Acknowledgments

In addition to our Associate Editors, the following referees have assisted the MAGAZINE *during the past year (May 2018 to May 2019). We thank them for their time and care.*

Elias Abboud, *Beit Berl College*
Edward Aboufadel, *Grand Valley State University*
William Adkins, *Louisiana State University*
David Aldous, *UC Berkeley*
Pieter Allaart, *University of North Texas*
Claudio Alsina, *Universitat Politécnica de Catalunya*
Michael Amspaugh, *Delta State University*
Marlow Anderson, *Colorado College*
Stephen Andrilli, *La Salle University*
Sigurd Angenent, *University of Wisconsin-Madison*
Dave Auckly, *Kansas State University*
Jathan Austin, *Salisbury University*
Eric Bach, *University of Wisconsin Madison*
Richard Bagby, *New Mexico State University*
Edward Barbeau, Jr., *University of Toronto*
Julia Barnes, *Western Carolina University*
Margaret Bayer, *University of Kansas*
Alan Beardon, *Centre for Mathematical Sciences*
Raymond Beauregard, *University of Rhode Island*
Bernard Beauzamy, *Société de Calcul Mathématique*
Paul Becker, *Pennsylvania State University*
Robert Beeler, *East Tennessee State University*
Jennifer Beineke, *Western New England University*
sarah-marie belcastro, *Smith College*
Allan Berele, *DePaul University*
Stephan Berendonk, *Universität Duisburg-Essen*
Ethan Berkove, *Lafayette College*
Ádám Besenyei, *Eötvös Loránd University*
Robert Bix, *University of Michigan Flint*
Harold Boas, *Texas A&M University*
Henry Boateng, *Bates College*
Mark Bollman, *Albion College*
Michael Bolt, *Calvin College*
Carmine Boniello, *Universita degli Studi di Salerno*
Michael Bosse, *Appalachian State University*
Khristo Boyadzhiev, *Ohio Northern University*
David Brink, *Akamai Technologies*
Marc Brodie, *Benedictine University Mesa*
Robert Brouzet, *Université Perpignan Via Domitia*
Eric Brussel, *California Polytechnic State University*
Charles Buehrle, *Notre Dame of Maryland University*
Kimberly Burch, *Indiana University of Pennsylvania*
Steve Butler, *Iowa State University*
Gunhan Caglayan, *New Jersey City University*
Andrés Caicedo, *Mathematical Reviews*
Thomas Cameron, *Davidson College*
Ann Cannon, *Cornell College*
Mindy Capaldi, *Valparaiso University*
Dean Carlson, *Mathematical Reviews*
Nathan Carlson, *California Lutheran University*

LETTER FROM THE EDITOR

The final issue of the Magazine for 2019 has something for everyone to read over the winter break. The first article is by Nathan Carter, in which he examines *RGB Express*, a puzzle game for mobile devices, and explains how digital computation can be represented through its puzzles. The puzzles involve designing a route for a truck to deliver packages under constraints.

Did your fall classes include grading schemes in which the best k out of n quizzes counted toward the final grade? Unless all the scores were the same, the end result is a higher average than if all n quizzes counted. In terms of grade inflation, what effect does this have on the final grade? In the second article of the issue, Peter Zizler and Mandana Sobhanzadeh provide a model for the expected grade inflation when lower scores are removed and the student's test-writing ability changes over time.

Stokes' theorem is a foundational result of several variable calculus. Iosif Pinelis focuses on different versions of Stokes' theorem to highlight how Stokes' theorem handles oriented and non-oriented surfaces. In doing so, he revisits Stokes' theorem on the Möbius strip.

Magic squares are well studied. But, can you create new ones from old ones? If M is a magic matrix (a real-valued matrix with the typical sum restrictions of a magic square), then what polynomials of M are also magical? Alan Beardon answers this question in his article.

Cut a circular pizza with n straight cuts that go all the way through the pizza. What is the maximum number of pieces you create, if you do not restrict the size and shape of the pieces? This is the pizza-cutter's problem, and Jean-Luc Baril and Céline Moreira Dos Santos answer this question by constructing Hamiltonian paths in an associated graph. Pieces are vertices in the graph and vertices are connected by an edge if they are adjacent pieces in the pizza.

There has been a recent resurgence in the analysis and study of nontransitive dice, including in THIS MAGAZINE. Along these lines, in their article "A Game of Nontransitive Dice," Artem Hulko and Mark Whitmeyer introduce a two-player, simultaneous zero-sum game in which each player's strategy is to select an n-sided die where each face consists of a positive integer and the face values sum to $n(n+1)/2$. The standard die with face values $1, 2, \ldots, n$ is one such die. Hulko and Whitmeyer provide a constructive algorithm to show that choosing the standard die is the unique Nash equilibrium in pure strategies (although mixed strategies may also exist).

Some of Ramanujan's well-known identities are related to a special class of 3rd degree polynomials, referred to as Ramanujan simple cubic polynomials. Gregory P. Dresden, Prakriti Panthi, Anukriti Shrestham, and Jiahao Zhang provide insights into the roots of cubic polynomials by showing that every monic polynomial of degree three with complex coefficients and no repeated roots is either a translation of $y = x^3$ or can be composed with a linear function to obtain a Ramanujan cubic. David Reimann's cover of this issue of the MAGAZINE is inspired by this article. See "About the cover" on the inside cover for more information.

The last article in this issue is by David Richeson and Tom Edgar. They give a visual proof of a lemma from which Gregory's result that bounds the area of a circle using inscribed and circumscribed regular polygons follows. They also recount some history about Gregory's attempt to prove that squaring the circle is impossible, Huygens' discovery of a flaw in Gregory's argument, and Huygens' accusation of plagiarism.

Brendan Sullivan has once again created a crossword puzzle to highlight the upcoming Joint Math Meetings. There is another TRIBUS puzzle in this issue, too. If the puzzles are not enough to keep you busy, head over to the Problems section to see if you can solve the quickies and problems. And, pay a visit to the Reviews section for some winter break reading. There is also a poem by Shashi Kant Pandey in this issue, as well as End Notes for some corrections from the last two years. Due to changes in privacy laws, the annual Referee Thank You has been moved to the publisher's website, available here. Without the care, knowledge, and generosity of the referees, the MAGAZINE would not exist. Thank you.

This issue is the last from my term as editor. My sincere thanks to the members of the Editorial Board, to Managing Editor Bonnie Ponce, and to Electronic Production and Publishing Manager Bev Ruedi. Special thanks to Problems Editor Eduardo Dueñez, who finishes up his term with this issue, Proposals Editor Eugen J. Ionaşcu, and Reviews Editor Paul Campbell. A final thanks to David Reimann who created cover art for each issue for the last five years. Best of luck to the new Editor Jason Rosenhouse and new Problems Editor Les Reid.

Michael A. Jones, Editor

Math. Mag. **92** (2019) 322. doi:10.1080/0025570X.2019.1683389

Circuits in RGB EXPRESS

NATHAN CARTER
Bentley University
Waltham, MA 02452
ncarter@bentley.edu

RGB Express is a puzzle game available for iOS and Android. Its press release says: "You are running RGB Express, the one and only company specialized in delivering colors. Your mission is to draw routes for your trucks, press play, and watch how the trucks deliver all the packages to their right destinations." [**1**] In this paper, we demonstrate the complexity of *RGB Express* by proving that its puzzles are capable of embodying arbitrary digital circuits.

Rules of the game

Play happens in two phases. First, the player specifies routes for trucks to follow. After initiating the second phase, the player watches while the trucks follow the routes. If they complete all the deliveries, the player has solved the puzzle. If not, she can learn from what went wrong and try again. Two example levels are shown in Figure 1.

A puzzle is a rectangular grid of cells, each containing one of the following.

Roads: Delivery trucks may drive only on roads.
Houses: Trucks deliver cargo to houses by passing in front of the house, as shown by small triangles in Figure 1. Houses sit adjacent to roads and come in several colors.
Bridges: They come in a variety of colors and can be raised or lowered as described below. Bridges can be driven over only when lowered.

Any one of the following secondary pieces of content can sit on top of a road cell. See Figure 1 for illustrations.

Trucks: Trucks come in the same set of colors as houses do.
Cargo: These, too, come in the same set of colors as houses do.
Buttons: These control bridges, and come in the same set of colors as the bridges.

Buttons may be up or down. Driving a truck over an "up" button instantly toggles all bridges and buttons of the same color: All up buttons become down and vice versa, while all raised bridges become lowered and vice versa. Driving a truck over a lowered button has no effect.

There is a single constant speed at which all trucks drive their assigned routes. They start, stop, and turn corners instantly. Any truck driving over any cargo automatically (and instantly) picks it up. Trucks can carry from zero to three items of cargo at once and newly acquired cargo is stacked on top of old. If a truck carrying three pieces of

Math. Mag. **92** (2019) 323–338. doi:10.1080/0025570X.2019.1614418
MSC: Primary 68Q05, Secondary 03-01
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/umma.

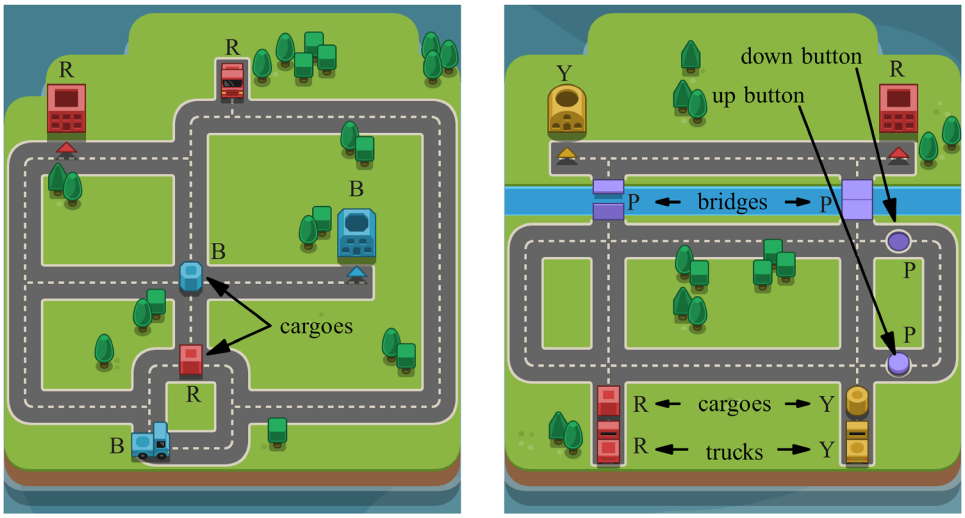


Figure 1 Two example game levels, Dallas C-3 on the left and Liverpool E-10 on the right. Images used with permission of *Bad Crane* [1]. To disambiguate colors, letters have been placed near colored items to indicate their colors (Red, Blue, Purple, Yellow). These letters do not appear in the game itself.

cargo drives over another piece of cargo, it does not pick it up, and proceeds as if the cargo were not there.

Trucks may deliver cargo only if the truck, cargo, and destination house are all the same color. A truck may not drive in front of a house unless the colors of house, truck, and topmost cargo match. If a truck arrives at a house and these three colors do not all match, the game stops and rejects the player's solution. One exception is a special white truck that can deliver any color cargo to a house matching the cargo's color.

Trucks do not stop when delivering cargo; they launch the top cargo from their stack into the house as they pass. Trucks' paths do not need to end at a house, but may.

No more than one truck may occupy any road cell at any given time. This sometimes requires the player to plan inefficient routes, to delay a truck so that it crosses an intersection at a later time than another truck. All trucks start driving at the same time, but may end at different times, depending on the lengths of the routes assigned.

For any two adjacent and connected road cells A and B, if any truck moves from A to B while driving a delivery route, then no truck (neither the same nor another) may move from A to B later, nor from B to A. In the game, this constraint is enforced by a user interface that prevents the player from drawing routes that overlap in this way.

Consider the two example route plans in Figure 2. Black cell boundaries have been added for clarity. On the left, the player has drawn an eight-step route for the red truck, and consequently cannot extend the route for the yellow truck any further. To do so would violate the rule about adjacent road cells, by requiring the yellow truck to drive between two cells through which the red truck is already scheduled to travel.

On the right, the two routes cross and later pass near one another, but never does a truck pass from cell A to cell B and then later another truck do the same, nor the reverse. The trucks will occupy the intersections at different times, preventing collisions.

Puzzle-loving readers may wish to pause and devise solutions for the levels shown in Figure 1. Solutions appear at the end of this article in Figure 15. Those two levels are a tiny sample. The game contains hundreds of levels and as the player progresses,

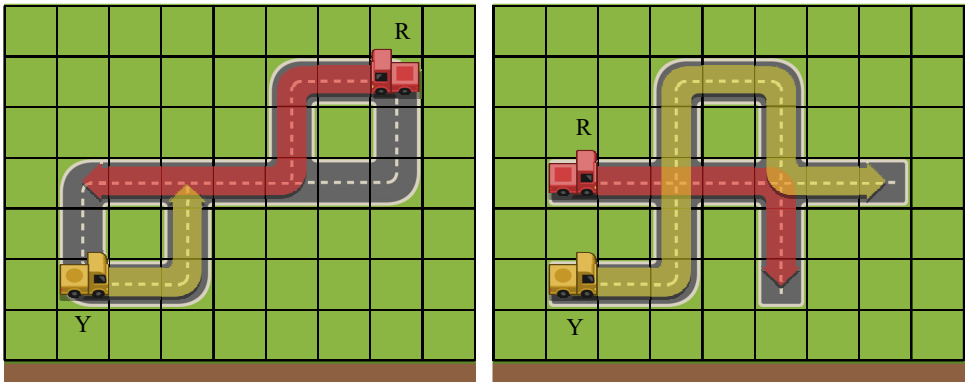


Figure 2 Example truck routes discussed in the first section of the article. Trucks are labeled R or Y to indicate their color (red or yellow).

new elements of the game are revealed that we do not consider in this paper. Consider downloading and playing the game! What mathematics does it suggest? The restriction against overlapping routes certainly brings up graph theory, but the rest of this paper considers the connection between *RGB Express* and circuits.

Representing computation

Mathematicians, computer scientists, and hobbyists have built computing devices out of toys for fun. A Tinkertoy computing device from the 1980s [**4**] plays tic-tac-toe. Minecraft players use a virtual resource to build circuits in a virtual world [**3**, **10**]. Hobbyists create computing devices out of dominos [**6**, **9**, **11**] and have learned which types of logic gates can be represented using dominos [**13**], showing the capabilities and limitations of dominoes as a medium for computation.

Creating a computing device in a new medium is a type of representation theorem. It embeds one class of objects, with all its complexity intact, into another class of objects. Many math students are familiar with representing groups as sets of permutations [**2**], implying that all the complexity of group theory can be found just within groups of permutations. Stone's representation theorem [**12**] represents any Boolean algebra as a field of sets.

Circuits are based on Boolean logic. The rest of this section briefly reviews that topic and additional detail is available in discrete mathematics texts such as [**5**]. In particular, Theorems 1 and 2 and Lemma 1 are quite well-known.

In mathematics, the order of operations tells us how to evaluate $3(2 + 9) - 7^2$; we do $2 + 9$ or 7^2 first and the subtraction last. We can represent this order using a tree, as shown on the left of Figure 3. A tree using logical operators is shown on the right of the same figure, using the standard symbols for and ($\wedge$), or ($\vee$), and not ($\neg$).

The trees should be read from the leaves (lowest nodes) upwards. A node containing the symbol for an operation (such as $+$) indicates that the operation should be applied to the values below it and the result propagated upward. This propagation happens physically in a digital circuit; wires transmit electricity in one of two states, on or off, realized by two voltage levels. One or two wires enter a gate, which embodies a logical operation and sends the result along an output wire, which may enter other gates.

The left side of Figure 4 shows the conventional illustrations for three types of logic gates common in digital circuits. An AND gate yields an "on" output iff both of

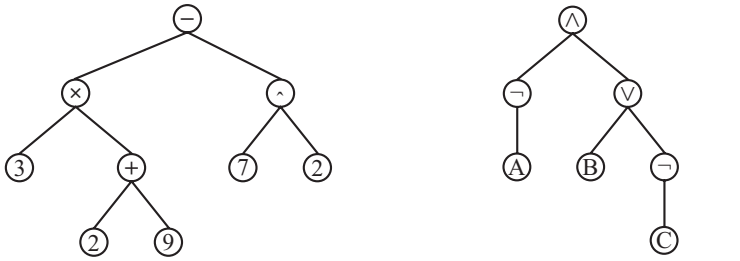


Figure 3 Trees representing the expressions $3(2+9)-7^2$ and $\neg A \wedge (B \vee \neg C)$

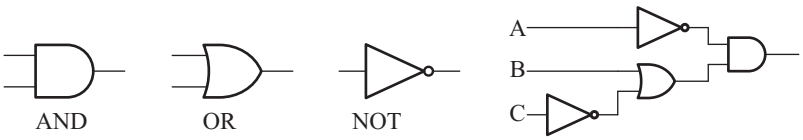


Figure 4 Conventional illustrations for three logic gates and a circuit that uses them

its inputs are on. An OR gate yields an "on" output iff either of its inputs is on. A NOT gate yields an "on" output iff its single input is off.

On the right side of that figure, the gates combine to form a circuit embodying the expression from the right of Figure 3. In all diagrams in this paper, electricity flows from the left side (where the inputs enter) to the right side (where the output exits). Not all circuits are trees; input wires are permitted to divide, thus sending their values to more than one gate.

This paper shows that it is possible to simulate all circuits as *RGB Express* puzzles.

Definition (Digital circuit)**.** Let B be the set of Boolean truth values $\{T, F\}$, standing for true and false. In this paper, a circuit with input wires $i_1, \ldots, i_n$ and output wires $o_1, \ldots, o_m$ is (a physical embodiment of) a function $f : B^n \to B^m$. Each wire transmits either T or F, with o_j transmitting the jth component of the vector $f(i_1, \ldots, i_n)$.

We start with the case when $m = 1$, then generalize to $m > 1$ in Theorem 2.

Theorem 1. *For any function $f : B^n \to B$, there is a digital circuit embodying f using only the gates AND, OR, and NOT.*

Proof. Take such an f and let $S = \{\vec{s} \in B^n \mid f(\vec{s}) = T\}$. Say $S = \{\vec{s}_1, \ldots, \vec{s}_k\}$.

To write a circuit that outputs T only on input $\vec{s}_1$ and outputs F on all other inputs, consider the input vector $\vec{s}_1$ described in English, such as "input 1 is T, input 2 is F, . . ., and input n is T." Such phrases are easy to write as conjunctions of (possibly negated) inputs, like $i_1 \wedge \neg i_2 \wedge \cdots \wedge i_n$, which can be converted easily into a circuit: Take the input wires $i_1, \ldots, i_n$, place NOT gates on those that are negated in the expression, and connect the results, two at a time, using $n - 1$ AND gates in any order.

For each $\vec{s}_i$, create such a circuit C_i that outputs T only on input $\vec{s}_i$. Beginning with C_1, we will add C_2 through C_k to form one larger circuit as follows.

1. Place circuit C_2 below C_1, and connect each input wire entering C_1 to its corresponding input in C_2, so that each input now sends its value to both C_1 and C_2.
2. Send the outputs of C_1 and C_2 into an OR gate, and make its output the output of the new circuit. Thus it outputs T iff its input vector is either $\vec{s}_1$ or $\vec{s}_2$.

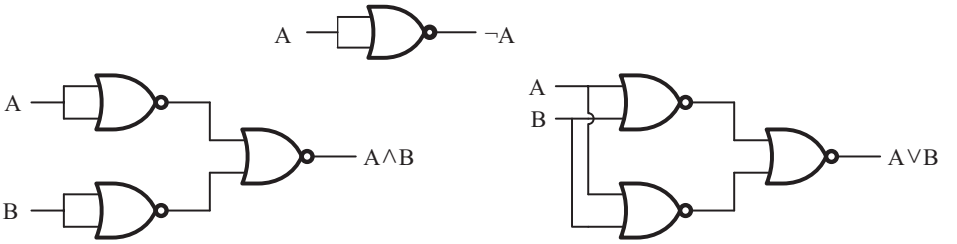


Figure 5 Simulating AND, OR, and NOT using only NOR, as in Lemma 1. In the $A \vee B$ circuit, the small curve means that one wire passes over the other without contact.

3. Repeat from step 1 to add C_3 to the growing circuit, yielding a new circuit that outputs T iff the input is in $\{\vec{s}_1, \vec{s}_2, \vec{s}_3\}$. Continue until all k circuits have been united, and the final circuit outputs T iff its input is in S. ■

If we add a new type of gate, we can strengthen Theorem 1. A NOR gate is so named because it computes "neither A nor B." It outputs true only when both of its inputs are false. Its symbol is that of an OR gate with a circle added on the right.

Lemma 1. *NOR gates alone are sufficient to build any circuit with one output wire.*

Proof. The diagrams in Figure 5 show that AND, OR, and NOT gates can be built from NOR gates. For each, the reader should verify that for all possible inputs, the output matches what the figure claims. By Theorem 1, these three gates are sufficient. ■

Theorem 2. *NOR gates alone are sufficient to build any circuit with m output wires.*

Proof. Consider a circuit C with m output wires. Consider any one such wire and trace it backwards, finding the set S_1 of all wires and gates that lead directly or indirectly into that output wire. Then S_1 is a circuit with one output wire, so we can apply Lemma 1, yielding some circuit S_1' that behaves like S_1 but uses only NOR gates.

Repeat this procedure for the next output wire, creating S_2 and S_2', which behave identically but S_2' uses only NOR gates. (While S_1 and S_2 may have gates and wires in common, S_1' and S_2' do not, as they have each been freshly constructed.) Continue until we have sets $S_1, \ldots, S_m$ and corresponding circuits $S_1', \ldots, S_m'$. Then the circuit shown in Figure 6 behaves exactly like C but uses only NOR gates. ■

Theorem 2 means that we have less work to do. We must show only that *RGB Express* levels can embody arbitrary combinations of NOR gates plus the techniques from Figures 5 and 6, that is, splitting and crossing wires.

Constraints

Before we construct an embedding of circuits into *RGB Express*, we define what we consider an acceptable result. We aim to encode any circuit C and input vector $\vec{i}$ into a winnable puzzle P such that whenever a player solves P, she is simultaneously computing the output of C on $\vec{i}$. More precisely:

1. **We must be able to create an *RGB Express* puzzle from any given circuit.** We will define a representation function $R(C)$ that converts any circuit C into a puzzle in which all roads entering from the left represent input wires and are initially empty, and m roads exit to the right, representing the output wires.

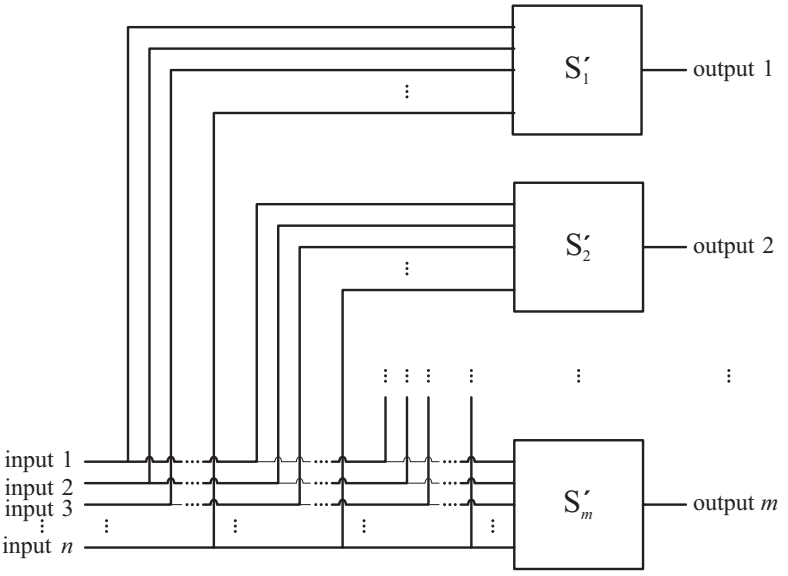


Figure 6 Assembling circuits S'_1 through S'_m from the proof of Theorem 2

2. **Given any vector of inputs, there is a simple modification to the puzzle that inserts those inputs.** We require an input-adding function $\alpha(P, \vec{i})$ that attaches input vector $\vec{i}$ (of correct length) to any puzzle P created by R, by modifying the n input roads only, in ways that are computationally trivial (say, $\alpha \in O(n)$).
3. **Solving the resulting puzzle (with inputs added) must make it obvious whether the circuit would output T or F on the inputs.** We require an output-reading function β that takes a completed play of the puzzle $\alpha(R(C), \vec{i})$ and tells us the output of C on $\vec{i}$ by a trivial inspection of the m output roads (say, $\beta \in O(m)$).

The force that makes our circuits flow is the player's effort toward winning. The computation is not driven by some automatic force, such as gravity toppling dominos.

RGB Express puzzles are limited by the size of the screen, so clearly we cannot represent *every* circuit. Thus we will permit arbitrary-sized game levels.

In each R constructed in this paper, the output puzzles are very easy to solve when they are small, but R could be applied to any size circuit and could yield arbitrarily large puzzles, beyond human grasp merely due to their size. So we assume an idealized player who can solve a puzzle of any size.

This is a great place to pause and try to design a representation theorem yourself! Download *RGB Express* and play to get inspired. Once you have done so—or if you are too curious for such a detour—read on to see my two representation theorems.

An acceptable solution

A natural choice is to represent each circuit wire as a single road in the puzzle. Trucks would drive along the roads the way current flows along the wires, and trucks carrying cargo would represent T while trucks without cargo would represent F.

Thus $R(C)$ will encode a circuit C with n inputs as a puzzle with n roads entering from the left, each containing an empty white truck, and m roads exiting to the right. And $\alpha(R(C), \vec{i})$ will place a red cargo item immediately in front of truck j iff $i_j = T$. Trucks with a red cargo item in front of them will, as soon as they move one cell, pick

up that cargo. Let us call this set of conventions *Convention Set 1*.

We must now design the functions R and β so that each puzzle R creates is winnable, β can read the correct output from a solved version of $R(C)$, and β is trivial. Unfortunately, we have a problem. Recall from the first section that each house must receive a delivery. Furthermore, I have not encountered any puzzle in which the game provides extra cargo. Respectively, these enforced and implied constraints are the following.

1. All houses must receive a delivery. (enforced)
2. All cargo must be delivered. (implied)

Theorem 3. *Convention Set 1 cannot satisfy constraints 1 and 2 simultaneously.*

Proof. Assume that Convention Set 1 and both constraints hold. Thus the number of houses must equal the number of cargoes. But the function α makes the number of cargoes vary with the input, while the number of houses remains constant across all inputs, so for some inputs, the two quantities will not be equal. ■

So under Convention Set 1, we must violate either constraint 1 or 2. Because the game enforces constraint 1 while only implying constraint 2, in this section we will violate constraint 2. But discarding constraints is irksome, so the next section provides another solution using a less natural convention set but satisfying both constraints.

For now, we assume that the player will be moving trucks rightwards as far as possible along roads, then in Lemma 6 prove that such an assumption is valid.

Lemma 2. *There exists a configuration that guarantees that T trucks take one path through it while F trucks take a different path.*

Proof. Consider the configuration shown in Figure 7, which we call a "separator," because it separates T and F trucks as per the statement of the lemma.

The only two valid routes through the separator are shown in the figure. A truck entering from the left must turn left or right at the fork at the top of the configuration. A truck turning to its own right at the fork presses the orange button, which raises both orange bridges. For that reason (and because it cannot retrace any path it has already traveled) its only possible path is the one shown on the top of Figure 7. Symmetrically, a truck choosing to go to its own left at the fork presses the purple button, and thus restricts itself to the path shown on the bottom of the figure.

Now consider a T truck entering the configuration from the left side. By Convention Set 1, it is carrying one red cargo. It cannot turn to its own right at the fork, as follows. The path on the top of Figure 7 would first send the truck over three yellow cargo items, of which it would pick up only two before being full. It would then pass three yellow houses. After dropping two yellow cargo items at the first two yellow houses, it would attempt to drop a red cargo item at the third yellow house. The game would stop and reject the solution.

If, however, it chose to go to its own left at the fork, it would follow the path on the bottom of the figure, which allows it to drop off its one red cargo item first, then pick up and drop off three yellow cargo items before picking up one red item again and exiting the configuration. Thus it exits the configuration in the state in which it entered and all houses receive cargo.

Now consider an F truck entering the configuration from the left side, which by Convention Set 1 contains no cargo. It cannot turn to its own left at the fork, because that path takes it directly to a red house, for which it has no cargo. Thus it is forced to turn to its own right, picking up and dropping off three yellow cargo items, then picking up and dropping off one red cargo item, then exiting the puzzle empty, the same state in which it entered. Again, all houses receive cargo.

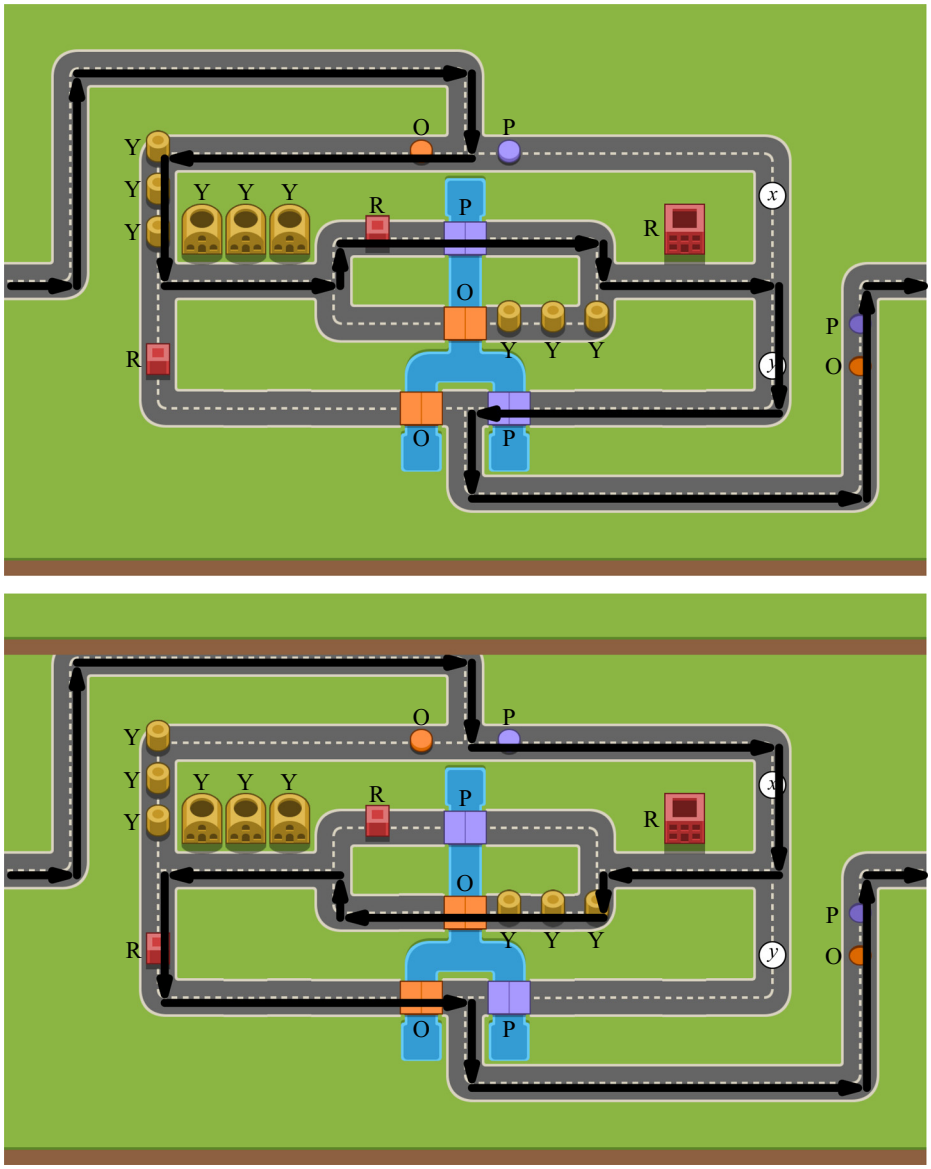


Figure 7 The two valid routes through a "separator," from Lemma 2.

Thus T and F trucks take different paths through the puzzle and both exit the configuration in the state in which they entered it, having delivered cargo to all houses. ■

Furthermore, T trucks always pass over the point marked x in the figure but not the point marked y, while F trucks pass over y but not x. This lets us use variations on the separator to have T and F trucks take different actions. For instance, buttons can be placed at x, or y, or both.

Thus we write $Sep(x, y)$ for this configuration, for various values of x and y. If we let Y^+ stand for a raised yellow button and $*$ stand for nothing, then $Sep(Y^+, *)$ is the configuration from Figure 7, but with a raised yellow button added at the location marked x, and nothing added at y. We use $Sep(x, y)$ for building gates.

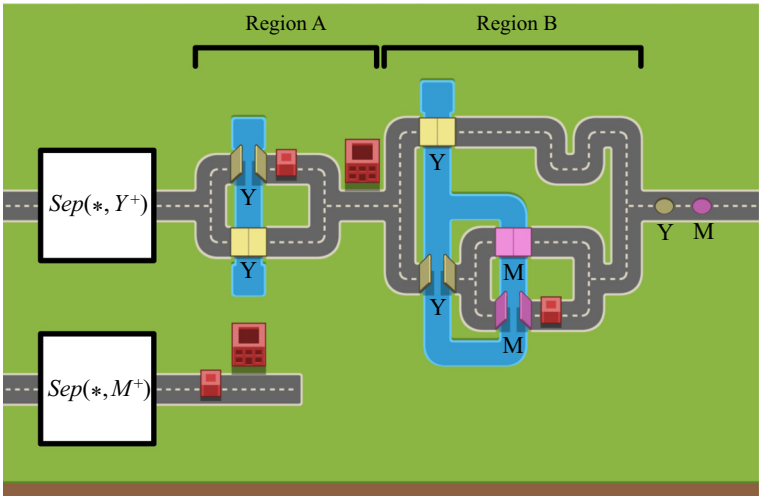


Figure 8 A NOR gate under Convention Set 1. The cargo and houses are red; buttons and bridges are labeled Y (M) to indicate yellow (magenta). The $Sep(x, y)$ notation is defined before Lemma 3, with Y^+ (M^+) meaning a raised yellow (magenta) button.

