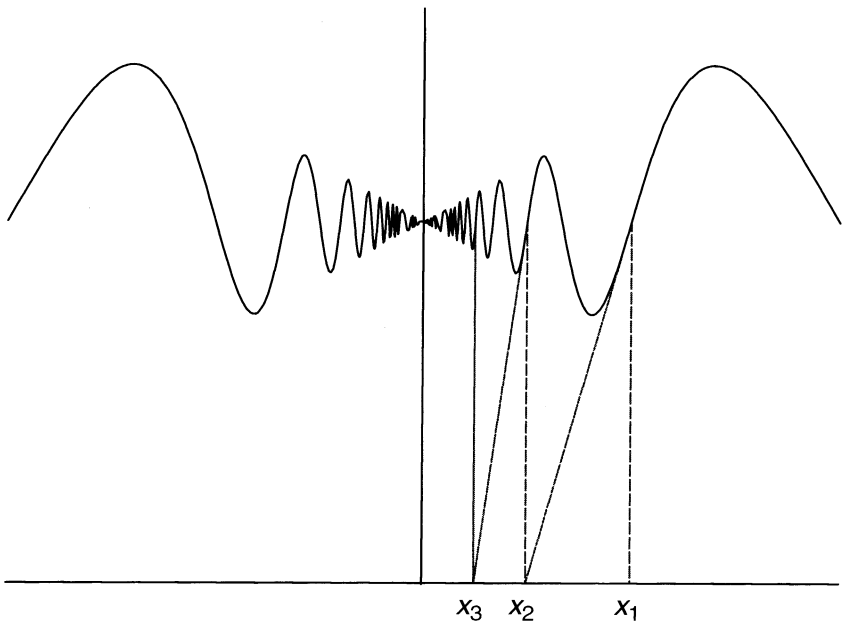
Vol. 80, No. 5, December 2007

MATHEMATICS MAGAZINE



Fooling Newton

- Tantalizing Steps into Semigroups
- Four Proofs of the Ballot Theorem
- A Property of Points on Quadratic Curves

An Official Publication of The MATHEMATICAL ASSOCIATION OF AMERICA

EDITORIAL POLICY

Mathematics Magazine aims to provide lively and appealing mathematical exposition. The *Magazine* is not a research journal, so the terse style appropriate for such a journal (lemma-theorem-proof-corollary) is not appropriate for the *Magazine*. Articles should include examples, applications, historical background, and illustrations, where appropriate. They should be attractive and accessible to undergraduates and would, ideally, be helpful in supplementing undergraduate courses or in stimulating student investigations. Manuscripts on history are especially welcome, as are those showing relationships among various branches of mathematics and between mathematics and other disciplines.

A more detailed statement of author guidelines appears in this *Magazine*, Vol. 74, pp. 75–76, and is available from the Editor or at www.maa.org/pubs/mathmag.html. Manuscripts to be submitted should not be concurrently submitted to, accepted for publication by, or published by another journal or publisher.

Submit new manuscripts to Allen Schwenk, Editor, Mathematics Magazine, Department of Mathematics, Western Michigan University, Kalamazoo, MI, 49008. Manuscripts should be laser printed, with wide line spacing, and prepared in a style consistent with the format of *Mathematics Magazine*. Authors should mail three copies and keep one copy. In addition, authors should supply the full five-symbol 2000 Mathematics Subject Classification number, as described in *Mathematical Reviews*.

The **cover image** illustrates a function for which Newton's method for seeking a root does converge nicely, but the point it converges to is nowhere near a root. For more about functions that "fool Newton," see the note by Peter Horton.

AUTHORS

Christopher Hollings is a research student in the Department of Mathematics at the University of York, UK. He completed a Master of Mathematics degree in York in 2004, at which point his love both of York and of semigroups compelled him to stay there to do a PhD in semigroup theory under the supervision of Dr. Victoria Gould. His particular interest is in the actions and partial actions of semigroups and monoids on sets. He is currently writing up his PhD thesis "Partial actions of monoids," which he expects to submit in mid-2007. The present article arose from his desire to bring the beauty of semigroup theory to a wider audience.

Marc Renault is an Associate Professor of Mathematics at Shippensburg University. He received his BS from Wake Forest University in 1994,and his PhD from Temple University in 2002. Though trained as an algebraist, his true passion is combinatorics, and he can be found whiling away the hours counting lattice paths. His greatest joy, however, comes from spending time with his wife, Tara, and two children, Olivia and Atticus.

Leonid Hanin is Professor of Mathematics at Idaho State University in Pocatello. He received his PhD in Analysis from Steklov Mathematical Institute, St. Petersburg, Russia, in 1985. He has published about 70 peer-reviewed works in functional analysis and function theory, biological applications of probability and stochastic processes, mechanical and thermal engineering, and theoretical biology. He coauthored a book on mathematical modeling in radiation biology and cancer radiotherapy that was published by CRC Press in 1994. When not busy with research and teaching mathematics he relishes hiking in the mountains of Idaho and reading.

Robert J. Fisher received his PhD in mathematics from the University of Massachusetts, Amherst in 1981; he is currently Professor and Chair of the Mathematics Department at Idaho State University in Pocatello, Idaho, where he has taught since 1989. His main research interest is in differential geometry in which he has published many articles; the most recent, entitled Generalized Immersions and the Rank of the Second Fundamental Form (with H. T. Laquer), appears in the June 2006 issue of the *Pacific Journal of Mathematics*. Outside of mathematics, he is a longtime distance runner. He is also an avid classical guitarist playing in the ISU guitar ensemble.

Boris Hanin is an undergraduate student at Stanford University majoring in Mathematics and Physics. In 2003, while still in high school, he asked his father, Leonid Hanin, for a math project to work on and was advised to think about the problem discussed on page 353 of this issue. His progress ignited an interest of the other two authors and led to far-reaching generalizations. For his project Boris received 2nd place in Mathematics at 2004 Intel International Science and Engineering Fair and in 2005 became a semifinalist of Siemens-Westinghouse Competition and Intel Science Talent Search.

Vol. 80, No. 5, December 2007

MATHEMATICS MAGAZINE

MATHEMATICS MAGAZINE (ISSN 0025-570X) is published by the Mathematical Association of America at 1529 Eighteenth Street, N.W., Washington, D.C. 20036 and Montpelier, VT, bimonthly except July/August.
The annual subscription price for *MATHEMATICS MAGAZINE* to an individual member of the Association is $131. Student and unemployed members receive a 66% dues discount; emeritus members receive a 50% discount; and new members receive a 20% dues discount for the first two years of membership.)

Subscription correspondence and notice of change of address should be sent to the Membership/ Subscriptions Department, Mathematical Association of America, 1529 Eighteenth Street, N.W., Washington, D.C. 20036. Microfilmed issues may be obtained from University Microfilms International, Serials Bid Coordinator, 300 North Zeeb Road, Ann Arbor, MI 48106.

Advertising correspondence should be addressed to

advertising@maa.org

Further advertising information can be found online at www.maa.org

Periodicals postage paid at Washington, D.C. and additional mailing offices.

Postmaster: Send address changes to Membership/ Subscriptions Department, Mathematical Association of America, 1529 Eighteenth Street, N.W., Washington, D.C. 20036-1385.

Printed in the United States of America

ARTICLES

Some First Tantalizing Steps into Semigroup Theory

CHRISTOPHER D. HOLLINGS
University of York
Heslington, York YO10 5DD, UK
cdh500@york.ac.uk

Semigroup theory is a thriving field in modern abstract algebra, though perhaps not a very well-known one. In this article, I will give a brief introduction to the theory of *algebraic* semigroups (there is a related topological theory which I won't touch on) and hopefully demonstrate that it has a flavor quite different from that of group theory.

The term *semigroup* was first coined in a French group theory textbook [**5**, p. 8] in 1904 (though with a more stringent definition than the modern one), before being introduced to the English-speaking mathematical world by Leonard Dickson the following year [**6**]. There then followed three decades, during which the only semigroup theory being done was that done in near-obscurity (at least from the Western perspective) by a Russian mathematician, Anton Kazimirovich Suschkewitsch, working in the Ukraine. Suschkewitsch was essentially doing semigroup theory before the rest of the world knew that there was such a thing, thus many of his results were rediscovered by later researchers who were unaware of his achievements.

The first tentative Western steps towards semigroup theory were taken during the 1930s (in, for example, [**1**] and [**2**]), and after the publication of a series of highly influential papers in the early 1940s ([**20**], [**3**], and [**7**]), the subject exploded. Semigroup theory developed rapidly in both East and West over the following decades to become the extremely fruitful area of research that it is today.

In the first few sections of this article, we will build up some basic semigroup theory, always with the group analogy at the back of our minds. Then, once we have established enough theory, we will break free of this restriction and see some truly "independent" semigroup theory. In the final section, we will consider the application of semigroups to the study of "partial symmetries."

In order that this article does not get too bogged down in detail, I have omitted most of the proofs of results. However, just in case you feel cheated by this, I have included references to J. M. Howie's book *Fundamentals of Semigroup Theory* [**12**], where the relevant proofs may be found. This textbook is also the one to consult if you find, after reading this article, that you want a more comprehensive introduction to semigroup theory!

In the following, I will assume some basic group theory, as well as a minimum of ring theory. It is also worth noting at this point that I will be following a convention often used in semigroup theory: that of writing a function on the right of its argument. For example, a function $\varphi : X \to Y$ will map $x \in X$ to $x\varphi \in Y$; we write $x\varphi$ for the value of the function φ at x, rather than the more usual $\varphi(x)$. We will also compose functions from left to right, so $x\varphi\psi$ will mean "do φ to x, then do ψ."

Definitions and Examples

The best place to start is, of course, the definition of a semigroup. Let S be a nonempty set and let $*$ be a binary operation on S. We say that the pair $(S, *)$ is a *semigroup* if the operation $*$ is *associative*: $a * (b * c) = (a * b) * c$, for all $a, b, c \in S$.

Whenever the operation in a semigroup is clear, we refer to "the semigroup S," rather than "the semigroup $(S, *)$." When discussing a general semigroup, we will denote the binary operation ("multiplication") by juxtaposition of elements; thus we will write ab rather than $a * b$.

Some semigroups contain an *identity* element, i.e., an element $1 \in S$ such that $1s = s = s1$, for all $s \in S$. Such semigroups are called *monoids*. It is easily shown that the identity of a monoid is unique. A monoid S in which every element $s \in S$ has a unique *inverse* $s^{-1} \in S$ (such that $ss^{-1} = 1 = s^{-1}s$) is clearly a *group*.

Semigroup examples are legion. From what we have just seen, every group is a semigroup; simply pick your favorite group (in my case, $\mathcal{S}_3$) and this will serve as an example of a semigroup. Some simple examples of semigroups which are not groups are the following:

1. The positive integers $\mathbb{N}$ form a semigroup under ordinary addition, whilst the nonnegative integers $\mathbb{N}^0$ form a monoid with identity 0.
2. The positive integers $\mathbb{N}$ also form a monoid under ordinary multiplication, this time with identity 1.
3. If $(R, +, \times)$ is a ring, then $(R, \times)$ is a semigroup. This is one of the historical origins of semigroups. Rather than taking the definition of a group and stripping it of identity and inverses, we can take a ring and strip it of an entire operation, its "addition." For example, in the thesis of Alfred Clifford, who went on to become one of the founders of modern semigroup theory, we find certain questions posed concerning rings. These questions were posed solely in terms of the ring's "multiplication," which led Clifford to comment that "it is natural to attempt a solution in the same terms" [**2**, p. 326]. This is also the principle behind some of Suschkewitsch's early papers (for example, [**25**]).

Further (more interesting!) examples of semigroups:

4. A *rectangular band*. This is the Cartesian product $I \times J$ of two nonempty sets I and J, together with the operation $(i, j)(k, l) = (i, l)$.
5. The *bicyclic semigroup*, $B = \mathbb{N}^0 \times \mathbb{N}^0$, with operation

$$(a, b)(c, d) = (a - b + \max\{b, c\}, d - c + \max\{b, c\}).$$

   This is a monoid with identity $(0, 0)$.
6. The *full transformation monoid* $\mathcal{T}_X$ on a set X. This is the monoid of all mappings of the set X to itself. The operation is composition of mappings. This is a very important semigroup because it is the semigroup analog of the symmetric group $\mathcal{S}_X$. For example, recall that Cayley's theorem tells us that every group can be embedded in some symmetric group; there is an analogous theorem for semigroups (originally proved by Suschkewitsch [**24**]) which tells us that every semigroup can be embedded in some full transformation monoid [**12**, Theorem 1.1.2.].
7. Let A be a nonempty set. A *word* (or *string*) over A is a finite sequence of elements of A; for $a_1, a_2, \ldots, a_n \in A$, we usually write the word $(a_1, a_2, \ldots, a_n)$ as $a_1a_2 \cdots a_n$. The *empty word*, denoted ε, is the word which contains no letters. Let A^+ denote the set of all nonempty words over A. We define a binary operation, called *concatenation*, on A^+ simply by writing one word after another and

demanding that

$$x(yz) = (xy)z,$$

for all $x, y, z \in A^+$. Under this operation, A^+ forms a semigroup—the *free semigroup on A*. If we put $A^* = A^+ \cup \{\varepsilon\}$ and make the further definition that

$$\varepsilon x = x = x\varepsilon,$$

for all $x \in A^*$, then A^* is the *free monoid on A*, with identity ε. Any semigroup S can be embedded in the free semigroup S^+ (where S is regarded simply as a set) via the mapping $\alpha : S \to S^+$ which sends any element $s \in S$ to the single-letter string $s = (s) \in S^+$. If S is a monoid, then we demand further that $1\alpha = \varepsilon$ and thereby embed S in S^*.

Unlike a group, a semigroup S can have a *zero element*, i.e., an element $z \in S$ such that $zs = z = sz$, for all $s \in S$. The zero of a semigroup is easily shown to be unique and, by analogy with the integers, it is usually denoted by the symbol 0. A semigroup (with zero) in which the product of any two elements is zero is called, unsurprisingly, a *zero semigroup*. For example:

8. The collection of matrices

$$\left\{\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}\right\}$$

forms a zero semigroup under matrix multiplication.

An element $a \in S$ is called a *right zero (element)* if $ba = a$, for all $b \in S$. A semigroup in which every element is a right zero is called a *right zero semigroup*. We will see an example of one of these shortly. Left zero elements and semigroups are defined in an analogous manner.

An *idempotent* in S is an element e such that $e^2 = ee = e$. A group G can only ever have one idempotent (its identity) but a semigroup can have many, indeed, there are semigroups which consist entirely of idempotents; these are called *bands*. As the name suggests, a rectangular band is an example of one of these. The subset of idempotents of a semigroup S is often denoted $E(S)$.

Observe that the semigroup $(\mathbb{N}, +)$ can be "embedded" in the group $(\mathbb{Z}, +)$ (i.e., there is a one-one homomorphism from $(\mathbb{N}, +)$ into $(\mathbb{Z}, +)$—see the next section). However, it is by no means the case that *every* semigroup arises as "half a group" in this way; the reasons for the name "semigroup" are purely historical. The first comprehensive discussion of whether a given semigroup is group embeddable was given by A. I. Mal'tsev in his 1939 paper [**16**], in which he proved, among other things, the highly nontrivial result that even a cancellative semigroup (that is, a semigroup in which $a = b$ if either $ac = bc$ or $ca = cb$) may fail to be group embeddable. Similarly, not every semigroup arises as the multiplicative semigroup of a ring (as in example 3), nor may every semigroup be embedded in such; if we recall that a ring of idempotents is necessarily commutative, then we see that a rectangular band (example 4) serves as a suitable counterexample.

Subsemigroups and Morphisms

With the definition of a semigroup established, it is natural to proceed to the definition of a *subsemigroup*. This is defined in just the way you would expect: a subset T of

a semigroup S is called a *subsemigroup* if it forms a semigroup under the same operation as S, i.e., if $a, b \in T$, then $ab \in T$. Further, if T contains an identity, then T is a *submonoid* of S. Note that in this case S need not be a monoid itself—it is only necessary for there to be an identity 1_T such that $1_T t = t = t1_T$ for all $t \in T$; 1_T need not be an identity for the whole of S. To go one step further, if every element of T has a unique inverse (with respect to 1_T), then T is called a *subgroup* of S.

9. A semigroup S is always a subsemigroup of itself. If $e \in E(S)$, then $\{e\}$ is a subgroup of S; in particular, if S has identity 1, then $\{1\}$ is a subgroup. If S has a zero 0, then $\{0\}$ is also a subgroup.
10. The set $2\mathbb{N}^0$ of all even nonnegative integers forms a submonoid of $(\mathbb{N}^0, +)$, with identity 0.
11. The subset $E(B) = \{(a, a) : a \in \mathbb{N}^0\}$ is a subsemigroup of the bicyclic semigroup.
12. Let X be a set. For each $a \in X$, define the mapping $c_a : X \to X$ by $xc_a = a$ for all $x \in X$—this is the *constant mapping* which sends everything in X to a. Each c_a is clearly an element of $\mathcal{T}_X$. We see also that $xc_ac_b = ac_b = b$ so $c_ac_b = c_b$, hence $R = \{c_a : a \in X\}$ is a subsemigroup of $\mathcal{T}_X$, but not a submonoid, since I_X, the identity mapping on X, does not belong to R. Note that R is a right zero semigroup.
13. The symmetric group $\mathcal{S}_X$ is a subgroup of $\mathcal{T}_X$.

Of particular interest are those semigroups whose idempotents form subsemigroups. One way in which this can be achieved is if the idempotents commute. To demonstrate, let S be a semigroup and let $e, f \in E(S)$. Then $(ef)^2 = efef = eeff = ef$. Thus, if $e, f \in E(S)$, then $ef \in E(S)$, so $E(S)$ is a subsemigroup of S. In this case, $E(S)$ is a commutative band, or *semilattice*.

Semigroup theory also has a concept of homomorphisms, or simply *morphisms*, as they are often termed. These are defined in exactly the same way as for groups: let S and T be semigroups; a function $\varphi : S \to T$ is called a *morphism* if $(s\varphi)(t\varphi) = (st)\varphi$, for all $s, t \in S$. If, in addition, φ is a bijection, we call it an *isomorphism*; the semigroups S and T are then said to be *isomorphic*, denoted $S \cong T$.

14. The mapping $\varphi : B \to \mathbb{Z}$, given by $(a, b)\varphi = a - b$ is a morphism, since $[(a, b)\varphi] + [(c, d)\varphi] = (a - b) + (c - d) = a - b + m - (d - c + m) = (a - b + m, d - c + m)\varphi = [(a, b)(c, d)]\varphi$, where $m = \max\{b, c\}$.
15. Let $I \times J$ be a rectangular band. Then $\alpha : I \times J \to \mathcal{T}_J$, given by $(i, j)\alpha = c_j$ is a morphism: $[(i, j)\alpha][(k, l)\alpha] = c_jc_l = c_l = (i, l)\alpha = [(i, j)(k, l)]\alpha$.

We will see shortly how semigroup morphisms give us a semigroup analog of the fundamental theorem of homomorphisms (a.k.a. the first isomorphism theorem) for groups, but first we need to remind ourselves of how *binary relations* work.

Interlude: Binary Relations

Formally, a binary relation between the elements of a set X is a subset ρ of $X \times X$. Thus $x \in X$ is "ρ-related" to $y \in X$ if, and only if, $(x, y) \in \rho$. In the interests of shortening the notation, we usually write $x \rho y$ to mean "x is ρ-related to y."

EXAMPLE. A familiar binary relation on $\mathbb{N}$ is the relation $\leq$, defined by the subset $\{(a, b) \in \mathbb{N} \times \mathbb{N} : a \text{ is less than or equal to } b\} \subseteq \mathbb{N} \times \mathbb{N}$. In this case, we can see that the "$a \rho b$" notation is better than the "$\rho \subseteq X \times X$" notation, for if we write $a \leq b$ in this latter notation, we end up with the inelegant $(a, b) \in \leq$.

Given a set X, there are two binary relations which we can always define. The *equality relation* ι is defined by $a \,\iota\, b \Leftrightarrow a = b$, whilst the *universal relation* ω is the relation via which all elements are related: $a \,\omega\, b$, for all $a, b \in X$. In the formal set notation of binary relations we can write $\omega = X \times X$.

One type of binary relation with which you are no doubt familiar is an *equivalence relation*, i.e., a binary relation $\sim$ between elements of a set X such that, for all x, y, $z \in X$,

(i) $x \sim x$ (reflexivity);

(ii) $x \sim y \Leftrightarrow y \sim x$ (symmetry);

(iii) $x \sim y, y \sim z \Rightarrow x \sim z$ (transitivity).

Associated with any equivalence relation are its *(equivalence) classes*; the equivalence class of $x \in X$ is the collection of all other elements of X which are $\sim$-related to x and is denoted by $[x]$, i.e., $[x] = \{y \in X : x \sim y\}$. The classes of an equivalence relation $\sim$ will often be referred to as "$\sim$-classes."

We are interested in defining binary relations on semigroups, rather than just sets. We would therefore like to be able to say something about the interaction between the relation and the semigroup's multiplication. Let ρ be a binary relation on a semigroup S. Then ρ is said to be *compatible* (with multiplication in S) if

$$a \,\rho\, b \text{ and } c \,\rho\, d \Rightarrow ac \,\rho\, bd.$$

A compatible equivalence relation is called a *congruence*. The equivalence classes of a congruence are, quite naturally, referred to as *congruence classes*.

EXAMPLE. Consider the relation δ on the bicyclic semigroup B, given by

$$(a, b) \,\delta\, (c, d) \Leftrightarrow a - b = c - d.$$

It is easily verified that δ is an equivalence relation. Further, suppose that $(a, b) \,\delta\, (c, d)$ and $(s, t) \,\delta\, (u, v)$, so $a - b = c - d$ and $s - t = u - v$. Then

$$a - b - (t - s) = c - d - (v - u).$$

We put $m = \max\{b, s\}$, $n = \max\{d, u\}$ and "add zero" to both sides to give

$$a - b + m - [t - s + m] = c - d + n - [v - u + n].$$

We therefore have

$$(a - b + m, s - t + m) \,\delta\, (c - d + n, u - v + n),$$

or $(a, b)(s, t) \,\delta\, (c, d)(u, v)$. Thus δ is a congruence on B.

The concept of a congruence on a semigroup enables us to construct *factor semigroups*. In group theory, we factor a group G by a normal subgroup N to obtain a factor group G/N; in ring theory, we factor a ring R by an ideal I to obtain a factor ring R/I. In semigroup theory, we factor a semigroup S by a congruence ρ to obtain a factor semigroup S/ρ. This factor semigroup is simply the collection of ρ-classes on S, $S/\rho = \{[s] : s \in S\}$, together with the operation $[s][t] = [st]$.

EXAMPLE. In the example of the congruence δ on B, the δ-classes are subsets of B in which all elements have the same difference in coordinates. Thus $B/\delta = \{[(a, b)] : (a, b) \in B\}$, where $[(a, b)] = \{(c, d) \in B : a - b = c - d\}$.

Binary relations play an important role in semigroup theory and crop up frequently. One example of their use is in *Green's relations*, which we will see shortly. Before that, however, we will consider the part which binary relations play in the *fundamental theorem of morphisms*.

The Fundamental Theorem of Morphisms

Let us first remind ourselves of the analogous theorem for groups:

THEOREM. *Let $\phi : G \to H$ be a homomorphism of groups G and H, and let* $\operatorname{Ker}\phi = \{g \in G : \phi(g) = 1_H\}$. *Then* $\operatorname{Ker}\phi$ *is a normal subgroup of* G, $\operatorname{im}\phi$ *is a subgroup of* H *and* $G/\operatorname{Ker}\phi$ *is isomorphic to* $\operatorname{im}\phi$.

In the last section, we developed semigroup analogs of both normal subgroups (congruences) and factor groups (factor semigroups). If we are to obtain a semigroup analog of the above theorem, we need the concept of the kernel of a morphism (the image of a semigroup morphism has the same definition as for a group homomorphism). Clearly, we cannot use the same definition as for groups, since we do not necessarily have an identity at our disposal. We will use "ker" to denote the kernel of a semigroup morphism in order to distinguish it from the kernel "Ker" of a group homomorphism.

Let $\varphi : S \to T$ be a morphism of semigroups S and T. We define the kernel, $\ker\varphi$, of φ not as a subset of S, but as a binary relation on S:

$$s\,(\ker\varphi)\,t \Leftrightarrow s\varphi = t\varphi,$$

for $s, t \in S$. Thus s and t are $(\ker\varphi)$-related if φ maps them to the same element of T. The relation $\ker\varphi$ is clearly an equivalence relation. Let $s, t, u, v \in S$ and observe further that if $s\,(\ker\varphi)\,t$ and $u\,(\ker\varphi)\,v$ (i.e., $s\varphi = t\varphi$ and $u\varphi = v\varphi$) then $(s\varphi)(u\varphi) = (t\varphi)(v\varphi)$, whence $(su)\varphi = (tv)\varphi$, since φ is a morphism. Thus $\ker\varphi$ is a congruence. Indeed, every congruence on a semigroup S can be realized in this way—as the kernel of some morphism with domain S. Specifically, a congruence ρ on S is the kernel of the morphism $\nu_\rho : S \to S/\rho$, given by $s\nu_\rho = [s]$. More importantly for our present interests, we can now factor by $\ker\varphi$, just as we could factor by $\operatorname{Ker}\phi$ in the group case.

The definition of the kernel as a congruence is still perfectly valid in a group theory context. Note that if $\phi : G \to H$ is a group homomorphism, then $\operatorname{Ker}\phi$ is in fact the $(\ker\phi)$-class of 1_G. Every normal subgroup is the Kernel of some homomorphism, to which we can associate a congruence, so whenever you factor a group by a normal subgroup, you are factoring by a congruence without realizing it! It is easy to show that the cosets of $\operatorname{Ker}\phi$ by elements of G are precisely the $(\ker\phi)$-classes in G.

We can finally state the desired theorem:

FUNDAMENTAL THEOREM OF MORPHISMS. [**12**, Theorem 1.5.2.] *Let $\varphi : S \to T$ be a morphism of semigroups S and T. Then* $\ker\varphi$ *is a congruence on* S, $\operatorname{im}\varphi$ *is a subsemigroup of* T *and* $S/\ker\varphi$ *is isomorphic to* $\operatorname{im}\varphi$.

EXAMPLE. Our congruence δ is clearly the kernel of the morphism $\varphi : B \to \mathbb{Z}$ of example 14. What is $\operatorname{im}\varphi$? Let $z \in \mathbb{Z}$ and observe that if $z \geq 0$, then $z = (z, 0)\varphi$, and if $z < 0$, then $z = (0, -z)\varphi$. Thus φ is onto, hence $\operatorname{im}\varphi = \mathbb{Z}$. By the fundamental theorem of morphisms, we have $B/\delta \cong \mathbb{Z}$ (cf. our previous expression for B/δ).

Ideals and Green's Relations

The aim of this section is to introduce the study of the structure of semigroups via a number of special equivalence relations: *Green's relations*. These are defined in terms a special type of subsemigroup: an *ideal*. To define an ideal in a semigroup, we simply take the definition of an ideal in a ring and delete all reference to addition. Thus a subsemigroup I of a semigroup S forms a *right ideal* if, whenever $i \in I$ and $s \in S$, we have $is \in I$. Similarly, I is a *left ideal* if $si \in I$. If, for $i \in I$ and $s, t \in S$, we have $sit \in I$, then I is called a *two-sided ideal* or, simply, an *ideal*.

The particular types of ideals which we need are so-called *principal ideals*, for which we need another new concept: a *semigroup with identity adjoined*. It is often an advantage to work with a monoid rather than a semigroup, so if the semigroup in question (let us call it S) has no identity, then we simply adjoin an extra symbol 1 to S and define 1 to behave like an identity. The semigroup with identity adjoined, denoted S^1, is then equal to S if S already has an identity, or $S \cup \{1\}$ if S does not have an identity. Principal ideals of S are defined in terms of S^1: given an element $a \in S$, the *principal right ideal generated by a* is simply the ideal $aS^1 = \{as : s \in S^1\}$. Similarly, $S^1a = \{sa : a \in S^1\}$ is the *principal left ideal generated by a* and $S^1aS^1 = \{sat : s, t \in S^1\}$ is the *principal (two-sided) ideal generated by a*. The reason for defining these in terms of S^1 rather than S is so that a belongs to the principal right/left/two-sided ideal which it generates.

16. The only ideal of a group G is itself.
17. Every ideal of $(\mathbb{N}, +)$ is principal and has the form $I_n = \{n, n+1, n+2, \ldots\}$, for each $n \in \mathbb{N}^0$.
18. If a semigroup S has a zero 0, then $\{0\}$ is an ideal of S.
19. Let $I \times J$ be a rectangular band. Then, for any $i \in I$, $\{i\} \times J$ is a right ideal of $I \times J$. Similarly, $I \times \{j\}$ is a left ideal, for each $j \in J$.
20. The subset $\{(x, y) : x \geq m, y \in \mathbb{N}^0\}$ is a right ideal of B for each fixed $m \in \mathbb{N}^0$. The bicyclic semigroup has no proper principal ideals.

