats, A Thing of Beauty, from *Endymion: A Poetic Romance*. London: Taylor and Hessey,1818. https://www.poetryfoundation.org/poems/44469/endymion-56d2239287ca5

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

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

[7] George Polya, *How to Solve It: A New Aspect of Mathematical Method.* Princeton, NJ: Princeton University Press, 1945.

[8] Richard Stanley, *Conversational Problem Solving*, Washington, DC: American Mathematical Society, 2020.

[9] James Stewart, *Calculus*, 8e. Boston, MA: Cengage Learning, 2016.

[10] Earl W. Swokowski, *Calculus with Analytic Geometry.* Boston, MA: Prindle, Weber & Schmidt, 1983.

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

7 Emmy Noether: Mother of Modern Algebra

Amartya Kumar Dutta and Neena Gupta

Indian Statistical Institute
Kolkata, India.
Email: amartrya.28@gmail.com
Email: rnanina@gmail.com

Amalie Emmy Noether (1882–1935) is one of the greatest mathematicians of the twentieth century. She was largely influential in giving the subject “Abstract Algebra”, one of the pillars of modern mathematics, its present shape. The ease with which the subject can now be introduced even at the undergraduate level has been made possible due to the formulation provided by the great visionary. Even the present axiomatic definition of a “Ring”, which has been accepted worldwide, is due to her. Besides, Emmy Noether is a founder of the subject “Commutative Algebra” which, apart from being a beautiful and deep branch of mathematics, provides the foundation for modern Algebraic Geometry and Algebraic Number Theory. She also made important contributions in areas of mathematics like Algebraic Invariant Theory, Inverse Galois Theory, Topology, Representation Theory, Non-Commutative Algebra and Number Theory. Again, Emmy Noether made stunning discoveries in Theoretical Physics which reveal a deep connection between the “symmetries” of Group Theory in Abstract Algebra and the all-important “conservation laws” in Physics. Her theorems revolutionized the way physicists analyze the universe.

Unfortunately, the name of this brilliant mathematician is not so well-known. Being a woman and a Jew, Emmy Noether had to struggle all her life in her pursuit of higher Mathematics. And yet she made her ground-breaking seminal contributions in diverse mathematical fields. We give below a brief account of her inspiring life — the story of a glorious triumph of the human spirit.

Emmy Noether was born on 23rd March 1882 in Erlangen, Germany, to a Jewish family. Her father Max Noether (1844–1921) is regarded as one of the finest mathematicians of the nineteenth century. One of her brothers Alfred Noether (1883-1928) was a doctorate in chemistry, another brother Fritz Noether (1884–1941) was a reputed applied mathematician.

The rules of the German society of those days did not allow women to teach in a University. Proficient in English and French, the initial aim of Emmy Noether, after completing high school, was to become a language teacher. In 1900, she became a certified teacher of English and French for girls' schools, obtaining very good grades in the qualifying examinations. However, her love for mathematics proved to be too strong. Although she was aware that the prevailing rules would never allow her to become a faculty member of a University like her father, she could not stay away from Mathematics. While her interest in languages continued, she also started attending mathematics lectures unofficially in the University of Erlangen, where her illustrious father was a faculty member. Due to the rules, she had to acquire the permission of the respective teachers. She was one of only two women among the thousand students studying in Erlangen. Overcoming all such obstacles, Emmy Noether graduated in July 1903. In the winter semester of 1903-04, she attended lectures by astronomer Karl Schwarzschild and mathematicians David Hilbert, Felix Klein, Otto Blumenthal and Hermann Minkowski at the University of Göttingen, the world-renowned centre for mathematical research.

Fortunately, in 1904, rules were changed in Erlangen and women were allowed to enroll in the PhD programme. Emmy Noether returned to Erlangen, did her doctorate under the guidance of Paul Gordan (1837–1912), a friend of Max Noether. Gordan was then regarded as the "King of Invariant Theory" and Emmy Noether's thesis was on algebraic invariants of biquadratic forms. The celebrated "Hilbert Basis Theorem" (1888), now a basic result in Algebra, had given an existential (i.e., non-constructive) result on the finiteness of invariants; Emmy Noether's thesis followed the constructive approach of Gordan — listing explicitly worked out invariants. Her thesis was well-received. After her PhD in 1907, Emmy Noether remained in Erlangen till 1915, doing research and teaching without any pay, and helping her ailing father.

In 1911, when Gordan retired, Ernst Fischer (1875–1954), famous for the Riesz-Fischer Theorem in Lebesgue integration, succeeded Gordan to the chair of mathematics. Emmy Noether and Fischer had lively discussions on mathematics. Fischer introduced Emmy Noether to the work of Hilbert. This had momentous consequences as Emmy Noether began imbibing the abstract algebraic approach of Hilbert. During 1913–1916, Emmy Noether published several papers in which she extended Hilbert's methods and applied them on fields of rational functions and invariants of finite groups. Thus began her involvement with "Abstract Algebra", the field of mathematics which she would revolutionize.

In 1915, Albert Einstein had published his General Theory of Relativity which describes the phenomenon of gravity as the consequence of the curvature of space and time caused by massive bodies. Hilbert, who was working on closely related ideas, observed a paradox about the principle of conservation of energy that arose out of general relativity. After discussing with Klein, Hilbert invited Emmy Noether to Göttingen. He had realized that Emmy Noether, with her expertise in algebraic invariant theory, could help develop a better understanding of general relativity. Hilbert had to overcome vociferous opposition from historians and philologists to the recruitment of a woman.

Emmy Noether arrived in Göttingen in 1915 and not only resolved the paradox but gifted to the world of modern theoretical physics two fundamental theorems known as "Noether's Theorems". Her first theorem established the connection between symmetries and conservation laws and the second theorem provided the "generally covariant" theories for the apparently strange type of conservation laws. The contributions were typical of Emmy Noether (that would be seen prominently in her subsequent works on Abstract Algebra): placing a specific concept in a broad mathematical framework where its features can be understood better. The American physicists Leon M. Lederman and Christopher T. Hill have remarked in their book "Symmetry and the Beautiful Universe" that Noether's theorem connecting conservation laws with symmetries is "certainly one of the most important mathematical theorems ever proved in guiding the development of modern physics". The connection explains properties of the universe that had earlier appeared arbitrary. Noether's theorem became a cornerstone for particle physics. For instance, the prediction of the existence of the Higgs-Boson particle was an outcome of the connection with symmetries.

We mention here that although Hilbert managed to bring Emmy Noether to Göttingen in 1915, she had no official position during her initial years and worked voluntarily, without any salary, for eight years. She did her teaching (that too in Hilbert's name) and pathbreaking research out of sheer passion for mathematics. Although she began receiving a salary from 1923, she never became a full-fledged Professor in Göttingen. In his memorial address on Emmy Noether, Hermann Weyl of Göttingen rued, "I was ashamed to occupy such a preferred position beside her whom I knew to be my superior as a mathematician in many respects."

Emmy Noether is best remembered for her contributions to Abstract Algebra, which began from 1919. Her abstract and conceptual approach led to several principles unifying topology, logic, geometry, algebra and linear algebra. As Nathan Jacobson wrote in his introduction to Noether's Collected Papers, "The development of abstract algebra, which is one of the most distinctive innovations of twentieth century mathematics, is largely due to her — in published papers, in lectures, and in personal influence on her contemporaries." Her work on the theory of ideals in commutative rings helped develop Ring Theory into the major mathematical topic that Commutative Algebra is today, with wide-ranging applications. She established several beautiful and important results; her proofs are elegant, short and conceptual. With her deep insight, she realized the paramount importance of chain conditions on ideals of rings, especially the "the ascending chain condition" (ACC) and revealed the general properties of all rings satisfying the condition, which included the rings which were being studied in isolation earlier, like polynomial rings in finitely many variables over a field, or rings of algebraic integers. As a tribute to her work, the term “Noetherian” was coined by Chevalley in 1943 for rings satisfying the ACC on ideals, a condition satisfied by an important class of rings arising in Algebraic Geometry. Her 1921 paper on ideals has been called “revolutionary” by the noted algebraist Irving Kaplansky. A major result in this paper was her theory of primary decomposition for all Noetherian rings, a deep generalization of the Fundamental Theorem of Arithmetic that every integer can be decomposed as product of prime integers. In 1905, the mathematician Emanuel Lasker, who was also the World Chess Champion for 27 years, had established the primary decomposition property for polynomial rings over a field. Noether’s approach not only generalized, but also enormously simplified Lasker’s theory — it is now accessible to even our Master’s students. Another paper of Emmy Noether gives several characterizations of rings called Dedekind domains in which unique factorization of ideals into prime ideals holds. The paper also contains the famous “isomorphism theorems” which are fundamental in Algebra and basic results on modules satisfying chain conditions.

From 1927, Emmy Noether achieved several landmarks in other areas of algebra, like the Skolem-Noether Theorem for central simple algebras and the Brauer-Noether Theorem for central division algebras. Along with Emil Artin, Richard Brauer and Helmut Hasse, she had founded the theory of central simple algebras. She gave the first general representation theory of groups and algebras, unifying the earlier scattered studies on group representations and associative algebras by placing them under the framework of the theory of Noetherian Modules.

Generous regarding sharing of her mathematical insights, Emmy Noether has been acknowledged for the great idea of studying topology algebraically, including the fundamental ideas that led to the development of algebraic topology, especially the idea of homology groups, the details of which were worked out by other mathematicians. She made several other fundamental contributions to mathematics, not all of which can be listed in a short article.

In 1932, Emmy Noether was awarded the Ackermann-Teubner Memorial Prize in Mathematics and became the the first woman plenary speaker in the International Congress of Mathematicians (ICM), since its inaugural in 1900. (It was only in 1990 that the ICM would witness another woman plenary speaker.) However, she was forced to leave Germany next year when the Nazis took control over the German Government. She then moved to U.S.A. to work as a guest lecturer at Bryn Mawr College in Pennsylvania, where she remained till her death from a post-operative infection in 1935.

Emmy Noether had accepted the decision of her expulsion from Göttingen calmly, providing support to others during the difficult days. Hermann Weyl later wrote that “Emmy Noether — her courage, her frankness, her unconcern about her own fate, her conciliatory spirit — was in the midst of all the hatred and meanness, despair and sorrow surrounding us, a moral solace.”

We conclude our tribute by recalling B.L. van der Waerden's observation that Emmy Noether's mathematical originality was "absolute beyond comparison" and Hermann Weyl's remark that she "changed the face of algebra by her work". A crater on the far side of the Moon is named after her.

8 The Nine-point Circle of a Triangle – Part 2

Shailesh Shirali

The Valley School KFI,
Thatguni Post, Kanakapura Road, Bengaluru – 560082

Email: shailesh.shirali@gmail.com

In Part 1 of this article we introduced the Euler line and the nine-point circle of a triangle. To recap: the Euler line passes through the circumcentre, the centroid, and the orthocentre of the triangle, and the nine-point circle passes through the midpoints of the three sides, the feet of the three altitudes, and three other significant points of the triangle. We briefly noted the history of this discovery, and gave a neat proof of the theorem using pure geometry and another (very compact) proof using vector algebra.

Now, in Part 2 of the article, we shall describe a most astonishing tangency property of the nine-point circle. It was first announced and proved by the German mathematician Karl Wilhelm Feuerbach in 1822. The property is not difficult to discover; indeed, one is likely to spot it the moment one draws an accurate picture of the nine-point circle. But proving it is another matter altogether. As earlier, we shall give two proofs of the result; the first is computational, but it uses only standard results from circle geometry; the second proof uses vector algebra.

Theorem 1 (Feuerbach)**.** *The nine-point circle of a triangle is internally tangent to the incircle of the triangle and externally tangent to the three ex-circles of the triangle.*

Figures 1 and 2 depict the statement of the theorem.

As the reader will readily appreciate, it is a challenge to display the full picture on a printed page. In Figure 2, we have shown only one of the three ex-circles.

What strategy can we use to prove that two circles are internally or externally tangent to each other? The most obvious approach, surely, would be through a consideration of

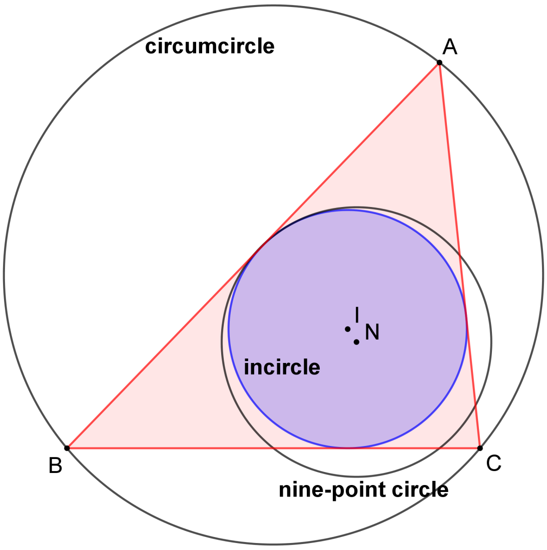


Figure 1: The nine-point circle is internally tangent to the incircle

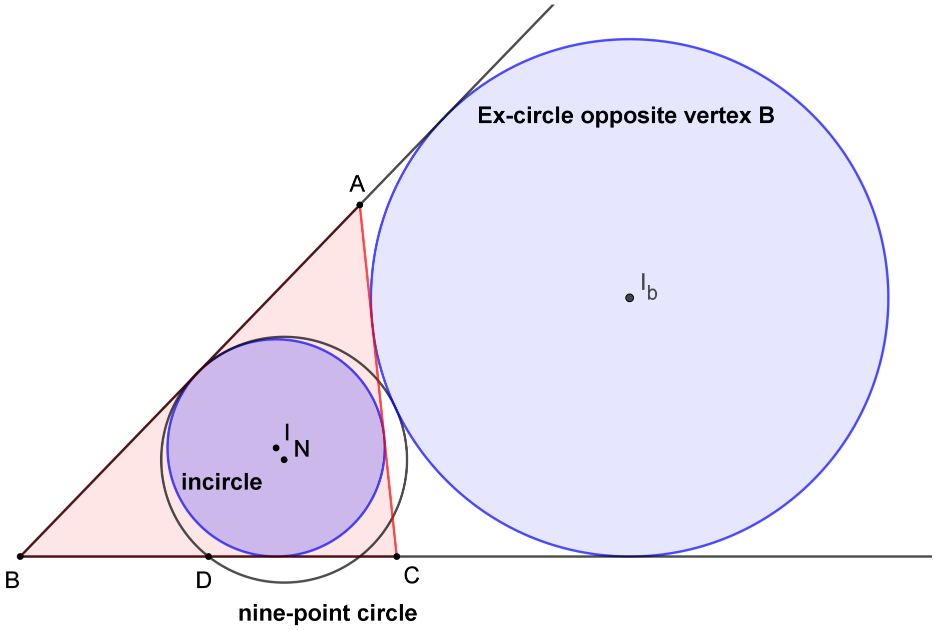


Figure 2: The nine-point circle is externally tangent to the three ex-circles

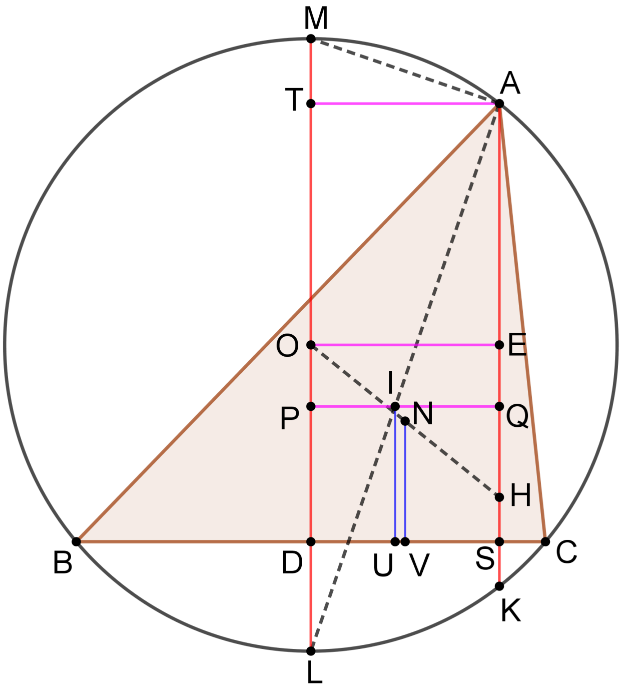
distance: we would have to show that the distance between the centres of the two circles is either equal to the difference between their radii or equal to the sum of their radii. We shall proceed to do just this.

Remark

This is not the only approach possible. Another approach uses *inversion*. This is a non-linear geometrical transformation that allows us to map circles to circles or to straight lines. Showing that some two circles touch each other can then reduce to showing that two straight lines are parallel to each other — which certainly sounds a simpler task! We will describe this proof in a later article.

Proof using pure geometry

The proof, which has been adapted from [6], requires a large number of auxiliary lines and points to be constructed, as shown in Figure 3.



- D: Midpoint of BC
- I: Incentre
- O: Circumcentre
- H: Orthocentre
- LDM: Diameter
- $AHSK \perp BC$
- $N = (O + H)/2$
- ONH: Euler line
- $IU \perp BC$
- $NV \perp BC$
- $PIQ \perp AHSK$
- $AT \perp ML$

Figure 3: K. J. Sanjana's proof, from [6]. We must show that $2\,IN = R - 2r$

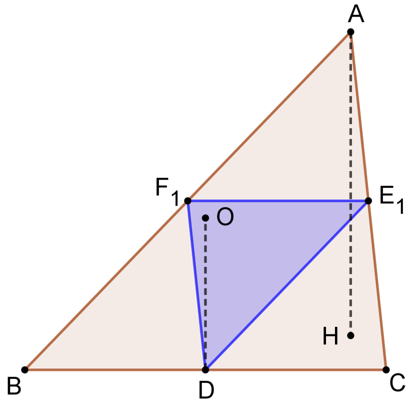


Figure 4: To show that $AH = 2\,OD$

We need to compute the distance IN. Let R and r be (respectively) the radii of the circumcircle and the incircle of the triangle. The radius of the nine-point circle is then $R/2$. So the first part of the theorem will be established if we can show that $IN = R/2 - r$.

To show this we establish a number of subsidiary results.

(a) *Claim.* In Figure 3, $HS = SK$, i.e., S is the midpoint of HK.

For, $\measuredangle KBC = \measuredangle KAC = 90° - \measuredangle C$, and $\measuredangle HBC = 90° - \measuredangle C$. Since $BH = BK$ and $HK \perp BS$, the assertion follows.