Lemma 3. *Under Convention Set 1, the configuration in Figure 8 is a NOR gate.*

Proof. Two trucks enter the left roads, possibly with cargo. We show that one truck exits the right road, with cargo iff neither of the entering trucks had cargo.

We assume that the player will move the top truck through the top configuration and out the right side of the puzzle; we prove this in Lemma 6. Because that truck will go through the $Sep(*, Y^+)$ structure, it toggles the yellow bridges iff it is an F truck. In Region A, the top truck will therefore pick up cargo iff it was an F truck and will get to the house at the end of Region A with cargo in all cases. Thus it will deliver cargo to that house and exit Region A empty in all cases.

Because the player aims to win, the truck on the bottom road must deliver cargo to the bottom house. So it must go through the $Sep(*, M^+)$ structure, toggling the magenta bridges iff it is an F truck.

In Region B, the top truck has only one path that allows it to pick up cargo, and that path can be accessed only if both the yellow and magenta bridges have been toggled. The $Sep(*, Y^+)$ and $Sep(*, M^+)$ structures guarantee that this happens only if the two trucks that entered the gate were both F trucks, and thus the top truck exits the configuration with cargo iff neither truck entered with cargo, as desired.

After Region B, the truck restores all buttons and bridges to their initial states. ■

Although Lemma 3 shows that we can create the one gate necessary to form any kind of circuit, we have not yet shown that we can assemble such gates into a larger puzzle. There are three capabilities we still need. First, we must be able to cross one "wire" over another, as we did for $A \vee B$ in Figure 5. (In *RGB Express*, roads cannot pass over or under one another.) Second, we need a way to divide a wire, as we do with all input wires in Figure 5, sending the same value along both paths. Third, we need a way to ensure that all the gates in a compound circuit behave independently of one another. This is not yet guaranteed, because our gate designs use buttons, which affect all other buttons and bridges of the same color throughout the puzzle.

Lemma 4 solves the first two of our three problems. It gives us a way to cross one "wire" over another, by introducing a new truck at any location we like, with this

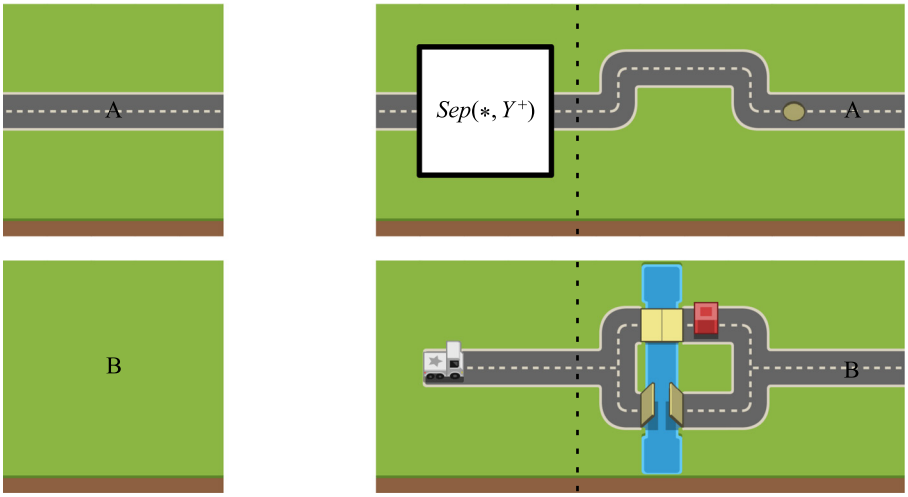


Figure 9 Duplicating a truck's meaning from cell A to cell B, as in Lemma 4. All cargo and houses in the figure are red; the buttons and bridges are yellow.

guarantee: The new truck is a T (F) truck iff the original truck was a T (F) truck. It also gives us a way to divide a wire, by making a copy next to the original.

Definition (copy of a road cell)**.** In a puzzle encoding a circuit using Convention Set 1, given any cell A on a road that represents a wire, we say that another road cell B elsewhere in the puzzle is a *copy* of A if every possible way to win the puzzle guarantees the following three properties at each point in time.

1. If there is no truck on A, then there is no truck on B.
2. If there is an empty truck on A, then there is an empty truck on B.
3. If there is a truck with cargo on A, then there is a truck with cargo on B.

Lemma 4. *Under Convention Set 1, take any encoded puzzle P and any cell A on a road representing a wire. For any grassy area B directly above or below A, we can modify P to add a copy of A at B without changing the puzzle's output behavior.*

Proof. Consider the situation shown in the left of Figure 9, where we have a road cell A and wish to create a copy of it at some empty location B in the same puzzle. The two sections are shown disconnected because they may not be near one another.

Modify the puzzle by inserting the configuration on the right of the figure. Extend to the left the road containing the white truck enough to ensure that the two trucks pass the dashed line at the same time. The time the top truck takes to get to the dashed line is constant because all paths through Figures 7–9 are the same length.

If the truck on the top road contains cargo, then based on the definition of $Sep(*, Y^+)$, it will not press a yellow button. The bottom truck will therefore be constrained to the top half of its road, and thus the trucks pass points A and B simultaneously and both with cargo. If the top truck enters the configuration empty, then $Sep(*, Y^+)$ forces it to toggle the yellow bridges, and thus the bottom truck does not gain cargo, and the trucks pass points A and B simultaneous and both empty.

In both cases, the top truck restores the yellow button (and bridges) to their initial states. The vertical distance between the top road and the bottom road can be as large as needed when building a circuit, including moving the bottom road above the top one if necessary. Thus A and B need not be near one another. ■

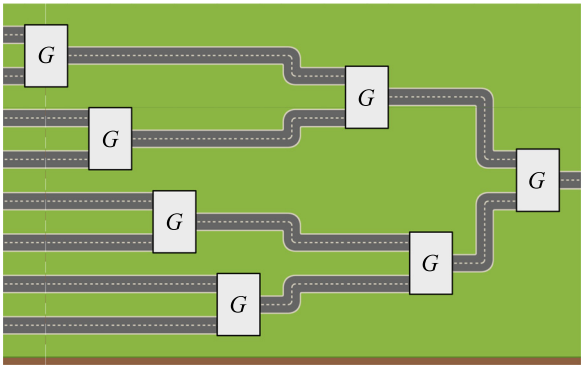


Figure 10 Arrangement of gates (each labeled G) ensuring only one is used at a time

The following lemma resolves the third and final problem stated before Lemma 4.

Lemma 5. *Under Convention Set 1, gates in a puzzle can be arranged so that all operate independently.*

Proof. The gates shown so far in the paper use buttons to toggle bridges, which may impact other buttons and bridges in the puzzle. If two gates were to operate at the same time, a button in one might impact buttons or bridges in the other. We can ensure that all gates operate at different times by stretching the circuit horizontally and ensuring that no gate overlaps another vertically, as shown in Figure 10. If some gates take longer to traverse than other gates or roads, extra horizontal space can be added to synchronize all trucks vertically after each gate. ■

Lemma 6. *Under Convention Set 1, there is a method for encoding any single-output circuit as a puzzle in* RGB Express.

Proof. The function R should apply Lemma 1 to convert a single-output circuit into exclusively NOR gates, apply Lemma 3 to convert each NOR gate into a puzzle fragment, and connect those fragments using Lemmas 4 and 5.

We now validate the assumption made throughout this section that players will be motivated to move trucks continually to the right in every situation. Consider any truck in the puzzle. The construction described in the previous paragraph gives every truck only one path it can follow to the right at any given time. If we follow that road to the right, it must eventually end, because the puzzle is finite and trucks cannot follow cyclic routes. There are only two ways that a road can end: It can be the lower road in a NOR gate or a dead end (such as the end of the entire circuit or an unused input wire). The lower road in a NOR gate ends with a cargo and a house, which together require the truck to proceed all the way to the end. To all other dead ends, we append that same configuration, a single red cargo followed by a single red house. Thus the player must move all trucks to the rightmost end of their roads.

Let α place inputs into a puzzle as per Convention Set 1 and β read output from the final road exiting the circuit as follows. The truck that delivers cargo to the last house on that road will indicate the output of the circuit by whether it is carrying cargo (T) or not (F). The cargo the truck just delivered will have increased its load by one and then immediately decreased it again, restoring its original value before β reads it. ■

Theorem 4. *Under Convention Set 1, there is a method for encoding any circuit as a puzzle in* RGB Express.

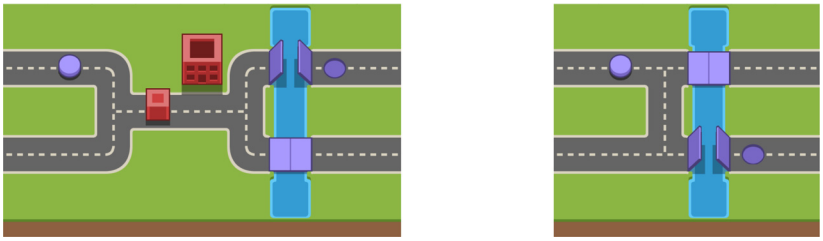


Figure 11 Left: A pattern that incentivizes players to move trucks further along a pair of roads. Right: A NOT gate encoded according to Convention Set 2. The cargo and house are shown in red; all buttons and bridges are purple.

Proof. As in the proof of Theorem 2, we consider separately each output wire of the circuit, tracing it back to find every wire and gate that leads to it. Apply Lemma 6 to that subcircuit, creating a puzzle. Repeat this procedure for each output wire. As Theorem 2 assembles the resulting circuits in the configuration shown in Figure 6, we do the same, using the ability to divide and cross input wires from Lemma 4. ■

Theorem 4 accomplishes our goal, but we had to concede that there would often be more cargo than houses (Theorem 3), which never happens (to my knowledge) in *RGB Express* itself. We also had to create complex configurations such as $Sep(x, y)$.

We therefore ask whether there is an alternative to Convention Set 1 that might yield a more elegant solution; the following section presents one. As before, the reader may pause to consider the question and read my solution thereafter.

A two-road solution

Let *Convention Set 2* represent each wire by two parallel roads in the puzzle. The function α now places one red truck on each pair of roads, using the top road to indicate T and the bottom road for F. Trucks begin the puzzle without cargo.

The left of Figure 11 shows a pattern that can be appended to any pair of roads to incentivize the player to move a truck along the pair and through the pattern. Trucks enter the pattern empty and exit it empty, but on the same road (top or bottom) on which they entered.

We assume that every configuration introduced in the remainder of this section has this pattern appended, forcing players to move trucks to the right when possible.

Lemma 7. *Under Convention Set 2, there exist configurations embodying AND, OR, and NOT gates.*

Proof. Although we could create a configuration for just a NOR gate, as in Lemma 3, it is nice to see the natural duality between AND and OR shown in their gates.

Consider the NOT gate on the right of Figure 11. If the truck enters the top road (T) then it presses the purple button, toggling the bridges, and forcing itself to the lower path (F) before toggling the bridges back again. If it enters on the lower path (F), the bridges are not toggled, and it must exit on the upper path (T). Both paths have the same length and negate the truck's meaning.

The reader should consider the four possible input cases for each of the AND and OR gates shown in Figure 12. The extra wiggles on some roads ensure that the time it takes the top truck to traverse each gate is the same in all cases. The OR gate was produced from the AND gate by vertically flipping each of the two two-road wires, thus interchanging the meanings of T and F for both the inputs and the output. This shows DeMorgan's duality between conjunction and disjunction. ■

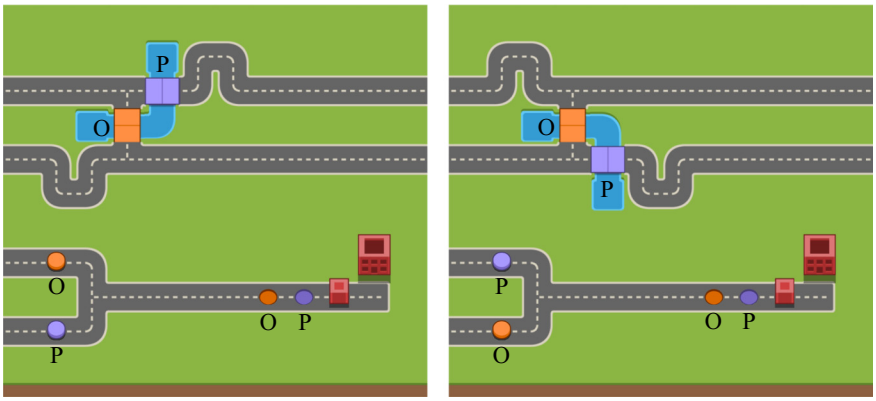


Figure 12 AND (left) and OR (right) gates encoded according to Convention Set 2 . All cargo and houses are red; buttons and bridges labeled P (O) are purple (orange).

We begin to see some of the benefits of the two-road approach. While it may be less natural to represent a single wire with two *RGB Express* roads, we have been able to avoid the complex "separator" circuit. Furthermore, the gates in Figures 11 and 12 are all simpler than the one in Figure 8 and do not require the special white truck. Also, we need not sprinkle our circuits with cargo that will not be delivered.

Assembling gates into circuits requires results analogous to Lemmas 4 and 5.

Definition (copy of a point on a pair of roads)**.** In a puzzle encoding a circuit using Convention Set 2, given any cell A sitting between two roads that represent a wire, we say that another point B between two roads representing a wire elsewhere in the puzzle is a copy of A if every possible way to win the puzzle guarantees the following two properties at every point in time.

1. There is a truck on the road above A iff there is a truck on the road above B.
2. There is a truck on the road below A iff there is a truck on the road below B.

Like Lemma 4, Lemma 8 lets us cross one "wire" over another or divide one.

Lemma 8. *Under Convention Set 2, take any encoded puzzle P and any point A between a pair of roads representing a wire. For any grassy area B directly above or below A, we can modify P to add a copy of A at B without changing the puzzle's output behavior.*

Proof. Consider the situation on the left of Figure 13, where a pair of roads pass around the cell A. We wish to create a copy of it at some other, empty location B in the same puzzle. We modify the puzzle by inserting the configuration shown in the right of that figure. The dashed line has the same meaning as in Lemma 4, and the bottom truck's road can be extended to the left as needed. The time the first truck takes to get to the dashed line is constant, because each path through a gate takes the same amount of time. The top truck governs which path the bottom truck must take by toggling the purple bridges iff the lower truck should take the top route. ■

Lemma 9. *Under Convention Set 2, gates in a puzzle can be arranged so that they all operate independently.*

Proof. Take the proof of Lemma 5 and double the input and output roads per gate. ■

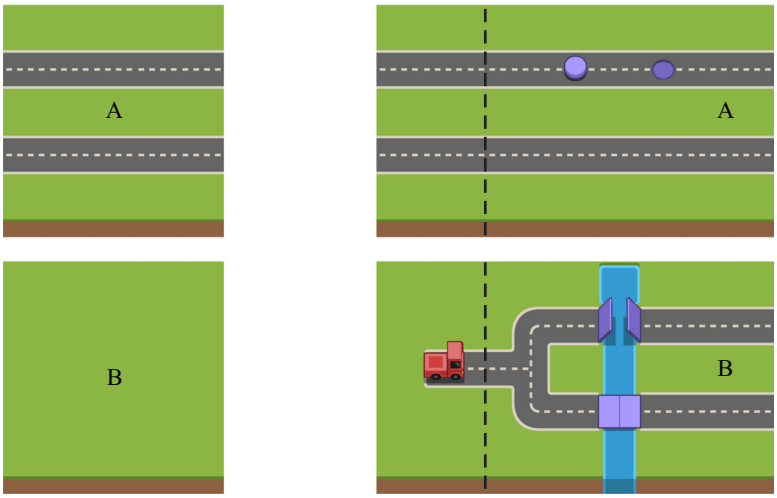


Figure 13 Method of duplicating an input signal from one wire (top) to another wire (bottom) using Convention Set 1 . The truck is red; the buttons and bridges are purple.

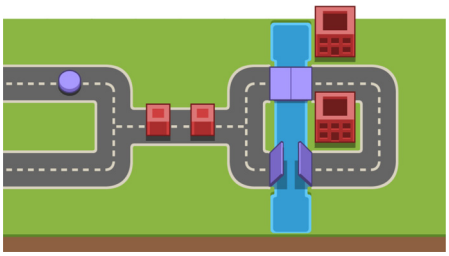


Figure 14 The final structure on the right of any puzzle created by Theorem 5. The cargo and houses are red; the button and bridges are purple.

Theorem 5. *Under Convention Set 2, there is a method for encoding arbitrary circuits as puzzles in* RGB Express*.*

Proof. Let R apply Theorem 1 to convert a circuit into AND, OR, and NOT gates, apply Lemma 7 to convert each gate into a puzzle fragment, assemble those fragments using Lemmas 8 and 9, and append to all outputs the configuration in Figure 14.

Let α place inputs into a puzzle as described in Convention Set 2. The function β reads each output by considering its copy of the configuration in Figure 14. If the truck entered on the top road (T), it toggled the bridges and was forced to deliver the two cargoes by following the final loop counterclockwise, ending at the top house. Otherwise it will do the reverse, ending at the bottom. Thus β can examine the truck's position (top or bottom) to discern its meaning (T or F, respectively). ■

Conclusion

We have demonstrated two solutions for encoding arbitrary computations as *RGB Express* puzzles. The first used a more natural encoding of wires as single roads, but required us to leave some cargo undelivered, used more intricate gates, and required a white truck. The second used the less natural encoding of a wire as a pair of roads, but avoided the other inelegances. These results demonstrate the inherent complexity in the game itself. But we can also learn a few other things from this representation.

The first was mentioned in the proof of Lemma 7; we have gained a new way to look at the duality between "and" and "or," as expressed by DeMorgan's laws. This is illustrated in the vertical inversions in the constructions in Figure 12.

Second, we came face-to-face with the importance of the flow of electricity. Our representation of circuits required some way to cause the trucks in the puzzle to move along the roads, imitating the flow of electrons along wires. We used the player's desire to solve the puzzle as a driving force, which brings up a related question:

In a computer, electricity is not pushed into the circuit once but flows continuously. The system can use that flow to power data storage and a ticking clock that permits the execution of operations in succession, each one using the data computed by previous operations. Could the work in this paper be extended similarly? Perhaps a puzzle could be partially reset after each win so that it was ready for another play, altered in some way by previous play to propagate information.

Finally, we can see a difference between the design of puzzles and the design of hardware and software systems. The sample puzzles in Figure 1 seem entertaining and inviting, while those created by the procedures in this paper are of interest only for their structural properties; they aren't fun to play. This is due to competing design goals. When designing hardware or software, good organization is a virtue. Keeping the components of the system largely separate and making their interaction simple enables others to understand the invention, maintain it, and improve it. But engaging puzzles are supposed to have complex interactions among their components, requiring the player to explore, experiment, and learn on the way to a solution.

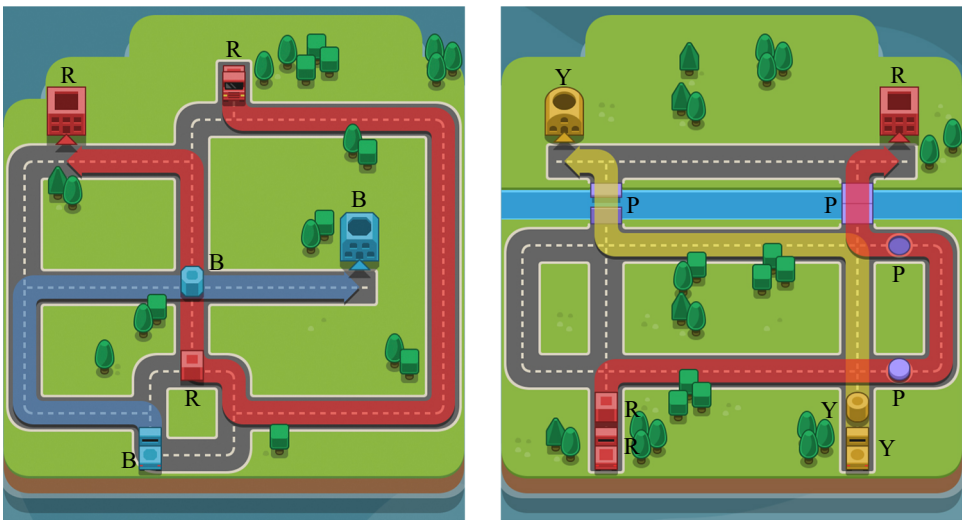


Figure 15 Solutions to puzzles shown in Figure 1. The letters overlaid on the figure do not appear in the game itself; they have the same meanings as in Figure 1.

Afterward. The computational complexity of games is still an active research area. As this article went to print, three authors submitted an article to the arXiv claiming to prove that the card game *Magic: The Gathering* has the same complexity as a Turing machine. It cites several related papers on complexity in various board games, video games, and puzzles.

REFERENCES

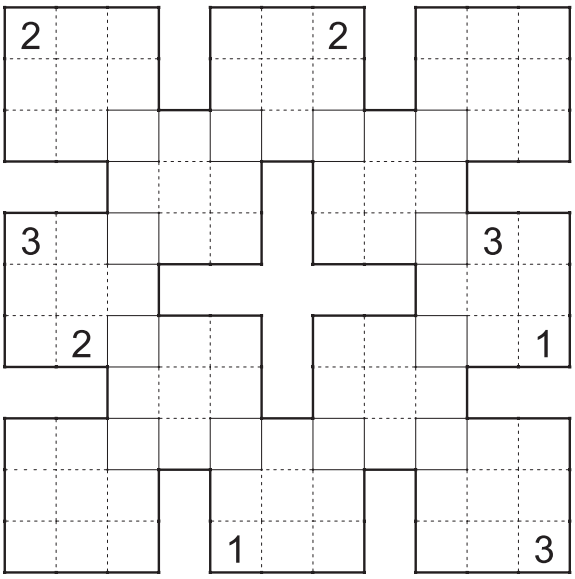
[1] Bad Crane. (2014). RGB Express Press Kit. badcrane.com/presskit/sheet.php?p=rgb_express. 2017-01-30 21:36:04.

[2] Cayley, A. (1854). On the theory of groups as depending on the symbolic equation $\theta^n = 1$. *Philos. Mag.* 7(42):40–47.
[3] Cordeiro, J. (2014). *Minecraft Redstone for Dummies*. Hoboken, NJ: John Wiley and Sons.
[4] Dewdney, A. K. (1989). A Tinker toy computer that plays tic-tac-tow. *Scient. Amer.* 261(4):120–123.
[5] Epp, S. S. (2010). *Discrete Mathematics with Applications*. 4th ed. Boston, MA: Brooks/Cole Publishing Co.
[6] Fraser, N. (2008). Somino logic. neil.fraser.name/news/2008/06/29/. 2017-01-30 21:19:39.
[7] IEEE Graphic Symbols for Logic Functions. *IEEE Std. 91/91a-1991*. doi.org/10.1109/IEEESTD.2000.92296.
[8] Jordan, C. (1870). *Traite des substitutions et des equations algebriques*. Paris, France: Guathier-Villars.
[9] Kybernetikos. (2007). Domino computation. kybernetikos.com/2007/03/01/domino-computation. 2017-01-30 21:22:39.
[10] Minecraft Wiki (2017). Redstone circuit. minecraft.gamepedia.com/Redstone_circuit. 2017-01-30 21:29:29.
[11] Newman, L. H. (2014). Building a computer out of dominoes. slate.com/blogs/future_tense/2014/04/09/numberphile_builds_circuits_out_of_dominoes_to_show_how_computers_do_basic.html. 2017-01-31 20:31:03.
[12] Stone, M. H. (1936). The theory of representation for boolean algebras. *Trans. Amer. Math. Soc.* 40(1):37—111.
[13] Think Maths. (2017). *Worksheets for learning domino circuits*. think-maths.co.uk/downloads/domino-computer-worksheets. 2017-01-30 21:39:40.

Summary. A recurring theme in mathematics and computer science is *representation,* using one class of objects to imitate or simulate another class. These include classic examples such as Stone's and Cayley's representation theorems, as well as recreational ones such as representing digital computation using dominos or Tinkertoys. This paper is of the latter kind, investigating how digital computation can be embedded in the mobile puzzle game *RGB Express*. I introduce the game first, then prove two representation theorems about it.

NATHAN CARTER (MR Author ID: 643096) is a professor of mathematics at Bentley University. He sits on the boundary between mathematics and computer science, doing things like mathematical software, mathematical visualizations in computer graphics, and data science.

Tribus Puzzle



How to play. Fill each of the three-by-three squares with either a 1, 2, or 3 so that each number appears exactly once in each column and row. Some cells apply to more than one square, as the squares overlap. Each of the three-by-three squares must be distinct. The solution can be found on page 395.

— David Nacin, William Paterson University, Wayne, NJ (nacind@wpunj.edu)

Grade Inflation Due to Selective Averaging

MANDANA SOBHANZADEH
Mount Royal University
Calgary, AB, Canada
msobhanzadeh@mtroyal.ca

PETER ZIZLER
Mount Royal University
Calgary, AB, Canada
pzizler@mtroyal.ca

Rather than taking the mean of all the tests written by a student during the term, some professors compute each student's final grade by taking the average of only the student's best scores. In general, for each student, the mean of the best k-out-of-n tests is counted towards her final grade. Naturally, there is a danger of grade inflation at the end of the semester in comparison to the standard method of taking the mean of all tests.

We are interested in estimating the expected grade inflation from this selective assessing prior to any tests being written. Clearly, for each student one can calculate the individual grade inflation after the tests are written, but this is not what we are interested in. We wish to have a priori knowledge about the expected grade inflation based on the procedure implemented.

We have found the expected grade inflation, with the fixed choice of the best k-out-of-n, depends only on the student test-writing inconsistency σ and not her test-writing ability. This holds under the assumption that the student test-writing inconsistency σ remains static during the course. However, we allow for the student test-writing ability to change over time during the term. In particular, we show that the expected grade inflation is the same whether the student changes her ability over time or not. This is not immediately obvious nor intuitive. We provide a formula for the expected grade inflation estimate in terms of k, n, and σ. Furthermore, we also allow for nesting of selective assessments, for example, choosing the best 4-out-of-5 questions within each test and then choosing the best 3-out-of-5 among tests. We provide expected grade inflation formulas for these different nesting options and observe that as we implement more nesting the expected grade inflation increases. However, as we keep increasing the levels of nesting the grade inflation eventually stabilizes.

A simplified model for this was done in [4], where it was assumed that the student does not change her ability over time. In particular, it was assumed that a student has a certain constant, measurable academic ability μ. When she writes a test we draw the resulting score x from a normally distributed random variable X with mean μ and standard deviation σ, which is also assumed to be constant. The value of σ can be thought of as the measurement of the student's test-writing inconsistency. We assume students do not intentionally skip tests, which in practice they might well decide to do. For the reader's convenience we review the set up and notation from [4].

Let Y be the mean of the best k-out-of-n tests written during the term. Then Y is a random variable with mean $\mathbf{ave}_{\mu,\sigma}(k,n)$ and standard deviation $\mathbf{std}_{\mu,\sigma}(k,n)$:

$$\mathbf{ave}_{\mu,\sigma}(k,n) = \sigma \cdot \mathbf{ave}_{0,1}(k,n) + \mu \text{ and } \mathbf{std}_{\mu,\sigma}(k,n) = \sigma \cdot \mathbf{std}_{0,1}(k,n), \qquad (1)$$

Math. Mag. **92** (2019) 339–344. doi:10.1080/0025570X.2019.1637676
MSC: Primary 62P99

where $\mathbf{ave}_{0,1}(k,n)$ and $\mathbf{std}_{0,1}(k,n)$ are constants independent of μ and σ that depend on k and n only. Equation 1 is justified as follows. The probability that all means of k elements (chosen out of n) are less than a given x is

$$F_{\mu,\sigma}(x) = \int_L g_{\mu,\sigma}(t_1)\cdots g_{\mu,\sigma}(t_n)\,dt_1\cdots dt_n,$$

where $g_{\mu,\sigma}(t)$ is the normal probability density function

$$g_{\mu,\sigma}(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(t-\mu)^2}{2\sigma^2}},$$

and L is the region in $\mathbb{R}^n$ defined by the inequalities

$$t_{n_1} + t_{n_2} + \cdots + t_{n_k} < kx,$$

where $\{n_1,\ldots,n_k\}$ is any subset of $\{1,\ldots,n\}$. Note that $F_{\mu,\sigma}(x) = F_{0,1}(z)$, where $z = \dfrac{x-\mu}{\sigma}$. Then the probability density function of Y is

$$f_{\mu,\sigma}(x) = \frac{dF_{\mu,\sigma}}{dx} = \frac{1}{\sigma} f_{0,1}(z), \text{ where } f_{0,1}(z) = \frac{dF_{0,1}}{dz}.$$

The expected value $\mathbf{ave}_{\mu,\sigma}(k,n)$ of Y is

$$\mathbf{ave}_{\mu,\sigma}(k,n) = \int_{\mathbb{R}} x f_{\mu,\sigma}(x)\,dx = \int_{\mathbb{R}} (\sigma z + \mu) f_{0,1}(z)\,dz = \sigma\cdot\mathbf{ave}_{0,1}(k,n) + \mu,$$

where $\mathbf{ave}_{0,1}(k,n)$ is given by

$$\mathbf{ave}_{0,1}(k,n) = \int_{\mathbb{R}} z f_{0,1}(z)\,dz.$$

Similarly, the standard deviation $\mathbf{std}_{\mu,\sigma}(k,n)$ of Y is given by

$$\begin{aligned}
\mathbf{std}^2_{\mu,\sigma}(k,n) &= \int_{\mathbb{R}} (x - \mathbf{ave}_{\mu,\sigma}(k,n))^2 f_{\mu,\sigma}(x)\,dx \\
&= \int_{\mathbb{R}} (\sigma z + \mu - (\sigma\cdot\mathbf{ave}_{0,1}(k,n) + \mu))^2 f_{0,1}(z)\,dz \\
&= \sigma^2 \int_{\mathbb{R}} (z - \mathbf{ave}_{0,1}(k,n))^2 f_{0,1}(z)\,dz \\
&= \sigma^2\cdot\mathbf{std}^2_{0,1}(k,n),
\end{aligned}$$

where $\mathbf{std}_{0,1}(k,n)$ is independent of μ and σ and depends only on k and n. Therefore, the expected grade inflation due to the selective averaging is given by $\mathbf{ave}_{0,1}(k,n)\sigma$, where σ is the measurement of the student test-writing inconsistency, assumed to be a constant during the term. Below we provide a few numerical simulations.

Figures 1 and 2 show plots for $\mathbf{ave}_{0,1}(k,n)$ and $\mathbf{std}_{0,1}(k,n)$ for fixed $k = 4$, with increasing n, as well as for fixed ratio $k/n = 3/5$, for increasing n. We provide some useful values for these quantities, with two decimal precision, that can be used in practice: $\mathbf{ave}_{0,1}(3,5) = 0.55$, $\mathbf{std}_{0,1}(3,5) = 0.50$, $\mathbf{ave}_{0,1}(3,4) = 0.34$, $\mathbf{std}_{0,1}(3,4) = 0.53$, $\mathbf{ave}_{0,1}(7,9) = 0.35$, $\mathbf{std}_{0,1}(7,9) = 0.35$, $\mathbf{ave}_{0,1}(4,5) = 0.29$, $\mathbf{std}_{0,1}(4,5) = 0.46$.

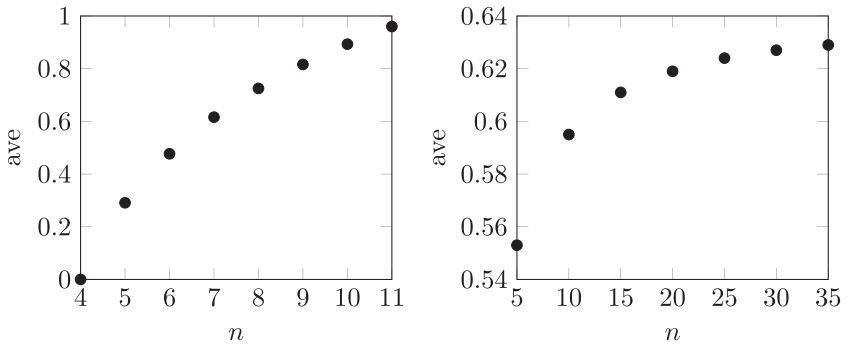


Figure 1 How the averages changes with respect to n with $k = 4$ (Left) and $k/n = 3/5$ (Right).

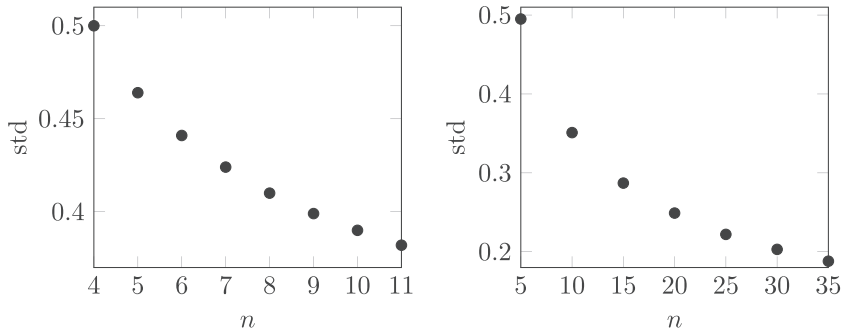


Figure 2 How the standard deviation changes with respect to n with $k = 4$ (Left) and $k/n = 3/5$ (Right).

Order statistics techniques are clearly hidden in the above model. The expectation of the i^{th} order statistic is given by

$$\mathrm{E}\left[X_{(i)}\right] = \frac{n!}{(i-1)!(n-i)!} \int_{\mathbb{R}} x F^{i-1}(x)\,(1-F(x))^{n-i}\, f(x)\,dx,$$

where $F(x)$ is the distribution in question. To get the standard deviation of the mean of the best k-out-of-n test scores requires computing the covariances between order statistics. We do not pursue this avenue as this closed form approach is very cumbersome for our purposes. A suitable resource text for order statistics is [2]. For further references the reader can consult [1] and [3], for example.