We are now in a position to define Green's relations. These relations, introduced explicitly in a paper by J. A. Green in 1951 [**9**] (though used implicitly before then), allow us to study the "large scale" structure of a semigroup via its principal ideals. The fundamental importance of Green's relations to the study of semigroups has led Howie to comment:

> ... on encountering a new semigroup, almost the first question one asks is 'What are the Green relations like?' [**13**, p. 9]

The two most basic of Green's relations are $\mathcal{R}$ and $\mathcal{L}$. The first of these is defined as follows, for $a, b \in S$:

$$a \mathcal{R} b \Leftrightarrow aS^1 = bS^1.$$

Thus a and b are $\mathcal{R}$-related if they generate the same principal right ideal. Similarly, a and b are $\mathcal{L}$-related if they generate the same principal left ideal: $S^1a = S^1b$.

As you might expect, there is also a relation $\mathcal{J}$ which is defined in terms of principal two-sided ideals:

$$a \mathcal{J} b \Leftrightarrow S^1aS^1 = S^1bS^1.$$

The relation $\mathcal{H}$ is defined in terms of $\mathcal{R}$ and $\mathcal{L}$: $\mathcal{H} = \mathcal{R} \cap \mathcal{L}$, so a and b are $\mathcal{H}$-related if they are both $\mathcal{R}$-related *and* $\mathcal{L}$-related.

21. In a group G, $\mathcal{L} = \mathcal{R} = \mathcal{H} = \mathcal{J} = \omega$.
22. In a commutative semigroup, $\mathcal{R} = \mathcal{L} = \mathcal{H}$.
23. In $(\mathbb{N}, +)$, $\mathcal{L} = \mathcal{R} = \mathcal{H} = \iota$.
24. In B, $(u, v)\,\mathcal{R}\,(p, q) \Leftrightarrow u = p$ and $(u, v)\,\mathcal{L}\,(p, q) \Leftrightarrow v = q$, so $\mathcal{H} = \iota$.

We see then from example 21 that Green's relations are semigroup theory tools which have no real use in group theory. We now consider some further "independent" semigroup theory.

Regular and Inverse Semigroups

In this penultimate section, we will look at two much-studied classes of semigroups and establish the connection between them. Hopefully, in this section you will detect more of the distinct flavor of semigroup theory. We will see more details of proofs in this section.

A *regular semigroup* is remarkably easy to define, yet the study of such semigroups has spawned much subsequent semigroup theory (see, for example, Chapter 4 of Howie [**12**]). The concept of a regular semigroup was first introduced by J. A. Green [**9**] and is defined thus: a semigroup S is *regular* if, for each $a \in S$, there is an $x \in S$ such that $a = axa$. (The word "regular" is the normal candidate for the title of "most overused word in mathematics"!)

One small point to note about regular semigroups is the following: for $a \in S$, we have $a = axa \in aS$, and similarly, $a \in Sa$, $a \in SaS$. Thus a is shown to belong to the principal ideals which it generates, without recourse to adjoining an identity. For regular semigroups, Green's relations can be expressed in terms of S, rather than S^1.

25. A rectangular band $I \times J$ is regular, since $(i, j) = (i, j)(k, l)(i, j)$, for all $(i, j), (k, l) \in I \times J$. Indeed, *every* band E is regular, since $e = e^2 = e^3 = eee$, for all $e \in E$.
26. The bicyclic semigroup B is regular, since $(a, b) = (a, b)(b, a)(a, b)$.
27. The full transformation semigroup $\mathcal{T}_X$ is regular.
28. The semigroup $(\mathbb{N}, +)$ is *not* regular, since there is no solution (in $\mathbb{N}$) to $1 + x + 1 = 1$.

When dealing with semigroups in general, we often discuss *inverses*. However, since we do not necessarily have an identity, we cannot define these inverses in the traditional (group) way. We need a more general concept of inverse, which we define as follows: the element s of a semigroup S has inverse $s' \in S$ if

$$ss's = s \qquad \text{and} \qquad s'ss' = s'.$$

The definition is symmetric in that we also say that s is an inverse for s'. A group inverse clearly satisfies these properties, so every group inverse is a semigroup inverse, but not conversely. Note that in the case when we *do* have an identity 1, it is still not necessarily true that $ss' = 1$; all we can say is that ss' is idempotent, since $(ss')^2 = (ss')(ss') = (ss's)s' = ss'$. Similarly, $s's$ is idempotent. Note that, in general, $ss' \neq s's$.

In a given semigroup S, an element s need not necessarily have a generalized inverse, or, if it does, it could have more than one. The most extreme situation can be seen in example 25, which tells us that every element of a rectangular band is an inverse for every other element! Given a semigroup S, we define the subset $V(a) \subseteq S$

by

$$V(a) = \{a' : a' \text{ is an inverse for } a\}.$$

In a regular semigroup S, every element has at least one inverse ($V(a) \neq \emptyset$). To see this, let us take an element $a \in S$. By definition, there is an $x \in S$ such that $a = axa$. In this case it is not x, but xax which is an inverse for a, since we have: $a = axa = ax(axa) = a(xax)a$ and $xax = x(axa)x = xax(axa)x = (xax)a(xax)$.

A semigroup S in which every element s has precisely one generalized inverse (denoted by s^{-1}) is called an *inverse semigroup*. To put this another way, $|V(a)| = 1$ for inverse semigroups.

29. Every group is an inverse semigroup.
30. The bicyclic semigroup is an inverse semigroup, with $(a, b)^{-1} = (b, a)$.
31. A rectangular band $I \times J$ is an inverse semigroup only if it is trivial, i.e., $|I| = |J| = 1$.
32. Every semilattice is an inverse semigroup.
33. The full transformation monoid $\mathcal{T}_X$ is *not* an inverse semigroup.

Inverse semigroups are an example of the Cold War duplication of results which was a consequence of the relative lack of mathematical communication across the iron curtain: they were introduced independently by Victor Vladimirovich Wagner [**26**] (who called them *generalized groups*) in the Soviet Union in 1952, and by Gordon Preston [**19**] in Britain in 1954.

There are two properties of inverse semigroups which interest us in particular. First of all, an inverse semigroup is regular. This is easy to see: given an element a, its inverse a^{-1} will play the role of x in the condition for being regular. The second property is that idempotents commute. With a careful argument, we can prove this. Suppose that S is an inverse semigroup and that $e, f \in E(S)$. Let $x \in S$ be an inverse for the product ef. We observe that the element fxe is idempotent ($(fxe)^2 = f[x(ef)x]e = fxe$) and also an inverse for ef:

$$ef(fxe)ef = (ef)x(ef) = ef \quad \text{and} \quad (fxe)ef(fxe) = (fxe)^2 = fxe.$$

By uniqueness of inverses, we have $x = fxe$, so x must be idempotent, in which case it is self-inverse. Since x is the inverse of ef, ef is the inverse of x. Thus $x = ef$ and ef is idempotent. We have shown that $E(S)$ is closed under multiplication. It only remains to show that idempotents do indeed commute. By closure of $E(S)$, we know that fe is idempotent. Notice that fe is an inverse for ef:

$$ef(fe)ef = (ef)(ef) = ef \quad \text{and} \quad fe(ef)fe = (fe)(fe) = fe.$$

Thus ef and fe are both inverses for ef. By uniqueness of inverses, $ef = fe$.

We have shown that an inverse semigroup is a regular semigroup whose idempotents form a semilattice. This begs the question: is the converse true? Given a regular semigroup S in which idempotents commute, is S necessarily an inverse semigroup? We know that every element of a regular semigroup has at least one inverse. We need to show that if idempotents commute, then that inverse is unique. We suppose that an element $a \in S$ has two inverses, b and c. Then

$$a = aba, \quad b = bab \qquad \text{and} \qquad a = aca, \quad c = cac,$$

by definition. Note that ab, ba, ac and ca are all idempotent. Thus

$$b = bab = bacab = bacacab \overset{(i)}{=} bacabac \overset{(ii)}{=} cababac = cabac = cac = c,$$

where at step (i), we have commuted the idempotents ab and ac, whilst at step (ii), we have commuted the idempotents ba and ca. We have therefore shown that inverses are unique if idempotents commute.

THEOREM. [**12**, Theorem 5.1.1.] *Let S be a semigroup. Then the following are equivalent:*

(i) *S is an inverse semigroup;*

(ii) *S is regular, with commuting idempotents.*

This theorem gives us immediate justification of example 32; every band is regular, hence every semilattice is an inverse semigroup. To demonstrate the result further, let us use our old friend, the bicyclic semigroup B. We have stated (example 30) that B is an inverse semigroup with $(a, b)^{-1} = (b, a)$ but we have not actually proved this; the proof that (b, a) is the *unique* inverse of (a, b) is somewhat tricky. However, we can now use the above theorem to make our work easier. We know that B is regular; in order to prove that B is an inverse semigroup, it only remains to show that its idempotents commute. We have already seen the idempotents of B in example 11; any idempotent of B has the form (a, a) and elements of this form commute:

$$(a, a)(b, b) = (a - a + m, b - b + m) = (b - b + m, a - a + m) = (b, b)(a, a),$$

where $m = \max\{a, b\}$. Thus B is an inverse semigroup.

As is clearly the case with the bicyclic semigroup, it is often much easier to show that a given semigroup is regular and that its idempotents commute, rather than proving that it is an inverse semigroup directly. Likewise, when showing that a semigroup is *not* an inverse semigroup. We can use our theorem to justify our earlier statement that $\mathcal{T}_X$ is not an inverse semigroup (example 33). Among the idempotents of $\mathcal{T}_X$ are the constant maps (see example 12). However, these do not commute: $c_x c_y = c_y \neq c_x = c_y c_x$. Hence $\mathcal{T}_X$ cannot be an inverse semigroup.

Application: Partial Symmetries

Semigroups have a number of applications, many of which lay in the realms of computer science. For example, semigroups are inextricably linked with finite state automata, and are used to classify formal languages—see [**10**]. However, since we have been considering inverse semigroups, it makes sense to look at an application of these. We will therefore consider the application of inverse semigroups to the study of *partial symmetries*.

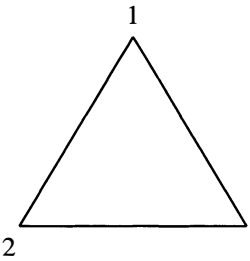


Figure 1 Equilateral triangle

Let us begin by considering the equilateral triangle (Figure 1). We are familiar with the "traditional" notion of the *symmetries* of a polygon and the fact that these symmetries form a group under composition of mappings. In the particular case of the

equilateral triangle, the symmetry group is, of course, $\mathcal{S}_3$, consisting of the identity map I_3, the rotations (123) and (132), and the reflections (12), (13) and (23). This is also known as the dihedral group, D_3 (or D_6, depending on your convention!).

Observe now that the equilateral triangle has further "partial" symmetries; specifically, the sides 12, 23 and 13 all "look like" each other. In order to describe such partial symmetries, we need the concept of a "partial mapping." Let X be an arbitrary set. We define a *partial mapping of* X to be a function $f : A \to B$, where A and B are subsets of X. We compose partial mappings f and g on X (from left to right) on the largest domain upon which it makes sense to do so, namely,

$$\operatorname{dom} fg = (\operatorname{im} f \cap \operatorname{dom} g)\, f^{-1}, \tag{1}$$

where f^{-1} denotes the preimage under f. For $x \in \operatorname{dom} fg$, we put $x(fg) = (xf)g$. Observe that $\operatorname{dom} fg$ is the subset of $\operatorname{dom} f$ consisting of all those elements x for which $xf \in \operatorname{dom} g$. If $\operatorname{dom} fg = \emptyset$, then we say that fg is the *empty transformation* on X, which we denote by ϵ.

Of particular interest are the partial *one-one* maps of a set X; any such map is clearly a bijection from its domain to its image, so we refer to the partial one-one mappings of X as its *partial bijections*. By analogy with the definition of the symmetries of a set as its bijective self-maps, we define the *partial symmetries* of X to be its partial bijections. The collection of all partial bijections of X is denoted by $\mathcal{I}_X$. Under the composition given in (1), $\mathcal{I}_X$ forms a monoid with identity I_X and zero ϵ. Notice that by virtue of the fact that every set is a subset of itself, we have $\mathcal{S}_X \subseteq \mathcal{I}_X$. Moreover, since $\operatorname{dom} f = X$, for all $f \in \mathcal{S}_X$, the partial composition (1) reduces to the ordinary composition of functions in $\mathcal{S}_X$, so $\mathcal{S}_X$ is in fact a subgroup of $\mathcal{I}_X$: every symmetry is a partial symmetry.

In the case of the equilateral triangle, we of course have $X = \{1, 2, 3\}$. We can therefore describe the partial symmetry between the edges 12 and 23, for example, by means of the following partial bijection, which we write using a modification of the "two-row" notation often used for permutations:

$$\alpha = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & \times \end{pmatrix},$$

i.e., $1\alpha = 2$ and $2\alpha = 3$ but 3α is undefined, as indicated by the $\times$. Alternatively, this partial symmetry could be described by the partial bijection

$$\alpha' = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & \times \end{pmatrix}.$$

Similarly, partial bijections can be found which send 12 to 13, 23 to 13, and so on.

Let us look at $\alpha \in \mathcal{I}_{\{1,2,3\}}$ in more detail. If we consider α as a bijection $\{1, 2\} \to \{2, 3\}$, then it is clear that α is invertible on $\{2, 3\}$, with inverse

$$\alpha^{-1} = \begin{pmatrix} 1 & 2 & 3 \\ \times & 1 & 2 \end{pmatrix}.$$

Observe, however, that if we compose α with α^{-1} in the sense of (1), then we obtain

$$\alpha\alpha^{-1} = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & \times \end{pmatrix}.$$

Thus $\alpha\alpha^{-1}$ is not equal to I_3, but to $I_{\{1,2\}}$, the restriction of I_3 to $\operatorname{dom}\alpha = \{1, 2\}$. Similarly, $\alpha^{-1}\alpha = I_{\operatorname{dom}\alpha^{-1}} = I_{\{2,3\}}$. We see then that α^{-1} is not an inverse for α in the usual (group) sense. However, given the content of the previous section, it should come as no surprise that we instead have $\alpha\alpha^{-1}\alpha = \alpha$ and $\alpha^{-1}\alpha\alpha^{-1} = \alpha^{-1}$. Hence α^{-1} is a *generalized* inverse for α. This argument is easily extended to show that every element of $\mathcal{I}_{\{1,2,3\}}$ has a (generalized) inverse. We next consider the idempotents of $\mathcal{I}_{\{1,2,3\}}$; these are the partial identity maps I_Z (for $Z \subseteq \{1, 2, 3\}$). It is very easy to see that these commute:

$$I_A I_B = I_{A\cap B} = I_{B\cap A} = I_B I_A,$$

where $A, B \subseteq \{1, 2, 3\}$. Therefore, by the theorem of the previous section, $\mathcal{I}_{\{1,2,3\}}$ forms an inverse monoid. More generally, we have the following result:

THEOREM. [**12**, Theorem 5.1.5.] *For any set X, $\mathcal{I}_X$ is an inverse monoid—the* symmetric inverse monoid *on X.*

We see then that inverse semigroups are an invaluable tool in the study of partial bijections and therefore of partial symmetries. Indeed, it was an investigation of the partial bijections of a set which led to the initial definition of an inverse semigroup in [**26**].

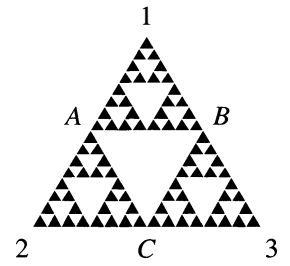


Figure 2 Sierpiński gasket

The equilateral triangle is a comparatively simple example to give; its partial symmetries are not too difficult to come to grips with. Other figures, however, have far more complicated (and more interesting) partial symmetries. Let us conclude this article by briefly considering the *Sierpiński gasket*. This is the fractal formed from a shaded equilateral triangle by repeatedly deleting the middle quarter of successively smaller triangles, ad infinitum (Figure 2). (For a discussion of the ubiquity of the Sierpiński gasket, see [**23**].)

The Sierpiński gasket displays some very obvious partial symmetries: not only those which we have already observed for the equilateral triangle, but also translational symmetries (between the region bounded by the triangle $1AB$ and that bounded by $A2C$, for example) and symmetries between parts and the whole (the rescaling which maps the region $1AB$ onto the whole gasket, for instance). Similar such partial symmetries can, of course, be found at every level of the gasket's construction.

We see then that when it comes to describing the symmetries of the Sierpiński gasket, group theory is unfortunately unequal to the task. The symmetry group of the gasket is simply that of the equilateral triangle, $\mathcal{S}_3$, but this does not (indeed, cannot) capture the complete structure of the figure. If we are to do the gasket justice, then we must step over into the realms of semigroup theory and consider a symmetry *monoid*—just one example of the utility of this most elegant of theories!

Conclusion

In an article such as this, it is only possible to scratch the surface of semigroup theory. My intention was to give you the basic definitions of the subject, as well as to demonstrate some of the techniques. I hope I have convinced you that it is field worthy of study, and that it is does not simply run parallel to group theory!

Further Reading

Mulcrone [**18**] is an excellent source of further examples. Those interested in learning more about semigroup theory can't go wrong with the book which is cited throughout the text: J. M. Howie's *Fundamentals of Semigroup Theory* [**12**]. An older textbook which is also worth consulting is Clifford & Preston's *Algebraic Theory of Semigroups* [**4**]. If you want to know why inverse semigroups are so exciting and learn more about partial symmetries, then Mark Lawson's *Inverse Semigroups* [**15**] is the book for you; the partial symmetries of the Cantor set are described explicitly in [**15**, §9.3]. Biographies are avaliable for many of the semigroup theorists mentioned in this article: Anton Kazimirovich Suschkewitsch (1889–1962) [**8**], Alfred Clifford (1908–1992) [**17**], Victor Vladimirovich Wagner [**21**] (1908–1981) and Gordon Preston (1925–) [**11**]. Some details on the early history of semigroup theory can be found in [**14**]; the origins of inverse semigroups are explored in [**22**].

Acknowledgment. Whilst compiling the examples given in this article, I drew heavily on lecture notes taken during the course on semigroup theory given by Dr. V. Gould at the University of York, UK, during the Autumn term of 2003. My thanks to Elizabeth Miller for useful comments and spotting of typos, to Adam McNaney for helpful discussions and also to the anonymous referees for many valuable suggestions.

REFERENCES

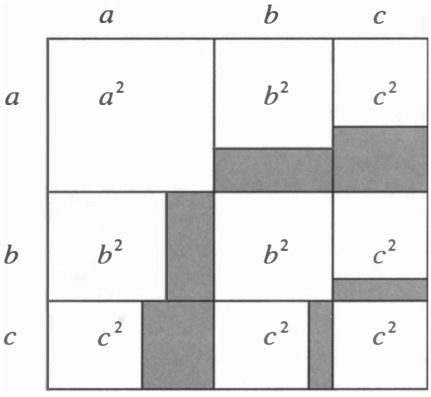
1. A. H. Clifford, A system arising from a weakened set of group postulates, *Ann. Math. (2)* **34** (1933) 865–871.
2. A. H. Clifford, Arithmetic and ideal theory of abstract multiplication, *Bull. Amer. Math. Soc.* **40** (1934) 326–330.
3. A. H. Clifford, Semigroups admitting relative inverses, *Ann. Math. (2)* **42** (1941) 1037–1049.
4. A. H. Clifford and G. B. Preston, *The Algebraic Theory of Semigroups, Volume 1, Mathematical Surveys of the American Mathematical Society, No. 7*, Providence, R. I., 1961.
5. J. A. de Séguier, *Théorie des Groupes Finis: Élements de la Théorie des Groupes Abstraits*, Paris, Gauthier-Villars, 1904.
6. L. E. Dickson, On semi-groups and the general isomorphism between infinite groups, *Trans. Amer. Math. Soc.* **6** (1905) 205–208.
7. P. Dubreil, Contribution à la théorie des demi-groupes, *Mèm. Acad. Sci. Inst. France (2)* **63** (1941) 1–52.
8. L. M. Gluskin and E. S. Lyapin, Anton Kazimirovič Suškevič, on his seventieth birthday, *Uspehi Mat. Nauk* **14** (1959) 255–260 (Russian).
9. J. A. Green, On the structure of semigroups, *Ann. Math. (2)* **54** (1951) 163–172.
10. J. M. Howie, *Automata and Languages*, Clarendon Press, Oxford, 1991.
11. J. M. Howie, Gordon Bamford Preston, *Semigroup Forum* **51** (1995) 270–271.
12. J. M. Howie, *Fundamentals of Semigroup Theory*, Clarendon Press, Oxford, 1995.
13. J. M. Howie, Semigroups, past, present and future, *Proceedings of the International Conference on Algebra and Its Applications 2002*, 6–20.
14. U. Knauer, Zur Entwicklung der algebraischen Theorie der Halbgruppen, *Simon Stevin* **54** (1980), 165–177.
15. M. V. Lawson, *Inverse Semigroups: The Theory of Partial Symmetries*, World Scientific, 1998.
16. A. I. Mal'tsev, On the immersion of associative systems in groups, *Mat. Sb. (N.S.)* **6** (1939) 331–336 (Russian).
17. D. D. Miller, A. H. Clifford: the first sixty-five years, *Semigroup Forum* **7** (1974) 4–9.
18. T. F. Mulcrone, Semigroup examples in introductory modern algebra, *Amer. Math. Monthly* **69** (1962) 296–301.
19. G. B. Preston, Inverse semi-groups, *J. London Math. Soc.* **29** (1954) 396–403.

20. D. Rees, On semi-groups, *Proc. Cam. Phil. Soc.* **36** (1940) 387–400.
21. B. M. Schein, Obituary: V. V. Vagner (1908–1981), *Semigroup Forum* **23** (1981) 189–200.
22. B. M. Schein, Prehistory of the theory of inverse semigroups, *Proceedings of the 1986 LSU Semigroup Conference (Kochfest 60)*, Louisiana State University, Baton Rouge, LA, 1986, 72–76.
23. I. Stewart, Four encounters with Sierpiński's gasket, *Math. Intell.* **17**(1) (1995) 52–64.
24. A. K. Suschkewitsch, Über die Darstellung der eindeutig nicht umkehrbaren Gruppen mittels der verallgemeinerten Substitutionen, *Mat. Sb.* **33** (1926) 371–374.
25. A. K. Suschkewitsch, Über die endlichen Gruppen ohne das Gesetz der eindeutigen Umkehrbarkeit, *Math. Ann.* **99** (1928) 30–50.
26. V. V. Wagner, Generalized groups, *Dokl. Akad. Nauk SSSR* **84** (1952) 1119–1122 (Russian).

Proof Without Words: An Algebraic Inequality

Problem 12 of the Leningrad Mathematics Olympiad, Grade 7, second round, 1989.

Let $a \geq b \geq c \geq 0$, and let $a + b + c \leq 1$. Prove $a^2 + 3b^2 + 5c^2 \leq 1$.



$$a^2 + 3b^2 + 5c^2 \leq (a + b + c)^2 \leq 1.$$

Similarly, if $a_1 \geq a_2 \geq \cdots \geq a_n \geq 0$, and $\sum_{i=1}^{n} a_i \leq 1$, then

$$\sum_{i=1}^{n} (2i - 1)a_i^2 \leq 1.$$

With $n = 2004$, this is problem 2 of the 2004 AMOC Senior Contest [1].

REFERENCE

1. Hans Lausch, A n Olympiad problem appeal, *The Australian Mathematical Society Gazette* **34**(2), 2007, 90–91.

Wei-Dong Jiang
Department of Information Engineering
Weihai Vocational College
Weihai 264200, Shangdon Province
P. R. China
jackjwd@163.com

Four Proofs of the Ballot Theorem

MARC RENAULT
Shippensburg University
Shippensburg, PA 17257
MSRenault@ship.edu

Introduction

One of the great pleasures in mathematics occurs when one considers several different proofs of a single result. In fact, when one considers the myriad proofs of the Pythagorean theorem and the irrationality of $\sqrt{2}$ constructed over the centuries, it seems we humans can never be satisfied with just one proof. Why do we continue to devise new approaches to known results? There is something in the reasoning itself that brings insight to the problem beyond what the result tells us, like looking at a sculpture from many different perspectives to appreciate it as fully as possible.

In this article we present four proofs of the ballot theorem, describe some of the history surrounding each of the proofs, and consider the different perspectives that each brings to the problem.

THE BALLOT PROBLEM. Suppose that in an election, candidate A receives a votes and candidate B receives b votes, where $a \geq kb$ for some positive integer k. Compute the number of ways the ballots can be ordered so that A maintains more than k times as many votes as B throughout the counting of the ballots.

THE BALLOT THEOREM. *The solution to the ballot problem is* $\dfrac{a-kb}{a+b}\binom{a+b}{a}$.

Let us call a permutation of the ballots *good* if A stays ahead of B by more than a factor of k throughout the counting of the ballots, and *bad* otherwise. Since the total number of distinct permutations of the $a+b$ ballots is

$$\binom{a+b}{a} = \frac{(a+b)!}{a!\,b!},$$

the theorem tells us that if all ballot permutations are equally likely, then the probability of a good permutation occurring is $(a-kb)/(a+b)$.

In 1887 Joseph Bertrand [**8**] introduced the ballot problem for the case $k=1$, gave its solution, outlined an inductive proof, and asked if a "direct solution" could be found. Almost immediately after Bertrand posed his question, Émile Barbier [**5**] stated and provided a solution to the ballot problem for arbitrary k, but without any proof. Very shortly after Barbier, Désiré André [**4**] produced a short combinatorial proof of the ballot theorem for $k=1$. In 1923 Aeppli [**2**] announced that he had the first proof of the ballot theorem for $k \geq 1$, and he directed interested readers to see his Ph.D. thesis [**3**, pp. 11–15] for the proof. Takács [**30**] supplies a nice account of the historical development of various ballot theorems, and gives several proofs of the ballot theorem, including the original proofs by André and Aeppli.

Proof 1: Count the Bad Ballot Permutations

André's approach for the case $k=1$ is to count the number of bad ballot permutations and subtract that from the number of all ballot permutations to obtain the number of

good ballot permutations. Briefly, André supposes that a ballots are marked "A" and b ballots are marked "B". He first notes that every ballot permutation starting with B is bad, and there are $\binom{a+b-1}{a}$ of these. Through a reversible procedure, he demonstrates a one-to-one correspondence between the bad ballot permutations starting with A, and all permutations consisting of a A's and $b-1$ B's. Again, these number $\binom{a+b-1}{a}$. He concludes that the number of bad ballot permutations is $2\binom{a+b-1}{a}$, and the ballot theorem then follows by simplifying $\binom{a+b}{a} - 2\binom{a+b-1}{a} = \frac{a-b}{a+b}\binom{a+b}{a}$.

The ballot problem and its solution caught the imagination of mathematicians, and many variations of André's proof have appeared throughout the years. For instance, Percy MacMahon [**18**] applied his deep theory of partitions to the problem. The most famous and elegant of these variations is the "reflection method" (often misattributed to André) in which ballot permutations are represented as lattice paths and portions of the bad paths are reflected across a line. This method was developed in the pair of papers [**1**] and [**19**] in 1923. Interestingly, the reflection method fails to generalize in a way that solves the ballot problem for $k > 1$. See [**23**] for more detail on André's original proof, the reflection method, and extending André's original proof to the case $k \geq 1$.

In 2003 Goulden and Serrano [**14**] provided a clever proof of the ballot theorem (for $k \geq 1$) using André's "count the bad ballot permutations" approach [**14**], and we present a variation of that proof here. Their proof *rotates* a portion of a lattice path instead of reflecting it.

Proof 1. We can think of a ballot permutation as a lattice path starting at $(0, 0)$ where votes for A are expressed as upsteps $(1, 1)$ and votes for B are expressed as downsteps $(1, -k)$. We seek the number of such paths with a upsteps and b downsteps where no step ends on or below the x-axis. Paths that remain above the x-axis (after the origin) are *good*, while those with steps that end on or below the x-axis are *bad*. A downstep that starts above the x-axis and ends on or below it is called a *bad step*.

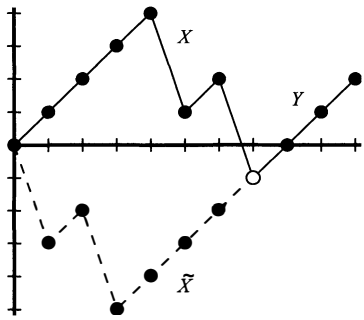


Figure 1 Example with $k = 3$. $XY \in \mathcal{B}_1$ and $\tilde{X}Y \in \mathcal{B}_3$.