(b) *Claim.* In Figure 3, $AH = 2\,OD$.

To see why, study Figure 4. $\triangle DE_1F_1$ (shaded blue) inscribed in $\triangle ABC$ has for its vertices the midpoints of the sides of ABC; it is the *medial triangle* of ABC. Hence the triangles are similar to one another, and the medial triangle has half the scale of the original. Also, the circumcenter O of $\triangle ABC$ is the orthocentre of $\triangle DE_1F_1$. It follows from this that $AH = 2\,DO$.

Remark

Another approach is to use vectors. With O as the centre of the coordinate system, let $\mathbf{OA} = \mathbf{a}$, $\mathbf{OB} = \mathbf{b}$, $\mathbf{OC} = \mathbf{c}$; then $\mathbf{OD} = \frac{1}{2}(\mathbf{b} + \mathbf{c})$. Also, $\mathbf{OH} = \mathbf{a} + \mathbf{b} + \mathbf{c}$; we had shown this in Part 1 of the article. Hence $\mathbf{AH} = \mathbf{b} + \mathbf{c}$. The result follows.

(c) *Claim.* In Figure 3, $AI \cdot IL = 2Rr$.

This is a significant result in itself. Like so many other results in mathematics, it was first proved by Euler. As it is a major result, we shall delay the proof till later.

(d) *Claim.* In Figure 3, $PI \cdot IQ = r \cdot MT$. To prove this we use triangle similarity. We have:

$$\frac{PI}{IL} = \frac{MT}{AM} \quad \text{and} \quad \frac{IQ}{AI} = \frac{AM}{ML}.$$

Hence

$$\frac{PI}{AI} \cdot \frac{IQ}{IL} = \frac{MT}{ML},$$

giving

$$\frac{PI \cdot IQ}{2Rr} = \frac{MT}{2R}.$$

It follows that $PI \cdot IQ = r \cdot MT$.

We are now in a position to compute the length of IN. For this, we consider the projections of IN on LM and BC.

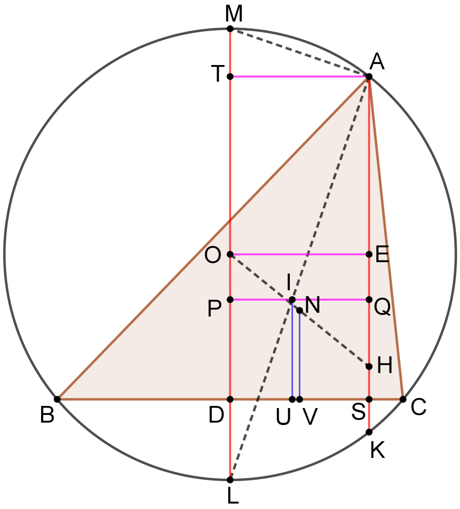


Figure 5: Computing the length of IN (we have repeated the figure for convenience).

Consider first the projection on LM:

$$\begin{aligned} \text{Projection of } IN \text{ on } LM &= IU - NV \\ &= r - \frac{1}{2}(OD + HS) \\ &= r - \frac{1}{4}(AH + HK) \\ &= r - \frac{1}{4}AK = r - \frac{1}{2}AE. \end{aligned} \tag{1}$$

Hence

$$\begin{aligned}(\text{Projection of } IN \text{ on } LM)^2 &= r^2 - r\cdot AE + \frac{1}{4}AE^2\\ &= r^2 - r\cdot TO + \frac{1}{4}AE^2. \qquad (2)\end{aligned}$$

Similarly we have for the projection of IN on BC:

$$\begin{aligned}(\text{Projection of } IN \text{ on } BC)^2 &= UV^2 = (DV - DU)^2\\ &= DV^2 - DU\cdot(2DV - DU)\\ &= DV^2 - DU\cdot US\\ &= \frac{1}{4}DS^2 - PI\cdot IQ\\ &= \frac{1}{4}OE^2 - r\cdot MT. \qquad (3)\end{aligned}$$

Add the above two; we get:

$$\begin{aligned}IN^2 &= \left(r^2 - r\cdot TO + \frac{1}{4}AE^2\right) + \left(\frac{1}{4}OE^2 - r\cdot MT\right)\\ &= \frac{1}{4}\left(AE^2 + OE^2\right) + r^2 - r\cdot(TO + MT)\\ &= \frac{1}{4}OA^2 + r^2 - r\cdot OM = \frac{1}{4}R^2 + r^2 - r\cdot R\\ &= \left(\frac{1}{2}R - r\right)^2, \quad \therefore\ IN = \left|\frac{1}{2}R - r\right|. \qquad (4)\end{aligned}$$

Thus the theorem follows for the incircle.

We will not elaborate on the proof of the corresponding statement for the ex-circles.

Proof of formula (c)

It remains to prove relation (c) quoted earlier, that $AI \cdot IL = 2Rr$. Note that this is essentially a statement about the "power of the point I with respect to the circumcircle."

We refer to Figure 5. Draw $IE \perp AB$; then E is the point of contact of the incircle with

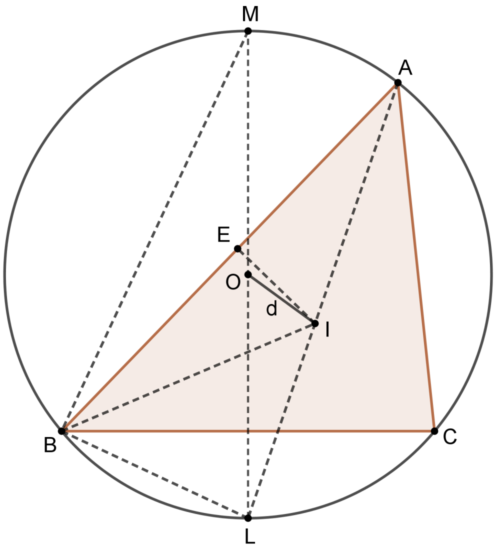


Figure 6: Proving Euler's formula for the distance d between I and O

side AB, and $IE = r$. Join BM. We now have:

$$\begin{aligned} \measuredangle BIL &= \frac{1}{2}\measuredangle A + \frac{1}{2}\measuredangle B \\ \text{and} \quad \measuredangle LBI &= \measuredangle LBC + \frac{1}{2}\measuredangle B, \\ &= \frac{1}{2}\measuredangle A + \frac{1}{2}\measuredangle B \\ &= \measuredangle BIL, \end{aligned}$$

hence $LI = LB$, and $AI \cdot IL = AI \cdot BL$.

Next, we have:

$$\frac{BL}{LM} = \sin\frac{A}{2}, \quad \therefore \ BL = 2R \cdot \sin\frac{A}{2},$$

and,

$$\frac{IE}{AI} = \sin\frac{A}{2}, \quad \therefore \ AI = \frac{r}{\sin A/2}.$$

It follows that $AI \cdot BL = 2Rr$.

Remark

It is well known that the power of I is $R^2 - d^2$, where $d = OI$. So what we have proved is $R^2 - d^2 = 2Rr$, i.e., $d^2 = R(R - 2r)$. This is the formula proved by Euler in 1765 (but apparently it had been proved earlier, in 1746, by British mathematician and surveyor William Chapple). Note an interesting inequality that is implied by this relation: $R \geq 2r$. □

A proof using vector algebra

We conclude with a proof of the tangency property using vectors. The proof is from [7].

Let the centre O of the circumcircle be the origin of the coordinate system. Let the position vectors of the vertices A, B, C be $\mathbf{a}, \mathbf{b}, \mathbf{c}$, respectively; then $|\mathbf{a}| = |\mathbf{b}| = |\mathbf{c}| =$ the radius R of the circumcircle. We know that the position vector $\mathbf{n}$ of the nine-point centre N is

$$\mathbf{n} = \frac{1}{2}(\mathbf{a} + \mathbf{b} + \mathbf{c}). \tag{5}$$

We also know that the position vector $\mathbf{i}$ of the incentre I is given by

$$\mathbf{i} = \frac{\alpha\mathbf{a} + \beta\mathbf{b} + \gamma\mathbf{c}}{\alpha + \beta + \gamma}.$$

This is a consequence of the well-known (and easily-proved) property that an angle bisector divides the opposite side in the ratio of the adjacent sides.

We need the following subsidiary result which enables us to compute distances between points.

(e) Given a triangle ABC with circumcentre O, let the coordinate system be set up as described above. Let X, Y be any two points in the plane of the triangle. Let the position vector $\mathbf{x}$ of X be expressed in terms of $\mathbf{a}, \mathbf{b}, \mathbf{c}$ as $\mathbf{x} = \alpha\mathbf{a} + \beta\mathbf{b} + \gamma\mathbf{c}$ where $\alpha + \beta + \gamma = 1$. Let the lengths of the sides of triangle ABC be a_1, b_1, c_1, respectively. (We have not named them a, b, c as these refer to the lengths of $\mathbf{a}, \mathbf{b}, \mathbf{c}$, respectively; in fact, $a_1 = |\mathbf{b} - \mathbf{c}|$, etc.) Then, we claim that:

$$XY^2 = \alpha\, AY^2 + \beta\, BY^2 + \gamma\, CY^2 - \left(\beta\gamma\, a_1^2 + \gamma\alpha\, b_1^2 + \alpha\beta\, c_1^2\right). \tag{6}$$

To prove this we note:

$$\begin{aligned} XY^2 &= |\mathbf{y} - \mathbf{x}|^2 = |\mathbf{y} - (\alpha\mathbf{a} + \beta\mathbf{b} + \gamma\mathbf{c})|^2 \\ &= |\alpha(\mathbf{y} - \mathbf{a}) + \beta(\mathbf{y} - \mathbf{b}) + \gamma(\mathbf{y} - \mathbf{c})|^2 \\ &= \alpha^2 AY^2 + \beta^2 BY^2 + \gamma^2 CY^2 \\ &\qquad + 2\alpha\beta(\mathbf{y} - \mathbf{a}) \cdot (\mathbf{y} - \mathbf{b}) + 2\alpha\gamma(\mathbf{y} - \mathbf{a}) \cdot (\mathbf{y} - \mathbf{c}) + 2\beta\gamma(\mathbf{y} - \mathbf{b}) \cdot (\mathbf{y} - \mathbf{c}). \end{aligned} \tag{7}$$

Next, we have:

$$\begin{aligned} c_1^2 &= |\mathbf{b}-\mathbf{a}|^2 \\ &= |(\mathbf{y}-\mathbf{a})-(\mathbf{y}-\mathbf{b})|^2 \\ &= AY^2+BY^2-2(\mathbf{y}-\mathbf{a})\cdot(\mathbf{y}-\mathbf{b}). \end{aligned} \tag{8}$$

Hence:

$$2\alpha\beta(\mathbf{y}-\mathbf{a})\cdot(\mathbf{y}-\mathbf{b}) = \alpha\beta(AY^2+BY^2-c_1^2).$$

There are similar relations for a_1^2 and b_1^2. Substitute these three expressions into (7); we find that the total coefficient of AY^2 becomes

$$\alpha^2+\alpha\beta+\alpha\gamma = \alpha(\alpha+\beta+\gamma) = \alpha.$$

After simplification we get (6). □

With this result in our possession, we progress rapidly.

(f) *Claim.* $2\mathbf{a}\cdot\mathbf{b} = 2R^2 - c_1^2.$

For, since $2\mathbf{n} = \mathbf{a}+\mathbf{b}+\mathbf{c}$,

$$c_1^2 = |\mathbf{a}-\mathbf{b}|^2 = |\mathbf{a}|^2+|\mathbf{b}|^2-2\mathbf{a}\cdot\mathbf{b} = 2R^2-2\mathbf{a}\cdot\mathbf{b}.$$

(g) *Claim.* $4AN^2 = R^2 - a_1^2 + b_1^2 + c_1^2$ (with similar expressions for BN^2 and CN^2).

For,

$$\begin{aligned} 4AN^2 &= |2\mathbf{a}-2\mathbf{n}|^2 = |\mathbf{a}-\mathbf{b}-\mathbf{c}|^2 \\ &= 3R^2-2\mathbf{a}\cdot\mathbf{b}-2\mathbf{a}\cdot\mathbf{c}+2\mathbf{b}\cdot\mathbf{c} \\ &= 3R^2-2R^2+c_1^2-2R^2+b_1^2+2R^2-a_1^2 \\ &= R^2-a_1^2+b_1^2+c_1^2. \end{aligned}$$

(h) *Claim.* $OI^2 = R^2 - 2Rr.$ This is Euler's relation, which we have already proved, but we shall show how it follows from the results established in this section. We use result (e) and the fact that

$$\mathbf{i} = \frac{a_1}{2s}\mathbf{a}+\frac{b_1}{2s}\mathbf{b}+\frac{c_1}{2s}\mathbf{c},$$

where $2s = a_1+b_1+c_1$ is the perimeter of the triangle.

Let $X = I$ and $Y = O$. Then result (e) reduces to

$$\begin{aligned} OI^2 &= \frac{a_1}{2s}R^2 + \frac{b_1}{2s}R^2 + \frac{c_1}{2s}R^2 - \left(\frac{b_1c_1}{(2s)^2a_1^2} + \frac{c_1a_1}{(2s)^2b_1^2} + \frac{a_1b_1}{(2s)^2c_1^2}\right) \\ &= R^2 - \frac{a_1b_1c_1}{2s}. \end{aligned}$$

Now we make use of two formulas for the area Δ of triangle ABC:

$$\Delta = \frac{a_1b_1c_1}{4R} = rs.$$

From this we get:

$$\frac{a_1b_1c_1}{2s} = \frac{a_1b_1c_1}{4R} \cdot \frac{2R}{s} = \Delta \cdot \frac{2R}{s} = rs \cdot \frac{2R}{s} = 2Rr.$$

Hence $OI^2 = R^2 - 2Rr = R(R - 2r)$. □

We now use result (e) again, with $X = I$ and $Y = N$, and thereby obtain an expression for IN. We get:

$$IN^2 = \sum_{\text{cyclic}} \frac{a_1}{2s}\frac{R^2 - a_1^2 + b_1^2 + c_1^2}{4} - \sum_{\text{cyclic}} \frac{b_1}{2s}\frac{c_1}{2s}a_1^2. \tag{9}$$

The second summation simplifies easily:

$$\begin{aligned} \sum_{\text{cyclic}} \frac{b_1}{2s}\frac{c_1}{2s}a_1^2 &= \frac{a_1b_1c_1}{4s^2}\sum_{\text{cyclic}} a_1 \\ &= \frac{a_1b_1c_1}{2s} = \frac{a_1b_1c_1}{4R} \cdot \frac{4R}{2s} \\ &= \frac{2\Delta R}{s} = 2Rr, \quad \text{since } \Delta = rs. \end{aligned}$$

We tackle the first summation now. Observe:

$$\begin{aligned} &\sum_{\text{cyclic}} a_1\left(R^2 - a_1^2 + b_1^2 + c_1^2\right) \\ &= R^2\sum_{\text{cyclic}} a_1 + \sum_{\text{cyclic}} \left(-a_1^3 + a_1b_1^2 + a_1c_1^2\right) \\ &= R^2 \cdot 2s + \left(-a_1^3 - b_1^3 - c_1^3 + a_1^2b_1 + a_1^2c_1 + a_1b_1^2 + a_1c_1^2 + b_1^2c_1 + b_1c_1^2\right). \end{aligned}$$

The expression $-a_1^3 - b_1^3 - c_1^3 + a_1^2b_1 + a_1^2c_1 + a_1b_1^2 + a_1c_1^2 + b_1^2c_1 + b_1c_1^2$ does not factor. Playing around with the expression, we find that by subtracting $2a_1b_1c_1$, we do obtain a factorization:

$$\begin{aligned} &- a_1^3 - b_1^3 - c_1^3 + a_1^2b_1 + a_1^2c_1 + a_1b_1^2 + a_1c_1^2 + b_1^2c_1 + b_1c_1^2 - 2a_1b_1c_1 \\ &= (b_1 + c_1 - a_1)(c_1 + a_1 - b_1)(a_1 + b_1 - c_1) \\ &= 8(s - a_1)(s - b_1)(s - c_1) = \frac{8\Delta^2}{s}. \end{aligned}$$

Hence we obtain from (9), recalling that $(a_1b_1c_1)/(2s) = 2Rr$, as shown above:

$$\begin{aligned} IN^2 &= \frac{R^2}{4} + \frac{\Delta^2}{s^2} - Rr \\ &= \frac{R^2}{4} + r^2 + \frac{a_1b_1c_1}{4s} - 2Rr \\ &= \frac{R^2}{4} + r^2 + Rr \\ &= \left(\frac{R}{2} - r\right)^2, \end{aligned}$$

and we get $IN = \left|\frac{1}{2}R - r\right|$, as earlier. □

Remark

The idea of subtracting and adding $2a_1b_1c_1$ may seem quite unmotivated and rather like cheating, but some educated guessing might just suggest it!

Closing remark

In Part 3 of the article we shall give a proof of the theorem using inversion, and in Part 4 we shall describe a generalization of this theorem to general conic sections.

Bibliography

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

[2] Wikipedia, "Feuerbach point." From https://en.wikipedia.org/w/index.php?title=Feuerbach_point&oldid=1132185936

[3] Weisstein, Eric W. "Feuerbach's Theorem." From *MathWorld*–A Wolfram Web Resource. https://mathworld.wolfram.com/FeuerbachsTheorem.html

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

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

[6] Sanjana, K J. "An Elementary Proof of Feuerbach's Theorem." https://tinyurl.com/244eba58

[7] Scheer, Michael J. G. “A Simple Vector Proof of Feuerbach’s Theorem.” *Forum Geometricorum*, Volume 11 (2011) 205–210. From https://forumgeom.fau.edu/FG2011volume11/FG201121.pdf OR https://arxiv.org/pdf/1107.1152

9 We are what we think we are!

Anisa Chorwadwala

IISER Pune
Dr. Homi Bhabha Road
Pune 411008
Email: anisa@iiserpune.ac.in

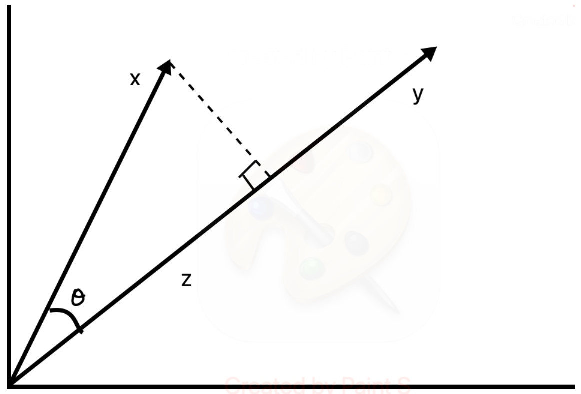
Recall that, for two vectors $x = (x_1, x_2, \dots, x_n)$ and $y = (y_1, y_2, \dots, y_n)$ in $\mathbb{R}^n$, their dot product $x \cdot y$ is defined as $x_1y_1 + x_2y_2 + \dots + x_ny_n$. Here, $n \geq 1$ is a natural number. Note that when $n = 1$, this dot product is just the product of the two real numbers x and y. Consider the Euclidean space $\mathbb{E}^n := (\mathbb{R}^n, \cdot)$, that is, our usual finite dimensional vector space $\mathbb{R}^n$ equipped with the dot product.

