Vol. 63, No. 4, October 1990

MATHEMATICS MAGAZINE

- Three-Colorings of Finite Groups
- How to Approach a Traffic Light
- A Historical Gem from Vito Volterra

An Official Publication of The MATHEMATICAL ASSOCIATION OF AMERICA

EDITORIAL POLICY

The aim of *Mathematics Magazine* is to provide lively and appealing mathematical exposition. This is not a research journal and, in general, the terse style appropriate for such a journal (lemma-theorem-proof-corollary) is not appropriate for an article 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. Articles on pedagogy alone, unaccompanied by interesting mathematics, are not suitable. Neither are articles consisting mainly of computer programs unless these are essential to the presentation of some good mathematics. Manuscripts on history are especially welcome, as are those showing relationships between various branches of mathematics and between mathematics and other disciplines.

The full statement of editorial policy appears in this *Magazine*, Vol. 54, pp. 44–45, and is available from the Editor. Manuscripts to be submitted should not be concurrently submitted to, accepted for publication by, nor published by another journal or publisher.

Send new manuscripts to: Martha Siegel, Editor-elect, Mathematics Magazine, Towson State University, Towson, MD 21204. Manuscripts should be typewritten and double spaced and prepared in a style consistent with the format of *Mathematics Magazine*. Authors should submit the original and two copies and keep one copy. In addition, authors should supply the full five-symbol Mathematics Subject Classification number, as described in *Mathematical Reviews*, 1980 and later. Illustrations should be carefully prepared on separate sheets in black ink, the original without lettering and two copies with lettering added.

AUTHOR

Robert Haas studied mathematics on a National Science Foundation graduate fellowship at Case Western Reserve University, receiving a Ph.D. for his work in homological algebra under the distinguished algebraic topologist Peter J. Hilton. He obtained his second Ph.D. from this university in the field of molecular biology and microbiology, and is currently a senior research associate there in the pharmacology department, while continuing to pursue mathematics as an avocation.

Cover: Vito Volterra (1860–1940) (Photograph from the collection of Donald J. Albers)

Vol. 63 No. 4 October 1990

MATHEMATICS MAGAZINE

The *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 the MATHEMATICS MAGAZINE to an individual member of the Association is $16 included as part of the annual dues. (Annual dues for regular members, exclusive of annual subscription prices for MAA journals, are $60. Student and unemployed members receive a 66% dues discount; emeritus members receive a 50% discount; and new members receive a 40% dues discount for the first two years of membership.) The nonmember/library subscription price is $64 per year. Bulk subscriptions (5 or more copies) are available to colleges and universities for classroom distribution to undergraduate students at $5.40 per copy (minimum order 5 copies).
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 Ms. Elaine Pedreira, Advertising Manager, The Mathematical Association of America, 1529 Eighteenth Street, N.W., Washington, D.C. 20036.

Second class postage paid at Washington, D.C. and additional mailing offices.
Postmaster: Send address changes to Mathematics Magazine Membership/Subscriptions Department, Mathematical Association of America, 1529 Eighteenth Street, N.W., Washington, D.C. 20077-9564.
PRINTED IN THE UNITED STATES OF AMERICA

ARTICLES

Three-Colorings of Finite Groups or an Algebra of Nonequalities

ROBERT HAAS
Department of Pharmacology, School of Medicine
Case Western Reserve University, Cleveland, Ohio 44106

Graph coloring theory is a remarkable branch of mathematics. It begins with one of the simplest of ideas, that of coloring a geographical map or the vertices of a graph in such a way that adjacent regions or vertices are colored differently. And yet, as the celebrated four color and Heawood map-coloring theorems show, the existence of such colorings reflects profoundly the fundamental topological properties of the graph. In general terms, one may say that the set of colorings of a graph contains *structural* information about the graph. This paper, going by the principle that a good way to understand a theory is to try to construct a parallel one, will develop a coloring theory in a quite different context, that of groups. The ultimate conclusion will be the same as it was for graphs: The set of colorings contains structural information about the group. Along the way to this conclusion, there will be a chance to think about what makes graph colorings so special, and also, we hope, a pleasant jaunt through elementary group theory, seen from a novel perspective.[1]

What is the unique essential ingredient in the idea of coloring a graph? Perhaps it is the fact that making one choice when defining a coloring does not dictate the next step, but merely limits its possibilities. Coloring point 1 in FIGURE 1 red does not force point 2 to be blue, but only constrains it not to be red. Algebraic constructions, in contrast, are usually much more deterministic—once the value of a homomorphism is defined at x, it is fixed forever at all x^n. There is a heady freedom in defining colorings, like dealing with a card sharper or a gambit player who smiles and says "Pick any card!" or "Now make any move at all!" And yet the freedom is often only temporary. Some point in FIGURE 1 eventually *must* be assigned a third color. There is, similarly, no 3-coloring at all of FIGURE 2. In colorings, as in cards or chess, a few moves later one may be locked in—or out—just as tightly as with the homomorphism!

An n-coloring of a graph is thus a mapping into a set of n colors, assigned with absolute freedom except for one special kind of constraint: If two points are directly related by the graph structure (that is, are adjacent), then the simplest choice (coloring them both the same) is forbidden. What would be a parallel construction for

[1] Harary's book [6] is a standard introduction to graph theory. Ringel, who played a major role in proving the Heawood theorem, describes it in detail in [9]. Appel and Haken's computer-assisted proof of the four color theorem [1] has been explained in several articles ([2], [11], and [12]), and a book [10].

The group theory used in the present paper is all covered in standard introductory abstract algebra textbooks, such as [4] and [8].

My group coloring theory is, to my knowledge, original. There is also, however, a classical, well-established, and very important procedure of Cayley which associates a "color graph" with any group. This is described in [6], [5] (elementary), and [13] (more advanced).

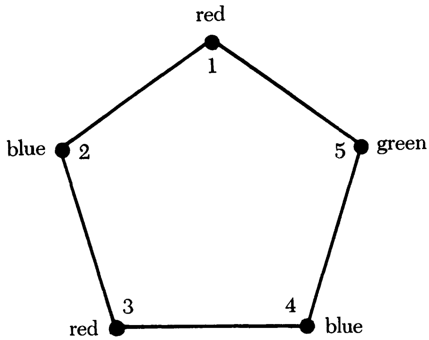


FIGURE 1
To color a graph, assign a color to each vertex in such a way that any two which are adjacent (joined by a line) receive different colors. One of the many ways to color this graph using three colors is shown; it's impossible to do it using only two.

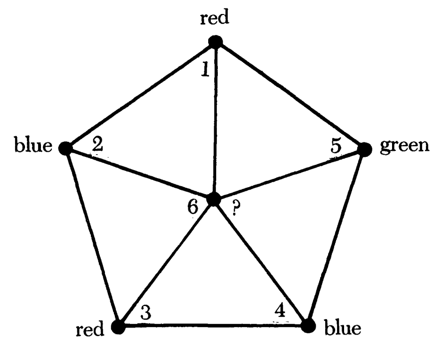


FIGURE 2
FIGURE 1 plus one point is no longer 3-colorable.

a group? An n-coloring f should be a special type of mapping to the canonical n-element group Z_n. If any element z is related by the group structure to x and y (that is, $z = x + y$), then the simplest choice [letting $f(z) = f(x) + f(y)$] must be forbidden. This suggests that an n-coloring of a group should be a "non-homomorphism" into Z_n:

DEFINITION[2]. An *n-coloring* of a group G is a function $f: G \to Z_n$ which for all $x, y \in G$ satisfies $f(x+y) \neq f(x) + f(y)$.

(The more general definition $f(x+y) \neq af(x) + bf(y)$, for some choice of $a, b \in Z_n$, can be shown when $n = 3$ to yield only equivalent forms, or else to reduce to trivialities.)

The reader can check instantly that no group has a 1-coloring. The 2-colorings that do exist are not very interesting, because $f(x+y) \neq f(x) + f(y) \in Z_2$ implies $f(x+y) = f(x) + f(y) + 1$, whence f', defined $f'(x) = f(x) + 1$, is an ordinary homomorphism. For $n \geqslant 4$, n-colorings are floppy, arbitrary, and exceedingly numerous: There are, respectively, 3, 10, 21, 44, and 83 4-colorings, and 4, 16, 52, 144, and 420 5-colorings of Z_1, Z_2, Z_3, Z_4, and Z_5. But at $n = 3$ there is an equilibrium point between the two extremes, and the following theorem holds:

THEOREM. *The 3-colorings of a finite group G form two families, each endowed with the group structure of G_{ab}, the abelianized quotient group of G.*

In other words, the set of 3-colorings of G contains the partial abelian *structure* of G. The remainder of this paper will be devoted to proving this result. Many of the steps will be a familiar review, just modified versions of the first propositions in elementary group theory. In addition, there are perhaps three other general conclusions:

---

[2]Notation: Z denotes the additive group of the integers, and Z_n the cyclic group of order n. The elements of Z_n will generally be written as $0, 1, \ldots, n-1$, with addition mod n. Let α_k: $Z_n \to Z_n$ and χ_k: $Z_n \to Z_{nk}$ be the homomorphisms determined by $\alpha_k(1) = k$ and $\chi_k(1) = k$; $\alpha_{(a,b)}$: $Z_n \times Z_m \to Z_n \times Z_m$ is, similarly, the endomorphism of the product defined by $\alpha_{(a,b)}(x, y) = (ax, by)$. The group operation in the arbitrary group G will be written as multiplication in section 2, and as addition everywhere else.

1) Algebra may be the science of equations and equalities, but one can do quite respectable algebra with non-equalities!

2) What does it really mean for a function to be a homomorphism? This paper marks out one boundary of the concept by showing how non-homomorphisms behave.

And, perhaps most importantly,

3) The set of colorings of an object contains structural information about the object.

1. Examples and Elementary Properties of n-Colorings

Consider the function r: $Z_5 \to Z_3$ defined as follows: $r(0) = r(1) = 1$, $r(2) = r(3) = 0$, and $r(4) = 2$. Then it is easy to check that r is a 3-coloring: For example,

$$r(0+0) = r(0) = 1 \quad \text{while} \quad r(0) + r(0) = 2$$
$$r(0+1) = r(1) = 1 \quad \text{while} \quad r(0) + r(1) = 2$$

and so forth.

Similarly, one can check that each of the other nine functions listed in Table 1A is a coloring. Note, incidentally, that $f(0)$ is never 0. The reason why is perhaps already obvious; if not, it will all be explained in Proposition 1.1 below.

The 3-coloring r of Z_5 follows a simple pattern: As x varies $0, 1, 2, \ldots, r(x)$ varies through one cycle $1, 1, \ldots, 0, 0, \ldots, 2, 2, \ldots$. A 3-coloring r may be constructed by this pattern for larger cyclic groups; for example, TABLE 1B gives r: $Z_{12} \to Z_3$. An n-coloring r for $n > 3$ can be defined similarly by letting $r(x)$ cycle once through $1, 1, \ldots, 0, 0, \ldots, n-1, n-1, \ldots, n-2, n-2, \ldots, \ldots, 2, 2, \ldots$; thus, TABLE 1C gives r: $Z_{11} \to Z_8$. The formal definition of r: $Z_m \to Z_n, m \geqslant n$, is this: Let $m = nk + b$, $0 \leqslant b < n$. Then $r(x) = 1 - i$ for $x \in A_i$, where

$$A_i = \begin{cases} \{i(k+1), \ldots, i(k+1) + k\} & \text{if } 0 \leqslant i < b \\ \{ik + b, \ldots, ik + b + k - 1\} & \text{if } b \leqslant i \leqslant n - 1. \end{cases}$$

TABLE 1. Examples of n-Colorings

A. The 3-colorings of Z_5

f \ x	0	1	2	3	4
r	1	1	0	0	2
$r \circ 2$	1	0	2	1	0
$r \circ 3$	1	0	1	2	0
$r \circ 4$	1	2	0	0	1
$r \circ 0$	1	1	1	1	1
$2 \circ r$	2	2	0	0	1
$2 \circ r \circ 2$	2	0	1	2	0
$2 \circ r \circ 3$	2	0	2	1	0
$2 \circ r \circ 4$	2	1	0	0	2
$2 \circ r \circ 0$	2	2	2	2	2

B. The 3-coloring r of Z_{12}

x	0	1	2	3	4	5	6	7	8	9	10	11
$r(x)$	1	1	1	1	0	0	0	0	2	2	2	2

C. The 8-coloring r of Z_{11}

x	0	1	2	3	4	5	6	7	8	9	10
$r(x)$	1	1	0	0	7	7	6	5	4	3	2

The most important 3-coloring of Z_5 is r, because it can be used to generate all of the others. For example, the value at x of the coloring labeled $r \circ 2$ is obtained by first multiplying x by 2 (in Z_5) and then applying r; in symbols, $(r \circ 2)(x) = r(2x)$. The first five colorings in TABLE 1A are in this way the "multiples" of r, namely r, $r \circ 2$, $r \circ 3$, $r \circ 4$, and $r \circ 0$. The last five colorings in TABLE 1A are derived in turn from the first five by subsequent multiplication (in Z_3) by 2. For example, the value at x of the coloring labeled $2 \circ r$ is obtained by multiplying $r(x)$ by 2 (in Z_3); in symbols, $(2 \circ r)(x) = 2r(x)$.

The table thus illustrates a particular case of the main theorem of this paper. Since Z_5 is abelian, $(Z_5)_{ab} = Z_5$, and consequently the theorem states that there are two families of five 3-colorings of Z_5. These ten colorings are exactly the ones shown in the table. It is easy to see how the single 3-coloring r of Z_n will, similarly, always generate n 3-colorings by pre-multiplication in Z_n, and n more by post-multiplying these by 2 in Z_3. The technically more difficult part of the theorem is to prove that these $2n$ colorings are the *only* ones; this will be done in section 3 below.

For the moment, though, we should consolidate our position with a formal proposition listing the properties of colorings suggested from the examples.

PROPOSITION 1.1. *If f: $G \to Z_n$ is a coloring, then*:

a) $f(0) \neq 0$,

b) $w \circ f \circ h$ *is a coloring for any one-to-one homomorphism* w: $Z_n \to Z_m$ *and homomorphism* h: $H \to G$.

Proof. How does one prove that a group homomorphism g preserves the identity? Just say $g(0) = g(0+0) = g(0) + g(0)$ and cancel $g(0)$ from both sides to conclude $g(0) = 0$. How does one prove a coloring never preserves the identity? It's fun and easy: just insert a non-equality sign in the same proof: $f(0) = f(0+0) \neq f(0) + f(0)$, so $f(0) \neq 0$. The reader should try the process to prove $f(x-y) \neq f(x) - f(y)$.

b) How does one prove a composite of group homomorphisms is a group homomorphism? One says $(g'g)(x+y) = g'g(x+y) = g'(gx+gy) = g'g(x) + g'g(y)$. Now insert a non-equality sign: $fh(x+y) = f(hx+hy) \neq fh(x) + fh(y)$. Since w is one-to-one, it preserves the non-equality, and so $wfh(x+y) \neq w[fh(x) + fh(y)] = wfh(x) + wfh(y)$.

Proposition 1.1b is basic and important. We've already seen how it can be used to classify the colorings of Z_5 (or of Z_n). In the language of category theory, it says that n-coloring is a contravariant functor. It can also be used to generate colorings of large groups via colorings of their smaller factor groups. For example, composing the canonical quotient map $Z \to Z_5$ with the 3-coloring r of Z_5 gives a 3-coloring of the integers Z which repeats with period 5:

x	...	−3	−2	−1	0	1	2	3	4	5	6	7	8	9	10	...
$f(x)$	...	0	0	2	1	1	0	0	2	1	1	0	0	2	1	...

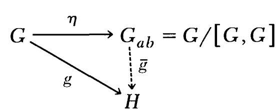
The full classification of the 3-colorings of the integers (are they all periodic?) remains an open problem!

2. Commutativity

In this section the whole issue of 3-colorings of a finite group G will be reduced to the abelian case. Since we will deal explicitly with nonabelian groups, we will change the notation a little and write G *multiplicatively* (while continuing to write Z_3

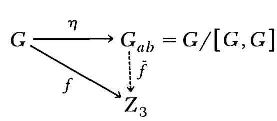
additively). The defining condition of a 3-coloring thus reads: $f(xy) \neq f(x)+f(y)$. The identity element of G will be denoted 1.

The reader will recall that any group G has a maximal abelian quotient group G_{ab}, which is obtained by factoring out the commutator subgroup $[G,G]$ generated by the commutators $[x,y]=x^{-1}y^{-1}xy$ in G. Since $[x,y]=1$ if and only if x and y commute with each other, factoring out $[G,G]$ eliminates precisely the non-commuting pairs of G. $G_{ab}=G/[G,G]$ is consequently abelian. It also has the universal property that any group homomorphism $g\colon G \to H$ mapping G to an *abelian* group H can be factored through G_{ab}; that is, there is a unique homomorphism $\bar{g}\colon G_{ab} \to H$ so that $g=\bar{g}\eta$, where $\eta\colon G \to G_{ab}$ is the quotient map.



The map $\bar{g}$ is defined on each coset $x[G,G]$ of G_{ab} by $\bar{g}(x[G,G])=g(x)$, and this definition works because g vanishes on elements of $[G,G]$: $g([x,y])=g(x^{-1}y^{-1}xy)$ $=g(x)^{-1}g(y)^{-1}g(x)g(y)=1$ because $g(x)$ and $g(y)$ are elements of the abelian group H.

In this section, we will show similarly that any 3-coloring $f\colon G \to Z_3$ of a finite group G factors through G_{ab}, that is, that there is a unique 3-coloring $\bar{f}\colon G_{ab} \to Z_3$ so that $f=\bar{f}\eta$.



The definition of $\bar{f}$ is again the obvious one: $\bar{f}(x[G,G])=f(x)$, and a critical step will be to show that this is well-defined, by demonstrating that f is constant on $[G,G]$.

Because of this factorization, the 3-colorings of G are completely determined by those of its abelian factor group G_{ab}. All we then need in order to complete the analysis is a description of the 3-colorings of abelian groups (sections 3 and 4 below).

To factor f through G_{ab}, we first demonstrate that the action of f is "commutative," that is, that $f(ab)=f(ba)$.

PROPOSITION 2.1. *If f is a 3-coloring of a finite group G, then $f(ab)=f(ba)$ for all $a,b \in G$.*

Proof. Suppose there are elements $a,b \in G$ so that $f(ab) \neq f(ba)$. We first show that without loss of generality we may assume (i) that $f(1)=1$, and (ii) that $f(ab)=f(a)+f(b)-1$ and $f(ba)=f(a)+f(b)+1$.

i) Since $f(1) \neq 0$ (Proposition 1.1a, in multiplicative notation), either $f(1)=1$ or $f(1)=2$. If $f(1)=2$, we could replace f throughout the proof by the coloring $f'=-1\circ f$, which also satisfies the hypothesis that $f'(ab) \neq f'(ba)$ and does satisfy $f'(1)=1$.

ii) Since $f(ab) \neq f(a)+f(b)$, $f(ab)$ is either $f(a)+f(b)+1$ or $f(a)+f(b)-1$; the same is true for $f(ba)$. Switching the names a and b if necessary, we may assume that $f(ab)=f(a)+f(b)-1$ and $f(ba)=f(a)+f(b)+1$.

Z_3 has only three elements, and so to calculate in this "algebra of non-equalities" one can use a characteristic type of argument: if $x,y \in Z_3$ with $x \neq y+1$ and $x \neq y-1$, then necessarily $x=y$. We use this argument to evaluate $f(aba)$ by breaking up the product in two different ways:

$$f(a\cdot b\cdot a)\neq f(a)+f(ba)=-f(a)+f(b)+1$$
$$f(a\cdot b\cdot a)\neq f(ab)+f(a)=-f(a)+f(b)-1.$$

Therefore, $f(aba)=-f(a)+f(b)$.

We continue evaluating higher and higher products by this procedure until we reach a contradiction: Breaking the products similarly in two ways at the dots, $f(b\cdot a\cdot b)=f(a)-f(b)$, $f(ab\cdot a\cdot b)=-f(a)-f(b)-1$, $f(ba\cdot b\cdot a)=-f(a)-f(b)+1$, $f(ab\cdot a\cdot ba)=-f(b)$, $f(ba\cdot b\cdot ab)=-f(a)$, $f(aba\cdot b\cdot ab)=2$, and $f(bab\cdot a\cdot ba)=1$. Inductively, $f[(ab)^{3n}]=2$ and $f[(ab)^{3n}a]=f(a)$ for all positive integers n, because $f[(ab)^{3n}\cdot a\cdot babab]=2$ and $f[(ab)^{3n}a\cdot babab\cdot a]=f(a)$. This contradicts the finiteness of G, since for some n, $[(ab)^3]^n=1$ and $f[(ab)^{3n}]=f(1)=1$.

We next apply Proposition 2.1 to conclude that f does not "see" commutators at all:

PROPOSITION 2.2. *If f is a 3-coloring of a finite group G, then $f(xy)=f(x)$ for all $x\in G$ and $y\in[G,G]$.*

Proof. Assume, as in Proposition 2.1, that $f(1)=1$. We first apply Proposition 2.1 and the rule $f(xy^{-1})\neq f(x)-f(y)$ (*cf.* proof of Proposition 1.1a) to conclude that f is never zero on commutators:

$$f([a,b])=f(a^{-1}b^{-1}ab)\neq f(a^{-1}b^{-1}a)-f(b^{-1})=f(b^{-1})-f(b^{-1})=0.$$

We next use this to prove that f is always 1 on commutators: If $f([a,b])=2$, then inductively $f([a^n,b])=2$, because $f([a^{n+1},b])\neq f(a^{-1}[a^n,b]a)+f([a,b])=f([a^n,b])+f([a,b])=1$. But $a^n=1$ for some n, because G is finite, a contradiction.

We now evaluate $f(x[a,b])$ by using our characteristic argument in Z_3 and the commutator identity $[a,b]=[b,a]^{-1}$: For all $a,b,x\in G$,

$$f(x[a,b])\neq f(x)+f([a,b])=f(x)+1$$
$$f(x[a,b])\neq f(x)-f([b,a])=f(x)-1$$
$$f(x[a,b])=f(x).$$

Since any $y\in[G,G]$ is a product of commutators, the conclusion $f(xy)=f(x)$ follows by iteration.

We can now finally factor f through G_{ab}:

THEOREM 2.1. *If f is a 3-coloring of a finite group G, there is a unique 3-coloring $\bar{f}$: $G/[G,G]\to Z_3$ so that $f=\bar{f}\eta$, where η: $G\to G/[G,G]$ is the canonical quotient map.*

Proof. Proposition 2.2 guarantees that $\bar{f}$, given by $\bar{f}(x[G,G])=f(x)$, is well-defined and it is easy to check that this $\bar{f}$ has the desired properties.

Both restrictions, that f be a 3-coloring and that G be finite, are necessary in Theorem 2.1:

COUNTEREXAMPLE 2.1 (A non-factoring 4-coloring). Let $G=S_3$ (the symmetric group on 3 letters) $=\langle a,b\colon a^3=b^2=1, bab=a^2\rangle$, and define a 4-coloring f by:

x	1	a	a^2	b	ab	a^2b
$f(x)$	1	1	3	1	1	3

(Verifying that this is a 4-coloring will show one reason why there are so many 4-colorings.) The commutator group $[G,G]$ is $\{1,a,a^2\}$, but a and a^2, in the same coset of $G/[G,G]$, are mapped to different values by f. Hence, f cannot factor through G_{ab}. The reader can check that this counterexample also violates Propositions 2.1 and 2.2.

COUNTEREXAMPLE 2.2 (A non-factoring 3-coloring of an infinite group). Let $G = \langle a, b \rangle$, the infinite free group on two generators. As described in Appendix 1, G can be partitioned into two disjoint sets, A and B, each closed under multiplication, with $ab \in A$ and $ba \in B$. The 3-coloring f defined by:

$$f(x) = \begin{cases} 1 & \text{if } x \in A \\ 2 & \text{if } x \in B \end{cases}$$

cannot factor through G_{ab}, because $f(ab) \neq f(ba)$ whereas ab and ba have the same image in G_{ab}.

3. Cyclic Groups

Theorem 2.1 cuts the problem of 3-colorings down to the commutative case. We will do the cyclic groups here, and arbitrary finite abelian groups in the next section. Since these groups are all abelian, we will return to additive notation [a coloring satisfies $f(x+y) \neq f(x) + f(y)$], and we will also assume from now on that $f(0) = 1$ (*cf.* proof of Proposition 2.1).