For $0 \leq i \leq k$, let $\mathcal{B}_i$ denote the set of bad paths whose first bad step ends i units below the x-axis. Clearly these $k+1$ sets are disjoint and their union is the set of *all* bad paths. Notice that the paths in $\mathcal{B}_k$ are exactly those paths that start with a downstep, and so $|\mathcal{B}_k| = \binom{a+b-1}{a}$. We now show that for any $i \neq k$ we actually have $|\mathcal{B}_i| = |\mathcal{B}_k|$.

Let P be a path in $\mathcal{B}_i$ $(i \neq k)$, and identify the first step of P that ends i units below the x-axis. Let X be the initial segment of P that ends with that step and write $P = XY$. Let $\tilde{X}$ denote the path that results from rotating X by 180°, exchanging its endpoints; see Figure 1. Since X ends with a downstep, $\tilde{X}$ starts with a downstep, and consequently $\tilde{X}Y \in \mathcal{B}_k$.

The same process converts a path in $\mathcal{B}_k$ into a path in $\mathcal{B}_i$ $(i \neq k)$. If $P \in \mathcal{B}_k$, then identify the first step that ends i units below the x-axis. Let X denote the initial segment of P that ends with that step and write $P = XY$. Since X necessarily ends with an upstep, we have $\tilde{X}Y \in \mathcal{B}_i$.

Thus each of the $k+1$ sets $\mathcal{B}_i$ have cardinality $\binom{a+b-1}{a}$, and the number of good paths is

$$\binom{a+b}{a} - (k+1)\binom{a+b-1}{a} = \frac{a-kb}{a+b}\binom{a+b}{a}. \qquad \blacksquare$$

Of particular interest in this proof is the fact that the sets $\mathcal{B}_i$ of bad ballot permutations all have the same cardinality, regardless of i. We say that these sets *uniformly partition* the set of bad paths. In Proof 4 we will see another instance of a uniform partition.

As often happens in mathematics, it appears that the above 2003 proof is essentially a rediscovery of the 1923 proof by Aeppli. Aeppli's proof of the ballot theorem appeared in his dissertation [**3**], and was not widely available until Takács [**30**] provided a "somewhat modified version" of the proof in 1997. In his proof, Aeppli uses no geometric reasoning, and instead of counting the number of good ballot permutations he computes the probability that a ballot permutation is good (provided, of course, that all ballot permutations are equally likely). He partitions the bad ballot permutations in exactly the same manner as the preceding proof does; moreover, to show a one-to-one correspondence he reverses an initial portion of a ballot permutation, which is geometrically equivalent to rotating an initial portion of a lattice path.

Proof 2: Induction

In 1887 Barbier stated the ballot theorem for $k \geq 1$ without proof. If he had a proof, one supposes it followed the inductive proof that Bertrand sketched for the case $k = 1$. An inductive proof is not difficult to construct, and no record seems to exist for the "first" such proof of the ballot theorem. The following proof is similar to that found in Takács [**30**].

Proof 2. Let $N_k(a, b)$ denote the number of ways the $a + b$ ballots $(a \geq kb)$ can be ordered so that candidate A maintains more than k times as many votes as B throughout the counting of the ballots. The conditions $N_k(a, 0) = 1$ for all $a > 0$ and $N_k(kb, b) = 0$ for all $b > 0$, are easily verified by considering the statement of the ballot problem, and they both satisfy $N_k(a, b) = \frac{a-kb}{a+b}\binom{a+b}{a}$.

For $b > 0$ and $a > kb$, we see that $N_k(a, b) = N_k(a, b-1) + N_k(a-1, b)$ by considering the last vote in a ballot permutation. By induction, this quantity is $\frac{a-k(b-1)}{a+b-1}\binom{a+b-1}{a} + \frac{a-1-kb}{a+b-1}\binom{a+b-1}{a-1}$ which simplifies to $\frac{a-kb}{a+b}\binom{a+b}{a}$ as needed. $\blacksquare$

Proof 3: The Cycle Lemma

In the ballot theorem we are given an expression where the total number of ballot permutations $\binom{a+b}{a}$ is multiplied by the fraction $(a - kb)/(a + b)$. Dvoretzky and Motzkin [**12**] solve the ballot problem by introducing the *cycle lemma* which makes evident the reason for the fraction. The cycle lemma provides a surprising result: for any ballot sequence of a votes for A and b votes for B, exactly $a - kb$ of the $a + b$ cyclic permutations of the sequence are good. Consequently, a fraction of $(a - kb)/(a + b)$ of all ballot permutations are good.

Dershowitz and Zaks [**11**] give two elegant proofs of the cycle lemma. Their first proof is a generalization (and simplification) of the proofs in [**7**], [**25**], and [**26**]; their second proof follows [**15**], [**22**], and [**32**]. In the following proof of the ballot theorem, we include what is essentially their first proof of the cycle lemma.

Proof 3. We can express a ballot permutation as a sequence of $a + b$ terms where each term is either 1 or $-k$; votes for A correspond to the 1's and votes for B correspond to the $-k$'s. A sequence is called *good* if every partial sum is positive, and *bad* otherwise. Observe that the sum of a sequence is $a - kb \geq 0$.

Let C be any circular arrangement of a 1's and b $-k$'s. We now prove the cycle lemma: of the $a + b$ terms in C, exactly $a - kb$ start good sequences when C is read once around clockwise.

By the pigeonhole principle there must exist a sequence $X = 1, 1, \ldots, 1, -k$ in C with k consecutive 1's. No term of X can start a good sequence, for when we get to the $-k$ we would have a partial sum less than or equal to zero.

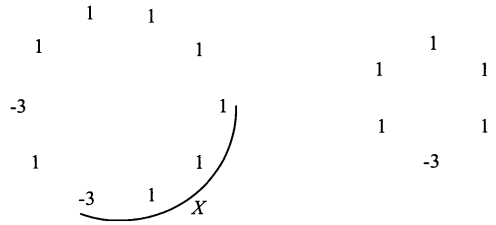


Figure 2 Example of C and C' with $k = 3$.

Let C' be the circular arrangement created from C by removing X. Since the sequence X has sum 0, it has no "net effect" on good sequences. Thus, a term of C starts a good sequence if and only if the corresponding term in C' starts a good sequence. Consequently, C and C' have exactly the same number of terms that start good sequences. Continuing in this manner, one removes sequences of the form $1, 1, 1, \ldots, 1, -k$ until a circular arrangement consisting only of 1's remains. At this stage, there are $a - kb$ 1's, and every term starts a good sequence. Hence there are exactly $a - kb$ good sequences in C, and the cycle lemma is proved.

If there is periodicity in C then not all $a - kb$ good sequences will be distinct. However, we can conclude that the *ratio* of good sequences to all sequences is $a - kb$ to $a + b$. Therefore, the number of good sequences is

$$\frac{a - kb}{a + b}\binom{a + b}{a}.$$

■

Dvoretzky and Motzkin [**12**] state and prove the cycle lemma as a means of solving the ballot problem, but Dershowitz and Zaks [**11**] point out that this is a "frequently rediscovered combinatorial lemma" and they provide two other applications of the lemma. They write,

> The Cycle Lemma is the combinatorial analogue of the Lagrange inversion formula; see Raney [**22**], Cori [**10**], and Gessel [**13**]. Other proofs of varying degree of generality may be found in Dvoretzky and Motzkin [**12**] (discussed in Grossman [**15**]), Motzkin [**21**] (two proofs), Hall [**16**], Raney [**22**], Yaglom and Yaglom [**32**], Takács [**29**], Silberger [**25**], Bergman [**7**] (three proofs), Sands [**24**] and Singmaster [**26**]. (The first paper [**12**] is not credited by the other authors,

but is referenced in Barton and Mallows [**6**] and Mohanty [**20**].) Dvoretzky and Motzkin, Motzkin, and Yaglom and Yaglom give the lemma in its general form; the other papers prove only the case $k = 1$ or $a - kb = 1$. Generalizations of the Cycle Lemma to non-integer k and sequences of reals may be found in Dvoretzky and Motzkin [**12**] and Spitzer [**27**], respectively.

(Reference numbers and notation in the above quote have been modified for consistency with this paper.)

Proof 4: A Uniform Partition

Consider the $\binom{2n}{n}$ possible lattice paths starting from the origin and consisting of n upsteps $(1, 1)$ and n downsteps $(1, -1)$. It turns out, surprisingly, that the number of these paths with i upsteps above the x-axis $(0 \le i \le n)$ is the same, regardless of the value of i. Consequently, the number of paths with all n upsteps above the x-axis must be $\binom{2n}{n}/(n+1)$. This fact is often called the Chung-Feller theorem [**9**, Thm. 2A]; however, it was actually given in 1909 by MacMahon [**18**, p. 167, §20] in the process of solving the ballot problem (for $k = 1$) via the theory of partitions.

In the following proof we apply a similar approach by creating a set Ψ with $(a - kb)\binom{a+b}{a}$ elements, and partitioning this set into $a + b$ subsets of equal size (that is, we uniformly partition Ψ into $a + b$ subsets). One of the subsets corresponds to the set of good ballot permutations, and from this we can conclude the ballot theorem. It appears that the following proof is the first to prove the ballot theorem by means of a uniform partition. It is based on and extends the proofs found in [**31**].

Proof 4. Consider lattice paths starting from the origin and consisting of a upsteps $(1, 1)$ and b downsteps $(1, -k)$, and assume the strict inequality $a > kb$. Let $\mathcal{A} = \mathcal{A}(a, b, k)$ be the set of all such paths. Given path $P \in \mathcal{A}$ we let $L(P)$ denote the set of x-values of the $a - kb$ "rightmost lowest" vertices of P; see Figure 3. More precisely, given path $P \in \mathcal{A}$, let y_0 denote the least y-value of all the vertices of P, and let $r(t)$ denote the x-value of the rightmost vertex of P along the line $y = t$; then $L(P) = \{r(t) \mid t \in \mathbb{Z},\ y_0 \le t \le y_0 + (a - kb) - 1\}$.

Let $\Psi = \{(P, j) \mid P \in \mathcal{A},\ j \in L(P)\}$ and note that $|\Psi| = (a - kb)\,|\mathcal{A}|$. Let $\Omega_i = \{(P, i) \in \Psi \mid i \in L(P)\}$, defined for $0 \le i \le a + b - 1$. The sets Ω_i partition Ψ into $a + b$ disjoint subsets.

Claim 1: There is a one-to-one correspondence between Ω_0 and the set of good paths. If $P \in \mathcal{A}$ is good, then $(0, 0)$ is the lowest vertex in P and it is the only vertex on the x-axis, so $(P, 0) \in \Omega_0$. Conversely, if $(P, 0)$ is in Ω_0, then no vertex of P can lie on the x-axis to the right of the origin, and so P is good.

Claim 2: The sets Ω_i uniformly partition Ψ. We show this by providing a one-to-one correspondence between Ω_i and Ω_0. If $(P, i) \in \Omega_i$, then write $P = XY$ where X is the initial path of P consisting of the first i steps, and Y consists of the remaining steps. Since $i \in L(P)$, we can observe that Y stays above the height of its initial vertex, and X never descends $a - kb$ or more units below the height of its terminal vertex. Consequently the path YX is good and $(YX, 0) \in \Omega_0$. Conversely, if $(Q, 0) \in \Omega_0$, then write $Q = YX$ where X consists of the final i steps of Q. The same qualities of X and Y hold as noted above, and the pair $(XY, i) \in \Omega_i$.

The two claims above imply that the number of good paths in $\mathcal{A}$ is

$$|\Omega_0| = \frac{|\Psi|}{a+b} = \frac{(a - kb)\,|\mathcal{A}|}{a+b} = \frac{a - kb}{a+b}\binom{a+b}{a}. \qquad \blacksquare$$

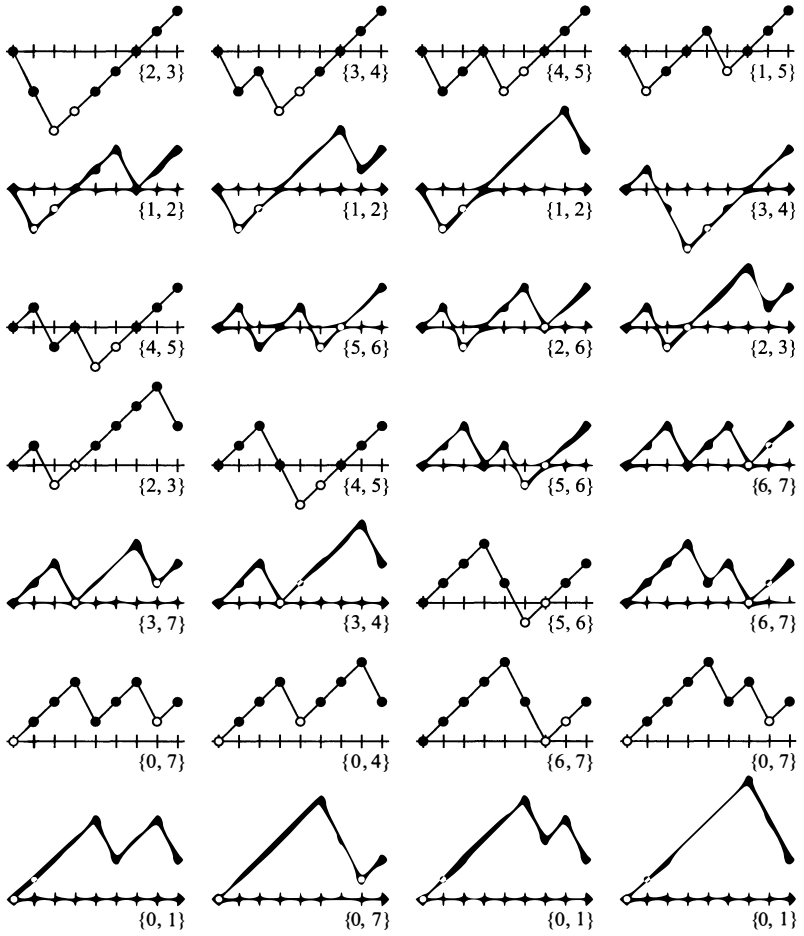


Figure 3 Example with $k = 2$, $a = 6$, $b = 2$. For each path $P \in \mathcal{A}$, the path P and the set $L(P)$ are shown. Observe that among all the sets $L(P)$, each number from 0 to 7 occurs exactly 7 times.

Suppose we let $\mathcal{A}_i$ denote the set of paths for which i is among the x-values of the $a - kb$ rightmost vertices, i.e., $\mathcal{A}_i = \{P \in \mathcal{A} \mid i \in L(P)\}$. When $a - kb = 1$, the sets $\mathcal{A}_0, \mathcal{A}_1, \ldots, \mathcal{A}_{a+b-1}$ are disjoint and all have the same cardinality. In other words, partitioning $\mathcal{A}$ according to the x-value of a path's rightmost lowest vertex creates a uniform partition of $\mathcal{A}$.

Curiously, when we allow $a - kb \geq 1$, the sets $\mathcal{A}_i$ continue to have the same cardinality. However, they are no longer disjoint. To the contrary, each path in $\mathcal{A}$ will be a member of precisely $a - kb$ of these sets.

The Weak Ballot Problem, Catalan Numbers

The ballot problem is often stated in a "weak" version: suppose that candidate A receives m votes and candidate B receives n votes, where $m \geq kn$ for some positive integer k, and compute the number of ways the ballots can be ordered so that A always has *at least* k times as many votes as B throughout the counting of the ballots.

Any ballot permutation in which A maintains *at least* k times the number of votes for B can be converted into one in which A maintains *more than* k times the number of votes for B by simply appending a vote for A to the beginning of the permutation. Clearly this process is reversible, and hence the solution to the weak version is the same as the "strict" version when A receives $m + 1$ votes and B receives n votes:

$$\frac{(m+1)-kn}{(m+1)+n}\binom{(m+1)+n}{m+1} = \frac{m+1-kn}{m+1}\binom{m+n}{m}.$$

Putting $k = 1$ and $m = n$ produces the well-known *Catalan numbers*

$$C_n = \frac{1}{n+1}\binom{2n}{n}.$$

Requiring only that $m = kn$ produces the *generalized Catalan numbers*, also called the *k-Catalan numbers*

$$C_n^k = \frac{1}{kn+1}\binom{(k+1)n}{n}.$$

The interested reader should see [**28**, pp. 219–229] and Stanley's website `http://www-math.mit.edu/~rstan/ec/` for an extensive list of combinatorial interpretations of the Catalan numbers. Furthermore, see [**17**] for several interpretations of the generalized Catalan numbers.

Acknowledgments. I would like to extend sincere thanks to the two referees who offered many valuable suggestions.

REFERENCES

1. J. Aebly, Démonstration du problème du scrutin par des considérations géométriques, *L'enseignement mathématique* **23** (1923) 185–186.
2. A. Aeppli, A propos de l'interprétation géométrique du problème du scrutin, *L'enseignement mathématique* **23** (1923) 328–329.
3. A. Aeppli, *Zur Theorie verketteter Wahrscheinlichkeitem, Markoffsche Ketten höherer Ordnung*, Ph.D. Thesis, Eidgenössische Technische Hochschule, Zürich, 1924.
4. D. André, Solution directe du problème résolu par M. Bertrand, *Comptes Rendus de l'Académie des Sciences*, Paris **105** (1887) 436–437.
5. É. Barbier, Généralisation du problème résolu par M. J. Bertrand, *Comptes Rendus de l'Académie des Sciences*, Paris **105** (1887) p. 407.
6. D.E. Barton and C.L. Mallows, Some aspects of the random sequence, *Ann. Math. Stat.* **36** (1965) 236–260.
7. G.M. Bergman, Terms and cyclic permutations, *Algebra Universalis* **8** (1978) 129–130.
8. J. Bertrand, Solution d'un problème, *Comptes Rendus de l'Académie des Sciences*, Paris **105** (1887) p. 369.
9. K.L. Chung and W. Feller, On fluctuations in coin tossing, *Proc. Natl. Acad. Sci. USA* **35** (1949) 605–608.
10. R. Cori, in: *Combinatorics on Words* (M. Lothaire, ed.), Encyclopedia of Mathematics and Its Applications, vol. 17, Addison-Wesley, Reading, Massachusetts, 1983.
11. N. Dershowitz and S. Zaks, The cycle lemma and some applications, *Europ. J. Combinatorics* **11** (1990) 35–40.
12. A. Dvoretzky and Th. Motzkin, A problem of arrangements, *Duke Math. J.* **14** (1947) 305–313.
13. I.M. Gessel, A combinatorial proof of the multivariate Lagrange inversion formula, *J. Combin. Theory Ser. A* **45** (1987) 178–195.
14. I.P. Goulden and L.G. Serrano, Maintaining the spirit of the reflection principle when the boundary line has arbitrary integer slope, *J. Comb. Theory, Ser. A* **104** (2003) 317–326.
15. H.D. Grossman, Fun with lattice points—21, *Scripta math.* **16** (1950) 120–124.
16. P. Hall, Some word problems, *J. London Math. Soc.* **33** (1958) 482–496.
17. P. Hilton and J. Pedersen, Catalan numbers, their generalizations, and their uses, *Math. Intell.* **13** (1991) 64–75.
18. P.A. MacMahon, Memoir on the theory of the partitions of numbers. part iv: on the probability that the successful candidate at an election by ballot may never at any time have fewer votes than the one who is

unsuccessful; on a generalization of this question; and on its connexion with other questions of partition, permutation, and combination, *Philosophical Transactions of the Royal Society of London, Series A* **209** (1909) 153–175. Also *Collected Papers Vol. 1* (G.E. Andrews, ed.), MIT Press, Cambridge, Mass 1978, 1292–1314.
19. D. Mirimanoff, A propos de l'interprétation géométrique du problème du scrutin, *L'enseignement mathématique* **23** (1923) 187–189.
20. S.G. Mohanty, *Lattice Path Counting and Applications*, Academic Press, New York, 1979.
21. Th. Motzkin, Relations between hypersurface cross ratios and a combinatorial formula for partitions of a polygon, for permanent preponderance, and for non-associative products, *Bull. Am. Math. Soc.* **54** (1948) 352–360.
22. G.M. Raney, Functional composition patterns and power series reversion, *Trans. Am. Math. Soc.* **94** (1960) 441–451.
23. M.S. Renault, Lost (and found) in translation: André's actual method and its application to the generalized ballot problem, *Amer. Math. Monthly* to appear. See `webspace.ship.edu/msrenault/ballotproblem/`.
24. A.D. Sands, On generalized Catalan numbers, *Discr. Math.* **21**(2) (1978) 219–221.
25. D.M. Silberger, Occurrences of the integer $(2n-2)!/n!\,(n-1)!$, *Roczniki Polskiego Towarzystwa Math. I* **13** (1969) 91–96.
26. D. Singmaster, An elementary evaluation of the Catalan numbers, *Am. Math. Monthly* **85** (1978) 366–368.
27. F. Spitzer, A combinatorial lemma and its application to probability theory, *Trans. Am. Math. Soc.* **82** (1956) 323–339.
28. R.P. Stanley, *Enumerative Combinatorics, Vol. 2*, Cambridge University Press, Cambridge, UK, 1999.
29. L. Takács, *Combinatorial Methods in the Theory of Stochastic Processes*, John Wiley, New York, 1967.
30. L. Takács, On the ballot theorems, *Advances in Combinatorial Methods and Applications to Probability and Statistics*, Birkhäuser, 1997.
31. W. Woan, Uniform partitions of lattice paths and Chung-Feller generalizations, *Amer. Math. Monthly* **108** (2001) 556–559.
32. A.M Yaglom and I.M. Yaglom, *Challenging Mathematical Problems with Elementary Solutions, vol 1: Combinatorial Analysis and Probability Theory*, Holden Day, San Francisco, 1964.

An Intriguing Property of the Center of Mass for Points on Quadratic Curves and Surfaces

LEONID G. HANIN
Idaho State University
Pocatello, ID 83209-8085
hanin@isu.edu

ROBERT J. FISHER
Idaho State University
Pocatello, ID 83209-8085
fishrobe@isu.edu

BORIS L. HANIN
Stanford University
Stanford, CA 94309
bhanin@stanford.edu

In 1972, a prominent Russian geometer Zalman A. Skopets (1917–1984), posed the following problem to the first author:

Let $X_1, X_2, \ldots, X_n$ be $n \geq 2$ points on a cirle, and C be their geometric center of mass. Denote by $Y_1, Y_2, \ldots, Y_n$ the second points of intersection of the lines $X_1C, X_2C, \ldots, X_nC$ with the circle, respectively (see Figure 1). Prove that

$$\frac{X_1C}{CY_1} + \frac{X_2C}{CY_2} + \cdots + \frac{X_nC}{CY_n} = n. \tag{1}$$

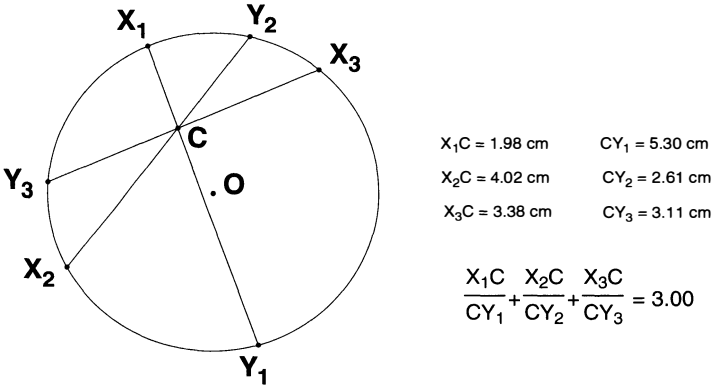


Figure 1 Illustration of equation (1) for $n = 3$ points on a circle.

This result is truly intriguing in many respects. First, the sum in (1) is independent of the locations of points $X_1, X_2, \ldots, X_n$; moreover, some of these points may coincide. Second, while for $n = 2$ the result is obvious, its direct verification using analytic geometry even in the case $n = 3$ is almost insurmountable. Finally, there seems to be no apparent geometric or physical reason as to why (1) should be true. Readers are encouraged to disprove this claim.

The historical origin of the problem remains unclear. Professor Skopets himself referred the problem "to an old German book on geometry." A careful examination of the two-volume *Gesammelte Werke* [**6**] of Jacob Steiner (1796–1863), a great Swiss/German geometer who discovered many important properties of the center of mass and moments of inertia, did not produce anything even remotely similar to the problem at hand.

The main body of the paper is divided into three parts. In the first part we prove property (1) of the center of mass, describe all points in the plane with the same property, and consider some natural generalizations. Our proof of property (1) for the circle serves as a primary driving force for all the results in this work. The first part of the paper forms the core of the third author's award winning high school research project [**4**] that was mentored by the second author. In the second part of the paper we extend our results to arbitrary quadratic curves in $\mathbb{R}^2$ and surfaces in $\mathbb{R}^3$. Finally, in the third part we discuss our findings from various angles and trace further generalizations.

Laying a Foundation

In what follows, points $X_1, X_2, \ldots, X_n$ on a circle (and other curves or surfaces of interest) will be assumed fixed. Note that the left-hand side of (1) can be defined for any point P inside the circle and thus becomes a function of P :

$$f(P) := \sum_{i=1}^{n} \frac{X_i P}{P Y_i}. \tag{2}$$

Then the initial problem is formulated as the following theorem.

THEOREM 1. *Let $X_1, X_2, \ldots, X_n$ be $n \geq 2$ points on a circle $\mathcal{S}$, and let C be their geometric center of mass. Then $f(C) = n$.*

Proof. Let r be the radius of $\mathcal{S}$ and O be its center. Pick any point P inside $\mathcal{S}$. For any two chords XY and AB of the circle $\mathcal{S}$ intersecting at the point P we have $XP \cdot PY = AP \cdot PB$, see e.g. [**2**, 116–118]. Apply this property to a chord XY and the diameter that both pass through P. Then (see Figure 2)

$$XP \cdot PY = (r + OP)(r - OP) = r^2 - OP^2. \tag{3}$$

This allows us to rewrite (2) as follows:

$$f(P) = \sum_{i=1}^{n} \frac{X_i P^2}{X_i P \cdot P Y_i} = \frac{1}{r^2 - OP^2} \sum_{i=1}^{n} X_i P^2. \tag{4}$$

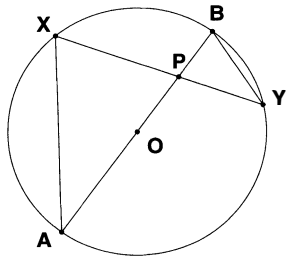


Figure 2 Triangles $\triangle XAP$ and $\triangle BYP$ are similar so that $XP \cdot PY = AP \cdot PB$.

Next, by vector algebra, we have for every point X on the circle

$$XP^2 = \mathbf{XP}^2 = (\mathbf{OP} - \mathbf{OX})^2 = \mathbf{OP}^2 - 2\mathbf{OP} \cdot \mathbf{OX} + \mathbf{OX}^2 = OP^2 + r^2 - 2\mathbf{OP} \cdot \mathbf{OX},$$

where $\mathbf{OP} \cdot \mathbf{OX}$ denotes hereafter the scalar product of vectors $\mathbf{OP}$ and $\mathbf{OX}$. Therefore,

$$\sum_{i=1}^{n} X_i P^2 = n(OP^2 + r^2) - 2\mathbf{OP} \cdot \sum_{i=1}^{n} \mathbf{OX}_i. \tag{5}$$

By definition of the geometric center of mass of the points X_i, $1 \leq i \leq n$,

$$\sum_{i=1}^{n} \mathbf{OX}_i = n\,\mathbf{OC}.$$

Then (5) becomes

$$\sum_{i=1}^{n} X_i P^2 = n\,(OP^2 + r^2 - 2\mathbf{OP} \cdot \mathbf{OC}). \tag{6}$$

Setting here $P = C$ yields

$$\sum_{i=1}^{n} X_i C^2 = n(r^2 - OC^2).$$

Substitution of this equation into (4) with $P = C$ completes the proof. ■

The solution set of the equation

$$f(P) = n \tag{7}$$

is not limited to the point $P = C$. For example, equation (7) is obviously met for $P = O$, the center of the circle. Surprisingly, if $C \neq O$ then there are points P inside the circle with property (7) that are different from C and O.