We have studied the following proposition in Mathematics:

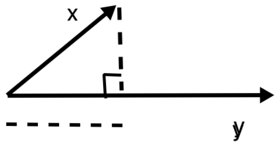
Statement 1. *If a vector x in $\mathbb{R}^n$ is such that $x \cdot y = 0$ for all y in $\mathbb{R}^n$, then x has to be the zero vector of $\mathbb{R}^n$.*

For $n = 1$, Statement 1 says that if the product of a real number x with every real number is zero then x has to be zero. For higher n, the statement says that if $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ is such that $x \cdot y = 0$ for each $y = (y_1, \dots, y_n)$ in $\mathbb{R}^n$ then x is the zero vector of $\mathbb{R}^n$. That is, $x_1 = x_2 = \dots = x_n = 0$.

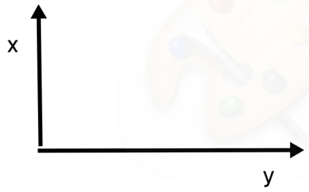
We know that the dot product $x \cdot y$ also has an interpretation closely related to the projection of the vector x on the one dimensional subspace generated by y. A philosophical perspective can be given to this mathematical result. Suppose we treat the value of the dot product of x with y as the opinion that y carries about the worth of x. For example, if $x \cdot y = \epsilon$ where ϵ is a small positive real number, it could mean that y thinks x has a low (but still a positive) worth. In other words, from the perspective of y, x is of low worth but still has some positive worth. In particular, $x \cdot y = 0$ means y thinks that x is worthless, that is, the projection of x to the world of y is zero. When $x \cdot y$ has a negative value, one can say that y thinks negatively about x. That is, x casts a negative shadow on y.



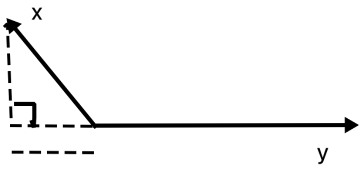
z is the projection of x onto y



Projection of x onto y
is positive



Projection of x onto y
is zero



Projection of x onto y
is negative

With this interpretation, Statement 1 means that if every y thinks that x is worthless then x is indeed worthless. Does that mean that our worth is determined by the opinion of others? We are told that we shouldn't care so much about the opinion of others about us, and that our worth is what we think and make of ourselves. If this teaching is to be believed then it looks like Mathematics is teaching us something wrong! How can it be! I have had faith in Mathematics, Maths can't be wrong. This is shattering!

There is a catch here. Let me explain. We know that:

Statement 2. *If o denotes the zero vector $(0, 0, \ldots, 0)$ of $\mathbb{R}^n$, then we have $x \cdot o = 0$ for every x in $\mathbb{R}^n$.*

Now, the contrapositive of Statement 1 states that:

Statement 3. *If x is a nonzero vector then there exists at least one nonzero $y \in \mathbb{R}^n$ such that $x \cdot y \neq 0$.*

This can be interpreted as "If at least one y sees some worth in x then x isn't worthless."

Please note that if $x = o$, the zero vector of $\mathbb{R}^n$, then clearly $x \cdot x = 0$. That is, if x is a zero vector (in other words, a loser) then clearly x knows that very well. x can't lie to itself and it gets reflected in the projection of x on itself.

Next, for a dot product, we have:

Statement 4.

(a) *$x \cdot x \geq 0$ for all $x \in \mathbb{R}^n$. (Meaning: the opinion of x about itself can never be negative).*

(b) *And that, if for some x in $\mathbb{R}^n$ if $x \cdot x = 0$ then x has to be the zero vector of $\mathbb{R}^n$. (Meaning: if x thinks of itself as worthless then it is worthless).*

This also means that

Statement 5. *If x is a vector in $\mathbb{R}^n$ such that $x \cdot x \neq 0$ then x can not be the zero vector. (Meaning: If x thinks it has some worth then it isn't worthless in reality.)*

In fact, the moral of the story is the following:

Statement 6. *For a vector x in $\mathbb{R}^n$, $x \cdot x = 0$ if and only if x is the zero vector. (Meaning: You are worthless when and only when you think you are!)*

This restores my faith in mathematics as now it aligns with my understanding of self worth. My worth has nothing to do with others' opinions about me, it is completely determined by my own projection on myself.

In fact, if we look at Statement 1 more carefully we realise that 'all' includes x itself, and therefore Statement 1 is not saying much or contradicting common sense. And it is impossible for "all others" to measure x's worth as 0 if x has a positive self-worth.

You are what you believe you are!

Or rather,

You are what you think you are!

Please remember that I am not saying that "You are what you say you are!" There is a difference. Thoughts and beliefs are honest in our heart. But a person can lie while saying/expressing his thoughts and beliefs to the outside world.

Reminds me of a couplet by the famous Urdu poet Sahir Ludhyanvi:

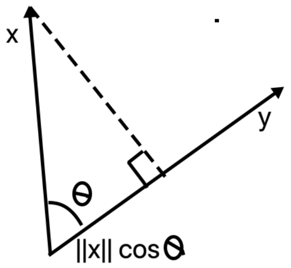
Le de ke apne paas fakat ek nazar to hai, Kyon dekhein zindagi ko kisi ki nazar se hum!

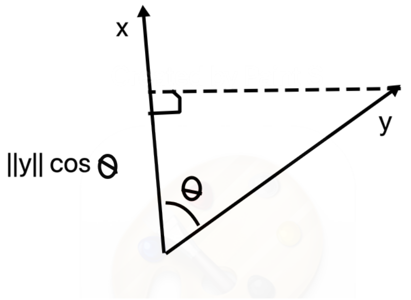
Why should we look at life through others’ lenses when we have our own worldview!

ले दे के अपने पास फ़क़त इक नज़र तो है
क्यूँ देखें ज़िंदगी को किसी की नज़र से हम

لے دے کے اپنے پاس فقط اک نظر تو ہے
کیوں دیکھیں زندگی کو کسی کی نظر سے ہم

We know that the projection of the vector x on the vector y denoted by $Proj_y(x)$ is given by the expression $\|x\| \cos\theta$. Here, θ is the angle between the vectors x and y in the two dimensional plane spanned by x and y. The norm of the vector x here plays a crucial role. Moreover, $x \cdot y = \|x\|\|y\| \cos\theta = \|y\|Proj_y(x)$. So, if y is a zero vector itself, its evaluation of every other x will be nil. So, please do not go by the judgements of worthless people.





In Sanskrit this is summarised by a line that says

Yatha Drishti, Tatha Shrishti!

यथा दृष्टि तथा सृष्टि

Meaning, we only see what lies within us. And, in Gujarati we say

KamaLo hoy tene piLu dekhhaay.

કમળો થયો હોય તેને બધું પીળું જ દેખાય

Meaning we see only yellow if we are suffering from jaundice.

For the readers who are familiar with abstract inner product spaces, we would like to remark that all the above statements hold true in every inner product space $(V, \langle , \rangle)$. Here, V is a vector space and $\langle , \rangle$ is an inner product on V.

The generalisation of the above statements for general inner product spaces are given below:

Statement 1′. *If a vector x in an Inner Product Space $(V, \langle , \rangle)$ is such that $\langle x, y\rangle = 0$ for all y in V, then x has to be the zero vector of V.*

Statement 2′. *If o denotes the zero vector of the inner product space $(V, \langle , \rangle)$, then we have $< x, o >= 0$ for every x in V.*

Statement 3′. *If x is a nonzero vector then there exists at least one nonzero $y \in V$ such that $\langle x, y\rangle \neq 0$.*

Statement 4′.

(a) $\langle x, x\rangle \geq 0$ *for all x in V.*

(b) *And that, if for some x in V if $\langle x, x\rangle = 0$ then x has to be the zero vector of V.*

Statement 5′. *If x is a vector in V such that $\langle x, x\rangle \neq 0$ then x can not be the zero vector.*

Statement 6′. *For a vector x in V, $\langle x, x\rangle = 0$ if and only if x is the zero vector.*

Definition: *An inner product space $(V, \langle , \rangle)$ is said to be a Hilbert space if it is complete with respect to the norm induced by the inner product.*

In the realm of Hilbert Spaces, there is another statement in Functional Analysis which states that

Statement 7. *Let $(u_\alpha)_{\alpha\in\Lambda}$ be an orthonormal basis of a Hilbert space $(H, \langle , \rangle)$. Then, $\langle x, u_\alpha\rangle = 0$ for all $\alpha \in \Lambda$ if and only if $x = o$.*

This gives us a feeling that if some important representatives of the system think that we are worthless then we are. But even in this case, it turns out that all that matters is our own perception of ourselves! This is because the proof essentially involves the inner product of x with itself. So, even if all the people at important positions feel that you are worthless, you aren't as long as you believe in yourself. This is nice, as we expected it to be. And this is possible since our inner product space is also "nice" as described in the definition of a Hilbert space given above. So, the system also has to be "nice" for such expected good things to happen!

10 A Neighbor-Elimination Random Walk on a Circle

Jyotirmoy Sarkar

Indiana University Indianapolis
Department of Mathematical Sciences
402 N Blackford Street
Indianapolis, IN 46202-3216, USA.
Email: jsarkar@iu.edu

Bikas K. Sinha

(Retired) Former Professor
Indian Statistical Institute
Kolkata, India.
Email: bikassinha1946@gmail.com

DEDICATION

This paper is dedicated to the memory of Krishna B. Athreya, Distinguished Professor Emeritus, Iowa State University, USA, a Fellow of the Indian Academy of Sciences, and an Institute of Mathematical Statistics Fellow. He communicated to the second author the problem we study here, including some manual computations for $n \leq 8$. He died at his home in Ames, Iowa, on March 24, 2023, before the research could proceed further. See his obituary at https://imstat.org/2023/05/16/obituary-krishna-b-athreya-1939-2023/

Abstract
A team of n players form a circle by holding hands. Player 1 tentatively wears the captain's hat. The other players are numbered as $2, 3, \ldots, n$ going clockwise from Player 1. At each stage, a biased coin is tossed. Heads, which occur with probability p, eliminate the tentative captain's clockwise neighbor; and tails, which occur with probability $q = 1 - p$, eliminate the counterclockwise neighbor. The eliminated player transfers the hat from the tentative captain on one side to the player on his other side and steps out of the circle. The previous and the current tentative captains join hands, reducing the circle by one player. Such neighbor elimination continues until only one player remains wearing the hat, and becomes the ultimate captain. We find the probability distribution of the serial number of the ultimate captain and document some mathematical properties.

Key Words and Phrases: Binomial Coefficients, Mathematical Induction, Palindrome, Stochastic Recursion, Rotation Symmetry

Mathematics Subject Classification: 60G07

1 Introduction

One of the authors recalls from his youth a strategy by which team captains were chosen when players assembled at impromptu games. Typically, the players would stand in a circle. One player would recite a rhyme of his choice. As each word is pronounced he would point at a player, beginning with himself and going clockwise or counterclockwise, whichever he desired. The player who was allocated the last word of the rhyme became the captain. After a few applications of this method, the players noticed that the reciter had too much sway to manipulate the outcome (by breaking a word into syllables, or by adding one or more irrelevant words, or by stopping earlier or later than normal) to influence the choice of captain—often trying to make himself the captain. So a revision of the method was called for. One such amendment involved eliminating the player on whom the last word fell, then re-forming the circle and letting the next person recite. This went on until only one player remained who became the ultimate captain.

In this paper, we are going to study an elimination random walk, reminiscent of the above captain selection process, but with a more precisely defined *random mechanism* to determine who gets eliminated. Let us elaborate.

At time $t = 0$, a team of n players form a circle by holding hands. One of the players tentatively wears the captain's hat and is numbered as Player 1, while the other players are numbered as $2, 3, \ldots, n$ going clockwise from Player 1. At each epoch $t = 1, 2, \ldots, n-$

1, a biased coin is tossed. If the coin lands heads, which occurs with probability p, then the tentative captain's clockwise neighbor (the player on the left) is eliminated. If the coin lands tails, which occurs with probability $q = 1 - p$, then the counterclockwise neighbor (the player on the right) is eliminated. The eliminated player transfers the hat from the tentative captain on one side to the player on the other side and steps out of the circle. The previous and the current tentative captains join hands, shrinking the circle down by one player. Thus at each stage, the tentative captain's left or right neighbor is eliminated and the hat moves further left or further right with probabilities p and q, respectively. Such elimination continues until only one player remains, who becomes the ultimate captain. We call this stochastic process the *neighbor-elimination random walk on a circle*.

To put this new problem in context, readers may review random walks on n points on a circle — without elimination. We refer them to Maiti and Sarkar [5] for the symmetric case when $p = q = 1/2$, and to Sarkar [8] for the asymmetric case when $p \neq q$.

Indeed, the reader need not assemble a team of players and watch them determine the captain in the *neighbor-elimination random walk on a circle*. Instead, the stochastic process can be imitated by spreading out chips numbered $1, 2, 3, \ldots, n$ clockwise on the circumference of a circle and placing a cup over Chip 1. Then the reader tosses a biased coin. If the coin lands heads, the reader moves the cup clockwise going over and eliminating the next chip in the clockwise direction and landing on the following chip as he continues in the same direction. Likewise, if the coin lands tails, the reader moves the cup in the counterclockwise direction, eliminates the neighbor, and places the cup on the next chip in the same direction. After each coin toss and cup movement, the reader is faced with a reduced problem with one chip fewer than before. The game continues until only one chip remains covered. Having eliminated all other chips, the last chip becomes the sole survivor. Henceforth, in this paper, the words 'survivor' and 'captain' are used interchangeably.

In this paper, we find the probability distribution of the serial number of the ultimate captain or survivor. Let X_n denote the serial number of the ultimate captain/survivor in an n-player/chip game. Let ${}^np_k, (1 \leq k \leq n)$ denote the probability that in an n-player game the ultimate captain is Player k; that is, let ${}^np_k = P\{X_n = k\}$. Let ${}^n\mathbf{p} = ({}^np_1, {}^np_2, {}^np_3, \ldots, {}^np_{n-1}, {}^np_n)$ denote the $1 \times n$ vector of probabilities associated with the random variable X_n. Clearly, ${}^n\mathbf{p}$ can be **approximated by running the simulation** mentioned in the preceding paragraph 10^6 times, say. But we will do far better than that by computing ${}^n\mathbf{p}$ exactly and efficiently.

The paper is organized as follows: We present the special case of the deterministic walk with $p = 1$ in Section 2. In Section 3, we derive the general recursive relation to obtain ${}^{n+1}\mathbf{p}$ starting from ${}^n\mathbf{p}$. Consequently, we obtain the distribution of X_{n+1} based on that of X_n. In Section 4, given the sequence of directions in which the hat moves, we determine the ultimate captain. If we repeat this determination for each possible sequence, we can construct the **exact distribution** of ${}^n\mathbf{p}$. But, of course, using the **recursive relation** is much more efficient. Section 5 recursively calculates the distribution in the symmetric case when $p = q = 1/2$, and documents some mathematical properties of ${}^n\mathbf{p}$. Section 6 does the same for the asymmetric case. Here, by letting $p \to 1$, we recover the result of the deterministic case of Section 2. Section 7 concludes the paper with a summary and some open problems.

2 Deterministic Walk (when $p = 1$)

The following is a well-known result. It appears in many books and on many websites—with or without proof. We recommend interested readers check out Levitin and Levitin [1], which is a treasure trove of many other attractive puzzles. Here we present a new proof.

Theorem 1. *Let there be $n \geq 2$ players. Write $n = 2^m + k$ where $0 \leq k < 2^m$ for the unique $m \geq 1$. Starting from Player 1 as tentative captain, if each tentative captain eliminates the neighbor in the clockwise direction and the next player in the clockwise direction becomes the tentative captain, then the ultimate captain will be $2k + 1$.*

Proof. The proof is by induction on m. If $m = 1$, then either $k = 0$ or $k = 1$. If $m = 1$ and $k = 0$, then $n = 2^1 + 0 = 2$. Player 1 eliminates Player 2, and the ultimate captain is Player 1. Indeed, $2k + 1 = 2 * 0 + 1 = 1$. If $m = 1$ and $k = 1$, then $n = 2^1 + 1 = 3$. Player 1 eliminates Player 2, Player 3 eliminates Player 1, and the ultimate captain is Player 3. Again, $2k + 1 = 2 * 1 + 1 = 3$. Thus the result holds for $m = 1$.

Assume that the result holds for $m - 1$. We will show that the result must hold for m. Let $n = 2^m + k$ for some $0 \leq k < 2^m$. The players have first-round serial numbers (FRSN) $1, 2, 3, \ldots, n$. We say the first round of elimination ends as soon as each player either eliminates someone or is eliminated by someone. We consider two cases according to the parity of n (equivalently, of k) and show that the result holds for m in both cases.

Case I: n (hence, k) is even. When Round 1 ends, all players with even FRSN are eliminated, and Player 1 is again the tentative captain. A player with an odd FRSN i receives a second round serial numbers (SRSN) $j = (i + 1)/2$, for $i = 1, 3, \ldots, n - 1$. In Round 2, the number of players is reduced to $n_2 = n/2 = 2^{m-1} + k/2$ where $0 \leq k/2 < 2^{m-1}$. Clearly, $n_2 < 2^m$. Therefore, by the induction hypothesis, the ultimate captain's SRSN is $2(k/2) + 1 = k + 1$. This ultimate captain's FRSN x is obtained by solving $(x + 1)/2 = k + 1$, or $x = 2k + 1$.