As discussed back in section 1, the standard coloring r of Z_n generates a family of colorings $r \circ \alpha_x$, where α_x: $Z_n \to Z_n$ is multiplication by $x \in Z_n$. The goal of this section of the paper is to prove that every 3-coloring of Z_n is in this family $\{r \circ \alpha_x\}$. Note that the colorings $r \circ \alpha_x$, $x = 0, 1, \ldots, n-1$, are all distinct, and so the representation $f = r \circ \alpha_x$ is unique. For, if $\alpha_x \neq \alpha_y$, then $r \circ \alpha_x = r \circ \alpha_y$ implies $r \circ \alpha_{(x-y)}(t) = r(xt - yt) \neq r(xt) - r(yt) = 0$ for all $t \in Z_n$. Thus r is nonzero on the whole subgroup $(x-y)Z_n$, a contradiction since r is by definition 0 on the "middle third" of Z_n.

We will begin with a number-theoretic lemma. The reader may recall that there is a famous unproved conjecture of Goldbach that any even number greater than 2 may be written as a sum of two primes.[3] Here we will state (and prove!) a finite analogue:

LEMMA 3.1 ("Finite Goldbach"). *For any positive integers n and m, there are integers a and b relatively prime to $2m$ so that $2n \equiv a + b \bmod 2m$.*

For example, when $m = 5$, the numbers relatively prime to $2m = 10$ are 1, 3, 7, and 9, and sums are: $0 \equiv 1 + 9$, $2 \equiv 1 + 1$, $4 \equiv 1 + 3$, $6 \equiv 3 + 3$, and $8 \equiv 1 + 7$.

Proof. Induction on the number of prime factors of $2m$: If $2m = 2^k$, then $2n \equiv (2n-1) + 1$ gives the required decomposition. Let $2m = 2cd$, d a power of a prime not dividing $2c$. Then $2n$ may be written $2n = t \cdot 2c + u \cdot d$ for some integers t and u (u even). Now decompose t and u: Inductively, u can be expressed as $u \equiv a' + b' \bmod 2c$, with a', b' prime to $2c$. One may write t as $t \equiv a'' + b'' \bmod d$, with a'' and b'' prime to d, simply by choosing $a'' = 1$ or 2. It now follows that $2n \equiv (a'' \cdot 2c + a'd) + (b'' \cdot 2c + b'd) \bmod 2m$, the desired form.

This lemma can be used to show that an arbitrary 3-coloring f is a "multiple" of a 3-coloring g having a standard form:

[3] Early work on this conjecture is described in Dickson's monumental history [3]; more recent developments are summarized in Hardy and Wright's classic text [7].

PROPOSITION 3.1. *Any 3-coloring f of Z_n $(n > 3)$ may be expressed in the form $f = g \circ \alpha_x$, where $g(1) = 1$.*

Proof. The proof is given in two steps. The first is to show that f is a "multiple" of a coloring g with $g(1) \neq 0$. We consider three cases:

i) If n is odd, $f(1) = 0 \Rightarrow f(2) \neq f(1) + f(1) = 0$. Define $g = f \circ \alpha_2$; then $g(1) = f(2) \neq 0$ and $f = g \circ \alpha_{(n+1)/2}$.

ii) If $f(y) \neq 0$ for some y which is relatively prime to n and thus has a multiplicative inverse x in Z_n, take $g = f \circ \alpha_y$, whence $g(1) \neq 0$ and $f = g \circ \alpha_x$.

iii) The remaining case, n is even and $f(y) = 0$ for every y relatively prime to n, is tailor-made for the finite Goldbach lemma. We prove $f = r \circ \alpha_{(n/2)}$: For any even element $2k$, there are a, b relatively prime to n with $2k = a + b$, so $f(2k) \neq f(a) + f(b) = 0$. Hence, f restricted to the even numbers of Z_n is never 0. This, in turn, implies that f on the even numbers is always 1. For if $f(2k) = 2$, then $f(4k) \neq f(2k) + f(2k) = 1 \Rightarrow f(4k) = 2$, and inductively $2 = f(2k) = f(4k) = \cdots = f(2nk)$, contradicting the fact that $f(2nk) = f(0) = 1$. Now f on the odd numbers is always 0, because $f(2k+1) \neq f(2k) + f(1) = 1 + 0 = 1$ and $2f(2k+1) \neq f(4k+2) = 1$. Thus

$$f(k) = \begin{cases} 0 & k \text{ odd} \\ 1 & k \text{ even} \end{cases}$$

which by definition of r implies $f = r \circ \alpha_{(n/2)}$.

The second step in proving the proposition is to find g with $g(1) = 1$. We just found a g with $g(1) \neq 0$, and there is no problem if $g(1) = 1$ (done!) or if $g(1) = 2$ but $g(-1) = 1$ (take $g \circ \alpha_{-1}$). Now $g(1) = 2$ implies $g(-1) \neq 2$, because $g(1) + g(-1) \neq g(0) = 1$, so the only bad case is when both $g(1) = 2$ and $g(-1) = 0$. In this situation,

$$g(k+1) \neq g(k) + g(1) = g(k) + 2$$
$$g(k+1) \neq g(k) - g(-1) = g(k).$$

Therefore,

$$g(k+1) = g(k) + 1.$$

The 3-coloring g thus has the form:

k	0	1	2	3	4	5	$\cdots$
$g(k)$	1	2	0	1	2	0	$\cdots$

The integer n must be a multiple of 3, in order that $g(n) = g(0)$. It now follows from the definition of r that $g = r \circ \alpha_{(2n/3)}$. This completes the proof.

According to this proposition, one may concentrate on the case $f(1) = 1$. Since $f(x+1) \neq f(x) + f(1) = f(x) + 1$, either $f(x+1) = f(x)$ or $f(x+1) = f(x) + 2$, that is, the value of f does not change as x increases except in jumps of 2. Thus the table of f is:

		c_1	c_2	c_3			c_m
x	0	$1 \cdots c_1$	$c_1+1 \cdots c_1+c_2$	$c_1+c_2+1 \cdots$		$\cdots$	
$f(x)$	1	$1 \cdots 1$	$0 \cdots 0$	$2 \cdots 2$	$1 \cdots 1$	$\cdots$	$2 \cdots 2$

Here $\Sigma c_i = n - 1$, and m must be a multiple of 3 so that $f(x)$ may progress properly through $m/3$ complete cycles $1 \ldots 0 \ldots 2 \ldots$. Note that r by definition has only one such cycle $(m/3 = 1)$; more generally, one may show that $r \circ \alpha_k$ has k cycles whenever $1 \leqslant k < n/3$.

Using the definition of a 3-coloring, one can derive tight constraints on the integers c_i which allow each c_i or cumulative sum of c_i's a leeway of only 1. For example,

$f(c_1+c_1+2)\neq f(c_1+1)+f(c_1+1)=0$, whence $c_1+2>c_2$, or $c_1+1\geqslant c_2$. An opposite bound comes from $f(c_1+x)\neq f(c_1)+f(x)=2$ for all $x=1,2,\dots,c_1$, whence $c_1\leqslant c_2$. Similar proofs establish that $c_1\leqslant c_k\leqslant 1+c_1$ for all $k=1,2,\dots,m$, and by induction on s one finds:

PROPOSITION 3.2.

a) $$\sum_{i=1}^{s} c_i \leqslant \sum_{i=k+1}^{k+s} c_i \leqslant 1+\sum_{i=1}^{s} c_i \qquad (k+s\leqslant m)$$

b) $$\sum_{i=1}^{s} c_{m-i+1} \leqslant \sum_{i=k+1}^{k+s} c_{m-i+1} \leqslant 1+\sum_{i=1}^{s} c_{m-i+1} \qquad (k+s<m).$$

These constraints then determine the c_i's completely. If, for example, numbers $\{d_i\}$ also satisfied them, with $\Sigma d_i=n=\Sigma c_i$ but with $d_1>c_1$, then $d_i\geqslant c_i$ for all $i>1$ because $d_i\geqslant d_1>c_1\geqslant c_i-1$, leading to the contradiction $n=\Sigma d_i>\Sigma c_i=n$. In the general case, $c_1=d_1,\dots,c_k=d_k$ and $c_{m-k+1}=d_{m-k+1},\dots,c_m=d_m$ follows by induction on k.

Combining all the information gives the main conclusion of this section:

THEOREM 3.1. *Any 3-coloring f of Z_n (with $f(0)=1$) has the form $f=r\circ\alpha_k$ for some unique $k\in Z_n$.*

Proof. The special cases $n=1$, 2, or 3 can easily be checked individually. When $n>3$, one may assume $f(1)=1$ (Proposition 3.1). Now we simply count the number of possible 3-colorings. Unless f is identically 1 (that is, $f=r\circ\alpha_0$), it is completely specified by the integers $\{c_i\}$; these are uniquely determined by m, a positive multiple of 3 which is less than n. Writing $n=3t+b$ for $0<b\leqslant 3$, there are t such multiples, hence at most t possibilities for f. But since r is defined $r(x)=1$ for $x=0,1,2,\dots,t$, there are also precisely t non-constant colorings $g_k=r\circ\alpha_k$ satisfying $g_k(1)=1$, namely those for $k=1,2,\dots,t$. These then exhaust the possible values for m, and $f=g_k=r\circ\alpha_k$ for some k.

4. 3-Colorings of Finite Abelian Groups

We will finish off the problem by analyzing the 3-colorings of arbitrary finite abelian groups. The program of attack is the same as was used in the previous section on cyclic groups:

1) Define a canonical 3-coloring $\bar{r}$.
2) Generate a family of multiples of $\bar{r}$.
3) Prove that any 3-coloring is such a multiple.

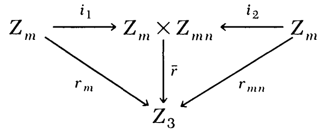
The reader will recall that any finite abelian group can be expressed in several standard forms as a product of cyclic groups [4, 8]. The one to be used here is:

$$G\cong Z_{t_1}\times\cdots\times Z_{t_j} \quad \text{with } t_1|t_2|\cdots|t_j,$$

that is, each t_i is a factor of the next term t_{i+1}. For simplicity, we illustrate the proof only in the special case $G=Z_m\times Z_{mn}$. The point of using the decomposition $G=Z_m\times Z_{mn}$ is that there is a natural embedding $\chi_n\colon Z_m\to Z_{mn}$ given by multiplication by n, that is, for $x\in Z_m$, $\chi_n(x)=nx\in Z_{mn}$. We can use this and the canonical 3-coloring $r=r_{mn}$ of Z_{mn} to define $\bar{r}$:

$$\bar{r}(x,y) = r(\chi_n x + y).$$

(It's obvious how to extend this to the general case $G = Z_{t_1} \times \cdots \times Z_{t_j}$: just embed everything in Z_{t_j}). The reader can check easily that $\bar{r}$ is a coloring. Further, it equals r when restricted to either component of G, that is,



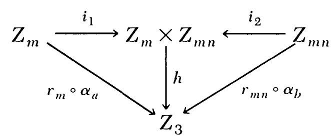
commutes, where i_1 and i_2 are inclusions into the first and second co-ordinate and r_m: $Z_m \to Z_3$ and r_{mn}: $Z_{mn} \to Z_3$ are the standard 3-colorings. It will follow from the theory we develop later (Appendix 2) that $\bar{r}$ is the *unique* 3-coloring with this property.

Just as for cyclic groups, we can now define multiples of $\bar{r}$: simply select $a \in Z_m$, $b \in Z_{mn}$, and premultiply componentwise

$$\bar{r} \circ \alpha_{(a,b)}(x,y) = \bar{r}(ax, by).$$

Checking on each component, one can then show that all $m \cdot mn$ 3-colorings $\bar{r} \circ \alpha_{(a,b)}$, $a \in Z_m$, $b \in Z_{mn}$, are distinct.

Suppose now that h: $Z_m \times Z_{mn} \to Z_3$ is an arbitrary 3-coloring. (As usual, assume $h(0) = 1$.)



When h is restricted to Z_m, it is a 3-coloring which can be analyzed by Theorem 3.1; thus $h \circ i_1 = r_m \circ \alpha_a$ for some unique $a \in Z_m$, where r_m is the standard 3-coloring of Z_m. Similarly, $h \circ i_2 = r_{mn} \circ \alpha_b$ for a unique $b \in Z_{mn}$. We simply use a and b as multipliers and define $\bar{h}$: $Z_m \times Z_{mn} \to Z_3$ by $\bar{h} = \bar{r} \circ \alpha_{(a,b)}$. This yields a 3-coloring $\bar{h}$ which equals $r_m \circ \alpha_a$ when restricted to Z_m and $r_{mn} \circ \alpha_b$ when restricted to Z_{mn}, just like h. A somewhat technical argument (Appendix 2) shows that the product representation is unique, and therefore $\bar{h} = h$.

We have, therefore, proved:

THEOREM 4.1. *Any 3-coloring h of $G = Z_{t_1} \times \cdots \times Z_{t_j}$ $(t_1|t_2|\cdots|t_j)$, with $h(0) = 1$, has the form $h = \bar{r} \circ \alpha_k$ for some $k \in G$.*

5. Summary and Conclusions

Graph coloring theory includes a number of remarkable concepts and results that represent its distinctive contribution to mathematics. Noteworthy is the profoundness of these results, which proceed from particularly "simple" hypotheses (*e.g.*, a graph is planar) to equally "simple" conclusions (*e.g.*, it is 4-colorable) by extraordinarily complex chains of reasoning. Also unique is the "non-deterministic" nature of the coloring process itself: One constructs a coloring not in a rigid, mechanical, syllogistic fashion, but rather with apparent freedom or elasticity of choice at each step. Often it is only at the final stages that the process may "crystallize" to a surprisingly restricted conclusion. Likewise unusual is the fact that the coloring is built up by a process of denial: at each step one is not told what color the next vertex must be, but rather what colors it may *not* be.

In this paper we have tried to mimic the graph coloring process in an algebraic context by using a similar rule of denial, $f(x+y) \neq f(x)+f(y)$. Just as the set of colorings of a graph reflects some of the topological structure of the graph, we have found that the set of colorings of a group partially reflects its algebraic structure: For a finite group, the 3-colorings give us two copies of the abelianized quotient. Our analysis proceeded in steps, in which we first factored the 3-colorings through the quotient group (section 2), then classified them for finite cyclic groups (sections 1 and 3) and thereafter for arbitrary finite abelian groups (section 4).

Here's the final conclusion:

THEOREM 5.1. *Any 3-coloring of a finite group G may be expressed uniquely in the form* $f = \pm \bar{r} \circ \alpha_x \circ \eta$, *where* $\eta: G \to G/[G,G] = G_{ab}$ *is the canonical quotient map,* $\alpha_x: G_{ab} \to G_{ab}$ *is componentwise multiplication determined by* $x \in G_{ab}$, $\bar{r}$ *is the standard 3-coloring of* G_{ab}, *and* $\pm$ *denotes multiplication by* $\pm 1 \in Z_3$. *Hence the 3-colorings form two families, according as* $f(0) = \pm 1$, *each endowed with the group structure of* G_{ab} *under the operation* $(\bar{r} \circ \alpha_x \circ \eta) + (\bar{r} \circ \alpha_y \circ \eta) = \bar{r} \circ \alpha_{(x+y)} \circ \eta$.

Proof. Combine Theorem 2.1, which factors f through G_{ab}, with Theorem 4.1, which classifies the 3-colorings of the finite abelian group G_{ab}.

Translation: The set of colorings of G contains structural information about G.

Acknowledgement. The author thanks Professor Peter J. Hilton, two referees, and the editor for their encouragement and advice concerning the manuscript.

REFERENCES

1. K. Appel and W. Haken, Every planar map is four colorable. Part I: Discharging, *Ill. J. Math.* 21 (1977), 429–490; K. Appel, W. Haken, and J. Koch, Every planar map is four colorable. Part II: Reducibility, *Ill. J. Math.* 21 (1977), 491–567.
2. Kenneth Appel and Wolfgang Haken, The solution of the four-color-map problem, *Scientific American* 237 (1977), 108–121.
3. Leonard Eugene Dickson, *History of the Theory of Numbers*, Chelsea, 1966 reprint of 1919 edition, vol. 1, pp. 421–424.
4. John B. Fraleigh, *A First Course in Abstract Algebra*, Addison-Wesley, 3d ed., 1982.
5. Israel Grossman and Wilhelm Magnus, *Groups and Their Graphs*, Random House, 1964.
6. Frank Harary, *Graph Theory*, Addison-Wesley, 1969.
7. G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers*, Oxford Univ. Press, 5th ed., 1979, pp. 415–416.
8. Saunders Mac Lane and Garrett Birkhoff, *Algebra*, Macmillan, 1967.
9. Gerhard Ringel, *Map Color Theorem*, Springer-Verlag, 1974.
10. Thomas L. Saaty and Paul C. Kainen, *The Four-Color Problem: Assaults and Conquest*, McGraw-Hill, 1977.
11. Lynn Arthur Steen, Solution of the four color problem, this MAGAZINE, 49 (1976), 219–222.
12. Joel G. Stemple, The four color problem, *Ann. N. Y. Acad. Sci.* 321 (1979), 91–101.
13. Arthur T. White, *Graphs, Groups and Surfaces*, North-Holland, 1973.

Appendix 1. Partition of the Free Group on Two Generators into Two Sets Closed under Multiplication

Counterexample 2.2, a "nonfactoring" 3-coloring of an infinite group, depended on partitioning the free group $G = \langle a, b \rangle$ into two sets A and B closed under multiplication. This partition will be constructed here. G consists of elements written (uniquely) as products of the form

$$x = a^{n_1}b^{m_1}a^{n_2}b^{m_2} \cdots a^{n_k}b^{m_k},$$

with n_1 or m_k possibly 0. We first partition G into three sets, according as the sum of the exponents of a is greater than, less than, or equal to the sum of the exponents of b:

$$A_0 = \{x : \textstyle\sum n_i > \sum m_i\}$$
$$B_0 = \{x : \textstyle\sum n_i < \sum m_i\}$$
$$C_0 = \{x : \textstyle\sum n_i = \sum m_i\}.$$

Clearly A_0, B_0, and C_0 are each closed under multiplication. Further, the elements of A_0 and B_0 remain in their respective sets when multiplied by elements of C_0.

We next divide C_0 into two sets which will be apportioned between A_0 and B_0 to produce A and B. To divide C_0, define a function H on the elements of C_0 by

$$H(x) = \sum_{i,j} (-1)_{j>i} m_i n_j$$

where

$$(-1)_{j>i} = \begin{cases} -1 \text{ if } j > i \\ 1 \text{ if } j \leqslant i. \end{cases}$$

H satisfies the rule $H(xy) = H(x) + H(y)$, as can be seen most easily by displaying the products occurring in $H(xy)$ in the form of a matrix. If, for instance, $x = a^{n_1}b^{m_1}a^{n_2}b^{m_2}$ and $y = a^{n_3}b^{m_3}a^{n_4}b^{m_4}$, then $H(xy)$ is the sum of the products in the array:

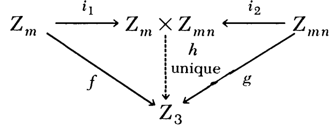
$$\begin{array}{cc|cc} m_1n_1 & -m_1n_2 & -m_1n_3 & -m_1n_4 \\ m_2n_1 & m_2n_2 & -m_2n_3 & -m_2n_4 \\ \hline m_3n_1 & m_3n_2 & m_3n_3 & -m_3n_4 \\ m_4n_1 & m_4n_2 & m_4n_3 & m_4n_4 \end{array}$$

where minus signs occur only above the main diagonal. The matrix can be divided into blocks as shown, and $H(x)$ is the sum of the products in the upper left quadrant, $H(y)$ of those in the lower right. We need only to verify that the terms in the remaining two quadrants sum to zero. But since $x \in C_0$, $n_1 + n_2 = m_1 + m_2$, and so the n_i's occurring in the sum from the lower left quadrant may be replaced by m_i's. The same happens in the upper right quadrant because $y \in C_0$, and these terms then precisely cancel those from the lower left quadrant.

Let $C_+ = \{x \in C_0: H(x) \geqslant 0\}$, $C_- = \{x \in C_o: H(x) < 0\}$, and define $A = A_o \cup C_+$, $B = B_o \cup C_-$. A and B are then closed and partition G as desired, with $ab \in C_+ \subset A$ and $ba \in C_- \subset B$.

Appendix 2. Uniqueness of a 3-Coloring of a Product

In this appendix we will prove that a 3-coloring h of the product $Z_m \times Z_{mn}$ is uniquely determined by its restrictions f and g on the factors Z_m and Z_{mn}.



Proofs of uniqueness in algebra are usually quite straightforward, and so we should begin by seeing what the problem is here. If f, g, and h were group homomorphisms rather than colorings, there would be at least three different easy approaches to prove uniqueness; it is illuminating to consider what properties homomorphisms have that we usually take for granted which make them so much easier to handle.

1) The homomorphism h is unique because it is given by a formula: $h(a,b) = h[(a,0)+(0,b)] = h(a,0)+h(0,b) = f(a)+g(b)$. For the coloring h, the second equality becomes instead a non-equality, and there is no such formula. We must use the trick of writing the product as $Z_m \times Z_{mn}$ and embedding Z_m in Z_{mn} even to *define* the product coloring $\bar{r}$.

2) The homomorphism h is unique because two homomorphisms coincide on a subgroup of their domain, and a subgroup which includes both factors of a product must be the entire group. In contrast, the set on which two colorings coincide need not be a group (see TABLE 1 of section 1 for examples).

3) The homomorphism h is unique because the difference of two abelian group homomorphisms is itself a homomorphism, and a homomorphism will be 0 on a product if it is 0 on the factors. The second half of this argument has a parallel: A coloring that is identically 1 on the factors is also 1 on the product. However, the first part fails because the difference of two colorings seems in no way related to a coloring itself.

For 3-colorings, we must therefore use a much more roundabout procedure, and prove the result first in three special cases (Proposition A2.1a, b, and c) before doing the general case (Proposition A2.1d).

PROPOSITION A2.1. *The* 3-*coloring* h (*indicated by a dotted arrow*) *is unique in each of the following diagrams*:

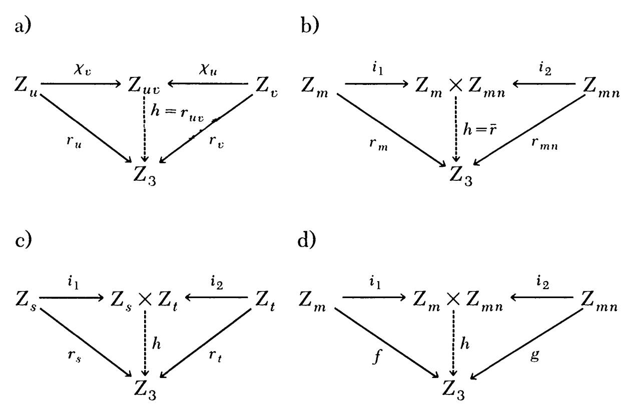


FIGURE 3

where m, n, s, *and* t *are arbitrary positive integers, and* u *and* v *are relatively prime.*

Proof. a) Note first that $h = r_{uv}$ will make the diagram commute for any positive integers u and v. This compatibility of the r's and the embeddings χ follows because, in the terminology of the proof of Theorem 3.1, r_u is the unique 3-coloring of Z_u for which $m = 3$.

The coloring h is unique when u and v are relatively prime because, by Theorem 3.1, $h = r_{uv} \circ \alpha_k$ for some $k \in Z_{uv}$. Thus $r_u = h \circ \chi_v = r_{uv} \circ \alpha_k \circ \chi_v = r_{uv} \circ \chi_v \circ \alpha_k = r_u \circ \alpha_k$, which implies $k \equiv 1 \bmod u$. Similarly, $k \equiv 1 \bmod v$, so $k \equiv 1 \bmod uv$, or $h = r_{uv}$.

b) Let $x=(1,-n)\in Z_m\times Z_{mn}$. For any $k\in Z$, $h(kx)=h(k,-nk)\neq h(k,0)-h(0,nk)=r_m(k)-r_{mn}(nk)=0$. It then follows that $h(kx)=1$: If $h(kx)=2$, $h(2kx)\neq h(kx)+h(kx)=1$, and, inductively, $2=h(kx)=h(2kx)=h(3kx)=\cdots$, contradicting the fact that $h(mkx)=h(0)=1$. Consequently,

$$\begin{aligned}h(a,b)=h[(0,b+na)+ax]&\neq h(0,b+na)+h(ax)=r_{mn}(b+na)+1\\&\neq h(0,b+na)-h(-ax)=r_{mn}(b+na)-1,\end{aligned}$$

and so $h(a,b)=r_{mn}(b+na)=\bar{r}(a,b)$.