Variable means

It is natural that the student will change her test-writing ability throughout the semester. In particular, we hope it will improve. If we still assume constant student test-writing inconsistency throughout the term we will show that the results above hold unchanged with the appropriate mean adjustment. Assume now we draw the best k tests out of n with variable means $\{\mu_i\}_{i=1}^{n}$ but a constant standard deviation σ for each test. This is quite plausible as we do not expect the student test-writing consistency to change too much over one semester.

The model remains similar. The probability that all means of k elements (chosen out of n) are less than a given x is

$$F_{\{\mu_i\},\sigma}(x) = \int_L g_{\mu_1,\sigma}(t_1)\cdots g_{\mu_n,\sigma}(t_n)\,dt_1\cdots dt_n,$$

where $g_{\mu_i,\sigma}(t)$, for $i \in \{1, 2, \ldots, n\}$, is the normal probability density function

$$g_{\mu_i,\sigma}(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(t-\mu_i)^2}{2\sigma^2}},$$

and L is the region in $\mathbb{R}^n$ defined by the inequalities

$$t_{n_1} + t_{n_2} + \cdots + t_{n_k} < kx,$$

where $\{n_1, \ldots, n_k\}$ is any subset of $\{1, \ldots, n\}$. We make the substitution

$$z_i = \frac{t_i - \mu_i}{\sigma} \text{ so that } t_i = \sigma z_i + \mu_i.$$

Observe that

$$\frac{1}{k}\sum_{i=1}^{n} t_i = \sigma\frac{1}{k}\sum_{i=1}^{n} z_i + \frac{1}{k}\sum_{i=1}^{n}\mu_i.$$

Setting $\mu_a = \dfrac{1}{k}\displaystyle\sum_{i=1}^{n}\mu_i$ and $z = \dfrac{1}{k}\displaystyle\sum_{i=1}^{n} z_i$, we obtain the expected value of Y:

$$\begin{aligned}\mathbf{ave}_{\{\mu_i\},\sigma}(k,n) &= \int_{\mathbb{R}} x f_{\{\mu_i\},\sigma}(x)\,dx \\ &= \int_{\mathbb{R}} (\sigma z + \mu_a) f_{0,1}(z)\,dz \\ &= \sigma\cdot\mathbf{ave}_{0,1}(k,n) + \mu_a.\end{aligned}$$

The standard deviation remains unchanged as we have

$$\begin{aligned}\mathbf{std}^2_{\{\mu_i\},\sigma}(k,n) &= \int_{\mathbb{R}} (x - \mathbf{ave}_{\{\mu_i\},\sigma}(k,n))^2 f_{\{\mu_i\},\sigma}(x)\,dx \\ &= \int_{\mathbb{R}} \left(\sigma z + \mu_a - \left(\sigma\cdot\mathbf{ave}_{0,1}(k,n) + \mu_a\right)\right)^2 f_{0,1}(z)\,dz \\ &= \sigma^2\int_{\mathbb{R}} (z - \mathbf{ave}_{0,1}(k,n))^2 f_{0,1}(z)\,dz \\ &= \sigma^2\cdot\mathbf{std}^2_{0,1}(k,n).\end{aligned}$$

Clearly, if we average all the tests during the term we expect the mean to be μ_a. Therefore, the expected grade inflation in the time varying student ability case is the same as in the time constant ability case, which is $\mathbf{ave}_{0,1}(k,n)\sigma$.

Nested best k-out-of-n

Some professors may go further and implement nested selective choices. This could be accomplished, for example, by choosing the best questions within a given test and

then choose the best tests within the term. Of course, this could easily be generalized to further nesting. We will model this process once again assuming a constant σ, meaning the student does not change her test-writing inconsistency during the term. Furthermore, without loss of generality, we will present our model with a constant μ. We can replace μ by the corresponding μ_a, if necessary, but the expected grade inflation remains unchanged.

In general, one can select best k-out-of-n events followed next by best l-out-of-m events and so on. First, we assume that the random variable $Y_{k,n}$, the mean of the best k-out-of-n questions, is very close to being normally distributed. Simulations have shown that for reasonable choices of k and n this is a reasonable approximation. Therefore $Y_{k,n}$ is assumed to be normally distributed with mean μ' and standard deviation σ' where

$$\mu' = \mu + \sigma s_{k,n} \text{ and } \sigma' = \sigma \tilde{s}_{k,n},$$

for $s_{k,n} = \mathbf{ave}_{0,1}(k, n)$ and $\tilde{s}_{k,n} = \mathbf{std}_{0,1}(k, n)$.

Now selecting the best l-out-of-m from the $Y_{k,m}$ we obtain a normally distributed variable with mean μ'' and standard deviation σ'' where

$$\mu'' = \mu' + \sigma' s_{l,m} = \mu + \sigma \left(s_{k,n} + \tilde{s}_{k,n} s_{l,m}\right) \text{ and } \sigma'' = \sigma' \tilde{s}_{l,m} = \sigma \tilde{s}_{k,n} \tilde{s}_{l,m}.$$

For example, if we choose the best 4-out-of-5 questions within a test and then choose the best 3-out-of-5 tests during the term we get

$$\mu'' = \mu + 0.54\sigma \; ; \sigma'' = 0.23\sigma.$$

However, if we reverse the procedure and choose first the best 3-out-of-5 questions within a test and then choose the best 4-out-of-5 tests during the term we get different results

$$\mu'' = \mu + 0.70\sigma \; ; \sigma'' = 0.23\sigma,$$

which results in significantly higher expected grade inflation than the case before. Note the standard deviation $\sigma'' = 0.23\sigma$ is the same in both cases.

Choosing first the best 3-out-of-5 questions within a test yields about 30 % higher expected grade inflation than the reversed case. This model works in the time varying student ability case as well as long as we assume a constant student test-writing inconsistency during the term.

There is yet another interesting, possibly counter intuitive observation. Suppose we only implement the best 3-out-of-5 selection for tests with no nested selection within the tests. Recall in this case we have the expected grade inflation given by 0.55σ. Now suppose we choose the best 4-out-of-5 questions within a test and then choose the best 3-out-of-5 tests during the term. It is interesting to note that the corresponding expected grade inflation under this paradigm is slightly less than 0.55σ (numerical experiments show 0.547σ) which is the grade inflation due to non nested best 3-out-of-5 test selection during the term. This might seem counter intuitive at first glance. However, the standard deviation for the non nested best 3-out-of-5 term test case is given by 0.50σ as opposed to the significantly smaller standard deviation in the case of the best 4-out-of-5 questions within a test followed by the best 3-out-of-5 test selection during the term.

Indefinite nesting

Consider now the scenario where the educator decides to nest the best k-out-of-n several times over. For example three nested selection might look something like this.

Take the best k-out-of-n parts within a question, then choose the best k-out-of-n questions within a given test and then choose the best k-out-of-n tests with the term. This can be clearly generalized to further nesting and one may even consider the hypothetical indefinite nesting with the best k-out-of-n. The closed form expression for the expected grade inflation under this paradigm (indefinite best k-out-of-n nesting) can be easily obtained. Once again we work, without loss of generality, with the unchanging μ to estimate the corresponding expected grade inflation. In particular, the expected grade inflation when implementing the indefinite best k-out-of-n nesting is given by

$$\frac{\sigma \cdot \mathbf{ave}_{0,1}(k,n)}{1-\mathbf{std}_{0,1}(k,n)}.$$

To see this we observe, by induction,

$$\mu^{(n)} = \mu + \sigma s_{k,n}\left(1+\tilde{s}_{k,n}+\cdots+\tilde{s}_{k,n}^{n-1}\right) \text{ and } \sigma^{(n)} = \sigma\tilde{s}_{k,n}^{n},$$

where $s_{k,n} = \mathbf{ave}_{0,1}(k,n)$ and $\tilde{s}_{k,n} = \mathbf{std}_{0,1}(k,n)$. Thus we have

$$\mu_{\lim} = \lim_{n\to\infty}\mu^{(n)} = \mu + \frac{\sigma s_{k,n}}{1-\tilde{s}_{k,n}} \text{ and } \lim_{n\to\infty}\sigma^{(n)} = 0,$$

provided $\mathbf{std}_{0,1}(k,n) < 1$. It is intuitive that the standard deviation approaches zero as we further nest with the best k-out-of-n selection.

For example, consider the best 4-out-of-5 nested indefinitely. We have

$$s_{4,5} = 0.29 \text{ and } \tilde{s}_{4,5} = 0.46, \text{ and thus } \mu_{\lim} = \mu + 0.54\sigma.$$

This inflates the grade by about a half of the student test-writing inconsistency. However, if we consider the best 3-out-of-5 nested indefinitely we have

$$s_{3,5} = 0.55 \text{ and } \tilde{s}_{3,5} = 0.50 \text{ and thus } \mu_{\lim} = \mu + 1.10\sigma,$$

yielding a significant increase in the expected grade inflation; this time more than the student test-writing inconsistency.

REFERENCES

[1] Mendenhall, W., Scheaffer R. L., Wackerly, D. D. (1981). *Mathematical Statistics with Applications*, 2nd ed. Boston, MA: Duxbury Press.
[2] Ross, S. M. (2005). *A First Course in Probability*, 7th ed. Upper Saddle River, NJ: Prentice-Hall.
[3] Wani, J. K. (1971). *Probability and Statistical Inference*. New York: Appleton-Century-Crofts, Educational Division, Meredith Corporation.
[4] Zizler, P. (2013). Quiz today: should I skip class? *College Math. J.* 44(3):166–170.

Summary. When assessing student performance, some educators choose to discard the worst tests written by the student. The mean of the remaining tests is taken as opposed to the mean of all tests written. Naturally, this process will result in some expected grade inflation. In our paper we provide a model for the expected grade inflation in the case when the student's test-writing ability changes over time. Furthermore, we provide results on the grade inflation when the educator decides to use nested selective averaging processes.

MANDANA SOBHANZADEH (MR Author ID: 1320403) is an Assistant Professor in the Department of General Education at Mount Royal University in Calgary. Her research work is in Science, Technology, Engineering and Mathematics (STEM) education.

PETER ZIZLER (MR Author ID: 600813) is an Associate Professor in the Department of Mathematics & Computing at Mount Royal University in Calgary. His research work has been in linear algebra, statistics and mathematics education. In his spare time he likes to ride his Harley Davidson motorcycle as well as his electric bike.

A Simple and More General Approach to Stokes' Theorem

IOSIF PINELIS
Michigan Technological University
Houghton, MI 49931
ipinelis@mtu.edu

In calculus texts (see, e.g., [4, §XII.6] or [12, §16.8]), Stokes' theorem is usually stated as follows: Let S be an oriented smooth enough surface in $\mathbb{R}^3$ bounded by a simple closed smooth enough curve C with positive orientation. Then for any smooth enough vector field $\mathbf{F}$

$$\oint_C \mathbf{F} \cdot d\mathbf{r} = \iint_S \operatorname{curl} \mathbf{F} \cdot \mathbf{n}\, dS, \tag{1}$$

where $\operatorname{curl} \mathbf{F}$ is the curl of the vector field $\mathbf{F}$ and $\mathbf{n}$ is a continuous field of unit normal vectors on S. The positive orientation of C is "defined" there as the condition for the surface S to remain on the left (with respect to $\mathbf{n}$) when C is traced out. In such elementary texts, it is of course impossible to rigorously define such notions as "remains on the left" (see, e.g., [11, Theorems 5.5 and 5.9], [5, §XXIII.4–XXIII.6], or [2, §4.5.6]). The very concept of a surface in general is not elementary. A surface may be defined as a two-dimensional manifold-with-boundary [11]—possibly with singularities, which need to be considered to cover the case of even such simple surfaces as the (convex hulls of) triangles or rectangles (see, e.g., [5, §XXIII]). Moreover, for the surface integral on the right-hand side of identity (1) to make sense, one has to ensure that the curl and the unit normal vector field have appropriate invariance properties with respect to the choice of an atlas for the manifold S.

Green's theorem is the special and comparatively very simple case of Stokes' theorem corresponding to the additional condition that the surface S is flat and thus may be assumed to coincide with a region $D \subset \mathbb{R}^2$. Informally, Green's theorem may be stated as follows:

Let D be a compact subset of $\mathbb{R}^2$ with a smooth enough boundary ∂D, and let P and Q be real continuously differentiable functions on an open neighborhood of D. Then

$$\oint_{\partial D} P\, du + Q\, dv = \iint_D \Big(\frac{\partial Q}{\partial u} - \frac{\partial P}{\partial v}\Big) du\, dv, \tag{2}$$

with the line integral $\oint_{\partial D} P\, du + Q\, dv$ appropriately defined.

One way to make this statement rigorous (and general enough) is to assume that the boundary ∂D of D is a rectifiable Jordan curve Γ oriented so as to make the winding number of Γ equal 1 (rather than -1) with respect to any point in the interior of Γ. These conditions on D and ∂D will be assumed in the sequel.

Even though Green's theorem is only a special case of Stokes', it is not easy to prove the just mentioned rigorous "Jordan curve" version of it, or even to show that the winding number can be consistently defined; see, e.g., [6] and references there to [1, 8, 9, 13]. The best practical approach to teaching Green's theorem in a calculus course would probably be to restrict the consideration to simple regions, of the form

Math. Mag. **92** (2019) 345–351. doi:10.1080/0025570X.2019.1660574
MSC: 26B20, 00A05, 00A06

$\{(x, y)\colon a \le x \le b, g_1(x) \le y \le g_2(x)\}$ or $\{(x, y)\colon a \le y \le b, g_1(y) \le x \le g_2(y)\}$, and possibly simple combinations thereof.

Anyway, in this note we shall show that, once an appropriate version of Green's theorem is established, it is then very easy and painless to derive versions of Stokes' theorem, which are even more general, in some aspects, than the conventional one.

Indeed, for some open neighborhood U of D and $(u, v) \in U$, let

$$(u, v) \longmapsto \mathbf{r}(u, v) \in \mathbb{R}^3$$

be a twice continuously differentiable map of U into $\mathbb{R}^3$. Let $C := \mathbf{r}(\partial D)$ be the image of the boundary ∂D of D under the map $\mathbf{r}$. Let

$$\mathbf{G}\colon U \to \mathbb{R}^3$$

by a continuously differentiable vector field on U. Since $d\mathbf{r} = \mathbf{r}_u\, du + \mathbf{r}_v\, dv$, we can write/define the line integral $\oint_C \mathbf{G} \cdot d\mathbf{r}$ as follows:

$$\oint_C \mathbf{G} \cdot d\mathbf{r} = \oint_{\partial D} (\mathbf{G} \cdot \mathbf{r}_u)\, du + (\mathbf{G} \cdot \mathbf{r}_v)\, dv. \tag{3}$$

Here, as usual, the subscripts $_u, _v, _x, \ldots$ denote the partial differentiation with respect to $u, v, x, \ldots$. One may say that formula (3) reduces the "line integral" $\oint_C \mathbf{G} \cdot d\mathbf{r}$ over the "image-curve" C (which does not have to be a simple curve, without self-intersections) to the well-defined line integral $\oint_{\partial D} (\mathbf{G} \cdot \mathbf{r}_u)\, du + (\mathbf{G} \cdot \mathbf{r}_v)\, dv$ over the properly oriented rectifiable Jordan curve $\Gamma = \partial D$ in the domain U of the map $\mathbf{r}$.

Letting now $P = \mathbf{G} \cdot \mathbf{r}_u$ and $Q = \mathbf{G} \cdot \mathbf{r}_v$, we have

$$\frac{\partial Q}{\partial u} - \frac{\partial P}{\partial v} = \mathbf{G}_u \cdot \mathbf{r}_v + \mathbf{G} \cdot \mathbf{r}_{vu} - \mathbf{G}_v \cdot \mathbf{r}_u - \mathbf{G} \cdot \mathbf{r}_{uv} = \mathbf{G}_u \cdot \mathbf{r}_v - \mathbf{G}_v \cdot \mathbf{r}_u,$$

since $\mathbf{r}_{vu} = \mathbf{r}_{uv}$. So, Green's theorem (2) immediately yields the following form of Stokes' theorem:

$$\oint_C \mathbf{G} \cdot d\mathbf{r} = \iint_D \big(\mathbf{G}_u \cdot \mathbf{r}_v - \mathbf{G}_v \cdot \mathbf{r}_u\big) du\, dv, \tag{4}$$

which, similarly to (1), reduces a line integral to a double one.

This may be compared with the more conventional form of Stokes' theorem:

$$\oint_C \mathbf{F} \cdot d\mathbf{r} = \iint_D (\nabla \times \mathbf{F}) \cdot (\mathbf{r}_u \times \mathbf{r}_v)\, du\, dv, \tag{5}$$

where $\mathbf{F}$ is a continuously differentiable 3D vector field defined on a neighborhood of the "surface"

$$S := \mathbf{r}(D),$$

and the line integral may be understood as the one in (4) (or, equivalently, in (3)) with

$$\mathbf{G} = \mathbf{F} \circ \mathbf{r}. \tag{6}$$

It is not hard to deduce (5) from (4). Indeed, let $(\mathbf{i}, \mathbf{j}, \mathbf{k})$ be any orthonormal basis in $\mathbb{R}^3$, in which the cross products $\nabla \times \mathbf{F}$ and $\mathbf{r}_u \times \mathbf{r}_v$ can be computed according to the standard determinant formulas. Let $\langle x(u, v), y(u, v), z(u, v)\rangle$ and $\langle f(x, y, z), g(x, y, z), h(x, y, z)\rangle$ be, respectively, the triples of the coordinates of $\mathbf{r}(u, v)$ and $\mathbf{F}(x, y, z)$ in the basis $(\mathbf{i}, \mathbf{j}, \mathbf{k})$. Since both sides of (5) are linear in $\mathbf{F}$,

without loss of generality $f = g = 0$, and then the integrands on the right-hand sides of (4) and (5) become, respectively,

$$\begin{aligned} I(u, v) &:= (h_x x_u + h_y y_u + h_z z_u) z_v - (h_x x_v + h_y y_v + h_z z_v) z_u \\ &= (h_x x_u + h_y y_u) z_v - (h_x x_v + h_y y_v) z_u \end{aligned}$$

and

$$\begin{aligned} J(u, v) &:= (h_y \mathbf{i} - h_x \mathbf{j}) \cdot \big((y_u z_v - z_u y_v)\mathbf{i} + (z_u x_v - x_u z_v)\mathbf{j} + (x_u y_v - y_u z_v)\mathbf{k} \big) \\ &= h_y (y_u z_v - z_u y_v) - h_x (z_u x_v - x_u z_v). \end{aligned}$$

Now it is quite easy to see that $I(u, v) = J(u, v)$, which completes the derivation of (5) from (4).

The main distinction of (4) and (5) from (1) is that the double integrals in (4) and (5) are taken over the flat preimage-region $D \subset \mathbb{R}^2$—whereas the "image-surface" $S = \mathbf{r}(D)$ plays no role in (4) and (5), except that the vector field $\mathbf{F}$ has to be defined on some neighborhood of S.

Therefore, as far as (4) and (5) are concerned, there is no need to talk about any properties of the image-surface S except for it being a subset of $\mathbb{R}^3$. In particular, there is no need to talk about the orientation of S or to use such hard to define terms as "remains on the left" or to care about the mentioned invariance properties of the curl and the unit normal vector field. Moreover, the image-surface $S = \mathbf{r}(D)$, as well as the image-curve $C = \mathbf{r}(\partial D)$, may be self-intersecting, and S does not have to be a manifold at all. As for the line integrals in (4) and (5), they are over the image-curve C only in form, as they immediately reduce to line integrals over the pre-image curve ∂D, according to (3). This reduction of integration in the image-space $\mathbb{R}^3$ to that in the flat preimage-space $\mathbb{R}^2$ is natural; it is even unavoidable—for how else would one actually compute the line and surface integrals in the conventional form (1) of Stokes' theorem?

(Of course, to do the integration in the preimage-space, we still need a proper orientation of the boundary ∂D of the *flat preimage* domain D, as was already pointed out in our discussion concerning a general "Jordan curve" version of Green's theorem and its elementary version for simple regions.)

Next, let us consider (4) versus (5).

We saw that (4) is very easy to obtain, modulo Green's theorem. Arguably, (4) is also easier to remember than (5).

An important advantage of (4) is that it is more general than (5). Indeed, on the one hand, (5) follows from (4); on the other hand, in (4) the composition-factorization (6) of the map $\mathbf{G}$ is not needed. That is, for (4) one does not need the implication $\mathbf{r}(u_1, v_1) = \mathbf{r}(u_2, v_2) \implies \mathbf{G}(u_1, v_1) = \mathbf{G}(u_2, v_2)$ (which necessarily follows from (6)).

Plus, one does not need the notion of the curl for (4). On the other hand, (4) by itself will not help when proving that a vector field is conservative if its curl is zero. Yet, as shown above, (5) is rather easy to get from (4).

We thus have three versions of Stokes' theorem, corresponding to the three formulas:

- the conventional version (1), which requires comparatively most stringent and even hard to define conditions, including an appropriate orientability of the image-surface together with the image-curve, and the invariance of the curl and the unit normal vector field;
- the "computational," less conventional version (5), which is not concerned with orientability of the images, but needs the composition-factorization condition (6);

- (4), which is the most general of the three versions and, at the same time, easiest to obtain—but not useful when, say, the curl is known to be zero.

The above discussion is illustrated by the following example.

Example. Consider the Möbius strip S_δ (of "radius" 1 and half-width $\delta > 0$), which is the image $\mathbf{r}(D_\delta)$ of the rectangle

$$D_\delta := [0, 2\pi] \times [-\delta, \delta] \tag{7}$$

under the map $\mathbf{r}\colon D_\delta \to \mathbb{R}^3$ given by the formula (see, e.g., [10])

$$\mathbf{r}(u, v) = \big((1 + v\cos\tfrac{u}{2})\cos u, (1 + v\cos\tfrac{u}{2})\sin u, v\sin\tfrac{u}{2}\big). \tag{8}$$

The image-curve $C_\delta := \mathbf{r}(\partial D_\delta)$ is self-intersecting, a reason being that for all $v \in [-\delta, \delta]$ one has $\mathbf{r}(0, v) = \mathbf{r}(2\pi, -v)[= (1 + v, 0, 0)]$—whereas $(0, v) \in \partial D_\delta$, $(2\pi, -v) \in \partial D_\delta$, and $(0, v) \neq (2\pi, -v)$. For this reason, the "surface" $S_\delta = \mathbf{r}(D_\delta)$ may be considered self-intersecting as well. If $\delta > 2$, then S_δ is self-intersecting in another, apparently more interesting manner (see [7, Theorem 1.9]).

Indeed, assume that $\delta > 2$ and let $u_1 = u$, $u_2 = u + \pi$, $v_1 = -2\cos\frac{u}{2}/\cos u$, and $v_2 = -2\sin\frac{u}{2}/\cos u$, where u is a small enough positive real number. Then $\mathbf{r}(u_1, v_1) = \mathbf{r}(u_2, v_2)[= (-1, -\tan u, -\tan u) \approx (-1, 0, 0)]$ – whereas $(u_1, v_1) \in D_\delta$, $(u_2, v_2) \in D_\delta$, and $(u_1, v_1) \neq (u_2, v_2)$. Similarly, letting $u_1 = 2\pi - u$, $u_2 = \pi - u$, $v_1 = 2\cos\frac{u}{2}/\cos u$, and $v_2 = 2\sin\frac{u}{2}/\cos u$ for small enough $u > 0$, we have $\mathbf{r}(u_1, v_1) = \mathbf{r}(u_2, v_2)[= (-1, \tan u, \tan u)]$.

Letting now, for instance, $\mathbf{G}(u, v) := (u^2, 0, 0)$ for all (u, v), we will have $\mathbf{G}(u_1, v_1) \neq \mathbf{G}(u_2, v_2)$ for all pairs of points (u_1, v_1) and (u_2, v_2) as above. Therefore, at the self-intersection points $\mathbf{r}(u_1, v_1) = \mathbf{r}(u_2, v_2) = (-1, \varepsilon\tan u, \varepsilon\tan u)$ with $\varepsilon = \pm 1$ and small enough $u > 0$, one will be unable to define a vector field $\mathbf{F}$ so that (6) hold. Thus, formula (5) is not applicable here; of course, formula (1) is not applicable either, because the Möbius strip is not orientable.

In contrast, (4) applies with no problem in this situation; each side of (4) evaluates here to $-160\delta/9$.

The problem of application of Stokes' theorem to the Möbius strip was previously considered in [3]. The version of the Möbius strip dealt with in [3] differs by a composition of an isometry and a homothety from the apparently more common version described by (8). So, in the subsequent discussion the considerations in [3] will be translated into terms corresponding to (8). Following [3], let us introduce here the vector field $\mathbf{F}$ defined by the formula

$$\mathbf{F}(x, y, z) = \Big(\frac{-y}{x^2 + y^2}, \frac{x}{x^2 + y^2}, 0\Big) \tag{9}$$

for $(x, y, z) \in \mathbb{R}^3$ such that $x^2 + y^2 \neq 0$. It is noted in [3] that (i) $\operatorname{curl}\mathbf{F} = \mathbf{0}$ wherever the vector field $\mathbf{F}$ is defined and (ii) the Möbius strip does not intersect the z-axis if $\delta < 1$. So, $\operatorname{curl}\mathbf{F} = \mathbf{0}$ on the Möbius strip.

It is then concluded in [3] that the surface integral on the right-hand side of (1) is 0. However, this conclusion is not quite correct, because the Möbius strip is not orientable and hence the normal vector field $\mathbf{n}$ cannot be appropriately defined on the strip, so that the surface integral is technically not defined either and thus has no value.

Also, it is observed in [3] that the boundary (say B) of the Möbius strip is the image of the interval $[0, 4\pi]$ under the map

$$u \mapsto \tilde{\mathbf{r}}(u) := \mathbf{r}(u, \delta), \tag{10}$$

and that the line integral $\oint_B \mathbf{F} \cdot d\tilde{\mathbf{r}}$ over the so-parameterized curve B is 4π, which differs from the presumed value 0 of the actually undefined surface integral. This discrepancy is ascribed in [**3**] to the non-orientability of the Möbius strip. Now one may wonder as follows:

We were told in this note that, as far as (4) and (5) are concerned, there is no need to talk about the orientation of the images S and C of D and ∂D under the map $\mathbf{r}$. If so, then the version (5) of Stokes' theorem must hold even for the Möbius strip and the vector field $\mathbf{F}$ as in (9). The double integral on the right-hand side of (5)—in contrast to that on the right-hand side of (1)—is well defined, and it must be 0, since $\operatorname{curl} \mathbf{F} = \mathbf{0}$ on the Möbius strip. But the line integral over the boundary B of the Möbius strip was found in [**3**] to be nonzero. Does this not contradict (5)?

In fact, there is no contradiction here. Recall that the line integral in (5) was to be understood as the one in (3) $\big($with $\mathbf{G}$ as in (6)$\big)$ and, in turn, the line integral in (3) over the "image-curve" C was defined as the corresponding line integral over the preimage ∂D of C, where ∂D is the boundary of the preimage-region D. In contrast with these conditions, the line integral in [**3**] was essentially taken over the segment $[0, 4\pi] \times \{\delta\}$, which is of course *not* the boundary of $D = D_\delta = [0, 2\pi] \times [-\delta, \delta]$. However, the value of the line integral in (5) computed indeed as the one in (3) with $\mathbf{G}$ as in (6) is 0, which is of course the same as the value of the double integral in (5).

At this point, one may still wonder:

The difficulty with the applicability of Stokes' theorem to the Möbius strip was ascribed in [**3**] to the non-orientability. However, the boundary B of the (non-orientable) Möbius strip can also be the boundary of an orientable surface. Then the line integral $\oint_B \mathbf{F} \cdot d\tilde{\mathbf{r}}$ as computed in [**3**] will have the same value $4\pi \neq 0$, whereas the corresponding surface integral on the right-hand side of (1) will still be 0, right? So, it seems we have another contradiction here.

The answer to this concern is as follows.

First here, indeed the boundary B of the Möbius strip can also be the boundary of an orientable surface. To see this, it is more convenient to use the interval $[-\pi, 3\pi]$ instead of $[0, 4\pi]$.

More specifically, recall (8) and note that, as u increases from $-\pi$ to π, the z-coordinate $\delta \sin \frac{u}{2}$ of the vector $\mathbf{r}(u, \delta)$ increases from its minimal value $-\delta$ to its maximal value δ; and as u increases further from π to 3π, the z-coordinate of $\mathbf{r}(u, \delta)$ decreases from δ back to $-\delta$. Moreover, each value of the z-coordinate of $\mathbf{r}(u, \delta)$ in the interval $[-\delta, \delta)$ is taken at exactly two points $u_1(z)$ and $u_2(z)$ in $[-\pi, 3\pi]$ such that $-\pi \leq u_1(z) < \pi < u_2(z) = 2\pi - u_1(z) \leq 3\pi$.

The value δ of the z-coordinate of $\mathbf{r}(u, \delta)$ for $u \in [-\pi, 3\pi]$ is taken only at $u = \pi$; accordingly, assume that $u_1(\delta) = u_2(\delta) = \pi$. Note also that $2\pi - u$ decreases from 3π to π as u increases from $-\pi$ to π. Connecting now, for each $z \in [-\delta, \delta]$, the points $\mathbf{r}(u_1(z), \delta)$ and $\mathbf{r}(u_2(z), \delta)$ by (say) a straight line segment, we do obtain an orientable surface, say $\hat{S}$, whose boundary is the same as the boundary B of the Möbius strip. The surface $\hat{S}$ is the image $\hat{\mathbf{r}}(\hat{D})$ of the rectangle $\hat{D} := [-\pi, \pi] \times [0, 1]$ under the map

$$\hat{D} \ni (u, t) \mapsto \hat{\mathbf{r}}(u, t) := (1 - t)\mathbf{r}(u, \delta) + t\, \mathbf{r}(2\pi - u, \delta).$$

To verify that $\hat{S}$ is orientable, note that the first two coordinates of the vector $\mathbf{c} := \hat{\mathbf{r}}_u(u, t) \times \hat{\mathbf{r}}_t(u, t)$ are $\delta \cos \frac{u}{2} \sin u$ and $-\delta^2 \cos^2 \frac{u}{2} \cos u$, respectively, whence $\mathbf{c} \neq \mathbf{0}$ for all (u, t) in the interior of $\hat{D}$, and so, one can define the unit normal vector field on the image of the interior of $\hat{D}$ under the map $\hat{\mathbf{r}}$ by the formula $\mathbf{n} = \mathbf{c}/|\mathbf{c}|$.

Moreover and more importantly, the image $\hat{\mathbf{r}}(\partial \hat{D})$ under this map of the boundary $\partial \hat{D}$ of the rectangle $\hat{D}$ is the boundary B of the Möbius strip. So, (5) will hold for any 3D vector field $\mathbf{F}$ which is smooth enough, in the sense of being continuously

differentiable on a neighborhood of the surface $\hat{S} = \hat{\mathbf{r}}(\hat{D})$. For instance, for the linear vector field defined by the formula $\mathbf{F}(x, y, z) = A(x\ y\ z)^\top$, where $A = (a_{i,j})_{i,j=1}^3$ is a constant 3×3 real matrix and $^\top$ denotes the transposition, both the left-hand side and right-hand side of (5) with $D = \hat{D}$ $\big($and $C = \hat{\mathbf{r}}(\partial\hat{D})\big)$ take the same value, $2\pi(2 + \delta^2)(a_{1,2} - a_{2,1}) + \pi\delta^2(a_{1,3} - a_{3,1})$.

Is the particular field $\mathbf{F}$ given by (9) continuously differentiable on a neighborhood of the surface $\hat{S}$? In other words, is the intersection of the surface $\hat{S}$ with the z-axis empty? If that were so, then the right-hand of (5) would be 0—whereas, as one can recheck, the value of the line integral in (5) is the same nonzero value, 4π, as the one found the other way in [**3**]. It follows that the surface $\hat{S}$ does intersect the z-axis. In fact, it is not hard to check directly that this intersection contains exactly two points, $(0, 0, \pm\delta/\sqrt{2})$.

A free bonus of this discussion is the fact that any orientable surface in $\mathbb{R}^3$ whose boundary coincides with the boundary of the Möbius strip given by (8) must necessarily intersect the z-axis. One may also note here that, for any such orientable surface and for $\mathbf{F}$ as in (9), the surface integral on the right-hand side of (1) will actually not be 0; rather, it will be undefined.

This discussion is partly illustrated in Figure 1.

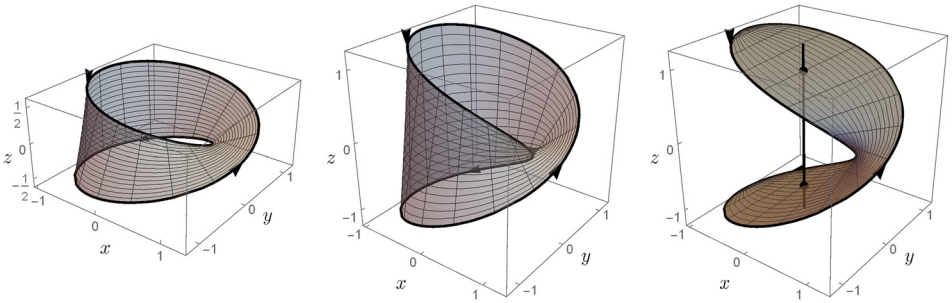


Figure 1 Left panel: The Möbius strip given by (8) (with $\delta = 3/10$), stretched vertically by the factor of 2. Middle panel: The same Möbius strip, now stretched vertically by the factor of 4. Right panel: The oriented surface $\hat{S}$ with the same boundary as the Möbius strip, stretched vertically by the factor of 4, for better viewing; shown here are also the two points of intersection of the surface $\hat{S}$ with the z-axis.