Case II: n (hence, k) is odd. When Round 1 ends, all players with even FRSN are eliminated and also Player 1 is eliminated. The tentative captain is Player 3. Each surviving player (with an odd FRSN $i > 1$) receives an SRSN $j = (i-1)/2$, for $i = 3, 5, \ldots, n$. In Round 2, the number of players have reduced to $n_2 = (n-1)/2 = 2^{m-1} + (k-1)/2$ where $(k-1)/2 < 2^{m-1}$. So, $n_2 < 2^m$. Therefore, by the induction hypothesis, the ultimate captain's SRSN is $2\{(k-1)/2\}+1 = k$. This ultimate captain's FRSN x is obtained by solving $(x-1)/2 = k$, or $x = 2k+1$. □

Remark 1. In binary notation we write $n = 2^m + k$ as $(a_m = 1, a_{m-1}, a_{m-2}, \ldots, a_2, a_1, a_0)$, where $a_i = 0, 1$ for $i = 0, 1, 2, \ldots, m-1$. Then the ultimate winner has a serial number $2k+1$, whose binary expansion is $(a_{m-1}, a_{m-2}, \ldots, a_2, a_1, a_0, 1)$. Thus, we simply remove the leftmost digit 1 of the binary expansion of n and augment it to the right of a_0. Note also that, as a single function of n, the last survivor is $2(n - 2^{\lfloor \log_2(n) \rfloor}) + 1$.

The deterministic problem, recounted in Theorem 1, is a special case of Josephus Problem, sometimes dubbed the most violent math problem. Herstein and Kaplansky [3] recount the following legend about the famous first-century historian Flavius Josephus (see [4]): In the Jewish revolt against Rome, Josephus and 39 of his comrades were holding out against the Romans in a cave. With defeat imminent, they resolved that they would rather die than be slaves to the Romans. They arranged themselves in a circle with serial numbers $1, 2, 3, \ldots, 40$, where 40 is to the right of 1. Starting from Number 1, they counted clockwise killing every *seventh* man. Josephus, an accomplished mathematician, instantly figured out that Number 24 would be the last survivor, where he positioned himself. But when the time came, instead of killing himself, he joined the Roman side and lived to tell the tale.

If, instead of killing every seventh man, the agreement was to kill every second man, then by Theorem 1, Josephus should have assumed position 17. If the agreement was to kill every third man, the position would be 31. For other variations and historical notes see Singmaster [9]. We leave it to the reader to discover the serial number of the last survivor among n people if going clockwise from Person 1, every d-th person is eliminated. Alternatively, they may see Halbeisen and Hungerbuhler [2]. We should also mention that, instead of studying only the last survivor, Robinson [7] studies the entire permutation of serial numbers of players in the order in which they are eliminated.

3 Recursive Evaluation of ^{n}p

Let us return to the neighbor-elimination random walk where the clockwise neighbor is eliminated with probability $p < 1$ or the counterclockwise neighbor is eliminated with probability $q = 1 - p$. Thus, we are studying a randomized version of the Josephus Problem when $d = 2$.

We proceed by increasing the number of players/chips one by one, by developing a recursive relation, and by establishing results using mathematical induction on n. Of course, when $n = 1$, Player 1 by default is the ultimate captain with probability 1. So, ${}^1\mathbf{p} = ({}^1p_1) = (1)$. Naturally also, when $n = 2$, there is no need to toss a coin, since irrespective of the outcome, Player 2 is eliminated, and he is forced to crown Player 1 as the ultimate captain with probability 1. Thus, ${}^2\mathbf{p} = ({}^2p_1, {}^2p_2) = (1, 0)$. When there are $n = 3$ players, the coin is tossed once. Then with probability p, the coin lands heads, Player 2 is eliminated, and Player 3 becomes the tentative captain in the presence of only Player 1. Now that the problem is reduced to two players only, in the follow-up step, Player 3 is guaranteed to become the ultimate captain. Therefore, in a three-player game, Player 3 becomes the ultimate captain with probability p. Likewise, starting from the original three players, with probability $q = 1 - p$, the coin lands tails, Player 3 is eliminated, and Player 2 becomes the tentative captain in the presence of only Player 1, who surely will be eliminated in the next step. Therefore, three-player game, Player 2 becomes the ultimate captain with probability q. Hence, ${}^3\mathbf{p} = ({}^3p_1, {}^3p_2, {}^3p_3) = p(0, \_, 1) + q(0, 1, \_) = (0, q, p)$. Here, a missing value symbol (underscore) indicates that the corresponding player has been eliminated, and hence is filled in by a zero probability to continue the computation further.

One can continue in this manner, to derive ${}^{n+1}\mathbf{p}$ based on ${}^n\mathbf{p}$. Recalling that in a 2-player game no coin toss is needed, by mathematical induction, for the n-player game we need to toss the coin exactly $(n - 2)$ times, since each toss eliminates one player. The outcomes of the tosses will determine the ultimate captain as will be explained in Theorem 2 below. Suppose that we have figured out how to calculate ${}^n\mathbf{p}$ in an n-player game based on the outcomes of $(n - 2)$ tosses. Let us explain how we calculate ${}^{n+1}\mathbf{p}$ in an $(n + 1)$-player game based on $(n - 1)$ tosses.

For an $(n + 1)$-player game, note that after the first coin toss, Player 2 is eliminated with probability p, and Player $(n + 1)$ is eliminated with probability q. If Player 2 is eliminated, we form a $1 \times (n + 1)$ vector ${}^{n+1}\mathbf{a}$ with position 2 marked with a missing symbol (underscore) and the probability vector ^{n} $\mathbf{p}$ re-written cyclically starting from position 3. Thus, the k-th element of ${}^n\mathbf{p}$, namely ${}^np_k$, becomes the $(k + 2)$-nd element [mod $(n + 1)$] of ${}^{n+1}\mathbf{a}$. In particular, the last element of ${}^n\mathbf{p}$, namely ${}^np_n$, becomes the first element of ${}^{n+1}\mathbf{a}$, namely ${}^{n+1}a_1$. This vector ${}^{n+1}\mathbf{a}$ is multiplied by p, which is the probability of eliminating Player 2 after the first toss.

Similarly, if Player $(n + 1)$ is eliminated, we form a second $1 \times (n + 1)$ vector ${}^{n+1}\mathbf{b}$ in which the probability vector ${}^n\mathbf{p}$ is re-written cyclically starting from position n and inserting in position $(n + 1)$ (the last position) a missing symbol. Thus, ${}^np_k$ becomes the $(k - 1)$-st element [mod (n)] of ${}^{n+1}\mathbf{b}$. In particular, the first element of ${}^n\mathbf{p}$, namely ${}^np_1$, becomes the second last element of ${}^{n+1}\mathbf{b}$, namely ${}^{n+1}b_n$. This second vector ${}^{n+1}\mathbf{b}$ is multiplied by q, which is the probability of eliminating Player $(n + 1)$ after the first toss.

Finally, these two vectors (multiplied by p and q, respectively) are added up (with the missing symbols replaced by zeroes) to obtain the probability vector for the $(n+1)$-player game ${}^{n+1}\ \mathbf{p} = p \cdot {}^{n+1}\mathbf{a} + q \cdot {}^{n+1}\mathbf{b}$. For instance,

$$\begin{aligned} {}^4\mathbf{p} &= p \cdot {}^4\mathbf{a} + q \cdot {}^4\mathbf{b} = p\ (p,\ \_\_, 0, q) + q\ (q, p, 0,\ \_\_) \\ &= (p^2+q^2, qp, 0, pq) = ({}^4p_1, {}^4p_2, {}^4p_3, {}^4p_4) \\ {}^5\mathbf{p} &= p \cdot {}^5\mathbf{a} + q \cdot {}^5\mathbf{b} = p\ (pq,\ \_\_, p^2+q^2, qp, 0) + q\ (qp, 0, pq, p^2+q^2,\ \_\_) \\ &= (p^2q+pq^2, 0, p^3+2pq^2, 2p^2q+q^3, 0) = ({}^5p_1, {}^5p_2, {}^5p_3, {}^5p_4, {}^5p_5) \end{aligned}$$

In general, for any $n \geq 2$, we have

$$\begin{aligned} {}^{n+1}\mathbf{p} &= p \cdot {}^{n+1}\mathbf{a} + q \cdot {}^{n+1}\mathbf{b} \\ &= \quad p\ ({}^np_n,\ \ \_\_, {}^np_1, {}^np_2, \ldots, {}^np_{n-3}, {}^np_{n-2}, {}^np_{n-1}) \\ &\quad + q\ ({}^np_2, {}^np_3, {}^np_4, {}^np_5, \ldots,\ {}^np_n,\quad {}^np_1,\quad \_\_\quad) \\ &= ({}^{n+1}p_1, {}^{n+1}p_2, {}^{n+1}p_3, {}^{n+1}p_4, \ldots, {}^{n+1}p_{n-1}, {}^{n+1}p_n, {}^{n+1}p_{n+1}) \end{aligned} \tag{1}$$

For a diagrammatic representation of (1), see Figure 1.

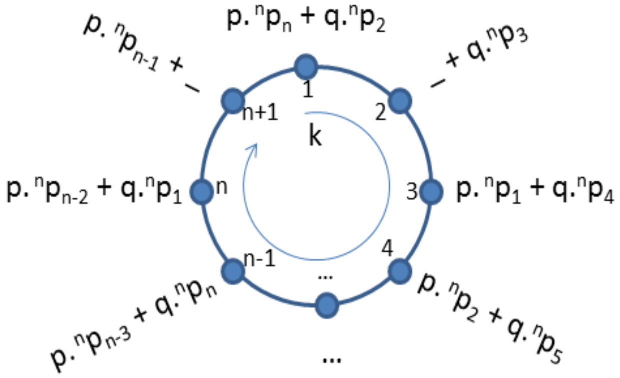


Figure 1. Inductive determination of ${}^{n+1}\mathbf{p}$ starting from ${}^n\mathbf{p}$ using (1), in an asymmetric neighbor-elimination random walk on a circle

The following theorem is a straight-forward consequence of (1).

Theorem 2. *For $n \geq 2$, X_{n+1} has the same distribution as $(X_{n+2})\ (\ mod\ n+1)$ with probability p, and $(X_{n-1})\ (\ mod\ n)$ with probability $q = 1-p$.*

To develop an algorithmic evaluation of (1), we adopt the following convention to extend the probability vector ${}^n\mathbf{p}$ by attaching two elements before it and two elements after it;

that is, we define

$$ {}^{n}p_{-1} = {}^{n}p_{n}, \quad {}^{n}p_{0} = 0, \quad {}^{n}p_{n+1} = {}^{n}p_{1}, \quad {}^{n}p_{n+2} = 0 \tag{2} $$

Thus, the first element of ${}^{n}\mathbf{p}$ is duplicated at the end followed by a zero, and the last element of ${}^{n}\mathbf{p}$ is duplicated at the front followed by a zero. Also, let D_j^n denote the operation on an $1 \times n$ vector that deletes the j-th element of the vector to produce an $1 \times (n-1)$ vector. Using (2), (1) can be written as

$$ {}^{n+1}\mathbf{p} = p \cdot ({}^{n}p_{n}, 0, D_n^n \circ {}^{n}\mathbf{p}) + q \cdot (D_1^n \circ {}^{n}\mathbf{p}, {}^{n}p_{1}, 0) \tag{3} $$

Equating the vector elements on the two sides of (3), we have

$$ {}^{n+1}p_{k} = p \cdot {}^{n}p_{k-2} + q \cdot {}^{n}p_{k+1} \qquad \text{for } 1 \leq k \leq n+1 \tag{4} $$

which is well defined in view of (2).

Using (4), one can proceed step-by-step to evaluate the probability vector ${}^{n}\mathbf{p}$. We shall, however, describe a more efficient way to obtain ${}^{n}\mathbf{p}$ in Section 6 by separating the coefficients (pure numbers) and the powers of p and q. First, in Section 4, let us discuss how a given sequence of outcomes of the coin toss determines the captain in an n-player game. Then we will study the special case of $p = 1/2$ in Section 5.

4 The Sequence of Tosses Determines the Captain

Recall from above that Player 1 is the captain among $n = 1, 2$ players, and in an n-player game (with $n \geq 3$) the captain is chosen by tossing the coin $(n-2)$ times, which of course results in a total of 2^{n-2} possible outcomes. Each sequence of $(n-2)$ tosses uniquely determines a captain. We give an explicit expression for this function in the next theorem, which follows directly from Theorem 1. Suppose that $W_j, j \geq 3$ denotes the outcome of the biased coin used to reduce the j-player problem to the $(j-1)$-player problem. That is, $W_n, W_{n-1}, \ldots, W_3$ are independent and identically distributed Bernoulli random variables taking values H or T with probabilities p and $q = 1 - p$, respectively.

Theorem 3. *Let ${}^{n}\boldsymbol{W} = (W_n, W_{n-1}, \ldots, W_3)$ be the vector of outcomes of $(n-2)$ independent tosses of a biased coin. Let f_n be the function that determines the ultimate captain in an n-player game. Then $f_1 \equiv 1$, $f_2 = 1$, and for any $n \geq 2$, we have*

$$ f_{n+1}({}^{n+1}\boldsymbol{W}) = \begin{cases} (f_n({}^{n}\boldsymbol{W}) + 2) \pmod{n+1} & \text{if } W_{n+1} = H \\ (f_n({}^{n}\boldsymbol{W}) - 1) \pmod{n} & \text{if } W_{n+1} = T \end{cases} $$

Applying Theorem 2, we have $f_3(H) = 1{+}2 \pmod 3 = 3$, and $f_3(T) = 1{-}1 \pmod 2 = 2$. Likewise, for four players, $f_4(HH) = 3{+}2 \pmod 4 = 1$, $f_4(HT) = 2{+}2 \pmod 4 = 4$, $f_4(TH) = 3 - 1 \pmod 3 = 2$, and $f_4(TT) = 2 - 1 \pmod 3 = 1$. And so on. Of course, in order to evaluate f_{n+1} for any one particular sequence of $(n - 1)$ outcomes, we do not need to evaluate f_n for all possible 2^{n-2} outcome sequences. It suffices to evaluate $f_3, f_4, \ldots, f_{n+1}$ by considering the outcomes of the coin toss one-by-one in reverse order. For example, in a 10-player game, if the outcomes are $HHTTHTHH$, which player becomes the captain? We answer it by successively evaluating $f_3(H) = 1 + 2 \;(\bmod 3) = 3$, $f_4(HH) = 3 + 2 \;(\bmod 4) = 1$, $f_5(THH) = 1 - 1 \;(\bmod 4) = 4$. Thereafter, (suppressing the arguments), we have $f_6 = 6, f_7 = 5, f_8 = 4, f_9 = 6, f_{10} = 8$. Hence, in a 10-player game, if the outcome sequence is $HHTTHTHH$, then Player 8 is the ultimate captain.

Note that several different sequences of $(n - 2)$ tosses determine the same captain, and there are players who are never chosen as the captain no matter what the outcome sequence is. For example, for $n = 4$, both HH and TT determine Player 1 as the ultimate captain, and Player 3 is never chosen as the captain. Thus the "ultimate captain determination" mapping from the set of outcome sequences of $(n - 2)$ tosses to the set of players is a many-to-one function, but it is not an onto function.

In an n-player game, originally Player k stands as far from Player 1 as Player $(n+2-k)$ stands from Player 1, but in the opposite direction. Hence, any sequence of tosses that declares Player k the ultimate caption, when complemented term by term, must declare Player $(n + 2 - k)$ the ultimate captain. Therefore, the following lemma holds.

Lemma 1. *Suppose that W_j^c denotes the complement of W_j; that is, $W_j^c = H$ iff $W_j = T$. Let ${}^n\boldsymbol{W}^c = (W_n^c, W_{n-1}^c, \ldots, W_3^c)$. Then $f_n({}^n\boldsymbol{W}^c) = n + 2 - f_n({}^n\boldsymbol{W})$.*

We showed above $f_{10}(HHTTHTHH) = 8$. Hence, by Lemma 1, $f_{10}(TTHHTHTT) = 10 + 2 - 8 = 4$. That is, in a 10-player game, if the outcome sequence is $TTHHTHTT$, then Player 4 is the ultimate captain.

Consider also a very special situation in which you have delegated the coin tossing to a third party who plays a prank on you by tossing a two-tailed coin (that is, $p = 0$). Who then will be the ultimate captain in an n-player game? In this case, all tosses result in tails. So (without writing the arguments), $f_2 \equiv 1, f_3 = 2, f_4 = 1, f_5 = 4, f_6 = 3, f_7 = 2, f_8 = 1, f_9 = 8, \ldots$. In general, if $n = 2^m + k$ where $1 \le k \le 2^m$, then $f_n = 2^m + 1 - k$. Likewise, if a two-headed coin is tossed (that is, $p = 1$), then $f_n = 2k + 1$. This result has been proved by induction on n in Section 2. In particular, in a 10-player game, a two-tailed coin declares Player 7 as the ultimate winner and a two-headed coin, Player 5. Of course, we mention these results not because we suspect the coin tosser will play such pranks, but because we want to establish the benchmark degenerate distribution of X_n as $p \to 0$ or $p \to 1$, stated below.

Theorem 4. *If $n = 2^m + k$ where $1 \leq k \leq 2^m$, then as $p \to 0$, X_n converges in distribution (and in probability) to $2^m + 1 - k$, and as $p \to 1$, X_n converges in distribution to $2k + 1$.*

Our goal is to determine ${}^n p_k$. We can determine the ultimate captain in all 2^{n-2} sequences of outcomes of $(n-2)$ tosses and collect the probabilities corresponding to all sequences that cause Player k to become the captain. This is the **method of complete enumeration**.

However, to determine ${}^n p_k$ more efficiently, we need to consider not only the subset of all 2^{n-2} outcome sequences that correspond to Player k being chosen as the ultimate captain, but also partition this subset into components corresponding to each distinct number of heads because the probability associated with each outcome sequence depends on the number of heads among the $(n-2)$ tosses. Fortunately, this partitioning is not essential in the symmetric case when $p = q = 1/2$, since then each outcome sequence has the same probability 2^{-n+2}. Therefore, let us first study the symmetric case in Section 5 before we shall return to the asymmetric case in Section 6.

5 The Symmetric Case: $p = q = 1/2$

In the symmetric case when $p = q = 1/2$, note that all elements of ${}^n\mathbf{p}$ have the same common denominator 2^{n-2}. Our task is to determine the numerator in each element of ${}^n\mathbf{p}$. Let us denote these numerators by ${}^n S_k = 2^{n-2} \cdot {}^n p_k$ for $1 \leq k \leq n$, and define the vector of numerators as

$$ {}^n\mathbf{S} = ({}^n S_1, {}^n S_2, \ldots, {}^n S_n) = 2^{n-2} \cdot {}^n\mathbf{p} \tag{5} $$

The question at hand is how many outcome sequences correspond to the same player chosen as the ultimate captain; that is, what are the values of ${}^n S_k$ for $1 \leq k \leq n$. To answer this question we proceed **recursively**.