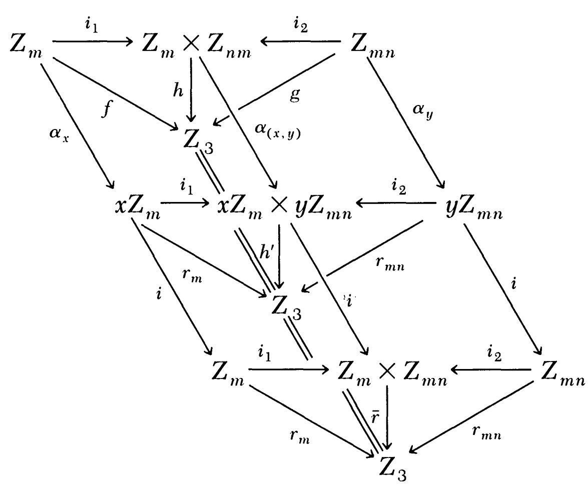
c) One must rearrange $Z_s\times Z_t$ into a product of the form $Z_m\times Z_{mn}\times\cdots$ to which part b can be applied. A specific example, $Z_{12}\times Z_{10}$, serves to indicate the process. We use a series of three isomorphisms:

$$Z_{12}\times Z_{10}\xrightarrow{\iota}(Z_3\times Z_4)\times(Z_2\times Z_5)\xrightarrow{\kappa}Z_2\times(Z_3\times Z_4\times Z_5)\xrightarrow{\tau}Z_2\times Z_{60},$$

where ι and τ are induced by isomorphisms of the form $Z_{12}\cong Z_3\times Z_4$, and κ denotes rearrangement of terms of the product.

On each factor, $h\circ\iota^{-1}$ is r because all r's are compatible (part a); $h\circ\iota^{-1}\circ\tau^{-1}$ is r because κ merely rearranges the factors; $h\circ\iota^{-1}\circ\kappa^{-1}\circ\tau^{-1}$ is r by part a; and now $h=\bar{r}\circ\tau\circ\kappa\circ\iota$ by part b.

d) By Theorem 3.1, $f=r_m\circ\alpha_x$ and $g=r_{mn}\circ\alpha_y$ for some $x\in Z_m$ and $y\in Z_{mn}$. We factor the maps α_x and α_y through their images as $Z_m\xrightarrow{\alpha_x}xZ_m\xrightarrow{i}Z_m$ and $Z_{mn}\xrightarrow{\alpha_y}yZ_{mn}\xrightarrow{i}Z_{mn}$, where the maps i are inclusions. The crux of the problem is that the images xZ_m and yZ_{mn} are in general proper subgroups of Z_m and Z_{mn}, the order of xZ_m may not divide that of yZ_{mn}, and so the proof of Theorem 4.1 cannot be applied directly.



We instead define a coloring $h'\colon xZ_m\times yZ_{mn}\to Z_3$ so that $h=h'\circ\alpha_{(x,y)}$, by setting $h'(xa,yb)=h(a,b)$. It is trivial to check that h' is a 3-coloring and $h=h'\circ\alpha_{(x,y)}$; the crucial point is to verify first that this h' is well-defined. If $xa=xa'$, then $r_m\circ\alpha_x(a-a')=r_m(0)=1$ and $r_m\circ\alpha_x(a'-a)=r_m(0)=1$, so

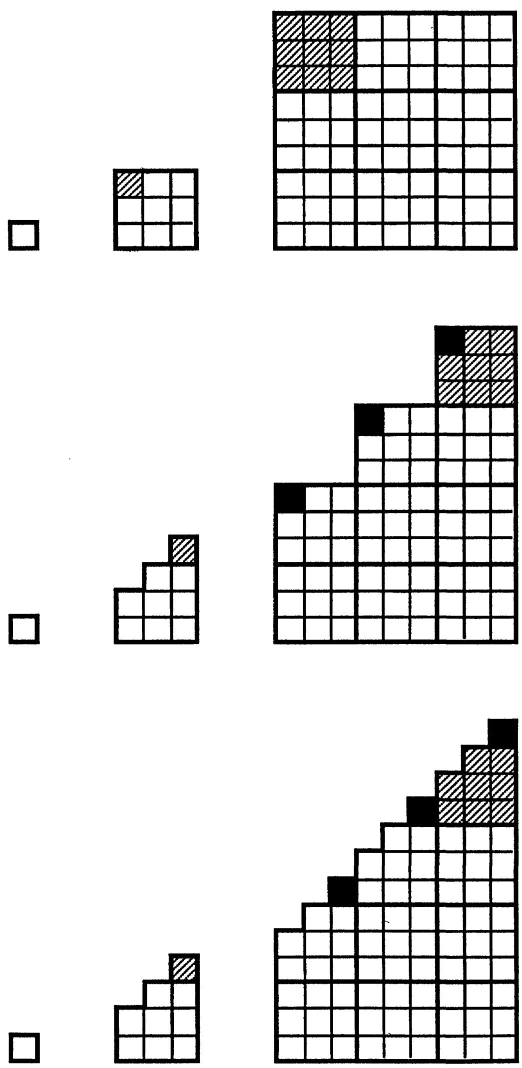
$$\begin{aligned}h(a,b)&=h(a'+(a-a'),b)\neq h(a',b)+h(a-a',0)=h(a',b)+r_m\circ\alpha_x(a-a')\\&=h(a',b)+1\neq h(a',b)-h(a'-a,0)\\&=h(a',b)-r_m\circ\alpha_x(a'-a)=h(a',b)-1,\end{aligned}$$

whence $h(a, b) = h(a', b)$; the proof is similar on the second coordinate.

By part c, $\bar{r} \circ i = h'$, since both maps restrict to r on each coordinate; hence $h = h' \circ \alpha_{(x,y)} = \bar{r} \circ i \circ \alpha_{(x,y)}$, and h is unique.

Please let me know if you find a faster proof!

Proof without Words: Sums of Consecutive Powers of Nine Are Sums of Consecutive Integers



$$1 + 9 + \cdots + 9^n = 1 + 2 + 3 + \cdots + (1 + 3 + \cdots + 3^n)$$

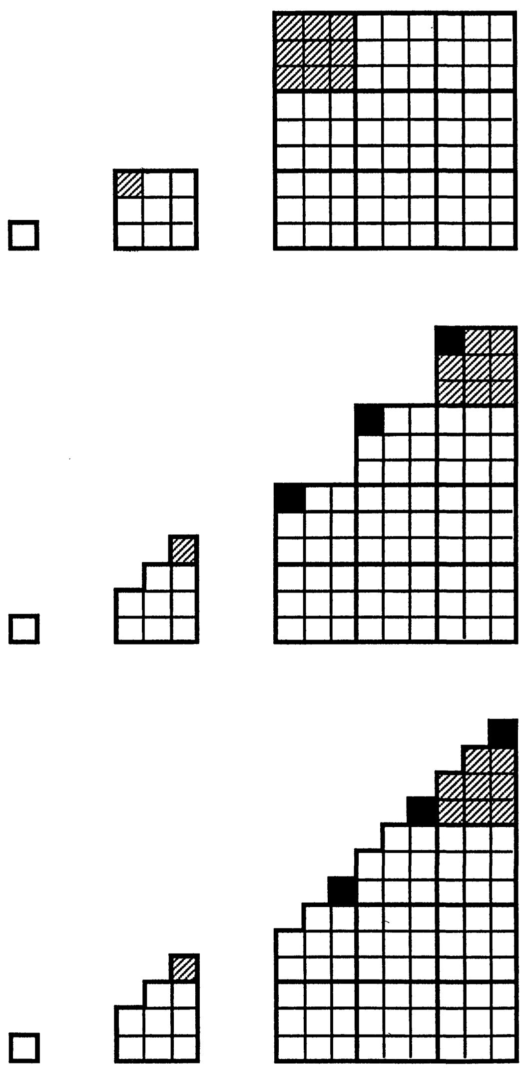
—Roger B. Nelsen
Lewis and Clark College

whence $h(a,b) = h(a',b)$; the proof is similar on the second coordinate.

By part c, $\bar{r} \circ i = h'$, since both maps restrict to r on each coordinate; hence $h = h' \circ \alpha_{(x,y)} = \bar{r} \circ i \circ \alpha_{(x,y)}$, and h is unique.

Please let me know if you find a faster proof!

Proof without Words: Sums of Consecutive Powers of Nine Are Sums of Consecutive Integers



$$1 + 9 + \cdots + 9^n = 1 + 2 + 3 + \cdots + (1 + 3 + \cdots + 3^n)$$

—ROGER B. NELSEN
LEWIS AND CLARK COLLEGE

NOTES

How to Approach a Traffic Light

J. I. KATZ
Department of Physics
Washington University
St. Louis, MO 63130

Introduction Anyone who has ridden a bicycle or driven an automobile has rounded a curve or crested a rise in the road to see a red traffic light some distance away, far enough that immediate braking is not required. The lazy bicyclist, or thrifty motorist, may then have asked himself whether he should maintain a constant speed in the hope that the light would turn green before his arrival, or at least until so close that prudence would require braking; alternatively, he may have considered immediately slowing his vehicle in order to increase the chance that he would arrive at a green light and not have to stop. His goal would have been the avoidance of needless dissipation of kinetic energy and thrift in the expenditure of muscular effort or fuel. A green traffic light presents a similar problem, for it may turn red before arrival, and one must decide if the expense of acceleration (supposing it to be legal and prudent) is justified by the increased chance of reaching the light before it turns red.

This problem at first sight appears simple. Closer examination reveals several parameter regimes and various complications; the analysis of the most important case, although nearly trivial, is illuminating. The answer, obvious in retrospect, was not apparent when I first posed the problem.

A quantitatively realistic statement of the problem would depend on many variables, several of them poorly known. I therefore make several simplifying assumptions: The road is level and the vehicle rolls without frictional, aerodynamic, or other losses. It is capable of arbitrary and instantaneous acceleration, positive or negative. The energy efficiency of its propulsion is constant, but kinetic energy lost in braking is irretrievable. No energy is consumed in engine idling or its human equivalent, which is equivalent to assuming that the goal is to optimize energy expenditure without regard to time. Traffic lights regularly repeat with a period T, shining red for a time t_r and green for a time t_g, with $T = t_r + t_g$. These parameters are assumed known, but when a traffic light is first sighted its position within this cycle is unknown and random except for the observation of the displayed color. The rider or driver is assumed to know his speed and distance from the light. This model is necessarily arbitrary, but is readily analyzed, and may provide a qualitative guide to more complex and realistic models.

An extensive literature exists on traffic flow problems (Montroll and Badger [2], Haberman [1]), but the single vehicle energy optimization problem does not appear to have been considered.

Calculation First consider the case in which the traffic light is observed to be red when it comes into view. The problem is best analyzed with the aid of a time line diagram, as in FIGURE 1. The vehicle is traveling at a speed v and comes into view of the light a distance d away. It is necessary to determine the probability P_g that the light will be a green a time t later, where $t = d/v$ is the time of arrival for unaccelerated motion. Note that t is a parameter of the problem, and not a random variable.

First replace t by $t' = t - nT \equiv t \bmod T$ (n is the largest integer for which $t' \geqslant 0$); this is permitted because of the assumed periodicity of the traffic light, and would not apply to demand-activated lights. It is then possible to consider only the short, finite time line shown in FIGURE 1. The origin is defined as the moment when the light turns from green to red.

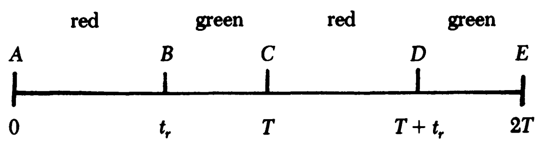


FIGURE 1
Time line diagram for the observation of a red traffic light. The variable is $t' + \tau$. Each of t' and τ lie between 0 and T, so the time line need only be considered between 0 and $2T$.

The light is sighted at an unknown time $0 < \tau < t_r$ uniformly distributed in the interval AB. The light will be reached at a time on the time line of $t' + \tau$. We must calculate the fraction of possible values of τ which yield $t' + \tau$ in BC (DE is not accessible for $t' < T$ because of the bounds on τ). Several cases must be distinguished:

a) $t' < t_r$ and $t' < t_g$: then $t' + \tau$ will be in BC for $t_r - t' < \tau < t_r$. This has probability $[t_r - (t_r - t')]/t_r = t'/t_r$.

b) $t' < t_r$ and $t' > t_g$: then $t' + \tau$ will be in BC for $t_r - t' < \tau < T - t'$. This has probability $[(T - t') - (t_r - t')]/t_r = t_g/t_r$.

c) $t' > t_r$ and $t' < t_g$: then $t' + \tau$ will be in BC for all allowable $0 < \tau < t_r$, which has probability 1.

d) $t' > t_r$ and $t' > t_g$: then $t' + \tau$ will be in BC for $0 < \tau < T - t'$. This has probability $(T - t')/t_r$.

Note that for a traffic light with $t_g > t_r$ case b) cannot occur, while if $t_g < t_r$ case c) cannot occur.

These four cases may be summarized in the probability table, TABLE 1. The cases in which t' equals t_r or t_g are of zero measure and may be neglected. Note that for the special case $t_g = T/2$ it is only possible to be in the upper left or lower right quadrants of the tables, the former applying if $t' < T/2$ and the latter if $t' > T/2$. Similar results for a light originally observed to be green are shown in TABLE 2.

TABLE 1
Probabilities P_g of reaching a green light if originally sighted red. The first row contains cases a) and c), and the second row contains b) and d)

	$t' < t_r$	$t' > t_r$
$t' < t_g$	$\frac{t'}{t_r}$	1
$t' > t_g$	$\frac{t_g}{t_r}$	$\frac{T - t'}{t_r}$

TABLE 2
Probabilities P_g of reaching a green light if originally sighted green.

	$t' < t_r$	$t' > t_r$
$t' < t_g$	$\frac{t_g - t'}{t_g}$	$\frac{t_g - t_r}{t_g}$
$t' > t_g$	0	$\frac{t' - t_r}{t_g}$

The parameter regimes are shown as functions of t_g and t' in FIGURE 2. This figure is a convenient graphical guide to the tables: the point corresponding to any desired values of t' and t_g lies in a triangular region labelled with the corresponding entries (UL = upper left, etc.) in the tables. FIGURE 3 shows the forms possible for $P_g(t)$ for a light observed red; the corresponding figure for a light observed green is similar and may easily be drawn from the entries in TABLE 2.

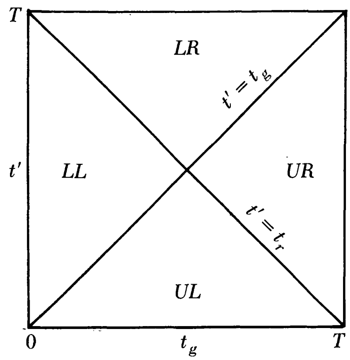


FIGURE 2
Diagram of parameter regimes as function of t_g, t'. Regions are labelled with entry (UL = upper left, etc.) in Tables 1, 2 corresponding to those values of t_g and t'.

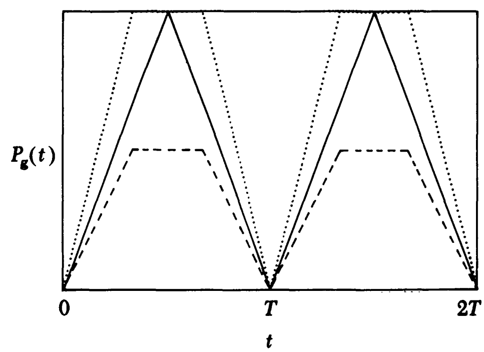


FIGURE 3
$P_g(t)$ for traffic light observed red. Solid curve is for $t_g = T/2$, dashed for $t_g = T/3$, dotted for $t_g = 2T/3$. The curves repeat indefinitely with period T.

Results Consider a traffic light first sighted red. We define the value KE of a strategy in which the vehicle instantaneously decelerates from its initial speed ν_o to a speed of ν, and then rolls to the light unaccelerated, as the mean kinetic energy surviving passage through the traffic light. For a vehicle of mass m this is

$$KE = \frac{1}{2} m \nu^2 P_g(t). \tag{1}$$

We wish to maximize KE. For values of t' in the upper left quadrant of TABLE 1, P_g is a decreasing function of ν (because t' is a decreasing function of ν), so only in this quadrant could small decelerations possibly increase KE. Substituting the expression for P_g (and using the definitions of t' and t) yields

$$KE = \frac{1}{2} \frac{m}{t_r} (\nu d - nT\nu^2). \tag{2}$$

In any of the other three quadrants P_g is a nondecreasing function of ν and small decelerations are always disadvantageous, for any n.

For $n = 0$ (the case encountered most often in practice) even in the upper left quadrant KE is an increasing function of ν and the conclusion is clear: do not slow down upon first sighting a red light, for the lost kinetic energy more than outweighs the increased chance of rolling through a green one upon reaching it.

For $n \geqslant 1$, corresponding to distant traffic lights (or slow vehicles), the results are different. In the upper left quadrant KE is easily seen to be a decreasing function of ν (and an increasing function of t), and it is advantageous to brake the vehicle immediately until t' increases to the lesser of t_g and t_r. This is not a large deceleration; it never exceeds its value for $t_g = T/2$ and $t' = 0$, in which case the fractional change in speed is $1/(2n+1)$. This deceleration places the parameters on the border of the other three quadrants, so further deceleration is disadvantageous (for any n).

If t' is initially in the lower right quadrant of the table (corresponding to negative slopes in the curves of FIGURE 3) then

$$KE = \frac{1}{2}\frac{m}{t_r}\left[(n+1)\nu^2 T - \nu d\right]. \tag{3}$$

It is apparent that $dKE/d\nu > 0$ and infinitesimal decelerations will not be advantageous. However, finite decelerations which increase n by 1 and place t in the upper left quadrant may be advantageous. This unexpected result comes about because of the minimum of $P_g(t)$ at $t = T$. A decrease in ν (and increase in t) may bring P_g to its maximum and, on balance, increase KE. Some algebra is required to decide whether this is so for any particular values of the parameters.

If t' is initially in the upper right or lower left quadrants of the table, then P_g has its maximum value, and the optimal course is no acceleration.

It is useful to check the theory against trivial cases such as the permanently green light, for which $t_g = T$, $t_r = 0$ and $P_g = 1$. Then (1) gives $KE = m\nu^2/2$, which is maximized by maintaining the initial speed, as must trivially be the case (because one will never be required to stop, deceleration would exact a completely unnecessary cost in energy).

The problem of a light initially sighted green is closely analogous to that just discussed; the probabilities are shown in TABLE 2. The most frequent case is that of the upper left quadrant with $n = 0$, in which case acceleration is colloquially known as racing to beat a light. Because P_g is here a decreasing function of t it is evident that deceleration is disadvantageous. In order to decide whether acceleration is advantageous we must consider the mean cost C in kinetic energy imposed on a vehicle coasting at speed ν by the risk of arrival at a red light

$$C = \frac{1}{2}m\nu^2\left[1 - P_g(t)\right]. \tag{4}$$

It is more convenient to deal with the cost of arrival at a red light than with the kinetic energy surviving passage through a green light because the former more conveniently allows for any extra kinetic energy acquired in accelerating from the initial speed ν_o to ν. For the usual case $n = 0$, and the expression for P_g in the upper left quadrant of TABLE 2 leads to

$$C = \frac{1}{2}\frac{m\nu d}{t_g}. \tag{5}$$

It is clear that increasing ν in order to increase the chance of reaching the light when it is still green is disadvantageous. It would be even more disadvantageous if the added kinetic energy had to be dissipated after passage through a green light, rather than being usefully applied, as tacitly assumed here. The problems of $n \geqslant 1$ and of other quadrants of TABLE 2 are slightly more complex, but may be treated just as were the analogous problems for the case of a traffic light initially sighted red.

For a light initially seen red, acceleration may be advantageous if t' and t_g correspond to the lower right quadrant of TABLE 1. This apparently paradoxical result (accelerating towards a red light) is the consequence of the increase of P_g with increasing ν (decreasing t) in this parameter regime. Defining the mean cost of acceleration as in (4) leads to the conclusion that within this quadrant acceleration is advantageous for $n = 0$ if $t < 2t_g$ and is always advantageous if $n \geqslant 1$.

An algebraically more complex problem is obtained if we allow for the accumulation of information after the traffic light is sighted. Once the light has been observed for a finite time as red or green the observer's position within its cycle is localized, and the probability of arrival at a green light differs from that upon the original observation. Use of this additional information may lead to somewhat decreased values of C, or increased values of KE, but does not change the principles of the original problem. Once a change of color is observed the problem becomes deterministic and simple algebra determines the optimal speed.

Conclusion The problem of determining the most energy-efficient strategy to use in approaching a traffic light which is sighted at an unknown phase in its cycle is remarkably complex, even given several simplifying physical assumptions, and consists of a maze of special cases. Fortunately, the case most often encountered in practice is the simplest. Under typical suburban conditions an informal survey found these values to be representative: $t_g = 40$ sec, $T = 100$ sec, $v = 30$ miles/hour and $d = 1/4$ mile ($t_r > t_g$ because there are periods in which left-turning traffic has the right of way, and through traffic in both directions halts). Then $t = 30$ sec, $n = 0$, and the upper left quadrants of the tables apply. Under urban conditions d and t are generally even smaller, while in rural areas larger d are usually accompanied by larger v (and traffic lights are infrequent in any case). Therefore, in most cases the applicable result is simple: do not slow down in anticipation of a sighted red light, or race to meet a green one. Instead, roll until the light (or a safe stopping distance from a red light) is reached, or until its color is observed to change.

In the real world the problem is complicated by friction and the presence of other vehicles. In most cases coasting should probably be interpreted as maintaining a constant speed; this is not the most energy-efficient strategy, which would be frictional slowing, but avoids the inconvenience of rolling to a halt in the middle of a block. The chief exception is the red light sighted from a distance less than the frictional stopping length, in which case coasting under the influence of friction is best, as maintaining a constant speed both costs energy and decreases P_g.

With our assumptions, strictly interpreted, the most energy-efficient trip would proceed at infinitesimal velocity and take infinitely long. In the real world we reject this and choose to pay the price of a cruising speed v; we maximize a "utility function" (in economists' language) which is a decreasing function of both time and fuel expended. Only when immediately faced with a traffic light and the imminent prospect of braking to a halt is the value of our kinetic energy worthy of conservation, and our model of minimizing energy expenditure without regard to time is useful.

There is an analogous problem, in which the goal is to minimize the trip duration without regard to energy expended. This has a completely trivial solution for the motorist: always travel at the maximum allowable speed. A bicyclist has complex physiological constraints on the rate at which he does work, so that wasted energy diminishes his subsequent acceleration or speed. If these constraints are quantified an optimum strategy may be found numerically. It is likely that experienced bicyclists have a good intuitive understanding of such a strategy, obtained from experience.

REFERENCES

1. R. Haberman, *Mathematical Models: Mechanical Vibrations, Population, Dynamics and Traffic Flow, An Introduction to Applied Mathematics*, Prentice-Hall, Englewood Cliffs, N.J., 1977.
2. E. W. Montroll and W. W. Badger, *Quantitative Aspects of Social Phenomena*, Gordon and Breach, London, 1974.