REFERENCES

[1] Apostol, T. M. (1957). *Mathematical Analysis: A Modern Approach to Advanced Calculus*. Reading, MA: Addison-Wesley Publishing Company, Inc.
[2] Federer, H. (1969). *Geometric Measure Theory*. Die Grundlehren der mathematischen Wissenschaften, Band 153. New York: Springer-Verlag New York Inc.
[3] Gabardo, J.-P. (2006). The Möbius strip and Stokes' theorem. ms.mcmaster.ca/gabardo/moebius.pdf
[4] Lang, S. (1987). *Calculus of Several Variables*, 3rd ed. New York: Springer.
[5] Lang, S. (1993). *Real and Functional Analysis*. Graduate Texts in Mathematics, Vol. 142, 3rd ed. New York: Springer-Verlag.
[6] MathOverflow. (2018). Proof of Green's formula for rectifiable Jordan curves. mathoverflow.net/questions/307713/proof-of-greens-formula-for-rectifiable-jordan-curves
[7] Pinelis, I. (2018). A simpler $\mathbb{R}^3$ realization of the Möbius strip. arxiv.org/abs/1808.03955
[8] Potts, D. H. (1951). A note on Green's theorem. *J. Lond. Math. Soc.* 26: 302–304.
[9] Ridder, J. (1941). Über den Greenschen Satz in der Ebene. *Nieuw Arch. Wiskunde (2)*. 21: 28–32.
[10] Schwarz, G. E. (1990). The dark side of the Moebius strip. *Amer. Math. Monthly*. 97(10): 890–897.
[11] Spivak, M. (1965). *Calculus on Manifolds. A Modern Approach to Classical Theorems of Advanced Calculus*. New York/Amsterdam: W. A. Benjamin, Inc.

[12] Stewart, J. (2016). *Calculus*, 8th ed. Boston, MA: Cengage.
[13] Verblunsky, S. (1949). On Green's formula. *J. Lond. Math. Soc.* 24: 146–148.

Summary. Oftentimes, Stokes' theorem is derived by using, more or less explicitly, the invariance of the curl of the vector field with respect to translations and rotations. However, this invariance—which is oftentimes described as the curl being a "physical" vector—does not seem quite easy to verify, especially for undergraduate students. An even bigger problem with Stokes' theorem is to rigorously define such notions as "the boundary curve remains to the left of the surface." Here an apparently simpler and more general approach is suggested.

IOSIF PINELIS (MR Author ID: 208523) has been with Michigan Technological University since 1992. He has also held visiting positions at the University of Illinois, Urbana–Champaign; the City University of New York; and Lehigh University. His most extensive expertise is in probability and statistics, including extremal problems, exact inequalities, and limit theorems of probability and statistics. He has also enjoyed doing work in machine learning, information theory, numerical analysis, operations research, combinatorics, mechanical engineering, biology, geometry, and physics.

L	A	G	S			G	A	G			B	F	F	
A	L	O	T		A	U	D	R	A		E	A	R	P
F	L	U	E		M	A	H	E	S	W	A	R	A	N
F	E	D	E	R	I	C	O		S	A	T	O	U	T
	N	A	P	A			C	L	A	S	S			
			L	I	A	R		A	D	A	M	S	O	N
E	L	L	E	N	B	E	R	G			E	I	N	E
X	I	I			A	G	N	E	W			L	E	T
A	M	I	D			G	A	R	I	B	A	L	D	I
M	B	I	R	I	K	A		S	T	E	M			
			A	R	I	E	S			S	E	A	S	
C	O	O	C	O	O		W	E	S	T	R	I	C	K
R	A	D	U	N	S	K	A	Y	A		I	K	E	A
I	S	I	L		K	E	Y	E	D		C	E	N	T
	T	E	A			L	S	D			A	N	T	E

Polynomials of Magic Matrices

ALAN F. BEARDON
University of Cambridge
Cambridge CB3 0WB, England
afb@dpmms.cam.ac.uk

A *magic square* M is an $n \times n$ matrix with entries $1, \ldots, n^2$ such that the sums over each of the n rows, each of the n columns, and each of the two diagonals, have the same value. There are many different subclasses of magic squares, and the construction, and enumeration, of all types of magic squares presents a substantial challenge, largely because of the lack of an algebraic structure on the set $\{1, \ldots, n^2\}$. A mathematical treatment of magic squares is more fruitful if we allow their entries to be real numbers, and then consider them to be elements of a vector space of real matrices, in which case we shall use the term *magic matrix*. For us, then, a magic matrix is a real $n \times n$ matrix M such that the sums over each of its n rows, each of its n columns, and each of its two diagonals, have the same value. We emphasize that, in contrast to magic squares, the elements of a magic matrix *need not be integers, nor distinct*. The word "magic" was first introduced in this context by Frenicle de Bessy [**6**] in 1629; more recently the term "magic matrix" rather than "magic square" was used in [**4**, **5**].

It is well known that if M is a 3×3 magic matrix, then $M^3, M^5, M^7, \ldots$ are also magic matrices. Since M^2 need not be a magic matrix, there have been many attempts to describe which powers of an $n \times n$ magic matrix are magic matrices (see, e.g., [**2**, **4**, **8**, **9**, **15**]). However, there seems to be no good reason why we should restrict ourselves to powers of M, and a more general discussion of polynomials in M allows us to make further progress. Here we shall show how, given an $n \times n$ magic matrix M, we can find all polynomials f such that $f(M)$ is a magic matrix. The cases $n = 1$ and $n = 2$ will left to the reader so, throughout, we shall assume that $n \geq 3$.

Some preliminary remarks

Throughout, let $\mathcal{I}$ be the $n \times n$ identity matrix, and $\mathcal{O}$ be the $n \times n$ zero matrix. Each $n \times n$ constant matrix is a scalar multiple of the $n \times n$ magic matrix $\mathcal{J}$, where

$$\mathcal{J} = \frac{1}{n}\begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{pmatrix},$$

$\mathcal{J}^2 = \mathcal{J}$. We should perhaps have attached a subscript n to $\mathcal{I}$, $\mathcal{O}$ and $\mathcal{J}$ but n will always be clear from the context.

The *trace* $\tau(M)$ of the $n \times n$ matrix M, where $M = (m_{ij})$ M, is the sum over the diagonal of M, namely $m_{11} + \cdots + m_{nn}$, and the *reversed trace* $\rho(M)$ of M is the sum over the reversed diagonal of M, namely $m_{1n} + \cdots + m_{n1}$. The trace function τ is a very special linear function; indeed, it is the only linear function θ on the space of real $n \times n$ matrices that satisfies the conditions $\theta(M_1M_2) = \theta(M_2M_1)$ and $\theta(\mathcal{I}) = n$; see [**12**, p. 99]. There does not seem to be analogous description for ρ, and most of the difficulties seem to arise from ρ.

Math. Mag. **92** (2019) 352–358. doi:10.1080/0025570X.2019.1660131
MSC: Primary 05B15, Secondary 11C99; 15B99

Semi-magic matrices

The key idea is to use semi-magic matrices. A square matrix M is a *semi-magic matrix* if its row sums and column sums (but not necessarily its diagonal sums) have the same value, and this common value is the *magic constant* $\mu(M)$ of M. Note that $\mu(\mathcal{I}) = 1$, $\mu(\mathcal{O}) = 0$ and $\mu(\mathcal{J}) = 1$.

It is clear that the magic constant function $\mu(M)$ of a semi-magic matrix M is linear: that is, $\mu(M_1 + M_2) = \mu(M_1) + \mu(M_2)$ and $\mu(kM) = k\mu(M)$. It is also important to note that *the product of semi-magic matrices is a semi-magic matrix*, and

$$\mu(M_1M_2) = \mu(M_1)\mu(M_2) \tag{1}$$

which we shall now prove. Let A and B be semi-magic matrices, and write $A = (a_{ij})$, $B = (b_{ij})$ and $AB = C = (c_{ij})$. Then, for each i,

$$\sum_j c_{ij} = \sum_{j,r} a_{ir}b_{rj} = \sum_r \left(a_{ir} \sum_j b_{rj} \right) = \sum_r a_{ir}\mu(B) = \mu(A)\mu(B).$$

A similar argument (which we omit) holds for the sums over the columns of C.

Finally, suppose that M is an $n \times n$ semi-magic matrix. Then N, defined by $N = M - \mu(M)\mathcal{J}$, is a semi-magic matrix with $\mu(N) = 0$. We call N the *null part* of M (for more information on the algebraic structure here, see [13, 16]). Obviously, if $M_1, \ldots, M_p$ are semi-magic matrices with decompositions $M_j = \mu_j\mathcal{J} + N_j$, where $\mu_j = \mu(M_j)$, then $N_1 + \cdots + N_p$ is the null part of $M_1 + \cdots + M_p$. We can also form products of semi-magic matrices (of the same size), and in this case we have

$$M_1 \cdots M_p = (\mu_1\mathcal{J} + N_1) \cdots (\mu_p\mathcal{J} + N_p) = \mu_1 \cdots \mu_p\mathcal{J} + N_1 \cdots N_p.$$

Indeed, (1) shows that $\mu(M_1 \cdots M_p) = \mu_1 \cdots \mu_p$, and the fact that the null part of $M_1 \cdots M_p$ is $N_1 \cdots N_p$ is a trivial consequence of the fact that if N is any semi-magic matrix with magic constant 0, then $\mathcal{J}N = N\mathcal{J} = \mathcal{O}$. In particular, if M is a semi-magic matrix then $M^k = \mu(M)^k\mathcal{J} + N^k$ for $k \geq 1$, while $M^0 = \mathcal{I}$. This leads easily to the following result.

Lemma 1. *Let f be a polynomial, and let M be an $n \times n$ semi-magic matrix with magic constant μ and null part N. Then $f(M) = f(N) + [f(\mu) - f(0)]\mathcal{J}$. In particular, $f(M)$ is a magic matrix if and only if $f(N)$ is.*

The main result

Given an $n \times n$ magic matrix M, and a polynomial f, we want to know whether $f(M)$ is magic or not. The first step is to simplify the matrix M; the second step is to simplify the polynomial f; the third step is to decide whether or not (in the simplified situation) $f(M)$ is a magic matrix.

Step 1: Lemma 1 allows us to replace M by its null part N which has magic constant 0, for $f(M)$ is a magic matrix if and only if $f(N)$ is.

Step 2: Let χ_N be the characteristic polynomial of N (the null part of M). Then χ_N has degree n and, by the Cayley–Hamilton theorem, $\chi_N(N) = \mathcal{O}$. We now write

$$f(x) = a(x)\chi_N(x) + b(x),$$

where $a(x)$ and $b(x)$ are polynomials, and where b has degree q with $q < n$. As $f(N) = b(N)$, we see that $f(M)$ is magic if and only if $b(N)$ is magic.

Step 3: Suppose that $b(x) = b_0 + b_1x + \cdots + b_qx^q$ with $q < n$. Because b_1N is a magic matrix, we let $B(x) = b(x) - b_1x$; thus $b(N)$ is a magic matrix if and only if $B(N)$ is a magic matrix. Then $f(M)$ is a magic matrix if and only if $B(N)$ is a magic matrix, where

$$B(N) = b_0\mathcal{I} + \big(b_2N^2 + \cdots + b_qN^q\big).$$

Now $B(N)$ is a semi-magic matrix with magic constant b_0; thus $B(N)$, hence also $f(M)$, is a magic matrix if and only if $\tau\big(B(N)\big) = b_0 = \rho\big(B(N)\big)$, or equivalently,

$$\tau\big(b_0\mathcal{I} + b_2N^2 + \cdots + b_qN^q\big) = b_0,$$
$$\rho\big(b_0\mathcal{I} + b_2N^2 + \cdots + b_qN^q\big) = b_0.$$

If we now let $\tau_j = \tau(N^j)$ and $\rho_j = \rho(N^j)$ for $j = 1, 2, \ldots$, and note that

$$\tau(\mathcal{I}) = n, \quad \rho(\mathcal{I}) = \varepsilon_n = \begin{cases} 0 & \text{if } n \text{ is even;} \\ 1 & \text{if } n \text{ is odd,} \end{cases}$$

we obtain the following result.

Theorem 1. *In the notation above, $f(M)$ is a magic matrix if and only if*

$$(n-1)b_0 + b_2\tau_2 + \cdots + b_q\tau_q = 0,$$
$$b_0(\varepsilon_n - 1) + b_2\rho_2 + \cdots + b_q\rho_q = 0.$$

The solutions $(b_0, b_2, \ldots, b_q)$ of these two equations depend only on the numbers

$$n, \tau(N^2), \ldots, \tau(N^{n-1}), \rho(N^2), \ldots, \rho(N^{n-1}),$$

and they provide all polynomials f of degree less than n (but ultimately all polynomials) such that $f(M)$ is a magic matrix. Note that, in general, the set of these solutions forms a vector space of dimension $q - 2$, so there should be many polynomials f such that $f(M)$ is a magic matrix. We illustrate these ideas with two examples.

Example 1. Let

$$M = \begin{pmatrix} 10 & 2 & 14 & 18 \\ 17 & 15 & 1 & 11 \\ 13 & 21 & 7 & 3 \\ 4 & 6 & 22 & 12 \end{pmatrix}$$

Then M (which appears in [**15**, p. 341]) is a magic matrix with $\mu(M) = 44$, and

$$N = \begin{pmatrix} -1 & -9 & 3 & 7 \\ 6 & 4 & -10 & 0 \\ 2 & 10 & -4 & -8 \\ -7 & -5 & 11 & 1 \end{pmatrix}.$$

Then we find (from a computer) that $\tau_2 = -536$, $\rho_2 = 8$, $\tau_3 = 432$ and $\rho_3 = -144$, so that the two equations in Theorem 1 become

$$3b_0 - 536b_2 + 432b_3 = 0,$$
$$-b_0 + 8b_2 - 144b_3 = 0.$$

These have the general solution $b_2 = 0$ and $b_0 = -144b_3$, and as $\chi_n(x) = x^4 + 268x^2 - 144x$, we find that $f(M)$ is a magic matrix if and only if for some real numbers b_1 and b_3, and some polynomial $p(x)$, we have

$$f(x) = b_1x + b_3(x^3 - 144) + p(x)(x^4 + 268x^2 - 144x).$$

Example 2. Let

$$M = \begin{bmatrix} 16 & 3 & 2 & 13 \\ 5 & 10 & 11 & 8 \\ 9 & 6 & 7 & 12 \\ 4 & 15 & 14 & 1 \end{bmatrix}$$

(this appears in Albrecht Dürer's engraving *Melencholia* in 1514). Then M has magic constant 34, and null part N, where

$$N = \frac{1}{2}\begin{pmatrix} 15 & -11 & -13 & 9 \\ -7 & 3 & 5 & -1 \\ 1 & -5 & -3 & 7 \\ -9 & 13 & 11 & 15 \end{pmatrix},$$

and $\chi_N(x) = x^2(x^2 - 64)$. However, N has minimal polynomial $x(x^2 - 64)$ (as can be seen by showing that $N^3 = 64N$). In this case we can argue as above except that now we use the minimal polynomial for N instead of χ_N. A calculation shows that $\tau_2 = 128$ and $\rho_2 = 0$, so the two equations in Theorem 1 are

$$3b_0 + 128b_2 = 0,$$
$$-b_0 + 0b_2 = 0.$$

So that $b_0 = b_2 = 0$. We conclude that $f(M)$ is a magic matrix if and only if $f(x)$ is divisible by $x(x^2 - 64)$. This implies that N^s, and hence M^s, is magic for every odd integer s; for example, N^5 is a magic matrix because

$$N^5 = N^2(N^3 - 64N) + 64(N^3 - 64N) + 64^2N = 64^2N.$$

The vector space of semi-magic matrices

The dimensions of the real vector spaces $\mathcal{S}^n$ of $n \times n$ semi-magic matrices, and $\mathcal{S}_0^n$ of semi-magic matrices with magic constant 0, have been known since at least 1938 [**3**] (but published again and again since then) and are, for $n \geq 3$, as follows:

$$\dim\left(\mathcal{S}_0^n\right) = (n-1)^2, \quad \dim\left(\mathcal{S}^n\right) = (n-1)^2 + 1. \tag{2}$$

To see this we observe that the map

$$\Theta : \begin{pmatrix} a & b \\ c & d \end{pmatrix} \mapsto \begin{pmatrix} a & b & -(a+b) \\ c & d & -(c+d) \\ -(a+c) & -(b+d) & a+b+c+d \end{pmatrix}$$

is an isomorphism from the space of 2×2 real matrices onto $\mathcal{S}_0^3$, so that $\dim\left(\mathcal{S}_0^3\right) = (3-1)^2$. An obvious generalization of this shows that $\dim\left(\mathcal{S}_0^n\right) = (n-1)^2$. Next, the map $\mu : \mathcal{S}^n \to \mathbb{R}$ (which takes a semi-magic matrix to its magic constant) is a linear map with kernel $\mathcal{S}_0^n$, and μ is surjective for if $k \in \mathbb{R}$, then $\Theta(k\mathcal{J}) = k$. Since $\dim\left(\mathcal{S}^n\right) = \dim\left(\mathcal{S}_0^n\right) + \dim(\mathbb{R})$, this completes our proof of (2).

Our next result gives a geometric characterization of matrices in $\mathcal{S}_0^n$ (a similar result has been proved for magic matrices, but only when $n = 3$ [14]). We shall regard $\mathbb{R}^n$ as the space of column vectors, and we let $\mathbf{0} = (0, 0, \ldots, 0)^\top$, $\mathbf{e} = (1, 1, \ldots, 1)^\top$, and $\mathbf{e}_1 = (1, 0, \ldots, 0)^\top, \ldots, \mathbf{e}_n = (0, \ldots, 0, 1)^\mathrm{t}$, where $\mathbf{x}^\top$ denotes the transpose of a row vector $\mathbf{x}$. Also, the subspace generated by a collection $\{\mathbf{u}_1, \ldots\}$ of vectors in $\mathbb{R}^n$ is denoted by $\langle \mathbf{u}_1, \ldots \rangle$. Our discussion is based on the two subspaces

$$E = \{r\mathbf{e} : r \in \mathbb{R}\} = \langle \mathbf{e} \rangle,$$

$$\Pi = \{(x_1, \ldots, x_n)^\top : x_1 + \cdots + x_n = 0\} = \{\mathbf{x} : \mathbf{x}\cdot\mathbf{e} = 0\}$$

of $\mathbb{R}^n$ of dimensions 1 and $n - 1$, respectively. Each $n \times n$ matrix M acts on $\mathbb{R}^n$ by matrix multiplication, and the linear transformation $\mathbf{x} \mapsto M\mathbf{x}$ induced by M is denoted by α_M. We shall now express the definition of a semi-magic matrix in terms of the action of α_M on E and on Π.

Theorem 2. *An $n \times n$ matrix M is a semi-magic matrix with magic constant 0 if and only if $E \subset \ker(\alpha_M)$ and $\alpha_M(\Pi) \subset \Pi$.*

Proof. Suppose that M is an $n \times n$ matrix. As the columns of M are $M\mathbf{e}_1, \ldots, M\mathbf{e}_n$, we see that each column of M sums to 0 if and only if $\alpha_M(\mathbf{R}^n) \subset \Pi$. Also, each row of M sums to 0 if and only if $M\mathbf{e} = \mathbf{0}$. Thus M is a semi-magic matrix with magic constant 0 if and only if $\alpha_M(\mathbf{e}) = \mathbf{0}$ and $\alpha_M(\Pi) \subset \Pi$. ∎

Theorem 2 provides an alternative proof that any finite product of semi-magic matrices with magic constant 0 is again a semi-magic matrix with magic constant 0. It also shows that the null part of a semi-magic matrix with magic constant 0 is, in effect, a linear transformation of Π into itself and, as $\dim(\Pi) = n - 1$, this provides an alternative proof that $\dim(\mathcal{S}_0^n) = (n - 1)^2$.

Inverses and adjugates

It is known that if M is a nonsingular 3×3 magic matrix, then M^{-1} is a magic matrix. However, there seems to be almost nothing in the literature about the inverse, or adjugate, of a general $n \times n$ magic matrix, and here we comment on the inverse (when it exists), and the adjugate (which always exists) of a general $n \times n$ magic matrix.

First, suppose that M is a non-singular $n \times n$ magic matrix, and let $p_0 + p_1 x + \cdots + p_{n-1}x^{n-1} + x^n$ be the characteristic polynomial of M. Then, from the Cayley–Hamilton theorem,

$$p_0 M^{-1} = -\big(p_1 \mathcal{I} + p_2 M + \cdots + p_{n-1}M^{n-2} + M^{n-1}\big).$$

Since M is non-singular, $\det(M) \neq 0$, so that $p_0 \neq 0$. This shows that M^{-1} is a polynomial in M and so, by using the ideas above, for any given M we can determine whether or not M^{-1} is a magic matrix.

We now consider the *adjugate* of a square matrix. Given an $n \times n$ matrix M, let Δ_{ij} be the determinant of the $(n - 1) \times (n - 1)$ matrix obtained by deleting the ith row and jth column from M. Then the *adjugate* $\mathrm{adj}(M)$ of M is the transpose of the matrix (M_{ij}) with entries of the form $M_{ij} = (-1)^{i+j}\Delta_{ij}$. The adjugate $\mathrm{adj}(M)$ is often called the *adjoint* (particularly in older texts), but the word "adjugate" is now commonly used to distinguish it from the idea of an adjoint operator. In any event, it is well known that, unlike M^{-1}, the adjugate $\mathrm{adj}(M)$ always exists and satisfies $M\,\mathrm{adj}(M) = \mathrm{adj}(M)\,M = \det(M)$. We refer the reader to [7, 10] for a discussion of

the adjugate matrix. In the context of our discussion here, it is known [7, p. 85] that $\mathrm{adj}(M)$ is a linear combination of $\mathcal{I}, M, \ldots, M^{n-1}$ so again we are, in principle, able to determine whether, given any square matrix M, the matrix $\mathrm{adj}(M)$ is a magic matrix or not.

The case $n = 3$

Finally, we consider the impact of our results in the much-studied case of $n = 3$. Our first lemma seems not to have been noticed before.

Lemma 2. *Let N be a 3×3 magic matrix with magic constant 0. Then N has eigenvalues of the form 0, λ and $-\lambda$, and $N^2 - \lambda^2\mathcal{I}$ is a magic matrix.*

Proof. It is well known that N is a 3×3 magic matrix with magic constant 0 if and only if we can write $N = aA + bB$, where A and B are the magic matrices given by

$$A = \begin{pmatrix} 1 & -1 & 0 \\ -1 & 0 & 1 \\ 0 & 1 & -1 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & -1 & 1 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \end{pmatrix}.$$

A calculation shows that $AB + BA = \mathcal{O}$, and $A^2 = -B^2$. Thus

$$N^2 = 3(a^2 - b^2)\mathcal{I} + 3(b^2 - a^2)\mathcal{J}.$$

However, a straightforward calculation shows that N has eigenvalues 0, λ and $-\lambda$, where $\lambda^2 = 3(a^2 - b^2)$, and this completes the proof. ∎

Lemma 2 shows that although N^2 may or may not be a magic matrix, $N^2 - \lambda^2\mathcal{I}$ is *always* a magic matrix; thus asking whether N^2 (for example) is a magic matrix is perhaps not the right question.

Finally, we consider $\mathrm{adj}(M)$, where M is a 3×3 magic matrix. Let k be the magic constant of M. Then k is an eigenvalue of M, and since $\tau(M)$ is both k and the sum of the eigenvalues, we see that M has eigenvalues k, λ and $-\lambda$, say (see [1, 11] for more information about this). Now $\tau_1 = k$ and $\tau_2 = k^2 + 2\lambda^2$. As

$$\mathrm{adj}(M) = \tfrac{1}{2}\left(\tau_1^2 - \tau_2\right)\mathcal{I} - \tau_1 M + M^2 = M^2 - \lambda^2\mathcal{I} - kM,$$

we find that

$$\begin{aligned} \mathrm{adj}(M) &= M^2 - \lambda^2\mathcal{I} - kM \\ &= (k\mathcal{J} + N)^2 - \lambda^2\mathcal{I} - kM \\ &= k^2\mathcal{J} - kM + (N^2 - \lambda^2\mathcal{I}), \end{aligned}$$

and this explains why $\mathrm{adj}(M)$ is magic when $n = 3$. This argument seems to depend entirely on Lemma 2.

REFERENCES

[1] Amir, A. R., Fredericks, G. A. (1984). Characteristic polynomials of magic squares. *Math. Mag*. 57(3): 220–221.

[2] Beardon, A. F. (2017). Products of 3×3 magic matrices. *Math. Gaz.* 101(550): 96–98.

[3] Chernick, J. (1938). Solution of the general magic matrix. *Amer. Math. Monthly*. 45: 172–175.

[4] Edwards, B., Hartman, J. (2011). Powers of magic matrices. *Math. Gaz*. 95(533): 284–292.

[5] Fox, C. (1956). Magic matrices. *Math. Gaz*. 40: 209–211.

[6] Frénicle de Bessy (1729). Des Quarrez ou Tables Magiques. *Mémoires de l'Académie Royale des Sciences depuis 1666 jusqu'a 1699*. V: 209.
[7] Gantmacher, F. R. (1959). *The Theory of Matrices*, Vol. 1. New York: Chelsea Pub. Co.
[8] Gauthier, N. (1997). Singular matrices applied to 3×3 magic matrices. *Math. Gaz*. 81(491): 225–230.
[9] Hill, R., Elzaidi S. M. (1996). Cubes and inverses of magic matrices. *Math. Gaz*. 80(489): 565–567.
[10] Hoffman K., Kunze, R. (1961). *Linear Algebra*. Englewood Cliffs, NJ: Prentice-Hall.
[11] Khan, N. A. (1957). Characteristic roots of semi-magic matrices. *Amer. Math. Monthly*. 64: 261–263.
[12] Loomis, L. H., Sternberg, S. (1968). *Advanced Calculus*. Reading, MA: Addison-Wesley.
[13] Murase, I. (1957). Semi-magic squares and non-semisimple algebras. *Amer. Math. Monthly*. 64: 168–173.
[14] Shapiro, D. B. (1999). A geometric view of magic matrices. *Math. Gaz*. 83(496): 108–109.
[15] Thompson, A. C. (1994). Odd magic powers. *Amer. Math. Monthly*. 101: 339–342.
[16] Weiner, L. M. (1955). The algebra of semi-magic matrices. *Amer. Math. Monthly*. 62: 237–239.

Summary. It is known that if M is a 3×3 magic matrix then every positive, odd power of M is a magic matrix, while an even power need not be. Given any $n \times n$ magic matrix M, we find all polynomials f such that $f(M)$ is a magic matrix.

ALAN BEARDON (MR Author ID: 33080) studied at Harvard and Imperial College, London, before teaching at the universities of Maryland, Kent at Canterbury, and Cambridge. He retired from Cambridge in 2007, and since then he has taught at the African Institute for Mathematical Sciences (AIMS) at Muizenberg, South Africa, and at several other AIMS centres in Africa.

The Pizza-Cutter's Problem and Hamiltonian Paths

JEAN-LUC BARIL
CÉLINE MOREIRA DOS SANTOS
Université de Bourgogne Franche-Comté
France
barjl@u-bourgogne.fr
celine.moreira@u-bourgogne.fr

The *pizza-cutter's problem* was introduced and solved by Steiner in 1826 (see [14]); it is considered as a doorstep to Euler's well-known formula $v + f - e = 2$ where v is the number of vertices, e the number of edges, and f the number of faces in a connected planar graph. The goal of the pizza-cutter's problem is to maximize the number of pieces that can be made with n straight cuts through a circular pizza, regardless of the size and shape of the pieces. Determining the maximum number of pieces of pizza is the same as determining the maximum number of regions formed by n lines in the plane, which appears in the literature as *Steiner's plane-cutting problem* [1, 2, 16, 17]. If ℓ_n denotes this number then it satisfies the recurrence relation $\ell_n = \ell_{n-1} + n$ for $n \geq 1$ anchored by $\ell_0 = 1$, which induces the closed form $\ell_n = \frac{n(n+1)}{2} + 1$ (see A000124 in [13]). Indeed, from a solution to the problem with $n - 1$ lines that forms ℓ_{n-1} regions, we add an nth line that is not parallel to any of the others, and such that $n - 1$ new intersection points are created. Then, this line crosses n different regions, and each of them is divided into two regions which induces the above recursive formula.

Historically, the problem of line arrangements in the plane is studied by considering oriented matroids, more specifically known as non degenerate dissection types (see [4, 5, 7, 11] for the literature and [6, 8] for some databases). In this paper, we consider this problem from the point of view of graph theory. We refer to a solution of the Steiner's plane-cutting problem as an *S-solution*. For each S-solution, we consider the associated graph $G = (V, E)$ with vertex set V and the edge set E such that

- V is the set of regions; and
- $(p, q) \in E$ if and only if the two regions p and q are *adjacent*, i.e., if they share a common boundary that is not a corner, where corners are the points shared by three or more regions.

Of course there are many ways to cut the plane into a maximal number of regions with n lines, but G always has $|V| = \ell_n$ and $|E| = n^2$. In the case where two solutions produce two isomorphic graphs, we say that these solutions are *isomorphic*; otherwise they are called *non-isomorphic*. See Figure 1 for an illustration of two non-isomorphic S-solutions with their corresponding graphs. Finding the number of classes of non-isomorphic solutions for the plane-cutting problem still remains an open problem for $n \geq 10$. For $1 \leq n \leq 9$, it is known that these numbers are given by the sequence A090338 in [13]: 1, 1, 1, 1, 6, 43, 922, 38609, 3111341; see [8].

Math. Mag. **92** (2019) 359–367. doi:10.1080/0025570X.2019.1664232
MSC: Primary 05C45, Secondary 05C10; 52C30
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/umma.

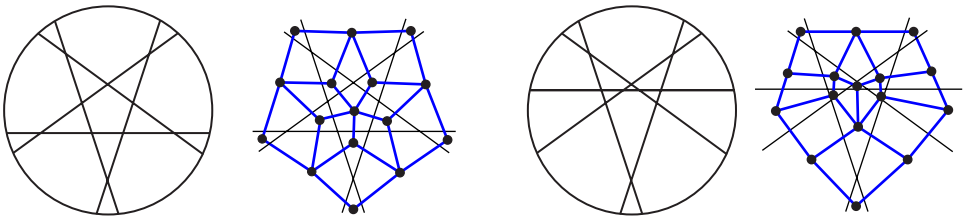


Figure 1 Two non-isomorphic S-solutions for $n = 5$ with their associated graphs drawn in blue.

A graph will be called *traceable* whenever it contains a *Hamiltonian path*, i.e., if there is a path that visits each vertex exactly once. This concept was introduced in 1856 in [15] to study whether a polyhedron contains a path that reaches each vertex once and only once. More generally, the problem of determining whether a graph is traceable is NP-complete and has many applications; see [9]. In particular, this problem appears in network theory where it is crucial to connect points so that the total length of connecting lines is a minimum. On the other hand, determining the traceability can often be a simple way to prove that two graphs are not isomorphic. Then it becomes natural to ask the following question. *For $n \geq 1$, does an S-solution exist such that its corresponding graph is traceable (respectively not traceable)?* A traceable solution to the pizza-cutter's problem means that we can eat up all pieces of the pizza such that any two pieces eaten consecutively are adjacent.

In the next section, we show how from an S-solution we can label each region with a binary string. This induces a graph where the vertex set is the set of labels, and two binary strings are adjacent if their Hamming distance is one. We prove that the traceability of the associated graph is equivalent to that of the graph on labels. Then, we construct an S-solution where the associated graph is not traceable for all $n \geq 5$. In the final section, we adapt this construction in order to obtain an S-solution for all n such that the graph is traceable. To our knowledge, no such precise constructions have previously been published. We conclude by formulating some open problems.

Binary string interpretation

A binary string s of length n is a word $s_1 s_2 \dots s_n$ on the alphabet $\{0, 1\}$. The value s_i, $1 \leq i \leq n$, will be called the ith digit of s. A substring t of s is a word made up of consecutive digits of s. A run of 1's in s is a maximal substring of s of the form 1^k where $k \geq 1$, i.e., a run of 1's is a substring constituted of 1's that cannot be extended to a larger substring of 1's in s. For a binary string set B, we denote by B' (respectively B'') the subset of B of strings with an odd (respectively even) number of 1's.

The Hamming distance between two n-length binary strings s and t is the number of i, $1 \leq i \leq n$, such that s_i is different from t_i. A Gray code for a set of binary strings $B \subseteq \{0, 1\}^n$ is an ordered list $\mathcal{B}$ for B, such that the Hamming distance between any two consecutive strings in $\mathcal{B}$ is exactly one. A Gray code $\mathcal{B}$ for the set B may be viewed as a Hamiltonian path in the restriction of the hypercube Q_n to the set B. Note that no Gray code is possible for B whenever $\big||B'| - |B''|\big| > 1$.

Now, let us consider an S-solution with n lines numbered from 1 to n. We label each region with a binary string of length n where the ith digit is either 0 or 1 depending on whether the region is on one side or the other of the ith line. See Figure 2 for three illustrations of such a labeling. Of course, there are $n!$ possible ways to label n lines from 1 to n, and two half-planes are delimited by each line. Therefore, for an

S-solution there are at most $2^n \cdot n!$ possible sets of labels. In the following, such a set will be called *admissible* for a given S-solution.

Lemma 1. Let us consider an S-solution for $n \geq 1$, $G = (V, E)$ its associated graph and W an admissible set of binary strings for this solution. Let $H = (W, F)$ be the graph where the vertex set is W and two elements are adjacent in H if and only if their Hamming distance is one. Then G and H are isomorphic; and thus, G is traceable if and only if H is traceable.

Proof. It is straightforward to see that the two following assertions are equivalent: (1) two regions r and s are adjacent; and (2) the Hamming distance of the binary strings labeling r and s is one. ■

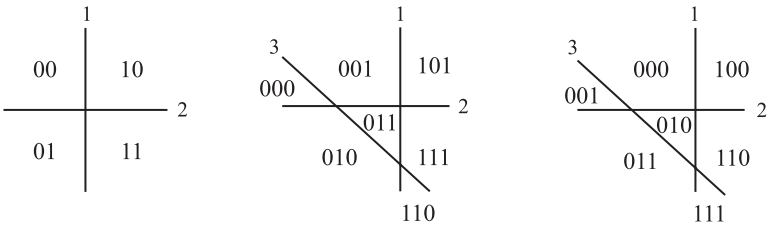


Figure 2 Regions labeled using admissible sets of binary strings. The leftmost and rightmost labeling provide the sets L_2 and L_3, while the central one does not generate L_3.