As in the previous section, we note that ${}^1\mathbf{S} = (1)$, ${}^2\mathbf{S} = (1, 0)$ and analogous to (3) for any $n \geq 2$, we have

$$ {}^{n+1}\mathbf{S} \quad = \quad ({}^n S_n,\ \_, D_n^n \circ {}^n\mathbf{S}) + (D_1^n \circ {}^n\mathbf{S}, {}^n S_1,\ \_) \tag{6} $$

where the first term is a right-shift by two positions [mod $(n+1)$], and the second term is a left-shift by one position [mod (n)].

Starting from ${}^2\mathbf{S} = (1, 0)$, we apply (6) repeatedly to get ${}^3\mathbf{S} = (0, \_, 1) + (0, 1, \_) = (0, 1, 1)$, ${}^4\mathbf{S} = (1, \_, 0, 1) + (1, 1, 0, \_) = (2, 1, 0, 1)$, and so on. Also, analogous to (2), we adopt the convention

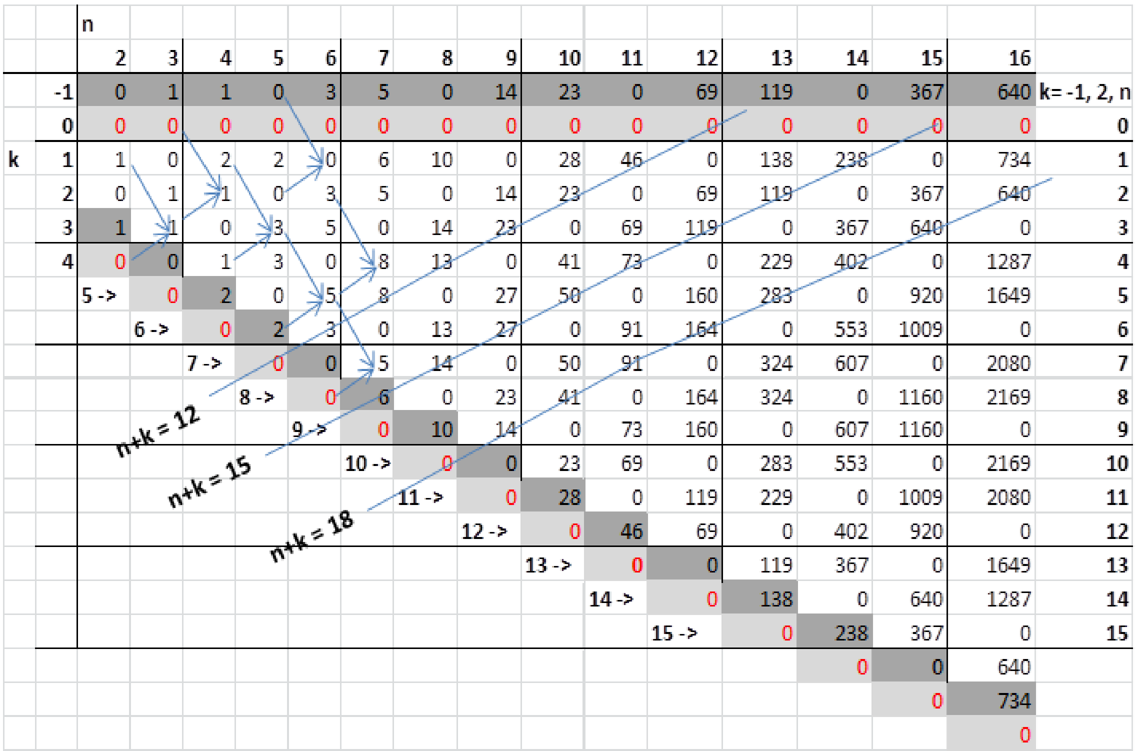
$$^{n}S_{-1} = {}^{n}S_{n}, \quad {}^{n}S_{0} = 0, \quad {}^{n}S_{n+1} = {}^{n}S_{1}, \quad \text{and } {}^{n}S_{n+2} = 0 \tag{7}$$

Then, analogous to (4), we have

$$^{n+1}S_{k} = {}^{n}S_{k-2} + {}^{n}S_{k+1} \qquad \text{for } 1 \leq k \leq n+1 \tag{8}$$

Starting from $^{2}\mathbf{S} = (1, 0)$, we use (8) recursively to obtain the values of $^{n}S_{k}$ for $1 \leq k \leq n$, $3 \leq n \leq 16$, and display $^{n}\mathbf{S}$ as column vectors in Table 1. For the readers' benefit, we give in the Appendix software codes using the freeware R.

Table 1. Numerators for the probability distribution of the survivor in a symmetric neighbor-elimination random walk on a circle, computed using (8). See R codes in the Appendix. The values in the shaded cells follow from (7). The denominator is always 2^{n-2}.



		n															
		2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
	-1	0	1	1	0	3	5	0	14	23	0	69	119	0	367	640	k= -1, 2, n
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
k	1	1	0	2	2	0	6	10	0	28	46	0	138	238	0	734	1
	2	0	1	1	0	3	5	0	14	23	0	69	119	0	367	640	2
	3	1	1	0	3	5	0	14	23	0	69	119	0	367	640	0	3
	4	0	0	1	3	0	8	13	0	41	73	0	229	402	0	1287	4
		5 ->	0	2	0	5	8	0	27	50	0	160	283	0	920	1649	5
			6 ->	0	2	3	0	13	27	0	91	164	0	553	1009	0	6
				7 ->	0	0	5	14	0	50	91	0	324	607	0	2080	7
					8 ->	0	6	0	23	41	0	164	324	0	1160	2169	8
			n+k = 12			9 ->	0	10	14	0	73	160	0	607	1160	0	9
							10 ->	0	0	23	69	0	283	553	0	2169	10
				n+k = 15				11 ->	0	28	0	119	229	0	1009	2080	11
						n+k = 18			12 ->	0	46	69	0	402	920	0	12
										13 ->	0	0	119	367	0	1649	13
											14 ->	0	138	0	640	1287	14
												15 ->	0	238	367	0	15
														0	0	640	
															0	734	
																0	

Many interesting properties can be found in Table 1. Some of these properties are listed below, and can be proved by mathematical induction on n.

Theorem 5. *The counting numbers $\{^{n}S_{k} : -1 \leq k \leq n+2,\ 2 \leq n\}$ displayed in white cells of the columns of Table 1, exhibit the following properties:*

a) *For* $1 \leq k \leq n+1$, ${}^{n+1}S_k = {}^{n}S_{k-2} + {}^{n}S_{k+1}$. *In particular,* ${}^{n+1}S_1 = 2 \cdot {}^{n}S_2$ *and* ${}^{n+1}S_2 = {}^{n}S_3$.

b) *For* $1 \leq k \leq n+1$, ${}^{n}S_k = 0$ *iff* $n + k \equiv 1$ (*mod* 3). *In particular, if* $n = 2m$ *is even then* ${}^{2m}S_{m+1} = 0$.

c) *For* $0 \leq k \leq n+2$, ${}^{n}S_k = {}^{n}S_{n+2-k}$*; that is, the extended vector* $({}^{n}S_0, {}^{n}\boldsymbol{S}, {}^{n}S_{n+1}, {}^{n}S_{n+2})$ *is a palindrome. Consequently,* $({}^{n}\boldsymbol{S}, {}^{n}S_{n+1})$ *and* $D_1^n \circ {}^{n}\boldsymbol{S}$ *are palindromes.*

d) *For* $1 \leq k \leq n+1$, ${}^{n+1}S_k = {}^{n}S_{k-2} + {}^{n}S_{n+1-k}$. *In particular,* ${}^{n+1}S_n = {}^{n+1}S_3 = {}^{n}S_1 + {}^{n}S_4$ *and* ${}^{n+1}S_{n+1} = {}^{n}S_{n-1}$.

e) *Given the values of* ${}^{m}\boldsymbol{u} = ({}^{m}S_m, {}^{m+1}S_{m-1}, {}^{m+2}S_{m-2}, \ldots, {}^{2m-1}S_1)$ *for any integer* m, *we obtain the values of*

$$({}^{m+1}S_{m+2}, {}^{m+2}S_{m+1}, {}^{m+3}S_m, {}^{m+4}S_{m-1}, \ldots, {}^{2m}S_3, {}^{2m+1}S_2, {}^{2m+2}S_1)$$

as follows: Start with ${}^{m+1}S_{m+2} = 2 \cdot {}^{m}S_m$, *which is double the first element of* $\boldsymbol{u}$, *and then cumulatively add to it the remaining elements of* $\boldsymbol{u}$ *and conclude with two additional terms* ${}^{2m+1}S_2 = {}^{2m}S_3$ *and* ${}^{2m+2}S_1 = 2 \cdot {}^{2m}S_3$.

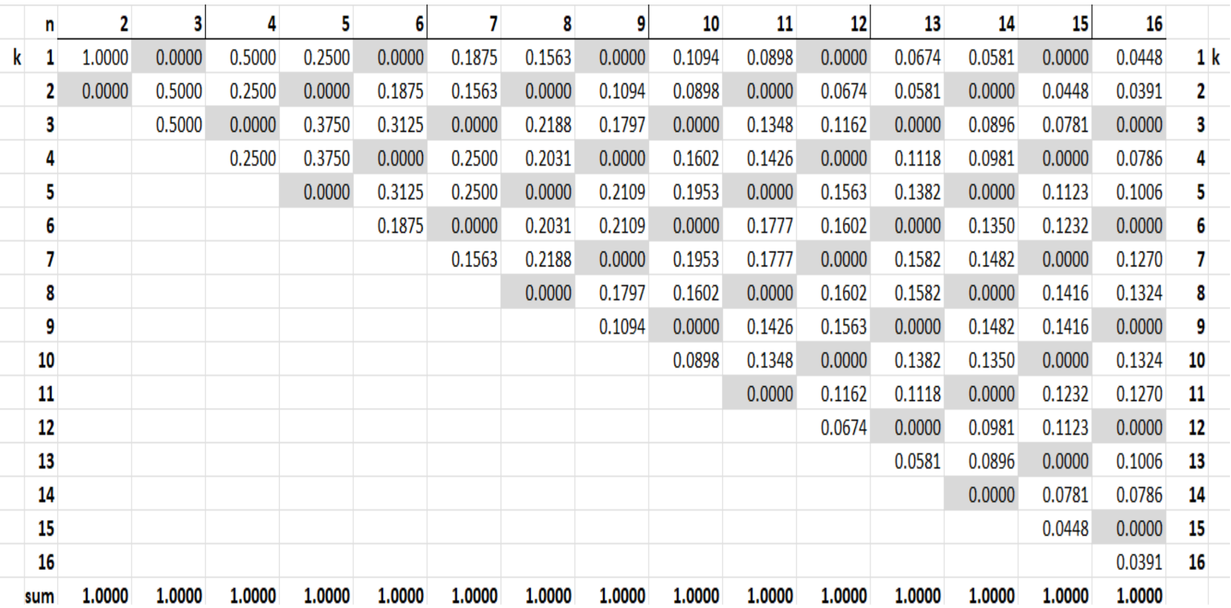
f) *Given the values of* ${}^{m}\boldsymbol{v} = ({}^{m+1}S_m, {}^{m+2}S_{m-1}, {}^{m+3}S_{m-2}, \ldots, {}^{2m}S_1)$ *for any integer* m, *we obtain the values of*

$$({}^{m+2}S_{m+2}, {}^{m+3}S_{m+1}, {}^{m+4}S_m, {}^{m+5}S_{m-1}, \ldots, {}^{2m+1}S_3, {}^{2m+2}S_2, {}^{2m+3}S_1)$$

as follows: Start with ${}^{m+2}S_{m+2} = {}^{m+1}S_m$, , *which is exactly the first element of* $\boldsymbol{v}$, *and then cumulatively add to it the remaining elements of* $\boldsymbol{v}$ *and conclude with two additional terms* ${}^{2m+2}S_2 = {}^{2m+1}S_3$ *and* ${}^{2m+3}S_1 = 2 \cdot {}^{2m+1}S_3$.

To construct the probability distribution of the survivor divide each element in column n of Table 1 by the column sum, which is 2^{n-2}.

Table 2. The probability distribution of the survivor in a symmetric neighbor-elimination random walk on a circle, computed from Table 1 by dividing each entry in column n by 2^{n-2}. Setting aside the first entry, the remaining entries in each column exhibit symmetry.



	n	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16		
k	1	1.0000	0.0000	0.5000	0.2500	0.0000	0.1875	0.1563	0.0000	0.1094	0.0898	0.0000	0.0674	0.0581	0.0000	0.0448	1	k
	2	0.0000	0.5000	0.2500	0.0000	0.1875	0.1563	0.0000	0.1094	0.0898	0.0000	0.0674	0.0581	0.0000	0.0448	0.0391	2	
	3		0.5000	0.0000	0.3750	0.3125	0.0000	0.2188	0.1797	0.0000	0.1348	0.1162	0.0000	0.0896	0.0781	0.0000	3	
	4			0.2500	0.3750	0.0000	0.2500	0.2031	0.0000	0.1602	0.1426	0.0000	0.1118	0.0981	0.0000	0.0786	4	
	5				0.0000	0.3125	0.2500	0.0000	0.2109	0.1953	0.0000	0.1563	0.1382	0.0000	0.1123	0.1006	5	
	6					0.1875	0.0000	0.2031	0.2109	0.0000	0.1777	0.1602	0.0000	0.1350	0.1232	0.0000	6	
	7						0.1563	0.2188	0.0000	0.1953	0.1777	0.0000	0.1582	0.1482	0.0000	0.1270	7	
	8							0.0000	0.1797	0.1602	0.0000	0.1602	0.1582	0.0000	0.1416	0.1324	8	
	9								0.1094	0.0000	0.1426	0.1563	0.0000	0.1482	0.1416	0.0000	9	
	10									0.0898	0.1348	0.0000	0.1382	0.1350	0.0000	0.1324	10	
	11										0.0000	0.1162	0.1118	0.0000	0.1232	0.1270	11	
	12											0.0674	0.0000	0.0981	0.1123	0.0000	12	
	13												0.0581	0.0896	0.0000	0.1006	13	
	14													0.0000	0.0781	0.0786	14	
	15														0.0448	0.0000	15	
	16															0.0391	16	
	sum	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000		

6 The Asymmetric Case: $p \neq 1/2$

Let us return to the general case when $p \neq 1/2$. In Section 4, we remarked that in order to determine ${}^n p_k$, we need to partition the set of ${}^n S_k$ outcome sequences that correspond to Player k becoming the ultimate captain into component subsets within which the outcomes have fixed number of heads. Suppose that there are exactly ${}^n m_{i,k}$ outcome sequences that have exactly $i, (0 \leq i \leq n-2)$ heads and $(n-2-i)$ tails among the $(n-2)$ tosses and that lead to Player k becoming the ultimate captain. Clearly, then

$$ {}^n S_k = \sum_{i=0}^{n-2} {}^n m_{i,k} \tag{9} $$

It remains to determine the counting numbers $\{{}^n m_{i,k} : 0 \leq i \leq n-2, 1 \leq k \leq n\}$. We already know when ${}^n m_{i,k} = 0$: From Theorem 3(b), whenever $n + k \equiv 1 \pmod 3$, we have ${}^n S_k = 0$; that is, no outcome set causes Player k to become the ultimate captain. Hence, whenever $n + k \equiv 1 \pmod 3$, we have ${}^n m_{i,k} = 0$ for all $i, (0 \leq i \leq n-2)$.

To determine all values of ${}^n m_{i,k}$, we proceed by induction on n. Let us denote by ${}^n\mathbf{M}$

the $(n-1)\times n$ matrix whose (i,k)-th element is ${}^{n}m_{i,k}$ for $0 \le i \le n-2, 1 \le k \le n$.

When $n=2$, Player 1 is the ultimate captain with probability 1. So, ${}^{2}\mathbf{M} = [{}^{2}m_{0,1}, {}^{2}m_{0,2}] = [1,0]$ and ${}^{2}\mathbf{p} = (1)\cdot{}^{2}\mathbf{M} = (1,0)$, as we noted earlier. When there are $n=3$ players, the coin is tossed once. Then with probability p, the coin lands heads, Player 2 is eliminated and Player 3 becomes the ultimate captain. This is expressed by writing down a 2×3 matrix, namely ${}^{3}\mathbf{A} = \begin{pmatrix} \_ & \_ & \_ \\ 0 & \_ & 1 \end{pmatrix}$, where the top missing row indicates that the toss resulted in a head (causing us to downshift the rows by one or add one more to the already accumulated number of heads), and the bottom portion indicates that the elements/columns of ${}^{2}\mathbf{M}$ have been right shifted by two positions (mod 3) with a missing element/column in position two. Likewise, with probability q, the coin lands tails, Player 3 is eliminated and Player 2 becomes the ultimate captain. This is expressed by writing a second 2×3 matrix, namely ${}^{3}\mathbf{B} = \begin{pmatrix} 0 & 1 & \_ \\ \_ & \_ & \_ \end{pmatrix}$, where the bottom missing row indicates that we did not get a head (hence, there is no need to downshift the rows), and the top portion indicates that the two elements/columns of ${}^{2}\mathbf{M}$ have been left shifted by one position (mod 2) and a third missing element/column has been augmented. Finally, writing zeroes for all missing values, we obtain ${}^{3}\mathbf{M} = {}^{3}\mathbf{A} + {}^{3}\mathbf{B}$ and

$$ {}^{3}\mathbf{p} = p\cdot{}^{3}\mathbf{A} + q\cdot{}^{3}\mathbf{B} = p\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} + q\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} = (q,p)\cdot\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = (q,p)\cdot\ {}^{3}\mathbf{M} $$

Similarly, ${}^{4}\mathbf{M} = {}^{4}\mathbf{A} + {}^{4}\mathbf{B}$ and

$$ \begin{aligned} {}^{4}\mathbf{p} &= p(\_,q,p)\cdot{}^{4}\mathbf{A} + q\cdot(q,p,\_)\cdot{}^{4}\mathbf{B} \\ &= p\cdot(\_,q,p)\cdot\begin{pmatrix} \_ & \_ & \_ & \_ \\ 0 & \_ & 0 & 1 \\ 1 & \_ & 0 & 0 \end{pmatrix} + q\cdot(q,p,\_)\cdot\begin{pmatrix} 1 & 0 & 0 & \_ \\ 0 & 1 & 0 & \_ \\ \_ & \_ & \_ & \_ \end{pmatrix} \\ &= (q^2,qp,p^2)\cdot\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix} = (q^2,qp,p^2)\cdot\ {}^{4}\mathbf{M} \end{aligned} $$

In general, we define a $1\times(n-1)$ row-vector

$$ {}^{n}\mathbf{r} = (q^{n-2}, q^{n-3}p, \ldots, qp^{n-3}, p^{n-2}) \tag{10} $$

and let A_{*j} denote the j-th column of matrix A, and $D_j^n \circ A$ denote the reduced matrix obtained by deleting column j of matrix A. Then analogous to (6) and (3), for any $n \ge 2$, we have

$$ \begin{aligned} {}^{n+1}\mathbf{p} &= p(\_,{}^{n}\mathbf{r})\cdot{}^{n+1}\mathbf{A} + q({}^{n}\mathbf{r},\_)\cdot{}^{n+1}\mathbf{B} \\ &= p(\_,{}^{n}\mathbf{r})\cdot\begin{pmatrix} \_ & \_ & \_ \\ {}^{n}M_{*n} & \mathbf{0} & D_n^n\circ{}^{n}\mathbf{M} \end{pmatrix} + q({}^{n}\mathbf{r},\_)\cdot\begin{pmatrix} D_1^n\circ{}^{n}\mathbf{M} & {}^{n}M_{*1} & \mathbf{0} \\ \_ & \_ & \_ \end{pmatrix} \\ &= {}^{n+1}\mathbf{r}\cdot\ {}^{n+1}\mathbf{M} \end{aligned} $$

where

$$ {}^{n+1}\mathbf{M} \;=\; {}^{n+1}\mathbf{A} + {}^{n+1}\mathbf{B} = \begin{pmatrix} \text{—} & \text{—} & \text{—} \\ {}^{n}M_{*n} & \mathbf{0} & D_n^n \circ {}^{n}\mathbf{M} \end{pmatrix} + \begin{pmatrix} D_1^n \circ {}^{n}\mathbf{M} & {}^{n}M_{*1} & \mathbf{0} \\ \text{—} & \text{—} & \text{—} \end{pmatrix} \quad (11) $$

In view of (11), we can **recursively compute** the matrices ${}^{n}\mathbf{M}$, consisting of pure numbers or coefficients. Then we can recover ${}^{n}\mathbf{p} = {}^{n}\mathbf{r} \cdot {}^{n}\mathbf{M}$, where ${}^{n}\mathbf{r}$ is given in (10). We do that in Table 3 for $n \leq 16$, and give the R codes in the Appendix.