Mathematical Motivation through Matrimony

MARTIN D. STERN
Manchester Polytechnic
Manchester, England

Introduction At the end of the Book of Common Prayer the following table of prohibited degrees of marriage is given:

A TABLE OF KINDRED AND AFFINITY

WHEREIN WHOSOEVER ARE RELATED ARE FORBIDDEN BY THE CHURCH OF ENGLAND TO MARRY TOGETHER

A man may not marry his:	*A woman may not marry her:*
1 Mother	1 Father
2 Daughter	2 Son
3 Father's mother	3 Father's father
4 Mother's mother	4 Mother's father
5 Son's daughter	5 Son's son
6 Daughter's daughter	6 Daughter's son
7 Sister	7 Brother
8 Father's daughter	8 Father's son
9 Mother's daughter	9 Mother's son
10 Wife's mother	10 Husband's father
11 Wife's daughter	11 Husband's son
12 Father's wife	12 Mother's husband
13 Son's wife	13 Daughter's husband
14 Father's father's wife	14 Father's mother's husband
15 Mother's father's wife	15 Mother's mother's husband
16 Wife's father's mother	16 Husband's father's father
17 Wife's mother's mother	17 Husband's mother's father
18 Wife's son's daughter	18 Husband's son's son
19 Wife's daughter's daughter	19 Husband's daughter's son
20 Son's son's wife	20 Son's daughter's husband
21 Daughter's son's wife	21 Daughter's daughter's husband
22 Father's sister	22 Father's brother
23 Mother's sister	23 Mother's brother
24 Brother's daughter	24 Brother's son
25 Sister's daughter	25 Sister's son

Though the prohibited marriages are arranged in parallel columns for men and women, it is not immediately obvious that there are no other implicit prohibitions. For example, according to line 15 in the first column, a man may not marry his mother's father's wife (obviously after her husband's demise!), and the question that might cross the reader's mind is whether this woman is explicitly prohibited from marrying him. In order to determine whether this is so, the relationship must be inverted and the purpose of this article is to describe a mathematical notation based on binary numbers that facilitates this inversion process.

The notation was originally developed for analyzing problems in Jewish marriage law [1], where the prohibited degrees are only stated from a man's standpoint in the standard codes. The problem considered was to list the corresponding prohibitions for a woman. In the latter system there was not a symmetrical relationship between the sexes. For example a man was allowed to marry his niece but not his aunt so, consequently, a woman was prohibited from marrying her nephew but not her uncle.

In the problem here considered we shall show that there are no further prohibited marriages other than those explicitly stated and hence that there is an element of symmetry between the sexes regarding the restriction in the choice of a marriage partner.

Notation We shall use the single digits 1 and 0 to denote male and female, respectively, and define the two digit codes for relationships between individuals:

00	spouse
01	parent
10	child
11	sibling

The choice of a particular code is, of course, arbitrary except that a symmetric relationship such as spouse or sibling must consist of two identical digits whereas parent or child may not. The reason for this will become evident when the problem of inverting relationships is discussed below.

To establish the relationship between one person and another, we write a digit for that person followed by a sequence of three digits, the first two representing the relationship and the third the sex of the second person, e.g., a man's wife would be expressed as

$$1\ 000$$

Some sequences are inadmissible since, for example, only heterosexual marriages are officially sanctioned by the Church of England, i.e., 1 001 would be inadmissible.

More complicated relationships can be expressed by appending further triplets. So, for example, a man's wife's father's brother can be written as

$$1\ 000\ 011\ 111.$$

Conversely any sequence of $3n+1$ binary digits can be interpreted as a relationship between two people. For example 1011000101 can be analyzed as

1	011	000	101
a man's	father's	wife's	son

i.e. a step brother. We shall assume that relationships will always be expressed in the most economical manner so that this is distinct from a brother 1 111 or a (paternal) half brother, i.e., a man's father's son which would be 1 011 101.

Inversion of relationships Using this notation it is easy to find the inverse of a given relationship by merely reversing the order of the digits and regrouping them as a single digit followed by a sequence of triplets. For example, a man's mother's father's sister's husband's mother would be expressed as

$$1\ 010\ 011\ 110\ 001\ 010.$$

When this procedure is followed we obtain

0 101 000 111 100 101,

which can be readily interpreted as a woman's son's wife's brother's daughter's son.

When this notation is applied to the table of kindred and affinity we obtain the results in the table below. In it we have given the prohibitions as listed for a man followed by its code, the inverse code and the corresponding prohibition for a woman. Since every prohibited relationship for a woman appears in this final column, we see that there are no further prohibitions implicit in the table. This contrasts with other relationship problems to which this notation has been applied, such as the case of the prohibition in rabbinic law of consanguinous relatives from acting as witnesses [**2**].

	Male	*Code*	*Inverse*		*Female*
	A man may not marry his:				A woman may not marry her:
1	Mother	1 010	0 101	2	Son
2	Daughter	1 100	0 011	1	Father
3	Father's mother	1 011 010	0 101 101	5	Son's son
4	Mother's mother	1 010 010	0 100 101	6	Daughter's son
5	Son's daughter	1 101 100	0 011 011	3	Father's father
6	Daughter's daughter	1 100 100	0 010 011	4	Mother's father
7	Sister	1 110	0 111	7	Brother
8	Father's daughter	1 011 100	0 011 101	8	Father's son
9	Mother's daughter	1 010 100	0 010 101	9	Mother's son
10	Wife's mother	1 000 010	0 100 001	13	Daughter's husband
11	Wife's daughter	1 000 100	0 010 001	12	Mother's husband
12	Father's wife	1 011 000	0 001 101	11	Husband's son
13	Son's wife	1 101 000	0 001 011	10	Husband's father
14	Father's father's wife	1 011 011 000	0 001 101 101	18	Husband's son's son
15	Mother's father's wife	1 010 011 000	0 001 100 101	19	Husband's daughter's son
16	Wife's father's mother	1 000 011 010	0 101 100 001	20	Son's daughter's husband
17	Wife's mother's mother	1 000 010 010	0 100 100 001	21	Daughter's daughter's husband
18	Wife's son's daughter	1 000 101 100	0 011 010 001	14	Father's mother's husband
19	Wife's daughter's daughter	1 000 100 100	0 010 010 001	15	Mother's mother's husband
20	Son's son's wife	1 101 101 000	0 001 011 011	16	Husband's father's father
21	Daughter's son's wife	1 100 101 000	0 001 010 011	17	Husband's mother's father
22	Father's sister	1 011 110	0 111 101	24	Brother's son
23	Mother's sister	1 010 110	0 110 101	25	Sister's son
24	Brother's daughter	1 111 100	0 011 111	22	Father's brother
25	Sister's daughter	1 110 100	0 010 111	23	Mother's brother

Conclusion While the notation presented here is not claimed to be particularly profound, it might provide a useful tool for motivating pupils to take a greater interest in mathematics by emphasizing the non-numerical aspects of the subject. In England, where we have in our inner cities a considerable mix of people of different religious and ethnic backgrounds, this notation has been used as a basis for projects in which the students have analysed one another's traditions and, hopefully, come to a closer understanding both of their fellows and the value of mathematics.

REFERENCES

1. J. Cashdan and M. D. Stern, Forbidden marriages from a woman's angle, *Math. Gaz.* 71, No. 456 (1987).
2. M. D. Stern, Consanguinity of witnesses—a mathematical analysis, *Teaching Mathematics and Its Applications*, Vol. 6.2 (1987).

A Historical Gem from Vito Volterra

WILLIAM DUNHAM
Hanover College and
The Ohio State University
Columbus, OH 43210

Sometime during the junior or senior year, most undergraduate mathematics students first examine the theoretical foundations of the calculus. Such an experience—whether it's called "analysis," "advanced calculus," or whatever—introduces the precise definitions of continuity, differentiability, and integrability and establishes the logical relationships among these ideas.

A major goal of this first analysis course is surely to consider some standard examples—perhaps of a "pathological" nature—that reveal the superiority of careful analytic reasoning over mere intuition. One such example is the function defined on (0, 1] by

$$g(x)=\begin{cases}1/q & \text{if } x=p/q \text{ in lowest terms}\\ 0 & \text{if } x \text{ is irrational}\end{cases}$$

This function, of course, is continuous at each irrational point in the unit interval and discontinuous at each rational point (see, for instance [2, p. 76]). It thus qualifies as extremely pathological, at least to a novice in higher mathematics. By extending this function periodically to the entire real line, we get a function—call it "G"—continuous at each irrational number in $\mathbb{R}$ and discontinuous at each rational in $\mathbb{R}$.

The perceptive student, upon seeing this example, will ask for a function continuous at each *rational* point and discontinuous at each *irrational* one, and the instructor will have to respond that such a function cannot exist.

"Why not?" asks our perceptive, and rather dubious, student.

In reply, the highly trained mathematician may recall his or her graduate work and begin a digression to the Baire Category Theorem, with attendant discussions of nowhere dense sets, first and second category, F_σ's and G_δ's, before finally demonstrating the non-existence of such a function (for example, see [6, p. 141]). Of course, it takes quite a while to set up all this sophisticated mathematical machinery, during which time the student probably will have lost interest or graduated.

It may come as a surprise, then, that this question was answered in a short and simple 1881 proof by the brilliant Vito Volterra (1860–1940). Discovered by Volterra when he was still a student at Pisa's Scuola Normale Superiore, his result predates René Baire's groundbreaking category theorem [1] by almost two decades yet uses only the relatively unsophisticated notions of "continuous function" and "dense set." As such, it provides a fine example of the advantages of studying the history of mathematics; for not only does it give a glimpse into the past but simultaneously satisfies the classroom needs of the present.

The argument appeared early in Volterra's paper "Alcune osservasioni sulle funzioni punteggiate discontinue" [8, pp. 7–8]. Central to this work was the concept of "pointwise discontinuous" functions, i.e., functions whose points of continuity form a dense set. The function $G(x)$ above is pointwise discontinuous, and other such functions had appeared on the scene by the mid-nineteenth century. Bernhard Riemann [5, p. 242], for instance, startled his 1854 audience with an example of an integrable function having discontinuities precisely at rationals of the form $m/2n$

where m and $2n$ are relatively prime. To some, this suggested an intimate link between pointwise discontinuity and the highly complicated matter of integrability.

One such mathematician was Hermann Hankel, who in 1870 made a detailed examination of the "punktirt unstetigen," i.e., "pointwise discontinuous," functions. (This rather odd-sounding term was introduced as a contrast to the "total unstetige," i.e., "totally discontinuous," functions, whose points of continuity were not dense.) Hankel gave a proof [3, pp. 89–90] purporting to show that a function is (Riemann) integrable if and only if it is pointwise discontinuous. This conclusion certainly thrust pointwise discontinuous functions into the limelight, and, as Thomas Hawkins observes in his excellent book *Lebesgue's Theory of Integration: Its Origins and Development* [4, p. 30], seemed to identify them as precisely "the functions amenable to mathematical analysis."

However, Hankel's reasoning was flawed, and the error was exposed in 1875 by Oxford professor H. J. S. Smith [7, p. 150]. While Smith agreed that any integrable function must be pointwise discontinuous, his explicit example of a pointwise discontinuous but non-integrable function destroyed the "if and only if" nature of Hankel's proof. Smith had established, of course, that the integrable functions form a *proper* subset of the pointwise discontinuous ones, but his paper was not widely read and its impact was minimal. Consequently, pointwise discontinuous functions occupied an important, although not necessarily well-understood, position in the research of the day. (Hawkins gives a detailed account of the situation in [4, Ch. 2].)

It was in this context that the young Volterra composed his 1881 paper. There, he stated the following key theorem:

THEOREM. *There do not exist pointwise discontinuous functions defined on an interval (a,b) for which the continuity points of one are the discontinuity points of the other, and vice versa.*

Beginning a proof by contradiction, Volterra assumed the existence of two such functions, f and ϕ. For notational ease, we shall let

$$C_f = \{x \in (a,b) | f \text{ is continuous at } x\}.$$

Thus, Volterra's assumption was that the dense sets C_f and C_ϕ partition (a,b) into disjoint subsets.

Let x_0 be any point in C_f and take $\alpha = 1$. Continuity guarantees the existence of a $\delta > 0$ such that $(x_0 - \delta, x_0 + \delta) \subset (a,b)$ and $|f(x) - f(x_0)| < 1/2$ for all $x \in (x_0 - \delta, x_0 + \delta)$. We then can choose $a_1 < b_1$ so that $[a_1, b_1]$ is a *closed* subinterval of $(x_0 - \delta, x_0 + \delta)$ and consequently, for any x and y in $[a_1, b_1]$,

$$\begin{aligned} |f(x) - f(y)| &\leqslant |f(x) - f(x_0)| + |f(x_0) - f(y)| \\ &< 1/2 + 1/2 = 1 = \alpha. \end{aligned}$$

Pointwise discontinuity now yields a continuity point of ϕ in the open interval (a_1, b_1), and by the preceding argument there exist $a_1' < b_1'$ with $[a_1', b_1'] \subset (a_1, b_1)$ and with

$$|\phi(x) - \phi(y)| < 1 \text{ for all } x \text{ and } y \text{ in } [a_1', b_1'].$$

To summarize, then, for all x and y in $[a_1', b_1'] \subset (a,b)$,

$$|f(x) - f(y)| < 1 \text{ and } |\phi(x) - \phi(y)| < 1$$

as well.

Volterra next repeated this argument, starting with the open interval (a_1', b_1') and the margin $\alpha = 1/2$, then $\alpha = 1/4$, and generally $\alpha = 1/2^n$. This generates a strictly descending sequence of closed intervals

$$(a,b) \supset [a_1', b_1'] \supset \cdots \supset [a_n', b_n'] \supset \cdots$$

such that, for all x and y in $[a_n', b_n']$, we have both

$$|f(x) - f(y)| < 1/2^n \text{ and } |\phi(x) - \phi(y)| < 1/2^n.$$

(Note how the pointwise discontinuity plays a central role at each step, providing at least one continuity point for f or ϕ in any open subinterval.)

But by the Nested Interval Theorem there exists at least one point A contained in *all* of the closed subintervals above, and thus both f and ϕ are continuous at A. In short, $C_f \cap C_\phi \neq \varnothing$, a contradiction. From this, Volterra concluded that no such functions f and ϕ exist. Q.E.D.

We should observe that Volterra was a bit vague about the intervals $[a_n', b_n']$ being CLOSED, as indeed they must be to guarantee a point in their intersection. Volterra was not alone among nineteenth century mathematicians in this vagueness. Hankel, in the paper cited above, failed to stress this pivotal detail, as did Baire in his proof of the "Category Theorem" that carries his name [1, p. 65]. Fortunately, Volterra's argument is easily repaired, as has been done above.

From this simple proof, Volterra then drew two interesting conclusions. The first, answering the question of our perceptive student, was that no function can have the set of rationals as its only points of continuity, for such a function would be pointwise discontinuous with continuity points corresponding to the discontinuity points of the "extended" pathological function G above, a situation whose impossibility he had just demonstrated.

Second, Volterra reasoned that there can be no continuous function ϕ mapping rationals to irrationals and vice versa. For, if such a ϕ existed, we could introduce the composite $h \equiv G \circ \phi$, where again G is as above. Clearly, if x is rational, $\phi(x)$ is irrational, so G is continuous at $\phi(x)$ and thus h is continuous at x.

On the other hand, h will be discontinuous at any irrational y. To see this, choose $\{x_n\}$, a sequence of *rationals* converging to y. Then,

$$\lim_n h(x_n) = \lim_n G(\phi(x_n)) = \lim_n 0 = 0,$$

since $\phi(x_n)$ is irrational for each n and G is zero on the irrationals. But $h(y) = G(\phi(y)) \neq 0$ because $\phi(y)$ is a rational number.

In short, the function h defined above has C_h equal to the set of rationals, an impossibility by Volterra's earlier observation. Thus, it is impossible continuously to transform rationals to irrationals and vice versa.

With these reasonably elementary arguments—a tribute to the genius of Vito Volterra—we can use some of yesterday's mathematics to answer today's questions. The inquisitive student should be both satisfied and, one hopes, impressed. And in this instance the history of mathematics comes alive by proving its value in the contemporary classroom.

REFERENCES

1. R. Baire, Sur les fonctions de variables réelles, *Annali di Matematica pura ed applicata* 3 (1899), 1–122.
2. C. Clark, *The Theoretical Side of Calculus*, Krieger, 1978.

3. H. Hankel, Untersuchungen über die unendlich oft oszillierenden und unstetigen Functionen, presented in March, 1870 at the University of Tübingen; reprinted in *Mathematische Annalen* 20 (1882), 63–112.
4. T. Hawkins, *Lebesgue's Theory of Integration: Its Origins and Development*, Chelsea, 1975.
5. B. Riemann, *Gesammelte Mathematische Werke und Wissenschaftlicher Nachlass*, 2nd ed., Leipzig, B. G. Teubner, 1902.
6. H. Royden, *Real Analysis*, 2nd ed., Macmillan, 1968.
7. H. J. S. Smith, On the integration of discontinuous functions, *Proc. London Math Soc.* 6 (1875), 140–153.
8. V. Volterra, Alcune osservasioni sulle funzioni punteggiate discontinue, in *Opere Matematiche* 1 (Rome 1954), 7–15 (originally in *Giornale di Matematiche* 19 (1881), 76–86).

Einstein's Principle

B. L. FOSTER
University of Montana
Missoula, MT 59812

In his remarkable Spencer Lecture, delivered at Oxford in 1933, Albert Einstein advanced an astonishing epistemological view:

> If then it is the case that the axiomatic basis of theoretical physics cannot be an inference from experience, but must be free invention, have we any right to hope that we shall find the correct way?... To this I answer with complete assurance, that in my opinion there is *the* correct path and, moreover, that it is in our power to find it. Our experience up to date justifies us in feeling sure that in Nature is actualized the ideal of mathematical simplicity. It is my conviction that pure mathematical construction enables us to discover the concepts and the laws connecting them which give us the key to the understanding of the phenomena of Nature... In a certain sense, therefore, I hold it to be true that pure thought is competent to comprehend the real, as the ancients dreamed.
>
> To justify this confidence of mine, I must necessarily avail myself of mathematical concepts. The physical world is represented as a four-dimensional continuum. If in this I adopt a Riemannian metric, and look for the simplest laws which such a metric can satisfy, I arrive at the relativistic gravitation-theory of empty space. If I adopt in this space a vector-field,..., and if I look for the simplest laws which such a field can satisfy, I arrive at the Maxwell equations for free space. [1, p. 167]

The elevation of mathematical simplicity from helpful adjunct to indispensable guiding principle for the discovery of natural law seems radical even today. That so many of his last years were spent in vain on a quest for the simple field which would comprise, as special cases, gravity's Riemannian metric and electromagnetism's vector-field, may have unduly obscured the basic soundness of the principle, which he believed to be founded on the rock of general relativity.

This note gives an elementary illustration of the principle applied to pre-relativity physics. If a vector-field is thought of as first rank, and a metric as second rank, then the next step down is a scalar-field, of rank zero. For it the following, itself based on a hint of Einstein's (in [2, p. 29]), holds:

3. H. Hankel, Untersuchungen über die unendlich oft oszillierenden und unstetigen Functionen, presented in March, 1870 at the University of Tübingen; reprinted in *Mathematische Annalen* 20 (1882), 63–112.
4. T. Hawkins, *Lebesgue's Theory of Integration: Its Origins and Development*, Chelsea, 1975.
5. B. Riemann, *Gesammelte Mathematische Werke und Wissenschaftlicher Nachlass*, 2nd ed., Leipzig, B. G. Teubner, 1902.
6. H. Royden, *Real Analysis*, 2nd ed., Macmillan, 1968.
7. H. J. S. Smith, On the integration of discontinuous functions, *Proc. London Math Soc.* 6 (1875), 140–153.
8. V. Volterra, Alcune osservasioni sulle funzioni punteggiate discontinue, in *Opere Matematiche* 1 (Rome 1954), 7–15 (originally in *Giornale di Matematiche* 19 (1881), 76–86).

Einstein's Principle

B. L. FOSTER
University of Montana
Missoula, MT 59812

In his remarkable Spencer Lecture, delivered at Oxford in 1933, Albert Einstein advanced an astonishing epistemological view:

> If then it is the case that the axiomatic basis of theoretical physics cannot be an inference from experience, but must be free invention, have we any right to hope that we shall find the correct way?... To this I answer with complete assurance, that in my opinion there is *the* correct path and, moreover, that it is in our power to find it. Our experience up to date justifies us in feeling sure that in Nature is actualized the ideal of mathematical simplicity. It is my conviction that pure mathematical construction enables us to discover the concepts and the laws connecting them which give us the key to the understanding of the phenomena of Nature... In a certain sense, therefore, I hold it to be true that pure thought is competent to comprehend the real, as the ancients dreamed.
>
> To justify this confidence of mine, I must necessarily avail myself of mathematical concepts. The physical world is represented as a four-dimensional continuum. If in this I adopt a Riemannian metric, and look for the simplest laws which such a metric can satisfy, I arrive at the relativistic gravitation-theory of empty space. If I adopt in this space a vector-field,..., and if I look for the simplest laws which such a field can satisfy, I arrive at the Maxwell equations for free space. [**1**, p. 167]

The elevation of mathematical simplicity from helpful adjunct to indispensable guiding principle for the discovery of natural law seems radical even today. That so many of his last years were spent in vain on a quest for the simple field which would comprise, as special cases, gravity's Riemannian metric and electromagnetism's vector-field, may have unduly obscured the basic soundness of the principle, which he believed to be founded on the rock of general relativity.

This note gives an elementary illustration of the principle applied to pre-relativity physics. If a vector-field is thought of as first rank, and a metric as second rank, then the next step down is a scalar-field, of rank zero. For it the following, itself based on a hint of Einstein's (in [**2**, p. 29]), holds:

And if in the ordinary three-dimensional continuum, time aside, one adopts a scalar-field, and looks for the simplest laws which such a field can satisfy, namely, spherical symmetry and Laplace's equation, then one arrives at Newton's inverse square law.

The argument is simple indeed. Let the scalar-field be $\phi(x, y, z)$. If it is spherically symmetric, then $\phi = \phi(r)$, where $r = \sqrt{x^2 + y^2 + z^2}$. The analogue of relativity's vanishing curvature condition is Lapace's equation: $\phi_{xx} + \phi_{yy} + \phi_{zz} = 0$. Computing gives

$$\phi_{xx} = \frac{\phi'}{r} + \frac{x^2\phi''}{r^2} - \frac{x^2\phi'}{r^3},$$

and similarly for ϕ_{yy} and ϕ_{zz}.

Thus Laplace's equation becomes $\phi'' + 2\phi'/r = 0$, which, upon integration, yields $\phi = -1/r$. This potential has force-field,

$$-(\phi_x, \phi_y, \phi_z) = -\left(\frac{x}{r^3}, \frac{y}{r^3}, \frac{z}{r^3}\right),$$

whose magnitude is $1/r^2$.

Therefore, Newtonian physics also may be regarded as satisfying Einstein's principle, that in Nature is actualized the ideal of mathematical simplicity.

REFERENCES

1. Einstein, A., On the Method of Theoretical Physics, *Philosophy of Science* 1 (1934), 163–169.
2. Einstein, A., *Autobiographical Notes*, Open Court, 1979.

Another Hysterical Gram

There was a statistics prof
Who went too oft to the trough
He acquired a mode
At his navel node
Which was more than he could doff.

A standard snore with a standard score.

DAVE LOGOTHETTI
Santa Clara University
Santa Clara, CA 95053

And if in the ordinary three-dimensional continuum, time aside, one adopts a scalar-field, and looks for the simplest laws which such a field can satisfy, namely, spherical symmetry and Laplace's equation, then one arrives at Newton's inverse square law.

The argument is simple indeed. Let the scalar-field be $\phi(x, y, z)$. If it is spherically symmetric, then $\phi = \phi(r)$, where $r = \sqrt{x^2 + y^2 + z^2}$. The analogue of relativity's vanishing curvature condition is Lapace's equation: $\phi_{xx} + \phi_{yy} + \phi_{zz} = 0$. Computing gives

$$\phi_{xx} = \frac{\phi'}{r} + \frac{x^2\phi''}{r^2} - \frac{x^2\phi'}{r^3},$$

and similarly for ϕ_{yy} and ϕ_{zz}.

Thus Laplace's equation becomes $\phi'' + 2\phi'/r = 0$, which, upon integration, yields $\phi = -1/r$. This potential has force-field,

$$-(\phi_x, \phi_y, \phi_z) = -\left(\frac{x}{r^3}, \frac{y}{r^3}, \frac{z}{r^3}\right),$$

whose magnitude is $1/r^2$.

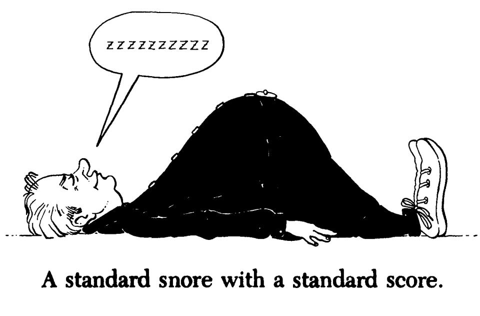
Therefore, Newtonian physics also may be regarded as satisfying Einstein's principle, that in Nature is actualized the ideal of mathematical simplicity.