Remark 1. With the hypotheses of Lemma 1, a necessary condition for the traceability of the graph G is that the cardinalities of W' and W'' differ by at most one.

We end this section by introducing a set that will be crucial in what follows. Let L_n be the set of binary strings of length n containing at most one run of 1's. Any string $s_1 s_2 \dots s_n \in L_n$, $n \geq 1$, can be written either $s = 0 s_2 \dots s_n$ where $s_2 \dots s_n \in L_{n-1}$, or $s = 1^k 0^{n-k}$ with $1 \leq k \leq n$. So, we have $|L_n| = |L_{n-1}| + n$ which induces $|L_n| = \ell_n$. Now, we denote by L'_n (respectively L''_n) the subset of L_n constituted of strings in L_n with an odd (respectively even) number of 1's. For instance, $L_3 = \{000, 001, 010, 100, 110, 011, 111\}$, $L'_3 = \{001, 010, 100, 111\}$ and $L''_3 = \{000, 110, 011\}$.

An S-solution where $G = (V, E)$ is not traceable

In this part, we construct an S-solution such that for each $n \geq 5$, its associated graph $G = (V, E)$ is not traceable. For this, we prove that the set L_n of n-length binary strings with at most one run of 1's is admissible for this solution and that the cardinality of their two subsets L'_n and L''_n differ by at least 2. Using Lemma 1 and Remark 1, we conclude that G is not traceable.

Lemma 2. For $n \geq 1$, there is an S-solution such that the set L_n of binary strings of length n with at most one run of 1's is admissible.

Proof. We proceed by induction on the number n of lines. The case $n = 1$ is trivial since we label the two half-planes by 0 and 1 and $L_1 = \{0, 1\}$.

Assume now that there is an S-solution of $n-1$ lines such that the regions can be labeled with the binary strings of the set L_{n-1}. Since there are exactly $n-1$ binary strings ending in a one in L_{n-1}, the $(n-1)$th line splits the plane into two half-planes such that one of them contains exactly the $n-1$ regions labeled $0^{n-2}1, 0^{n-3}1^2, \ldots, 01^{n-2}, 1^{n-1}$. Then, we necessarily have the leftmost configuration illustrated in Figure 3 where all previous binary strings appear on the same half-plane defined by the line $n-1$ (line 5 in Figure 3). Now, it suffices to place the nth line (line 6 in Figure 3) such that: it crosses the region 0^{n-1} and all other regions labeled 0^k1^{n-1-k} for $0 \leq k \leq n-2$ (the process is illustrated in Figure 3). Note that we can always add this line since it can be obtained from the $(n-1)$th line by a rotation centered on a point placed on the border between the regions 0^{n-1} and $0^{n-2}1$, and with an angle small enough to allow the nth line to intersect the first $(n-2)$ lines (as the $(n-1)$th line). Then, the labels of the newly created regions are obtained by adding 1 to the right of 0^{n-1} and 0^k1^{n-1-k} for $0 \leq k \leq n-2$, and adding 0 to the right of all other labels in L_{n-1}. Finally, the set of the obtained labels is exactly the set L_n of binary strings of length n with at most one run of 1's, and the proof is obtained by induction. ■

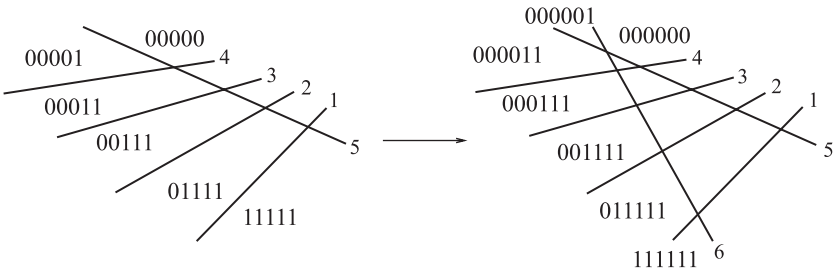


Figure 3 An illustration for the induction in the proof of Lemma 2.

Let $\{\phi_n\}_n \geq 0$ be the parity difference integer sequence corresponding to the binary strings with at most one run of 1's, i.e., $\phi_n = |L'_n| - |L''_n|$ for $n \geq 0$.

Lemma 3. For $n \geq 1$, we have $\phi_n = \lfloor \frac{n-1}{2} \rfloor$.

Proof. For $1 \leq i \leq n$, we denote by L_n^i the subsets of L_n made of strings with exactly i ones. Thus, it follows trivially that $|L_n^i| = n - i + 1$ for $1 \leq i \leq n$, and $|L_n^0| = 1$. Moreover, for i odd, $1 \leq i \leq n-1$, we have $|L_n^i| - |L_n^{i+1}| = 1$. Since $L'_n = \bigcup_{i=1}^{\lfloor \frac{n+1}{2} \rfloor} L_n^{2i-1}$ and $L''_n = \bigcup_{i=0}^{\lfloor \frac{n}{2} \rfloor} L_n^{2i}$, we distinguish two cases. If n is odd, then $\phi_n = |L'_n| - |L''_n| = |L_n^n| - |L_n^0| + \sum_{i=1}^{\lfloor \frac{n-1}{2} \rfloor} (|L_n^{2i-1}| - |L_n^{2i}|) = \lfloor \frac{n-1}{2} \rfloor$. If n is even, then $\phi_n = \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} (|L_n^{2i-1}| - |L_n^{2i}|) - |L_n^0| = \lfloor \frac{n-1}{2} \rfloor$. ■

Theorem 1. For each $n \geq 5$, there exists an S-solution such that its associated graph is not traceable.

Proof. Figure 4 demonstrates a graph for $n = 5$ that is not traceable. For $n \geq 5$, Lemma 3 implies that $\phi_n = \lfloor \frac{n-1}{2} \rfloor \geq 2$. The combination of Remark 1 and Lemma 2 extends the result for $n > 5$. ■

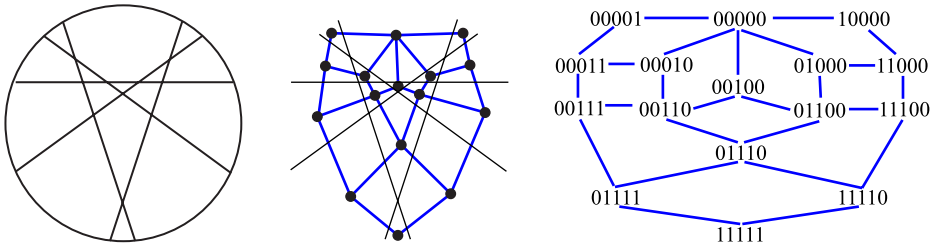


Figure 4 An S-solution where its associated graph is not traceable.

An S-solution where $G = (V, E)$ is traceable

In this part, for each $n \geq 1$, we construct an S-solution such that its associated graph is traceable.

From the set L_n defined previously (as the set of binary strings o flength n containing at most one run of 1's), we define the set K_n by replacing all strings $0^{4i}00100^{n-4(i+1)} \in L_n$ with $0^{4i}01010^{n-4(i+1)}$ for $0 \leq i \leq \lfloor \frac{n}{4} \rfloor - 1$. For instance, we obtain K_5 (respectively K_8) from L_5 (respectively L_8) by replacing 00100 (respectively 00100000 and 00000010) with 01010 (respectively 01010000 and 00000101).

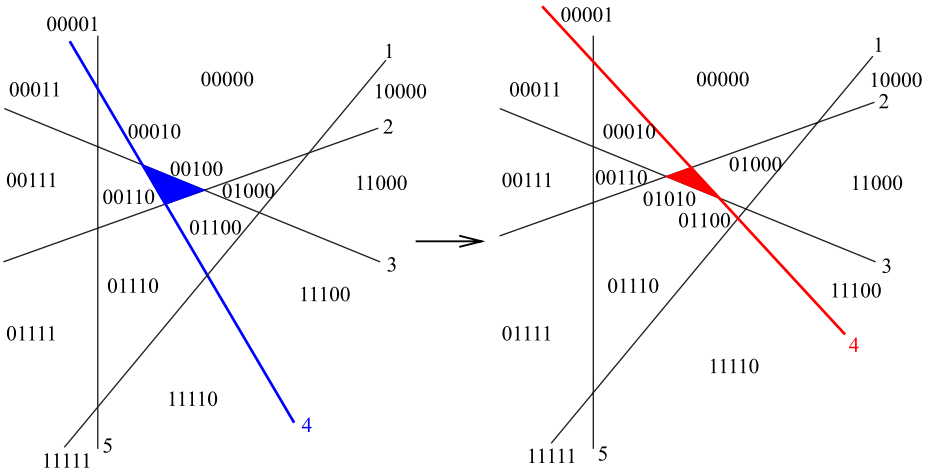


Figure 5 Construction in the proof of Lemma 4.

Lemma 4. For $n \geq 1$, there is an S-solution such that the set K_n is admissible.

Proof. Let us take the S-solution constructed in the proof of Lemma 2. Then we modify the position of each line labeled $4i$, $1 \leq i \leq \lfloor \frac{n}{4} \rfloor$ in the following way. For i from 1 to $\lfloor \frac{n}{4} \rfloor$, the line labeled $4i$ is moved so that in this new position, the half-plane delimited by this line and containing the point of intersection of the lines $4i - 1$ and $4i - 2$ does not contain any other points of intersection between lines from 1 to $4i - 1$. See Figure 5 for an illustration of the process. Then, a simple observation provides that the labels in L_n are preserved up to the labels $0^{4i}00100^{n-4(i+1)}$, $0 \leq i \leq \lfloor \frac{n}{4} \rfloor - 1$, that are replaced with $0^{4i}01010^{n-4(i+1)}$ which transforms the set L_n into the set K_n. Thus K_n is admissible. ■

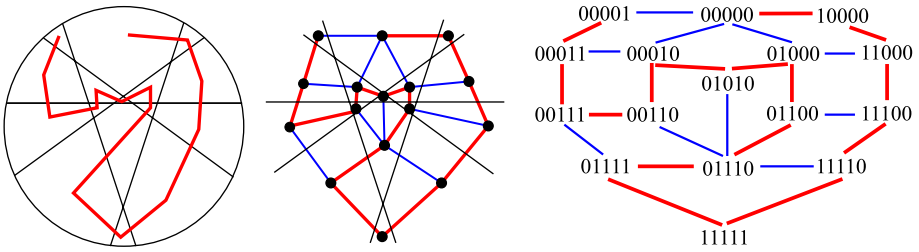


Figure 6 An S-solution where its associated graph is traceable. The red edges constitute a Hamiltonian path.

Theorem 2. For $n \geq 1$, there exists an S-solution such that its associated graph is traceable.

Proof. Due to Lemmas 1 and 4, it suffices to prove that the set K_n can be ordered in a list $\mathcal{K}_n$ such that two consecutive elements differ by one digit, i.e., $\mathcal{K}_n$ is in Gray code order. In order to facilitate the reading of the (somewhat theoretical) proof, we invite the reader to follow it by setting $n = 7$ or $n = 8$ before referring to Table 1. (We use color, different fonts, and boxed and underlined items to make proof easier to follow.)

Let $\mathcal{S}_n$, $n \geq 0$, be the list of the $n+1$ binary strings defined as follows: the ith binary element of the list is $1^{i-1}0^{n-i+1}$, $1 \leq i \leq n+1$. For instance, the list $\mathcal{S}_4$ is 0000, 1000, 1100, 1110, 1111. For $n = 0$, the list $\mathcal{S}_n$ is reduced to the empty string. Obviously, two consecutive elements of $\mathcal{S}_n$ differ by exactly one digit and the first and last elements of $\mathcal{S}_n$ are respectively 0^n and 1^n.

Using the lists $\mathcal{S}_n$, $n \geq 0$, we define an ordered list $\mathcal{L}_n$ of the set L_n by

$$\mathcal{L}_n = 0^n \odot \bigodot_{i=0}^{n-1} 0^i 1 \cdot \mathcal{S}^i_{n-i-1},$$

where $\odot$ is the concatenation operator of lists, and where $\mathcal{S}^i_n$ is the reverse list of $\mathcal{S}_n$ (i.e., the list $\mathcal{S}_n$ considered from the last to the first element) whenever i is odd, and the list $\mathcal{S}_n$ otherwise. See Table 1 for an illustration of the two lists $\mathcal{L}_7$ and $\mathcal{L}_8$.

In the list $\mathcal{L}_n$, it is straightforward to see that two consecutive elements differ by at most one digit except for the transitions between the sublists $0^i1 \cdot \mathcal{S}^i_{n-i-1}$ and $0^{i+1}1 \cdot \mathcal{S}^{i+1}_{n-i-2}$ for i odd and $1 \leq i \leq n-2$. In these cases, the transitions move two digits since (when i is odd) the last element of $0^i1 \cdot \mathcal{S}^i_{n-i-1}$ is $\mathbf{0^i10^{n-i-1}}$ and the first element of $0^{i+1}1 \cdot \mathcal{S}^{i+1}_{n-i-2}$ is $\boxed{0^{i+1}10^{n-i-2}}$. Moreover, the first and last elements of the list $\mathcal{L}_n$ are respectively 0^n and $0^{n-1}1$. Now we modify the list $\mathcal{L}_n$ in order to construct a list $\mathcal{K}_n$ in Gray code order for the set K_n.

For all odd i such that $i = 1 \mod 4$, $1 \leq i \leq n-3$, we replace the string $\boxed{0^i0100^{n-i-3}}$ with $\boxed{0^i1010^{n-i-3}}$ and we change the place of $\underline{0^i0010^{n-i-3}}$ by inserting it just after $\boxed{0^i1010^{n-i-3}}$ and thus just before 0^i0110^{n-i-3}. See Table 1 for an illustration of this process for the lists $\mathcal{K}_7$ and $\mathcal{K}_8$.

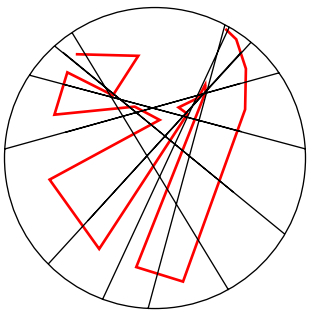
By construction, the four binary strings $\mathbf{0^i1000^{n-i-3}}$, $\boxed{0^i1010^{n-i-3}}$, $\underline{0^i0010^{n-i-3}}$ and 0^i0110^{n-i-3} are consecutive in the list $\mathcal{K}_n$ and the three transitions differ by only one digit.

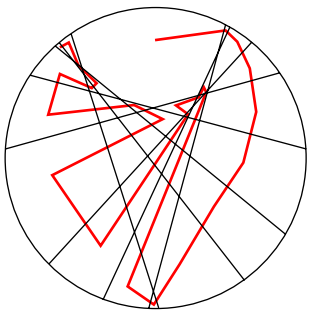
On the other hand, since we change the position of the binary strings of the form $\underline{0^i0010^{n-i-3}}$, for $i = 1 \mod 4$, $1 \leq i \leq n-3$, we create a new transition between its

TABLE 1: The Lists $\mathcal{L}_n$ and $\mathcal{K}_n$ for $n = 7, 8$ and the Hamiltonian path on the S-solutions associated to the sets K_7 and K_8.

	$\mathcal{L}_7$	$\mathcal{K}_7$
	0000000	1000000
$i=0$	1000000	1100000
	1100000	1110000
	1110000	1111000
	1111000	1111000
	1111000	1111000
	1111000	1111110
	1111110	1111111
	1111111	0111111
$i=1$	0111111	0111110
	0111110	0111100
	0111100	0111000
	0111000	0110000
	0110000	**0100000**
	0100000	*0010000*
$i=2$	*0010000*	0001000
	0011000	0011000
	0011100	0011100
	0011110	0011110
	0011111	0011111
$i=3$	0001111	0001111
	0001110	0001110
	0001100	0001100
	0001000	0000100
$i=4$	0000100	0000110
	0000110	0000111
	0000111	0000111
$i=5$	0000011	0000010
	0000010	0000000
$i=6$	0000001	0000001

	$\mathcal{L}_8$	$\mathcal{K}_8$
	00000000	00000000
$i=0$	10000000	10000000
	11000000	11000000
	11100000	11100000
	11110000	11110000
	11111000	11111000
	11111100	11111100
	11111110	11111110
	11111111	11111111
$i=1$	01111111	01111111
	01111110	01111110
	01111100	01111100
	01111000	01111000
	01110000	01110000
	01100000	01100000
	01000000	**01000000**
$i=2$	*00100000*	*01010000*
	00110000	00010000
	00111000	00110000
	00111100	00111000
	00111110	00111100
$i=3$	00011111	00111111
	00011110	00011111
	00011100	0011110
	00011000	00011100
	00010000	00011000
$i=4$	00001000	00001000
	00001100	00001100
	00001110	00001110
	00001111	00001111
$i=5$	00000111	00000111
	00000110	00000110
	00000100	**00000100**
$i=6$	*00000010*	*00000101*
	00000011	00000001
$i=7$	00000001	00000011





predecessor 0^i00110^{n-i-4} and its successor 0^i00010^{n-i-4} (if it exists) that moves only one digit. Notice that if $i = n-3$ then the string 0^i0010^{n-i-3} has no successor in the list $\mathcal{L}_n$ and after moving its position, the last element of $\mathcal{K}_n$ becomes $0^{n-2}11$.

If n is even, then the last transition of two digits in $\mathcal{L}_n$ occurs between $0^{n-3}100$ and $0^{n-2}10$ which means that all transitions of two digits have been treated above, and the list $\mathcal{K}_n$ is in Gray code order. So, the first and last elements are respectively 0^n and $0^{n-2}11$ for $n = 0 \mod 4$, and 0^n and $0^{n-1}1$ for $n = 2 \mod 4$.

If n is odd, then the last transition of two digits in $\mathcal{L}_n$ occurs between $0^{n-2}10$ and $0^{n-1}1$. We distinguish two subcases. If $n \neq 3 \mod 4$, then the string $0^{n-2}10$ is moved by the above process and the obtained list is in Gray code order. So, the first and last elements are respectively 0^n and $0^{n-1}1$ (the Gray code is cyclic). However, if $n = 3 \mod 4$, then we insert the first element 0^n between $0^{n-2}10$ and $0^{n-1}1$ and we obtain a Gray code. Here, the first and last elements are respectively 10^{n-1} and $0^{n-1}1$.

Finally, for all $n \geq 1$ the constructed list $\mathcal{K}_n$ is in Gray code order. ■

Remark 2. For $n = 1, 2 \mod 4$, the Hamming distance between the first and last elements of the list $\mathcal{K}_n$ is one. Thus the associated graph becomes Hamiltonian (see Figure 6).

Going further

In this paper, we use a constructive method in order to prove that the pizza-cutter's problem admits an S-solution where its associated graph is traceable. Is it possible to provide a similar result using probabilistic method as studied in [**3**, **12**]? For a given n, can we find the number of isomorphism classes of S-solutions for $n \geq 10$? How many classes induce a traceable graph? For a given S-solution, can we characterize its corresponding admissible sets? More generally, can we make the same study for the space-cutting problem where the dimension of the space is greater than two?

Acknowledgment We would like to thank the anonymous referees for their very careful reading of this paper and their helpful comments and suggestions.

REFERENCES

[1] Alexanderson, G.L., Wetzel, J.E. (1978). Simple partitions of space. *Math. Mag.* 51(3): 220–225.
[2] Banks, R.B. (1999). *Slicing Pizzas, Racing Turtles, and Further Adventures in Applied Mathematics*. Princeton, NJ: Princeton University Press.
[3] Ben-Shimon, S., Krivelevich, M., Sudakov, B. (2011). On the resilience of Hamiltonicity and optimal packing of Hamilton cycles in random graphs. *SIAM J. Discrete Math.* 25(3): 1176–1193.
[4] Björner, A., Las Vergnas, M., Sturmfels, B., White, N., Ziegler, G. (1993). *Oriented Matroids*, 2nd ed. Cambridge: Cambridge University Press.
[5] Bokowski, J., Guedes de Oliveira, A. (2000). On the generation of oriented matroids. *Discret. Comput. Geom*. 24(2/3): 197–208.
[6] Christ, T. (2019). Database of combinatorially different simple line arrangements. https://geometry.inf.ethz.ch/christt/linearr/.
[7] Finschi, L. (2001). A graph theoretical approach for reconstruction and generation of oriented matroids. Thesis. https://www.math.ethz.ch/ifor/publications.html, Zurich.
[8] Finschi, L., Fukuda, K. (2019). Oriented matroids database. http://www.om.math.ethz.ch.
[9] Garey, M.R., Johnson, D.S. (1979). *Computers and Intractability*, Vol. 174. New York: Freeman.
[10] Graham, R.L., Knuth, D.E., Patashnik, O. (1990). *Concrete Mathematics*. New York: Addison Wesley.
[11] Ringel, G. (1956). Teilungen der ebene durch geraden oder topologische geraden. *Math. Z.* 64: 79–102.
[12] Shang, Y. (2015). On the Hamiltonicity of random bipartite graphs. *Indian J. Pure Appl. Math.* 46(2): 163–173.
[13] Sloane, N.J.A. (2019). The on-line encyclopedia of integer sequences. https://oeis.org.

[14] Steiner, J. (1826). Einige Gesetze über die Theilung der Ebene und des Raumes. *J. für die Reine Angewandte Mathematik*. 1: 349–364.
[15] Kirkman, T.P. (1856) On the enumeration of x-edra having triedral summits and an $(x-1)$-gonal base. *Philos. Trans. R. Soc. London*. 146: 399–411.
[16] JWetzel, J.E. (1978). On the division of the plane by lines. *Amer. Math. Monthly*. 85: 647–656.
[17] Zimmerman, S. (2001). Slicing space. *College Math. J*. 32: 126–128.

Summary. The pizza-cutter's problem is to determine the maximum number of pieces that can be made with n straight cuts through a circular pizza, regardless of the size and shape of the pieces. For a solution to this problem, we consider the graph $G = (V, E)$ where the vertex set V is the set of pieces and $(p, q) \in E$ if and only if the two pieces p and q are adjacent. We prove that there exists a solution where the graph G contains (respectively does not contain) a Hamiltonian path. Finally we present some open questions.

JEAN-LUC BARIL (MR Author ID: 709339) received his Ph.D. degree in pure mathematics from Bordeaux University, France, in 1996. He is currently full professor at the University of Burgundy, France, and member of LIB EA 7534. His main research involves combinatorial problems.

CÉLINE MOREIRA DOS SANTOS (MR Author ID: 695593) received her Ph.D. degree in mathematics from Caen University, France, in 2002. She is currently associate professor at the University of Burgundy, France, and member of LIB EA 7534. Her main research involves combinatorial problems.

A Game of Nontransitive Dice

ARTEM HULKO
Tusculum University
Greeneville, TN 37745
artemhulko@gmail.com

MARK WHITMEYER
University of Texas at Austin
Austin, TX 78712
mark.whitmeyer@utexas.edu

If Hercules and Lychas play at dice
Which is the better man, the greater throw
May turn by fortune from the weaker hand.

William Shakespeare,
The Merchant of Venice

Sets of nontransitive dice are fascinating mathematical objects that have attracted the curiosity of many for nearly fifty years. They first came into the limelight in one of Martin Gardner's column [**8**] and are one of a larger class of nontransitivity "paradoxes" (see [**2**, **16**]), which also include the well-known Condorcet voting paradox, as described in [**7**].

The past few years have seen a surge in interest in the topic, including [**1**, **4**, **15**]. The underlying ideas have not been limited merely to dice; one notable work is [**10**], which instead reinterprets the scenario through throws of unfair *coins*. Nontransitive dice have even been the subject of investigation by the well-known polymath project (see [**12**]), and indeed we borrow some terminology from that paper.

When one speaks of dice, one usually speaks also of a game, and when a game is the topic at hand, a natural question is how it should be *played*. Along those lines, several papers have investigated how nontransitive dice should be thrown in a strategic interaction of two or more players. Rump [**14**] was the first one to do so: he explores a two-player game in which each player may choose one of the four six-sided Efron dice and finds the set of equilibria, before extending the analysis to cover the situation in which each player chooses two such dice.

Here, we investigate a broader problem: we consider a two-player, simultaneous-move game in which each player selects a general n-sided die and throws it. The player with the highest face showing wins a reward, normalized to 1, and each player receives $1/2$ in the event of a tie. For our game, we use the Nash equilibrium solution concept. Note that this game is a constant-sum game; therefore it is equivalent to a zero-sum game for which a Nash equilibrium is a saddle point. We show that, for $n > 3$, there is a single, unique, pure-strategy Nash equilibrium in which both players play the *standard* n-sided die where each possible value, $1, 2, \ldots, n$, occurs with probability $1/n$. There may be additional mixed strategy equilibria; however, in this analysis we focus exclusively on pure strategy equilibria and henceforth, by Nash equilibrium or equilibrium, we mean only those in pure strategies.

Moreover, our proof of uniqueness is constructive and contains an algorithm that, for any nonstandard die, generates a die that beats it. We introduce the idea of a *one-step die*: a die that is the result of a simple modification of the standard die in which one dot is moved from one face to another. Intuitively, such a die is merely "one-step"

Math. Mag. **92** (2019) 368–373. doi:10.1080/0025570X.2018.1662263
MSC: Primary 91A05, Secondary 91A60; 60C05

away from the standard die. We show that for any nonstandard die, there is always at least one one-step die that beats it.

Two additional papers bear mention. The closest paper to this one, [6], considers the same problem, where for some fixed integer n, two players each choose a die and throw against each other. The authors show that the standard die ties every other die, and that every nonstandard die loses to some other die. Another paper, [5], also explores dice games though in a slightly more general setting, and the existence and uniqueness of an equilibrium in which both players each throw the standard die follows from their Propositions 6 and 8.

Our paper differs from [5,6] in the following key ways. We provide different proofs of the existence and uniqueness of the Nash equilibrium in the game, and we are able to do so exclusively using elementary mathematics. Additionally, our proof is constructive and we formulate a simple algorithm that allows us, for any nonstandard die, to generate a die that beats it. Moreover, our last result—that for any nonstandard die, there is a one-step die that beats it—is also novel.

Finally, dice games can be placed in a more general context, as a member of the family of Colonel Blotto games. First developed by E. Borel in 1921 (see [3]), a burgeoning literature has resulted, due to the game's general applications in economics, operations research, political science, and other areas. Some recent papers include [9,13]. In another paper [11], we explore an n-player continuous version of this game played on the interval [0, 1], which is then extended in [17] to a dynamic setting. For two players, the unique equilibrium is the continuous analog of the unique equilibrium here, the uniform distribution.

The basic game

Fix a positive integer n and define an n-tuple $D = (d_1, d_2, \ldots, d_n)$ with $1 \le d_1 \le d_2 \le \cdots \le d_n \le n$ such that $\sum d_i = n(n+1)/2$. We then define a general n-sided die (henceforth just a "die") as a random variable, D, that takes values in the finite set $\{1, 2, \ldots, n\}$, provided the distribution satisfies the following conditions:

1. For each $i = 1, 2, \ldots, n$, the probability that a certain value occurs, $\mathbb{P}[D = i] = d_i$, is a multiple of $1/n$.
2. The expectation of the random variable is $\mathbb{E}[D] = \sum_{i=1}^{n} i \cdot d_i = \dfrac{n+1}{2}$.

Denote the set of all n-sided dice by $\mathcal{D}_n$. As mentioned above, the standard n-sided die S is the die where each possible value occurs with probability $1/n$.

Example 1. Five 4-sided dice appear in the set $\mathcal{D}_4$:

$$\mathcal{D}_4 = \{[1, 1, 4, 4], [2, 2, 2, 4], [1, 3, 3, 3], [2, 2, 3, 3], S = [1, 2, 3, 4]\}.$$

The game Two players, Amy and Bob, play the following one shot game for fixed n. Amy and Bob each independently select and roll any n-sided die, $A, B \in \mathcal{D}_n$. The payoff to a player is the expected gain, where the reward is 1 for throwing the higher number, 1/2 for a tie, and 0 for a lower number. Amy's expected payoff is the probability that the realization of her throw is higher than the realization of Bob's throw (with ties settled by a fair coin flip), and Bob's payoff is Amy's mirror.

A *strategy* for Amy (and analogously for Bob) is simply a choice of die $A \in \mathcal{D}_n$. For any pair of strategies, (A, B), Amy's expected payoff and Bob's expected payoff,

are, respectively,

$$U_{\text{Amy}}(A, B) = \Pr(A > B) + (1/2)\Pr(A = B) \text{ and}$$
$$U_{\text{Bob}}(A, B) = \Pr(B > A) + (1/2)\Pr(A = B),$$

where $U_{\text{Amy}}(A, B) + U_{\text{Bob}}(A, B) = 1$ and $U_{\text{Amy}}(B, A) = U_{\text{Bob}}(A, B)$.

Example 2. Suppose Amy and Bob choose dice $A = [1, 1, 4, 4]$ and $B = [2, 2, 2, 4]$ from Example 1, respectively. Then $U_{\text{Amy}}(A, B) = 7/16$ and $U_{\text{Bob}}(A, B) = 9/16$.

The main result of this paper is the following theorem.

Theorem. *For any n, the unique Nash equilibrium of the two-player game is where both players play the standard die S.*

We prove this theorem through two propositions by first showing that the strategy pair (S, S) is a Nash equilibrium, and then, in the case where $n \geq 4$, by proving that (S, S) is the unique equilibrium. Before we proceed, we remind the reader of the definition of a Nash equilibrium.

Definition. A pair of strategies (A, B) is a (pure strategy) Nash equilibrium if neither player, holding his or her opponent's strategy fixed, has a profitable deviation to any other strategy. Formally, (A, B) is a pure strategy Nash equilibrium if $U_{\text{Amy}}(A, B) \geq U_{\text{Amy}}(A', B)$ for all $A' \in \mathcal{D}_n$, and $U_{\text{Bob}}(A, B) \geq U_{\text{Bob}}(A, B')$ for all $B' \in \mathcal{D}_n$.

Now, our first proposition.

Proposition 1. *The strategy pair (S, S) is a Nash equilibrium.*

Proof. We begin by showing that for either $i \in \{\text{Amy}, \text{Bob}\}$ and for all $D \in \mathcal{D}_n$, $U_i(S, D) = U_i(D, S) = 1/2$.

Note that it suffices to prove that $U_{\text{Amy}}(D, S) = 1/2$, since that clearly implies $U_{\text{Bob}}(S, D) = 1/2$. Suppose that Bob chooses the standard die S and that Amy chooses an arbitrary die D. If the realization of D is d_i (i.e., when D is rolled and "lands" showing face d_i), then with probability $(d_i - 1)/n$, D beats the standard die, and with probability $1/n$, D ties the standard die. Hence,

$$U_{\text{Amy}}(D, S) = \sum_{i=1}^{n} \left(\frac{1}{n}\right)\left(\frac{d_i - 1}{n} + \frac{1}{n} \cdot \frac{1}{2}\right) = \left(\frac{1}{n}\right)\left(\mathbb{E}[D] - \frac{1}{2}\right) = \frac{1}{2}.$$

Since (S, S) gives Amy a payoff of $1/2$, she cannot profit by deviating from S. ■

It remains to show uniqueness, which we accomplish in the following proposition. For any two dice $A, B \in \mathcal{D}_n$, we say that A *beats* B if the number of pairs (a_i, b_j) with $a_i > b_j$ exceeds the number of pairs with $a_i < b_j$.

Proposition 2. *The Nash equilibrium (S, S) is unique for $n \geq 4$.*

Proof. Clearly, for any strategy pair, player i has a profitable deviation if and only if there is a die that beats her opponents die. Our proof is constructive and we show that for any die $B \neq S$, we can construct a die, G, that beats it.

To that end, let $B = [b_1, b_2, \ldots, b_n]$ and recall that $S = [1, 2, \ldots, n]$. For $k = 1, 2, \ldots, n$, define γ_k by $\gamma_k = |\{b_i | b_i = k\}|$. By construction of the dice,

$$\sum_{k=1}^{n} \gamma_k = n \quad \text{and} \quad \sum_{k=1}^{n} k\gamma_k = \frac{n(n+1)}{2}.$$

Next, for $k = 1, 2, \ldots, n-1$, define ξ_k as $\xi_k = \gamma_k + \gamma_{k+1}$. To construct a die G that beats B, we need simply find a pair (ξ_i, ξ_j) with $\xi_i > \xi_j$ (and so clearly $j \neq i$) and $i \neq j+1$. To see this, we take a look at what happens when we match the standard die with a die D represented by $(\gamma_1, \gamma_2, \ldots, \gamma_n)$. Face i of the standard die defeats $\gamma_1 + \cdots + \gamma_{i-1}$ faces of D and loses to $\gamma_{i+1} + \cdots + \gamma_n$ faces of D.

Observe what happens when we move a dot from face $j+1$ to face i on the standard die: the number of wins changes by $\gamma_i - \gamma_j$ and the number of losses changes by $\gamma_{j+1} - \gamma_{i+1}$. This new die is a one-step die, and for this die to dominate D, we need $\gamma_i - \gamma_j > \gamma_{j+1} - \gamma_{i+1}$; equivalently, $\gamma_i + \gamma_{i+1} > \gamma_j + \gamma_{j+1}$. Before we return to the proof, let's consider an example, and then we prove a necessary lemma.

Example 3. Suppose player A chooses die X from our previous examples, $X = [1, 1, 4, 4]$. We have $\gamma_1 = 2$, $\gamma_2 = \gamma_3 = 0$, and $\gamma_4 = 2$, and so $\xi_1 = 2$, $\xi_2 = 0$, and $\xi_3 = 2$. Evidently, $\xi_1 > \xi_2$ and $1 = i \neq j+1 = 2+1 = 3$. Hence, adjusting the standard die S, we can add 1 to s_1 and subtract 1 from s_3 to yield the die $Y = [2, 2, 2, 4]$, which is a one-step die that beats X. Indeed, should player B choose Y she would achieve a payoff of $9/16 > 1/2$.