Table 3 is on the next page

Some interesting properties of matrices ${}^{n}\mathbf{M}$ are listed below. These can be proved by mathematical induction on n.

Theorem 6. *The matrices ${}^{n}\boldsymbol{M}$ displayed in Table 3, exhibit the following properties:*

a) *The columns of ${}^{n+1}\boldsymbol{M}$ are obtained from the columns of ${}^{n}\boldsymbol{M}$ by using (11).*

b) *For $1 \leq k \leq n$, the k-th column of ${}^{n}M$ is $\boldsymbol{0}$ iff $n + k \equiv 1 \pmod 3$. In particular, ${}^{2m}M_{*,(m+1)} = \boldsymbol{0}$, and ${}^{3m}M_{*,1} = \boldsymbol{0}$.*

c) *For $0 \leq i \leq n-2$, ${}^{n}m_{i,1} = {}^{n}m_{n-2-i,1}$; that is, the first column of ${}^{n}\boldsymbol{M}$ is a palindrome. This is a consequence of Lemma 1.*

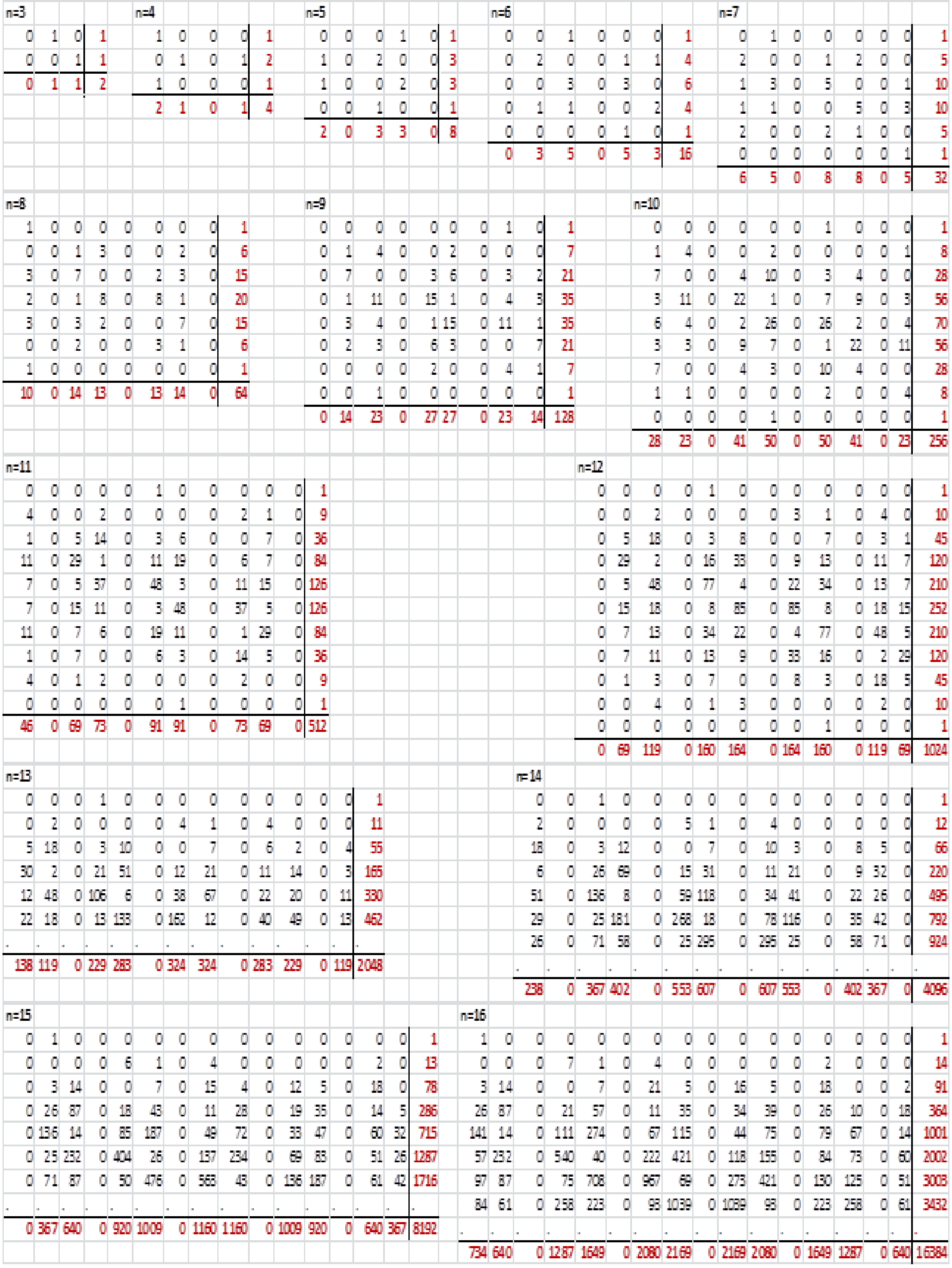
d) *For $0 \leq i \leq n-2$ and $2 \leq k \leq n$, ${}^{n}m_{i,k} = {}^{n}m_{n-2-i,n+2-k}$; that is, the matrix $D_1^n \circ {}^{n}\boldsymbol{M}$ remains invariant under a 180° rotation, or the elements of $D_1^n \circ {}^{n}\boldsymbol{M}$ read column by column (or row by row) form a palindrome. This is a consequence of Lemma 1. Therefore, for $n \geq 13$, we did not print the lower half of ${}^{n}\boldsymbol{M}$.*

e) *The column-vector of row sums of ${}^{n}\boldsymbol{M}$ yield the binomial coefficients in the expansion of $(a+b)^{n-2}$, because each row represents the number of heads among $(n-2)$ tosses.*

f) *The row-vector of column sums of ${}^{n}\boldsymbol{M}$ agree with ${}^{n}\boldsymbol{S}$, the columns of Table 1. This holds in view of (9) and because each column k signifies that Player k is the ultimate captain.*

What is the effect of a biased coin on the probability distribution of the ultimate captain among n players? We show it in Table 4 below for $p = .50(.05)1.00$. For values of $p < .5$,

Table 3. Computing ${}^n\mathbf{M} = [{}^n m_{i,k} : 0 \leq i \leq n-2, 1 \leq k \leq n]$ using (11) for $n \leq 16$, and showing the row and the column sums of each ${}^n\mathbf{M}$. See R codes in the Appendix. Since ${}^nM_{*1}$ is a palindrome, and $D_1^n \circ {}^n\mathbf{M}$ is invariant under a 180^o rotation, we need not display the lower 'half' of ${}^n\mathbf{M}$.



n=3

0	1	0	1
0	0	1	1
0	1	1	2

n=4

1	0	0	0	1
0	1	0	1	2
1	0	0	0	1
2	1	0	1	4

n=5

0	0	0	1	0	1
1	0	2	0	0	3
1	0	0	2	0	3
0	0	1	0	0	1
2	0	3	3	0	8

n=6

0	0	1	0	0	0	1
0	2	0	0	1	1	4
0	0	3	0	3	0	6
0	1	1	0	0	2	4
0	0	0	0	1	0	1
0	3	5	0	5	3	16

n=7

0	1	0	0	0	0	0	1
2	0	0	1	2	0	0	5
1	3	0	5	0	0	1	10
1	1	0	0	5	0	3	10
2	0	0	2	1	0	0	5
0	0	0	0	0	0	1	1
6	5	0	8	8	0	5	32

n=8

1	0	0	0	0	0	0	0	1
0	0	1	3	0	0	2	0	6
3	0	7	0	0	2	3	0	15
2	0	1	8	0	8	1	0	20
3	0	3	2	0	0	7	0	15
0	0	2	0	0	3	1	0	6
1	0	0	0	0	0	0	0	1
10	0	14	13	0	13	14	0	64

n=9

0	0	0	0	0	0	0	1	0	1
0	1	4	0	0	2	0	0	0	7
0	7	0	0	3	6	0	3	2	21
0	1	11	0	15	1	0	4	3	35
0	3	4	0	1	15	0	11	1	35
0	2	3	0	6	3	0	0	7	21
0	0	0	0	2	0	0	4	1	7
0	0	1	0	0	0	0	0	0	1
0	14	23	0	27	27	0	23	14	128

n=10

0	0	0	0	0	0	1	0	0	0	1
1	4	0	0	2	0	0	0	0	1	8
7	0	0	4	10	0	3	4	0	0	28
3	11	0	22	1	0	7	9	0	3	56
6	4	0	2	26	0	26	2	0	4	70
3	3	0	9	7	0	1	22	0	11	56
7	0	0	4	3	0	10	4	0	0	28
1	1	0	0	0	0	2	0	0	4	8
0	0	0	0	1	0	0	0	0	0	1
28	23	0	41	50	0	50	41	0	23	256

n=11

0	0	0	0	0	1	0	0	0	0	0	1
4	0	0	2	0	0	0	0	2	1	0	9
1	0	5	14	0	3	6	0	0	7	0	36
11	0	29	1	0	11	19	0	6	7	0	84
7	0	5	37	0	48	3	0	11	15	0	126
7	0	15	11	0	3	48	0	37	5	0	126
11	0	7	6	0	19	11	0	1	29	0	84
1	0	7	0	0	6	3	0	14	5	0	36
4	0	1	2	0	0	0	0	2	0	0	9
0	0	0	0	0	0	1	0	0	0	0	1
46	0	69	73	0	91	91	0	73	69	0	512

n=12

0	0	0	0	1	0	0	0	0	0	0	0	1
0	0	2	0	0	0	0	3	1	0	4	0	10
0	5	18	0	3	8	0	0	7	0	3	1	45
0	29	2	0	16	33	0	9	13	0	11	7	120
0	5	48	0	77	4	0	22	34	0	13	7	210
0	15	18	0	8	85	0	85	8	0	18	15	252
0	7	13	0	34	22	0	4	77	0	48	5	210
0	7	11	0	13	9	0	33	16	0	2	29	120
0	1	3	0	7	0	0	8	3	0	18	5	45
0	0	4	0	1	3	0	0	0	0	2	0	10
0	0	0	0	0	0	0	0	1	0	0	0	1
0	69	119	0	160	164	0	164	160	0	119	69	1024

n=13

0	0	0	1	0	0	0	0	0	0	0	0	0	1
0	2	0	0	0	0	4	1	0	4	0	0	0	11
5	18	0	3	10	0	0	7	0	6	2	0	4	55
30	2	0	21	51	0	12	21	0	11	14	0	3	165
12	48	0	106	6	0	38	67	0	22	20	0	11	330
22	18	0	13	133	0	162	12	0	40	49	0	13	462
.	.	.	.	.	.	.	.	.	.	.	.	.	.
138	119	0	229	283	0	324	324	0	283	229	0	119	2048

n=14

0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
2	0	0	0	0	5	1	0	4	0	0	0	0	0	12
18	0	3	12	0	0	7	0	10	3	0	8	5	0	66
6	0	26	69	0	15	31	0	11	21	0	9	32	0	220
51	0	136	8	0	59	118	0	34	41	0	22	26	0	495
29	0	25	181	0	268	18	0	78	116	0	35	42	0	792
26	0	71	58	0	25	295	0	295	25	0	58	71	0	924
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
238	0	367	402	0	553	607	0	607	553	0	402	367	0	4096

n=15

0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	6	1	0	4	0	0	0	0	0	2	0	13
0	3	14	0	0	7	0	15	4	0	12	5	0	18	0	78
0	26	87	0	18	43	0	11	28	0	19	35	0	14	5	286
0	136	14	0	85	187	0	49	72	0	33	47	0	60	32	715
0	25	232	0	404	26	0	137	234	0	69	83	0	51	26	1287
0	71	87	0	50	476	0	563	43	0	136	187	0	61	42	1716
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
0	367	640	0	920	1009	0	1160	1160	0	1009	920	0	640	367	8192

n=16

1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	7	1	0	4	0	0	0	0	0	2	0	0	0	14
3	14	0	0	7	0	21	5	0	16	5	0	18	0	0	2	91
26	87	0	21	57	0	11	35	0	34	39	0	26	10	0	18	364
141	14	0	111	274	0	67	115	0	44	75	0	79	67	0	14	1001
57	232	0	540	40	0	222	421	0	118	155	0	84	73	0	60	2002
97	87	0	75	708	0	967	69	0	273	421	0	130	125	0	51	3003
84	61	0	258	223	0	93	1039	0	1039	93	0	223	258	0	61	3432
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
734	640	0	1287	1649	0	2080	2169	0	2169	2080	0	1649	1287	0	640	16384

shown in the second row, holding Player 1 fixed, reverse the serial numbers of Players $2, 3, \ldots, n-1, n$ as $n, n-1, \ldots, 3, 2$, shown in the rightmost column, and replace p by $q = 1 - p > .5$, shown in the top row.

Table 4. The probability mass functions $\{{}^n p_k : 1 \leq k \leq n\}$ for $3 \leq n \leq 9$. For $p = .50(.05)1.00$, read [k] in the left margin; for $p < .5$, read [k] in the right margin.

```
  p->   .50    .55    .60    .65    .70    .75    .80    .85    .90    .95  1.00
n=3            .45    .40    .35    .30    .25    .20    .15    .10
.05  0.00   <- p
 [1] 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000      0     [1]
 [2] 0.5000 0.4500 0.4000 0.3500 0.3000 0.2500 0.2000 0.1500 0.1000 0.0500      0     [3]
 [3] 0.5000 0.5500 0.6000 0.6500 0.7000 0.7500 0.8000 0.8500 0.9000 0.9500      1     [2]
n=4
 [1] 0.5000 0.5050 0.5200 0.5450 0.5800 0.6250 0.6800 0.7450 0.8200 0.9050      1     [1]
 [2] 0.2500 0.2475 0.2400 0.2275 0.2100 0.1875 0.1600 0.1275 0.0900 0.0475      0     [4]
 [3] 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000      0     [3]
 [4] 0.2500 0.2475 0.2400 0.2275 0.2100 0.1875 0.1600 0.1275 0.0900 0.0475      0     [2]
n=5
 [1] 0.2500 0.2475 0.2400 0.2275 0.2100 0.1875 0.1600 0.1275 0.0900 0.0475      0     [1]
 [2] 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000      0     [5]
 [3] 0.3750 0.3891 0.4080 0.4339 0.4690 0.5156 0.5760 0.6524 0.7470 0.8621      1     [4]
 [4] 0.3750 0.3634 0.3520 0.3386 0.3210 0.2969 0.2640 0.2201 0.1630 0.0904      0     [3]
 [5] 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000      0     [2]
n=6
 [1] 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000      0     [1]
 [2] 0.1875 0.1751 0.1632 0.1519 0.1407 0.1289 0.1152 0.0979 0.0747 0.0431      0     [6]
 [3] 0.3125 0.2996 0.2848 0.2664 0.2433 0.2148 0.1808 0.1414 0.0973 0.0496      0     [5]
 [4] 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000      0     [4]
 [5] 0.3125 0.3254 0.3408 0.3616 0.3913 0.4336 0.4928 0.5736 0.6813 0.8214      1     [3]
 [6] 0.1875 0.1999 0.2112 0.2201 0.2247 0.2227 0.2112 0.1871 0.1467 0.0859      0     [2]
n=7
 [1] 0.1875 0.1887 0.1920 0.1962 0.1995 0.1992 0.1920 0.1737 0.1395 0.0837      0     [1]
 [2] 0.1562 0.1348 0.1139 0.0932 0.0730 0.0537 0.0362 0.0212 0.0097 0.0025      0     [7]
 [3] 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000      0     [6]
 [4] 0.2500 0.2427 0.2342 0.2253 0.2159 0.2051 0.1907 0.1692 0.1354 0.0820      0     [5]
 [5] 0.2500 0.2547 0.2554 0.2502 0.2377 0.2168 0.1869 0.1483 0.1022 0.0515      0     [4]
 [6] 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000      0     [3]
 [7] 0.1562 0.1790 0.2045 0.2351 0.2739 0.3252 0.3942 0.4876 0.6132 0.7803      1     [2]
n=8
 [1] 0.1562 0.1591 0.1683 0.1854 0.2136 0.2573 0.3226 0.4176 0.5528 0.7414      1     [1]
 [2] 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000      0     [8]
 [3] 0.2188 0.2130 0.2089 0.2064 0.2044 0.2007 0.1917 0.1730 0.1391 0.0836      0     [7]
 [4] 0.2031 0.1888 0.1705 0.1482 0.1224 0.0945 0.0663 0.0403 0.0190 0.0049      0     [6]
 [5] 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000      0     [5]
 [6] 0.2031 0.2140 0.2223 0.2287 0.2333 0.2351 0.2314 0.2170 0.1831 0.1169      0     [4]
 [7] 0.2188 0.2250 0.2300 0.2313 0.2263 0.2124 0.1879 0.1521 0.1060 0.0531      0     [3]
 [8] 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000      0     [2]
```

```
n=9
 [1] 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000     0    [1]
 [2] 0.1094 0.0959 0.0836 0.0722 0.0613 0.0502 0.0383 0.0260 0.0139 0.0042     0    [9]
 [3] 0.1797 0.1725 0.1692 0.1724 0.1863 0.2166 0.2714 0.3610 0.4994 0.7046     1    [8]
 [4] 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000     0    [7]
 [5] 0.2109 0.2135 0.2143 0.2142 0.2131 0.2093 0.1997 0.1796 0.1435 0.0853     0    [6]
 [6] 0.2109 0.2051 0.1943 0.1773 0.1536 0.1240 0.0906 0.0570 0.0277 0.0073     0    [5]
 [7] 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000     0    [4]
 [8] 0.1797 0.1893 0.2007 0.2136 0.2274 0.2407 0.2497 0.2471 0.2201 0.1482     0    [3]
 [9] 0.1094 0.1238 0.1380 0.1503 0.1584 0.1593 0.1503 0.1293 0.0954 0.0504     0    [2]
```