REFERENCES

1. Einstein, A., On the Method of Theoretical Physics, *Philosophy of Science* 1 (1934), 163–169.
2. Einstein, A., *Autobiographical Notes*, Open Court, 1979.

Another Hysterical Gram

There was a statistics prof
Who went too oft to the trough
He acquired a mode
At his navel node
Which was more than he could doff.



A standard snore with a standard score.

DAVE LOGOTHETTI
Santa Clara University
Santa Clara, CA 95053

A Generating Function for the Distribution of the Scores of all Possible Bowling Games

CURTIS N. COOPER
ROBERT E. KENNEDY
Central Missouri State University
Warrensburg, MO 64093

1. Introduction Assuming that you know the terminology, rules of play, and the method of scoring, we will determine the number of ways that any particular score can occur in an ordinary game of ten-pin bowling. For example, it is clear that there is exactly one way that a perfect score of 300 can be obtained (all strikes) and exactly one way that a score of 0 can be obtained (all gutter balls). It may not be easy to see, however, that there are exactly 50613244155051856 ways of obtaining a score of 100.

Consider the following "line" of a bowling game:

5 4	9 /	X	X	X	7 1	X	6 /	4 2	3 / X
9	29	59	86	104	112	132	146	152	172

(1)

The lower number in each frame is the cumulative score. The upper numbers in each frame are the number of pins knocked down by the first and second ball respectively. The "/" indicates that the remaining pins were knocked down by the second ball and is called a "spare." The "X" indicates that all ten pins were knocked down by the first ball and is called a "strike."

Note that the game given in (1) can be represented by the sequence of nine ordered pairs and an ordered triple,

$$(5,4),(9,1),(10,0),(10,0),(10,0),(7,1),(10,0),(6,4),(4,2),(3,7,10) \quad (2)$$

With (2) as a model, we define a "bowling game" as a sequence of the form

$$(x_1,y_1),(x_2,y_2),\dots,(x_9,y_9),(x_{10},y_{10},z_{10}), \quad (3)$$

where the terms of (3) represent the ten frames of a game and each component of a term denotes the number of pins knocked down by that ball. Here, x_i, y_i and z_{10} are nonnegative integers where

$$x_i + y_i \leqslant 10 \qquad \text{for} \quad i = 1,\dots,9, \quad (4)$$

and with somewhat more involved conditions given in (5) on x_{10}, y_{10}, and z_{10}.

It was shown in [1] that the number of all possible bowling games is $(66^9)(241)$ which is approximately 5.7 billion billion and that the mean of all possible bowling games is approximately 80.

Thus, to determine the exact distribution of all possible bowling games by a computer generation of all possibilities would require in excess of 180 years even if every computer operation would take less than one-billionth of a second to perform. We will avoid this problem by constructing a generating function which determines the distribution of the scores of all possible bowling games. To do this, we will use the following sets where all components are nonnegative integers:

$$\begin{aligned}A &= \{(x,y): x+y \leqslant 9\}\\ B &= \{(x,y,0):(x,y) \in A\}\\ &\quad \cup \{(x,10-x,z): x \leqslant 9; z \leqslant 10\}\\ &\quad \cup \{(10,y,z): y \leqslant 9; y+z \leqslant 10\}\\ &\quad \cup \{(10,10,z): z \leqslant 10\}. \end{aligned} \tag{5}$$

Therefore, A is the set of frames in which a "mark" (spare or strike) is not made, and the set B is the set of all possibilities for the tenth frame. We wish to determine a polynomial function of the form

$$P(t) = \sum_{i=0}^{300} s_i t^i, \tag{6}$$

where s_i is the number of ways that a score of i is made. For example, $s_0 = 1$, $s_{300} = 1$, and $s_{100} = 50613244155051856$. Thus, $P(t)$ is a generating function for the set of all possible bowling scores.

2. States To aid in finding the function given in (6), we define four "states" in the process of calculating the score of a game. They are called the "OPEN state", the "SPARE state", the "STRIKE state", and the "DOUBLE state." These are defined as follows:

1. The OPEN state describes that the current frame is open. (7)
2. The SPARE state describes that a spare has been made in the current frame.
3. The STRIKE state describes that a strike has been made in the current frame and that either an open or spare was made in the previous frame.
4. The DOUBLE state describes that a strike was made in both the current and previous frames.

We observe here that all capital letters are used in reference to a state so as to emphasize that the state of a frame is not the same as what was rolled in that frame. For example, a frame in which a strike is rolled is not necessarily in the STRIKE state since the previous frame may not have been an open or a spare.

As we bowl a game and keep score, we pass from one state to another state. By convention, we will assume that each game starts with a 0th frame which is in the OPEN state with an accumulated score of 0. To clarify the above terminology, consider the bowling line given in (1). The 0th, 1st, 6th, and 9th frames are in the OPEN state, the 2nd and 8th frames are in the SPARE state, the 3rd and 7th frames are in the STRIKE state, while the 4th and 5th frames are in the DOUBLE state. Since the current state of a frame will determine the contribution of the next frame to the accumulated score, we do not need to define a state for the 10th frame.

3. Generating functions of transitions Here, we will determine the generating functions for the 16 possible transitions from one state to another state. First, it is clear that the transitions

OPEN	to	DOUBLE,	
SPARE	to	DOUBLE,	(8)
STRIKE	to	STRIKE,	

and

DOUBLE to STRIKE

cannot occur and hence may be considered as having a generating function of 0. To determine the generating function of a transition where (x, y) is the second state, we define the "value" of this transition as the contribution of the second state to the accumulated score. The following list gives the value of each of the other 12 transitions in terms of x and y.

OPEN	to OPEN	has a value of	$x+y$
OPEN	to SPARE	has a value of	10
OPEN	to STRIKE	has a value of	10
SPARE	to OPEN	has a value of	$2x+y$
SPARE	to SPARE	has a value of	$x+10$
SPARE	to STRIKE	has a value of	20
STRIKE	to OPEN	has a value of	$2x+2y$
STRIKE	to SPARE	has a value of	20
STRIKE	to DOUBLE	has a value of	20
DOUBLE	to OPEN	has a value of	$3x+2y$
DOUBLE	to SPARE	has a value of	$x+20$
DOUBLE	to DOUBLE	has a value of	30.

(9)

Thus, considering the information in (8) and (9), we see that the generating functions for the 16 transitions are given by the following transition matrix, T.

$$\begin{pmatrix} \sum\limits_{(x,y)\in A} t^{x+y} & 10t^{10} & t^{10} & 0 \\ \sum\limits_{(x,y)\in A} t^{2x+y} & \sum\limits_{x=0}^{9} t^{x+10} & t^{20} & 0 \\ \sum\limits_{(x,y)\in A} t^{2x+2y} & 10t^{20} & 0 & t^{20} \\ \sum\limits_{(x,y)\in A} t^{3x+2y} & \sum\limits_{x=0}^{9} t^{x+20} & 0 & t^{30} \end{pmatrix} \tag{10}$$

The rows of the matrix T represent the first state while the columns of T represent the second state in the order OPEN, SPARE, STRIKE, and DOUBLE. For example, the entry in the third row and second column of T is the generating function for the transition STRIKE to SPARE.

In addition, the column matrix

$$C = \begin{pmatrix} \sum\limits_{(x,y,z)\in B} t^{x+y+z} \\ \sum\limits_{(x,y,z)\in B} t^{2x+y+z} \\ \sum\limits_{(x,y,z)\in B} t^{2x+2y+z} \\ \sum\limits_{(x,y,z)\in B} t^{3x+2y+z} \end{pmatrix} \tag{11}$$

represents the contribution made by the 10th frame depending on whether the 9th frame is in the OPEN, SPARE, STRIKE, or DOUBLE state.

Since T^9 will be the matrix that gives the generating functions for all possible scores commencing with a given state and terminating with another state through nine transitions, it follows that the generating function $P(t)$ in (6) will be the entry in the one-by-one matrix

$$RT^9C \tag{12}$$

where $R = (1,0,0,0)$.

Fortunately, we do not have to actually calculate and simplify the matrix expression in (12). An Apple IIe Pascal program was written which uses (12) and determines the coefficient of each term of (6) and hence finds the exact number of ways that each bowling score can occur. This program is available upon request. The distribution of all possible bowling scores generated by this program is listed in appendix A.

APPENDIX A

Distribution of Bowling Scores

0	1	50	11193770355829009	100	50613244155051856
1	20	51	13810930667765157	101	45887089510794122
2	210	52	16878453276117746	102	41483436078768079
3	1540	53	20435326129713654	103	37397371704961189
4	8855	54	24515635362932954	104	33621048067136846
5	42504	55	29146610869639549	105	30144388614623696
6	177100	56	34346628376654913	106	26955619314626157
7	657800	57	40123251227815383	107	24041709119775647
8	2220075	58	46471404549689351	108	21388640692533960
9	6906900	59	53371780703441318	109	18981680119465910
10	20030010	60	60789577452586487	110	16805547548715206
11	54627084	61	68673668434334934	111	14844654231857239
12	141116637	62	76956298564663402	112	13083276623221517
13	347336412	63	85553384395717227	113	11505812292077067
14	818558424	64	94365480254213528	114	10096971927616045
15	1854631380	65	103279445170253902	115	8842020009154293
16	4053948342	66	112170812747354087	116	7726929590817265
17	8574134256	67	120906827121834566	117	6738528470417086
18	17590903116	68	129350064451661348	118	5864552560171552
19	35084425512	69	137362512979745598	119	5093653838062639
20	68153183370	70	144809940796620325	120	4415377510495980
21	129156542039	71	151566341291631624	121	3820097597373727
22	239128282128	72	157518221668013078	122	3298981687014508
23	433093980298	73	162568486673578693	123	2843905747206868
24	768175029950	74	166639683923175378	124	2447444695948898
25	1335679056261	75	169676402232105648	125	2102793053565659
26	2278764308864	76	171646676234883305	126	1803790254604935
27	3817721269708	77	172542309343731946	127	1544848145184291
28	6285424931278	78	172378125687965848	128	1320992367181792
29	10176048813473	79	171190226627438257	129	1127775864826813
30	16210652213304	80	169033430825208027	130	961294388171457
31	25423690787719	81	165978103316094584	131	818085023387881
32	39274771758064	82	162106654714921075	132	695128788327698
33	59789973730461	83	157509948809043576	133	589753122859383
34	89736657900900	84	152283892386077931	134	499630252931260
35	132834787033075	85	146526364181517039	135	422696870992462
36	194006223597572	86	140334651650668803	136	357151976811922
37	279661205716974	87	133803399444707801	137	301400973036441
38	398018151390200	88	127023103852577896	138	254052574077937
39	559449136091831	89	120079021507938035	139	213889601295347
40	776838931567572	90	113050455155943519	140	179862464456172
41	1065940588576732	91	106010240661754449	141	151065169242834
42	1445705502357343	92	99024411737621323	142	126722015973414
43	1938561121705315	93	92151904402003308	143	106169469752641
44	2570605432880903	94	85444345654857875	144	88840622360686
45	3371684590465908	95	78945863453573001	145	74252067274687
46	4375319099346208	96	72693023944120045	146	61990415093876
47	5618445228564793	97	66714881583314335	147	51701385089887
48	7140942201229333	98	61033240145235763	148	43082666091665
49	8984922304030443	99	55663091133973346	149	35870481552300

150	29843343433392	200	1526313637	250	37965
151	24808172866872	201	1239515641	251	31193
152	20607116162379	202	1007719386	252	26131
153	17101443169235	203	818568928	253	21406
154	14181008701762	204	666193896	254	17422
155	11747089496422	205	542061609	255	13613
156	9723545122578	206	442072320	256	10696
157	8040378083433	207	360234562	257	7975
158	6644452641044	208	293886739	258	6005
159	5486702080236	209	239045260	259	4374
160	4529003381568	210	194337731	260	3534
161	3736165201688	211	157306293	261	3016
162	3081105018158	212	127325163	262	2635
163	2539255963377	213	102799565	263	2264
164	2091793858275	214	83194097	264	1933
165	1721930513702	215	67300605	265	1603
166	1416734360140	216	54691522	266	1323
167	1164733232308	217	44477808	267	1045
168	957190045595	218	36317458	268	810
169	785911852914	219	29606794	269	585
170	645295369580	220	24117404	270	406
171	529489941608	221	19554213	271	277
172	434606120455	222	15820964	272	258
173	356481490646	223	12736481	273	227
174	292487050484	224	10258846	274	206
175	239755303889	225	8244157	275	173
176	196550315542	226	6659561	276	150
177	160954253448	227	5381526	277	115
178	131791387388	228	4385243	278	90
179	107847709116	229	3576841	279	53
180	88241591630	230	2930385	280	26
181	72162948863	231	2376760	281	15
182	59038079745	232	1924226	282	15
183	48284335855	233	1541327	283	14
184	39509743432	234	1231527	284	14
185	32308399043	235	975760	285	13
186	26423428886	236	777090	286	13
187	21582203262	237	617547	287	12
188	17624621529	238	498228	288	12
189	14368737009	239	404981	289	11
190	11720626558	240	335065	290	11
191	9552812749	241	275998	291	1
192	7790240907	242	226966	292	1
193	6351933169	243	183727	293	1
194	5185250585	244	148442	294	1
195	4232118751	245	117291	295	1
196	3457204258	246	93525	296	1
197	2821392492	247	73010	297	1
198	2302090127	248	57960	298	1
199	1874802017	249	45826	299	1
				300	1

REFERENCE

1. C. N. Cooper and R. E. Kennedy. Is the Mean Bowling Score Awful? *J. Rec. Math.* 18(3) (1985-86).

Some Mathematical Reminiscences

MORRIS MARDEN
Milwaukee, WI 53217

1. Introduction It has been my good fortune, during over sixty years in mathematics, to have met and befriended many mathematicians*, some quite prominent in their day. In what follows I report not on their mathematical accomplishments and discoveries but relate, rather, my personal experiences with a few of them.

2. Cambridge, Massachusetts Harvard is where I received my basic training: a bachelor's degree in 1925, two years as a part-time instructor in 1925–27, and a Ph.D. in 1928. In those days the mathematics department was relatively small so that there was a close relationship between students and faculty. My first account concerns three Harvard faculty members, William F. Osgood, George D. Birkhoff and Julian Coolidge; and one visitor, Constantin Carathéodory.

Professor Osgood was a tall, black-bearded man, an 1886 graduate of Harvard College. In those days a young American desiring to pursue mathematics further had to go to Europe since mathematics in America was still in its infancy. Osgood chose Göttingen and Erlangen, Germany, where in 1890 he obtained a Ph.D. under the renowned Felix Klein. Returning to the United States and a position at Harvard, he made some distinguished contributions to complex variable theory, including results on conformal mapping and a two-volume treatise entitled, *Lehrbuch der Funktionentheorie*. By the way, this treatise was written in German, because English was not as yet regarded as an appropriate language for a mathematical treatise.

I met Osgood in 1920. I was a 15-year-old high school junior, eager to take advantage of the "anticipatory examinations" which a student could take if he were entering with more subjects than needed for regular admission to Harvard. The student could, thereby, earn in advance up to a year's college credit. I resolved to do just that, in particular to cover by myself the Harvard freshman course in analytic geometry and calculus. Osgood, then the mathematics department chairman, advised me as to the texts used in the course. I did well on the examination and was given an A. This success, plus Osgood's apparent interest in me, then persuaded me at age 16 to aspire to become a mathematician.

Osgood served as my college advisor, invited me to his home and visited me in the student infirmary. He was a superb teacher who struck a good balance between giving complete details and leaving matters to the student's initiative and intuition, frequently motivating a new subject through physical applications. My last meeting with him was during September 1932 in a Harvard Square cafeteria. Greeting me there was a clean shaven man whose voice I recognized as Osgood's. Since that was only a few days before my marriage, I invited him and his new young wife to the wedding, and both came.

Professor Birkhoff had the distinction of being among the first American-trained mathematicians to receive international recognition. As a sophomore I was hired to be problem reader in his calculus course, and subsequently took his various courses on differential equations, including one on the three-body problem. His lecture style was

*Over thirty of them are pictured in the *Pólya Picture Album* published by Birkhäuser, Boston and Basel, 1987.

quite different from Osgood's. He did not seem to come well prepared to class; however, when he got stuck on the proof of a difficult theorem, it was always interesting to watch this brilliant scholar work his way out. Every now and then he would propose what he alleged to be an unsolved problem. For one such problem, I developed a relatively simple solution. However, my hopes of thereby publishing my first mathematical article at age 19 were dashed when Birkhoff finally recalled that he had developed, several years earlier, a similar solution.

Professor Coolidge remains in my memory especially because of the following incident. When our small research club invited him to talk on his geometrical specialties, he accepted on the condition that we first have dinner at his home. Of course, we agreed. After seating us around his dining room table, each of us separated by our host's daughters or governesses, Coolidge noticed that he was lacking one female. He then excused himself, only to return in a few minutes wearing a bathrobe and declaring that he would substitute for the missing female. After dinner, as we bid good night to the ladies, he stood in the corner of the room throwing peanuts into the air and catching them in his mouth—quite a feat for a dignified Harvard professor.

Professor Carathéodory, a visiting professor at Harvard during 1926–27, was probably the most famous Greek mathematician since antiquity, yet he held a position not in Greece but in Munich. Physically, he was somewhat plump, of medium height, but cross-eyed so that it was hard to determine whether or not he was looking at you. After attending his lectures, I would often walk him home, choosing preferably a route passing some gardens containing tulips, his favorite flower. I visited him around Christmas, 1929, at his home in Munich, but my final contact with him occurred during 1936–37 when he was serving as visiting professor at the University of Wisconsin-Madison. During our walks, I tried to get his views on the Nazi regime but —I suppose for his own protection—he was completely silent about the matter. However, I could deduce his probable views from the fact that he had sent both his son and daughter back to Greece for their education.

3. Zürich, Switzerland I was awarded a National Research Fellowship, which allowed me to do research under E. B. Van Vleck at Madison, Wisconsin, during the summer of 1928, under Einar Hille at Princeton during the academic year of 1928–29, and under George Pólya in Zürich during fall, spring and summer of 1929–30. While in Zürich, I renewed my acquaintance with the famous Hermann Weyl, whose lectures I had attended in Princeton.

Professor Pólya, a native of Hungary, had done postdoctoral work at Göttingen. Though he was 40 and I was only 24, we developed a friendship, taking walks and swims together. I remember, for instance, that when we walked on a hot summer day, he carried his hat in his hand, but would don it hastily when he sighted a friend. This he did so as to be able to conform to the custom of tipping one's hat to any passing acquaintance. I found he was quite different from typical European professors. For example, at picnics he would mix and chat with the students, whereas the other professors remained aloof.

During my stay in Zürich, Pólya was collaborating with Hardy and Littlewood, the English mathematicians, to write the book *Inequalities*. Hardy would occasionally consult with Pólya in Zürich, following which Pólya would ask me to play tennis with Hardy (whose lectures I had attended at Princeton). I found Hardy to be quite a good player. In fact, he was an enthusiastic sports fan in general, including American baseball.

In 1940, when the Nazis seemed to be closing in on Switzerland, Pólya emigrated to the US, eventually taking a professorship at Stanford University. Among his

Professor and Mrs. Morris Marden with George Pólya (right).

colleagues there were three other refugees from Europe: Gábor Szegő, Stefan Bergman, and Charles Loewner. The presence of these four men greatly enhanced the reputation of the Stanford Mathematics Department, especially in analysis. Pólya remained active almost to his death in 1985 at age 97. He wrote several books and hundreds of research papers, in recognition of which he received several honorary degrees including one from the University of Wisconsin-Milwaukee. Furthermore, Stanford named a building in his honor.

Professor Hermann Weyl was among several faculty in Zürich who were quite friendly to me, an American student. He used to invite me to his home for dinner and ping-pong. Later on, when he was appointed to the prestigious Hilbert chair in Göttingen, he gave a public lecture entitled, "Die Stufen der Unendlichkeit" (the Levels of Infinity), which attracted a number of theologians who, however, were dismayed on finding that the lecture was on mathematics. Subsequently, a farewell picnic was given for him by the mathematics students and faculty at a Zürich open-air resort. On the grounds was a see-saw that Weyl seemed delighted to use, but when he was to receive a gift, he could not be found. Evidently, engrossed in a conversation with someone, he had wandered off.

4. Paris My visit to Zürich was interrupted by a stay in Paris, from January to April 1930, for study under Professor Paul Montel. However, before then, I did make short visits to Paris, and I begin by recounting one such visit.

I chose to return to Zürich indirectly by way of Strassburg (France) and Fribourg (Germany). I reached Fribourg late one evening, quite tired. I asked the railroad stationmaster for the name of a small, quiet hotel. When I arrived long after regular dinner hours, I was at first alone in the dining room, then a small dog held on a leash by an elderly woman entered, followed soon by a tall man with a long gray beard. I was amazed to discover that he was none other than Ferdinand Lindemann, who had made history about fifty years earlier as the first to prove that π is transcendental, that is, that π is not the root of any algebraic equation with whole number coefficients. When he said that he was about to take a walk, I asked (even though I was pretty tired) to accompany him. He took me through the darkened, local university campus, where he had once taught, pointing with his cane at the more important buildings. I finally got to ask him what he thought of the then modern mathematics and his answer was "Zehr compliziert." This is a reply that I must remember to use when some younger mathematician describes his research to me.

During a longer stay in Paris I made many friends among the American and French mathematicians. My closest friend was the eventually very well-known French mathematician, Jean Dieudonné, whom I had gotten to know during the period 1928–29 when we were both in residence at Princeton University. He and I visited the castles at Fontainebleau and Versailles and had a date every Thursday evening to have dinner in a provincial restaurant of Paris and then to attend a symphony. At dinner we always consumed two bottles of wine so that, though the evening began sedately, it was not so sedate when we headed for the concert hall.

5. Providence, Rhode Island During the summers of 1943 and 1944, I attended lectures at Brown University, which had received a large government grant to convert pure mathematicians into the applied mathematicians needed for the war effort. The staff at Brown was augmented by such European mathematicians as Stefan Bergman, Lipman Bers, Stefan Warshawski, and J. D. Tamarkin.

Whereas the first three men were refugees from Nazi-held Europe, the fourth, Tamarkin, had fled from the Soviet Union by tramping across the muddy fields of Russia and Poland. When he and his companion reached the German border, they were so covered with mud as to be unrecognizable. Doubtful of their identity, the border guard called the local American consul, who proceeded to test Tamarkin's knowledge of mathematics. However, the consul had gone only as far as calculus in college so that Tamarkin had no difficulty in passing the test. Tamarkin was a somewhat stout man who walked slowly, often taking with him his little Scotch terrier. He was very near-sighted so that when Arnold Ross and I invited him and his visitor, Antoni Zygmund, to the movies, we all had to sit in the front row.

6. Milwaukee I close my account by describing three distinguished visitors to the University of Wisconsin-Milwaukee during the postwar period: the Polish mathematician, Casimir Kuratowski; the Japanese mathematician, Akitsugu Kawaguchi, and the English mathematician, J. E. Littlewood.

Professor Kuratowski held two official positions: university professor and vice-president of the Polish Academy of Science. I met him in 1958 while attending the meetings of the Mathematics Institute of the Polish Academy, held in Lublin. After we gave our scheduled talks, he invited me to ride back to Warsaw in his chauffeured car. On the way, besides viewing the ruins of old palaces, we visited a former Nazi extermination camp maintained as a museum by the Polish government in memory of the thousands of innocent men, women and children who were murdered there.

Kuratowski described his narrow escape from the Gestapo. One day when the Gestapo visited him in his Warsaw apartment, he told them his profession. They replied that he was no longer a professor. Kuratowski sensed trouble brewing and had to think quickly. He showed the Gestapo a letter he had just received from the editor of the *Mathematische Zeitschrift* asking him to collaborate on the editorial work of the journal. The word "collaborate" seemed to appease the Gestapo men, who then left whereupon Kuratowski immediately went into hiding for the rest of the war.

I last saw Kuratowski when he gave a colloquium in Milwaukee. On the way back to the airport, I drove him along a route which passed many stores with good Polish nameplates. He was amazed to learn that more people of Polish ancestry lived in Milwaukee than in most Polish cities.