PROPOSITION 1. *The set of points P inside the circle $\mathcal{S}$ that satisfy the equation $f(P) = n$ is the circle with diameter OC.*

Proof. The proof of Theorem 1 up to and including equation (6) holds for every point P inside the circle. In particular, it follows from (4) and (6) that equation (7) is equivalent to $\mathbf{OP}^2 = \mathbf{OP} \cdot \mathbf{OC}$. This means that

$$\mathbf{OP} \cdot \mathbf{PC} = \mathbf{OP} \cdot (\mathbf{OC} - \mathbf{OP}) = \mathbf{OP} \cdot \mathbf{OC} - \mathbf{OP}^2 = 0$$

or, in other words, that vectors $\mathbf{OP}$ and $\mathbf{PC}$ are perpendicular. This condition defines the circle with diameter OC (see Figure 3), as claimed. ■

Are there solutions of equation (7) among points P outside the circle? To answer this question, it is convenient to extend the definition of function f. Specifically, when computing the function f for points P outside the circle $\mathcal{S}$ we will assume that the ratio XP/PY is *negative* and thus take into account the opposite orientations of segments XP and PY. This leads to the following commonly used agreement that we will adopt in the rest of the paper:

SIGN CONVENTION. *If points X, P, Y in a Euclidean space are collinear and $P \neq Y$ then the* ***signed ratio*** $\overline{XP}/\overline{PY}$ *of segments XP and PY is the unique real number λ such that $\mathbf{XP} = \lambda\mathbf{PY}$.*

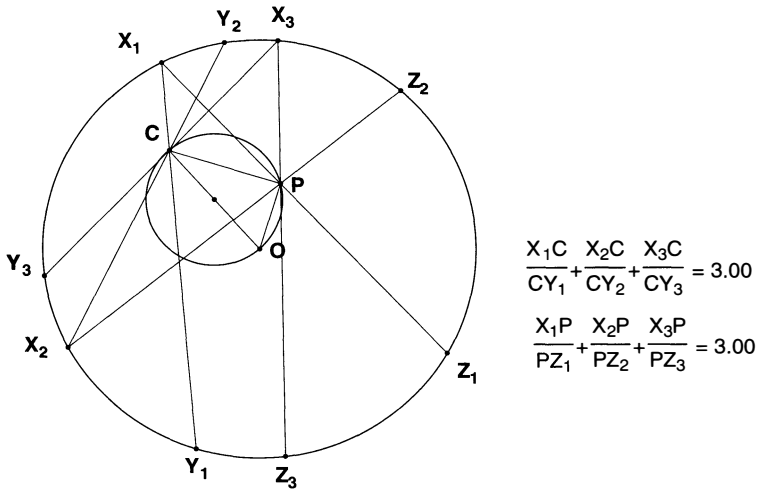


Figure 3 Proposition 1 in the case of $n = 3$ points: the solution set of the equation $f(P) = 3$ is the circle with diameter OC.

Observe that if point P is between X and Y then the signed ratio is the usual ratio of the lengths of segments XP and PY while if $Y = X$ then $\overline{XP}/\overline{PY} = -1$. Finally, if point Y is "at infinity" then $\overline{XP}/\overline{PY} = 0$. Similarly, one can define the signed product of the lengths of segments XP and PY. An important property of signed ratios of segments is that they are invariant under affine isomorphisms of the Euclidean space including translations, dilations, rotations and reflections.

Thus, for points P outside the circle, equation (7) becomes

$$f(P) = -n, \tag{8}$$

where the extended function f is defined by

$$f(P) := \sum_{i=1}^{n} \frac{\overline{X_i P}}{\overline{PY_i}}. \tag{9}$$

A key to solving equation (8) is to use inversion in the circle $\mathcal{S}$ [**2**, p. 452]. Recall that, given a circle $\mathcal{S}$ with center O and radius r, the *inversion* of a point $P \neq O$ in $\mathcal{S}$ is the point P' on the ray $\overrightarrow{OP}$ such that

$$OP \cdot OP' = r^2.$$

Note that inversion in a circle fixes the circle and interchanges its interior and exterior.

PROPOSITION 2. *A point P outside a circle $\mathcal{S}$ satisfies the equation $f(P) = -n$ if and only if P is the inversion in $\mathcal{S}$ of a point on the circle with diameter OC.*

Proof. For any two secants XY and AB through a given point P outside a circle we have $PX \cdot PY = PB \cdot PA$ [**2**, p. 116–118]. Applying this property to the secants PX and PO (see Figure 4) we find that (3) holds if the product $XP \cdot PY$ is assumed to be a signed product. Then the argument in the proof of Theorem 1 leading to equation (6) is true without any changes for the point P. Taken together, equations (4) and (6) imply that, for any point P outside the circle, equation $f(P) = -n$ is equivalent to

$$\mathbf{OP} \cdot \mathbf{OC} = r^2.$$

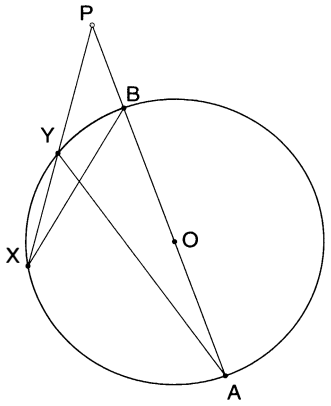


Figure 4 Triangles $\triangle PXB$ and $\triangle PAY$ are similar so that $PX \cdot PY = PA \cdot PB$.

Suppose now that point P is the inversion of a point P' that lies on the circle with diameter OC, so that $\mathbf{OP'} \cdot \mathbf{P'C} = 0$. Hence also $\mathbf{OP} \cdot \mathbf{P'C} = 0$, because vectors $\mathbf{OP}$ and $\mathbf{OP'}$ are collinear. Therefore,

$$\mathbf{OP} \cdot \mathbf{OC} = \mathbf{OP} \cdot (\mathbf{OP'} + \mathbf{P'C}) = \mathbf{OP} \cdot \mathbf{OP'} = OP \cdot OP' = r^2$$

implying that $f(P) = -n$. Conversely, suppose $f(P) = -n$ for a point P outside the circle so that $\mathbf{OP} \cdot \mathbf{OC} = r^2$. Together with the definition of inversion this yields

$$\mathbf{OP} \cdot \mathbf{P'C} = \mathbf{OP} \cdot (\mathbf{OC} - \mathbf{OP'}) = \mathbf{OP} \cdot \mathbf{OC} - \mathbf{OP} \cdot \mathbf{OP'} = r^2 - r^2 = 0.$$

In view of the collinearity of vectors $\mathbf{OP}$ and $\mathbf{OP'}$ this implies $\mathbf{OP'} \cdot \mathbf{P'C} = 0$. Therefore, point P' belongs to the circle with diameter OC. ■

Combining Propositions 1 and 2, we obtain the following theorem (see Figures 3 and 5).

THEOREM 2. *Let $X_1, X_2, \ldots, X_n$ be $n \geq 2$ points on a circle $\mathcal{S}$ centered at a point O, and let C be their geometric center of mass. Then the set of all points P in the plane that satisfy the equation $f(P) = n$, where function f is defined by* (9)*, is the circle with diameter OC. Furthermore, if $C \neq O$ then the solution set for the equation $f(P) = -n$ is the line through C' (the image of the point C under the inversion in $\mathcal{S}$) that is perpendicular to OC.*

Since the function f is invariant under translations, dilations and rotations of the plane, Theorem 2 extends to ellipses. In fact, every ellipse can be obtained from the standard ellipse

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1, \quad a, b > 0,$$

by a combination of translation and rotation, while the standard ellipse is the image of the unit circle centered at the origin under the linear transformation $(x, y) \longmapsto (ax, by)$. This observation and Theorem 2 lead to the following more general result (see Figure 6).

THEOREM 3. *Let $X_1, X_2, \ldots, X_n$ be $n \geq 2$ points on an ellipse $\mathcal{S}$ centered at a point O, and let C be their geometric center of mass. Then:*

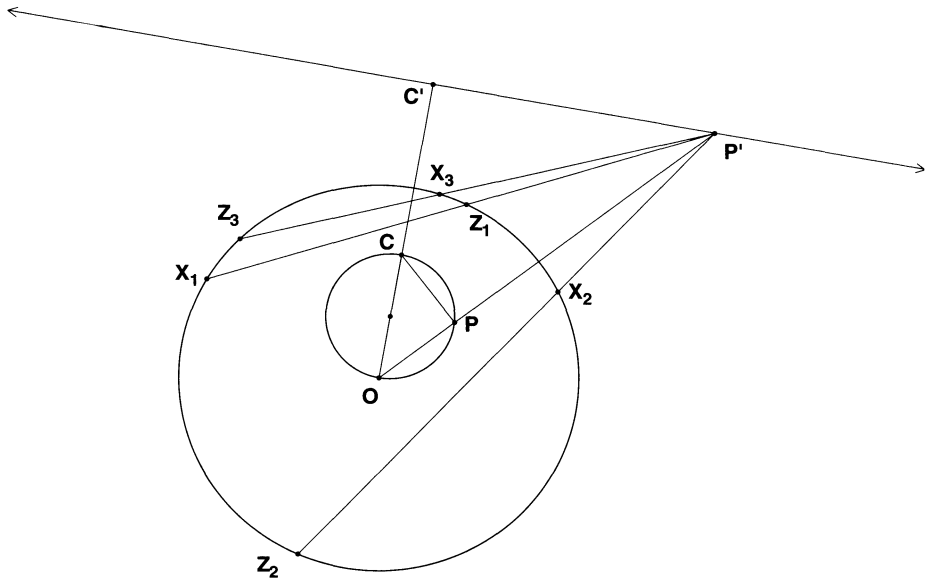


Figure 5 Proposition 2 for $n = 3$ points on a circle: the solution set of the equation $f(P) = -3$ is the inversion of the circle with diameter OC in the original circle. The inversion is the line through C' that is perpendicular to the line OC.

1. *The set of all points P in the plane that satisfy the equation $f(P) = n$ is the ellipse that is centered at the midpoint of OC, is similar to $\mathcal{S}$, passes through points O and C, and has its axes parallel to the respective axes of the ellipse $\mathcal{S}$.*
2. *If $C \neq O$ then the solution set of the equation $f(P) = -n$ is a line outside the ellipse $\mathcal{S}$ that is parallel to the tangent line to $\mathcal{S}$ at the point of intersection of the ray $\overrightarrow{OC}$ with $\mathcal{S}$.*

A quick analysis of the proofs of Theorems 1 and 2 suggests that these theorems hold with obvious modifications for spheres. Furthermore, application of the same transformation approach that allowed us to extend Theorem 2 to ellipses would show that an analogue of Theorem 3 is true for ellipsoids. This leads to the following questions:

1. *Can Theorems 1 and 2 be extended to planar curves other than circles and ellipses, in particular, to hyperbolas and parabolas?*
2. *For what surfaces in $\mathbb{R}^3$ do Theorems 1 and 2 generalize?*

These questions will be answered in the next section. As a motivation, we give a heuristic argument for the validity of formula (1) for points on a parabola. Recall that circles, ellipses, parabolas and hyperbolas are conic sections, that is, can be obtained by intersecting a conic surface in $\mathbb{R}^3$ by a plane that passes through (and rotates around) a fixed point on the axis of the cone different from its vertex. In particular, a parabola can be viewed as a limiting case of ellipses and arises when the secant plane is parallel to a generator of the cone. Then, for n points on a parabola, we rotate the secant plane to get an ellipse, project the points onto the ellipse along the conic surface from its vertex, apply formula (1) to these new n points on an ellipse, and finally take limit in (1) by rotating the plane back to its initial position to obtain property (1) for the parabola. (As an exercise, the reader is encouraged to make this a rigorous argument). Observe, however, that this approach would never work for hyperbolas.

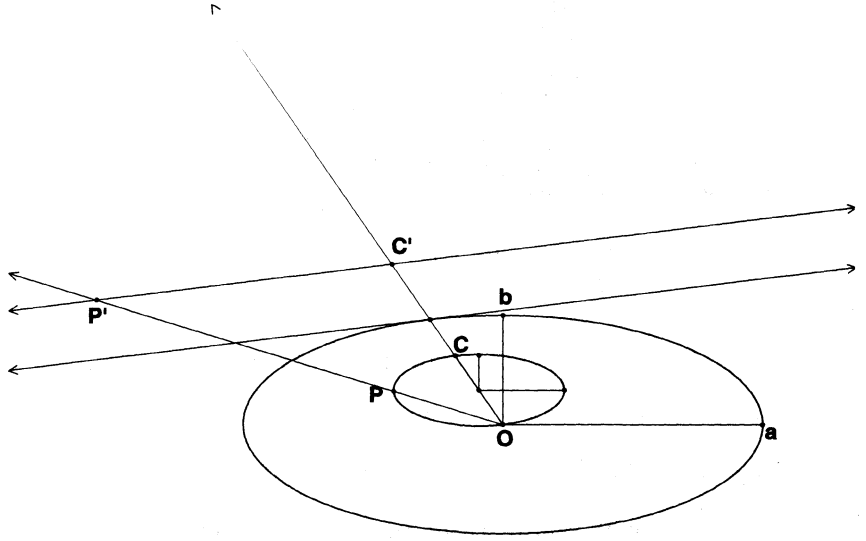


Figure 6 Illustration of Theorem 3.

Finally, if a curve or surface bounds a convex set (think of circles, ellipses and parabolas or the corresponding cylindrical surfaces or spheres and ellipsoids) then all ratios in (9) are positive for interior points P and negative for exterior points so that in this case consideration of signed ratios of segments is not strictly necessary. However, for non-convex curves and surfaces including those that are disconnected or self-intersecting (such as hyperbolas and pairs of intersecting or parallel lines in $\mathbb{R}^2$ or conic surfaces and pairs of intersecting or parallel planes in $\mathbb{R}^3$), the ratios in (9) may have different signs. In this case signed ratios are absolutely indispensable (to see this, construct a counterexample to formula (1) for three points on a two-branched hyperbola $y = 1/x$)!

Erecting a Building

Quadratic curves in $\mathbb{R}^2$ and surfaces in $\mathbb{R}^3$ will be called *quadrics*. To define a general quadric, we need some standard notation. For vectors in $\mathbb{R}^2$ and $\mathbb{R}^3$ with coordinate representations $x = (x_1, x_2)$ and $x = (x_1, x_2, x_3)$, we denote by $x \cdot y = \sum_{i=1}^{d} x_i y_i$ the Euclidean scalar product and write $x^2 = x \cdot x$. The position vector of a point X in $\mathbb{R}^d$ will be denoted by the corresponding lower case letter x, and we will alternate freely between the two notations.

A *quadric* in $\mathbb{R}^d$, $d = 2, 3$ (as well as in Euclidean spaces of higher dimensions), is defined as the set $\mathcal{S}$ of all points $x \in \mathbb{R}^d$ such that

$$Q(x) := Ax \cdot x + 2b \cdot x + \alpha = 0, \tag{10}$$

where A is a non-zero $d \times d$ real symmetric matrix, $b \in \mathbb{R}^d$ and $\alpha \in \mathbb{R}$ (see e.g. [**3**, p. 285]). Without restricting generality it will be assumed that quadric $\mathcal{S}$ does not degenerate into a line or plane.

The following elementary proposition carries a significant part of the computational effort required for the proof of our main result.

PROPOSITION 3. *Let $\mathcal{S}$ be a quadric* (10)*, $X \in \mathcal{S}$, and $P \notin \mathcal{S}$. Suppose that the line through the points X and P intersects $\mathcal{S}$ at a point $Y \neq X$. Then point Y is unique*

and

$$\frac{\overline{XP}}{\overline{PY}} = -\frac{A(p-x)\cdot(p-x)}{Q(p)}. \tag{11}$$

Proof. For the point Y we have $y = x + t(p-x)$ for some $t \neq 0$. Taking the difference of equations $Q(x) = 0$ and $Q(y) = 0$ we find that

$$\beta t + \gamma = 0, \tag{12}$$

where

$$\beta = A(p-x)\cdot(p-x) \quad \text{and} \quad \gamma = 2(Ax+b)\cdot(p-x).$$

Since $X \in \mathcal{S}$ and $P \notin \mathcal{S}$, we have

$$\beta + \gamma = (Ap + Ax + 2b)\cdot(p-x) = (Ap+2b)\cdot p - (Ax+2b)\cdot x = Q(p) \neq 0, \tag{13}$$

so that β and γ cannot both vanish. Therefore, equation (12) has a solution $t \neq 0$ if and only if $\beta, \gamma \neq 0$, in which case

$$t = -\gamma/\beta, \tag{14}$$

and thus the point Y is unique. From $y - p = (t-1)(p-x)$ we find using the definition of the signed ratio of segments and equations (13) and (14) that

$$\frac{\overline{XP}}{\overline{PY}} = \frac{1}{t-1} = -\frac{\beta}{\beta+\gamma} = -\frac{A(p-x)\cdot(p-x)}{Q(p)}. \tag{15}$$

■

Proposition 3 provides a formula for the signed ratio $\overline{XP}/\overline{PY}$ in the generic case $Y \neq X$. It remains to define this signed ratio in the case when X is the only point of intersection of the line XP with $\mathcal{S}$. This degeneracy may be of the following two kinds:

(i) Line XP is not tangent to the surface $\mathcal{S}$ at point X. In this case point Y is "at infinity," and we must set $\overline{XP}/\overline{PY} = 0$.

(ii) Line XP is tangent to the surface $\mathcal{S}$ at point X. Approximating the tangent line by a secant line and taking limit we conclude that in this case $Y = X$, that is, $\overline{XP}/\overline{PY} = -1$.

Notice that in case (i) $\gamma \neq 0$ and hence $\beta = 0$, see the proof of Proposition 3. Also, in case (ii) $\gamma = 0$ and hence by (13) $\beta \neq 0$. Applying (15) we find that in both cases formula (11) produces the same values for the signed ratio $\overline{XP}/\overline{PY}$ that we imputed geometrically. Thus, *formula (11) holds for all points $X \in \mathcal{S}$ and $P \notin \mathcal{S}$.*

We are now ready to formulate and prove our main result.

THEOREM 4. *Let $\mathcal{S}$ be a quadric* (10) *in $\mathbb{R}^d$, $d = 2, 3$, and let $X_1, X_2, \ldots, X_n$ be any $n \geq 2$ points on $\mathcal{S}$. Suppose that their geometric center of mass C does not belong to $\mathcal{S}$. Then*

$$f(C) = n. \tag{16}$$

Furthermore, the set of all points P in $\mathbb{R}^d$ such that $f(P) = n$ is a quadric

$$Ap\cdot p + (b - Ac)\cdot p - b\cdot c = 0. \tag{17}$$

Finally, if the set of all points P in $\mathbb{R}^d$ such that $f(P) = -n$ is non-empty then it is a line or plane with the equation

$$(Ac + b) \cdot p + b \cdot c + \alpha = 0. \tag{18}$$

Proof. Let $P \notin \mathcal{S}$. Setting in (11) $X = X_i$, $Y = Y_i$, $i = 1, 2, \ldots, n$, we obtain

$$f(P) = \sum_{i=1}^{n} \frac{\overline{X_iP}}{\overline{PY_i}} = -\frac{1}{Q(p)} \sum_{i=1}^{n} A(p - x_i) \cdot (p - x_i). \tag{19}$$

For $i = 1, 2, \ldots, n$, we expand $A(p - x_i) \cdot (p - x_i)$ and use $Q(x_i) = 0$ to have

$$A(p - x_i) \cdot (p - x_i) = Ap \cdot p - 2(Ap \cdot x_i) - 2b \cdot x_i - \alpha.$$

By definition of the geometric center of mass, $\sum_{i=1}^{n} x_i = nc$. Therefore, in view of (19)

$$f(P) = -n \, \frac{Ap \cdot p - 2(Ap \cdot c) - 2b \cdot c - \alpha}{Q(p)}. \tag{20}$$

In particular, setting here $p = c$ we obtain (16). Also, it follows from (20) that $f(P) = n$ if and only if point P satisfies (17) while $f(P) = -n$ is equivalent to equation (18). Finally, observe that because $C \notin \mathcal{S}$

$$(Ac + b) \cdot c + b \cdot c + \alpha = (Ac + 2b) \cdot c + \alpha = Q(c) \neq 0.$$

Therefore, the vector $Ac + b$ and number $b \cdot c + \alpha$ cannot be both zero. Thus, a non-empty solution set of equation (18) must be a line in $\mathbb{R}^2$ or plane in $\mathbb{R}^3$. ∎

Adding a Finishing Touch

Theorem 4 implies Theorems 2 and 3 for circles and ellipses and extends these theorems to hyperbolas, parabolas and pairs of intersecting or parallel lines in the plane. It also gives rise to a plethora of more specific results for spherical, elliptic, hyperbolic, parabolic, conic and cylindrical quadrics as well as pairs of intersecting or parallel planes in $\mathbb{R}^3$. Furthermore, Theorem 4 and its proof lead to the following general conclusions.

1. Theorem 4 is true for quadrics (10) in Euclidean spaces of any dimension $d \geq 2$. Moreover, it holds not only for the geometric center of mass where masses at points $X_1, X_2, \ldots, X_n$ are equal but also for the center of mass with arbitrary masses at these points. Finally, an analogue of Theorem 4 is true for any finite mass distribution on a quadric, most notably, for the surface area on a bounded quadric.

2. It is curious to mention that solution sets (17) and (18) for the equations $f(P) = \pm n$ depend on the points $X_1, X_2, \ldots, X_n$ only through their geometric center of mass C, compare with Theorems 2 and 3, and that the solution set $\mathcal{S}_c$ for the equation $f(P) = n$ does not depend on the parameter α in (10). We note also that $\mathcal{S}_c$ is a quadric of the same type as the original quadric $\mathcal{S}$, see equation (17).

3. For many quadrics $\mathcal{S}$, the assumption $C \notin \mathcal{S}$ is met automatically. Specifically, this is the case for circles, ellipses, parabolas, spheres, ellipsoids and circular or elliptic paraboloids while this is not necessarily true for hyperbolas. Examination of the canonical forms [**1**, Ch. XIII] of quadrics shows that all other quadratic surfaces contain lines or hyperbolas so that in this case the geometric center of mass of points

$X_1, X_2, \ldots, X_n$ on $\mathcal{S}$ may occasionally belong to $\mathcal{S}$. Furthermore, the sets of points P with the equations (17) and (18) may intersect the original surface $\mathcal{S}$. Although we were assuming initially that point P does not belong to $\mathcal{S}$, the intersection points may be included in the solution sets of the equations $f(P) = \pm n$ "by continuity." It is easy to see that any such point must satisfy *both* equations (17) and (18).

4. If matrix A is non-singular then one may assume that $b = 0$ in which case quadric $\mathcal{S}$ is symmetric about the origin. Then equations (17) and (18) take on, respectively, the following simpler forms:

$$Ap \cdot (p - c) = 0$$

and

$$Ac \cdot p + \alpha = 0. \tag{21}$$

In particular, the set $\mathcal{S}_c$ contains the origin. Furthermore, if matrix A is positive definite then equation (10) becomes

$$Ax \cdot x = a, \quad \text{where} \quad a = |\alpha| > 0, \tag{22}$$

and hence represents either an ellipse (or a circle) in $\mathbb{R}^2$ or an ellipsoid (or a sphere) in $\mathbb{R}^3$. We define the *inversion* with respect to the surface (22) by mapping any point $P \neq O$ to a point P' on the ray $\overrightarrow{OP}$ such that $(Ap \cdot p) \cdot (Ap' \cdot p') = a^2$. Then the sets (17) and (18) can be obtained from each other by inversion, compare with Proposition 2. Notice also that in the case $c \neq 0$ equation (21) in $\mathbb{R}^3$ represents a plane parallel to the tangent plane to the surface (22) at the point θc on this surface, where $\theta = \sqrt{a/(Ac \cdot c)}$, compare with Theorems 2 and 3.

Finally, the reader is invited to visit the web site `http://www.isu.edu/math/links.shtml` and view dynamic constructions on *The Geometer's Sketchpad* **[5]** that illustrate the validity of Theorem 4 for all the conic sections.

REFERENCES

1. M. Berger, *Geometry*, vol. 1–2, Springer-Verlag, New York, 1987.
2. R. Bix, *Topics in Geometry*, Academic Press, Boston, 1994.
3. W.H. Greub, *Linear Algebra*, Springer-Verlag, New York, 3rd ed., 1967.
4. B.L. Hanin, Geometric center of mass for points on conic sections: Properties, generalizations, applications, and mysteries. *Reflections of Young Scientists. Essays by winners of 2005 Intel Science Talent Search*, 11 pp. `http://mazziotti.uchicago.edu/journal/main.htm`
5. J. King, *Geometry Through the Circle with The Geometer's Sketchpad*, Key Curriculum Press, Berkeley, CA, 1996.
6. J. Steiner, *Gesammelte Werke*, vol. 1–2, Chelsea Publishing Company, 2nd ed., 1971.

NOTES

Another Approach to Solving $A = mP$ for Triangles

TOM LEONG
The University of Scranton
Scranton, PA 18510
thomas.leong@scranton.edu

DIONNE T. BAILEY
Angelo State University
San Angelo, TX 76909
dionne.bailey@angelo.edu

ELSIE M. CAMPBELL
Angelo State University
San Angelo, TX 76909
elsie.campbell@angelo.edu

CHARLES R. DIMINNIE
Angelo State University
San Angelo, TX 76909
charles.diminnie@angelo.edu

PAUL K. SWETS
Angelo State University
San Angelo, TX 76909
paul.swets@angelo.edu

The problem (and solution) of finding all integer-sided triangles whose area A and perimeter P are numerically equal appear as far back as 1865 in [**2**] and more recently in [**1**] and [**8**]. Although differing in details, all of these solutions required a good measure of trial and error. So, not surprisingly, the generalization of solving $A = mP$ for integers $m \geq 2$ wasn't introduced (and partially solved) until 1985 in [**3**]. Coincidentally, in 2006, [**4**] posed the problem of solving $A = 3P$ with $P > 2000$ and, a few months later, a recent article [**5**] in this MAGAZINE gave an algorithm for completely solving $A = mP$. We present another approach which gives a relatively simple characterization of all solutions and essentially relies only on the creative manipulation of a single equation.

A Special Diophantine Equation

We first consider the Diophantine equation

$$xyz = n(x + y + z) \tag{1}$$

where n is a fixed positive integer. By symmetry, we may assume $x \leq y \leq z$. This first lemma establishes the range of values of x required to find all solutions of (1).

LEMMA. *If $xyz = n(x + y + z)$ with $x \leq y \leq z$, then $x \leq \sqrt{3n}$.*

Proof. Since $x \le y \le z$, we immediately get

$$x^2 z \le xyz = n(x + y + z) \le 3nz$$

and the result follows. ∎

Thus we need only consider $x = 1, 2, \ldots, \left\lfloor \sqrt{3n} \right\rfloor$, where $\lfloor t \rfloor$ denotes the greatest integer not exceeding t. Our main Theorem describes how to find the solutions for y and z which go with each value of x.

THEOREM. *For each* $x = 1, 2, \ldots, \left\lfloor \sqrt{3n} \right\rfloor$, *let* $\overline{x} = x/\gcd(n, x)$, *and* $\overline{n} = n/\gcd(n, x)$. *The integral values of* y *and* z *(if any) which make* (x, y, z) *a solution of* (1) *are*

$$y = \frac{w + \overline{n}}{\overline{x}}, \qquad z = \frac{\hat{w} + \overline{n}}{\overline{x}} \tag{2}$$

where $w\hat{w} = \overline{n}(\overline{n} + \overline{x}x)$ *and* $w \le \hat{w}$ *such that*

$$w \ge \max\{\overline{x}x - \overline{n}, 1\} \qquad \textit{and} \qquad w \equiv -\overline{n} \pmod{\overline{x}}.$$

Proof. A proof easily follows if we rewrite (1) as

$$(\overline{x}y - \overline{n})(\overline{x}z - \overline{n}) = \overline{n}(\overline{n} + \overline{x}x).$$

Thus for w, $\hat{w}$ satisfying $w\hat{w} = \overline{n}(\overline{n} + \overline{x}x)$ and $w \le \hat{w}$, we can put

$$\overline{x}y - \overline{n} = w, \qquad \overline{x}z - \overline{n} = \hat{w},$$

that is,

$$y = \frac{w + \overline{n}}{\overline{x}}, \qquad z = \frac{\hat{w} + \overline{n}}{\overline{x}}.$$

To make $x \le y$, we need additionally that

$$x \le \frac{w + \overline{n}}{\overline{x}} \qquad \text{or} \qquad w \ge \overline{x}x - \overline{n}.$$

Of course, this is replaced by $w \ge 1$ if $\overline{x}x - \overline{n} \le 0$. Finally, if $w \equiv -\overline{n} \pmod{\overline{x}}$, then

$$(-\overline{n})\hat{w} \equiv w\hat{w} \equiv \overline{n}(\overline{n} + \overline{x}x) \equiv \overline{n}^2 \equiv (-\overline{n})(-\overline{n}) \pmod{\overline{x}}$$

and, consequently, $\hat{w} \equiv -\overline{n} \pmod{\overline{x}}$ since $\gcd(\overline{x}, -\overline{n}) = 1$. Hence, the condition $w \equiv -\overline{n} \pmod{\overline{x}}$ ensures that (2) yields integral solutions for y and z. ∎

As a result of this Theorem, (1) always has solutions: taking $x = w = 1$ yields $y = n + 1$ and $z = n^2 + 2n$. Furthermore, the Theorem and the Lemma guarantee that the number of solutions of (1) is finite because y and z depend on x and w which, in turn, are bounded by n.