Let us explain how to compute the probability mass function of the ultimate captain in a neighbor-elimination random walk involving a biased coin. Suppose that $n = 10$ and $p = .6$. Then the probability that Player k (for $1 \leq k \leq 10$) is the ultimate captain is given by ${}^{10}\mathbf{p} = {}^{10}\mathbf{r} \cdot {}^{10}\mathbf{M}$ and the computation is shown below (see R codes in the Appendix):

```
^{10}p (p=.6) = (.1162,  .0677,  0,  .1358,  .1792,  0,  .2088,  .1718,  0,  .1204)
```

To see the effect of the biased coin ($p = .6$), contrast this probability vector against that in the symmetric case (when $p = 1/2$) given by ${}^{10}\mathbf{p} = {}^{10}\mathbf{S}/2^8$ and shown below:

```
^{10}p (p=.5) = (.1094,  .0898,  0,  .1602,  .1953,  0,  .1953,  .1602,   0,  .0898)
```

Notice that the probability distribution is symmetric for $p = .5$, but not for $p \neq .5$. Also, notice that any player who has no chance of being the ultimate captain for $p = .5$ (that is, Players 3, 6, 9), still has no chance of being the ultimate captain for $p \neq .5$. Thus, a biased coin does not change the fate of losers.

Finally, we compute ${}^{10}p_1$ and ${}^{10}p_5$, the probability that Player 1 or Player 5 is the ultimate captain, for some values of p closer and closer to 1. The computations show that ${}^{10}p_1 \to 0$ and ${}^{10}p_5 \to 1$ as $p \to 1$, which we already know as a corollary to Theorem 1.

```
         p  ->      .90      .95      .99      .999     .9999     .99999
^{10}p_1  ->    .0872    .0481    .0100    .0010    .0001    .0000
^{10}p_5  ->    .4523    .6697    .9230    .9920    .9992    .9999
```

7 Summary and Open Problems

We have studied a randomized version of the Josephus Problem when $d = 2$. That is, starting with Player 1, either the clockwise neighbor is eliminated with probability $p < 1$, or the counterclockwise neighbor is eliminated with probability $1 - p$. We have recursively calculated the probability distribution of the last survivor. In particular, the survival probability is 0 for every third position starting from 1, 2, or 3 according to $n \pmod 3$ is 0, 2, or 1, respectively. To determine which position has the best chance of survival, one can compute the probability mass function ${}^n\mathbf{p} = {}^n\mathbf{r} \cdot {}^n\mathbf{M}$, where ${}^n\mathbf{r}$ is given in (10) and ${}^n\mathbf{M}$ is given in Table 3. As claimed in Theorem 6(f), Table 1 can be recovered from Table 3 by taking the column sums of ${}^n\mathbf{M}$.

Interested readers may identify other properties in Tables 1 and 3. Astute readers can amend the recursive algorithm to study an elimination random walk in which either the d-th neighbor in the clockwise direction is eliminated with probability p, or the d-th neighbor in the counterclockwise direction is eliminated with probability $q = 1-p$, where $d \geq 3$, and the captain's hat moves to the next player in the same direction beyond the eliminated player. As an illustration, we show the R codes for $d = 3$ in the Appendix. A more ambitious reader may study an elimination random walk in which one member of a designated subset is eliminated with an associated probability mass function. For instance, starting from Player 1 wearing the captain's hat, eliminate one of the players $\{2, 3, n\}$ with probabilities (p, q, r), respectively, and move the captain's hat to Player 3, 4 or $(n - 1)$, respectively.

Research-minded readers may try to discover a mathematical formula for the probability mass function ${}^n\mathbf{p}$ in Sections 5 and 6, and in any of the more general elimination random walk problems mentioned in the above paragraph.

Acknowledgment

The second author thanks Arizona State University where he and Professor Athreya were visiting professors during which time the problem was formulated.

Bibliography

[1] Levitin, Anany and Levitin, Maria, *Algorithmic Puzzles*, Oxford University Press, 2011, Puzzle #141.

[2] Halbeisen, Lorenz and Hungerbuhler, Norbert (1997), The Josephus Problem, *Journal de Théorie des Nombres de Bordeaux* **9(2)**, 303–318.

[3] Herstein, I. N. and Kaplansky, Irving, *Matters Mathematical*, New York: Chelsea Publishing, 1978.

[4] Josephus, Flavius, *The Jewish War, Book III*, Translated by H. S. Thackeray, Heinemann (1927), 341–366, 387–391.

[5] Maiti, Saran Ishika and Sarkar, Jyotirmoy (2019), Symmetric walks on paths and cycles, *Mathematics Magazine*, **92(4)**, 252-268, DOI: 10.1080/0025570X.2019.1611166

[6] R Core Team (2021), R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[7] Robinson, W. J. (1960), The Josephus Problem, *Math Gazette* **44(347)** 47–52. DOI: https://doi.org/10.2307/3608532

[8] Sarkar, Jyotirmoy (2006), Random walk on a polygon, In: Recent Developments in Nonparametric Inference and Probability, J. Sun, A. DasGupta, V. Melfi, C. Page, Eds., *IMS Lecture Notes–Monograph Series*, **50**, 31-43, Beachwood, OH: Inst. Math. Statist.

[9] Singmaster, David, *Assorted Articles on Recreational Mathematics and the History of Mathematics*. https://www.puzzlemuseum.com/singma/singma-index.htm

Appendix

Here are the codes (using freeware R) to compute Table 1, Table 3 and Table 4.

```
### If the neighbor to the left or right is eliminated (d=2)
### Table 1. When p=1/2, compute ^{n}S iteratively
### PMF = ^{n}S/2^{n-2}
n=16    # change to any positive integer
a=c(1, 0)
for (i in 3:n){
   l=length(a); l1=l-1
   bm=c(a[2:l],  a[1], 0)
   bp=c(a[l], 0, a[1:l1])
   a=bm+bp
print(c(i, a, sum(a)))  # sum = 2^{n-2} }

### Table 3: When p != 1/2 compute ^{n}M iteratively
n=10    # Change to any positive integer
a=matrix(c(1, 0), nrow=1)
for (i in 3:n){   l=ncol(a)
   zc=matrix(rep(0,l-1), ncol=1)
   zr=matrix(rep(0,l+1), nrow=1)
   bm=cbind(a[,2:l], a[,1], zc)
   bp=cbind(a[,l], zc, a[,1:l-1])
   a=rbind(bm,zr) + rbind(zr,bp)
   lc=matrix(rep(l+1,l), ncol=1)
print( a )  # matrix M
oner=matrix(rep(1,l ), nrow=1)
onec=matrix(rep(1,l+1), ncol=1)
print(c(l+1, oner%*%a, sum(oner%*%a))) # col sum
print(a%*%onec) # row sum = binomial coeff
}

## Table 4: Compute PMF = ^{n}r * ^{n}M when n=10 and p>1/2
p=.6; q=1-p     # Change the value of p as needed
r=c(q^(8), q^7)*p, q^(6)*p^(2), q^(5)*p^(3), q^(4)*p^(4),
    q^(3)*p^(5), q^(2)*p^(6), q^(1)*p^(7), p^(8) )
round(r%*%a, 4)
round(print(oner%*%a)/2^(8),4)

### Discussion: If the 3rd person on the left or right is eliminated (d=3)
### when p=1/2, compute ^{n}S iteratively
### PMF = ^{n}S/2^{n-2}
n=16    # change to any positive integer
a=c(0, 1, 1)
for (i in 4:n){
   l=length(a); l1=l-1; l2=l-2
  bm=c(a[3:l], a[1], 0, a[2])
  bp=c(a[l1:l], 0, a[1:l2])
  a=bm+bp
  print(c(i, a, sum(a)))    # sum = 2^{n-2}
}
```

11 Historical Roots of Calculus – 5: James Gregory

Shailesh Shirali

The Valley School KFI,
Thatguni Post, Kanakapura Road,
Bengaluru – 560082, India.
Email : shailesh.shirali@gmail.com

In past instalments of this series, we have studied the works of many remarkable individuals — Roberval, Descartes, Fermat, Mercator, Leibniz, ...— and the relationship of their work with the calculus we know and use today. Now we shall take up the work of the Scottish mathematician James Gregory (1638–1675) whose name is permanently associated with the series

$$\frac{\pi}{4} = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{2n-1} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots . \qquad (1)$$

As described in the previous instalment of this article, Leibniz discovered the same series. So did the Kerala mathematician Madhava (two centuries earlier); this will be the subject of a later article in this series. These discoveries were all made independently.

The work of James Gregory

James Gregory was a mathematician of considerable native ability. He died young, but in his short life he accomplished a great deal. In [2] we read:

> *For a long time the light of James Gregory did not shine as brightly as did that of John Wallis, Isaac Barrow and Isaac Newton, the other three great British mathematicians of the seventeenth century. Only recently, through the endeavours of several Scottish mathematicians, ...Gregory's genius is revealed and fills with admiration all those interested in the development of modern mathematics.*

And in [4]:

> *Most of James Gregory's research notes were lost ...after his premature death in 1675 in Edinburgh at the age of 36, and thus the mathematician who was second only to Newton in the early 1670s was almost forgotten for more than two and a half centuries. Not until 1939 were many of his important contributions to mathematics revealed to the world ...It was then made clear that Gregory had proved the rules of differentiation, that he was in possession of the chain rule before Leibniz, ...that he had discovered Taylor's theorem more than forty years before Taylor, and that he had derived an important interpolation formula, and then from it derived the binomial theorem before Newton made it public (but Newton knew it first).*

We are clearly dealing here with a major figure in the history of mathematics. One reason for his not being as well-known as might have been is that many of his findings lay hidden in private letters or unpublished notes. Some of these notes have been found scribbled in the blank spaces of letters he received from correspondents! (One such was John Collins (1625–1683); he played a vital role in making widely known the discoveries made by mathematicians in the UK and Europe, much like Marin Mersenne (1588–1648) did in Europe.) Fortunately, these letters are now well-preserved in library archives.

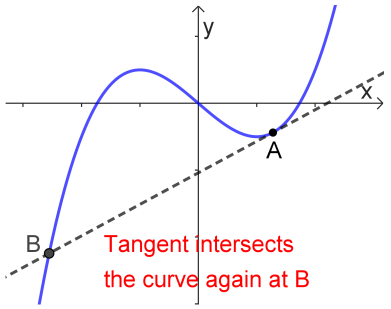
We shall describe three of these findings in this article: his discovery of the Fundamental Theorem of Calculus (FTC for short), his work on Taylor series (from which (1) follows immediately), and his work on Interpolation and on the Binomial Theorem.

We remark at the outset that it is not easy for us to understand Gregory's work. In keeping with the practice of his time, he wrote everything in the language of Euclidean geometry. (Keep in mind that the formalism of limits appeared much later.) Moreover, much of the writing is in long-form prose. Brought up as we are on the pithy language of limits, with everything couched in compact algebra, we find the notation and verbose arguments challenging to follow. But we shall soldier on!

James Gregory's formulation of the FTC in the language of Euclidean geometry

Gregory's analysis clearly reveals the assumptions he makes about the nature of functions and their associated curves. Firstly, the curves are strictly rising, and they pass through the origin.

More crucial is the way he regards a tangent to a curve. Extrapolating from what a tangent to a circle looks like, the early notion of tangency was this: *A tangent to a* *curve touches the curve at just one point, and the curve lies entirely on one side of the tangent.*



Clearly, Gregory's thinking was consistent with this viewpoint.

But this viewpoint ignores possibilities like the one shown here (where the tangent intersects the curve again at another point).

This is a serious limitation of the analysis.

Figure 1: Early notions of tangent to a curve

These assumptions are not stated explicitly; they are assumed implicitly to be so. That is simply how curves and tangents were thought of at the time.

Figure 2 shows a pair of curves $y = f(x)$ and $y = F(x)$. The curve $y = f(x)$ is arbitrary, but as noted above, Gregory wants it to be strictly increasing and passing through the origin O.

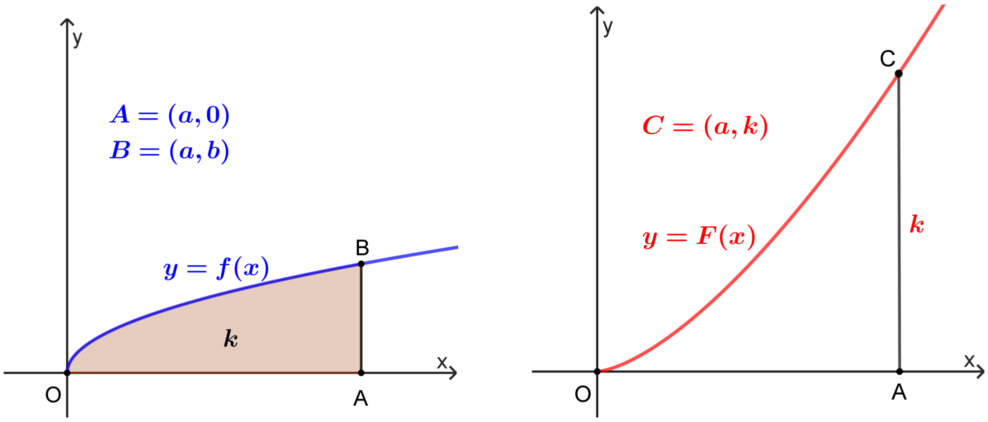


Figure 2: Statement of the fundamental theorem of calculus (FTC)

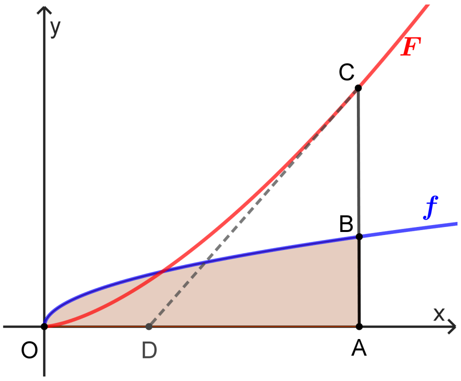
The curve $y = F(x)$ is defined as follows. Let AB be an arbitrary ordinate of the curve $y = f(x)$; say $A = (a, 0)$ and $B = (a, b)$. Let k be the area of the region (shown shaded) enclosed by the curve, the x-axis and the ordinate AB; then the point C has coordinates (a, k). Expressed in modern terms,

$$F(x) = \int_0^x f(t)\,dt. \tag{2}$$

If $a = 0$ the area is zero, so the curve $y = F(x)$ passes through the origin. The FTC states that

$$F'(x) = f(x). \tag{3}$$

Geometrically, this amounts to the claim shown in Figure 3.



Let the tangent at C to the curve $y = F(x)$ intersect the x-axis at D.

Then the claim is that

$$\frac{CA}{DA} = AB.$$

Figure 3: Gregory's proof of the FTC. Here, F is the area-under-f curve, in the sense that $CA = \text{Area}(OAB)$, etc.

Gregory proceeds as follows. To prove that $\frac{CA}{DA} = AB$, he supposes that the line through C with slope AB is **not** tangent to the F-curve. This is ultimately going to lead to a contradiction. (Here, he draws on the old notion of tangency described above.) The fact that this is essentially a proof by contradiction illustrates how Gregory is drawing on the traditions of Euclidean geometry.

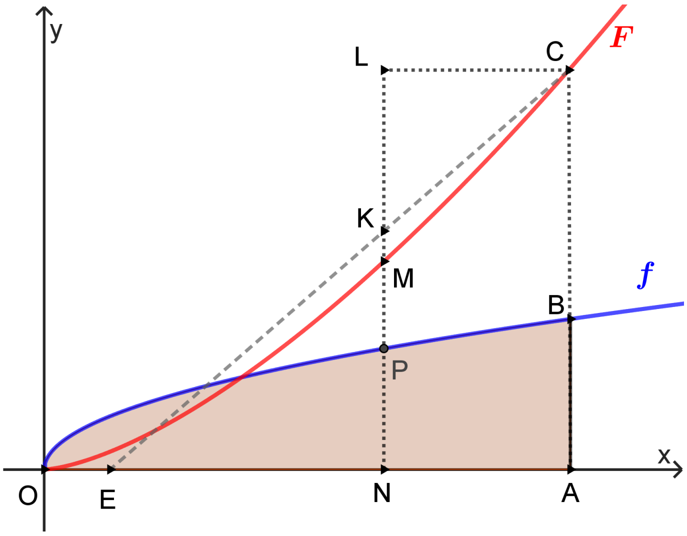


Figure 4: CE is the line through C whose slope is equal to AB.

As line CE is not a tangent to the F-curve, it "cuts" across the curve and passes through some point K that lies *above* the curve. The picture is as shown again in Figure 4. (We have repeated this figure as it is referred to many times.)