Professor Kawaguchi served as chairman of two large mathematics departments, one in Tokyo and the other in Sapporo, many miles from Tokyo. I met him too in 1958 at the meetings of the Polish Academy. As we both had gone to the meetings alone, we took walks together and otherwise provided mutual companionship. Several years later, he invited me to give a lecture at the meetings of the Japanese Mathematics Society in Tokyo and then to give several lectures at his university in Sapporo. While in Sapporo, my wife and I were entertained in a geisha house. Because my wife was accompanying me, Mrs. Kawaguchi went along with her husband. It was her first visit ever to a geisha house. Other mathematicians in our party had left their wives at home, as was the Japanese tradition. The geisha girls, attractive and talented, provided conversation at the table and then enacted a short play for our entertainment.

In appreciation of this hospitality, I invited Professor and Mrs. Kawaguchi to Milwaukee, where he gave a colloquium. We saw the Kawaguchis once more at the International Mathematical Congress in August 1970 in Nice. Thereafter, we continued to exchange New Year cards until their deaths.

Whereas Professors Kawaguchi and Kuratowski remained in Milwaukee only for a day or two, the third visitor, Professor Littlewood, spent nearly a semester there. Initially declining an invitation to spend September to mid-January in Milwaukee, he agreed to come provided he could leave by Christmas for the skiing season in Switzerland. My last contact with Littlewood was at Trinity College, Cambridge University. I understand that the University guaranteed him occupancy of his rooms after his retirement, along with an adequate supply of his favorite sherry. (In fact, soon after arriving in Milwaukee he made a deal for a case of sherry with the manager of his hotel.) His quarters at Trinity included a large living room overfilled with books and papers as well as a small piano and an open fireplace.

During my visit Littlewood took me on a tour of the college library. He seemed especially proud of the large collection of beautifully bound volumes of mathematics dating over several centuries. He would have been equally proud of the inscription on a plaque on the wall of the college chapel. It listed the many distinguished mathematicians (including Newton) who had been associated with Trinity College. Eventually this plaque would list his name alongside that of Godfrey H. Hardy, his close friend and colleague.

A Remark on Boolean Rings

DESMOND MAC HALE
University College
Cork, Ireland

Let S be a non-empty set and P its power set, i.e., the set of all subsets of S. Define two binary operations denoted by $+$ and $\cdot$ on P as follows: for A, $B \in P$ let $A+B=(A \cup B)-(A \cap B)$ and $A \cdot B = A \cap B$. The operation $+$ is often called the symmetric difference operation. It is an easy exercise to show that the operation $+$ is closed, associative, and commutative on P and that the empty set Φ is the identity element for $\{P, +\}$. In addition, since for each A, $A+A=\Phi$, A is its own inverse later, he invited me to give a lecture at the meetings of the Japanese Mathematics Society in Tokyo and then to give several lectures at his university in Sapporo. While in Sapporo, my wife and I were entertained in a geisha house. Because my wife was accompanying me, Mrs. Kawaguchi went along with her husband. It was her first visit ever to a geisha house. Other mathematicians in our party had left their wives at home, as was the Japanese tradition. The geisha girls, attractive and talented, provided conversation at the table and then enacted a short play for our entertainment.

In appreciation of this hospitality, I invited Professor and Mrs. Kawaguchi to Milwaukee, where he gave a colloquium. We saw the Kawaguchis once more at the International Mathematical Congress in August 1970 in Nice. Thereafter, we continued to exchange New Year cards until their deaths.

Whereas Professors Kawaguchi and Kuratowski remained in Milwaukee only for a day or two, the third visitor, Professor Littlewood, spent nearly a semester there. Initially declining an invitation to spend September to mid-January in Milwaukee, he agreed to come provided he could leave by Christmas for the skiing season in Switzerland. My last contact with Littlewood was at Trinity College, Cambridge University. I understand that the University guaranteed him occupancy of his rooms after his retirement, along with an adequate supply of his favorite sherry. (In fact, soon after arriving in Milwaukee he made a deal for a case of sherry with the manager of his hotel.) His quarters at Trinity included a large living room overfilled with books and papers as well as a small piano and an open fireplace.

During my visit Littlewood took me on a tour of the college library. He seemed especially proud of the large collection of beautifully bound volumes of mathematics dating over several centuries. He would have been equally proud of the inscription on a plaque on the wall of the college chapel. It listed the many distinguished mathematicians (including Newton) who had been associated with Trinity College. Eventually this plaque would list his name alongside that of Godfrey H. Hardy, his close friend and colleague.

A Remark on Boolean Rings

DESMOND MAC HALE
University College
Cork, Ireland

Let S be a non-empty set and P its power set, i.e., the set of all subsets of S. Define two binary operations denoted by $+$ and $\cdot$ on P as follows: for $A, B \in P$ let $A+B=(A \cup B)-(A \cap B)$ and $A \cdot B = A \cap B$. The operation $+$ is often called the symmetric difference operation. It is an easy exercise to show that the operation $+$ is closed, associative, and commutative on P and that the empty set Φ is the identity element for $\{P, +\}$. In addition, since for each A, $A+A=\Phi$, A is its own inverse under $+$. Next, we can verify that $\{P, \cdot\}$ is closed, associative, and commutative and finally that the distributive law $A \cdot (B + C) = (A \cdot B) + (A \cdot C)$ holds for all A, B, $C \in P$. Thus $\{P, + \cdot\}$ is a commutative ring, and we note that the set S itself is the unity element for $\{P, \cdot\}$.

A ring $\{R, +, \cdot\}$ in which $x^2 = x$ for all x is called *Boolean*. In such a ring, $x = x^2 = (-x)^2 = -x$, so $x + x = 0$, for all x. In addition, for all x and y, $(x + y)^2 = x + y$, so on cancellation we get $xy + yx = 0$. Combining this with $xy + xy = 0$, we get $xy = yx$, so a Boolean ring is commutative. Note that we do not require a Boolean ring to have a multiplicative identity element.

In the ring $\{P, +, \cdot\}$ we see that $A \cdot A = A$ for all A so that $\{P, +, \cdot\}$ is a Boolean ring.

The purpose of this note is to investigate other algebraic conditions which give rise to Boolean rings.

THEOREM. *Let R be a ring such that for all $x \in R$, $2x = 0$ and $x^{2^n+1} = x$, for a fixed integer $n > 1$. Then R is Boolean.*

We need the following result which one might suspect from a perusal of the Pascal triangle.

LEMMA. *Let k be a power of 2. Then the binomial coefficients $\binom{k}{r}$ are even numbers for $k > r > 1$.*

This result is easily established by induction and the binomial theorem using the fact that

$$(1 + x)^{2^m}(1 + x)^{2^m} = (1 + x)^{2^{m+1}}.$$

We can now prove the theorem quoted. For convenience, we put $2^n = k$. For all $x \in R$, $(x^2 + x) = (x^2 + x)^{k+1} = (x^2 + x)^k(x^2 + x) = (x^{2k} + x^k)(x^2 + x)$, by the lemma. With simplification and cancellation we reduce this equation to

$$x^{2k+1} = x^{k+2},$$

so

$$x^2 = x^{2k+2} = x^{k+3} = (x^{k+1})(x^2) = x^3.$$

Thus $x^2 = x^3 = x^4 = \cdots = x^{k+1} = x$, and so R is Boolean. Finally, we note that it is *not* true in general that the conditions $2x = 0$ and $x^m = x$ for all x and a fixed *odd* integer $m > 1$ force a ring to be Boolean. The smallest odd integer for which this fails is 7, so we pose the following problems for the reader:

(i) Find a non-Boolean ring R that satisfies the identities $x^7 = x$ and $2x = 0$ for all x. Of course by the well-known theorem of Jacobson (see [1]) R must be commutative.

(ii) For which odd integers t not of the form $2^n + 1$ do the conditions $x^t = x$, $2x = 0$ for all $x \in R$ force R to be Boolean?

REFERENCES

1. I. N. Herstein, *Non-Commutative Rings*, Carus Mathematical Monographs, MAA, 1968.

Averages Still on the Move

P. S. BULLEN
University of British Columbia
Vancouver, British Columbia, Canada V6T 1Y4

1. Introduction Let **a** and **w** denote n-tples of positive numbers, $n \geqslant 1$; $\mathbf{a} = (a_1, \dots, a_n)$; $\mathbf{w} = (w_1, \dots, w_n)$ say, and write $W = \sum_{i=1}^n w_i$. Let $-\infty \leqslant r \leqslant \infty$ then there is an immense literature on what is called the rth power mean (or average) of **a** with weight (or weights) **w**. This mean is defined as follows, the actual form of definition depending on r;

$$
\begin{aligned}
M_n^{[r]}(\mathbf{a};\mathbf{w}) &= \left\{ \frac{1}{W} \sum_{i=1}^n w_i a_i^r \right\}^{1/r}, \quad \text{if } -\infty < r < \infty, \quad r \neq 0, \\
&= \left\{ \prod_{i=1}^n a_i w_i \right\}^{1/W}, \quad \text{if } r = 0, \\
&= \max \mathbf{a} = \max\{a_1, \dots, a_n\}, \quad \text{if } r = \infty, \\
&= \min \mathbf{a} = \min\{a_1, \dots, a_n\}, \quad \text{if } r = -\infty.
\end{aligned} \tag{1}
$$

All these symbols may seem formidable so let us spend a little time with them. Some special cases of the rth power means have their own names and notations.

(i) $r = -1$; the harmonic mean;

$$H_n(\mathbf{a};\mathbf{w}) = \frac{W}{\dfrac{w_1}{a_1} + \cdots + \dfrac{w_n}{a_n}}; \tag{2}$$

(ii) $r = 0$; the geometric mean;

$$G_n(\mathbf{a};\mathbf{w}) = \left(a_1^{w_1} \cdots a_n^{w_n}\right)^{1/W}; \tag{3}$$

(iii) $r = 1$; the arithmetic mean;

$$A_n(\mathbf{a};\mathbf{w}) = \frac{w_1 a_1 + \cdots + w_n a_n}{W}; \tag{4}$$

(iv) $r = 2$; the quadratic mean;

$$Q_n(\mathbf{a};\mathbf{w}) = \sqrt{\frac{w_1 a_1^2 + \cdots + w_n a_n^2}{W}}. \tag{5}$$

To discuss these further it is convenient to introduce a shorthand for use with n-tples. Given an n-tple $\mathbf{a} = (a_1, \dots, a_n)$ and a function f we will let $f(\mathbf{a})$ denote the n-tple $(f(a_1), \dots, f(a_n))$; for instance if $-\infty < r < \infty$, $\mathbf{a}^r = (a_1^r, \dots, a_n^r)$, $\ln \mathbf{a} = (\ln a_1, \dots, \ln a_n)$.

Using this notation some simple relations between various means can be written in a convenient form: thus from (2) and (4),

$$\frac{1}{H_n(\mathbf{a};\mathbf{w})} = \frac{1}{W}\left(\frac{w_1}{a_1} + \cdots + \frac{w_n}{a_n}\right) = A_n(\mathbf{a}^{-1};\mathbf{w}). \tag{6}$$

This simple relation leads to the neglect in the theoretical discussions of the harmonic mean; its properties follow, using (6), from those of the arithmetic mean.

Now looking at (3) and (4) we see that

$$\ln\{G_n(\mathbf{a};w)\} = \frac{1}{W}(w_1 \ln a_1 + \cdots + w_n \ln a_n) = A_n(\ln \mathbf{a};\mathbf{w}) \tag{7}$$

or

$$G_n(\mathbf{a};\mathbf{w}) = \exp(A_n(\ln \mathbf{a};\mathbf{w})). \tag{7'}$$

In fact all the rth power means can be written in a form similar to (7'); for if $-\infty < r < \infty$, $r \neq 0$, and $f(x) = x^r$, $x > 0$, then f has an inverse function $f^{-1}(x) = x^{1/r}$ and

$$M_n^{[r]}(\mathbf{a};\mathbf{w}) = f^{-1}(A_n(f(\mathbf{a});\mathbf{w}));$$

in the case $r = 0$ we get this form, (7'), by taking $f = \ln$, $f^{-1} = \exp$.

The name mean is easily justified since it is readily seen that

$$\min \mathbf{a} \leqslant M_n^{[r]}(\mathbf{a};\mathbf{w}) \leqslant \max \mathbf{a}; \tag{8}$$

for instance if $-\infty < r < \infty$, $r \neq 0$ to obtain the right-hand inequality in (8),

$$M_n^{[r]}(\mathbf{a};\mathbf{w}) = \left\{\frac{1}{W}\sum_{i=1}^{n} w_i a_i^r\right\}^{1/r}$$

$$\leqslant \left\{\frac{1}{W}\sum_{i=1}^{n} w_i (\max \mathbf{a})^r\right\}^{1/r} = \max \mathbf{a}$$

and, in fact, $M_n^{[r]}(\mathbf{a};\mathbf{w}) < \max \mathbf{a}$ unless $a_1 = \cdots = a_n$.

The rather eccentric definitions (1) in the cases $r = 0, \pm\infty$ are justified since

$$\lim_{r \to 0} M_n^{[r]}(\mathbf{a};\mathbf{w}) = G_n(\mathbf{a};\mathbf{w}), \tag{9}$$

$$\lim_{r \to \infty} M_n^{[r]}(\mathbf{a};\mathbf{w}) = \max \mathbf{a}, \tag{10}$$

$$\lim_{r \to -\infty} M_n^{[r]}(\mathbf{a};\mathbf{w}) = \min \mathbf{a}. \tag{11}$$

All these results can be obtained by nice uses of L'Hôpital's rule, although a little extra care is needed for (10) and (11). To see (9) consider

$$\ln\left(M_n^{[r]}(\mathbf{a};\mathbf{w})\right) = \frac{n(r)}{r},$$

where

$$n(r) = \ln\left(\frac{1}{W}\sum_{i=1}^{n} w_i a_i^r\right).$$

Since $\lim_{r \to 0} n(r) = 0$ we can apply L'Hôpital's rule as follows,

$$\begin{aligned}\lim_{r\to 0}\ln\big(M_n^{[r]}(\mathbf{a};\mathbf{w})\big) &= \lim_{r\to 0}\frac{n(r)}{r}\\ &= \lim_{r\to 0} n'(r)\\ &= \lim_{r\to 0}\left(\frac{\sum_{t=1}^{n} w_i a_i^r \ln a_i}{\sum_{i=1}^{n} w_i a_i^r}\right)\\ &= \frac{1}{W}\sum_{i=1}^{n} w_i \ln a_i = A_n(\ln\mathbf{a};\mathbf{w}),\end{aligned}$$

and (9) follows from (7′).

Inequality (8) is a particular case of a more general and basic result; if $-\infty \leqslant r < s \leqslant \infty$, then

$$M_n^{[r]}(\mathbf{a};\mathbf{w}) \leqslant M_n^{[s]}(\mathbf{a};\mathbf{w}) \tag{12}$$

with equality only when $a_1 = \cdots = a_n$.

While this result appears difficult to establish it can be rewritten by methods used above into a form that makes it almost trivial. To avoid considering lots of special cases let us assume $0 < r < s < \infty$ when what we have to prove is that

$$\left(\frac{1}{W}\sum_{i=1}^{n} w_i a_i^r\right)^{1/r} \leqslant \left(\frac{1}{W}\sum_{i=1}^{n} w_i a_i^s\right)^{1/s}; \tag{12′}$$

put $\mathbf{b} = \mathbf{a}^r$, i.e., $b_1 = a_1^r, \ldots, b_n = a_n^r$ when $\mathbf{a}^s = \mathbf{b}^{s/r}$, i.e. $a_1^s = b_1^{s/r} = (a_1^r)^{s/r}$ etc.; now let $t = s/r$, when $t > 1$ and (12′) is just

$$\left(\frac{1}{W}\sum_{i=1}^{n} w_i b_i\right)^t \leqslant \frac{1}{W}\sum_{i=1}^{n} w_i b_i^t,$$

or letting $f(x) = x^t$, $x > 0$

$$f\left(\frac{1}{W}\sum_{i=1}^{n} w_i b_i\right) \leqslant \frac{1}{W}\sum_{i=1}^{n} w_i f(b_i). \tag{13}$$

In this form this is equivalent to saying that f is convex and is known as Jensen's inequality; and, of course, if $f(x) = x^t$, $(x > 0)$, $f''(x) = t(t-1)x^{t-2} \geqslant 0$ since $t > 1$, and so f is convex, which proves that (13) holds and so also (12′). In fact $f'' > 0$ so f is what is called strictly convex, and (13) is strict unless $b_1 = \cdots = b_n$, which means (12′) is strict unless $a_1 = \cdots = a_n$.

This kind of argument is typical of the theory of inequalities where many a well-known result lurks in impenetrable disguise and is seemingly impossible to prove.

Inequality (12), besides including (8), includes as a special case the famous inequality between the more well-known means (2)–(5):

$$\min\mathbf{a} \leqslant H_n(\mathbf{a};\mathbf{w}) \leqslant G_n(\mathbf{a};\mathbf{w}) \leqslant A_n(\mathbf{a};\mathbf{w}) \leqslant Q_n(\mathbf{a};\mathbf{w}) \leqslant \max\mathbf{a}, \tag{14}$$

all these inequalities being strict unless $a_1 = \cdots = a_n$.

The above results are discussed in many books on inequalities; see in particular [1], [3], [5].

2. A result of Hoehn and Niven In a very interesting article [4], these authors considered how the classical means, (2) (3) (4) and (5), with equal weights, that is, $\mathbf{w} = (1/n, \dots, 1/n)$ behave when the terms in the n-tple $\mathbf{a}$ increase at a uniform rate. Later Brenner [2] extended their results to the general rth power means, again with $\mathbf{w} = (1/n, \dots, 1/n)$. It is the object of this note to give a very simple proof of this result, and to allow general $\mathbf{w}$.

For this purpose let us introduce some more notation; $\mathbf{e} = (1, \dots, 1)$ so $t\mathbf{e} = (t, \dots, t)$, $\mathbf{a} + t\mathbf{e} = (a_1 + t, \dots, a_n + t)$.

If now $-\infty < r < \infty$ put

$$M_r(t) = M_n^{[r]}(\mathbf{a} + t\mathbf{e}; \mathbf{w}), \tag{15}$$

and in the special case $r = 1$ write

$$A(t) = M_1(t) = A_n(\mathbf{a} + t\mathbf{e}; \mathbf{w}).$$

Then, using (4),

$$\begin{aligned} A(t) &= \frac{1}{W} \sum_{i=1}^{n} w_i(a_i + t) \\ &= A_n(\mathbf{a}; \mathbf{w}) + t, \end{aligned} \tag{16}$$

and

$$\begin{aligned} \max(\mathbf{a} + t\mathbf{e}) = \max\{a_1 + t, \dots, a_n + t\} &= t + \max \mathbf{a}, \\ \min(\mathbf{a} + t\mathbf{e}) &= t + \min \mathbf{a}. \end{aligned} \tag{17}$$

If now $t \to \infty$ the behaviour of $M_n(t)$ gives that of $M_n^{[r]}(\mathbf{a}; \mathbf{w})$ when the elements of $\mathbf{a}$ increase at a uniform rate. The result of Hoehn and Niven is quite a reasonable one —in fact, it is what one might expect; if only t is large enough all the power means approach the arithmetic mean; equivalently, from (16), sooner or later the value of the rth power mean, r finite, can be obtained from the original value of the arithmetic mean plus a factor depending only on how much the entries in $\mathbf{a}$ have grown. This is stated formally in the following theorem.

THEOREM 1. *With the notation* (15)

$$\lim_{t \to \infty} \{M_r(t) - A(t)\} = 0 \tag{18}$$

or, equivalently

$$\lim_{t \to \infty} \{M_r(t) - t\} = A_n(\mathbf{a}; \mathbf{w}). \tag{19}$$

The equivalence of (18) and (19) is an immediate consequence of (16). The existence of the limit in (18) or (19) follows from the proof of the theorem, but it is of some interest to see it directly. For this let us write

$$f(t) = M_r(t) - A(t). \tag{20}$$

Then by (8)

$$\min(\mathbf{a} + t\mathbf{e}) - A(t) \leqslant f(t) \leqslant \max(\mathbf{a} + t\mathbf{e}) - A(t),$$

and so by (16) and (17) f is bounded since we see that

$$\min \mathbf{a} - A_n(\mathbf{a};\mathbf{w}) \leqslant f(t) \leqslant \max \mathbf{a} - A_n(\mathbf{a};\mathbf{w}).$$

Now let us show f is monotonic, from which it will follow that $\lim_{t\to\infty} f(t)$ exists. Clearly f has a derivative and

$$\begin{aligned} f'(t) &= M_r'(t) - A'(t) \\ &= M_r'(t) - 1, \quad \text{by (16)}. \end{aligned} \tag{21}$$

To calculate $M_r'(t)$ let us first suppose $r \neq 0$, so that

$$M_r^r(t) = \frac{1}{W}\sum_{i=1}^n w_i(a_i + t)^r.$$

Hence,

$$\begin{aligned} M_r^{r-1}(t)M_r'(t) &= \frac{1}{W}\sum_{i=1}^n w_i(a_i+t)^{r-1} \\ &= \left(M_n^{[r-1]}(\mathbf{a}+t\mathbf{e};\mathbf{w})^{r-1}\right) \\ &= M_{r-1}^{r-1}(t), \end{aligned}$$

which, on simplifying, gives,

$$M_r'(t) = \left(\frac{M_{r-1}(t)}{M_r(t)}\right)^{r-1}. \tag{22}$$

If now $r = 0$, we have from (7), that

$$M_0(t) = \exp\{A_n(\ln(\mathbf{a}+t\mathbf{e});\mathbf{w})\}$$

so

$$M_0'(t) = M_0(t)\cdot A_n(\mathbf{b};w),$$

where $\mathbf{b}$ is the n-tple $(1/(a_1+t),\ldots,1/(a_n+t))$, or by (6),

$$M_0'(t) = \frac{M_0(t)}{M_{-1}(t)},$$

which is just (22) with $r = 0$.

These calculations lead to the following result which generalizes a lemma of Hoehn and Niven [4, p. 152].

LEMMA 2. *With the notation of* (15) *if* $\mathbf{a}$ *is not constant then*

(i) $M_r'(t) = 1$, *if* $r = 1$,
(ii) $M_r'(t) > 1$, *if* $r < 1$,
(iii) $M_r'(t) < 1$, *if* $r > 1$.

Proof. By (12), $M_{r-1}(t) < M_r(t)$, and so the lemma follows from (22). If $\mathbf{a}$ is constant, $M_r'(t) = 1$ for all r.

COROLLARY 3. *The function* f, *defined by* (20) *is monotonic.*

Proof. More precisely, from Lemma 2 and (21), if $\mathbf{a}$ is not constant f is strictly increasing if $r < 1$, strictly decreasing if $r > 1$.

If $r = 1$ or $\mathbf{a}$ is constant, f is constant and zero.

COROLLARY 4. *With the notation of* (15) *if* a *is not constant and* $-\infty < r < 1 < s < \infty$, *then*

$$M_r'(t) > M_s'(t). \tag{23}$$

Proof. This is an immediate consequence of Lemma 2.

If we assume only $-\infty < r < s < \infty$, an example of Hoehn and Niven [4, p. 153 and pp. 154–155] shows that (23) need not hold.

3. Proof of Theorem 1 It is clear that if either a is constant or $r = 1$, then from (20)

$$f(t) = M_r(t) - A(t) = 0,$$

and the result is trivial, so we can assume $r \neq 1$ and a is not constant.

First suppose $r \neq 0$. Then by the definitions (20), (16), (15), (1),

$$\begin{aligned} f(t) &= t\left\{\left(\frac{1}{W}\sum_{i=1}^{n} w_i\left(1+\frac{a_i}{t}\right)^r\right)^{1/r} - 1 - \frac{1}{t}A_n(\mathbf{a};\mathbf{w})\right\} \\ &= t\left\{\left(\frac{1}{W}\sum_{i=1}^{n} w_i\left(1+\frac{ra_i}{t}+\frac{r(r-1)}{2}\frac{a_i^2}{t^2}+\cdots\right)\right)^{1/r} - 1 - \frac{1}{t}A_n(\mathbf{a};\mathbf{w})\right\} \\ &= \frac{(r-1)}{2t}\left(A_n(\mathbf{a}^2;\mathbf{w}) + A_n^2(\mathbf{a};\mathbf{w})\right) + \cdots \\ &= O\left(\frac{1}{t}\right) \to 0, \quad t \to \infty, \end{aligned}$$

as had to be proved.

If now $r = 0$, then by (14) and the definitions (15)

$$0 > M_0(t) - A(t) > M_{-1}(t) - A(t)$$

and as $t \to \infty$ the last term on the right tends to zero, if we use the case $r = -1$, which has just been considered.