The $A = mP$ Problem

Let a, b, c denote the side lengths of a triangle. For convenience, we assume that $a \le b \le c$. By Heron's Formula, the condition $A = mP$ can be rewritten

$$P(P - 2a)(P - 2b)(P - 2c) = 16A^2 = 16m^2P^2$$

or

$$(a+b-c)(a-b+c)(-a+b+c) = 16m^2(a+b+c). \tag{3}$$

Since the expressions $(a+b-c)$, $(a-b+c)$, and $(-a+b+c)$ are all even or all odd, (3) implies that all three must be even. Then, we may let

$$a+b-c = 2x, \quad a-b+c = 2y, \quad -a+b+c = 2z$$

to get

$$a = x+y, \quad b = x+z, \quad c = y+z, \quad \text{and} \quad a+b+c = 2(x+y+z). \tag{4}$$

Note that $a \le b \le c$ implies that $x \le y \le z$. With this substitution, (3) reduces to

$$xyz = 4m^2(x+y+z), \tag{5}$$

which is (1) with $n = 4m^2$. Hence, to solve $A = mP$, we need only solve (5) and then use (4) to find a, b, c. An interesting geometric interpretation of x, y, z can be found in [**6**] (although the notation will have to be adjusted somewhat).

Example. We find all 45 triangles satisfying $A = 3P$. (Solutions to $A = 2P$ can be found in [**5**].) We need to solve $xyz = 36(x+y+z)$.

Since $\lfloor\sqrt{3 \cdot 36}\rfloor = 10$, we limit ourselves to $x = 1, 2, \ldots, 10$. All solutions are given in the table on the next page. For instance, to find the solutions when $x = 6$, we have $y = w + 6$ and $z = \hat{w} + 6$ where $w\hat{w} = 72$ and $w \le \hat{w}$. So

$$(w, \hat{w}) = (1, 72), (2, 36), (3, 24), (4, 18), (6, 12), (8, 9)$$

and consequently

$$(y, z) = (7, 78), (8, 42), (9, 30), (10, 24), (12, 18), (14, 15).$$

Some Properties of Solutions

We close with some properties of the solutions to $A = mP$. A trigonometric proof of the first Corollary can be found in [**5**].

COROLLARY 1. *The resulting triangles for $x < 2m$, $x = 2m$, and $x > 2m$ are obtuse, right, and acute respectively.*

Proof. We show that $x < 2m$ if and only if $c^2 > a^2 + b^2$. Using $a = x+y$, $b = x+z$, $c = y+z$, the inequality $c^2 > a^2 + b^2$ reduces to $yz > x(x+y+z)$ which, in turn, reduces to $x < 2m$ if we use $xyz = 4m^2(x+y+z)$. The acute and right triangle cases are similar. ∎

The Theorem and Corollary 1 immediately imply

COROLLARY 2. *The Pythagorean triple $a = 4m + w$, $b = 4m + \hat{w}$, $c = 4m + w + \hat{w}$, where $w\hat{w} = 8m^2$ and $w \le \hat{w}$, satisfies $A = mP$. Moreover, every Pythagorean triple satisfying $A = mP$ is obtained in this fashion.*

Proof. Take $x = 2m$ in the main Theorem. ∎

The next Corollary settles a conjecture made in [**9**].

COROLLARY 3. *Of all the integer-sided triangles which satisfy $A = mP$, the largest (in area and hence perimeter) is attained by the triangle with sides $a = 4m^2 + 2$, $b = (4m^2+1)^2$, $c = 16m^4 + 12m^2 + 1$.*

x	y	z	a	b	c	P	x	y	z	a	b	c	P
1	37	1368	38	1369	1405	2812	3	15	72	18	75	87	180
1	38	702	39	703	740	1482	3	16	57	19	60	73	152
1	39	480	40	481	519	1040	3	17	48	20	51	65	136
1	40	369	41	370	409	820	3	18	42	21	45	60	126
1	42	258	43	259	300	602	3	21	32	24	35	53	112
1	45	184	46	185	229	460	3	22	30	25	33	52	110
1	48	147	49	148	195	392	3	24	27	27	30	51	108
1	54	110	55	111	164	330	4	10	126	14	130	136	280
1	72	73	73	74	145	292	4	12	48	16	52	60	128
2	19	378	21	380	397	798	4	18	22	22	26	40	88
2	20	198	22	200	218	440	5	8	117	13	122	125	260
2	21	138	23	140	159	322	5	9	56	14	61	65	140
2	22	108	24	110	130	264	6	7	78	13	84	85	182
2	23	90	25	92	113	230	6	8	42	14	48	50	112
2	24	78	26	80	102	208	6	9	30	15	36	39	90
2	26	63	28	65	89	182	6	10	24	16	30	34	80
2	27	58	29	60	85	174	6	12	18	18	24	30	72
2	28	54	30	56	82	168	6	14	15	20	21	29	70
2	30	48	32	50	78	160	7	8	27	15	34	35	84
2	33	42	35	44	75	154	7	10	18	17	25	28	70
2	36	38	38	40	74	152	8	9	17	17	25	26	68
3	13	192	16	195	205	416	8	12	12	20	20	24	64
3	14	102	17	105	116	238							

Proof. It suffices to maximize $x + y + z$ in the solution of (5). Let $k = \gcd(4m^2, x)$, $\overline{x} = x/k$, and $\overline{n} = 4m^2/k$. First note that if $r, s \geq 1$, then $(r-1)(s-1) \geq 0$, or, after rearranging, $r + s \leq rs + 1$. We employ this inequality a couple of times. For instance, if $w\hat{w} = \overline{n}(\overline{n} + \overline{x}x)$, we have

$$kw + k\hat{w} \leq k^2 w\hat{w} + 1 = k^2\overline{n}(\overline{n} + \overline{x}x) + 1 = 16m^4 + 4m^2x^2 + 1.$$

Now by the Theorem,

$$x + y + z = x + \frac{w + \overline{n}}{\overline{x}} + \frac{\hat{w} + \overline{n}}{\overline{x}} = x + \frac{8m^2}{x} + \frac{kw + k\hat{w}}{x}$$

$$\leq x + \frac{8m^2}{x} + \frac{16m^4 + 4m^2x^2 + 1}{x}$$

$$= (4m^2+1)\left(x + \frac{4m^2+1}{x}\right)$$
$$\leq (4m^2+1)(4m^2+2).$$

Now since $x = 1$, $y = 4m^2+1$, $z = 4m^2(4m^2+2)$ is a solution to (5) satisfying $x+y+z = (4m^2+1)(4m^2+2)$, we are done. ∎

From the Theorem and Corollary 1, it is clear that there are many obtuse and right triangle solutions if m has many divisors. A trivial lower bound is given by

COROLLARY 4. *Let $d(n)$ denote the number of positive divisors of n. There are $d(8m^2)/2$ right triangles satisfying $A = mP$. Further, for $m \neq 1, 2, 4$, there are at least $d(4m^2) + d(2m^2) + d(m^2) + d(4m)$ obtuse triangles satisfying $A = mP$.*

Proof. From the Theorem, taking $x = 2m$ (right triangles), every divisor w of $8m^2$ with $w \leq \sqrt{8m^2}$ yields a distinct solution. There are $d(8m^2)/2$ such divisors. For $x < 2m$ (obtuse triangles), we can take $x = 1, 2, 4, m$. Then each divisor of $\hat{x} = 4m^2, 2m^2, m^2, 4m$ produces a distinct solution. ∎

Thus obtuse and right triangle solutions to $A = mP$ can be as plentiful as desired. However, as remarked in [**5**], it appears that acute and isosceles triangles are rare among all solutions to $A = mP$; for instance, there is only one acute and no isosceles solution among the 80 solutions to $A = 7P$. We do however have the following curious construction.

COROLLARY 5. *For any positive integer N, there is an m such that at least N of the solutions to $A = mP$ are acute isosceles triangles.*

Proof. Let (r_i, s_i, t_i), $i = 1, 2, \ldots, N$ denote N Pythagorean triples with $r_i < s_i < t_i$, $s_i/r_i < \sqrt{3}$ and distinct values of s_i/r_i. (There are infinitely many primitive Pythagorean triples satisfying $|s_i/r_i - 1| < \varepsilon$ for any $\varepsilon > 0$. See, e.g., [**7**].) Put $m = \prod_{i=1}^{N} r_i s_i$ and $x_k = 2ms_k/r_k$ for $k = 1, 2, \ldots, N$. Then each x_k divides $4m^2$ and $2m < x_k < 2\sqrt{3}m$. If we let $\hat{x}_k = 4m^2/x_k$, then the Theorem guarantees that there is a solution for each divisor w of $\hat{x}_k(\hat{x}_k + x_k)$ such that $w \geq \max\{x_k - \hat{x}_k, 1\}$. Since

$$4m^2 = 4\prod_{i=1}^{N} r_i^2 s_i^2 \qquad \text{and} \qquad \hat{x}_k = 2r_k^2 \prod_{i \neq k} r_i s_i,$$

it follows that

$$\hat{x}_k(\hat{x}_k + x_k) = 4m^2 + \hat{x}_k^2 = 4r_k^2(r_k^2 + s_k^2)\prod_{i \neq k} r_i^2 s_i^2 = 4r_k^2 t_k^2 \prod_{i \neq k} r_i^2 s_i^2$$

and hence, $\hat{x}_k(\hat{x}_k + x_k)$ is a perfect square. Thus we may choose the solution of (2) for which $w = \sqrt{\hat{x}_k(\hat{x}_k + x_k)}$ and it is easily seen that $y_k = z_k$ (and therefore, $a = b$) in this case. Since each $x_k > 2m$, we get distinct isosceles solutions which are also acute for $i = 1, 2, \ldots, N$. ∎

The final result shows that there are few solutions if we reverse the role of A and P. An alternative proof can be found in [**10**].

COROLLARY 6. *The equation $P = mA$ has no solutions for $m > 2$. The only solution to $P = 2A$ is the $(3, 4, 5)$ triangle.*

Proof. If $P = mA$, a similar approach to above leads to the equation

$$m^2xyz = 4(x + y + z),$$

which can be rewritten

$$(m^2xy - 4)(m^2xz - 4) = 4(m^2x^2 + 4). \tag{6}$$

If $m \geq 2$, then $(m^2xy - 4)$ and $(m^2xz - 4)$ are non-negative and we get

$$4(m^2x^2 + 4) = (m^2xy - 4)(m^2xz - 4) \geq (m^2x^2 - 4)^2.$$

Since this simplifies to $mx \leq 2\sqrt{3}$, we must have $x = 1$ and $m = 2$ or 3. If $m = 2$, (6) reduces to

$$(y - 1)(z - 1) = 2$$

which yields the (3, 4, 5) triangle. If $m = 3$, (6) reduces to

$$(9y - 4)(9z - 4) = 52$$

which has no integral solutions. ■

Acknowledgment. This paper is a combination of two similar approaches developed separately by the two teams of authors in response to [**4**] and [**5**]. The authors wish to thank the editor for introducing the two groups and suggesting that we merge our submissions into a single work. Thanks are also extended to the anonymous referees whose suggestions improved the presentation of this paper.

REFERENCES

1. C. J. Bradley, *Challenges in Geometry*, Oxford University Press, Oxford, 2005.
2. L. Dickson, *History of the Theory of Numbers*, Vol. II, Dover Publications, Inc., New York, 2005 (reprint from the 1923 edition), p. 195.
3. J. Goehl, Area = k(perimeter), *Math. Teach.* **76** (1985) 330–332.
4. K. Korbin, Problem 4907, *School Science and Mathematics* **106** (2006) 106.
5. L. P. Markov, Pythagorean Triples and the Problem $A = mP$ for Triangles, this MAGAZINE **79** (2006) 114–121.
6. R. B. Nelsen, Proof without words: Padoa's Inequality, this MAGAZINE **79** (2006) 53.
7. I. Niven, H. Zuckerman and H. Montgomery, *An Introduction to the Theory of Numbers*, 5th ed., John Wiley and Sons, Inc., 1991, p. 234.
8. D. O. Shklarsky, N. N. Chentzov and I. M. Yaglom, *The USSR Olympiad Problem Book*, W. H. Freeman and Co., San Francisco, 1962.
9. D. Stone and J. Hawkins, Solution to Problem 4907, *School Science and Mathematics* **106** (2006).
10. M. Subbarao, Perfect Triangles, *Amer. Math. Monthly* **72** (1971) 384–385.

On the Number of Self-Avoiding Walks on Hyperbolic Lattices

JONATHAN D. BARRY
Penn State University, Beaver Campus
Monaca, PA 15061

C. CHRIS WU
Penn State University, Beaver Campus
Monaca, PA 15061

Imagine that you are standing at an intersection in a city where the street system is a square grid. You choose a street at random and begin walking away. At each intersection, you choose either to continue straight ahead or to turn left or right. There is only one rule: you must not return to any intersection you have already visited. In other words, your path should be self-avoiding. A fundamental question is: if you walk n blocks, how many possible paths could you have followed?

What we have described above is the self-avoiding walk on the square lattice $\mathcal{Z}^2$. Before given a specific definition of self-avoiding walks on general lattices, we first describe the lattices on which we will study self-avoiding walks.

Euclidean and Hyperbolic Lattices

Suppose that you want to remodel your kitchen floor with ceramic tiles. Suppose you want the tiles to be identical size and identical shape, with the shape being a regular polygon (that is, equal-angled and equilateral). Then you have only three choices of shapes to choose: the square, the triangle and the hexagon (see Figure 1). For example, pentagonal tiles cannot tile your kitchen floor. The three (infinite) lattices in Figure 1 are called the square lattice, the triangular lattice and the hexagonal lattice respectively. Each of these lattices can be characterized by two integers, both greater than two: v, the number of neighbors of each vertex (two vertices are called neighbors if there is an edge connecting them); and p, the number of sides of each polygon. We denote such a lattice by $\mathcal{G}(v, p)$. For example, the triangular lattice is denoted by $\mathcal{G}(6, 3)$. Translating the statement that the kitchen floor can be covered by square, triangular and hexagonal tiles into mathematical terms, we say that the Euclidean plane R^2 can be tessellated (covered without gaps or overlaps) by $\mathcal{G}(4, 4)$, $\mathcal{G}(6, 3)$, and $\mathcal{G}(3, 6)$. See Chapter 2 of Singer [6] for a detailed proof that these are the only regular tessellation for R^2. We call these lattices *Euclidean lattices.*

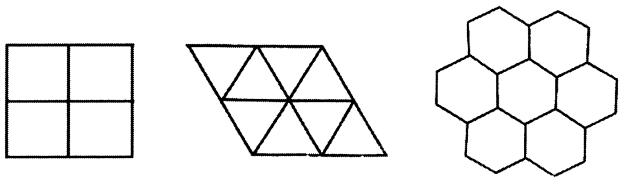


Figure 1

More generally, consider the tessellation of the following three 2-dimensional spaces: the sphere S^2, the Euclidean plane R^2, and the hyperbolic plane H^2. (For an

intuitive tutorial of the hyperbolic plane H^2, we recommend an undergraduate web site at `http://cs.unm.edu/~joel/NonEuclid/`, although it is not necessary to have any knowledge about H^2 in order to read this paper.) Then $\mathcal{G}(v, p)$ can tessellate S^2, R^2, or H^2 respectively if the quantity $(v-2)(p-2)$ is smaller than, equal to, or larger than 4 respectively. This can be seen as follows. One of the main differences between the geometries on S^2, R^2, or H^2 is that the sum of the angles in a triangle is greater than, equal to, or less than π. This implies that the interior angle θ of a regular p-polygon on S^2, R^2, or H^2 is respectively greater than, equal to, or less than $(p-2)\pi/p$. On the other hand, to fit v copies of a p-polygon together so that they share a common vertex, requires that $\theta v = 2\pi$, or $\theta = 2\pi/v$. Therefore, $2\pi/v$ is greater than, equal to, or less than $(p-2)\pi/p$ respectively on S^2, R^2 or H^2. This is equivalent to what needs to be shown. See Sections 2.2, 3.2, and 4.1 of Singer [**6**] for more details.

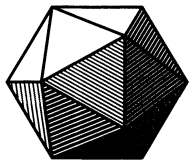
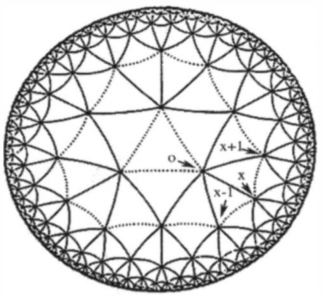


Figure 2 $G(5, 3)$, the logo of the MAA.

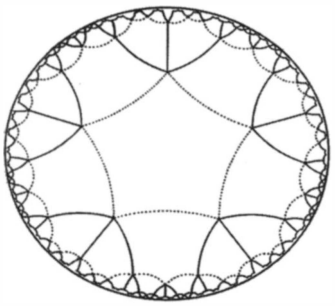
Recall that v and p are integers strictly greater than two, so there are exactly five pairs of (v, p) satisfying $(v-2)(p-2) < 4$ and, hence, that there are five regular tessellations on S^2, which are $\mathcal{G}(3, 3)$, $\mathcal{G}(3, 4)$, $\mathcal{G}(3, 5)$, $\mathcal{G}(4, 3)$, and $\mathcal{G}(5, 3)$—the five *Platonic solids*. For example, $\mathcal{G}(5, 3)$ is the logo of the Mathematical Association of America (Figure 2). There are exactly three pairs of (v, p) satisfying $(v-2)(p-2) = 4$. So, as we have seen previously, there are three lattices that tessellate R^2: $\mathcal{G}(4, 4)$, $\mathcal{G}(6, 3)$ and $\mathcal{G}(3, 6)$. By convention, we will write $\mathcal{Z}^2$ for the square lattice $\mathcal{G}(4, 4)$. When tessellating the hyperbolic plane H^2, there are infinitely many choices of (v, p) satisfying $(v-2)(p-2) > 4$. In this case the lattices are called *hyperbolic lattices* and will be denoted by $\mathcal{H}(v, p)$ (in stead of $\mathcal{G}(v, p)$). The hyperbolic lattices $\mathcal{H}(7, 3)$, $\mathcal{H}(5, 5)$ and $\mathcal{H}(6, 4)$ are shown in Figure 3. The famous *Circle Limit III* created by artist M. C. Escher and the 2003 Math Awareness Month poster are both related to hyperbolic tessellations. The later can be viewed at `http://www.mathaware.org/mam/03/`.

Self-avoiding Walks

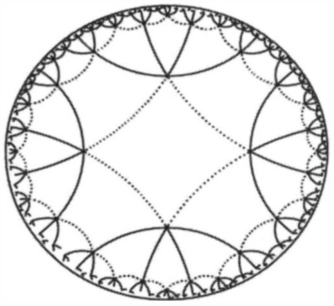
We now describe self-avoiding walks on the lattices defined above. An n-step self-avoiding walk ω on $\mathcal{H}(v, p)$, beginning at vertex x, is defined as a sequence of vertices $\omega(0), \omega(1), \ldots, \omega(n)$, where $\omega(0) = x$, $\omega(i)$ and $\omega(i+1)$ are neighbors for $0 \le i \le n-1$, and $\omega(i) \ne \omega(j)$ when $i \ne j$. The study of self-avoiding walks arose in chemical physics as a model for long polymer chains. Roughly speaking, a polymer is a macromolecule composed of a large number of smaller molecules, called monomers. The monomers are linked together randomly except that they cannot overlap: the presence of a monomer at position x prohibits any other part of the polymer from getting too close to x. This restriction is modelled by a self-repulsion term. The simplest mathematical model to state with such a self-repulsion term is the self-avoiding walk. It was originally introduced on $\mathcal{Z}^d$, the d-dimensional Euclidean lattice, where $d \ge 2$. But in recent years, the study on hyperbolic lattices of this model



(a) $H(7, 3)$



(b) $H(5, 5)$



(c) $H(6, 4)$

Figure 3

and other statistical mechanics models has received increasing attention from physicists and mathematicians. See references [5] and [7].

Let c_n denote the number of n-step self-avoiding walks beginning at vertex x. Notice that c_n is independent of x. Our basic question is to ask for the value of c_n. For example, on the square lattice $\mathcal{Z}^2$, $c_1 = 4$, $c_2 = 4 \times 3$, $c_3 = 4 \times 3^2$, and $c_4 = 4 \times 3^3 - 4 \times 2$. However, the combinatorics quickly becomes difficult as n increases and soon becomes intractable. It was proved in 1954 by Hammersley and Morton [2] that the limit

$$\mu = \lim_{n\to\infty} (c_n)^{1/n} \tag{1}$$

exists on any connected infinite lattice, where μ is called the *connective constant* of the self-avoiding walk. At the first glance, one may think (1) would imply

$$c_n \approx \mu^n, \tag{2}$$

where $a_n \approx b_n$ means $\lim_{n\to\infty} a_n/b_n = 1$. But this is not the case, since for any sequence g_n satisfying $\lim_{n\to\infty}(g_n)^{1/n} = 1$ (for instance, $g_n = e^{\sqrt{n}}$ or a polynomial of n), $c_n = g_n\mu^n$ satisfies (1). In fact, it is conjectured that for large n

$$c_n \approx An^{\gamma}\mu^n, \tag{3}$$

where $A > 0$ and $\gamma \geq 0$ are constants and γ is called a *critical exponent*. For self-avoiding walks on the Euclidean lattice $\mathcal{Z}^d$, Hara and Slade [3] have shown that (3) is valid and $\gamma = 0$ when $d > 4$; while for self-avoiding walks on the hyperbolic lattice $\mathcal{H}(v, p)$, (3) is also valid and $\gamma = 0$ when $(v - 5)(p - 2) > 4$, proved by Madras and Wu [4]. Therefore on these lattices, the growth of c_n is indeed given by $c_n \approx A\mu^n$.

The exact value of μ is not known for any lattice described in the last subsection, although for the hexagonal lattice there is nonrigorous argument showing that $\mu =$

$\sqrt{2+\sqrt{2}}$. Even obtaining a good numerical value of μ is very difficult. On the square lattice, Guttmann and Enting [**1**] showed in 1988 that $\mu \approx 2.6381585$. It is beyond the scope of this paper to study conjecture (3). The main purpose of this paper is to give a lower and upper bound for μ for the self-avoiding walk on the hyperbolic lattice $\mathcal{H}(v, p)$.

We first look at a trivial lower and upper bound for μ on the square lattice $\mathcal{Z}^2$. Starting from any given vertex, say the origin of $\mathcal{Z}^2$, the self-avoiding walk has 4 choices to move at the first step and at most 3 choices to move at each step after. So $c_n \leq 4 \times 3^{n-1}$. On the other hand, if the walk is restricted only to move eastward or northward at each step, then it is certainly self-avoiding, and at each step it has exactly 2 choices to move. So $c_n \geq 2^n$. Therefore we have

$$2^n \leq c_n \leq 4 \times 3^{n-1} \tag{4}$$

which together with (1) implies the following trivial bound of μ on $\mathcal{Z}^2$:

$$2 \leq \mu \leq 3. \tag{5}$$

An upper bound for μ on the hyperbolic lattice $\mathcal{H}(v, p)$ is also easy. Starting at a given vertex, the walk has v choices to move at the first step since each vertex has v neighbors. It then has at most $v-1$ choices to move at each step after. So $c_n \leq v(v-1)^{n-1}$ and hence

$$\mu \leq v - 1. \tag{6}$$

A crude lower bound for μ is given in Proposition 1. We then improve it in Proposition 2. We want to demonstrate that much more work is required to get the improved lower bound.

PROPOSITION 1. *For the self-avoiding walk on the hyperbolic lattice $\mathcal{H}(v, p)$, we have*

(a) *for $p = 3$ (that is, when the faces are triangles), $v - 1 \geq \mu \geq v - 4$,*

(b) *for $p \geq 4$, $v - 1 \geq \mu \geq v - 3$.*

PROPOSITION 2. *For the self-avoiding walk on the hyperbolic lattice $\mathcal{H}(v, p)$, we have*

(a) *for $p = 3$, $v - 1 \geq \mu \geq \sqrt{(v-2)(v-3)}$,*

(b) *for $p = 4$, $v - 1 \geq \mu \geq \sqrt[3]{(v-1)(v-2)^2}$,*

(c) *for $p > 4$, $v - 1 \geq \mu \geq \sqrt{(v-1)(v-2)}$. More generally, $v - 1 \geq \mu \geq ((v-1)^{p-4}(v-2))^{1/(p-3)}$.*

REMARK. *Since $\lim_{p\to\infty}((v-1)^{p-4}(v-2))^{1/(p-3)} = v - 1$, part (c) of Proposition 2 implies that μ approaches $v - 1$ as $p \to \infty$.*

In light of the remark, we see that although the upper bound given in (6) is so easily obtained, it is actually not bad, and it will be difficult to improve it, especially for large p.

It is not easy to carry out numerical calculation for μ if v or p is large. On the other hand, from the remark, part (c) of Proposition 2 gives a good bound for μ for large p. For example, on $\mathcal{H}(5, 30)$, we have $4 \geq \mu \geq 3.9576$. In fact, even for small v and p, the only numerical result we are aware is on the graph $\mathcal{H}(7, 3)$ and $\mathcal{H}(5, 5)$, which was obtained by Swierczak and Guttmann [**7**] in 1996. For example, on $\mathcal{H}(5, 5)$, the numerical calculation of Swierczak and Guttmann is that $\mu \approx 3.9746$, while our rigorous bound gives $4 \geq \mu \geq \sqrt{12} = 3.4641$.

Proof of propositions. The hyperbolic lattice $\mathcal{H}(v, p)$ is both *vertex-transitive* and *polygon-transitive*. This means, informally speaking, each vertex plays the same role as all the other vertices, and each polygon plays the same role as all the other polygons. If you stand on two different polygonal-tiles (respectively, vertices) and views the entire lattice, you cannot see any difference. We single out one polygon, say the central one in the figure, and call it the *original polygon*. Call one of its p vertices the *origin* or *root*.

To prove Propositions 1 and 2, we need to partition the vertices of $\mathcal{H}(v, p)$ into different layers. Let L_1 be the original polygon. Define L_{n+1} inductively as the union of L_n and the polygons that have a vertex in common with L_n. See Figure 3 where the boundary of L_1 and L_2 is drawn in dotted edges. Denote by ∂L_n the boundary of L_n. Then $L_{n+1} - (L_n - \partial L_n)$ consists of those polygons that have either two vertices and one edge in common with ∂L_n or one vertex and no edge in common with ∂L_n. We call the set of vertices on ∂L_n the nth layer of vertices.

Proof of Proposition 1. The upper bound $v - 1$ is given in (6). We only need to show the lower bound. For $p \geq 4$, each vertex on any layer has at least $v - 3$ neighbors in the next layer. Starting at a given vertex x, if the walk chooses to move only to the next layer, then it has at least $v - 3$ choices at each step and the resulting path is certainly self-avoiding. So $c_n \geq (v - 3)^n$, which implies $\mu \geq v - 3$. For $p = 3$, the same argument shows $\mu \geq v - 4$, because each vertex on any layer has at least $v - 4$ neighbors in the next layer. ∎

A sketch of a proof for part (c) of Proposition 2 can be found in reference [**4**]. We give a proof for all three parts here.

Proof of Proposition 2. We first prove part (c). Again, we only need to show the lower bound. For a given vertex x on the nth layer ∂L_n, denote by $x + i$ (respectively, $x - i$) the ith vertex counting from x along ∂L_n counter-clockwise (respectively, clockwise). For $p > 4$, each vertex x on any layer other than the first one has either zero or one neighbor in the previous layer, which we will call the parent of x and denote it by $\tilde{x}$. See Figure 3(b). Starting from a vertex in the first layer, define a walk according to the following rules. At the first step, the walk is allowed to move to a neighbor vertex in the next layer; so it has $v - 2$ choices to do so. Each time when the walk reaches a vertex x on a new layer, it is allowed to have $v - 1$ choices to move at its next step: $v - 3$ choices to move to the next layer and two choices to move within the same layer—to $x + 1$ or $x - 1$. Notice that $x + 1$ and $x - 1$ have no neighbor in the previous layer and $v - 2$ neighbors in the next layer. (In fact, each time when the walk reaches a vertex x on a new layer ∂L_n, vertices $x + i$ and $x - i$, for $i = 1, 2, \ldots, p - 4$, have no neighbor in the previous layer ∂L_{n-1} and $v - 2$ neighbors in the next layer ∂L_{n+1}. The next paragraph gives an explanation.) If the walk reaches $x + 1$ (respectively, $x - 1$), then from there it is allowed to have $v - 2$ choices to move—to the next layer. The walk defined above is self-avoiding and has the property that at each step the walk has either $v - 1$ or $v - 2$ choices to move and there are no two consecutive steps at which the walk has $v - 2$ choices to move; that is, if the walk has $v - 2$ choices to move at a step, then it must have $v - 1$ choices to move at the previous (and the next) step. It follows from this property that

$$c_{2n} \geq \overbrace{(v-2)(v-1)(v-2)(v-1)\cdots(v-2)(v-1)}^{2n} = (v-2)^n(v-1)^n.$$

Therefore

$$\mu = \lim_{n\to\infty} (c_n)^{1/n} = \lim_{n\to\infty} (c_{2n})^{1/(2n)} \geq \sqrt{(v-2)(v-1)},$$

where the last inequality is from substituting the lower bound $(v-2)^n(v-1)^n$ for c_{2n}. This proves the first statement of part (c).