We now invoke the definition of the F-curve repeatedly and argue as follows. We have:

$$\begin{aligned} CA &= \text{Area}(OABO), \\ MN &= \text{Area}(ONPO), \\ \therefore\quad \frac{MN}{CA} &= \frac{\text{Area}(ONPO)}{\text{Area}(OABO)}, \qquad (4) \\ \therefore\quad \frac{KN}{CA} &> \frac{\text{Area}(ONPO)}{\text{Area}(OABO)} \qquad \text{(since } KN > MN). \qquad (5) \end{aligned}$$

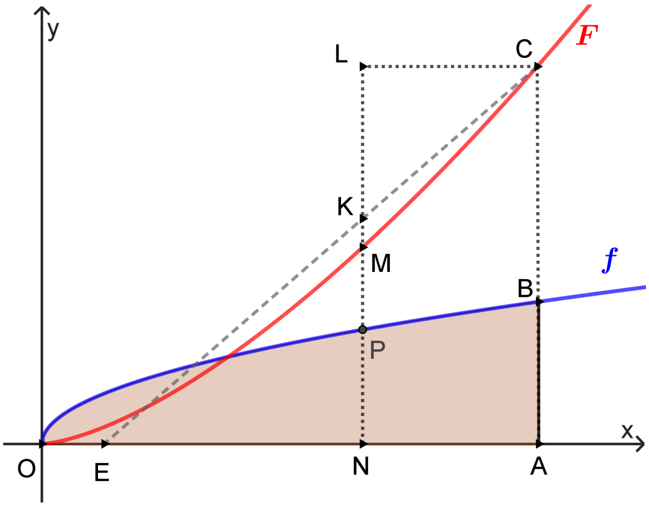


Figure 5:

Drawing on the similarity of triangles KEN and CEA, we have:

$$\frac{KN}{CA} = \frac{EN}{EA}, \qquad \therefore\ \frac{EN}{EA} > \frac{\text{Area}(ONPO)}{\text{Area}(OABO)}. \qquad (6)$$

Hence

$$\frac{EN \cdot AB}{EA \cdot AB} > \frac{\text{Area}(ONPO)}{\text{Area}(OABO)}. \qquad (7)$$

It is not difficult to read Gregory's mind when he writes (7) in place of (6) — a step which otherwise might look strange. The fraction on the left side of (6) features *lengths*, whereas the fraction on the right side of (6) features *areas*. He would like the left side to feature only areas, so that it will be possible to compare the quantities on the two sides directly. Hence the multiplication of both numerator and denominator by AB.

□

From (7) follows:

$$
\begin{aligned}
1 - \frac{EN \cdot AB}{EA \cdot AB} &< 1 - \frac{\text{Area}(ONPO)}{\text{Area}(OABO)}, \\
\therefore \quad \frac{NA \cdot AB}{EA \cdot AB} &< \frac{\text{Area}(PNAB)}{\text{Area}(OABO)}. \qquad (8)
\end{aligned}
$$

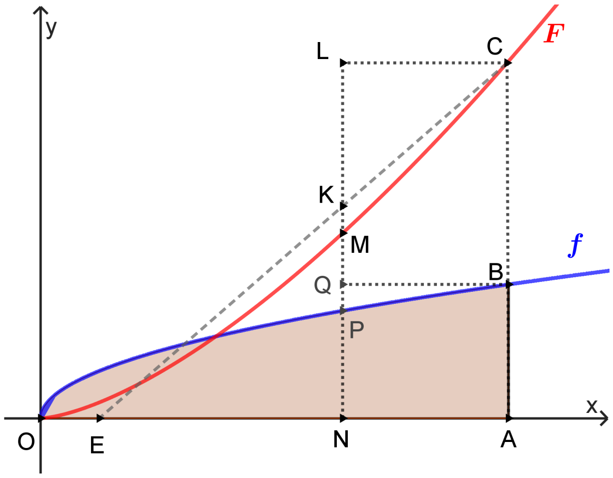


Figure 6:

Now recall the construction of line CE: its slope is equal to AB. That is,

$$\frac{CA}{EA} = AB, \qquad \therefore \; EA \cdot AB = CA. \qquad (9)$$

Hence (8) reduces to

$$\frac{NA \cdot AB}{CA} < \frac{\text{Area}(PNAB)}{\text{Area}(OABO)}. \qquad (10)$$

But by the definition of the F-curve, $CA = \text{Area}(OABA)$. So the denominators on the two sides of (10) are equal! This implies that

$$
\begin{aligned}
NA \cdot AB &< \text{Area}(PNAB), \\
\text{that is,} \quad \text{Area}(QNAB) &< \text{Area}(PNAB). \qquad (11)
\end{aligned}
$$

But this inequality is not possible, given the increasing nature of f (true by assumption). So we reach a contradiction, which has arisen because of what we had supposed at the start — that the line through C with slope AB is **not** tangent to the F-curve. Withdrawing the supposition, we conclude that the line through C with slope AB **is** tangent to the F-curve. This is just the statement of the fundamental theorem of calculus.

Comments on the proof

There are a few obvious limitations of Gregory's approach. We list these below with a few comments.

- He assumes (as noted earlier) that f is an increasing function. This naturally limits the scope of the work.

  Note that if f is an increasing function, then F is convex. This ensures the existence of a suitable point K on the tangent line (i.e., a point on CE that lies *above* the F-curve). Gregory is well aware of this fact. The increasing nature of f is thus pivotal to his proof.

  On the other hand, the requirement that $f(0) = 0$ does not lead to any loss of generality.

- The analysis presented above covers the case when AB is less than the slope at C. A similar analysis can be carried out for the case when AB is greater than the slope at C. The diagram will be slightly different now but much the same reasoning works in this case too.

- Gregory's proof gives the impression of being dependent on the diagram. While to some extent this seems natural (after all, such thinking has been part of mathematical culture for a long time), we must also note that the practice of "arguing from a diagram" or being overly dependent on a diagram is something to be wary of. Over the past few centuries, many geometrical fallacies have been devised that highlight the danger of this habit.

- We remarked earlier on Gregory's understanding of the notion of tangency. This surely cannot be described as his contribution; it is a view of tangency that he has inherited from Greek times.

- We may view the above as limitations of Gregory's approach, but during his time, such reasoning would have been highly regarded and not flawed at all. Standards of rigour have changed greatly over the centuries. In relation to the topic we are studying here, let us keep in mind that the notion of limit took well over a century before taking the form it has today.

- Gregory is clearly a gifted geometer — he relishes the geometrical approach! In this too, he is following an ancient Greek tradition.

- Lastly, we remark that it is not clear whether Gregory realized the significance and generality of his finding. There is no evidence that he tried to apply the result in more general settings. In contrast, Newton and Leibniz seemed fully aware of the power of their findings and made full use of them. This fact alone has contributed to Gregory's not being as well-known a figure as he should be.

Gregory's work on power series

The theorem we refer to today as "Taylor's theorem" on power series — and the fact that some functions can be represented as power series in the independent variable — was discovered by Gregory much before the discoveries made by Brook Taylor (1685–1731) and by Colin Maclaurin (1698–1746).

The manner in which this came about is quite curious ([9]):

> *In late 1670, James Gregory was shown in a letter from John Collins several Maclaurin series (*$\sin x$, $\cos x$, $\arcsin x$ *and* $x \cot x$*) derived by Isaac Newton, and told that Newton had developed a general method for expanding functions in series. Newton had in fact used a cumbersome method involving long division of series and term-by-term integration, but Gregory did not know it and set out to discover a general method for himself. In early 1671 Gregory discovered something like the general Maclaurin series and sent a letter to Collins including series for* $\arctan x$, $\tan x$, $\sec x$, $\ln \sec x$, *...*$\ln \tan \frac{1}{2}(\frac{1}{2}\pi + x)$*, ... However, thinking that he had merely redeveloped a method by Newton, Gregory never described how he obtained these series, and it can only be inferred that he understood the general method by examining scratch work he had scribbled on the back of another letter from 1671.*

The available evidence suggests the following.

- Perhaps prompted by the news from Collins about Mercator's derivation of the logarithmic series,

  $$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots , \tag{12}$$

  Gregory most likely did something similar for the function $1/(1+x^2)$ (i.e., he must have used term-by-term integration and obtained two different expressions for the area under the curve between $x = 0$ and $x = 1$) and thereby arrived at the series for the arctangent:

  $$\arctan x = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \cdots . \tag{13}$$

  From this, the expression for $\pi/4$ is immediate, but Gregory does not bother to write it down explicitly.

- Alongside the arctangent series, Gregory also gives (correctly, except for a small computational error in one series; but this appears to be just a copying error) the power series for a number of other functions ($\tan x$, $\sec x$, $\ln \sec x$, ...$\ln \tan \frac{1}{2}(\frac{1}{2}\pi + x)$, ...). These could not have been obtained in the same manner (using term-by-term integration), so he must have possessed a more general method for tackling such functions. An examination of his scribblings on letters and envelopes suggests that he was using an approach much like what Taylor used more than forty years later; i.e., through the use of successive derivatives; for, on one of these letters, a table is found giving the successive derivatives of just those functions for which he had given the power series, and with just as many derivatives as are needed for the terms that he had computed.

  We remark in passing that Gregory must have been (among many other things) also a pioneer in the art of back-of-the-envelope calculations!

Gregory's work on interpolation

Closely related to his work on power series expressions for functions is his work on interpolation and on the binomial theorem.

The Newton-Gregory Interpolation Formula is a method for making estimates about (in-between) function values when the input consists of values of the function at evenly-spaced data points. The formula was independently found by Gregory and Newton during 1668–1670. (Their work was most likely based on work done by John Wallis a decade earlier.) The simplest example of this is *linear interpolation*: Given a pair of function values $f(a)$ and $f(a+h)$, then for $0 < t < 1$, writing $\Delta f = f(a+h) - f(a)$,

$$f(a+th) \approx f(a) + t \cdot \Delta f. \tag{14}$$

The formula comes from approximating the portion of the f-curve joining the pair of points $(a, f(a))$ and $(a+h, f(a+h))$ by the straight line segment joining the same two points. If h is sufficiently small, then (14) can give very good results. It must have proved extremely useful in the computation of tables of logarithms and the trigonometric functions.

But we can get even better results if we have more data points and we approximate the given curve using polynomial curves of higher degree. Say we are given a data set consisting of f-values at $(n+1)$ evenly-spaced points $x_0 = a$, $x_1 = a+h$, $x_2 = a+2h$, ..., $x_n = a+nh$. Let the corresponding function values be y_0, y_1, y_2, ..., y_n. Then for $0 < t < 1$,

$$f(x_j + th) \approx y_j + \frac{t}{1!}\Delta y_j + \frac{t(t-1)}{2!}\Delta^2 y_j + \frac{t(t-1)(t-2)}{3!}\Delta^3 y_j + \cdots, \tag{15}$$

where Δy_j, $\Delta^2 y_j$, $\Delta^3 y_j$, ...are the *successive differences* (see [10] for details) given by

$$\Delta^0 y_k = y_k, \qquad k = 0, \ldots, n \tag{16}$$

$$\Delta^j y_k = \Delta^{j-1} y_{k+1} - \Delta^{j-1} y_k, \qquad k = 0, \ldots, n-j,\ j = 1, \ldots, n. \tag{17}$$

We may put (15) in a more familiar form by considering the particular case $a = 0$, $h = 1$. We are given the function values $f(0)$, $f(1)$, $f(2)$, ..., $f(n)$. We now have:

$$f(t) \approx f(0) + \frac{t}{1!}\Delta f(0) + \frac{t(t-1)}{2!}\Delta^2 f(0) + \frac{t(t-1)(t-2)}{3!}\Delta^3 f(0) + \cdots. \tag{18}$$

If f is a polynomial of degree n, then (18) is an identity. (For example, take the function $f(x) = x^2$. We have: $f(0) = 0$; $f(1) = 1$; $f(2) = 4$; $\Delta f(0) = 1$; $\Delta f(1) = 3$; $\Delta^2(0) = 2$; later differences are 0. So the sum on the right side is $0+t+t(t-1) = t^2$, as it should be.) It seems plausible to suppose that this is how the discoverers of the formula stumbled upon it.

One of those who discovered the formula was Gregory. By today's standards, he did not provide a proof as such; rather, he extrapolated from patterns seen in tables of successive differences and then showed how the formula worked, using examples.

Gregory's work on the binomial theorem

Gregory also made important contributions to the binomial theorem for fractional exponents. For the case when the exponent is a positive integer, the result was already known at the time. Gregory succeeded in finding a generalization of the theorem to fractional exponents by making clever use of his interpolation formula. Here's how he went about it. He appears to have stumbled upon the result while trying to solve the 'anti-logarithm' problem: given the logarithm x of an unknown number y to some base b, to find that number. In other words, given that $y = b^x$ (for some known b and x), to find y.

Write the base b as $1 + d$. We need to compute $(1 + d)^x$, for known d and x. Define the function f thus: $f(x) = (1 + d)^x$. Write down the 'values' of $f(x)$ at the known points, $x = 0, 1, 2, 3, \ldots$, i.e., $f(0) = 1$, $f(1) = 1+d$, $f(2) = 1+2d+d^2$, $f(3) = 1+3d+3d^2+d^3$, and so on. Since

$$\begin{aligned} f(x+1) - f(x) &= (1+d)^{x+1} - (1+d)^x \\ &= d \cdot (1+d)^x = d \cdot f(x), \end{aligned} \tag{19}$$

the successive differences $\Delta f(0)$, $\Delta^1 f(0)$, $\Delta^2 f(0)$, $\Delta^3 f(0)$, $\Delta^4 f(0)$, ...follow an extremely simple pattern (as seen in Table 1, which is full of appealing patterns):

x	$f(x) = (1+d)^x$	$\Delta f(x)$	$\Delta^2 f(x)$	$\Delta^3 f(x)$
0	1			
		d		
1	$1+d$		d^2	
		$d+d^2$		d^3
2	$1+2d+d^3$		d^2+d^3	
		$d+2d^2+d^3$		d^3+d^4
3	$1+3d+3d^2+d^3$		$d^2+2d^3+d^4$	
		$d+3d^2+3d^3+d^4$		$d^3+2d^4+d^5$
4	$1+4d+6d^2+4d^3+d^4$		$d^2+3d^3+3d^4+d^5$	

Table 11.1:

So we have,

$$f(0) = 1, \quad \Delta^1 f(0) = d, \quad \Delta^2 f(0) = d^2, \quad \Delta^3 f(0) = d^3, \quad \Delta^4 f(0) = d^4, \quad \ldots \tag{20}$$

It is an easy task (using induction) to show that $\Delta^k f(0) = d^k$ for all positive integers k.

From this, (18) yields the generalization of the binomial theorem immediately:

$$(1+d)^t = 1 + \frac{t}{1!}d + \frac{t(t-1)}{2!}d^2 + \frac{t(t-1)(t-2)}{3!}d^3 + \cdots \tag{21}$$

Observe that this derivation proceeds in a very different way from the approach we follow now in schools and colleges.

Gregory no doubt noticed that when t is not a positive integer, (21) yields an infinite series. But this was not an era where questions of convergence troubled mathematicians. As [2] notes,

> *Neither Gregory nor Newton tried to prove the convergence of the series. Such a proof was not, at this time, believed to be necessary; but certainly they had the feeling that these infinite sums determined definite numbers.*

Closing remarks

We have not touched on all the areas that Gregory took up for examination. In [2] we see a detailed analysis of the many themes that he explored in depth. We see Gregory coming up with a unified approach for finding the area of a sector of an ellipse and a sector of a hyperbola: the first time ever that this had been done. We see him exploring what he calls the *terminatio* of a series (much later, it would be called the 'limit' of a sequence). We see him exploring a sequence of arithmetic, geometric and harmonic means formed from two different sequences (much later, such investigations by Gauss would yield the notion of the AGM or *arithmetic geometric mean* — a rich area of modern mathematics), and he speculates that the terminatio is not expressible in terms of the elementary operations of arithmetic and algebra. He is attempting a proof of impossibility here, but his efforts fall short; the problem is far too difficult. It is a matter of wonder that he is even attempting such a proof! — for what he is trying to prove, in effect, is the transcendence of π. And there are other such explorations.

Keeping all these facts as a backdrop , we may wonder why James Gregory is not as well-known as he ought to be. In [2], the authors wonder "why this great man did not exert more influence on the actual development of mathematics." They explain that it is mostly due to a series of circumstantial factors: his living in an old university, with virtually no contact with the leading scholars of the day; his living in the same era as Newton and Leibniz, whose results on the calculus were of such power and generality that they overshadowed nearly everything else done in that area during that period; his habit of storing his findings in private letters and unpublished notes; and other such factors.

In 1675, James Gregory suffered a stroke and died soon after, having not yet reached the age of 37.

Acknowledgement

The author would like to record his deep appreciation to an anonymous referee for many helpful suggestions and for drawing attention to an argument presented in the 1981 book by Marsden & Weinstein, *Calculus Unlimited* [11]. This is in connection with Gregory's notion of a tangent line as "a line that touches the curve just once and does not cross it." We commented above that Gregory's definition is not consistent with the modern notion, which is based on the limit approach. However, based on the path taken by Marsden and Weinstein in their book *Calculus Unlimited* (pages 184-185), it is possible to refine this intuitive notion so that it is consistent with the limit definition.

Bibliography

[1] MacTutor. “James Gregory.” https://tinyurl.com/2ht3tyub

[2] Max Dehn and E. D. Hellinger, “Certain Mathematical Achievements of James Gregory.” *The American Mathematical Monthly*, Vol. 50, No. 3 (Mar., 1943), pp. 149-163. https://doi.org/10.2307/2302394

[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. https://tinyurl.com/2kox2rfx

[4] Enrique A. González-Velasco, “James Gregory’s Calculus in the ‘Geometriœ Pars Universalis’.” *The American Mathematical Monthly*, Aug.-Sep., 2007, Vol. 114, No. 7, pp. 565-576. https://www.jstor.org/stable/27642272

[5] Andrew Leahy, “A Euclidean Approach to the FTC.” https://tinyurl.com/2zany62s

[6] Wikipedia, “James Gregory (mathematician).” https://en.wikipedia.org/wiki/James_Gregory_(mathematician)

[7] Wikipedia, “History of calculus.” https://en.wikipedia.org/wiki/History_of_calculus

[8] Wikipedia, “Arctangent series.” https://en.wikipedia.org/wiki/Arctangent_series

[9] Wikipedia, “Taylor series.” https://en.wikipedia.org/wiki/Taylor_series

[10] Wikipedia, “Divided differences.” https://en.wikipedia.org/wiki/Divided_differences

[11] Jerrold Marsden & Alan Weinstein, *Calculus Unlimited*, Copyright © 1981 by The Benjamin/Cummings Publishing Company, Inc. https://authors.library.caltech.edu/records/7ejp1-f6z72/files/CalcUnlimited.pdf?download=1