It should perhaps be noted that the discussion of this section assumes and needs r to be finite; in particular it is easy to check that (19) is false if $r = \pm\infty$.

REFERENCES

1. E. F. Beckenbach and R. Bellman, *Inequalities*, Springer-Verlag, Berlin-Heidelberg-New York, 1961.
2. J. L. Brenner, Limits of means for large values of the variables, *Pi Mu Epsilon J.* 8(1985), 160–163.
3. G. M. Hardy, J. E. Littlewood and G. Pólya, *Inequalities*, Cambridge, 1934.
4. L. Hoehn and I. Niven, Averages on the move, this MAGAZINE 58(1985), 151–156.
5. D. S. Mitrinović and P. M. Vasić, *Analytic Inequalities*, Springer-Verlag, Berlin-Heidelberg-New York, 1970.

A New Proof of the Double Butterfly Theorem

LARRY HOEHN
Austin Peay State University
Clarksville, TN 37044

The butterfly problem, which dates back to at least 1815, seems to hold a continual interest. It has had a variety of proofs and extensions in recent years. A synopsis of this celebrated problem is given in [4] and a very comprehensive update is given in [1]. In this note, we provide a new proof of one of its extensions—namely, the double butterfly theorem.

This theorem, as given by D. Jones in [3], is

DOUBLE BUTTERFLY THEOREM. *Let PQ be a fixed chord of a circle and let "butterfly R" and "butterfly S" be inscribed in the circle and oriented such that their wings cut PQ (in order from left to right) at* R_4, R_3, R_2, R_1 *and* S_1, S_2, S_3, S_4 *respectively. If* $PR_1 = QS_1$, $PR_2 = QS_2$, *and* $PR_3 = QS_3$, *then* $PR_4 = QS_4$ *(see* FIGURE 1).

For our proof we use a lemma attributed to Hiroshi Haruki [2].

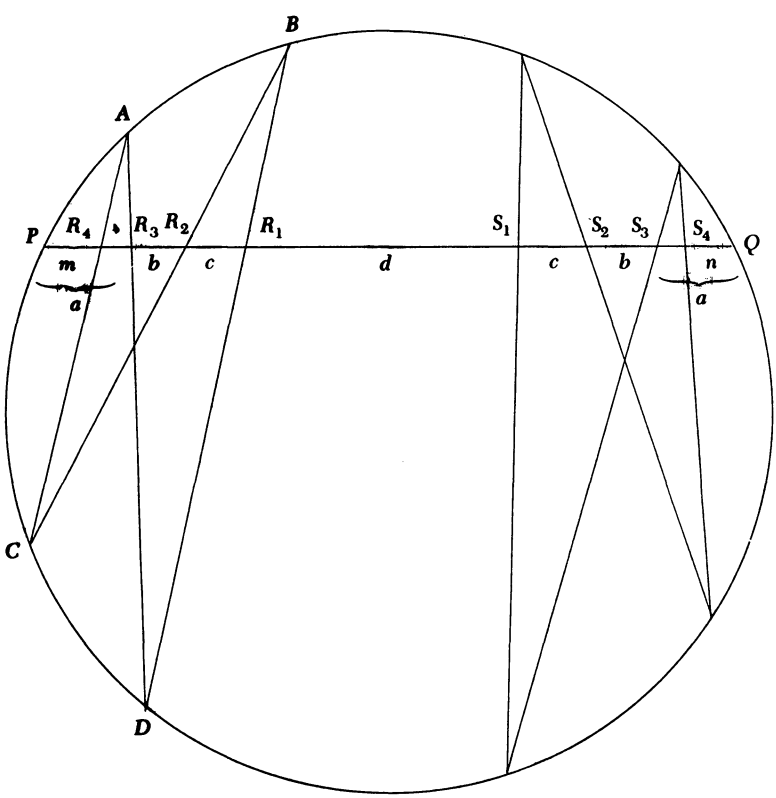


FIGURE 1

LEMMA. *Suppose PQ and CD are non-intersecting chords in a circle and that B is a variable point on the arc PQ remote from C and D. Then for each position of B, the lines BC and BD cut PQ into segments of lengths x, y, z where xz/y is a constant* (*see* FIGURE 2).

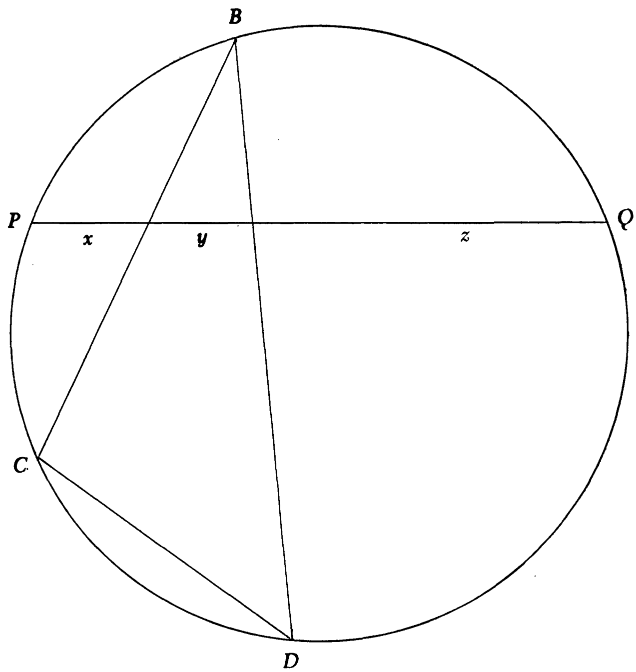


FIGURE 2

This lemma, which deserves a much wider recognition, is proved in [2] by elementary geometry and in [1] by cross ratios. For our proof of the theorem we let $m = PR_4$, $n = QS_4$, $a = PR_3 = QS_3$, $b = R_3R_2 = S_3S_2$, $c = R_2R_1 = S_2S_1$, $d = R_1S_1$, and $e = a + b + c + d$ as seen in FIGURE 1. By applying Haruki's lemma twice to points A and B and fixed chords PQ and CD of butterfly R, we have

$$\frac{m(b+c+e)}{a-m} = \frac{(a+b)e}{c},$$

since both are equal to the same constant. Similarly

$$\frac{n(b+c+e)}{a-n} = \frac{(a+b)e}{c}$$

for butterfly S. By solving these two equations simultaneously, we obtain $m = n$ which completes the proof.

REFERENCES

1. Leon Bankoff, The metamorphosis of the butterfly problem, this MAGAZINE 60 (1987), 195–210.
2. Ross Honsberger, The butterfly problem and other delicacies from the noble art of Euclidean geometry, Part 1: *TYCMJ* 14 (1983) 2–8; Part 2: *TYCMJ* 14 (1983), 154–158.
3. Dixon Jones, A double butterfly theorem, this MAGAZINE 49 (1976), 86–87.
4. Leo Sauvé, The celebrated butterfly problem, *EUREKA* (now *CRUX MATHEMATICORUM*) (1976), 2–5.

Reading a Bus-Connected Processor Array

DAN KALMAN
The Aerospace Corporation
P.O. Box 92957
Los Angeles, CA 90009

An area of active research in computer science is concerned with the design and programming of parallel computing devices. In one design, an M by N array of processing elements communicates with a host computer using a series of busses. Each row and column of the array has a bus. Although a processing element can write a value to either its row bus or its column bus (see FIGURE 1), only one value can be output to each bus at a time. It is therefore possible to transfer to the host one entry from each row and one entry from each column in a single operation cycle. Now imagine that an M by N matrix resides in the processor array with one entry stored in each processing element. This might be the situation, for example, if the processor array has just finished computing a matrix product. How many operation cycles are required to transmit the entire M by N matrix to the host? Since there are MN memory entries and $M+N$ can be read at once, it appears that approximately $MN/(M+N)$ cycles should suffice to transmit the matrix. However, in order to be sure that all $M+N$ busses are used at each step, some care is required. For example, transmitting all the entries from the first row and column at the first step would only yield $M+N-1$ matrix entries. Sending the remaining entries of the second row and column at the next step would only produce $M+N-3$ entries. Continuing in this fashion would require as many steps to finish as the smaller of M and N. This is no better than using only the larger set of busses (the column busses if $N>M$). A little exploration with special cases should convince you that it *is* possible to schedule the reading operations so that $N+M$ entries are obtained at each step, although it may not be clear how to establish a generally applicable procedure. Our goal is to present a simple scheduling algorithm that uses all the busses at each step (except possibly the last) for any rectangular array.

First, impose an order on the processing elements as follows. The $(1,1)$ element is first. The successor of element (i,j) is $(i+1,j+1)$ where the addition in the first component is performed modulo M, and the addition in the second component is performed modulo N. This shall be referred to as the *diagonal order*. TABLE 1 lists the ordered elements for a 3 by 8 array. The ordering can also be understood geometrically as a path that traverses all the entries of the array. That is, begin in the $(1,1)$ position and travel down the diagonal, wrapping from the bottom edge of the matrix to the top, and from the right side to the left as if the matrix were a torus. FIGURE 2 shows this geometric interpretation for the ordering listed in TABLE 1. Note that it is possible for the ordering procedure to repeat the $(1,1)$ position without traversing all the array entries. In this case only a subset of the entries are ordered. Otherwise, we may say that the array is exhausted by the diagonal order. Note further that among the ordered elements, any M consecutive entries come from different rows, while any N consecutive entries come from different columns. When it exhausts an array, the diagonal order can be used to schedule reading operations, as formalized in the following lemma.

LEMMA 1. *If an M by N processor array is exhausted by the diagonal ordering, a scheduling can be determined which makes use of all busses at each step, save possibly the last.*

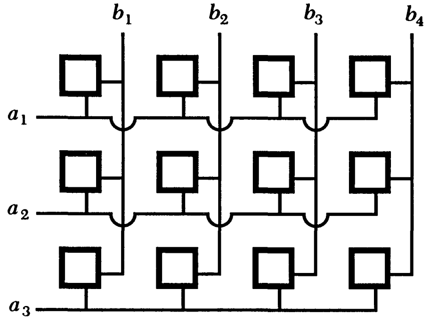


FIGURE 1
A 3 by 4 bus-connected processor array.

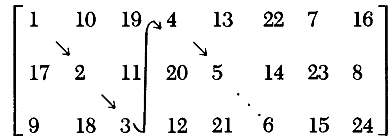


FIGURE 2
Diagonal order diagram for a 3 by 8 array.

TABLE 1. Diagonal order for a 3 by 8 array

Processor Number	Matrix Position	Processor Number	Matrix Position	Processor Number	Matrix Position
1	(1, 1)	9	(3, 1)	17	(2, 1)
2	(2, 2)	10	(1, 2)	18	(3, 2)
3	(3, 3)	11	(2, 3)	19	(1, 3)
4	(1, 4)	12	(3, 4)	20	(2, 4)
5	(2, 5)	13	(1, 5)	21	(3, 5)
6	(3, 6)	14	(2, 6)	22	(1, 6)
7	(1, 7)	15	(3, 7)	23	(2, 7)
8	(2, 8)	16	(1, 8)	24	(3, 8)

Proof. To schedule the output operations, follow the diagonal ordering *in both directions*. Specifically, output sequentially from the beginning of the order using the row busses and from the end of the order using the column busses. Thus, at the first time step, transmit the first M entries using the row busses and the last N entries with the column busses; at the second time step, send the next M entries with row busses and the next to the last N entries with column busses; and so on. Continuing in this fashion, $N + M$ entries can be obtained in each time step until the array is exhausted, or fewer than $N + M$ entries remain. Indeed, if NM is divisible by $N + M$, the array is completely read after $NM/(N + M)$ time steps. Otherwise, after $\text{INT}(NM/(N + M))$ time steps, fewer than $N + M$ entries remain. As these are still consecutive in the diagonal ordering, up to M may be output using row busses, and if need be, any remaining entries can be transmitted using column busses. This completes the proof.

Under what conditions does the diagonal order fail to exhaust an array, and how can the output of data from the processing elements then be scheduled? These questions are easily addressed in the context of group theory. Let the index pair (i, j) be considered an element of the group $Z_M \oplus Z_N$, the direct sum of the integers modulo M and N, respectively. The diagonal order is generated by following the multiples of $(1, 1)$ in this group. Let $\langle(1, 1)\rangle$ denote the cyclic subgroup generated by $(1, 1)$. If $\langle(1, 1)\rangle$ is the whole group, then the diagonal order exhausts the array. It is straightforward to verify that this will occur if and only if M and N are relatively prime. In fact, if s is the order of $\langle(1, 1)\rangle$, then $(1 + s, 1 + s) = (1, 1)$, and consequently, s must be divisible by both M and N. (What is s exactly?) If M and N are not relatively prime, $\langle(1, 1)\rangle$ is a proper subgroup of $Z_M \oplus Z_N$ so that the diagonal order does not apply to all the array elements. However, it is possible to extend the order to the entire array in such a way that the output algorithm remains valid. The extended order is presented in the next lemma.

LEMMA 2. *The elements of any M by N array can be ordered in a sequence $e_1, e_2, \ldots, e_{MN}$ such that the following properties hold:*

(1) *If $k \equiv 1 \bmod M$, the M consecutive entries beginning with e_k come from different rows; and*
(2) *If $k \equiv 1 \bmod N$, the N consecutive entries beginning with e_k come from different columns.*

Proof. If M and N are relatively prime, the diagonal order may be used. Otherwise, let $C_1, C_2, \ldots, C_r$ be the cosets of $\langle(1,1)\rangle$ with $C_1 = \langle(1,1)\rangle$. An ordering of the array entries is produced by ordering the elements of each coset, and then ordering the cosets. Thus, the successor of the last element of C_k is the first element of C_{k+1} for $k = 1, 2, \ldots, r-1$. To order elements within a coset, express the coset in the form $(p, q) + \langle(1,1)\rangle$, and follow the order imposed by $\langle(1,1)\rangle$. That is, order the elements (p, q), $(p+1, q+1)$, $(p+2, q+2)$, and so on.

Now consider a sequence of M consecutive elements starting with e_k, where $k \equiv 1 \bmod M$. The number of elements in each coset is s, which is divisible by M. Therefore, the M elements are contained in a single coset. Furthermore, if $e_k = (i, j)$, the row indices of the M elements may be listed as $i, i+1, \ldots, M, 1, 2, \ldots, i-1$. This shows that the elements come from different rows, and verifies property (1). Property (2) is established by a parallel argument, completing the proof. The extended order defined in Lemma 2 supports a general algorithm for scheduling the output of a bus-connected processor array.

THEOREM. *For any M by N array, a scheduling exists which makes full use of all busses at each step save possibly the last.*

Proof. Order the elements $e_1, e_2, \ldots, e_{MN}$ as in Lemma 2. At the kth operation cycle, $k = 1, 2, \ldots, \mathrm{INT}(MN/(M+N))$, transmit the M consecutive elements beginning with $e_{(k-1)M+1}$ to row busses, and the N consecutive elements beginning with $e_{(M-k)N+1}$ to column busses. This is simply sending the kth block of M entries to row busses, and the kth from the last block of N entries to column busses. Properties (1) and (2) assure that there is only one entry assigned to each bus. If any entries remain, they may be obtained in one additional cycle. Up to M of these entries are output to row busses, and the remainder go to column busses. This completes the proof.

As an example, consider a 4 by 6 array, as illustrated in FIGURE 3. The ordering begins with the entries of $\langle(1,1)\rangle$, labeled G_1 through G_{12} in the figure. Taking $(1,2)$ as the initial element of the remaining coset, the order continues through the entries labeled C_1 through C_{12}. On the first operation cycle, output G_1 through G_4 with row busses, and C_7 through C_{12} with column busses. On the next cycle, G_5 through G_8 transmit to row busses, and C_1 through C_6 transmit to column busses. The final four

G_1	C_1	G_9	C_9	G_5	C_5
C_6	G_2	C_2	G_{10}	C_{10}	G_6
G_7	C_7	G_3	C_3	G_{11}	C_{11}
C_{12}	G_8	C_8	G_4	C_4	G_{12}

FIGURE 3
Extended order for a 4 by 6 array.

entries may be output to row busses at the third cycle. These operations are illustrated in FIGURE 4. In each part of the figure, a horizontal arrow indicates that an entry will be sent to the row bus, a vertical arrow indicates a column bus, and a shaded cell denotes an entry output at a previous cycle.

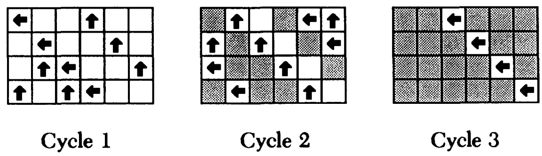


FIGURE 4
Three output cycles for a 4 by 6 array.

The scheduling algorithm has an interesting recursive formulation. Note that the order s of $(1,1)$ is the lcm of M and N, and so is easily computed. Thus, the number of cosets is $r = MN/s$, which is the GCD of M and N. We now show that the elements $(1,1)$, $(1,2),\dots,(1,r)$ are from different cosets, so that the first r elements of row 1 may be taken as the initial points for the ordering within the cosets. If two of these elements are from the same coset, the difference, $(0,k)$, is in $\langle(1,1)\rangle$, and $0<k<r$. Setting $(0,k)=(t,t)$, we have the conditions $t=aM$ and $t=bN+k$ for non-negative integers a and b. This, in turn, implies that $k=aM-bN$, and, hence, k is divisible by r. Since this last conclusion is impossible, the elements must come from distinct cosets, as asserted.

Using the entry $(1,k)$ as the initial element of the coset C_k, observe that we may obtain the first element of a coset by adding $(0,1)$ to the first entry of the preceding coset. Thus, for the forward ordering, the successor of an element is obtained by adding $(1,1)$, with an extra increment of $(0,1)$ as required to reach a new coset. A similar device works for the reverse ordering. Note that the last element of a coset is found by subtracting $(1,1)$ from the first element of the coset. In reverse order, the last elements of the cosets are therefore $(M,r-1)$, $(M,r-2),\dots,(M,1)$, and (M,N). To move from the last element of a coset to the last element of the preceding coset, simply *subtract* $(0,1)$. Accordingly, the reverse order is generated by starting at (M,N), and successively subtracting $(1,1)$, with an extra decrement of $(0,1)$ as needed to move to a new coset.

Because the recursive generation is essentially the same for both forward and reverse orders, a single algorithm can be used. A description of the algorithm appears in FIGURE 5. There are three parameters, STEP, SIDE-STEP, and FIRST. STEP is the normal increment, set to $(1,1)$ for the forward order, $(-1,-1)$ for the reverse. SIDE-STEP is used to jump between cosets, and is assigned to be $(0,1)$ for the forward order, $(0,-1)$ for the reverse. FIRST is simply the starting point for the order, and so is $(1,1)$ or (M,N). The only other variables in the algorithm are (i,j), the current array position in the order, and (i_0,j_0), which stores the first element of the current coset. When adding the increment STEP to an element results in the initial element of the coset containing that element, the SIDE-STEP value is used to reach the next coset.

The reader may find it interesting to consider some variations on this problem. For example, if the whole array can be ordered without repetition (i.e., if $Z_M \oplus Z_N$ is cyclic), then it is not necessary to read rows and columns from opposite ends of the ordering. Simply read the first N elements with column busses and the next M elements with row busses at the first stage, the next $N+M$ elements at the next

```
(i_0, j_0) ← FIRST
(i, j)   ← FIRST
DO NM times:
    OUTPUT (i, j)
    (i, j) ← (i, j) + STEP
    IF (i, j) = (i_0, j_0) DO:
            (i_0, j_0) ← (i_0, j_0) + SIDESTEP
            (i, j) ← (i_0, j_0)
             END DO
    END DO
```

FIGURE 5
Procedure to print matrix locations (i, j) in forward or reverse order.

stage, and so on. How might this approach be extended to the case of several cosets? Or, could the reading operations be scheduled to alternate between row and column busses for each set of $N+M$ elements? A different point of view occurs if the point of control is assumed to lie within the processor array element. Then, if the element knows its address (i, j) and the array dimensions M and N, how long should it wait before placing its contents on a bus, and should it use the row or column bus?

In conclusion, here are a few observations about how this problem might contribute to undergraduate education. The problem and its solution may be presented as an example in an abstract algebra class. It is credible as a real-world problem, requires no special background, and is clearly simplified by group theory. Alternatively, this example is well suited for use in a problem solving context. It is easily accessible to students, and a little experimentation will likely lead to some progress on the problem. In fact, for any fixed matrix size, ad hoc scheduling patterns should be easily constructed. The real difficulty is formalizing a general scheme, and verifying its validity. The machinery of group theory is quite useful as discussed above, but the problem can be handled without any mention of groups by making the equivalent arguments in terms of number theory. The problem might be given for recreation or enrichment in an abstract algebra or discrete mathematics course, with no hint that group theory should be used. In the same way, the problem might be of interest to a mathematics club or problem group.

The solution presented here should be considered an existence proof for the proposition that the scheduling problem possesses a slick mathematical solution. A comparison of some student solutions with one based on group theory should inspire appreciation of the utility and beauty of mathematical constructs. And if a student comes up with a different solution that is as simple, or simpler, all the better!

This work was supported under the Aerospace Sponsored Research program.

The American folksinger, Pete Seeger, composed a song whose chorus goes "But if two and two and fifty make a million, we'll see that day come round... ." The sentiment is more appealing than the mathematics. Of course he does not say that two and two and fifty make a million, but one may be tempted to try, with the help of such devices as scientific notation and the greatest integer function, to see whether one can write one million in terms of 2 and 2 and 50. Perhaps the readers would like to try.

MERRILL BARNEBEY
Kerrville, TX 78028

Some Interesting Fixed Points

JOHN MATHEWS
SOO TANG TAN
California State University
Fullerton, CA 92634

Introduction In this note we look at the solution of two problems drawn from the field of mathematics of finance. These problems are interesting because they provide us with real-world examples of fixed-point problems. (Recall that a fixed point x_0 of a function f is a point in the domain of f such that $f(x_0) = x_0$.) The second of the two examples to be considered also shows that, in practice, the "most obvious" way of writing an equation in the form $x = f(x)$ may not produce a convergent sequence; it is sometimes necessary to recast the equation in some other form in order to guarantee convergence of the sequence generated by the resulting iterative scheme. Finally, because the numerical computations are easily performed with a pocket calculator, a student can generate the required sequences readily and thus reinforce his understanding of the mathematical concepts involved.

Two problems of interest

Problem A. Suppose a borrower secures a conventional 30-year fixed-rate mortgage when purchasing a home. Besides paying interest for the loan (interest payments constitute a portion of the monthly installment) a borrower is charged a certain number of *points* for the loan. A point is the equivalent of 1 percent of the amount of the loan and must be paid by the borrower, in cash, when the loan closes. Currently, lenders are commanding between two and four points on conventional 30-year fixed-rate mortgage loans. Using points, lenders can effectively raise the price they charge borrowers for a home loan without actually increasing the mortgage rate. How does this practice of tacking on points affect the total price of a mortgage? More specifically, what is the effective (true) rate of interest paid for the loan?

Let us formulate the problem mathematically. To this end, let

A = amount of the loan
n = life of the loan in years
r = nominal annual interest rate
P = points charged for the loan
k = amount of each (equal) monthly installment

An expression for k is normally obtained by treating the present amortization problem as a special case of the annuity problem (see [1]). Equivalently, we can derive it by observing that the amount of money due the finance company at the end of year n is

$$A\left(1+\frac{r}{12}\right)^{12n} - k\left[1+\left(1+\frac{r}{12}\right)+\cdots+\left(1+\frac{r}{12}\right)^{12n-1}\right]. \tag{1}$$

Setting this expression equal to zero (we wish to amortize the loan by the end of year n) and using the formula for summing a geometric series yields

$$k = \frac{Ar}{12\left[1 - \left(1 + \frac{r}{12}\right)^{-12n}\right]}. \tag{2}$$

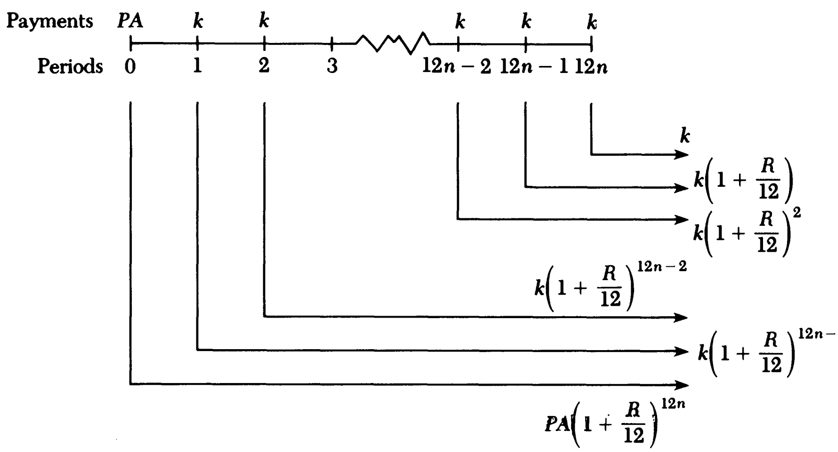


FIGURE 1

Next, let R denote the nominal rate of interest per annum corresponding to r, convertible monthly, for a loan with a surcharge of P points. Then, we have the following relationship

$$PA\left(1 + \frac{R}{12}\right)^{12n} + k\left[\left(1 + \frac{R}{12}\right)^{12n-1} + \left(1 + \frac{R}{12}\right)^{12n-2} + \cdots + 1\right] = A\left(1 + \frac{R}{12}\right)^{12n},$$

where the expression on the left gives the total amount the financier would have finished with if he had reinvested each repayment collected over the remaining term of the loan (see FIGURE 1).