In order to prove the second statement of part (c), we need more detailed observations. As stated in the previous paragraph, each time when the walk reaches a vertex x on a new layer ∂L_n, vertices $x+i$ and $x-i$, for $i = 1, 2, \ldots, p-4$, have no neighbor in the previous layer ∂L_{n-1} and $v-2$ neighbors in the next layer ∂L_{n+1}. This is because the unique polygon with $\tilde{x}$, x and $x+1$ as part of its vertices has either two vertices (and one edge) in ∂L_{n-1} and $p-2$ vertices in ∂L_n, or one vertex (and no edge) in ∂L_{n-1} and $p-1$ vertices in ∂L_n. In the former case, $x+i$ for $i = 1, 2, \ldots, p-4$ has no neighbor in ∂L_{n-1} (and x and $x+(p-3)$ have one neighbor in ∂L_{n-1}). In the later case, $x+i$ for $i = 1, 2, \ldots, p-3$ has no neighbor in ∂L_{n-1} (and x and $x+(p-2)$ have one neighbor in ∂L_{n-1}). Therefore, in any case, $x+i$ for $i = 1, 2, 3, \ldots, p-4$ has no neighbor in ∂L_{n-1}. A similar argument shows the statement for $x-i$ for $i = 1, 2, \ldots, p-4$.

Now, modify the rules of the walk defined in the proof of the first statement of part (c) as follows. Starting at a vertex in the first layer, the walk is allowed to move to a neighbor vertex in the next layer; so it has $v-2$ choices to do so. Each time when the walk reaches a vertex x on a new layer, it is allowed to have $v-1$ choices to move at its next step: $v-3$ choices to move to the next layer and two choices to move within the same layer—to $x+1$ or $x-1$. If the walk reaches $x+1$ (respectively, $x-1$), then from there it is allowed to have $v-1$ choices to move: $v-2$ choices to the next layer and one choice to vertex $x+2$ (respectively, $x-2$). In general, if the walk reaches vertex $x+(i-1)$ (respectively, $x-(i-1)$) for $i = 2, 3, \ldots, (p-4)$, then it is allowed to have $v-1$ choices to move at the next step: $v-2$ choices to move to the next layer and one choice to move to vertex $x+i$ (respectively, $x-i$). When the walk reaches vertex $x+(p-4)$ (respectively, $x-(p-4)$), then at the next step it is only allowed to move to the next layer—$v-2$ choices.

The walk defined in the last paragraph is self-avoiding and has the property that if at a step it has $v-2$ choices to move, then it must have $v-1$ choices to move at the next $p-4$ steps. It follows that $c_{(p-3)n} \geq ((v-2)(v-1)^{p-4})^n$. Therefore

$$\mu = \lim_{n\to\infty} c_n^{1/n} = \lim_{n\to\infty} (c_{(p-3)n})^{1/((p-3)n)} \geq \sqrt[p-3]{(v-2)(v-1)^{p-4}}.$$

This proves part (c).

For the proof of part (b) where $p = 4$, first note that as in the case $p > 4$, each vertex on any layer other than the first one has either zero or one neighbor in the previous layer. See Figure 3(c). Starting at a vertex in the first layer, define a walk according to the following rules. At the first step, the walk is allowed to move to a neighbor vertex in the next layer; so it has $v-2$ choices to do so. Each time when the walk reaches a vertex x on a new layer ∂L_n, it is allowed to have $v-1$ choices to move at its next step: $v-3$ choices to move to the next layer and two choices to move within the same layer—to $x+1$ or $x-1$. Recall that the polygon containing $\tilde{x}$, x, and $x+1$ as part of its vertices has either one vertex and no edge in ∂L_{n-1} or two vertices and one edge in ∂L_{n-1}. In the former case, $x+1$ has no neighbor in ∂L_{n-1}. In the later case, $x+1$ has one neighbor in ∂L_{n-1} but $x+2$ has no neighbor in ∂L_{n-1}. Therefore, either

(A) $x+1$ (respectively, $x-1$) has no neighbor in the previous layer, or

(B) $x+1$ (respectively, $x-1$) has one neighbor in the previous layer but $x+2$ (respectively, $x-2$) has no neighbor in the previous layer.

If the walk reaches $x+1$ (respectively, $x-1$), then in case (A), it is allowed to have $v-2$ choices to move to the next layer. In case (B), it is also allowed to have $v-2$

choices to move: $v-3$ choices to move to the next layer and one choice to move to $x+2$ (respectively, $x-2$). If the walk reaches $x+2$ (respectively, $x-2$), then from there it is allowed to have $v-2$ choices to move to the next layer.

The walk defined in the last paragraph is self-avoiding, and at each step it has either $v-1$ or $v-2$ choices to move. Furthermore, there are no three consecutive steps at which the walk has $v-2$ choices to move, that is, if it has $v-2$ choices to move at steps k and $k+1$, then at step $k-1$ it must have $v-1$ choices to move. Therefore $c_{1+3n} \geq (v-2)((v-1)(v-2)^2)^n$ which implies that

$$\mu = \lim_{n\to\infty} (c_n)^{1/n} = \lim_{n\to\infty} (c_{1+3n})^{1/(1+3n)} \geq \sqrt[3]{(v-1)(v-2)^2}.$$

To prove part (a), notice that for $p=3$, each vertex x on any layer other than the first one has either one or two neighbors in the previous layer, which we call the parent(s) of x. So in other words, each vertex x in any layer other than the first one has either one or two parents. Starting at a vertex, say O, the origin, in the first layer, define a walk according to the following rules. At the first step, the walk is allowed to move to a neighbor vertex in the next layer; so it has $v-2$ choices to do so. Each time when the walk reaches a vertex x on a new layer, as just explained, one of the two cases occurs:

(A) x has two parents, or

(B) x has only one parent.

In case (A), the walk is allowed to have $v-2$ choices to move from x: $v-4$ choices to move to the next layer and two choices to move within the same layer—to $x+1$ or $x-1$. Notice that $x+1$ and $x-1$ have only one parent and have $v-3$ neighbors in the next layer. If the walk reaches $x+1$ (respectively, $x-1$), then from there it is allowed to have $v-3$ choices to move—to the next layer. Therefore, in the two steps after x, the walk has at least

$$(v-2)(v-3) \tag{7}$$

choices to move.

In case (B), one of the following four sub-cases occurs:

(B1) both $x+1$ and $x-1$ have only one parent,

(B2) $x-1$ has two parents and $x+1$ has one parent,

(B3) $x-1$ has one parent and $x+1$ has two parents, or

(B4) both $x+1$ and $x-1$ have two parents.

In case (B1), the walk is allowed to have $v-1$ choices to move from x (see Figure 3(a)): $v-3$ choices to move to the next layer and two choices to move to $x+1$ or $x-1$. If the walk moves to $x+1$ (respectively, $x-1$), then from there it is allowed to have $v-3$ choices to move to the next layer. Therefore, in the two steps after x, the walk has at least

$$(v-1)(v-3) > (v-2)(v-3) \tag{8}$$

choices to move.

In case (B2), the walk is allowed to have $v-2$ choices to move from x: $v-3$ choices to move to the next layer and one choice to move to $x+1$ (it is not allowed to move to $x-1$). If the walk moves to $x+1$, then it is allowed to have $v-3$ choices to move from there to the next layer. Therefore, in the two steps after x, the walk has at

least

$$(v-2)(v-3) \tag{9}$$

choices to move.

Case (B3) is similar to case (B2) but with $x+1$ and $x-1$ interchanged.

In case (B4), the walk is only allowed to have $v-3$ choices to move to the next layer. But note that case (B4) can only occur when $v=7$. And even for $v=7$, it only occurs if the parent of x (call it $\tilde{x}$) has two parents. This means that the walk at vertex $\tilde{x}$ (which belongs to case (A)) has $v-2$ choices to move. Therefore, in the two steps around x—one before x and another after x, the walk has

$$(v-2)(v-3) \tag{10}$$

choices to move.

The walk defined above is self-avoiding and has the following properties:

(1) at each step it has either $v-3$ choices to move or at least $v-2$ choices to move (at some step it has $v-1$ choices to move), and

(2) there are no two consecutive steps at which the walk has $v-3$ choices to move; that is, if the walk has $v-3$ choices to move at a step, then it must have at least $v-2$ choices to move at the previous step.

Therefore, from (7)—(10), we have that $c_{2n} \geq (v-2)^n(v-3)^n$, which implies the lower bound in part (a). ■

Acknowledgments. We thank the former editor, Dr. Frank A. Farris, and two anonymous referees for a careful reading of the manuscript and for detailed suggestions that helped to improve the presentation of the paper. In addition, we thank Dr. Farris for the extensive markings about matters of style that he made on the manuscript. The material of this paper is from talks given in the Undergraduate Research Fair at Penn State University, Beaver Campus in March of 2002 and in the Regional Campus Mathematics Conference at Kent State University in May of 2002. We thank the organizers of these meetings. The work of C. C. W. was partially supported by NSF Grants DMS-01-03994 and DMS-05-05484.

REFERENCES

1. A. Guttmann and I. Enting, The size and number of rings on the square lattice, *J. Phys. A.* Math. Gen. **21** (1988) L165–L172.
2. J. Hammersley and K. Morton, Poor man's Monte Carlo, *J. Roy. Stat. Soc. B* **16** (1954) 23–38.
3. T. Hara and G. Slade, Self-avoiding walk in five or more dimensions I: the critical behavior, Commun. Math. Phys. **147** (1992) 101–136.
4. N. Madras and C. Wu, Self-avoiding walks on hyperbolic graphs, In preparation.
5. R. Rietman, B. Nienhuis, and J. Oitmaa, The Ising model on hyperlattices, *J. Phys. A.* Math. Gen. **25** (1992) 6577–6592.
6. D. A. Singer, *Geometry: Plane and Fancy,* Springer, New York, (1998).
7. E. Swierczak and A. Guttmann, Self-avoiding walks and polygons on non-Euclidean lattices. *J. Phys. A.* Math. Gen. **29** (1996) 7485–7500.

Uncountable Sets and an Infinite Real Number Game

MATTHEW H. BAKER
Georgia Institute of Technology
Atlanta, GA 30332-0160
mbaker@math.gatech.edu

The Game

Alice and Bob decide to play the following infinite game on the real number line. A subset S of the unit interval $[0, 1]$ is fixed, and then Alice and Bob alternate playing real numbers. Alice moves first, choosing any real number a_1 strictly between 0 and 1. Bob then chooses any real number b_1 strictly between a_1 and 1. On each subsequent turn, the players must choose a point strictly between the previous two choices. Equivalently, if we let $a_0 = 0$ and $b_0 = 1$, then in round $n \geq 1$, Alice chooses a real number a_n with $a_{n-1} < a_n < b_{n-1}$, and then Bob chooses a real number b_n with $a_n < b_n < b_{n-1}$. Since a monotonically increasing sequence of real numbers that is bounded above has a limit (see [**8**, Theorem 3.14]), $\alpha = \lim_{n\to\infty} a_n$ is a well-defined real number between 0 and 1. Alice wins the game if $\alpha \in S$, and Bob wins if $\alpha \notin S$.

Countable and Uncountable Sets

A non-empty set X is called *countable* if it is possible to list the elements of X in a (possibly repeating) infinite sequence $x_1, x_2, x_3, \ldots$ Equivalently, X is countable if there is a surjective function from the set $\{1, 2, 3, \ldots\}$ of natural numbers onto X. The empty set is also deemed to be countable. For example, every finite set is countable, and the set of natural numbers is countable. A set that is not countable is called *uncountable*. Cantor proved using his famous *diagonalization argument* that the real interval $[0, 1]$ is uncountable. We will give a different proof of this fact based on Alice and Bob's game.

PROPOSITION 1. *If S is countable, then Bob has a winning strategy.*

Proof. The conclusion is immediate if $S = \emptyset$. Otherwise, since S is countable, one can enumerate the elements of S as $s_1, s_2, s_3, \ldots$. Consider the following strategy for Bob. On move $n \geq 1$, he chooses $b_n = s_n$ if this is a legal move, and otherwise he randomly chooses any allowable number for b_n. For each n, either $s_n \leq a_n$ or $s_n \geq b_n$. Since $a_n < \alpha < b_n$ for all n, we conclude that $\alpha \notin S$. This means that Bob always wins with this strategy! ∎

If $S = [0, 1]$, then clearly Alice wins no matter what either player does. Therefore we deduce:

COROLLARY 1. *The interval $[0, 1] \subset \mathbb{R}$ is uncountable.*

This argument is in many ways much simpler than Cantor's original proof!

Perfect Sets

We now prove a generalization of the fact that [0, 1] is uncountable. This will also follow from an analysis of our game, but the analysis is somewhat more complicated. Given a subset X of [0, 1], we make the following definitions:

- A *limit point* of X is a point $x \in [0, 1]$ such that for every $\epsilon > 0$, the open interval $(x - \epsilon, x + \epsilon)$ contains an element of X other than x.
- X is *perfect* if it is non-empty* and equal to its set of limit points.

For example, the famous middle-third *Cantor set* is perfect (see [**8**, §2.44]). If $L(X)$ denotes the set of limit points of X, then a nonempty set X is closed $\Leftrightarrow L(X) \subseteq X$, and is perfect $\Leftrightarrow L(X) = X$. It is a well-known fact that every perfect set is uncountable (see [**8**, Theorem 2.43]). Using our infinite game, we will give a different proof of this fact. We recall the following basic property of the interval [0, 1]:

($\star$) Every non-empty subset $X \subseteq [0, 1]$ has an *infimum* (or *greatest lower bound*), meaning that there exists a real number $\gamma \in [0, 1]$ such that $\gamma \leq x$ for every $x \in X$, and if $\gamma' \in [0, 1]$ is any real number with $\gamma' \leq x$ for every $x \in X$, then $\gamma' \leq \gamma$. The infimum γ of X is denoted by $\gamma = \inf(X)$.

Let's say that a point $x \in [0, 1]$ is *approachable from the right*, denoted $x \in X^+$, if for every $\epsilon > 0$, the open interval $(x, x + \epsilon)$ contains an element of X. We can define *approachable from the left* (written $x \in X^-$) similarly using the interval $(x - \epsilon, x)$. It is easy to see that $L(X) = X^+ \cup X^-$, so that a non-empty set X is perfect $\Leftrightarrow X = X^+ \cup X^-$.

The following two lemmas tell us about approachability in perfect sets.

LEMMA 1. *If S is perfect, then* $\inf(S) \in S^+$.

Proof. The definition of the infimum in ($\star$) implies that $\inf(S)$ cannot be approachable from the left, so, being a limit point of S, it must be approachable from the right. ■

LEMMA 2. *If S is perfect and $a \in S^+$, then for every $\epsilon > 0$, the open interval $(a, a + \epsilon)$ also contains an element of S^+.*

Proof. Since $a \in S^+$, we can choose three points $x, y, z \in S$ with $a < x < y < z < a + \epsilon$. Since $(x, z) \cap S$ contains y, the real number $\gamma = \inf((x, z) \cap S)$ satisfies $x \leq \gamma \leq y$. If $\gamma = x$, then by ($\star$) we have $\gamma \in S^+$. If $\gamma > x$, $\gamma \in S = L(S)$ and $(x, \gamma) \cap S = \emptyset$, so that $\gamma \notin S^-$ and therefore $\gamma \in S^+$. ■

From these lemmas, we deduce:

PROPOSITION 2. *If S is perfect, then Alice has a winning strategy.*

Proof. Alice's only constraint on her nth move is that $a_{n-1} < a_n < b_{n-1}$. By induction, it follows from Lemmas 1 and 2 that Alice can always choose a_n to be an element of $S^+ \subseteq S$. Since S is closed, $\alpha = \lim a_n \in S$, so Alice wins! ■

From Propositions 1 and 2, we deduce:

COROLLARY 2. *Every perfect set is uncountable.*

*Some authors consider the empty set to be perfect.

Further Analysis of the Game

We know from Proposition 1 that Bob has a winning strategy if S is countable, and it follows from Proposition 2 that Alice has a winning strategy if S contains a perfect set. (Alice just chooses all of her numbers from the perfect subset.) What can one say in general? A well-known result from set theory [**1**, §6.2, Exercise 5] says that every uncountable *Borel set*[†] contains a perfect subset. Thus we have completely analyzed the game when S is a Borel set: Alice wins if S is uncountable, and Bob wins if S is countable. However, there do exist non-Borel uncountable subsets of [0, 1] which do not contain a perfect subset [**1**, Theorem 6.3.7]. So we leave the reader with the following question (to which the author does not know the answer):

Question. Do there exist uncountable subsets of [0, 1] for which: (a) Alice does not have a winning strategy; (b) Bob has a winning strategy; or (c) neither Alice nor Bob has a winning strategy?

Related games. Our infinite game is a slight variant of the one proposed by Jerrold Grossman and Barry Turett in [**2**] (see also [**6**]). Propositions 1 and 2 above were motivated by parts (a) and (c), respectively, of their problem. The author originally posed Propositions 1 and 2 as challenge problems for the students in his Math 25 class at Harvard University in Fall 2000.

A related game (the "Choquet game") can be used to prove the Baire category theorem (see [**5**, §8.C] and [**3**, p. 22]). In Choquet's game, played in a given metric space X, Pierre moves first by choosing a non-empty open set $U_1 \subseteq X$. Then Paul moves by choosing a non-empty open set $V_1 \subseteq U_1$. Pierre then chooses a non-empty open set $U_2 \subseteq V_1$, and so on, yielding two decreasing sequences U_n and V_n of non-empty open sets with $U_n \supseteq V_n \supseteq U_{n+1}$ for all n, and $\bigcap U_n = \bigcap V_n$. Pierre wins if $\bigcap U_n = \emptyset$, and Paul wins if $\bigcap U_n \neq \emptyset$. One can show that if X is *complete* (i.e., if every Cauchy sequence converges in X), then Paul has a winning strategy, and if X contains a nonempty open set O that is a countable union of closed sets having empty interior, then Pierre has a winning strategy. As a consequence, one obtains the *Baire category theorem:* If X is a complete metric space, then no nonempty open subset of X can be a countable union of closed sets having empty interior.

Another related game is the Banach-Mazur game (see [**7**, §6] and [**5**, §8.H]). A subset S of the unit interval [0, 1] is fixed, and then Anna and Bartek alternate play. First Anna chooses a closed interval $I_1 \subseteq [0, 1]$, and then Bartek chooses a closed interval $I_2 \subseteq I_1$. Next, Anna chooses a closed interval $I_3 \subseteq I_2$, and so on. Together the players' moves determine a nested sequence I_n of closed intervals. Anna wins if $\bigcap I_n$ has at least one point in common with S, otherwise Bartek wins. It can be shown (see Theorem 6.1 of [**7**]) that Bartek has a winning strategy if and only if S is *meagre*. (A subset of X is called *nowhere dense* if the interior of its closure is empty, and is called *meagre*, or of the *first category*, if it is a countable union of nowhere dense sets.) It can also be shown, using the axiom of choice, that there exist sets S for which the Banach-Mazur game is undetermined (i.e., neither player has a winning strategy).

For a more thorough discussion of these and many other *topological games*, we refer the reader to the survey article [**9**], which contains an extensive bibliography. Many of the games discussed in [**9**] are not yet completely understood.

Games like the ones we have been discussing play a prominent role in the modern field of *descriptive set theory*, most notably in connection with the *axiom of determinacy* (AD). (See Chapter 6 of [**4**] for a more detailed discussion.) Let X be a given

[†]A Borel set is, roughly speaking, any subset of [0, 1] that can be constructed by taking countably many unions, intersections, and complements of open intervals; see [**8**, §11.11] for a formal definition.

subset of the space ω^ω of infinite sequences of natural numbers, and consider the following game between Alice and Bob. Alice begins by playing a natural number, then Bob plays another (possibly the same) natural number, then Alice again plays a natural number, and so on. The resulting sequence of moves determines an element $x \in \omega^\omega$. Alice wins if $x \in X$, and Bob wins otherwise. The axiom of determinacy states that this game is determined (i.e., one of the players has a winning strategy) for *every* choice of X.

A simple construction shows that the axiom of determinacy is inconsistent with the axiom of choice. On the other hand, with Zermelo-Fraenkel set theory plus the axiom of determinacy (ZF+AD), one can prove many non-trivial theorems about the real numbers, including: (i) every subset of $\mathbb{R}$ is Lebesgue measurable; and (ii) every uncountable subset of $\mathbb{R}$ contains a perfect subset. Although ZF+AD is not considered a "realistic" alternative to ZFC (Zermelo-Fraenkel + axiom of choice), it has stimulated a lot of mathematical research, and certain variants of AD are taken rather seriously. For example, the axiom of *projective determinacy* is intimately connected with the continuum hypothesis and the existence of large cardinals (see **[10]** for details).

Acknowledgment. The author was supported by NSF Grant DMS-0300784.

REFERENCES

1. K. Ciesielski, *Set Theory for the Working Mathematician*, London Mathematical Society Student Texts **39**, Cambridge University Press, 1997.
2. J. W. Grossman and B. Turett, Problem #1542, this MAGAZINE **71** (1998) 67.
3. F. Hirsch and G. Lacombe, *Elements of Functional Analysis*, Graduate Texts in Mathematics **192**, Springer-Verlag, 1999.
4. A. Kanamori, *The Higher Infinite* (2nd ed.), Springer-Verlag, 2003.
5. A. Kechris, *Classical Descriptive Set Theory*, Springer-Verlag, 1995.
6. W. A. Newcomb, Solution to Problem #1542, this MAGAZINE **72** (1999) 68–69.
7. J. Oxtoby, *Measure and Category* (2nd ed.), Springer-Verlag, 1980.
8. W. Rudin, *Principles of Mathematical Analysis* (3rd ed.), McGraw-Hill, 1976.
9. R. Telgársky, Topological games: On the 50th anniversary of the Banach-Mazur game, *Rocky Mountain J. Math.* **17** (1987) 227–276.
10. H. Woodin, The Continuum Hypothesis, Part I, *Notices of the AMS* **48** (2001) 567–576.

When Multiplication Mixes Up Digits

DAVID WOLFE
Gustavus Adolphus College
Saint Peter, MN 56082
davidgameswolfe@gmail.com

My daughter, Lila, has been learning about counting, and I wrote out the digits 1 through 9 on our whiteboard. She asked, "What number is that?". "Why that's 123 million, 456 thousand, 789." "That's a very big number. Can you make a bigger one?" I doubled the number, and got:

$$123456789 \times 2 = 246913578$$

Wow! The product has the same digits 1 through 9 reordered. Before long, I found myself doubling it over and over again:

$$123456789 \times 2 = 0246913578$$
$$246913578 \times 2 = 0493827156$$
$$493827156 \times 2 = 0987654312$$
$$987654312 \times 2 = 1975308624$$
$$1975308624 \times 2 = 3950617248$$
$$3950617248 \times 2 = 7901234496 \text{ (first exception)}$$

Notice that every product is *pandigital,* until the last product which has two 4's. A pandigital number in base b contains all the base-b digits. The literature varies about whether 0 needs to be one of the digits and whether digits may appear multiple times. Here we say a number is pandigital if it contains either all the digits *or* all the non-zero digits, with no digit repeated.

The fact that multiples of 123456789 and 987654321 are pandigital has long been observed. See, for example, David Wells, *The Penguin Dictionary of Curious and Interesting Numbers,* or the *The Nine Digits* web page.

Why So Many Pandigital Multiples?

It turns out that the doubling process is a red herring, for lots of multiples of 123456789 are pandigital. If you list the numbers under 10 which, when multiplied by 123456789, are pandigital, you find

$$123456789 \times 2 = 246913578$$
$$123456789 \times 4 = 493827156$$
$$123456789 \times 5 = 617283945$$
$$123456789 \times 7 = 864197523$$
$$123456789 \times 8 = 987654312$$

The multipliers are all the single digit numbers which do not have a prime factor of 3.

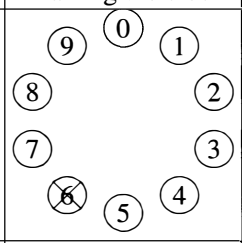
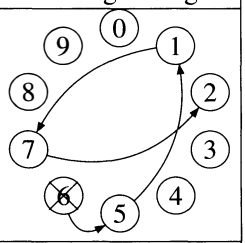
To investigate what exactly is going on, let's generalize the question to base b:

THEOREM 1. *Let x be the base-b number* $123\ldots(b-1)$, *and choose an n between* 1 *and b. The product $n \cdot x$ is pandigital if and only if $b-1$ and n are relatively prime.*

In particular, in base $b = 10$, we have $b - 1 = 9$ which has a single prime factor 3. The theorem says that for values of n which *don't* have a factor of 3, i.e., when n is 2, 4, 5, 7 , or 8, multiplication by n results in a pandigital product.

Using a diagram, we can compute the product another way by "walking around a clock," and in so doing shed light on the theorem. After describing this clock method, we'll see why it explains the theorem, and then why the clock method correctly computes the product.

Before explaining the clock method, we'll walk through an example of multiplying 123456789 by $n = 4$. On the left below is the usual method for multiplying we learned in grade school:

Grade school method	Drawing the clock	Counting off digits
carries: 01122333 x: 123456789 n: x 4		
product: 493827156	56	27156

The last two digits of the product are $b - n$ (in the example, $b - n = 6$) and $b - n - 1$ (which is 5). To see why, note that the last two digits of x are $b - 1$ and $b - 2$. The last digit of the product is generated by multiplying $n(b - 1) = bn - n = b(n - 1) + (b - n)$, i.e., the last digit will be $b - n$ with a carry of $n - 1$. The second to last digit is generated by the product plus carry $(b - 2)n + (n - 1) = b(n - 1) + (b - n - 1)$, i.e., the second to last digit is $b - n - 1$, with a carry of $n - 1$.

For the clock method, first write the digits 0 through $b - 1$ in a circle. Write down the last two digits of the product $b - n$ and $b - n - 1$ (again, 56 in the example). Cross out the $b - n = 6$. Now, beginning with $b - n - 1 = 5$, count counter-clockwise $n = 4$ positions around the circle to read off the digits in the product. In the example, you'll go from 5 to 1 to 7 to 2 to 8 to 3 to 9 to 4 to 0. Always skip past the crossed out $b - n = 6$ without counting it. Upon recording the 9-digit product, stop; the tenth digit will be 0.

This process hits all the non-zero digits if and only if $b - 1$ (the number of digits *not* crossed out) and n share no common factors. (Otherwise, you'd repeat digits as you used the clock method.)

To see why this alternative way of computing the product works, let us compare what happens when you multiply 123456789 by small numbers like 4 by both the clock method and the grade school method. First, look at the carries. The carry from the last digit is $n - 1 = 3$, and the carries stay the same or decrease proceeding leftward from digit to digit. In our example on the left, $123456789 \cdot 4$, the digit-products with carries are, in order from *right to left,*

$$36, 35, 31, 27, 22, 18, 13, 9, 4$$

Suppose while computing the product by the grade school method, we forgot to carry. Since (working from right to left) the digits of 123456789 decrease by 1, each digit's product by $n = 4$ would decrease by $n = 4$. Reintroducing the carries, since the rightmost digit has no carry but generates a carry of $n - 1 = 3$, the last 2 digits of the product will differ by $n - (n - 1) = 1$. Doing arithmetic mod b, proceeding from right to left, the digits-products decrease by 1 and then by either n or $n + 1$ depending on whether the carry stayed the same or decreased. Further, the carry decreases when the ten's (or, in general, b's) digit decreases and the units digit increases.

This brings us to why we cross out the 6. Returning to the circle of digits, counting by 4 reflects the fact that consecutive digit products differ by 4 or 5. They differ by 5 when the previous carry decreased, and that's exactly when the previous count around the circle passed 0. If you are currently on digits 7 through 9, the carry must have just dropped, and the next product should decrease by 5 rather than 4. Crossing out the 6 is tantamount to counting down 5 rather than 4 when the current digit is 7 through 9, since the next count will skip the 6.

THEOREM 2. *Let x be the base-b number $(b-1)\ldots321$ and choose an n between 1 and b. The product $n \cdot x$ is pandigital if and only if $b-1$ and n are relatively prime.*

Here, use a slightly different circle process to generate the product. Start with the same circle of digits. Cross out n and write it down as the rightmost digit. Then count by n's (skipping the crossed out n) *clockwise* until all b digits are recorded.

Returning to Theorem 1, 123456789 times n is pandigital for lots of larger values of n, too. In particular, n can be any of 10, 11, 13, 14, 16, 17, 20, 22, 23, 25, 26, 31, 32, 34, 35, 40, 41, 43 44, 50, 52, 53, 61, 62, 70, 71, or 80. For instance,

$$12345789 \times 71 = 8765432019$$

Note that this list includes no numbers which are a multiple of 3 (which comes as no surprise) but also omits other numbers such as 19. It remains open to generalize this example to base b.

Acknowledgment. Thanks to David Molnar who identified that the multipliers yielding pandigital numbers are relatively prime to $b-1$ in Theorem 1.