If $\xi_a \neq \xi_b$ for some a, b, then there must be some i, j with $\xi_i > \xi_j$. Thus, we establish the following lemma:

Lemma. *If $n \geq 4$, then for any nonstandard n-sided die there exists a pair $a, b \in \{1, 2, \ldots, n\}$, for which $\xi_a \neq \xi_b$.*

Proof. The equality $\xi_a = \xi_b$ holds for all $a, b \in \{1, 2, \ldots, n\}$ if and only if

$$\gamma_1 + \gamma_2 = \gamma_2 + \gamma_3 = \gamma_3 + \gamma_4 = \cdots = \gamma_{n-1} + \gamma_n$$

which holds if and only if

$$\begin{aligned} &\gamma_1 = \gamma_3 = \cdots = \gamma_k \quad \text{for all odd integers} \quad k \in \{1, 2, \ldots, n\} \text{ and} \\ &\gamma_2 = \gamma_4 = \cdots = \gamma_j \quad \text{for all even integers} \quad j \in \{1, 2, \ldots, n\}. \end{aligned} \tag{1}$$

We also have the following two relationships:

$$\sum_{k \text{ odd}}^{n} \gamma_k + \sum_{j \text{ even}}^{n} \gamma_j = n \tag{2}$$

and

$$\sum_{k \text{ odd}}^{n} k\gamma_k + \sum_{j \text{ even}}^{n} j\gamma_j = \frac{n(n+1)}{2}. \tag{3}$$

Note that any nonstandard die must have some $\gamma_i = 0$. Then either $\gamma_i = 0$ for all odd i or $\gamma_i = 0$ for all even i.

Suppose n is odd and that $\gamma_1 = 0$. From equation (2) we have $(n-1)\gamma_2 = 2n$, which does not have a solution in integers n, γ_2 for $n > 3$. Next, suppose n is odd and that $\gamma_2 = 0$. From equation (2) we have $(n+1)\gamma_1 = 2n$, which does not have a solution in integers n, γ_1 for $n > 1$. Thus, we conclude that n cannot be odd.

Suppose n is even and that $\gamma_1 = 0$. From equations (1) and (2) we must have $\gamma_2 = 2$, and from equations (1) and (3) we get that $n + 2 = 2(n+1)$, which is obviously a contradiction. Finally, suppose n is even and that $\gamma_2 = 0$. From equations (1) and (2) we must have $\gamma_1 = 2$, and from equations (1) and (3) we must have that $2(n+1) = n+1$, which is also a contradiction. Thus, we have proved the lemma. ∎

To wrap up the proof of Proposition 2 we need to verify that we cannot have the situation in which the only pair ξ_i, ξ_j that satisfies $\xi_i > \xi_j$ occurs when $i = j + 1$. To that end, suppose $\xi_i > \xi_j$ for $i = j + 1$. First, let $j \neq 1$. Then, if $\xi_{j-1} \leq \xi_j$, relabel $j - 1$ as j', which yields $\xi_i > \xi_{j'}$ for $i \neq j' + 1$. On the other hand, if $\xi_{j-1} > \xi_j$, relabel $j - 1$ as i', implying $\xi_{i'} > \xi_j$ for $i' \neq j + 1$. Next, let $j = 1$. If $\xi_{i+1} \geq \xi_i$, relabel $i + 1$ as i', which yields $\xi_{i'} > \xi_j$ for $i' \neq j + 1$. If, instead, $\xi_{i+1} < \xi_i$, relabel $i + 1$ as j', and thus we have $\xi_i > \xi_{j'}$ for $i \neq j' + 1$. ■

We may also write the following corollary, which we have proved along the way.

Corollary. *Let $n \geq 4$. Then, for any die $B \neq S$, there exists a one-step die G that beats B.*

Note that given some die $B \neq S$, the algorithm developed in our proof yields *every* winning one-step die (i.e., a one-step die that beats B). Moreover, it is easy to see how by "flipping" the algorithm we could also obtain the set of losing one-step dice. Finally, the algorithm also enables us to find the "best" (and "worst") one-step dice to play versus B: the die (or dice) that have the greatest (or least) chance of beating B.

This last result is somewhat surprising. One natural notion of "closeness" of dice is that two dice D and D' are close if the dice are one step away from each other. That is, if we could move a dot from a particular face on dice D to another face and thereby obtain die D' (and vice-versa). Hence, the corollary can be interpreted as saying that for any nonstandard die, there is a die close to the standard die that beats it.

REFERENCES

[1] Angel, L., Davis, M. (2016). A direct construction of non-transitive dice sets. ArXiv e-prints 1610.08595. arxiv.org/pdf/1610.08595

[2] Blyth, C. (1972). Some probability paradoxes in choice from among random alternatives. *J. Am. Stat. Assoc*. 67: 366–373.

[3] Borel, E. (1921). La théorie du jeu et les équations integrales à noyau symétrique. *C. R. Acad. Sci*. 173: 1304–1308; English translation by Savage, L. (1953). The theory of play and integral equations with skew symmetric kernels. *Econometrica*. 21: 97–100.

[4] Conrey, B., Gabbard, J., Grant, K., Liu, A., Morrison, K. (2016). Intransitive dice. *Math. Mag*. 89(2): 133–143.

[5] De Schuymer, B., De Meyer, H., De Baets, B. (2006). Optimal strategies for equal-sum dice games. *Discrete Appl. Math*. 154: 2565–2576.

[6] Finkelstein, M., Thorp, E. O. (2007). Nontransitive dice with equal means. In: Ethier, S. N., Eadington, W. R., eds. *Optimal Play: Mathematical Studies of Games and Gambling*. Reno: Institute for the Study of Gambling and Commercial Gaming. www.math.uci.edu/ mfinkels/dice9.pdf

[7] Fishburn, P. C., Brams, S. J. (1983). Paradoxes of preferential voting. *Math. Mag*. 56(4): 207–214.

[8] Gardner, M. (1970). The paradox of the nontransitive dice. *Sci. Amer*. 223: 110–111.

[9] Hart, S. (2008). Discrete Colonel Blotto and General Lotto games. *Int. J. Game Theory*. 36(3–4): 441–460.

[10] Hetyei, G. (2016). Efron's coins and the linial arrangement. *Discrete Math*. 339(12): 2998–3004.

[11] Hulko, A., Whitmeyer, M. (2017). A game of random variables. *Mimeo*. arxiv.org/abs/1712.08716

[12] Polymath, D. H. J. (2017). The probability that a random triple of dice is transitive. *Mimeo*. gowers.files.wordpress.com/2017/07/polymath131.pdf

[13] Roberson, B. (2006). The Colonel Blotto game. *Econ. Theory*. 29(1): 1–24.

[14] Rump, C. (2001). Strategies for rolling the Efron dice. *Math. Mag*. 74(3): 212–216.

[15] Savage, R. P., Jr. (1994). The paradox of nontransitive dice. *Amer. Math. Monthly*. 101(5): 429–436.

[16] Tenney, R. L., Foster, C. C. (1976). Non-transitive dominance. *Math. Mag*. 49(3): 115–120.

[17] Whitmeyer, M. (2018). A game of martingales. *Mimeo*. arxiv.org/abs/1811.11664

Summary. We consider a two-player, simultaneous-move game where each player selects any permissible n-sided die for a fixed integer n. For any $n > 3$, there is a unique Nash equilibrium in pure strategies in which each player throws the standard n-sided die. Our proof of uniqueness is constructive, and we introduce an algorithm with which, for any nonstandard die, we can generate another die that beats it. For any nonstandard die there exists a one-step die—a die that is obtained by transferring one dot from one side to another on the standard die—that beats it.

ARTEM HULKO (MR Author ID: 1223837) is an assistant professor of Mathematics at Tusculum University. He received his B.A. from USC Upstate and his Ph.D. from UNC Charlotte. His primary areas of research are spectral analysis, scattering theory, and game theory.

MARK WHITMEYER (MR Author ID: 1236225) received his B.A. from Lehigh University and is a Ph.D. Candidate in Economics at UT Austin. He is primarily interested in game theory and information economics.

Prime Time

Try all the tactics but you'll always find the gap of one,
Whenever you increase the gap, Our appearance goes to none.
We are very basic but you know little about us
Do you really think you can explore the empire that belongs to us?
People out there know our values and worth,
Like Ramanujan, Mersenne, Fermat and Aryabhata.
Weakness in calculation had supported Fermat falsification,
The invention of contraption produces Prime size competition.
We are unbreakable because of our indivisible quality.
It's promising you, That's why your security is our guarantee.
How you count our existence between two numbers each?
We can typify any number if some of us planned to stitch.
Our counting has been building the theory of numbers,
Thus Kronecker told number theorists are like lotus eaters.
Try all the tactics but you'll always find the gap of one,
Whenever you increase the gap, Our appearance goes to none.

—Submitted by Shashi Kant Pandey
University of Delhi, India

Math. Mag. **92** (2019) 373. doi:10.1080/0025570X.2019.1445931

Cubic Polynomials, Linear Shifts, and Ramanujan Simple Cubics

GREG DRESDEN
PRAKRITI PANTHI
ANUKRITI SHRESTHA
JIAHAO ZHANG
Washington and Lee University
Lexington, VA 24450
dresdeng@wlu.edu
panthip20@mail.wlu.edu
shresthaa19@mail.wlu.edu
zhangj20@mail.wlu.edu

It's hard to pick out a favorite from Ramanujan's nearly uncountable collection of delightful identities, but these two have to be near the top of anyone's list:

$$\sqrt[3]{1/9} - \sqrt[3]{2/9} + \sqrt[3]{4/9} = \sqrt[3]{\sqrt[3]{2} - 1} \tag{1}$$

$$\sqrt[3]{\cos\frac{2\pi}{9}} + \sqrt[3]{\cos\frac{4\pi}{9}} - \sqrt[3]{\cos\frac{\pi}{9}} = \sqrt[3]{\frac{3}{2}\left(\sqrt[3]{9} - 2\right)}. \tag{2}$$

Both of these equations appear in Ramanujan's notebooks [2], and they have been studied in a number of papers. Landau [8, 9] treated the first equation as an example of how a "nested radical" like $\sqrt[3]{\sqrt[3]{2} - 1}$ can be "de-nested" into a sum of simple cube roots. Berndt and Bhargava [3] gave a proof of the second equation using only elementary methods. Shevelev [15] provided an elementary proof of both formulas (and quite a few other involving similar sums of cube roots). It turns out that both (1) and (2) are related to the roots of a special class of degree-3 polynomials, and in keeping with past work [1, 16, 17], we will define a *Ramanujan simple cubic* (RSC) to be a polynomial with (possibly complex) coefficients of the form

$$p_B(x) = x^3 - \left(\tfrac{3+B}{2}\right)x^2 - \left(\tfrac{3-B}{2}\right)x + 1.$$

We will use a technique from the mid-1800s to prove that almost every cubic is just a linear shift away from a Ramanujan simple cubic, and this will allow us to recapture the two formulas above, and also to come up with new identities, such as the deceptively simple formula

$$2\sqrt{6}\cos\frac{11\pi}{36} + 6\cos\frac{10\pi}{36} = \left(3\sqrt{2} + \sqrt{6}\right)\cos\frac{\pi}{36} \tag{3}$$

and the rather surprising fact that

$$\frac{1}{2}\left(-5 + \sqrt{13} + 2\sqrt{26 - 6\sqrt{13}}\cos\frac{\pi}{26}\right) \tag{4}$$

is a solution to $x^3 + x^2 - 4x + 1 = 0$.

In the next section, we discuss the properties of these RSC polynomials, and in the following section we prove our main result. We finish up with many nice examples.

Math. Mag. **92** (2019) 374–381. doi:10.1080/0025570X.2019.1655310
MSC: Primary 11C08, Secondary 11B83

Ramanujan simple cubics

Surprisingly, these RSC's have been studied in one form or another for over a hundred years. In 1911, Dickson [4] discussed integral solutions to $p_B(x) = 0$ modulo a prime. More recently, a number of authors [1,16,17] have studied a slightly more general class of polynomials they call *Ramanujan cubics*, which are simply our RSC polynomials $p_B(x)$ but with x replaced by x/s. Similarly, if we replace the x in $p_B(x)$ with $-x$, we get the *Shanks polynomials*, so called because they generate what Shanks called the "simplest cubic fields" [14]. Foster's paper [6] reviews earlier work on the Shanks polynomials and the simplest cubic fields; he also proved that every degree-three cyclic extension of the rationals is generated by a Shanks polynomial (which implies the same for our RSC); this was done earlier by Kersten and Michaliček [7]. Also, Lehmer [11] and Lazarus [10] have shown that the minimal polynomials for so-called *cubic Gaussian periods*, when composed with some $x - a$ for a an integer, will equal one of the Shanks polynomials (and thus are related to our RSC's).

The following theorem illustrates some of the remarkable properties of Ramanujan simple cubics (RSC).

Theorem 1. *For* $p_B(x) = x^3 - \left(\frac{3+B}{2}\right)x^2 - \left(\frac{3-B}{2}\right)x + 1$ *the Ramanujan simple cubic defined earlier,*

1. *The roots* r_1, r_2, r_3 *of* $p_B(x)$ *are always permuted by the order-three map* $n(x) = \frac{1}{1-x}$.
2. *The roots* r_1, r_2, r_3 *satisfy*

$$\sqrt[3]{r_1} + \sqrt[3]{r_2} + \sqrt[3]{r_3} = \sqrt{\sqrt[3]{\left(\tfrac{3+B}{2}\right)} - 6 + 3\sqrt[3]{\tfrac{27+B^2}{4}}} \tag{5}$$

   so long as, for complex arguments, we choose the appropriate values for the cube roots.
3. *If we define the elements of the set* $\{s_1, s_2, \ldots, s_6\}$ *as*

$$s_k = \tfrac{1}{3}\left(\left(\tfrac{3+B}{2}\right) + \sqrt{27+B^2}\,\cos\left(\tfrac{k\pi}{3} + \tfrac{1}{3}\arctan\tfrac{3\sqrt{3}}{B}\right)\right) \tag{6}$$

   then for $B \geq 0$ *the roots of* $p_B(x)$ *are* $\{s_2, s_4, s_6\}$ *and for* $B \leq 0$ *the roots of* $p_B(x)$ *are* $\{s_1, s_3, s_5\}$.

A more complicated version of equation (6) for the roots of a general cubic has been known since the times of Viète and Descartes; our presentation serves to illustrate how much simpler this formulation can be when dealing with the Ramanujan simple cubics. Also, while (6) is not actually defined at $B = 0$, we can interpret it at that value by simply taking the limit of (6) as B approaches 0. Surprisingly, whether we have B approach 0 from above or from below, the three values of $\{s_2, s_4, s_6\}$ and the three values of $\{s_1, s_3, s_5\}$ coincide at $\{-1, 1/2, 2\}$, which are indeed the three roots of $p_0(x) = x^3 - 3/2\,x^2 - 3/2\,x + 1$.

The above theorem takes in a Ramanujan simple cubic and gives results about its roots (e.g., the roots are permuted by $n(x) = 1/(1-x)$). We note that the converse is true: if r_1, r_2, r_3 are permuted by $1/(1-x)$, then by multiplying and simplifying we find that

$$(x - r_1)(x - r_2)(x - r_3) = (x - r_1)\left(x - \frac{1}{1 - r_1}\right)\left(x - \frac{r_1 - 1}{r_1}\right)$$

is the Ramanujan simple cubic $p_B(x)$ with $B = (2 - 3r_1 - 3r_1^2 + 2r_1^3)/(r_1^2 - r_1)$.

Proof of Theorem 1. For part 1, a quick calculation gives us that $-p_B\left(\frac{1}{1-x}\right)\cdot(1-x)^3 = p_B(x)$. Since 1 is never a root of $p_B(x)$, this shows that if r_1 is a root of p_B then so also is $\frac{1}{1-r_1}$. This is enough to show that $n(x)$ permutes the roots so long as $\frac{1}{1-r_1}$ is different from r_1. As for the case when $\frac{1}{1-r_1}$ equals r_1, this implies that r_1 is a primitive sixth root of unity, which means $B = \pm i\sqrt{27}$ and all three roots of $p_B(x)$ are identical and hence, technically, are still "permuted" by $\frac{1}{1-x}$. A similar (and slightly more general) proof can also be found in [**1**].

For part 2, there is an elementary proof in [**3**, p. 652] of a nearly identical statement for the roots of the Shanks polynomial $x^3 - ax^2 - (a+3)x - 1$; the roots of this Shanks polynomial are the negatives of the roots of the Ramanujan polynomial $p_B(x) = x^3 + ax^2 - (a+3)x + 1$ with $B = -2a - 3$ and so the identity follows. This proof also appears in [**2**, p. 22].

For part 3, we refer the reader to the similar proof for Shanks polynomials in [**1**, Theorem 7]; another version of this formula (without proof, and for just one root) can be found in [**12**]. ■

Example 1. We can now easily show that equations (1) and (2) arise from equation (5) of Theorem 1. For equation (1), we take $p_B(x)$ with $B = 0$ which has roots $1/2, -1, 2$, and so equation (5) gives us

$$\sqrt[3]{1/2} + \sqrt[3]{-1} + \sqrt[3]{2} = \sqrt[3]{\left(\tfrac{3}{2}\right) - 6 + 3\sqrt[3]{\tfrac{27}{4}}}$$

and after multiplying through by $\sqrt[3]{2/9}$ and doing some simplifying on the right, we get the desired equation.

As for (2), we note that the minimal polynomial for $2\cos 2\pi/9$ is $x^3 - 3x + 1$, a Ramanujan simple cubic with $B = -3$. It's easy to show that the other two roots are $2\cos 4\pi/9$ and $-2\cos \pi/9$, and so equation (5) gives us

$$\sqrt[3]{2\cos 2\pi/9} + \sqrt[3]{2\cos 4\pi/9} + \sqrt[3]{-2\cos \pi/9} = \sqrt[3]{\left(\tfrac{3-3}{2}\right) - 6 + 3\sqrt[3]{\tfrac{27+9}{4}}}$$

and after simplifying the right and dividing by $\sqrt[3]{2}$ we obtain the desired formula.

In the previous example, we began with a particular Ramanujan simple cubic and then derived statements about its roots. We can reverse the process, as seen next.

Example 2. Suppose we wish to create a Ramanujan simple cubic with roots that are related to $\sqrt{3}$. A bit of experimentation suggests that we should choose $r_1 = \sqrt{3} - 1$ to be one of its roots. We know that the other two roots must satisfy $r_2 = n(r_1)$ and $r_3 = n(r_2)$, where $n(x) = \frac{1}{1-x}$. This leads to $r_2 = 2 + \sqrt{3}$ and $r_3 = (1 - \sqrt{3})/2$, and the polynomial $(x - r_1)(x - r_2)(x - r_3)$ is easily calculated to be a Ramanujan simple cubic with $B = 3\sqrt{3}$. This leads to a particularly nice formulation of equation (5); after some simplification (and after multiplying through by $\sqrt[3]{2}$ on both sides) we obtain the following unexpected equation:

$$\sqrt[3]{2\sqrt{3}-2} - \sqrt[3]{\sqrt{3}-1} + \sqrt[3]{2\sqrt{3}+4} = \sqrt{3}\cdot\sqrt[3]{1+\sqrt{3}\left(\sqrt[3]{4}-1\right)}.$$

In the previous example, we note that other choices for r_1 such as $\sqrt{3}+1$ or $2\sqrt{3}$ would still lead to valid formulas, but they are not nearly as aesthetically appealing as the one seen above.

Main result

For $f(x) = x^3 + Px^2 + Qx + R$ a polynomial with (possibly) complex coefficients, we note that its discriminant is

$$\Delta = P^2Q^2 - 4Q^3 - 4P^3R + 18PQR - 27R^2,$$

and we recall that a polynomial has no repeated roots if and only if its discriminant Δ is not zero. With this in mind, we define the following two values (taken from their original definitions in [13, p. 468]):

$$a = \frac{\sqrt{\Delta} - (9R - PQ)}{2\sqrt{\Delta}} \text{ and } c = \frac{6Q - 2P^2}{2\sqrt{\Delta}}.$$

We now state our main result.

Theorem 2. *Let $f(x) = x^3 + Px^2 + Qx + R$ have non-repeated roots t_1, t_2, t_3, and let a and c be as defined above.*

1. *If $c = 0$, then there exists h and k such that $f(x) = (x - h)^3 + k$. In other words, $f(x)$ is a translation of x^3 (by h units horizontally and k units vertically).*
2. *If $c \neq 0$, then $f\left(\frac{a-x}{c}\right) \cdot (-c)^3$ equals the Ramanujan simple cubic $p_B(x) = x^3 - \left(\frac{3+B}{2}\right)x^2 - \left(\frac{3-B}{2}\right)x + 1$, with $B = 6a + 2cP - 3$. In particular, the set of roots of $p_B(x)$ are $\{a - c \cdot t_1,\ a - c \cdot t_2,\ a - c \cdot t_3\}$.*

Proof. First, suppose $c = 0$. Then $Q = P^2/3$ and so $f(x)$ can be written as $(x - h)^3 + k$ with $h = -P/3$ and $k = R - P^3/3$.

Next, suppose $c \neq 0$. It is possible to use brute force to show that $f\left(\frac{a-x}{c}\right) \cdot (-c)^3$ equals $p_B(x)$, but that does not provide much insight into the problem. Instead, we offer the following more detailed explanation. The key can be found in Serret's classic algebra textbook [13, p. 468] from the mid nineteenth century. In pursuit of an entirely unrelated problem, Serret defined the a and c seen above, along with the following:

$$b = \frac{2Q^2 - 6PR}{2\sqrt{\Delta}} \text{ and } d = 1 - a.$$

Serret showed that $m(x) = \frac{ax+b}{cx+d}$ is of order three under composition, permutes the roots t_1, t_2, t_3 of the cubic $f(x) = x^3 + Px^2 + Qx + R$, and has the property that $ad - bc = 1$. Now, we would like to transform the cubic $f(x)$ into a new cubic whose roots are permuted by $n(x) = \frac{1}{1-x}$, and one way to do that is to first find a linear map $q(x)$ such that $(q^{-1} \circ m \circ q)(x) = n(x)$, and then to consider the composition $(f \circ q)(x)$. This composition would have as roots the numbers $q^{-1}(t_1), q^{-1}(t_2), q^{-1}(t_3)$, and furthermore these roots would be permuted by $(q^{-1} \circ m \circ q)(x) = \frac{1}{1-x}$. We can then show this composition must be a Ramanujan simple cubic.

With this in mind, it remains to find our $q(x)$ such that $(q^{-1} \circ m \circ q)(x) = n(x)$. This is a fairly easy task if one uses the language of Möbius transforms (see, e.g., [5]). Since $n(x)$ takes ∞ to 0 to 1 back to ∞, and $m(x)$ takes ∞ to a/c to $-d/c$ back to ∞, we can choose $q(x)$ to take ∞ to ∞, and 0 to a/c, and 1 to $-d/c$. This gives us $q(x) = \frac{a-x}{c}$ and $q^{-1}(x) = a - cx$. We can verify that indeed $(q^{-1} \circ m \circ q)(x) = n(x)$, and that $f(q(x)) = f\left(\frac{a-x}{c}\right)$ has roots $a - c\,t_1$, $a - c\,t_2$, and $a - c\,t_3$ as desired. Since these are permuted by $\frac{1}{1-x}$, it's easy to show that the monic polynomial $f(q(x)) \cdot (-c)^3$ has the appropriate coefficients to match the desired form of a Ramanujan simple cubic. ∎

We can now combine Theorem 2 with Theorem 1 to give us the following results.

Corollary 1. *Let $f(x) = x^3 + Px^2 + Qx + R$ have non-repeated roots t_1, t_2, t_3, and let a, B, and c be as defined in Theorem 2, with $c \neq 0$. Then,*

1. *The order-three map $n(x) = \frac{1}{1-x}$ permutes the set $\{a - ct_1,\ a - ct_2,\ a - ct_3\}$.*
2. *We have the Ramanujan-style equation*

$$\sqrt[3]{a - ct_1} + \sqrt[3]{a - ct_2} + \sqrt[3]{a - ct_3} = \sqrt[3]{\left(\tfrac{3+B}{2}\right) - 6 + 3\sqrt[3]{\tfrac{27+B^2}{4}}}, \qquad (7)$$

   so long as, for complex arguments, we choose the appropriate values for the cube roots.
3. *If we define the elements of the set $\{u_1, u_2, \ldots, u_6\}$ as*

$$u_k = \frac{a}{c} - \frac{1}{3c}\left(\left(\tfrac{3+B}{2}\right) + \sqrt{27 + B^2}\,\cos\left(\tfrac{k\pi}{3} + \tfrac{1}{3}\arctan\tfrac{3\sqrt{3}}{B}\right)\right) \qquad (8)$$

   then for $B \geq 0$ the roots of $f(x)$ are $\{u_2, u_4, u_6\}$ and for $B \leq 0$ the roots of $f(x)$ are $\{u_1, u_3, u_5\}$.

We note that a similar version of formula (7) was presented (without proof) by forum user Tito Piezas III on math.stackexchange.com.

Examples

With such appealing formulas at hand, we cannot help but apply them to various cubics with rational (and irrational) coefficients.

Example 3. Here's a rather lovely formula which we believe has not been seen before:

$$\sqrt[3]{3 - \sqrt{21} + 8\cos\frac{2\pi}{21}} + \sqrt[3]{3 - \sqrt{21} + 8\cos\frac{8\pi}{21}} + \sqrt[3]{3 - \sqrt{21} + 8\cos\frac{10\pi}{21}}$$
$$= \sqrt[3]{-1 - \sqrt{21} + 6\sqrt[3]{28 - 4\sqrt{21}}}.$$

To obtain this, we begin with $x^6 - x^5 - 6x^4 + 6x^3 + 8x^2 - 8x + 1$, the minimal polynomial for $t_1 = 2\cos 2\pi/21$. This is not of degree three, but a common factoring trick is to adjoin the square root of the polynomial's discriminant to the rationals and then to seek a factorization over this larger set of coefficients. Sure enough, our degree-six polynomial factors over $\mathbb{Q}(\sqrt{21})$ as two cubics, and we choose the one which still has $2\cos(2\pi/21)$ as a root. This cubic is $x^3 + \frac{1}{2}\left(-1 - \sqrt{21}\right)x^2 + \frac{1}{2}\left(\sqrt{21} - 1\right)x + \frac{1}{2}\left(\sqrt{21} - 5\right)$, and its other two roots are $t_2 = 2\cos(8\pi/21)$ and $t_3 = 2\cos(10\pi/21)$, and after doing the computations in Theorem 2 we obtain $a = \frac{1}{2}\left(3 - \sqrt{21}\right)$, $c = -2$, and $B = 8 - \sqrt{21}$. We then plug these values into formula (7), multiply through by $\sqrt[3]{2}$, and apply a few simplifications to obtain the above expression.

Example 4. We can do similar calculations for $2\cos(\pi/18)$. This has a minimal polynomial of degree 6, but it factors in $\mathbb{Q}(\sqrt{3})[x]$ and we choose the degree-three factor $g(x) = x^3 - 3x - \sqrt{3}$. One root of $g(x)$ is indeed $2\cos(\pi/18)$, and the other two roots

are $2\cos(11\pi/18)$ and $2\cos(13\pi/18)$. Calculating a and c as defined in Theorem 2, we get $a = 2$ and $c = -\sqrt{3}$. Thus, by Theorem 2, we have $g(\frac{a-x}{c}) \cdot (-c)^3 = x^3 - 6x^2 + 3x + 1$ which is a particularly nice Ramanujan simple cubic with $B = 9$. (We will return to this cubic in Example 6.) By Corollary 1, we get a nice identity:

$$\sqrt[3]{2 + 2\sqrt{3}\cos\frac{\pi}{18}} + \sqrt[3]{2 + 2\sqrt{3}\cos\frac{11\pi}{18}} + \sqrt[3]{2 + 2\sqrt{3}\cos\frac{13\pi}{18}} = \sqrt[3]{9}. \tag{9}$$

Furthermore, by Theorem 1, we know the roots of $x^3 - 6x^2 + 3x + 1$ are permuted by $1/(1-x)$. Therefore, by choosing our roots carefully, we get

$$2 + 2\sqrt{3}\cos\frac{13\pi}{18} = \frac{1}{1 - \left(2 + 2\sqrt{3}\cos\frac{\pi}{18}\right)}$$

and this simplifies to

$$2\cos\frac{\pi}{18} + \cos\frac{13\pi}{18} + \sqrt{3}\cos\frac{14\pi}{18} = 0$$

which reduces to the simple but not trivial identity

$$\cos\frac{5\pi}{18} = 2\cos\frac{\pi}{18} - \sqrt{3}\cos\frac{4\pi}{18}. \tag{10}$$

Example 5. In an effort to find more equations like (10), we look at the minimal polynomials for $2\cos(\pi/36)$ and $2\cos(\pi/42)$. Both have minimal polynomials of degree 12, and both can be factored down into degree three polynomials by adjoining appropriate square roots of discriminants to the rationals. By following the same steps as in the previous example we can arrive at the following two identities:

$$2\sqrt{6}\cos\frac{11\pi}{36} + 6\cos\frac{10\pi}{36} - \left(3\sqrt{2} + \sqrt{6}\right)\cos\frac{\pi}{36} = 0 \tag{11}$$

$$(\sqrt{3} - \sqrt{7})\cos\frac{\pi}{42} - 2\sqrt{7}\cos\frac{25\pi}{42} - 8\cos\frac{\pi}{42}\cos\frac{25\pi}{42} = 3. \tag{12}$$

It's probably just a coincidence, but $\left(3\sqrt{2} + \sqrt{6}\right)\cos(\pi/36)$ from formula (11) is almost identical (to six decimal places) to 20/3. Also, note that (11) is equation (3) from the beginning of the article.

Example 6. Returning our attention to Example 4, we note that Theorem 2 gave us the particularly nice cubic $x^3 - 6x^2 + 3x + 1$ and gave us that one of its roots is $2 + 2\sqrt{3}\cos(\pi/18)$. Likewise, if we begin with $2\cos(\pi/26)$, we can factor its minimal (degree-12) polynomial down to a degree-3 polynomial with irrational coefficients, apply Theorem 2, and end up with another particularly nice polynomial, this time $x^3 + x^2 - 4x + 1$, one of whose roots is given in equation (4) at the beginning of this paper.

It turns out that, as seen in [**11**], these two polynomials are also just an integer shift from the minimal polynomials for certain *cubic Gaussian periods*. The exact nature of these objects is beyond the scope of this article; for our purposes, we can consider them to be sums of roots of unity with their inverses. Suffice it to say that this

recognition leads us to discover that $x^3 - 6x^2 + 3x + 1$ is the minimal polynomial for the following three numbers:

$$\left\{2 + 2\cos\frac{\pi}{9} + 2\cos\frac{2\pi}{9},\ \ 2 + 2\cos\frac{4\pi}{9} + 2\cos\frac{7\pi}{9},\ \ 2 + 2\cos\frac{5\pi}{9} + 2\cos\frac{8\pi}{9}\right\}.$$

Comparing these with $2 + 2\sqrt{3}\cos(\pi/18)$ leads us to the identity

$$2 + 2\sqrt{3}\cos\frac{\pi}{18} \ = \ 2 + 2\cos\frac{\pi}{9} + 2\cos\frac{2\pi}{9}.$$

Unfortunately, this simplifies to a triviality. However, along these lines, we also discover that $x^3 + x^2 - 4x + 1$ is the minimal polynomial for the following three numbers:

$$\left\{2\cos\frac{2\pi}{13} + 2\cos\frac{10\pi}{13},\ \ 2\cos\frac{4\pi}{13} + 2\cos\frac{6\pi}{13},\ \ 2\cos\frac{8\pi}{13} + 2\cos\frac{12\pi}{13}\right\}.$$

After comparing to the solution in equation (4), we obtain this (non-trivial) identity,

$$-5 \ + \ \sqrt{13} \ + \ 2\sqrt{26 - 6\sqrt{13}}\cos\frac{\pi}{26} \ = \ 4\cos\frac{4\pi}{13} \ + \ 4\cos\frac{6\pi}{13},$$

and this really is a lovely formula.

Example 7. We finish with an example that does not involve cosines. Consider the polynomial $f(x) = (x-1)(x-\sqrt{2})(x+\sqrt{3})$. This is not Ramanujan, but when we apply the methods of Theorem 2 we obtain a Ramanujan polynomial $p_B(x)$ with $B = -6 - \sqrt{2} - 5\sqrt{3} + \sqrt{6}$, and one of its roots is $-1 - \sqrt{2} - \sqrt{3} - \sqrt{6}$. After trying various values of k with formula (6), we find that

$$-1 - \sqrt{2} - \sqrt{3} - \sqrt{6} = \frac{1}{3}\left[\left(\frac{3+B}{2}\right) + \sqrt{27 + B^2}\cos\left(-\pi + \frac{1}{3}\arctan\frac{3\sqrt{3}}{B}\right)\right]$$

and after applying our value of B and simplifying, we obtain the following formula:

$$3 + 5\sqrt{2} + \sqrt{3} + 7\sqrt{6} = 2\sqrt{2\left(73 - 9\sqrt{2} + 28\sqrt{3} - \sqrt{6}\right)} \cdot \cos\left(\frac{1}{3}\tan^{-1}\left(\frac{3\sqrt{3}}{-6 - \sqrt{2} - 5\sqrt{3} + \sqrt{6}}\right)\right)$$

and this is surprising if for no other reason than the relatively small size of the coefficients.

REFERENCES

[1] Barbero, S., Cerruti, U., Murru, N., Abrate, M. (2013). Identities involving zeros of Ramanujan and Shanks cubic polynomials. *J. Integer Seq.* 16(8): Article 13.8.1, 13 pp.

[2] Berndt, B. C. (1994). *Ramanujan's Notebooks. Part IV*. New York: Springer-Verlag.

[3] Berndt, B. C., Bhargava, S. (1993). Ramanujan—for lowbrows. *Amer. Math. Monthly*. 100(7): 644–656.

[4] Dickson, L. E. (1911). Note on cubic equations and congruences. *Ann. Math. (2)*. 12(3): 149–152.