REFERENCES

1. Martin Gardner, *Magic Numbers of Dr Matrix,* Prometheus Books, 1985.
2. Patrick De Geest, *The nine digits page*, `http://www.worldofnumbers.com/ninedigits.htm`.
3. David Wells, *The Penguin Book of Curious and Interesting Numbers*, Penguin, revised edition, 1998.

No Fooling! Newton's Method Can Be Fooled

PETER HORTON
North Harris College
Houston, Texas 77073
peter.w.horton@nhmccd.edu

You might think that if the Newton sequence of a function converges to a number, that this number must be a zero of the function. At least that's what a group of first semester calculus students thought a couple of years ago.

Most calculus textbooks give examples where the Newton sequence gets stuck (oscillates), hits a horizontal tangent line and fails, or simply converges to a different zero than the one intended, but I don't see calculus textbooks give examples of Newton sequences converging to nonzeros.

Let's look at the Newton iteration and see why the students were so optimistic about the success of Newton's Method:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}.$$

Typically, the Newton sequence converges to a number L, and the function and its derivative are continuous at L with $f'(L) \neq 0$, so we can let $n \to \infty$ in the Newton formula, arriving at

$$L = L - \frac{f(L)}{f'(L)}$$
$$\Rightarrow \frac{f(L)}{f'(L)} = 0$$
$$\Rightarrow f(L) = 0.$$

So we can conclude that the Newton sequence converges to a zero of f.

If we can create a Newton sequence $\{x_n\}$ which converges to a number L with the property that $\{|f'(x_n)|\}$ diverges to ∞, then L might not be a zero of f. The reason is that in this case, $\frac{f(x_n)}{f'(x_n)}$ can converge to zero without having $f(x_n) \to 0$ or $f(L) = 0$.

When those overly optimistic calculus students wouldn't believe that such a situation was possible, I sketched a function whose graph and associated Newton sequence were similar to Figure 1.

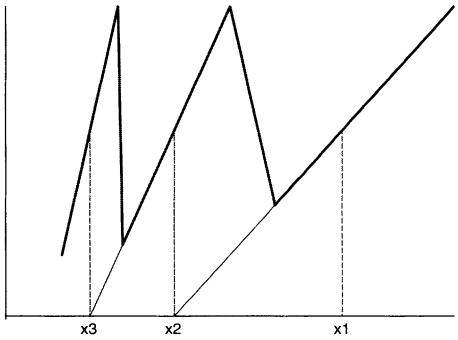


Figure 1 Sketched function with Newton sequence

Well, those calculus students wanted a formula. So I came up with the following formula for a discontinuous version of the previous function:

$$f(x) = \begin{cases} \frac{1}{5}(8x-3); \frac{1}{2} < x \leq 1 \\ \frac{1}{5}(16x-3); \frac{1}{4} < x \leq \frac{1}{2} \\ \vdots \\ \frac{1}{5}(2^{n+2}x-3); \frac{1}{2^n} < x \leq \frac{1}{2^{n-1}} \\ \vdots \\ \frac{1}{5}; x = 0 \end{cases}$$

If you choose any x_1 in $[0, 1]$ where $f'(x_1)$ exists, the Newton sequence will converge to 0, even though $f(0) \neq 0$. The reason is that each segment of the graph of f directs the subsequent term of the Newton sequence beneath the next closer segment to the vertical axis.

Figure 2 is a partial graph of this function f along with the beginning of an associated Newton sequence.

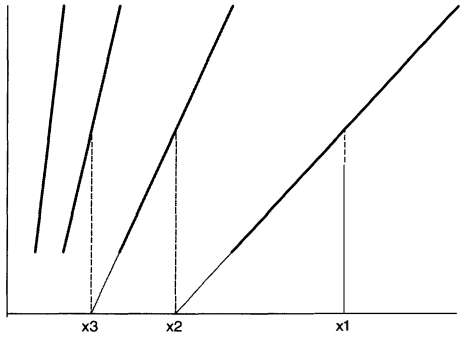


Figure 2 First function with a formula and a Newton sequence.

The convergence of the Newton sequence in the previous example is monotonic. I also constructed an example with nonmonotonic convergence:

Let

$$g(x) = \begin{cases} \frac{8}{3}\left(x - \frac{1}{2}\right) + \frac{7}{3}; \frac{1}{2} \le x < 1 \\ \frac{16}{3}\left(x - \frac{1}{4}\right) + \frac{7}{3}; \frac{1}{4} \le x < \frac{1}{2} \\ \vdots \\ \frac{2^{n+2}}{3}\left(x - \frac{1}{2^n}\right) + \frac{7}{3}; \frac{1}{2^n} \le x < \frac{1}{2^{n-1}} \\ \vdots \\ \frac{7}{3}; x = 0 \end{cases}$$

and let

$$f(x) = \begin{cases} g(x); x \ge 0 \\ g(-x); x \le 0 \end{cases}.$$

If you choose any x_1 in $(-1, 1)$ where $f'(x_1)$ exists, the Newton sequence will converge to 0, even though $f(0) \neq 0$. The reason in this case is that each segment of the graph of f directs the subsequent term of the Newton sequence beneath the next closer segment to the vertical axis on the opposite side of the vertical axis as illustrated in Figure 3.

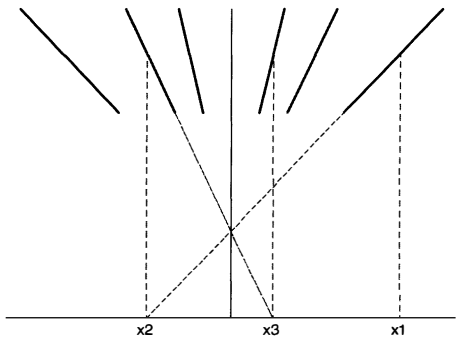


Figure 3 Second function with a formula and a Newton sequence.

Even with the formulas, the previous examples still seem very artificial to most first semester calculus students since the functions are discontinuous, but this can be remedied. The functions can easily be modified into smooth functions by truncating the disconnected segments of their graphs and then interpolating among them with smooth functions. Even better, the function $f(x) = \pi - 2x \sin\left(\frac{\pi}{x}\right)$ for $x \neq 0$, and $f(0) = \pi$, which is a modification of an exercise in the wonderful book *Microcomputers and Mathematics* [**1**, p. 57], does the trick. The function f is continuous everywhere and continuously differentiable except at 0. The formula for generating the Newton sequence is

$$x_{n+1} = x_n - \frac{\pi - 2x_n \sin\left(\frac{\pi}{x_n}\right)}{\frac{2\pi}{x_n}\cos\left(\frac{\pi}{x_n}\right) - 2\sin\left(\frac{\pi}{x_n}\right)}.$$

If we start with $x_1 = \frac{1}{2}$, the Newton sequence proceeds as follows:

$$\begin{aligned} x_2 &= \frac{1}{2} - \frac{\pi}{4\pi} = \frac{1}{4} \\ x_3 &= \frac{1}{4} - \frac{\pi}{8\pi} = \frac{1}{8} \\ x_4 &= \frac{1}{8} - \frac{\pi}{16\pi} = \frac{1}{16} \\ &\vdots \\ x_n &= \frac{1}{2^{n-1}} - \frac{\pi}{2^n\pi} = \frac{1}{2^n} \end{aligned}$$

So $x_n \to 0$, but $f(0) \neq 0$. If we examine the derivative values of f at the terms of the Newton sequence, we arrive at the following: $f'\left(x_n = \frac{1}{2^n}\right) = 2^{n+1}\pi\cos(2^n\pi) - 2\sin(2^n\pi) = 2^{n+1}\pi$. This means that $\{|f'(x_n)|\}$ diverges to ∞.

Figure 4 is a graph of this function f and the first few terms of this Newton sequence.

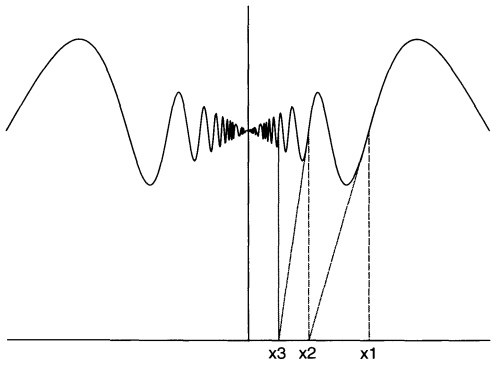


Figure 4 Smooth function and a Newton sequence.

Similarly, an example of nonmonotonic convergence is provided by the function $f(x) = 3\pi - 2x \sin\left(\frac{\pi}{x}\right)$, for $x \neq 0$, and $f(0) = 3\pi$. Starting with $x_1 = \frac{1}{2}$, we find, as

the reader may verify, that

$$x_2 = \frac{1}{2} - \frac{3\pi}{4\pi} = -\frac{1}{4}$$

$$x_3 = -\frac{1}{4} - \frac{3\pi}{-8\pi} = \frac{1}{8}$$

$$\vdots$$

$$x_n = \frac{(-1)^n}{2^{n-1}} - \frac{(-1)^n 3\pi}{2^n \pi} = \frac{(-1)^{n+1}}{2^n}.$$

Thus, $\{x_n\}$ converges nonmonotonically to 0, but $f(0) \neq 0$.

Conclusion

So even if a Newton sequence of a function converges, there's no guarantee that it converges to a zero of that function, and that's no fooling!

Acknowledgement. I would like to thank the referees for their helpful suggestions—both grammatical and mathematical. And even though the editor thinks he is only doing his job, he deserves a lot of thanks for great suggestions to improve the smooth example functions.

REFERENCE

1. J. W. Bruce, P. J. Giblin, P. J. Rippon, *Microcomputers and Mathematics*, Cambridge University Press, Cambridge, England, 1993.

On Antiderivatives of the Zero Function

R. MICHAEL RANGE
State University of New York at Albany
Albany, NY 12222
range@math.albany.edu

We all know that a function whose derivative is zero everywhere must be constant. Have you ever wondered about this fact? Perhaps it appears so intuitively obvious to you that no further thought is needed. What else could a function be whose graph has everywhere a horizontal tangent? Yet the detailed proofs presented in most calculus texts do not attempt to fill in the gaps to turn a vague intuitive argument into a correct proof. Instead, the proofs invariably involve an elaborate and completely unmotivated detour via the existence of extrema, Rolle's Theorem and the Mean Value Theorem. Professional mathematicians and committed mathematics students may appreciate the elegance and logic of such a proof, yet the vast majority of our students and anyone else who tries to understand the basic ideas of calculus might wonder about the lack of a direct and correct intuitive argument that validates what is surely one of the important truths of calculus.

The trouble with the apparently so "obvious" statement is that it ultimately relies on the completeness of the real numbers. Note, for example, that there exist strictly monotonic differentiable functions f on [0, 1] which satisfy $f'(r) = 0$ at all rational

numbers r ([**8**, p. 216]). Therefore, any intuitive argument that does not acknowledge this fact—at least in the background, clearly visible to anyone familiar with completeness—is bound to misrepresent the situation.

There are of course proofs of this result and of other closely related basic calculus theorems that avoid the Mean Value Theorem. (See, for example, references [**1**]–[**5**], [**7**]). Surprisingly, at least to my knowledge, none of these arguments has entered mainstream calculus texts, where the Mean Value Theorem continues to dominate. So it might be useful to present an intuitive and correct argument that should be convincing to anyone with a crude understanding of derivatives, and that may easily be strengthened into a rigorous proof. At the technical level, this note does not add any novel ideas. In particular, in different formulations, the simple observation at the heart of the argument has been used in this context before, for example in [**4**] and [**7**].

A popular intuitive explanation of derivatives involves the concept of instantaneous velocity, easily understood as the "limit" of average velocities over shorter and shorter time intervals. The speedometer in a car provides a familiar instrument that displays the velocity at any given moment. Rephrasing the result in this framework simply says that if the instantaneous velocity is always zero, then there is no motion at all, in other words, the average velocity over any time interval is zero. Equivalently, if the average velocity is nonzero over some time interval, then, at some time, the instantaneous velocity must be nonzero as well. Everyone would probably agree that either version is consistent with one's experience and hence "obvious". Yet the "obvious" straightforward attempt to turn this argument into a proof typically fails, because nonzero average velocities may very well approximate an instantaneous velocity that is zero.

What is needed is a sequence of nonzero average velocities over shrinking time intervals that manifestly has nonzero limit.

The following simple observation provides the key to the construction of such a sequence.

If during each of two successive time periods $[t_0, t_1]$ *and* $[t_1, t_2]$ *the average velocity is less than or equal to* v*, then the average velocity over the combined period* $[t_0, t_2]$ *is also less than or equal to* v.

Again, this statement is consistent with our experience and completely "obvious". Moreover, in contrast to the earlier statements, it also has the great advantage that it is easily verified straight from the definitions by simple algebra—there is no need to invoke any subtle properties of numbers here. Just try it! I therefore skip the short proof.

It is now clear how to proceed. Suppose the average velocity v_0 over the interval $[c_0, d_0]$ is nonzero, say $v_0 > 0$. Divide the interval in half. By the observation just made, v_0 is less than or equal to the maximum of the average velocities over each half interval. In other words, the average velocity v_1 over at least one of these half intervals must be at least as large as v_0, i.e., $v_1 \geq v_0$. Label that half by $[c_1, d_1]$. Continue this process. At the nth step one obtains an interval $[c_n, d_n] \subset [c_{n-1}, d_{n-1}] \subset [c_0, d_0]$ of length $(d_0 - c_0)/2^n$, so that the average velocity v_n over $[c_n, d_n]$ satisfies $v_n \geq v_0$. Let T be a point (in fact the only point) contained in all these intervals. (For the technically minded: this is precisely the place where the completeness of $\mathbb{R}$ must be invoked!) Then the instantaneous velocity $v(T)$, being the limit of average velocities $v_n \geq v_0$ over shorter and shorter time intervals shrinking to T, must also be greater than or equal to $v_0 > 0$, and hence $v(T) \neq 0$ as needed. If desired, the last (intuitive) argument can be made rigorous by invoking the precise limit definition of derivatives combined with the observation above to pass from average velocities over $[c_n, d_n] = [c_n, T] \cup [T, d_n]$ to intervals with one endpoint at T.

What if $v_0 < 0$? Then the preceding argument still gives $v(T) \geq v_0$, although now this does not imply $v(T) \neq 0$. Yet it seems perfectly reasonable that the whole argu-

ment can be modified to also find T^* with $v(T^*) \leq v_0$. Most students will be willing to accept this; if not, they should work through the exercise, thereby demonstrating that they really understand what is going on. Alternatively, one may refer to the following slightly more abstract discussion.

Let us summarize the standard consequences that are central to this circle of ideas. For a function f defined on an interval I, we denote the average rate of change of f over $[a, b] \subset I$, where $a < b$, by

$$\Delta(f, [a, b]) = \frac{f(b) - f(a)}{b - a}.$$

The argument we just went through proves the second inequality in the following theorem. The first inequality follows by applying that result to $-f$ in place of f.

THEOREM. *Assume that f is differentiable on I. If $[a, b] \subset I$, then there exist x_{low} and $x_{\text{high}} \in [a, b]$ such that*

$$f'(x_{\text{low}}) \leq \Delta(f, [a, b]) \leq f'(x_{\text{high}}).$$

The result that prompted the whole discussion follows immediately.

COROLLARY 1. *If $f' \equiv 0$ on I, then f is constant.*

Proof. By the Theorem, the hypothesis implies $\Delta(f, [a, b]) = 0$, i.e. $f(b) = f(a)$ for all $a, b \in I$. ∎

COROLLARY 2. *If $f' \geq 0$ on I, then f is increasing on I.*

Proof. $\Delta(f, [a, b]) \geq f'(x_{\text{low}}) \geq 0$ implies $f(b) \geq f(a)$. ∎

COROLLARY 3. *If $f' > 0$ on I, then f is strictly increasing on I.*

Proof. $\Delta(f, [a, b]) \geq f'(x_{\text{low}}) > 0$ implies $f(b) > f(a)$. ∎

Corresponding statements hold for $f' \leq 0$ or $f' < 0$.

COROLLARY 4. *(Mean Value Theorem for continuously differentiable functions.) If f' is continuous on I, then there exists $p \in [a, b]$ such that $f'(p) = \Delta(f, [a, b])$.*

Proof. Apply the Intermediate Value Theorem to f' and the value $\Delta(f, [a, b])$ which lies between $f'(x_{\text{low}})$ and $f'(x_{\text{high}})$. ∎

Our discussion shows that the important and natural question of finding *all* antiderivatives of the zero function can be answered in a well motivated and direct way. The resulting mean value *inequality* easily handles all the other related elementary calculus results. Perhaps this should become the preferred approach in introductory calculus texts. (See [7], for example.) Purists may argue that the additional continuity requirement in the proof of the standard Mean Value Theorem is a major flaw, but I agree with L. Bers [1] that the version of that theorem given here is all that's needed at this level. Differentiable functions with discontinuous derivative are an anomaly of interest mainly to mathematicians. Let us keep matters simple and transparent for our main audience.

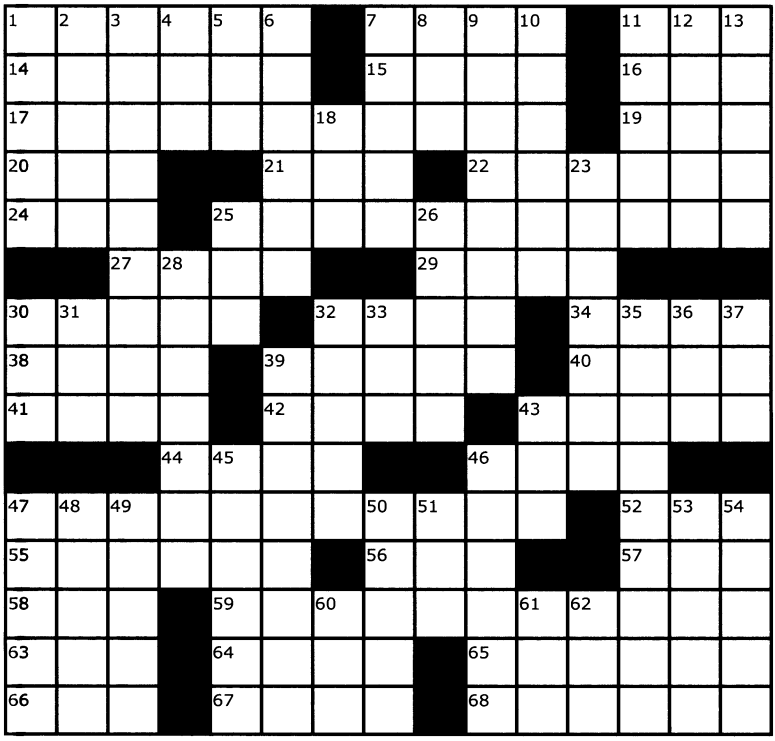
REFERENCES

1. L. Bers, On avoiding the mean value theorem, *Amer. Math. Monthly* **74** (1967) 583.
2. L. W. Cohen, On being mean to the mean value theorem, *Amer. Math. Monthly* **74** (1967) 581–582.
3. D. Desbrow, On zero derivatives. *Amer. Math. Monthly* **103** (1996) 410–411.

4. I. Halperin, A fundamental theorem of the calculus. *Amer. Math. Monthly* **61** (1954) 122–123.
5. M. Powderly, A simple proof of a basic theorem of the calculus. *Amer. Math. Monthly* **70** (1963) 544.
6. R. M. Range, *Calculus in One and in Several Variables: Basic Concepts and Applications.* Manuscript, 512 + x pp. (Under review)
7. D. E. Richmond, An elementary proof of a theorem of calculus. *Amer. Math. Monthly* **92** (1985) 589–590.
8. K. Stromberg, *Introduction to Classical Real Analysis.* Wadsworth Int. Group, Belmont, CA, 1981.

Six Famous Mathematicians

—Ronald E. Prather
78 Gleneden Ave.
Oakland, CA 94611



Solution on page 398

Clues

ACROSS

1 Place to see primates
7 Robert or Alan
11 Dog sound
14 Identity
15 End of maca
16 Actress Aishwarya
17 Mathematician
19 Part of speech; abbr.
20 Building extension
21 Bay Area airport
22 Badge of a fed?
24 Inlet
25 Mathematician
27 Uproar
29 Unadorned
30 Satisfied
32 Kind of transport
34 Work
38 Sills solo
39 "Aren't _____ ?"
40 Huge
41 It's found in pockets
42 Letter insert; abbr.
43 Low card
44 Wind instrument
46 Actor Malden
47 Mathematician
52 Small change; abbr.
55 "_____ a break!"
56 Wrestling place
57 Be indebted
58 British bus. suffix
59 Mathematician
63 Pitching stat.
64 Space org.
65 Mason's secretary
66 "Dracula" director Browning
67 Film pooch
68 Soc. Sci. class asst.

DOWN

1 Rage
2 Physicist Wolfgang
3 Mathematician
4 Alphabet ending to British
5 Tic-tac-toe winner
6 Kind of printing
7 Tree place
8 Law to Henri
9 "_____ us, we'll call you"
10 Kind of cracker
11 Funding for research
12 Television antecedent
13 Hudson parkway, abbr.
18 Alien craft
23 Maryland city
25 Deity
26 Grand
28 Alchemist proclamation?
30 She's a gal
31 Jackie's second
32 Take out again
33 WWII flying unit
35 Mathematician
36 Pac. Ten sch.
37 Fr. anointed one
39 60's group debt to China?
43 "_____ Kapital"
45 Central Illinois town
46 A kind of sugar
47 Shoelace tip
48 Explosive, for short
49 Movie promo
50 Ming introduces himself?
51 Computer syst.
53 Bird sound
54 Sealy rival
60 Denver clock abbr.
61 And so on
62 You might get one on a stalker; abbr.

PROBLEMS

ELGIN H. JOHNSTON, *Editor*
Iowa State University

Assistant Editors: RĂZVAN GELCA, Texas Tech University; ROBERT GREGORAC, Iowa State University; GERALD HEUER, Concordia College; VANIA MASCIONI, Ball State University; BYRON WALDEN, Santa Clara University; PAUL ZEITZ, The University of San Francisco

Proposals

To be considered for publication, solutions should be received by May 1, 2008.

1781. *Proposed by Paul Bracken, University of Texas, Edinburg, TX.*

Let γ be Euler's constant and for positive integer n define

$$\gamma_n = \sum_{k=1}^{n} \frac{1}{k} - \log n \qquad \text{and} \qquad \alpha_n = 2n(\gamma_n - \gamma).$$

Prove that the sequence $\{\alpha_n\}$ is monotonically increasing and bounded above. In addition, determine $\lim_{n\to\infty} \alpha_n$.

1782. *Proposed by Stephen J. Herschkorn, Highland Park, NJ.*

Lines $\overleftrightarrow{AB}$ and $\overleftrightarrow{AC}$ are perpendicular, D lies on $\overline{BC}$, and E and F lie on $\overline{AC}$. In addition, $\overleftrightarrow{AD}$ and $\overleftrightarrow{DF}$ are perpendicular, $AB = AD = 1$, and $AE = DE = x$. Find CF.

1783. *Proposed by Ovidiu Bagasar, Babes Bolyai University, Cluj Napoca, Romania.*

Let n be a positive integer and let $x_1, x_2, \ldots, x_n$ be positive real numbers. Let $S = x_1^n + x_2^n + \cdots + x_n^n$ and $P = x_1 x_2 \cdots x_n$. Prove that

$$\sum_{k=1}^{n} \frac{1}{S - a_k^n + P} \leq \frac{1}{P}.$$

1784. *Proposed by Ovidiu Furdui, Western Michigan University, Kalamazoo, MI.*

Let $\alpha > 0$ and let p be a positive integer. Prove that

$$\sum_{n=1}^{\infty} \frac{\alpha^{n-1}}{(\alpha + p)(2\alpha + p) \cdots (n\alpha + p)} = e \int_0^1 x^{p-1+\alpha} e^{-x^\alpha}\, dx.$$

We invite readers to submit problems believed to be new and appealing to students and teachers of advanced undergraduate mathematics. Proposals must, in general, be accompanied by solutions and by any bibliographical information that will assist the editors and referees. A problem submitted as a Quickie should have an unexpected, succinct solution.

Solutions should be written in a style appropriate for this MAGAZINE. Each solution should begin on a separate sheet.

Solutions and new proposals should be mailed to Elgin Johnston, Problems Editor, Department of Mathematics, Iowa State University, Ames IA 50011, or mailed electronically (ideally as a LaTeX file) to ehjohnst@iastate.edu. All communications should include the reader's name, full address, and an e-mail address and/or FAX number. Please make sure your name appears on all pages, including electronic pages.

1785. *Proposed by Mihaly Bencze, Brasov, Romania.*

Let k be a positive integer, let x a real number, and let $\{x\}$ denote the fractional part of x. Prove that

a. $\displaystyle\sum_{j=1}^{n} \left\lfloor x + \frac{j-1}{n} \right\rfloor^k = n\lfloor x \rfloor^k + \left((\lfloor x \rfloor + 1)^k - \lfloor x \rfloor^k\right) \lfloor n\{x\} \rfloor.$

b. $\displaystyle\sum_{j=1}^{n} \left\lfloor x + \frac{2j-1}{2n} \right\rfloor^k = n\lfloor x \rfloor^k + \left((\lfloor x \rfloor + 1)^k - \lfloor x \rfloor^k\right) \left\lfloor n\{x\} + \frac{1}{2} \right\rfloor.$

Quickies

Answers to the Quickies are on page 398.

Q975. *Proposed by Michael W. Botsko, Saint Vincent College, Latrobe, PA.*

Let f be a positive continuous function and g the derivative of a real valued function, both defined on an interval $[a, b]$. Prove that fg has the intermediate value property on $[a, b]$.

Q976. *Proposed by Ovidiu Furdui, Western Michigan University, Kalamazoo, MI.*

Let $m + 1 > 2k > 0$ and let α, β be positive numbers with $\alpha\beta = \pi^2$. Show that

$$\alpha^{\frac{m+1}{2}} \int_0^\infty \frac{x^m e^{-\alpha x^2}}{(e^{2\pi x} - 1)^k (e^{2\alpha x} - 1)^k}\, dx = \beta^{\frac{m+1}{2}} \int_0^\infty \frac{x^m e^{-\beta x^2}}{(e^{2\pi x} - 1)^k (e^{2\beta x} - 1)^k}\, dx.$$

Solutions

Factorial factors **December 2006**

1756. *Proposed by Courtney H. Moen and William P. Wardlaw, U. S. Naval Academy, Annapolis, MD.*

For which nonnegative integers n do there exist nonnegative integers a and b such that $n! = 2^a(2^b - 1)$?

Solution by John Christopher, California State University, Sacramento, CA.

It is easy to check that any triple (n, a, b) from the set

$$\{(0, 0, 1), (1, 0, 1), (2, 1, 1), (3, 1, 2), (4, 3, 2), (5, 3, 4)\}$$

is a solution. We show that there are no solutions for $n \geq 6$.

First observe that for $n = 6, 7, 8$, the equation $n! = 2^a(2^b - 1)$ has no solutions in positive integers a, b. Now let $n \geq 9$ and let m be the largest positive integer i such that $3^i | n!$. If $n! = 2^a(2^b - 1)$, then $(2^b - 1) \equiv 0 \pmod{3^m}$. Because 2 is a primitive root of the prime 3 and $2^2 - 1 \not\equiv 0 \pmod{3^2}$, it follows that 2 is a primitive root for 3^m. (See David M. Burton, *Elementary Number Theory*, McGraw Hill, 2007, pages 160–161.) Thus, because $(2^b - 1) \equiv 0 \pmod{3^m}$, it must be the case that $\phi(3^m) | b$, and hence that $b \geq \phi(3^m) = 2 \cdot 3^{m-1}$.

Now write $n \geq 9$ as $n = 3k + j$, $k \geq 3$ and $j = 0$, 1 or 2. Then by counting the factors of 3 in $n!$ we conclude that $m \geq k + 1$ and hence that $b \geq 2 \cdot 3^k$. If $k = 3$, then

$n = 9$, 10, or 11 and $b \geq 2 \cdot 3^3 = 54$. But then $2^b - 1 \geq 2^{54} - 1 > 11!$, showing there is no solution to $n! = 2^a(2^b - 1)$ for $k = 3$ (e.g., $n = 9$, 10, 11.)

Assume that for some $k = i \geq 3$ we have $2^{2 \cdot 3^i} - 1 \geq (3i + 2)!$, so $n! = 2^a(2^b - 1)$ is impossible for $k = i$. Now consider the case $k = i + 1$. Then $m \geq i + 2$, $b \geq 2 \cdot 3^{i+1}$, and

$$\begin{aligned} 2^b - 1 &\geq 2^{2 \cdot 3^{i+1}} - 1 > 2^{2 \cdot 3^{i+1}} - 2^{4 \cdot 3^i} \\ &= 2^{4 \cdot 3^i}(2^{2 \cdot 3^i} - 1) > 2^{4 \cdot 3^i}(3i + 2)! > (3(i + 1) + 2)!, \end{aligned}$$

because

$$2^{4 \cdot 3^i} = 2^{3^i} 2^{3^i} 2^{3^i} 2^{3^i} > 3(i + 1)\,(3(i + 1) + 1)\,(3(i + 1) + 2)\,.$$

Thus $n! = 2^a(2^b - 1)$ is impossible for $k = i + 1$ and hence, by induction, for all $k \geq 3$, that is, for all $n \geq 9$.