[5] Dresden, G. P. (2004). There are only nine finite groups of linear fractional transforms with integer coefficients. *Math. Mag.* 77(3): 211–218.

[6] Foster, K. (2013). HT90 and "simplest" number fields. *Ill. J. Math.* 55(4):1621–1655.

[7] Kersten, I., Michaliček, J. (1987). A characterization of Galois field extensions of degree 3. *Commun. Algebra*. 15(5): 927–933.
[8] Landau, S. (1992). Simplification of nested radicals. *SIAM J. Comput.* 21(1): 85–110.
[9] Landau, S. (1994). How to tangle with a nested radical. *Math. Intell*. 16(2): 49–55.
[10] Lazarus, A. J. (1992). Gaussian periods and units in certain cyclic fields. *Proc. Amer. Math. Soc.* 115(4): 961–968.
[11] Lehmer, E. (1988). Connection between Gaussian periods and cyclic units. *Math. Comput.* 50(182): 535–541.
[12] Louboutin, S. (2002). The exponent three class group problem for some real cyclic cubic number fields. *Proc. Amer. Math. Soc.* 130(2): 353–361.
[13] Serret, J-A. (1992). *Cours d'algèbre supérieure. Tome II*. Les Grands Classiques Gauthier-Villars. [Gauthier-Villars Great Classics]. Éditions Jacques Gabay, Sceaux. Reprint of the fourth (1879) edition.
[14] Shanks, D. (1974). The simplest cubic fields. *Math. Comput.* 28: 1137–1152.
[15] Shevelev, V. S. (1999). Three formulas of Ramanujan [Kvant **1988**, no. 6, 52–55]. In: Tabachnikov, S., ed. *Kvant Selecta: Algebra and Analysis, I*, Vol. 14 of *Math. World*. Providence, RI: American Mathematical Society, pp. 139–144.
[16] Shevelev, V. (2009). On Ramanujan cubic polynomials. *South East Asian J. Math. Math. Sci.* 8(1): 113–122.
[17] Wituła, R. (2010). Full description of Ramanujan cubic polynomials. *J. Integer Seq.* 13(5): Article 10.5.7, 8 pp.

Summary. We show that every monic polynomial of degree three with complex coefficients and no repeated roots is either a (vertical and horizontal) translation of $y = x^3$ or can be composed with a linear function to obtain a Ramanujan cubic. As a result, we gain some new insights into the roots of cubic polynomials.

GREG DRESDEN (MR Author ID: 623871) received his Ph.D. from the University of Texas in 1997 and now teaches at Washington and Lee University in the mountains of Virginia. When not thinking about math, he enjoys playing piano and spending time with family.

PRAKRITI PANTHI (MR Author ID: 1326984) is a junior at Washington and Lee working toward a B.S. degree in Mathematics and Economics. She enjoys playing badminton, hiking, reading fiction, and rereading the Harry Potter series.

ANUKRITI SHRESTHA (MR Author ID: 1326985) is a senior at Washington and Lee majoring in Integrated Engineering (Chemistry) and Mathematics. She plans to pursue a Ph.D. in Chemical Engineering after graduation. She likes spending her free time reading mystery novels and hanging out with dogs.

JIAHAO ZHANG (MR Author ID: 1326986) is a Math major and East Asian Study minor at Washington and Lee University. He likes playing basketball and reading ancient Chinese poems in his spare time.

ACROSS

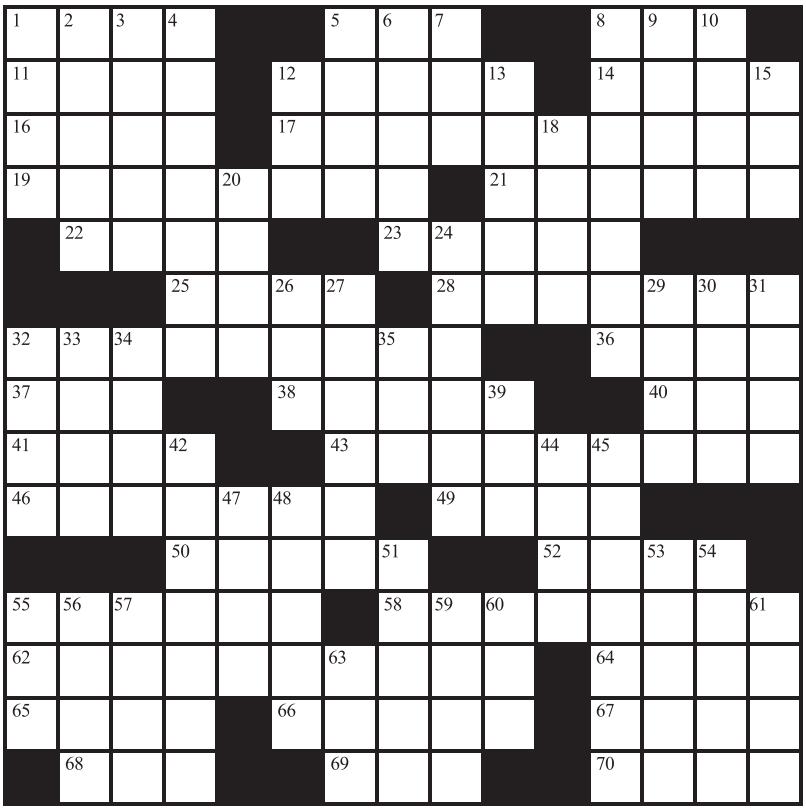
1. Infuriating gaming delays
5. ___ order: public commentary ban
8. Closest pal, in text-speak
11. "Thanks ___ !"
12. Actress McDonald with six Tony Awards
14. Old West lawman Wyatt
16. Chimney channel
17. * Rajiv, of Second Spectrum, who will give the Porter Public Lecture about mathematics and sports
19. * Ardila, of SFSU, who will give the MAA Project Next Lecture on creating a belonging environment in the mathematics classroom
21. Warmed the bench, say
22. California wine valley
23. Math term for a collection of sets
25. When doubled, a 1997 Jim Carrey film
28. * Scott, of Chandler-Gilbert CC, who will give an MAA Invited Address about riddles and zombies
32. * Jordan, of U. Wisconsin-Madison, who will give a Current Events Bulletin Lecture about gerrymandering
36. "___ kleine Nachtmusik"
37. Twelve, to Julius Caesar
38. He resigned the Vice Presidency in 1973
40. Hilarious math joke: "___ $\epsilon < 0$."
41. Surrounded by
43. * Skip, of Institute for Defense Analyses, who will give an AMS-MAA Invited Address about lotteries
46. * aBa, of U. Wisconsin-Eau Claire, who will give an MAA Invited Address about undergraduate research projects in recreational mathematics
49. Mathematics is the final letter in this acronym
50. Astrological ram
52. Caribbean and others
55. Dove's call, doubled
58. * Linda Brown, of Penn State, who will give an ASL Invited Address about Borel sets
62. * Last name of the professor from Pomona College who will give the MAA Lecture for Students about musical scales
64. Place for furniture and meatballs
65. Another acronym for Daesh or ISIS
66. Excited, with "up"
67. Penny
68. Mad Hatter's drink
69. "Acid" or expy in Chicago
70. Poker state

DOWN

1. TV network featuring sitcom reruns
2. John ___ Paulos, author of "Innumeracy"
3. Cheese from the Netherlands
4. "Here is the church, here is the___, . . . "
5. Taqueria dip, for short
6. Like some faculty committees
7. Test you might take during your last year of univ. or coll.
8. "Who knows?"
9. Card game popular before the invention of poker
10. German honorific for an adult woman
12. * First name of 62-Across
13. Syrian leader
15. Acronym for a mathematical result that says $\pi(N) \sim \frac{N}{\log(N)}$
18. "There once ___ . . ."
20. Drops from the sky
24. Beers made from bottom-fermenting yeast
26. Lawyer's org.
27. Musical genre for Bob Marley and Toots & the Maytals
29. Window ledge
30. Like a line, not a plane, briefly
31. ___ pot: mechanism for cleansing sinuses
32. It could be oral or written
33. Arm or leg
34. XXXIV + XIX
35. Genetic material
39. Cleverness in humor
42. Bram Stoker creation
44. Top notch
45. "A Horse with No Name" band
47. Golf club likely used from the fairway
48. Sales booth
51. Influences
53. "American Idol" runner-up Clay
54. Bloodhound's clue
55. ___ de coeur: passionate exclamation
56. Kiln in which you might make 24-Down
57. "Garfield" dog
59. Looked at
60. Bummed
61. Actress Winslet
63. Childhood comedy partner of Kenan Thompson

Joint Mathematics Meetings 2020

BRENDAN SULLIVAN
Emmanuel College
Boston, MA
sullivanb@emmanuel.edu



Clues start at left, on page 385. The Solution is on page 351.

Extra copies of the puzzle can be found at the MAGAZINE's website, www.maa.org/mathmag/supplements.

Crossword Puzzle Creators

If you are interested in submitting a mathematically themed crossword puzzle for possible inclusion in MATHEMATICS MAGAZINE, please contact the editor at mathmag@maa.org.

A Visual Proof of Gregory's Theorem

TOM EDGAR
Pacific Lutheran University
Tacoma, WA 98447
edgartj@plu.edu

DAVID RICHESON
Dickinson College
Carlisle, PA 17013
richesod@dickinson.edu

The Scottish mathematician James Gregory published the short book *Vera Circuli et Hyperbolae Quadratura* (*The True Squaring of the Circle and the Hyperbola*) in 1667. He continued the tradition started by Archimedes of using regular polygons to approximate a circle. Whereas Archimedes used the perimeters of the polygons to bound the circumference of the circle (to obtain his famous bounds, $223/71 < \pi < 22/7$), Gregory used the areas of the polygons to obtain ever-tightening bounds on the area of the circle. He proved the following recursive formulas for obtaining these bounds.

Theorem. *Let I_k and C_k denote the areas of regular k-gons inscribed in and circumscribed around a given circle. Then I_{2n} is the geometric mean of I_n and C_n, and C_{2n} is the harmonic mean of I_{2n} and C_n; that is,*

$$I_{2n} = \sqrt{I_n C_n} \qquad \text{and} \qquad C_{2n} = \frac{2C_n I_{2n}}{C_n + I_{2n}} = \frac{2}{\frac{1}{I_{2n}} + \frac{1}{C_n}}.$$

Alsina and Nelsen [1] provide a visual proof of the nearly identical theorem about the perimeters of inscribed and circumscribed regular polygons. Here, we give a short, visual proof of the following lemma from which the theorem follows.

Lemma. *For all n,*

$$\frac{I_{2n}}{I_n} = \frac{C_n}{I_{2n}} = \frac{C_n - C_{2n}}{C_{2n} - I_{2n}}.$$

Proof. Suppose we have a circle of radius r with inscribed and circumscribed regular n- and $2n$-gons. Let a be the length of the apothem of the inscribed n-gon, b be the radius of the circumscribed n-gon, and c and d be half the side lengths of the inscribed n-gon and circumscribed $2n$-gon, respectively. Then, as we see in Figure 1,

$$\frac{I_{2n}}{I_n} = \frac{2n \cdot \frac{1}{2}rc}{2n \cdot \frac{1}{2}ac} = \frac{r}{a}, \quad \frac{C_n}{I_{2n}} = \frac{2n \cdot \frac{1}{2}bc}{2n \cdot \frac{1}{2}rc} = \frac{b}{r}, \quad \text{and}$$

$$\frac{C_n - C_{2n}}{C_{2n} - I_{2n}} = \frac{2n \cdot \frac{1}{2}(b-r)d}{2n \cdot \frac{1}{2}(r-a)d} = \frac{b-r}{r-a}.$$

Math. Mag. **92** (2019) 384–386. doi:10.1080/0025570X.2019.1685321
MSC: Primary 01A45, Secondary 01A55
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/umma.

And, by similar triangles (see Figure 2),

$$\frac{r}{a} = \frac{b}{r} = \frac{b-r}{r-a}.$$

■

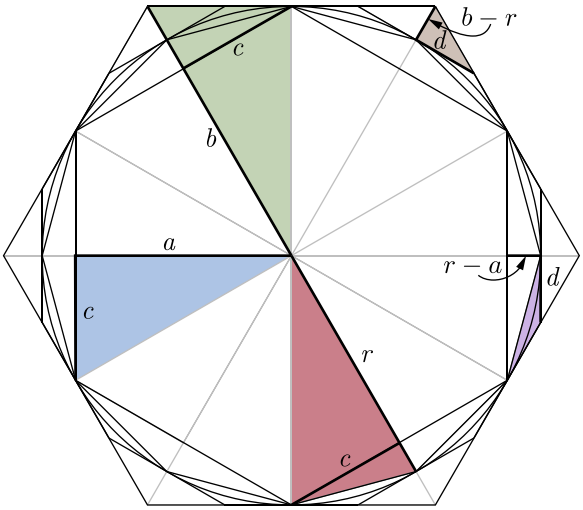


Figure 1 A circle with inscribed and circumscribed regular n- and $2n$-gons.

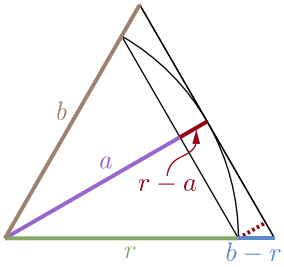


Figure 2 A sector of the circle and sides of the inscribed and circumscribed regular n-gons.

In fact, *Vera Circuli* contained more than this result. Gregory proved that for all n, $|C_{2n} - I_{2n}| < \frac{1}{2}|C_n - I_n|$, and thus as $k \to \infty$, C_k and I_k converge to the area of the circle (in fact, Gregory coined the term *convergent*). So, if the circle has radius 1, we can use these values to approximate π. For instance, inscribed and circumscribed squares yield $I_4 = 2$ and $C_4 = 4$. Applying the formulas produces the tighter bounds, $I_8 = 2\sqrt{2} = 2.8284\ldots$ and $C_8 = 8\sqrt{2} - 8 = 3.3137\ldots$, and Table 1 shows the next several bounds. Moreover, Gregory proved a more general version of this theorem that applies to ellipses and hyperbolas.

Vera Circuli also contained Gregory's proof that the ancient Greek problem of squaring the circle is impossible. In particular, he claimed that the circumference of the circle cannot be obtained from the radius using addition, subtraction, multiplication, division, and the extraction of roots.

n	I_n	C_n	$C_n - I_n$
4	2	4	2
8	2.8284	3.3137	0.4852
16	3.0614	3.1825	0.1211
32	3.1214	3.1517	0.0302
64	3.1365	3.1441	0.0075
128	3.1403	3.1422	0.0018
256	3.1412	3.1417	0.0004

TABLE 1: Bounds for π from inscribed and circumscribed n-gons.

Gregory sent his manuscript to Christian Huygens who was 10 years his senior and a leading mathematician of the day. Rather than replying to Gregory directly, Huygens published a review of *Vera Circuli* identifying a flaw in Gregory's argument and asserting that some of Gregory's results had previously appeared his own work. This review initiated an unpleasant dispute between the two mathematicians.

Although Gregory was correct that it is impossible to square the circle, the mathematical community had to wait over two centuries for a rigorous proof—Lindemann's 1882 proof that π is transcendental. Despite the flaw in Gregory's work, twentieth century mathematicians Max Dehn and E. D. Hellinger wrote, "A modern mathematician will highly admire Gregory's daring attempt of a 'proof of impossibility' even if Gregory could not attain his aim." [**2**]

The disagreement between Gregory and Huygens reveals more than just the issue of the correctness of Gregory's proof and Huygens's accusation of plagiarism. Huygens was at heart a mathematical traditionalist—a geometer. Whereas Gregory was one of the new breed—an algebraist and a pioneer of the new field of calculus. As Scriba noted, Gregory "was one of the wild young men who wanted to tear down the barriers of traditional mathematics at almost any price, who wanted to view hitherto uncultivated areas. Inspired by hopes for as yet unheard-of results, he freely introduced new methods while at times he neglected necessary care for details and exactness." [**3**]

REFERENCES

[1] Alsina, C., Nelsen, R. B. (2010). *Charming Proofs: A Journey Into Elegant Mathematics*. Dolciani Mathematical Expositions, Vol. 42. Washington, DC: Mathematical Association of America.

[2] Dehn, M., Hellinger, E. D. (1943). Certain mathematical achievements of James Gregory. *Amer. Math. Monthly*. 50: 149–163.

[3] Scriba, C. J. (1983). Gregory's converging double sequence: A new look at the controversy between Huygens and Gregory over the "analytical" quadrature of the circle. *Hist. Math*. 10(3): 274–285.

Summary. We visually demonstrate recursive formulas for areas of certain regular polygons.

TOM EDGAR (MR Author ID: 821633) is an associate professor of mathematics at Pacific Lutheran University and is the editor-elect of *Math Horizons.* He enjoys searching for and learning about visual proofs.

DAVID RICHESON (MR Author ID: 642588) is a professor of mathematics at Dickinson College and is editor of *Math Horizons.* He is the author of *Tales of Impossibility: The 2000-Year Quest to Solve the Mathematical Problems of Antiquity* (Princeton University Press, 2019).

PROBLEMS

EDUARDO DUEÑEZ, *Editor*
University of Texas at San Antonio

EUGEN J. IONAŞCU, *Proposals Editor*
Columbus State University

JOSÉ A. GÓMEZ, Facultad de Ciencias, UNAM, Mexico; *CODY PATTERSON*, Texas State University; *RICARDO A. SÁENZ*, Universidad de Colima, Mexico; *ROGELIO VALDEZ*, Centro de Investigación en Ciencias, UAEM, Mexico; *Assistant Editors*

Proposals

To be considered for publication, solutions should be received by May 1, 2020.

2081. *Proposed by Ioan Băetu, Botoşani, Romania.*

Finitely many distinct complex numbers $z_1, z_2, \ldots, z_n$ are called *associate over* $\mathbb{Q}$ if they are the n roots of a degree-n irreducible polynomial in $\mathbb{Q}[X]$. Assume that $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ are collections of complex numbers that are associate over $\mathbb{Q}$. If $y_1 - x_1$, $y_2 - x_2$, $\ldots$, $y_n - x_n$ are rational, prove that $y_1 - x_1 = y_2 - x_2 = \cdots = y_n - x_n$.

2082. *Proposed by Yves Nievergelt, Department of Mathematics, Eastern Washington University, Cheney, WA.*

Let

$$L = \begin{pmatrix} 1 & & & \\ & \ddots & & \\ & & 1 & \\ z_1 & \ldots & z_n & 1 \end{pmatrix}$$

be any $(n+1) \times (n+1)$ real matrix differing from the identity matrix only on the nondiagonal entries $\mathbf{z} = (z_1, \ldots, z_n)$ of its last row. Let L^T be the transpose of L. Find the largest and smallest eigenvalues of the matrices $L^T L$ and LL^T in terms of $\mathbf{z}$.

2083. *Proposed by Paul Bracken, University of Texas Rio Grande Valley, Edinburg, TX.*

Math. Mag. **92** (2019) 387–395. doi:10.1080/0025570X.2019.1673627

We invite readers to submit original problems appealing to students and teachers of advanced undergraduate mathematics. Proposals must always be accompanied by a solution and any relevant bibliographical information that will assist the editors and referees. A problem submitted as a Quickie should have an unexpected, succinct solution. Submitted problems should not be under consideration for publication elsewhere.

Proposals and solutions should be written in a style appropriate for this MAGAZINE.

Authors of proposals and solutions should send their contributions using the Magazine's submissions system hosted at mathematicsmagazine.submittable.com. *More detailed instructions are available there. We encourage submissions in PDF format, ideally accompanied by LaTeX source. General inquiries to the editors should be sent to* mathmagproblems@maa.org.

Let a_1 be a positive real number. Define the sequence $\{a_n\}$ recursively by $a_{n+1} = n^2/a_n$ for $n = 1, 2, \ldots$. Evaluate

$$\lim_{n\to\infty} \frac{1}{\ln n} \sum_{k=1}^{n} \frac{1}{a_k}$$

in terms of a_1.

2084. *Proposed by Andrei Ionescu, Lucretiu Patrascanu High School, Romania.*

A *random n-tournament* is a simple complete directed graph on n vertices in which the direction of each edge is chosen uniformly and independently at random. A vertex of a random tournament is called a *Rome* if it can be reached from every other vertex: "*all roads lead to Rome*." Let R be the number of Romes, and let $\mathbb{E}_n[R]$ be the expected number of Romes in a random n-tournament. Find

$$\lim_{n\to\infty} \frac{\mathbb{E}_n[R]}{n}.$$

2085. *Proposed by Florin Stanescu, Serban Cioculescu School, Gaesti, Romania.*

Solve the equation

$$3^x \cdot 4^y + 5^z = 7^w$$

in nonnegative integers w, x, y, z.

Quickies

1095. *Proposed by Moubinool Omarjee, Lycée Henri IV, Paris, France.*

Is there a nonnegative real sequence (x_n) such that $\sum_{n=1}^{\infty} |\sin n| \cdot x_n$ converges, but $\sum_{n=1}^{\infty} x_n/2^n$ diverges?

1096. *Proposed by George Stoica, New Brunswick, Canada.*

Given positive integers $n \geq 2$ and $k \leq n$, prove that there exists a degree-$(n+1)$ real polynomial $P(x)$ such that $P(x)$ and $d^jP(x)/dx^j$ have at least one common root for all positive $j \leq n$ such that $j \neq k$.

Solutions

A Fibonacci oddity **December 2018**

2056. *Proposed by Armend Sh. Shabani, University of Prishtina, Republic of Kosovo.*

Let n and k be positive integers. Regard the $2k$ numbers $1, 2, 3, \ldots, 2k$ as letters of an alphabet. Find the number of length-n words (using letters from said alphabet with repetitions allowed) in which no two odd letters are adjacent.

Solution by José Heber Nieto, Universidad del Zulia, Maracaibo, Venezuela.
The number of such words is $k^n F_{n+2}$, where F_n is the nth Fibonacci number (defined by the well-known recurrence $F_0 = 0$, $F_1 = 1$, and $F_{n+2} = F_{n+1} + F_n$ for $n \geq 0$). Of the totality of length-n allowable words, i.e., words without adjacent odd letters, let a_n

be the number of those whose last letter is odd, and b_n the number of those whose last letter is even. Clearly, $a_1 = b_1 = k$ (the number of even, or of odd, letters is k), $a_2 = k^2$ (length-2 allowable words ending in any of k odd letters must start with any of k even letters), $b_2 = 2k^2$ (length-2 allowable words ending with an even letter do start with any of the $2k$ letters) and, for $n \geq 1$, $a_{n+1} = kb_n$ and $b_{n+1} = kb_n + ka_n$ (length-$(n+1)$ admissible words ending in an odd letter are obtained appending any of k letters to an admissible length-n word ending in an even letter; those ending in an even letter are obtained by appending one of k even letters to either of the $a_n + b_n$ admissible words of length n). The recursive relations above continue to hold true for $n = 0$ if we let $a_0 = 0$ and $b_0 = 1$. Thus, for $n \geq 0$, we have $b_{n+2} = kb_{n+1} + ka_{n+1} = kb_{n+1} + k^2 b_n$, and it follows that $k^{-(n+2)} b_{n+2} = k^{-(n+1)} b_{n+1} + k^{-n} b_n$. Hence, the sequence $c_n = k^{-n} b_n$ satisfies the Fibonacci recurrence $c_{n+2} = c_{n+1} + c_n$ for $n \geq 0$. Since $c_0 = k^{-0} b_0 = 1$ and $c_1 = k^{-1} b_1 = 1$, a straightforward inductive argument shows that $c_n = F_{n+1}$, so $b_n = k^n c_n = k^n F_{n+1}$. Thus, for $n \geq 1$, we see that $a_n = kb_{n-1} = k \cdot k^{n-1} F_n = k^n F_n$, and the number of length-n admissible words is $a_n + b_n = k^n (F_n + F_{n+1}) = k^n F_{n+2}$. This number may be expressed in closed algebraic form using Binet's formula $F_n = (\phi^n - \varphi^n)/\sqrt{5}$ as

$$a_n + b_n = k^n F_{n+2} = \frac{k^n}{\sqrt{5}} (\phi^{n+2} - \varphi^{n+2}),$$

where $\phi = (1 + \sqrt{5})/2$ is the golden ratio, and $\varphi = (1 - \sqrt{5})/2$ its algebraic conjugate.

Also solved by Armstrong Problem Solvers Georgia Southern University, Michel Bataille (France), Vincent Blevins, Robert Calcaterra, Kyle Draghi, Natacha Fontes-Merz, Gregory Dresden, Kyle Gatesman, GWstat Problem Solving Group at The George Washington University, The Iowa State Undergraduate Problem Solving Group, Kathleen E. Lewis (Gambia), Elias Lampakis (Greece), Peter McPolin (Northern Ireland), Northwestern University Math Problem Solving Group, Randy K. Schwartz, Jacob Siehler, Enrique Treviño, Edward and Roberta White and the proposer. There were three incomplete or incorrect solutions.

Limit averages of sequences alternating at factorials **December 2018**

2057. *Proposed by Enrique Treviño, Lake Forest College, Lake Forest, IL.*

Let the sequence $x_1, x_2, \ldots, x_n, \ldots$ be defined by $x_1 = 0$, $x_2 = x_3 = x_4 = x_5 = 1$, $x_6 = x_7 = \cdots = x_{23} = 0$, $x_{24} = x_{25} = \cdots = x_{119} = 1, \ldots, x_{(2k-1)!} = \cdots = x_{(2k)!-1} = 0$, $x_{(2k)!} = \cdots = x_{(2k+1)!-1} = 1, \ldots$ (for $k = 1, 2, 3, \ldots$).

(*i*) For $n \geq 1$, let

$$a_n = \frac{1}{n} \sum_{i=1}^{n} x_i$$

be the nth arithmetic average of (x_n). Is the sequence $a_1, a_2, \ldots, a_n, \ldots$ convergent? If so, find its limit.

(*ii*) For $n \geq 1$, let

$$b_n = \frac{1}{H_n} \sum_{i=1}^{n} \frac{x_i}{i}$$

be the nth harmonically weighted average of (x_n), where

$$H_n = \sum_{i=1}^{n} \frac{1}{i}$$

is the nth harmonic sum. Is the sequence $b_1, b_2, \ldots, b_n, \ldots$ convergent? If so, find its limit.

Solution by The Iowa State Undergraduate Problem Solving Group, Iowa State University, Ames, IA.

(*i*) The sequence (a_n) is divergent. On the one hand,

$$a_{(2n+1)!-1} = \frac{1}{(2n+1)!-1}\sum_{j=1}^{n}[(2j+1)!-(2j)!] \geq \frac{(2n+1)!-(2n)!}{(2n+1)!}$$

$$= \frac{1}{1+1/2n},$$

so $\limsup_{k\to\infty} a_k \geq \limsup_{n\to\infty} 1/(1+1/2n) = 1$. On the other hand,

$$a_{(2n+2)!-1} = \frac{1}{(2n+2)!-1}\sum_{j=1}^{n}[(2j+1)!-(2j)!] \leq \frac{n[(2n+1)!-(2n)!]}{(2n+2)!-1}$$

$$= \frac{n\cdot 2n\cdot(2n)!}{(2n+2)!-1} = \frac{2n^2}{(2n+1)(2n+2)-1/(2n!)},$$

so $\liminf_{k\to\infty} a_k \leq \liminf_{n\to\infty} 2n^2/[(2n+1)(2n+2)-1/(2n)!] = 1/2$. It follows that (a_n) is divergent.

(*ii*) We prove that the sequence (b_n) converges to $1/2$.

Since $H_n = \ln n + \gamma + O(1/n)$ (where $\gamma \approx 0.577$ is the Euler–Mascheroni constant), it follows that $H_{n!-1} = H_{n!} - 1/n! = \ln(n!) + \gamma + O(1/n!)$, so

$$H_{(2k+1)!-1} - H_{(2k)!-1} = \ln((2k+1)!) - \ln((2k)!) + O\left(\frac{1}{(2k)!}\right)$$

$$= \ln(2k+1) + O\left(\frac{1}{(2k)!}\right);$$

hence,

$$\sum_{k=1}^{n}\left[H_{(2k+1)!-1} - H_{(2k)!-1}\right] = \sum_{k=1}^{n}\left[\ln(2k+1) + O\left(\frac{1}{(2k)!}\right)\right]$$

$$= \ln\left(\frac{(2n+1)!}{2^n n!}\right) + O(1).$$

From the definitions of (x_n) and (b_n), we have

$$b_{(2n+1)!-1} = \frac{1}{H_{(2n+1)!-1}}\sum_{k=1}^{n}\sum_{j=(2k)!}^{(2k+1)!-1}\frac{1}{j} = \frac{1}{H_{(2n+1)!-1}}\sum_{k=1}^{n}\left[H_{(2k+1)!-1} - H_{(2k)!-1}\right].$$

From the expressions above, we find

$$\lim_{n\to\infty} b_{(2n+1)!-1} = \lim_{n\to\infty}\frac{\ln((2n+1)!) - n\ln 2 - \ln(n!) + O(1)}{\ln((2n+1)!)}$$

$$= 1 - \lim_{n\to\infty}\frac{n\log n}{(2n+1)\ln(2n+1)} = \frac{1}{2}.$$

Note that

$$b_n - b_{n-1} = \frac{H_{n-1}\sum_{i=1}^{n}\frac{x_i}{i} - H_n\sum_{i=1}^{n-1}\frac{x_i}{i}}{H_{n-1}H_n} = \frac{\frac{x_n}{n}\sum_{i=1}^{n-1}\frac{1}{i} - \frac{1}{n}\sum_{i=1}^{n-1}\frac{x_i}{i}}{H_{n-1}H_n}.$$

Since (x_n) takes only the values $0, 1$, the equation above shows that $b_n \geq b_{n-1}$ if $x_n = 1$, while $b_n \leq b_{n-1}$ if $x_n = 0$. It follows that

$$\limsup_{k\to\infty} b_k = \lim_{n\to\infty} b_{(2n+1)!-1} = \frac{1}{2}.$$

A similar argument shows that $\liminf_{k\to\infty} b_k = \lim_{n\to\infty} b_{(2n)!-1} = 1/2$; this shows that $\lim_{n\to\infty} b_n = 1/2$.

Also solved by Robert Calcaterra, Dmitry Fleischman, Russell Gordon, Eugene A. Herman, Elias Lampakis (Greece), José Heber Nieto (Venezuela), and the proposer. There was one incomplete or incorrect solution.

A sextic with Galois group S_3 **December 2018**

2058. *Proposed by Gregory Dresden, Saimon Islam (student) and Jiahao Zhang (student), Washington & Lee University, Lexington, VA.*

Let a be a rational number such that the polynomial

$$f(x) = x^6 + 3x^5 - ax^4 - (2a+5)x^3 - ax^2 + 3x + 1$$

is irreducible over $\mathbb{Q}$, and let F be the splitting field for $f(x)$ over $\mathbb{Q}$. Find the Galois group $\mathrm{Gal}(F/\mathbb{Q})$ (up to isomorphism).

Solution by Robert Calcaterra, University of Wisconsin-Platteville, Platteville, WI.
The Galois group $G = \mathrm{Gal}(F/\mathbb{Q})$ is isomorphic to the symmetric group S_3. Observe that $f(x)$ is palindromic, so $x^6 f(1/x) = f(x)$; also, $f(-x-1) = f(x)$. It follows that z is a zero of $f(x)$ if and only if $\iota(z) := 1/z$ is a zero thereof, if and only if $\tau(z) := -z - 1$ is. Thus, ι and τ are involutions (i.e., ι^2 and τ^2 are both the identity transformation) acting on the set of zeros of f. (They may be regarded formally as projective transformations of $\widehat{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$.) Let $\sigma = \iota\tau$ be the transformation $z \mapsto -1/(1+z)$. The group of transformations generated by ι and τ is evidently the same as that generated by σ and τ (since $\sigma = \iota\tau$ and $\iota = \sigma\tau$). It is easy to check that σ^3 is the identity transformation, and $\tau\sigma = \sigma^2\tau$. Therefore, the group $\mathfrak{T} = \langle \iota, \tau\rangle = \langle \sigma, \tau\rangle$ generated by ι, τ (or by σ, τ) is isomorphic to the dihedral group D_6 (i.e., to the symmetric group S_3); it consists of the elements $\mathrm{id}, \sigma, \sigma^2, \tau, \sigma\tau, \sigma^2\tau$. Since ι, τ act on the set of roots of $f(x)$, so does $\mathfrak{T}$.

Lemma. *$\mathfrak{T}$ acts on the set of roots of $f(x)$ simply, i.e., given a root z of $f(x)$ and transformations $\alpha \neq \beta$ in $\mathfrak{T}$, we have $\alpha(z) \neq \beta(z)$.*

Proof. Note that the group $\mathfrak{T}$ consists of degree-1 projective transformations with coefficients in $\mathbb{Q}$, i.e., $\alpha(z)$ and $\beta(z)$ are quotients of polynomials of degree at most 1 in z, not both constant, with coefficients in $\mathbb{Q}$. We see that $\alpha(z)$ and $\beta(z)$ are roots of $f(x)$ since $\mathfrak{T}$ acts on roots of $f(x)$ and z is one such root. If we had $\alpha(z) = \beta(z)$, clearing denominators in this equation one sees that z would be root of a linear or quadratic equation with rational coefficients, contradicting the hypothesis that z is a root of the degree-6 polynomial $f(x)$ that is irreducible over $\mathbb{Q}$. ■

Fix a root z of $f(x)$. By the lemma above, the set of six distinct roots of $f(x)$ is $\{\alpha(z) : \alpha \in \mathfrak{T}\}$. For all $\alpha \in \mathfrak{T}$, the complex number $\alpha(z)$ is a rational expression in z

with rational coefficients; thus, every field automorphism $g \in G$ satisfies $g(\alpha(z)) = \alpha(g(z))$. Since $f(x)$ is irreducible, G acts transitively on the set of these six roots; in particular, for each $\alpha \in \mathfrak{T}$ there is $g \in G$ such that $\alpha(z) = g(z)$. If $g, h \in G$ satisfy $g(z) = \alpha(z) = h(z)$, then the images under g, h of any fixed root z' of $f(x)$ (necessarily of the form $z' = \beta(z)$ for some $\beta \in \mathfrak{T}$) must coincide: $g(z') = g(\beta(z)) = \beta(g(z)) = \beta\alpha(z) = \beta(h(z)) = h(\beta(z)) = h(z')$. It follows that, given $\alpha \in \mathfrak{T}$, a unique automorphism $g = g_\alpha$ of F is determined by the condition $g(z) = \alpha(z)$. Given $\alpha, \beta \in \mathfrak{T}$, we have $g_{\alpha\beta} = g_\beta g_\alpha$, since $g_{\alpha\beta}(z) = \alpha\beta(z) = \alpha(g_\beta(z)) = g_\beta(\alpha(z)) = g_\beta(g_\alpha(z))$. Therefore, $\alpha \mapsto g_\alpha$ is an isomorphism between G and the opposite group of $\mathfrak{T}$, which is still isomorphic to S_3.

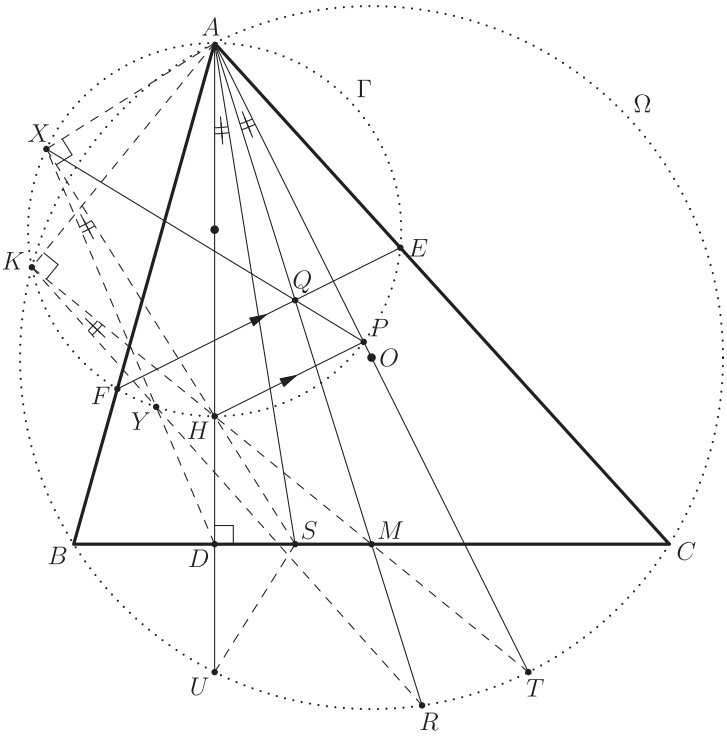
Also solved by Anthony Bevelacqua, Peter McPolin (Northern Ireland), Michael Reid, and the proposer.

A canonical similarity transformation of a given triangle **December 2018**

2059. *Proposed by Andrew Wu, St. Albans School, McLean, VA.*

Let triangle $\triangle ABC$ be acute and scalene with orthocenter H, altitudes $\overline{AD}$, $\overline{BE}$, and $\overline{CF}$, and circumcircle Ω. Let Γ be the circle with diameter $\overline{AH}$. Circles Γ and Ω intersect at A and at a second point K. Let point P lie on Γ so that $\overline{HP}$ is parallel to $\overline{EF}$. Let M be the midpoint of $\overline{BC}$. Let $\overleftrightarrow{AM}$ intersect Ω at $R \neq A$, and $\overline{EF}$ at Q. Let $\overleftrightarrow{PQ}$ meet Γ again at $X \neq P$. Show that $\overline{DX}$ and $\overline{KR}$ concur on Γ.

Solution by Kyle Gatesman (student), Johns Hopkins University, Baltimore, MD.



Let Y be the intersection of $\overline{DX}$ and $\overline{KR}$. Let U be the reflection of H on $\overleftrightarrow{BC}$; thus, $\angle USD = \angle HSD$, and it is well known that U lies on Ω. Let O be the circumcenter of

$\triangle ABC$, and let T be the point diametrically opposite to A on Ω. The angles of $\triangle ABC$ will be denoted $\angle A = \angle BAC$, $\angle B = \angle ABC$, $\angle C = \angle ACB$.

Since $\overline{BE}$ and $\overline{CF}$ are altitudes of $\triangle ABC$, the angles $\angle BEA = \angle HEA$ and $\angle CFA = \angle HFA$ are both right; hence, E and F lie on the circle Γ with diameter $\overline{AH}$, and moreover $BCEF$ is cyclic. It follows that $\angle AEF = \angle B$ and $\angle AFE = \angle C$, so we have a similarity $\triangle ABC \sim \triangle AEF$.

Let $\mathfrak{I}$ be the similarity transformation taking $\triangle ABC$ to $\triangle AEF$. ($\mathfrak{I}$ is the composition of a homotethy with center A and the reflection on the bisector of $\angle A$.) Then $\mathfrak{I}$ takes Ω to Γ and the diameter $\overline{AT}$ of Ω to the diameter $\overline{AH}$ of Γ; thus, $\angle FAH = \angle CAT$. On the other hand, $\overleftrightarrow{HP}$ is parallel to $\overleftrightarrow{EF}$ by construction, so the arcs $\widehat{FH}$, $\widehat{EP}$ of Γ are equal, hence $\angle CAP = \angle EAP = \angle FAH$, being angles inscribed in equal arcs of Γ. It follows that $\angle CAP = \angle CAT$, so P lies on the diameter $\overline{AT}$ of Ω.

Let $\overleftrightarrow{AS}$ be the symmedian of $\triangle ABC$ though A, i.e., S lies on $\overline{BC}$ so that $\overleftrightarrow{AS}$ is the image of the median $\overleftrightarrow{AM}$ of $\triangle ABC$ on the bisector of angle $\angle BAC$. Clearly, $\overleftrightarrow{AS}$ is the image of $\overleftrightarrow{AM}$ under $\mathfrak{I}$. It follows that the similarity $\mathfrak{I}$ transforms A, B, C, S, T, U, respectively, into A, E, F, Q, H, P. Since $\overline{AH}$ is a diameter of Γ, the arcs $\widehat{AX}$, $\widehat{XH}$ are supplementary, so the inscribed angles $\angle HPX$, $\angle AHX$ are complementary. On the other hand, the angles $\angle DSH$, $\angle DHS$ of the right triangle $\triangle DHS$ are also complementary. Since similarity preserves angles and $\overleftrightarrow{HP}$ is parallel to $\overleftrightarrow{EF}$, we have $\angle HPX = \angle FQX = \angle EQP = \angle BSU = \angle DSH$. Thus, angles $\angle AHX$ and $\angle DHS$ have equal complements, so they must be equal. We conclude that X, H, S are collinear.

Since angles $\angle AKH$ and $\angle AKT$ are both right (each being inscribed in a half-circle), they are equal, so K, H, and T are collinear. Similarly, angle $\angle AXS = \angle AXH$ is right, as is $\angle ADS = \angle ADC$ ($\overline{AD}$ is an altitude of $\triangle ABC$). Thus, quadrilateral $ASDX$ is cyclic. Now, we have

$$\begin{aligned}
\angle HKY &= \angle TKR && (K, H, T \text{ and } K, Y, R \text{ collinear})\\
&= \angle TAR && (\text{both inscribed in } \widehat{TR} \text{ of } \Omega)\\
&= \angle PAQ = \angle UAS = \angle DAS && (P, Q \text{ correspond to } U, S \text{ under } \mathfrak{I})\\
&= \angle DXS && (\text{quadrilateral } ASDX \text{ is cyclic})\\
&= \angle HXY && (H, S, X \text{ are collinear}).
\end{aligned}$$

We conclude that H, K, X, Y are concyclic, so Y lies on Γ.

Also solved by Herb Bailey, Dixon Jones, and the proposer.

Positive matrices in a Gaussian Orthogonal Ensemble **December 2018**

2060. *Proposed by Su Pernu Mero, Valenciana GTO, Mexico.*

Let A, B, C be independent standard normal random variables, i.e., the PDF of each is $f(x) = \exp(-x^2/2)/\sqrt{2\pi}$. Consider the random matrix

$$M = \begin{pmatrix} A & B/\sqrt{2} \\ B/\sqrt{2} & C \end{pmatrix}.$$

What is the probability that M is positive definite?

Solution by GWstat Problem Solving Group, The George Washington University, Washington, DC.

The probability is $p = (2-\sqrt{2})/4$ (about 14.6%). The event that matrix M be positive definite is characterized by the inequalities $A > 0$ and $0 < \det M = AC - B^2/2$; thus, $p = \mathbb{P}\{A > 0, 2AC > B^2\}$. Since the probability distribution of standard normal variables, and hence of M, are invariant under sign changes, we see that $p = \mathbb{P}\{A < 0, 2AC > B^2\}$ is also the probability that M be negative definite. It follows that $p = \frac{1}{2}\mathbb{P}\{2AC > B^2\}$ (the event $\{A = 0\}$ has zero probability). Thus,

$$p = \frac{1}{2} \cdot \mathbb{P}\{2AC - B^2 > 0\} = \frac{1}{2} \cdot \mathbb{P}\left\{\left(\frac{A+C}{\sqrt{2}}\right)^2 - \left[\left(\frac{A-C}{\sqrt{2}}\right)^2 + B^2\right] > 0\right\}.$$

Lemma. *The variables $X = (A+C)/\sqrt{2}$, $Y = (A-C)/\sqrt{2}$ and B are standard normal and pairwise independent.*

Proof of the lemma. Since B is standard normal and independent of A, C, it is enough to show that X and Y are standard normal and pairwise independent. Since the transformation $(x, y) = T(a, c) = ((a+c)/\sqrt{2}, (a-c)/\sqrt{2})$ is orthogonal, it preserves the quadratic form $a^2 + c^2$ and hence the joint density $f(a, c) = f(a)f(c) = \exp[-(a^2 + c^2)/2]/(2\pi)$ of A, C. Thus, X, Y are jointly identically distributed to A, C, so they are standard normal and independent. ■

It follows from the lemma that $p = \frac{1}{2}\mathbb{P}\{\chi_1^2 - \chi_2^2 > 0\}$, where $\chi_1^2 = X^2$ and $\chi_2^2 = Y^2 + B^2$ are independent chi-square variables having 1 and 2 degrees of freedom, respectively. These nonnegative variables have density $f_1(x) = x^{-1/2}e^{-x/2}/\sqrt{2\pi}$ and $f_2(z) = e^{-z/2}/2$, respectively, supported on $(0, \infty)$.

The density function of $W = \chi_1^2 - \chi_2^2$ is given by:

$$g(w) = \begin{cases} \displaystyle\int_0^\infty f_1(z+w)\, f_2(z)\, dz, & w \geq 0; \\ \displaystyle\int_0^\infty f_1(x)\, f_2(x-w)\, dx, & w \leq 0. \end{cases}$$

For $w \leq 0$:

$$g(w) = \int_0^\infty \frac{1}{\sqrt{2\pi}} x^{-1/2} e^{-x/2} \cdot \frac{1}{2} e^{-(x-w)/2}\, dx = \frac{e^{w/2}}{2\sqrt{2}} \int_0^\infty \frac{1}{\sqrt{\pi}} x^{-1/2} e^{-x} dx$$
$$= \frac{e^{w/2}}{2\sqrt{2}}.$$

In conclusion, we find

$$p = \frac{\mathbb{P}\{W > 0\}}{2} = \frac{1 - \mathbb{P}\{W \leq 0\}}{2} = \frac{1}{2}\left(1 - \int_{-\infty}^0 \frac{e^{w/2}}{2\sqrt{2}}\, dw\right) = \frac{1}{2}\left(1 - \frac{1}{\sqrt{2}}\right).$$

Editor's Note. The *Gaussian Orthogonal Ensemble GOE(2)* consists of 2×2 real symmetric matrices M with the given probability distribution. In the analogous ensemble GOE(n) of $n \times n$ real symmetric matrices M, it is known that the probability that a random such M be positive definite decays asymptotically as $\exp(-cn^2)$ when n tends to infinity, where $c = \ln\sqrt{3} = 0.5493\ldots$. (D. S. Dean, S. N. Majumdar, Extreme value statistics of eigenvalues of Gaussian random matrices, *Phys. Rev. E* **77** (2008) 041108, arxiv.org/abs/0801.1730.)

Also solved by Robert A. Agnew, Robert Calcaterra, Bruce E. Davis, John Fitch, J. A. Grzesik, Northwestern University Math Problem Solving Group, and the proposer. There were two incomplete or incorrect solutions.

Answers

Solutions to the Quickies from page 302.

A1095. Such a sequence (x_n) does not exist. Using classical estimates, due to Mahler, of the irrationality measure of π, Christopher Stuart has shown that $|\sin n| \geq \alpha^{-n}$ for all positive integers n, where $\alpha = \sqrt[3]{\sin 3} = 1.9207\ldots$ (Christoper Stuart, An Inequality Involving $\sin(n)$, *Amer. Math. Monthly* **125** (2018) 173–174). Since $\alpha < 2$, we have $\sum_{n=1}^{\infty} |\sin n| \cdot x_n \geq \sum_{n=1}^{\infty} x_n/\alpha^n \geq \sum_{n=1}^{\infty} x_n/2^n$. Thus, it is not simultaneously possible for $\sum_{n=1}^{\infty} |\sin n| \cdot x_n$ to converge and $\sum_{n=1}^{\infty} x_n/2^n$ to diverge.

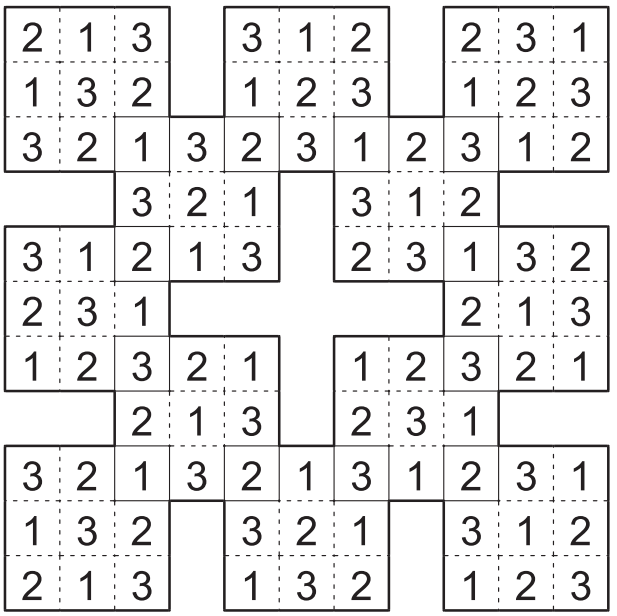
A1096. Given a real $(n-1)$-tuple $\mathbf{a} = (a_1, \ldots, a_{n-1})$ in $[0, 1]$, let $P_{\mathbf{a}}(x)$ be the degree-$(n+1)$ polynomial $x(x-1)(x-a_1)\ldots(x-a_{n-1})$. Clearly, $P_{\mathbf{a}}$ has all its $n+1$ roots in $[0, 1]$; therefore, for $j = 1, \ldots, n$, the derivative $P_{\mathbf{a}}^{(j)} = d^j P_{\mathbf{a}}/dx^j$ also has all its $n+1-j = \deg P_{\mathbf{a}}^{(j)}$ roots $b_{j,1} \leq b_{j,2} \leq \cdots \leq b_{j,n+1-j}$ (listed with multiplicity) in $[0, 1]$.

For $j = 1, \ldots, n$, fix an arbitrary integer i_j such that $1 \leq i_j \leq n+1-j$. Let $f : [0, 1]^{n-1} \to [0, 1]^{n-1}$ be the mapping

$$\mathbf{a} = (a_1, a_2, \ldots, a_{n-1}) \mapsto \mathbf{b} = (b_{1,i_1}, b_{2,i_2}, \ldots, \widehat{b_{k,i_k}}, \ldots, b_{j_n,i_n}),$$

where $b_{j,i}$ is the ith root of $P_{\mathbf{a}}^{(j)}(x)$ as above, and the entry b_{k,i_k} is omitted to obtain the $(n-1)$-tuple $\mathbf{b}$. An elementary argument shows that the mapping f is continuous. The $(n-1)$-cube $[0, 1]^{n-1}$ is homeomorphic to the unit $(n-1)$-ball; thus, by Brouwer's fixed-point theorem, the mapping f fixes some point $\mathbf{a} \in [0, 1]^{n-1}$. From the definition of f, the polynomial $P(x) = P_{\mathbf{a}}(x)$ shares the root $a_j = b_{j,i_j}$ (resp., the root $a_{j-1} = b_{j,i_j}$) with $d^j P/dx^j$ as j ranges over $1, 2, \ldots, k-1$ (resp., over $k+1, \ldots, n$). Note that we are not merely ensuring that P and $P^{(j)}$ have a common root for $j \neq k$, but in fact specifying that they share precisely the ith root of $P^{(j)}$ (since the choice of i_j given j is otherwise arbitrary).

TRIBUS Solution



REVIEWS

PAUL J. CAMPBELL, *Editor*
Beloit College

Assistant Editor: Eric S. Rosenthal, West Orange, NJ. Articles, books, and other materials are selected for this section to call attention to interesting mathematical exposition that occurs outside the mainstream of mathematics literature. Readers are invited to suggest items for review to the editors.

Granville, Andrew, and Jennifer Granville, with illustrations by Robert J. Lewis, *Prime Suspects: The Anatomy of Integers and Permutations*, Princeton University Press, 2019; 229 pp, $22.95. ISBN 978-0-691-14915-8.

Renowned number-theorist Andrew Granville (U. de Montréal) here explores the graphic novel as a format to popularize mathematical discoveries. This absolutely brilliantly illustrated book arose from an earlier play and an accompanying commissioned musical piece. The mathematics too is astonishing—but explained understandably. It is about the "anatomical" similarities of the structures of "elements" (integers, permutations, polynomials in finite fields) in different areas of mathematics: the proportions of "indecomposable parts" (primes, cycles, irreducible polynomials) among such elements, the typical number of indecomposable parts in an element, the size of those parts, the size of the smallest part or of the largest part, and the proportion of elements with a particular number of indecomposable parts. The plot of the novel is about the similarities of two murders, the detective characters are take-offs on famous mathematicians, and the graphic frames are chock full of mathematical allusions. The last will be appreciated only by mathematical sophisticates, as will be the details in a 26-page appendix by Granville on the mathematics—which professors are going to need to learn so that they can offer explanations to students who read the novel (ever heard of Buchstab's function?). The endpapers are a treasure in themselves. [A few typos occur: "Roslyn" (p. 218), "co-incidences" (189), "K.F. Gauss" (185), and (intended?) "Leider ohne worte."]

Tao, Terence, What's new: Almost all Collatz orbits attain almost bounded values, terrytao.wordpress.com/2019/09/10/almost-all-collatz-orbits-attain-almost-bounded-values/.

Tao, Terence, Almost all orbits of the Collatz map attain almost bounded values, arxiv.org/abs/1909.03562.

The recurrence $a_{n+1} = a_n/2$ if a_n is even and $a_{n+1} = 3a_n + 1$ if a_n is odd, started from an initial integer $a_0 > 0$, generates an orbit through the integers. The Collatz conjecture is that the orbit of every initial a_0 contains 1. The conjecture has been verified for all integers through 10^{20}. Let f be any function that goes off to infinity as $n \to \infty$. Tao shows: For almost all a_0, the smallest value in the Collatz orbit of a_0 is smaller than $f(a_0)$. To interpret "almost bounded," think of $f(n) = n^\theta$ for some positive $\theta \approx 0$. The "almost all" is not in the sense of asymptotic density but of logarithmic density; still, it means that if there are any exceptions to the conjecture, they get rarer as $N \to \infty$. Further, "almost all numbers lie outside of periodic orbits."

Grossman, David, After 65 years, supercomputers finally solve this unsolvable math problem, popularmechanics.com/science/math/a28943849/unsolvable-math-problem/.

Can you find integers x, y, and z to solve $x^3 + y^3 + z^3 = k$, for k a positive integer? The answer for $k = 1, \ldots, 100$ is now known to be yes, with the cases $k = 33$ and $k = 42$ resolved at last. The latter required one million hours on 500,000 home Windows computers made available at charityengine.com.

Math. Mag. **92** (2019) 396–397. doi:10.1080/0025570X.2019.1673628

Harvey, David, and Joris van der Hoeven, Integer multiplication in time $O(n \log n)$, hal.archives-ouvertes.fr/hal-02070778/document.

Li, Mengxin, Mathematicians discover the perfect way to multiply, *Quanta Magazine*, quantamagazine.org/mathematicians-discover-the-perfect-way-to-multiply-20190411.

The method taught in schools to multiply two n-digit numbers requires $O(n^2)$ single-digit multiplications, plus additions. Divide-and-conquer methods, which trade (slow) multiplications for (faster) additions and subtractions, have been known since 1960; Li's article gives two example calculations of a method that requires $O(n^{1.58})$ multiplications. There has been progress since, culminating in this year's announcement by Harvey and van der Hoeven of an $O(n \log n)$ method. If a certain conjecture is true, then $O(n \log n)$ is the best that can be done.

Bressoud, David M., *Calculus Reordered: A History of the Big Ideas*, Princeton University Press, 2019; xvi+224 pp, $29.95. ISBN 978-0-691-18131-8.

"This book will not show you how to do calculus" but it will "explain how and why it arose." Author Bressoud criticizes the standard progression in a calculus course—limits, derivatives, integrals, and finally series—and favors the historical progression: accumulation (as the foundation for integration), differentiation, series, and then limits. He attributes the standard order to the need for rigor by research mathematicians in the 19th century, and notes that calculus as now taught "is appropriate for the student who wants to verify that calculus is logically sound. However, that describes very few students in first-year calculus." (Nevertheless, in my opinion there is scientific and liberal-arts value in students' experiencing the structure and mathematical method of starting from precise definitions, deducing sound theorems, and applying them carefully to applications.) The book follows the historical progression, is far from being a textbook, but serves as a "proof of concept." Bressoud suggests that the book requires little more than "mathematical curiosity," but only a reader who has already learned calculus can appreciate it. It would be a great companion for students studying analysis, and calculus instructors will find it an enriching experience.

Bartlett, Jonathan, and Asatur Zh. Khurshudyan, Extending the algebraic manipulability of differentials, *Dynamics of Continuous, Discrete and Impulsive Systems Series A: Mathematical Analysis* 26 (2019) 217–230. online.watsci.org/abstract_pdf/2019v26/v26n3a-pdf/4.pdf.

Have you—or your students—ever been bothered by the "shorthand" Leibniz notation for the second derivative,

$$\frac{d^2y}{dx^2}, \qquad \text{short for } \frac{d}{dx}\left(\frac{dy}{dx}\right) \qquad ?$$

The authors object that the shorthand version looks like a quotient but is rarely treated as such. Although d^2y suggests a second differential of y, what is dx^2? This difficulty does not arise with the Lagrange notation y'', nor with dy/dx, where we are happy to multiply through by the differential dx. The authors reject the ambiguity of writing dx but not meaning $d \times x$, which they (like some contemporary authors) avoid by writing the d not in math italic but in roman, as is customary for functions such as sin and log; using $\mathrm{d}(x)$ (some authors now use parentheses with sin and log) would be still clearer, but that would result in proliferation of parentheses. The authors suggest revising notation both for algebraic manipulability and to prevent student misconceptions and calculational mistakes. They point out that the Leibniz notation should mean the ratio of the second differential of y, viz., $\mathrm{d}(\mathrm{d}(y))$, to the square of $\mathrm{d}(x)$, viz., $(\mathrm{d}(x))^2$. Applying the quotient rule to find the differential of $\mathrm{d}y/\mathrm{d}x$ results (after division by $\mathrm{d}x$) in

$$y'' = \frac{\mathrm{d}^2y}{(\mathrm{d}x)^2} - \frac{\mathrm{d}y}{\mathrm{d}x}\frac{\mathrm{d}^2x}{(\mathrm{d}x)^2},$$

where $(\mathrm{d}x)^2$ replaces the authors' still-ambiguous $\mathrm{d}x^2$. "It is not very pretty or compact, but it works algebraically"—particularly for the chain rule if x is a function of some t. The authors make a strong case for a clear and honest notation that facilitates working with differentials, but the formula above is probably just as confusing for students as the shorthand expression.

END NOTES

In these pages, we offer corrections and notes for articles published in MATHEMATICS MAGAZINE. Unfortunately, these may not address all of the published mistakes, only the ones that we are aware of that are more problematic than typographic or similar errors.

Detecting Deficiencies: An Optimal Group Testing Algorithm, Seth Zimmerman, MATHEMATICS MAGAZINE **90** (3) (2017): 167–178, doi.org/10.4169/math.mag.90.3.167.

Yaakov Malinovsky had a number of observations about and a few corrections to Zimmerman's article. Zimmerman's algorithm, which we will denote as SZA, is an improvement of Sobel and Groll's dynamic programming algorithm, which they call Procedure R_3 [**3**]. The dynamic programming structure of SZA follows from the fact that the optimal design for stage t is constructed from optimal designs at stages $t-1,\ldots,1$.

SZA is a member of a nested class of group-testing algorithms defined by the property that if a positive subset I is identified, the next subset I_1 to be tested is a proper subset of I. Due to the additional restriction of assumption (vi), SZA is not optimal in the nested class. The optimal nested algorithm from this class is Procedure R_1 from [**3**]. To demonstrate that SZA is not optimal, using Zimmerman's example (on pages 172-173) with $q = 0.9999$, $n = 6765$, both Procedure R_3 and SZA give 12.94809 as the expected number of tests, while Procedure R_1 gives 10.14778. Although Procedure R_1 is optimal in the nested class, it is not optimal when non-nested procedures are considered [**2**]. An optimal group-testing procedure, for general n, is unknown.

Procedure R_3 has computational complexity proportional to the square of the population size, i.e., $O(n^2)$, where n is the population size. Theorems 2 and 3 show that SZA reduces this complexity by at least half, which amounts to an improvement in the speed of R_3. It follows that SZA also has complexity $O(n^2)$. For comparison, based on the work of Hwang [**1**], Procedure R_1 has computational complexity $O(n)$ (without sorting effort).

Finally, Zimmerman incorrectly claims on page 172, where for $q = 0.9999$, that

> For a population with $n > 6765$, a test that includes more than 6765 samples would always be disadvantageous.

For $n = 10000$, SZA yields 19.20284 as the expected number of tests. However, if the group of $n = 10000$ is divided into two groups of size 6765 and 3235, then the expected number of tests is $12.94809 + 6.34621 = 19.2943$, which is greater than the result for $n = 10000$.

REFERENCES

[1] Hwang, F. K. (1976). An optimal nested procedure in binomial group testing. *Biometrics*. 32: 939–943.

[2] Sobel, M. (1960). Group testing to classify efficiently all defectives in a binomial sample. In: Machol, R. E., ed. *Information and Decision Processes*. New York: McGraw-Hill, pp. 127–161.

[3] Sobel, M., Groll, P. A. (1959). Group testing to eliminate efficiently all defectives in a binomial sample. *Bell Syst. Tech. J.* 38: 1179–1252.

Math. Mag. **92** (2019) 398–400. doi:10.1080/0025570X.2019.1696125

From Chebyshev to Jensen and Hermite-Hadamard, Dan Ştefan Marinescu and Monea Mihai, Mathematics Magazine **91** (3) (2018): 213–217, doi.org/10.1080/0025570X.2018.1445376.

Michael Maltenfort noticed an error in the proof of Proposition 2. Earlier in the proof, the authors show that F is convex and that F is nonincreasing on $(0, 1]$. The authors then apply convexity (incorrectly) to show that F is nonincreasing on $[0, 1]$.

Because $x = x \cdot 1 + (1 - x) \cdot 0 \in (0, 1)$ and F is a convex function, then it follows that $F(x) \leq xF(1) + (1 - x)F(0)$, as opposed to, as the authors write, $F(x) \leq xF(0) + (1 - x)F(1)$. However, the correct inequality can still be used to finish the proof. Since $F(1) \leq F(x)$ for $x \in (0, 1)$, then $F(x) \leq xF(x) + (1 - x)F(0)$. This implies that $F(x) \leq F(0)$, since $(1 - x)$ is positive. This completes the proof that F in nonincreasing on $[0, 1]$.

Math Bite: When an Average of Averages is the Average, Tristen Pankake-Sieminski and Raymond Viglione, Mathematics Magazine **92** (2) (2019): 122, doi.org/10.1080/0025570X.2019.1573651.

Michael Maltenfort and Paul Stockmeyer both commented on the authors' claim that "If you want to find the mean of a data set, you would not say, split the data in half, find the average of each half, and average those results." However, that is precisely what one could do to find the average if the set of data has an even number of elements! Suppose the set is $S = \{a_1, \ldots, a_{2n}\}$ and $S = A \cup B$ where $|A| = |B| = n$. If $\bar{a}$ and $\bar{b}$ are the averages of the data in sets A and B, respectively, then the average of the data in set S, $\bar{s}$, satisfies $\bar{s} = (\bar{a} + \bar{b})/2$. As Stockmeyer writes, "More generally, whenever the size n of a data set has k as a proper divisor, one can find the mean by partitioning the data set into k subsets each of size n/k, averaging each of the k subsets, then averaging the results."

If S has an odd number of elements, then "half" would not be possible. But, as Maltenfort mentions, one could always use a weighted average. Suppose that $S = \{a_1, \ldots, a_5\}$ is partitioned into $A = \{a_1, a_2\}$ and $B = \{a_3, a_4, a_5\}$, so that $\bar{a} = \dfrac{a_1 + a_2}{2}$ and $\bar{b} = \dfrac{a_3 + a_4 + a_5}{3}$. Then, $\bar{s}$ is equal to the weighted average of $\bar{a}$ and $\bar{b}$ where $\bar{s} = (2/5)\bar{a} + (3/5)\bar{b}$. To be able to take the (nonweighted) average of the averages (of subsets of S) and have it equal the average ($\bar{s}$) suggests taking the average of subsets of equal size.

For $|S| = 5$ and subsets of size 3, then using the authors' idea to take the average of the averages of all subsets of size 3 yields

$$\frac{\bar{s}_{123} + \bar{s}_{124} + \bar{s}_{125} + \bar{s}_{134} + \bar{s}_{135} + \bar{s}_{145} + \bar{s}_{234} + \bar{s}_{235} + \bar{s}_{245} + \bar{s}_{345}}{10} = \frac{6(a_1 + \cdots + a_5)}{3 \cdot 10} = \bar{s},$$

where $\bar{s}_{ijk} = (a_i + a_j + a_k)/3$. The last equality holds because each a_i appears six times in a subset of size 3. But, taking the average of the averages from *all* subsets of size 3 is still not the *only* way to write $\bar{s}$ as the average of averages of sets of size 3. The key is that each element of S has to appear an equal number of times in the subsets. Another way to get the average from sets of size 3 is:

$$\frac{\bar{s}_{123} + \bar{s}_{234} + \bar{s}_{345} + \bar{s}_{451} + \bar{s}_{512}}{5} = \frac{3(a_1 + \cdots + a_5)}{3 \cdot 5} = \bar{s},$$

because each number a_i appears three times.

However, there are other families of subsets—not all of the same size—that still have the averaging property. For $|S| = 5$, then there are ten 2-element subsets and ten 3-element subsets. Taking the average of the averages of all 2- and 3-element subsets still yields the average. In general, for $|S| = n$, averaging the averages of all k- and $(n - k)$-element subsets yields the average of S.

Determining a set of k-element subsets of S so that the average of the averages of the subsets is equal to the average of S is related to combinatorial design theory. Given any positive integer t, a t-design B is a class of k-element subsets of S, called blocks, such that every element $a \in S$ appears in exactly r blocks, and every t-element subset T appears in exactly λ blocks.

The Instructor's Guide to Real Induction, Pete Clark, MATHEMATICS MAGAZINE **92** (2) (2019): 136–150, doi.org/10.1080/0025570X.2019.1549902.

Although Michael Maltenfort found this article to be the most enjoyable in the April issue, he was distracted by a number of errors. The most salient issues appear below.

In the proof of Theorem 3 that shows that a continuous function $f : [a, b] \to \mathbb{R}$ is bounded, when verifying condition RI2, it is stated that, if $x \in S$, then $y \in [x - \delta, x + \delta]$. Because x could equal a (a requirement of RI2), then y would have to be in the interval $[x, x + \delta]$ instead.

When verifying RI2 in the proof of an integrability theorem (Theorem 4), the assumption that $[a, x] \subseteq S(\epsilon)$ should instead be $x \in S(\epsilon)$. Further, near the end of the proof of condition RI3, there exists a partition $P_{x-\delta}$ that satisfies $U(f, P_{x-\delta}) - L(f, P_{x-\delta}) < (x - \delta - a)\epsilon$, not $U(f, P_{x-\delta}) = L(f, P_{x-\delta}) = (x - \delta - a)\epsilon$.

Triphos: A World Without Subtraction, Keely Grossnickle, Brian Hollenbeck, Jeana Johnson, and Zhihao Sun, MATHEMATICS MAGAZINE **92** (4) (2019): 272–285, doi.org/10.1080/0025570X.2019.1609293.

Norman Hill noticed a typographical error that could be "a slight stumbling block for someone lacking confidence in their own understanding." On page 275, the definitions of R and G were switched. The corrected passage is below.

> Being more methodical in our search for a satisfactory definition of Triphosian multiplication, observe that, by the definition of addition and scalar multiplication, any Triphosian number, ${}^{g}_{b}r$, can be expressed using an alternative notation using the *primaries* $G = {}^{1}_{0}0$, $R = {}^{0}_{0}1$, and $B = {}^{0}_{1}0$, so that
>
> $$ {}^{g}_{b}r = {}^{0}_{0}r + {}^{g}_{0}0 + {}^{0}_{b}0 = rR + gG + bB. $$

Real Mathematics Fact

Snapple includes "Real Facts" on the underside of the caps to its beverages. A number of them have been about mathematics, including the following.

Snapple's "Real Fact #1399" states that:

> "Never odd or even" spelled backwards is "Never odd or even."
>
> — Submitted by Nick Roopas
> Ann Arbor, MI