Also solved by Stefan Chatadus (Poland), Con Amore Problem Group (Denmark), Fejéntaláltulka Szeged Problem Solving Group (Hungary), G.R.A.20 Problem Solving Group (Italy), Douglas E. Iannucci, Lenny Jones, Mark Krusemeyer, Peter W. Lindstrom, Raúl A. Simon (Chile), Gary L. Walls, and the proposers. There were three incorrect submissions.

The birthday problem revisited — December 2006

1757. *Proposed by Ken Ross, University of Oregon, Eugene, OR.*

Consider $M \geq 3$ equally likely properties, like birthdays, that objects, such as people, can possess. We assume that each object possesses exactly one of the M properties. For $n \leq M$, let $P(n;\ M)$ be the probability that in a random sample of n objects, all of the objects possess different properties, and let $\beta(M)$ be the least n such that $P(n;\ M) < \frac{1}{2}$. (The familiar birthday problem is based on the fact that $\beta(365) = 23$.) It is well known, based on estimations, that $\beta(M)$ is close to $\sqrt{2M \ln 2}$. Show that, in fact, $\beta(M)$ is one of the two integers in the interval $(\sqrt{2M \ln 2},\ \sqrt{2M \ln 2} + 2)$.

Solution by Albert Stadler, Dübendorf, Switzerland.

First note that because $\ln 2$ is irrational, $\sqrt{2M \ln 2}$ is not an integer for any positive integer M. Thus there are two integers in the interval $(\sqrt{2M \ln 2},\ \sqrt{2M \ln 2} + 2)$. Furthermore, note that

$$\beta(3) = \beta(4) = \beta(5) = 3, \qquad \beta(6) = \beta(7) = \beta(8) = \beta(9) = 4,$$

and

$$\beta(10) = \beta(11) = \beta(12) = \beta(13) = 5,$$

and that in each of these cases $\beta(M)$ lies in the interval $(\sqrt{2M \ln 2},\ \sqrt{2M \ln 2} + 2)$.

Now assume that $M \geq 14$. It is well known that

$$P(n;\ M) = \prod_{k=1}^{n-1} \left(1 - \frac{k}{M}\right).$$

We seek $\beta(M)$, the smallest positive integer n for which

$$-\ln P(n;\ M) = -\sum_{k=1}^{n-1} \ln\left(1 - \frac{k}{M}\right) > \ln 2.$$

By Taylor's Theorem there are θ_k, $k = 1, 2, \ldots, n-1$ with $0 < \theta_k < 1$ and

$$-\ln P(n;\ M) = \sum_{k=1}^{n-1} \left(\frac{k}{M} + \frac{1}{2\left(1 - \theta_k \frac{k}{M}\right)^2} \left(\frac{k}{M}\right)^2 \right).$$

It then follows that there is a θ, $0 < \theta < 1$ such that

$$-\ln P(n;\ M) = \sum_{k=1}^{n-1} \frac{k}{M} + \frac{1}{\left(1 - \theta \frac{n-1}{M}\right)^2} \sum_{k=1}^{n-1} \frac{k^2}{2M^2} = \frac{n(n-1)}{2M} + \frac{n(n-1)(2n-1)}{12M^2\left(1 - \theta \frac{n-1}{M}\right)^2}.$$

Now let $\lambda \in [0,\ 1]$ be fixed and let

$$f_\lambda(x) = \frac{x(x-1)}{2M} + \frac{x(x-1)(2x-1)}{12M^2\left(1 - \lambda \frac{x-1}{M}\right)^2}.$$

Note that $f_\lambda(x)$ is increasing for $1 < x < M$ and that for fixed x, $f_\lambda(x)$ is an increasing function of $\lambda \in [0,\ 1]$. Thus

$$f_\theta\left(\sqrt{2M \ln 2} + 1\right) \geq f_0\left(\sqrt{2M \ln 2} + 1\right) > \ln 2$$

and

$$f_\theta\left(\sqrt{2M \ln 2}\right) \leq f_1\left(\sqrt{2M \ln 2}\right) < \ln 2 - \sqrt{\frac{\ln 2}{2M}} + \frac{\sqrt{2}(\ln 2)^{3/2}}{3\sqrt{M}\left(1 - \sqrt{\frac{2 \ln 2}{M}}\right)^2} < \ln 2,$$

where the last inequality holds for $M \geq 14$. The desired result follows.

Note. With a deeper look at estimates for $P(n;\ M)$, Allen Schwenk was able to prove that $\beta(M)$ lies in the interval $\left(\sqrt{2M \ln 2} + 0.5 - \frac{1}{3}\ln 2,\ \sqrt{2M \ln 2} + 1.28\right)$.

Also solved by Allen Schwenk, and the proposer. There were two partial solutions submitted.

A Fibonacci product **December 2006**

1758. *Proposed by Michael Goldenberg and Mark Kaplan, The Ingenuity Project, Baltimore Polytechnic Institute, Baltimore, MD.*

Let F_n be the nth Fibonacci number, that is, $F_0 = 0$, $F_1 = 1$, and $F_n = F_{n-1} + F_{n-2}$ for $n \geq 3$. Prove that

$$\prod_{n=2}^{\infty} \frac{F_{2n} + 1}{F_{2n} - 1} = 3.$$

Solution by CMC 328, Carleton College, Northfield, MN.

Let L_n denote the nth Lucas number, where $L_0 = 2$, $L_1 = 1$, and, for $n \geq 2$, $L_n = L_{n-1} + L_{n-2}$. The following identities can be proved "as a package" by induction:

$$F_{2n} = F_n L_n, \qquad F_{2n+1} - F_N L_{n+1} = (-1)^n, \qquad F_{2n+1} - F_{n+1} L_n = (-1)^{n+1},$$
$$F_{2n} - F_{n-1} L_{n+1} = (-1)^{n+1}, \qquad F_{2n} - F_{n+1} L_{n-1} = (-1)^n.$$

Applying the last two of these identities to the factors in the given product, we find that $\prod_{n=2}^{N} \frac{F_{2n+1}}{F_{2n-1}}$ telescopes to

$$\frac{F_1 L_2}{F_2 L_1} \cdot \frac{F_N L_{N+1}}{F_{N+1} L_N} \quad \text{for } N \text{ even, and to} \quad \frac{F_1 L_2}{F_2 L_1} \cdot \frac{F_{N+1} L_N}{F_N L_{N+1}} \quad \text{for } N \text{ odd.}$$

Because

$$\lim_{N\to\infty}\frac{F_{N+1}}{F_N}=\lim_{N\to\infty}\frac{L_{N+1}}{L_N}=\frac{1+\sqrt{5}}{2},$$

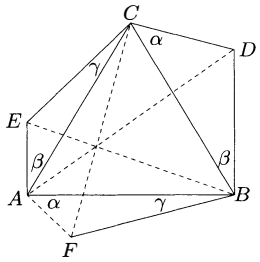
it follows that the infinite product converges to $\frac{F_1L_2}{F_2L_1}=3$.

Also solved by Armstrong Problem Solvers, Michel Bataille (France), Brain Bradie, Stefan Chatadus (Poland), Charles K. Cook, Knut Dale (Norway), Jim Delany, Charles R. Diminnie, Marty Getz and Dixon Jones, G.R.A.20 Problem Solving Group (Italy), Matthew Hudelson, Douglas E. Iannucci, Harris Kwong, Reiner Martin, Kim McInturff, Northwestern University Math Problem Solving Group, Herman Roelants (Belgium), Kenneth Sanderson and Hansun To, Volkhard Schindler (Germany), Albert Stadler (Switzerland), Dave Trautman, and the proposers. There were two incorrect submissions.

A matter of concurrence — December 2006

1759. *Proposed by Larry W. Cusick and Maria Nogin, California State University, Fresno, CA.*

In the accompanying figure, $\triangle ABC$ is equilateral. In addition, $\angle FAB\cong\angle DCB$, $\angle FBA\cong\angle ECA$, and $\angle EAC\cong\angle DBC$. Prove that segments $\overline{AD}$, $\overline{BE}$, and $\overline{CF}$ are concurrent.



Solution by Larry Hoehn, Austin Peay State University, Clarksville, TN.

We prove that

$$\frac{\sin\angle CAD}{\sin\angle DAB}\cdot\frac{\sin\angle ABE}{\sin\angle EBC}\cdot\frac{\sin\angle BCF}{\sin\angle FCA}=1.$$

It will the follow from the trigonometric form of Ceva's theorem that $\overline{AD}$, $\overline{BE}$, and $\overline{CF}$ are concurrent. Applying the law of sines in $\triangle CAD$ and $\triangle BAD$ we have,

$$\frac{\sin\angle CAD}{CD}=\frac{\sin\angle ACD}{AD}\quad\text{and}\quad\frac{\sin\angle DAB}{BD}=\frac{\sin\angle ABD}{AD}.$$

Therefore,

$$\frac{\sin\angle CAD}{\sin\angle DAB}=\frac{CD\sin\angle ACD}{BD\sin\angle ABD}=\frac{CD\sin(60+\alpha)}{BD\sin(60+\beta)}=\frac{\sin\beta}{\sin\alpha}\cdot\frac{\sin(60+\alpha)}{\sin(60+\beta)},$$

where we have used $\frac{CD}{\sin\beta}=\frac{BD}{\sin\alpha}$. Similarly,

$$\frac{\sin\angle ABE}{\sin\angle CBE}=\frac{\sin\gamma}{\sin\beta}\cdot\frac{\sin(60+\beta)}{\sin(60+\gamma)}\quad\text{and}\quad\frac{\sin\angle BCF}{\sin\angle ACF}=\frac{\sin\alpha}{\sin\gamma}\cdot\frac{\sin(60+\gamma)}{\sin(60+\alpha)}.$$

Thus

$$\frac{\sin\angle CAD}{\sin\angle DAB}\cdot\frac{\sin\angle ABE}{\sin\angle EBC}\cdot\frac{\sin\angle BCF}{\sin\angle FCA}=1,$$

and the desired result follows.

Also solved by Michel Bataille (France), Gordon Crandall, Chip Curtis, Con Amore Problem Group (Denmark), Apostolos Demis (Greece), Fejéntaláltulka Szeged Problem Solving Group (Hungary), John Ferdinands, Marty Getz and Dixon Jones, Michael Goldenberg and Mark Kaplan, G.R.A.20 Problem Solving Group, Peter Gressis and Dennis Gressis, Geoffrey A. Kandall, L. R. King, Victor Y. Kutsenok, Elias Lampakis (Greece), Vijayaprasad Nalluri (India), Robin Oakapple, Samuel Otten, John Rigby (Wales), Chad Ryan and Brian Thompson, Robert A. Russell, Volkhard Schindler (Germany), Seshadri Sivakumar, Raúl Simón (Chile), Albert Stadler (Switzerland), R. S. Tiberio, Michael Vowe (Switzerland), Stuart V. Witt, John B. Zacharias, and the proposers. There was one submission with no name.

An inequality of means **December 2006**

1760. *Proposed by Péter Ivády, Budapest Hungary.*

Prove that for $0 < a < b < \infty$,

$$\sqrt{\frac{a^2+b^2}{2}}+\sqrt{ab}-\frac{a+b}{2}>\frac{b-a}{\ln b-\ln a}.$$

Solution by Donald Jay Moore, Wichita KS.

Let $b = a(1+x)^2$. The inequality can then be written in the equivalent form

$$\ln(1+x)>\frac{x^2+2x}{\sqrt{2(1+x)^4+2}-x^2},\qquad 0<x<\infty.$$

The expressions on the left and right side of the inequality are equal (to 0) when $x = 0$. Furthermore, both have positive derivatives on $x > 0$. Thus, to prove the inequality, it suffices to prove that

$$\frac{d}{dx}\ln(1+x)>\frac{d}{dx}\frac{x^2+2x}{\sqrt{2(1+x)^4+2}-x^2},\qquad 0<x<\infty,$$

that is, that

$$\frac{1}{1+x}>\frac{\left(\sqrt{2(1+x)^4+2}-x^2\right)(2x+2)-(x^2+2x)\left(\dfrac{4(1+x)^3}{\sqrt{2(1+x)^4+2}}-2x\right)}{\left(\sqrt{2(1+x)^4+2}-x^2\right)^2},$$

for $0 < x < \infty$. After some algebraic manipulation this is seen to follow from

$$64x^2+320x^3+768x^4+1152x^5+1224x^6+984x^7+582x^8$$
$$+228x^9+54x^{10}+8x^{11}+x^{12}>0,$$

$0 < x < \infty$. This last inequality is valid because all coefficients are positive.

Also solved by Paul Bracken and N. Nadeau, Brian Bradie, Robert Calcaterra, Chip Curtis, Michael Goldenberg and Mark Kaplan, G.R.A.20 Problem Solving Group (Italy), Kee-Wai Lau (China), Elias Lampakis (Greece) Phil McCartney, Perfetti Paolo, Volkhard Schindler (Germany), Albert Stadler (Switzerland), Marian Tetiva (Romania), and the proposer. There was one incorrect submission.

Answers

Solutions to the Quickies from page 393.

A975. Let $a \le c < d \le b$ with $f(c)g(c) \ne f(d)g(d)$ and μ be between $f(c)g(c)$ and $f(d)g(d)$. Without loss of generality we may assume that $f(c)g(c) < \mu < f(d)g(d)$. Let h be a function with $h' = g$ on $[a, b]$. Define the function ϕ by

$$\phi(x) = h(x) - \int_a^x \frac{u}{f(t)}\,dt, \qquad x \in [a, b],$$

where the integral is the Riemann integral. Then,

$$\phi'(x) = h'(x) - \frac{\mu}{f(x)}.$$

Because

$$\phi'(c) = g(c) - \frac{\mu}{f(c)} < 0 < g(d) - \frac{\mu}{f(d)} = \phi'(d),$$

and derivatives have the intermediate value property, there is an $x_0 \in (c, d)$ with

$$0 = \phi'(x_0) = g(x_0) - \frac{\mu}{f(x_0)}.$$

The result follows.

A976. Making the substitution $x = \frac{\beta y}{\pi}$ we see

$$\int_0^\infty \frac{x^m e^{-\alpha x^2}}{(e^{2\pi x} - 1)^k (e^{2\alpha x} - 1)^k}\,dx = \left(\frac{\beta}{\pi}\right)^{m+1} \int_0^\infty \frac{y^m e^{-\beta y^2}}{(e^{2\beta y} - 1)^k (e^{2\pi y} - 1)^k}\,dy.$$

The result follows after noting that $\frac{\beta}{\pi} = \sqrt{\frac{\beta}{\alpha}}$.

Note. The condition $m + 1 > 2k$ is necessary for convergence of the integrals.

Solution to Puzzle on page 391

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

REVIEWS

PAUL J. CAMPBELL, *Editor*
Beloit College

Assistant Editor: Eric S. Rosenthal, West Orange, NJ. Articles and books 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.

Klarreich, Erica, Sensor sensibility: The mathematics of shapes is illuminating the structure of wireless sensor networks, *Science News* 171 (5 May 2007) 282–285.

Engineers predict a future of "smart dust," dust-size sensors of temperature, vibration, noise, light, water levels, and who knows what else. Coordinating a field of sensors presents mathematical problems, such as determining whether the sensors cover the entire field. Grouping the sensors into simplexes, such as triangles, brings topology to bear: Euler characteristic, holes, homology, and more. Applications include combining data from the network and shutting off redundant sensors. Says one researcher, "We want to make sure that when these networks come into existence, the math is there and ready to use."

Posamentier, Alfred S., and Ingmar Lehmann, *The (Fabulous) Fibonacci Numbers*, Prometheus Books, 2007; 385 pp, $28. ISBN 978-1-59102-475-0.

This delightful book for a general audience contains all of the usual applications, curiosities, and "sightings" of the Fibonacci numbers and the golden section, plus ones that I have not seen elsewhere.

Albert, Michael H., Richard J. Nowakowski, and David Wolfe, *Lessons in Play: An Introduction to Combinatorial Game Theory*, A K Peters, 2007; xvi + 288 pp, $49. ISBN 978-1-56881-277-9.

This book is a refreshing "reframing" in textbook form of much of the material in *Winning Ways for Your Mathematical Plays* by Berlekamp, Conway and Guy (2nd ed., 2001), with updating and new developments, by and for a new generation of games aficionados. Included are exercises and their solutions, plus notes to the instructor. Not treated in this volume are loopy games, misère-play, and the computer science approach to games. (Note to publisher: Please use thicker paper, to prevent distracting see-through to subsequent pages.)

Bailey, David H., Jonathan Borwein, et al., *Experimental Mathematics in Action*, A K Peters, 2007; xii + 322 pp, $49. ISBN 978-1-56881-271-7.

The base of mathematical results expands continually, despite the theoretical limitation (due to Gödel) that there are true statements that cannot be proved from axiom systems. "Experimental" mathematicians have come up against practical limitations, in the form of "striking conjectures with no known proof strategy.... [W]e can now have (near) certainty without proof." After citing examples, the authors assert that "mathematics is about *secure knowledge* not proof.... Proofs are often out of reach—but understanding, even certainty, is not." This book is full of wonders, not the least of which is 75 pp of exercises (mainly problems from the *American Mathematical Monthly* and Putnam competitions) for which "a few lines of computer algebra code either provides the solution, suggests an approach, or at least confirms the answer." The illustrious Paul Halmos (1916–2006) died only a year ago, but this book shows that his aphorism "computers are important, but not to mathematics" died an earlier death.

Albert, Jim, Jay Bennett, and James J. Cochran (eds.), *Anthology of Statistics in Sports*, SIAM, 2005; x +322 pp, $65 (P). ISBN 0-89871-587-3.

What's your favorite sport, and what does statistics have to say about its play, strategy, or player and team ratings? This compendium has sections of about half a dozen articles each on football, baseball, basketball, hockey, miscellaneous sports, and topics relevant to multiple sports. Introductory essays describe the origins and scope of the book, plus ways to use sports in teaching statistics. (It's difficult to see why a paperback of reprints, with virtually no permissions costs—all articles are from ASA publications—should cost so much.)

Davis, Philip J., *Mathematics and Common Sense: A Case of Creative Tension*, A K Peters, 2006; xlvi + 242 pp, $34.95. ISBN 978-1-56881-270-1.

"Where is mathematical knowledge lodged, and where does it come from? When should one add two numbers? Why do I believe a theorem? When is a problem solved? What is meant by the word 'random'?" Author Davis has assembled an eminently readable book of answers to these and other "frequently asked questions" about mathematics, from his philosophical perspective that mathematical knowledge is socially constructed.

Havil, Julian, *Nonplussed! Mathematical Proof of Implausible Ideas*, Princeton University Press, 2007; xv + 196 pp, $24.95. ISBN 978-0–691-12056-0.

When can you reuse a calendar? (For 2007 I have been using a black plastic dodecahedron, with one month per face, that is dated 1973.) A welcome change from experimental mathematics's certainty without proof is this book's focus on calculational proof of initially dubious propositions (such as that the 13th of a month is most likely to be a Friday), plus further investigations (such as the pattern for calendar reuse). Topics include tennis paradoxes, rolling a cone uphill, birthday paradoxes, derangements, Buffon's needle, nontransitive dice, Parrondo's paradoxes, calendars, and more. Even the well-worn topics are treated with a fresh perspective or twist. The author uses algebra freely and calculus occasionally.

Steiglitz, Ken, *Snipers, Shills, and Sharks: eBay and Human Behavior*, Princeton University Press, 2007; xix + 276 pp. ISBN 978-0-691-12713-2.

From the title of this book, you might think that it concerns psychology and economics, offers observational data, and perhaps provides recommendations on how to bid in an online auction. It does all those, but one-third of the book is an appendix on the mathematics of auction theory. The author explains why and how eBay has adapted the second-price auction (the highest bidder pays only the second-highest amount bid), how Amazon and Yahoo auctions differ, why "sniping" is a good idea, and how sellers should set opening bids. Little do most participants in online auctions know that mathematics could enlighten their strategies!

Simonson, Andrew J., *Hesiod's Anvil: Falling and Spinning through Heaven and Earth*, Mathematical Association of America, 2007; xv + 344 pp, $54.95. ISBN 978-0-88385-336-8.

Literature through the ages, from Dante to Jules Verne, has described motion—of the Earth, the heavens, and projectiles of all kinds. In doing so, authors have devised implicit models of gravitation, falling, and trajectories. This book whimsically analyzes those models in quantitative terms, together with current models, such as for Galileo's cannon ball drop from the Leaning Tower, Verne's moon shot, Poe's pendulum, H.G. Wells's journey to the center of the Earth, and playing ball on the spaceship in Arthur C. Clarke's film *2001: A Space Odyssey*. The reader needs calculus, and there are exercises, with solutions to some.

Nahin, Paul J., *Chases and Escapes: The Mathematics of Pursuit and Evasion*, Princeton University Press, 2007; xiv + 253 pp, $24.95. ISBN 978-0-691-12514-5.

The first two episodes of the TV series *Numb3rs* in the 2006–07 season introduced the mathematics of pursuit problems. Here is a colorful book dedicated to just that exciting topic, suitable for students who are comfortable with vector calculus and differential equations.

NEWS AND LETTERS

Acknowledgments

The following referees have assisted the MAGAZINE *during the past year. We thank them for their time and care.*

Aboufadel, Edward F., *Grand Valley State University, Allendale, MI*

Abbott, Steve, *Middlebury College, Middlebury, VT*

Albert, Jim, *Bowling Green State University, Bowling Green, OH*

Alsina, Claudi, *University Politecnica de Catalunya, Barcelona, Spain*

Ash, J. Marshall, *De Paul University, Chicago, IL*

Barnes, Julia, *Western Carolina University, Cullowhee, NC*

Battle, Laurie, *Knoxville, TN*

Benjamin, Arthur T., *Harvey Mudd College, Claremont, CA*

Bennett, Curtis D., *Loyola Marymount University, Los Angeles, CA*

Bigelow, Stephen, *University of California Santa Barbara, Santa Barbara, CA*

Boas, Harold, *Texas A & M University, College Station, TX*

Borwein, Jonathan, *Dalhousie University, Halifax, NS, Canada*

Brawner, James N., *Armstrong Atlantic State University, Savannah, GA*

Bremner, Andrew, *Arizona State University, Tempe, AZ*

Brewer, Patrick, *Lebannon Valley College, Annville, PA*

Bressoud, David M., *Macalester College, Saint Paul, MN*

Burton, David, *Franciscan University of Steubenville, Steubenville, OH*

Cairns, Grant, *LaTrobe University, Melbourne, Victoria, Australia*

Caldwell, Chris, *Rives, TN*

Campbell, Paul, *Beloit College, Beloit, WI*

Canada, Daniel, *Spokane, WA*

Cervone, David, *Union College, Schenectady, NY*

Cobb, George, *Mount Holyoke College, South Hadley, MA*

Cox, Jonathan, *SUNY Fredonia, Fredonia, NY*

Crannell, Annalisa, *Franklin & Marshall College, Lancaster, PA*

Cullen, David, *University of Alberta, Edmonton, AB, Canada*

Curran, Steve, *University of Pittsburgh at Johnstown, Johnstown, PA*

Dauben, Joseph, *Herbert H. Lehman College (CUNY), Bronx, NY*

Dunbar, Jean, *Converse College, Spartanburg, SC*

Dunham, William W., *Muhlenberg College, Allentown, PA*

Egge, Eric, *Carleton College, Northfield, MN*

Eisenberg, Bennet, *Lehigh University, Bethlehem, PA*

Fjelstad, Paul T., *Northfield, MN*

Fredrickson, Grey, *Purdue University, West Lafayette, IN*

Goodson, Geoffrey, *Towson University, Towson, MD*

Gordon, Russell, *Whitman College, Walla Walla, WA*

Grinstead, Charles, *Swarthmore College, Swarthmore, PA*

Grossman, Jerrold W., *Oakland University, Rochester, MI*

Guichard, Richard, *Whitman College, Walla Walla, WA*

Hanson, Denis, *University of Regina, Regina, SK, Canada*

Harper, James D., *Central Washington University, Ellensburg, WA*

Haunsperger, Deanna, *Carleton College, Northfield, MN*

Henle, James, *Smith College, Northampton, MA*

Howie, John M., *University of Saint Andrews, St. Andrews, Fife, Scotland*

James, David M., *Howard University, Washington, DC*

Jayawant, Pallaus, *Bates College, Lewiston, ME*

Johnson, Warren, *Connecticut College, New London, CT*

Johnson, Wells, *Bowdoin College, Brunswick ME*

Johnston, Elgin, *Iowa State University, Ames, IA*

Kallaher, Michael, *Pullman, WA*

Katz, Victor J., *Silver Springs, MD*

Kauri, Manmohan, *Benedictine University, Lislie, IL*

Kendig, Keith M., *Cleveland State University, Cleveland, OH*

Kiltinen, John O., *Northern Michigan University, Marquette, MI*

Kim, Jon-Lark, *University of Louisville, Louisville, KY*

Kimer, Chawne, *Lafayette College, Easton, PA*

Kung, Sidney, *Cupertino, CA*

Lamken, Esther, *San Francisco, CA*

Langer, Robert W., *Eau Claire, WI*

Levine, Alan, *Franklin and Marshall College, Lancaster, PA*

Little, John B., *College of the Holy Cross, Worcester, MA*

Lyons, David W., *Lebanon Valley College, Annville, PA*

McCarthy, John E., *Washington University, St. Louis, MO*

McCooey, Michael P., *Franklin and Marshall College, Lancaster, PA*

McCleary, John H., *Vassar College, Pouhkeepsie, NY*

Miller, Steven J., *Brown University, Providence, RI*

Mills, Mark A., *Central College, Pella, IA*

Monson, Barry R., *University of Brunswick, Fredericton, NB, Canada*

Neidinger, Richard D., *Davidson College, Davidson, NC*

Neugenbauer, Christoph J., *Purdue University, West Lafayette, IN*

Nievergelt, Yves, *Eastern Washington University, Cheney, WA*

Nelsen, Roger, *Lewis and Clark College, Portland, OR*

O'Leary, Michael, *College of DuPage, Glen Ellyn, IL*

Pedersen, Jean J., *Santa Clara University, Santa Clara, CA*

Ridenhour, Jim R., *Austin Peay State University, Clarksville, TN*

Robinson, Margaret M., *Mount Holyoke College, South Hadley, MA*

Rose, David A., *Florida Southern College, Lakeland, FL*

Rosenstein, George, *Lancaster, PA*

Ross, Kenneth A., *University of Oregon, Eugene, OR*

Salwach, Chester, *Lafayette College, Easton, PA*

Sandifer, Edward, *Western Connecticut State University, Danbury, CT*

Schuette, Paul H., *Meredith College, Raleigh, NC*

Scott, David R., *University of Puget Sound, Tacoma, WA*

Shell-Gellash, Amy, *Pacific Lutheran University, Tacoma, WA*

Sondow, Jonathan, *New York, NY*

Spivey, Michael, *University of Puget Sound, Tacoma, WA*

Stahl, Saul, *University of Kansas, Lawrence, KS*

Stanhope, Elizabeth, *Lewis and Clark College, Portland, OR*

Stanton, William G., *Gambier, OH*

Stravov, Iva, *Lewis and Clark College, Portland, OR*

Stenson, Catherine, *Juniata College, Huntingdon, PA*

Stewart, Sarah, *Belmont University, Nashville, TN*

Stockmeyer, Paul, *College of William and Mary, Williamsburg, VA*

Suzuki, Jeff, *Brooklyn College (CUNY), Brooklyn, NY*

Tanton, James, *St. Marks Institute of Mathematics, Southborough, MA*

Teague, Daniel, *North Carolina School of Science and Mathematics, Durham, NC*

Towse, Christopher, *Scripps College, Claremont, CA*

Veenstra, Tamara, *University of Redlands, Redlands, CA*

Wagon, Stanley, *Macalester College, Saint Paul, MN*

Wardlaw, William, *United States Naval Academy, Annapolis, MD*

Watkins, John J., *Colorado College, Colorado Springs, CO*

White, Arthur T., *Western Michigan University, Kalamazoo, MI*

Williams, Gordon, *Moravian College, Bethlehem, PA*

Wilson, John H., *Centre College, Danville, KY*

Yiu, Paul Y., *Florida Atlantic University, Boca Raton, FL*

Zhu, Qiji J., *Western Michigan University, Kalamazoo, MI*

Index to Volume 80

AUTHORS

TITLES

PROBLEMS

The letters P, Q, and S refer to Proposals, Quickies, and Solutions, respectively; page numbers appear in parentheses. For example, P1774 (231) refers to Proposal 1774, which appears on page 231.

CONTENTS

THE MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth Street, NW
Washington, DC 20036