Using the formula for summing a geometric series, and rearranging the terms, we obtain

$$R = f(R) \tag{3}$$

where

$$f(R) = \frac{12k}{A(1-P)}\left[1 - \left(1 + \frac{R}{12}\right)^{-12n}\right]. \tag{4}$$

Once R has been determined, the effective (or true) rate for the loan may be found from the equation

$$1 + \rho = \left(1 + \frac{R}{12}\right)^{12} \tag{5}$$

where ρ denotes the effective rate of interest. (Recall that the effective rate of interest is the rate which when compounded annually yields the same accumulated amount as the nominal rate compounded monthly.)

Problem B. Recall that an *annuity* is a sequence of payments made at regular time intervals. Now, consider an annuity which is *ordinary*, *simple*, and *certain*. Such an annuity fulfills the following conditions:

1. The payments are made at the end of each payment period.
2. The payment period coincides with the interest conversion period.
3. The term (the time period in which payments are made) is fixed.

We will also assume that the annuity payments are equal.

To find the amount A of such an annuity, let

$$P = \text{size of each payment in the annuity}$$
$$r = \text{interest rate per period}$$
$$n = \text{number of payment periods}.$$

Then the amount of the annuity is given by

$$A = P + P(1+r) + P(1+r)^2 + \cdots + P(1+r)^{n-1} \tag{6}$$

(see [1]) or, equivalently,

$$A = \frac{P}{r}\left[(1+r)^n - 1\right]. \tag{7}$$

An elementary exercise involving the use of formula (7) is one of finding the value of A given the values of P, r, and n. Now, suppose instead that we want to find the interest rate r given the values of A, P, and n. In this instance there is no simple way of solving the problem.

Fixed point solutions Thanks to equation (3), a rather obvious approach to the solution of Problem A is to find the fixed point of the function f defined by equation (4). Similarly, observe that equation (7) is readily put in the form

$$r = g(r), \tag{8}$$

where

$$g(r) = \frac{P}{A}\left[(1+r)^n - 1\right] \tag{9}$$

so that it appears too that Problem B can be solved by finding the fixed point of the function g defined by equation (9).

Iterative solutions

Solution of Problem A. The function $f(R)$ has the following properties:

(a) $f(0) = 0$
(b) $f(R)$ is monotonically increasing on $[0, \infty)$
(c) $f(R)$ is bounded above
(d) $f(R)$ is strictly concave downwards on the interval $[0, \infty)$
(e) $f'(0) > 1$.

In fact, (a) and (c) follow immediately from (4), and (b) follows from considering the derivative of $f(R)$,

$$f'(R) = \frac{12nk}{A(1-P)}\left(1 + \frac{R}{12}\right)^{-12n-1}. \tag{10}$$

Also, (d) follows from observing that

$$f''(R) = -\frac{nk(12n+1)}{A(1-P)}\left(1+\frac{R}{12}\right)^{-12n-2} < 0$$

for all $R \in [0,\infty)$. To prove (e), first note that

$$12n\left(1+\frac{r}{12}\right)^{12n-1} > 1+\left(1+\frac{r}{12}\right)+\left(1+\frac{r}{12}\right)^2+\cdots+\left(1+\frac{r}{12}\right)^{12n-1}$$
$$= \frac{\left(1+\frac{r}{12}\right)^{12n}-1}{\frac{r}{12}},$$

from which we see that

$$\left(1+\frac{r}{12}\right)^{12n}-1 < nr\left(1+\frac{r}{12}\right)^{12n-1}. \tag{11}$$

Then using equations (10), (2), and (11), we find

$$\begin{aligned} f'(0) &= \frac{12nk}{A(1-P)} \\ &= \frac{nr}{(1-P)\left[1-\left(1+\frac{r}{12}\right)^{-12n}\right]} \\ &= \frac{nr\left(1+\frac{r}{12}\right)^{12n}}{(1-P)\left[\left(1+\frac{r}{12}\right)^{12n}-1\right]} \\ &> \frac{\left(1+\frac{r}{12}\right)^{12n}}{(1-P)\left(1+\frac{r}{12}\right)^{12n-1}} \\ &> 1+\frac{r}{12} \\ &> 1 \end{aligned}$$

as was to be shown.

The above observations and computations show that for points sufficiently close to the origin, at least, the graph of $y = f(R)$ lies above the graph of $y = R$. But then the boundedness of $y = f(R)$ implies that these two curves intersect at at least one point. That there is in fact at most one solution (fixed point), R^*, to equation (3) follows from the concavity of the graph defined by $y = f(x)$ in $[0,\infty)$. This analysis suggests an iteration of the form

$$R_{i+1} = f(R_i) \qquad (i = 0,1,2,\ldots) \tag{12}$$

with a starting point (initial guess) $R_0 > 0$. The convergence of the iteration (12), independent of the choice of R_0 (as long as it is positive), is assured by the monotonicity and concavity of $y = f(R)$ in the interval $[0,\infty)$ (see fig. 2).

Solution of Problem B. Equation (8) suggests that we might attempt to find the solution to the problem at hand—the fixed point of the function g—by iterating

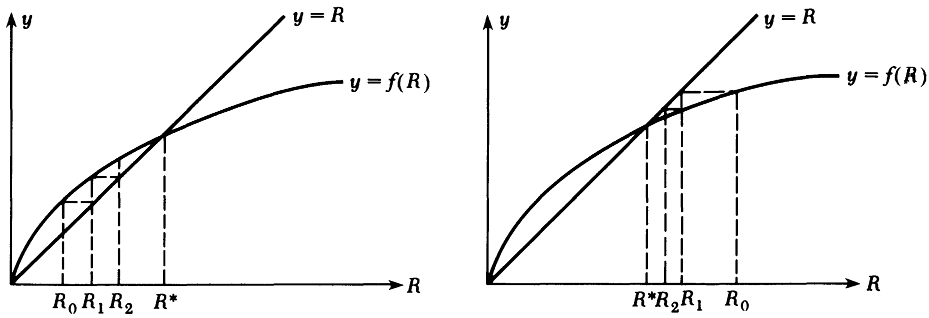


FIGURE 2
Convergence of the Fixed Point Iteration.

$$r_{i+1} = g(r_i) \qquad (i = 0, 1, 2, \ldots) \tag{13}$$

with some suitable initial guess r_0. Unfortunately, the fixed-point iteration defined by equation (13) using $g(r)$ of equation (9) fails to converge for any initial guess r_0 which is not equal to the fixed point r^* of g.

To see this, we compute

$$g'(r) = \frac{nP}{A}(1+r)^{n-1}. \tag{14}$$

Next, observe that

$$(1+r)^k < (1+r)^{n-1} \qquad (k = 0, 1, 2, \ldots, n-2) \tag{15}$$

so that equation (6) implies

$$A < nP(1+r)^{n-1}.$$

Now, using this inequality, we see that equation (14) implies

$$\begin{aligned} g'(r^*) &= \frac{nP}{A}(1+r^*)^{n-1} \\ &> 1 \end{aligned}$$

and we invoke the following theorem (see [2]) to conclude that the fixed-point iteration (13) fails to converge to r^* as asserted.

THEOREM 1. *Let g and g' be continuous on an interval (a, b) that contains the fixed point P of g.*

(a) *If $|g'(P)| < 1$ and the starting value P_0 is chosen sufficiently close to P, then the fixed point iteration $P_{k+1} = g(P_k)$ converges to P.*

(b) *If $|g'(P)| > 1$, then the iteration $P_{k+1} = g(P_k)$ does not converge to P unless the starting value P_0 is taken to be P itself.*

Let us rewrite equation (7) in a slightly different manner. Straightforward algebraic manipulations yield

$$r = G(r) = \left(1 + \frac{Ar}{P}\right)^{1/n} - 1. \tag{16}$$

Observe that the functions g and G are inverses. Thus, if r^* is a solution of equation (8), and therefore a solution of $r = G(r)$ as well, then we have

$$G'(r^*) = \frac{1}{g'(r^*)}. \tag{17}$$

Since $g'(r^*) > 1$ as demonstrated earlier, equation (17) implies that $G'(r^*) < 1$. This tells us that the fixed point iteration (see Theorem 1)

$$r_{i+1} = G(r_i) \qquad (i = 0, 1, 2, \dots) \tag{18}$$

will converge to r^* provided the initial guess r_0 is sufficiently close to r^*.

We now show that, in fact, the fixed-point iteration (18) converges for any initial guess $r_0 > 0$. To see this, observe that the function G has the following properties:

(a) $G(0) = 0$
(b) $G(r)$ is monotonically increasing on $[0, \infty)$
(c) $G(r)$ is bounded above
(d) $G(r)$ is strictly concave downwards on $[0, \infty)$
(e) $G'(0) = A/nP > 1$.

As a consequence of this, a sequence $\{r_i\}$ generated by the iteration (18) with $0 < r_0 < r^*$ will be monotonically increasing and have limit r^*, whereas a sequence $\{\tilde{r}_i\}$ generated by (18) with $r_0 > r^*$ will be monotonically decreasing and converge to r^*. Geometrically, this situation is similar to the case encountered in the solution of Problem A (see FIGURE 2).

Numerical examples

Example of Problem A. Suppose you secure a mortgage in the amount of \$80,000 to be amortized over 30 years with equal monthly installments. The interest rate is 10 percent per annum convertible monthly and you are charged 2 points for the loan. Determine the true rate of interest you are paying for the mortgage.

Solution. Here $A = 80{,}000$, $n = 30$, $r = 0.10$, and $P = 0.02$. Using equation (2), we find $k = 702.0573$. Using these values, the iteration (3) becomes

$$R_{i+1} = 0.10746\left[1 - \left(1 + \frac{R_i}{12}\right)^{-360}\right].$$

With the initial guess $R_0 = 0.12$, we obtain the sequence

$$R_1 = 0.10447, \qquad R_2 = 0.10272, \qquad R_3 = 0.10246,$$

$$R_4 = 0.10243, \qquad R_5 = 0.10242, \qquad R_6 = 0.10242.$$

Thus, we may take $R = R^* = 0.1024$. Using equation (5), we compute the true rate of interest to be 10.74 percent per annum.

Example of Problem B. Suppose a deposit of \$200 is made at the end of every month into a tax-sheltered time-deposit bank account paying interest compounded monthly. At what rate of interest should the money be invested if an accumulated amount of \$200,000 is to be realized at the end of 25 years?

Here $A = 200{,}000$, $P = 200$, and $n = (12)(25) = 300$. Using the iteration

$$r_{k+1} = \left(1 + \frac{Ar_k}{P}\right)^{1/n} - 1$$
$$= (1 + 1000r_k)^{1/300} - 1,$$

with initial guess $r_0 = 0.006$, we obtain the sequence

$r_1 = 0.00650745$, $r_2 = 0.00674228$, $r_3 = 0.00684565$, $r_4 = 0.00689016$,

$r_5 = 0.00690915$, $r_6 = 0.00691721$, $r_7 = 0.00692063$, $r_8 = 0.00692208$,

$r_9 = 0.00692270$, $r_{10} = 0.00692296$, $r_{11} = 0.00692307$, $r_{12} = 0.00692312$,

$r_{13} = 0.00692314$, $r_{14} = 0.00692315$, $r_{15} = 0.00692315$.

We conclude that the money should be invested at a rate of (12)(0.00692315) or 8.31 percent per annum.

Finally, we remark that if one is only interested in solving the problem at hand, then there are many other approaches. For example, we can solve equation (7) by writing it as a polynomial equation of degree n. However, since n is normally quite large, this is certainly not a desirable method. The Bisection Method of Bolzano will do the job even if it does so rather inefficiently. Newton's Method works very well provided we choose the initial guess $r_0 > r^*$ carefully, lest the sequence generated by the Newton iteration converge to 0 which is another root of both equations (3) and (8).

The authors wish to thank the referees for their useful comments and suggestions.

REFERENCES

1. Soo Tang Tan, *Applied Finite Mathematics*, 2nd ed., PWS-Kent Publishers, Boston, MA, 1987.
2. John H. Mathews, *Numerical Methods for Computer Science, Engineering, and Mathematics*, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1987.

Groups Formed From 2 × 2 Matrices Over Z_p

GREGOR OLŠAVSKÝ
Pennsylvania State University
Behrend College
Erie, PA 16563

It is well known that the set of all 2×2 matrices with real entries forms a group of infinite order under addition and the invertible ones form a group under multiplication. However, if the entries come from a finite field, say Z_p (where p is prime), the additive and multiplicative groups thus formed are necessarily of finite order and merit detailed study. My objective is to exhibit some nontrivial examples of finite, nonabelian groups and to investigate some of the more interesting subgroups. Also, I'd like to find the order of each subgroup and show how to construct a simple group (one having no proper normal subgroups except the trivial subgroup).

First, I will discuss the additive case. Let's denote the set of all 2×2 matrices with entries from Z_p by $OL(2, Z_p)$ (ordinary linear group of 2×2 matrices over Z_p). Let's find the order of $OL(2, Z_p)$, where $|OL(2, Z_p)|$ denotes the order of this group.

Now any matrix $M \in OL(2, Z_p)$ is of the form $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$, where $a, b, c, d \in Z_p$. There are p choices for a, b, c, and d and thus p^4 possible matrices. Therefore, we have constructed a group of order p^4. It is abelian, since addition in Z_p is abelian.

Let's investigate some subgroups of $OL(2, Z_p)$. $OL(2, Z_p)$ has proper subgroups of

$r_1 = 0.00650745$, $r_2 = 0.00674228$, $r_3 = 0.00684565$, $r_4 = 0.00689016$,

$r_5 = 0.00690915$, $r_6 = 0.00691721$, $r_7 = 0.00692063$, $r_8 = 0.00692208$,

$r_9 = 0.00692270$, $r_{10} = 0.00692296$, $r_{11} = 0.00692307$, $r_{12} = 0.00692312$,

$r_{13} = 0.00692314$, $r_{14} = 0.00692315$, $r_{15} = 0.00692315$.

We conclude that the money should be invested at a rate of (12)(0.00692315) or 8.31 percent per annum.

Finally, we remark that if one is only interested in solving the problem at hand, then there are many other approaches. For example, we can solve equation (7) by writing it as a polynomial equation of degree n. However, since n is normally quite large, this is certainly not a desirable method. The Bisection Method of Bolzano will do the job even if it does so rather inefficiently. Newton's Method works very well provided we choose the initial guess $r_0 > r^*$ carefully, lest the sequence generated by the Newton iteration converge to 0 which is another root of both equations (3) and (8).

The authors wish to thank the referees for their useful comments and suggestions.

REFERENCES

1. Soo Tang Tan, *Applied Finite Mathematics*, 2nd ed., PWS-Kent Publishers, Boston, MA, 1987.
2. John H. Mathews, *Numerical Methods for Computer Science, Engineering, and Mathematics*, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1987.

Groups Formed From 2 × 2 Matrices Over Z_p

GREGOR OLŠAVSKÝ
Pennsylvania State University
Behrend College
Erie, PA 16563

It is well known that the set of all 2×2 matrices with real entries forms a group of infinite order under addition and the invertible ones form a group under multiplication. However, if the entries come from a finite field, say Z_p (where p is prime), the additive and multiplicative groups thus formed are necessarily of finite order and merit detailed study. My objective is to exhibit some nontrivial examples of finite, nonabelian groups and to investigate some of the more interesting subgroups. Also, I'd like to find the order of each subgroup and show how to construct a simple group (one having no proper normal subgroups except the trivial subgroup).

First, I will discuss the additive case. Let's denote the set of all 2×2 matrices with entries from Z_p by $OL(2, Z_p)$ (ordinary linear group of 2×2 matrices over Z_p). Let's find the order of $OL(2, Z_p)$, where $|OL(2, Z_p)|$ denotes the order of this group.

Now any matrix $M \in OL(2, Z_p)$ is of the form $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$, where $a, b, c, d \in Z_p$. There are p choices for a, b, c, and d and thus p^4 possible matrices. Therefore, we have constructed a group of order p^4. It is abelian, since addition in Z_p is abelian.

Let's investigate some subgroups of $OL(2, Z_p)$. $OL(2, Z_p)$ has proper subgroups of order p^k for positive integers k such that $0 \leqq k < 4$. Therefore, it has 4 distinct subgroups of the form

$$\begin{cases} OL_3(2, Z_p) = \left\{ \begin{bmatrix} a & b \\ c & 0 \end{bmatrix} : a, b, c \in Z_p \right\} \cong Z_p \times Z_p \times Z_p \\ OL_2(2, Z_p) = \left\{ \begin{bmatrix} a & b \\ 0 & 0 \end{bmatrix} : a, b \in Z_p \right\} \cong Z_p \times Z_p \\ OL_1(2, Z_p) = \left\{ \begin{bmatrix} a & 0 \\ 0 & 0 \end{bmatrix} : a \in Z_p \right\} \cong Z_p \\ OL_0(2, Z_p) = \left\{ \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \right\} \cong \text{the trivial group } \{0\}. \end{cases}$$

There are 4 types of $OL_3(2, Z_p)$ subgroups (by permutation of the zero element) and there are 6 types of $OL_2(2, Z_p)$ subgroups. These are, of course, basically the same subgroup, up to an isomorphism. In particular, $OL_3(2, Z_p)$ is called a Sylow p-subgroup of $OL(2, Z_p)$ since $|OL_3(2, Z_p)| = p^3$ and this is the highest proper prime-power divisor of $|OL(2, Z_p)|$. Since $OL(2, Z_p)$ is abelian, all its subgroups are abelian.

The general linear group $GL(2, Z_p)$ is the group of all nonsingular matrices in $OL(2, Z_p)$ under matrix multiplication. To find out which matrices from $OL(2, Z_p)$ are also in $GL(2, Z_p)$, we can use the determinant function to discard any $M \in OL(2, Z_p)$ such that $\det(M) = 0$. What is the order of $GL(2, Z_p)$?

Let $\alpha \in GL(2, Z_2)$. Then $\alpha = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, where $a, b, c, d \in Z_p$ such that $\det(\alpha) \neq 0$. What elements may be in the first row of a nonsingular matrix? Clearly any elements may be in row 1 except that both a and b cannot be 0. Thus there are $p^2 - 1$ possibilities. Once row 1 is determined, row 2 can contain any elements except multiples of row 1. Therefore, there are $p^2 - p$ possibilities for row 2. Hence $|GL(2, Z_p)| =$

$$(p^2 - 1)(p^2 - p) = p(p-1)^2(p+1).$$

The smallest such group, $GL(2, Z_2)$, has order $(2^2 - 1)(2^2 - 2) = 6$. The six elements of $GL(2, Z_2)$ are

$$\varepsilon = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \alpha = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}, \quad \beta = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix},$$
$$\gamma = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \quad \theta = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}, \quad \rho = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

$GL(2, Z_2)$ is nonabelian and is isomorphic to S_3.

$GL(2, Z_p)$ has many interesting subgroups. The subgroup of all matrices of the form $\begin{bmatrix} 1 & a \\ 0 & 1 \end{bmatrix}$ is an abelian subgroup isomorphic to Z_p, called the unitriangular group. It is denoted by $UL(2, Z_p)$. The order of $UL(2, Z_p)$ is p and it is a Sylow p-subgroup of $GL(2, Z_p)$.

A second subgroup of $GL(2, Z_p)$ consists of the set of all diagonal matrices of the form $\begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix}$, denoted by $DL(2, Z_p)$. Since there are $p - 1$ choices for a and b, the order of this subgroup is $(p-1)^2$. If $p = 2$, $DL(2, Z_2)$ is the trivial subgroup. However, for $p \geq 3$, $DL(2, Z_2)$ is nontrivial. For example, $|DL(2, Z_3)| = (3-1)^2 = 4$ and is, therefore, abelian. The center of $GL(2, Z_p)$ happens to be a subgroup of $DL(2, Z_p)$ of all matrices of the form

$$\begin{bmatrix} a & 0 \\ 0 & a \end{bmatrix} = a\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = aI_2.$$

These are called scalar matrices and commute with any $M \in GL(2, Z_p)$. The order of $C(GL(2, Z_p))$ is $p-1$ and $C(GL(2, Z_p)) \approx Z_p^*$ (all non-zero elements in Z_p).

If $p \geqq 3, GL(2, Z_p)$ has a fourth subgroup formed from all matrices of the form $\begin{bmatrix} \pm 1 & a \\ 0 & 1 \end{bmatrix}$, denoted $STL(2, Z_p)$, the special triangular group. For each $a \in Z_p$, there are 2 possible matrices; thus $|STL(2, Z_p)| = 2p$. Since $STL(2, Z_p)$ is not cyclic, it is isomorphic to D_p, the dihedral group (group of symmetries of a regular p-gon). For example, $STL(2, Z_3) \approx D_3$.

For any $p \geqq 2, GL(2, Z_p)$ has a subgroup of all triangular matrices of the form

$$\begin{bmatrix} a & b \\ 0 & c \end{bmatrix} \left(\text{or} \begin{bmatrix} a & 0 \\ b & c \end{bmatrix}\right).$$

Since there are $p-1$ choices for a and c and p choices for b, the order of this group $TL(2, Z_p)$ is $p(p-1)^2$. For example, if $p = 3$, $|TL(2, Z_p)| = 12$.

The largest subgroup of $GL(2, Z_p)$ is the subgroup denoted $SL^*(2, Z_p)$, where $SL^*(2, Z_p) = \{M\colon \det(M) = \pm 1\}$. If $N \in GL(2, Z_p)$ and A is in the coset $N \cdot SL^*(2, Z_p)$, then $\det(A) = \det(NM)$, where $M \in SL^*(2, Z_p)$. But $\det(NM) = \det(N) \cdot \det(M)$ which is $\pm \det N$.

Since there are only $p-1$ possibilities for $\det(A)$ there are $(p-1)/2$ distinct cosets of $SL^*(2, Z_p)$ in $GL(2, Z_p)$; if $p = 2, SL^*(2, Z_2) \cong GL(2, Z_2)$ since $M \in GL(2, Z_2)$ implies $\det M = 1$. So $|SL^*(2, Z_p)| = p(p-1)^2(p+1)/((p-1)/2) = 2p(p^2-1)$ if $p \geqq 3$. Note that $SL^*(2, Z_p)$ is a proper subgroup only if $p \geqq 5$. For example, $|SL^*(2, Z_5)| = 240$.

If $p \geqq 3, GL(2, Z_p)$ has a proper subgroup of all matrices M such that $\det(M) = 1$. Such a matrix is called unimodular and this subgroup is called the special linear group, denoted $SL(2, Z_p)$.

To determine the structure and order of $SL(2, Z_p)$, consider the homomorphism $D | GL(2, Z_p) \to Z_p^*$ such that $D(M) = \det(M)$. The kernel of D is given by $\ker(D) = \{M\colon \det(M) = 1\}$. Therefore, $\ker(D) = SL(2, Z_p)$ by definition. The first isomorphism theorem says $\ker(D)$ is a normal subgroup of $GL(2, Z_p)$ and the factor group $GL(2, Z_p)/\ker(D) \approx \text{image}(D)$.

But $\text{image}(D) = Z_p^*$ so $GL(2, Z_p)/\ker(D) \approx Z_p^*$. Therefore,

$$\left| \frac{GL(2, Z_p)}{\ker(D)} \right| = |Z_p^*| = p - 1, \text{ so } \left| \frac{GL(2, Z_p)}{SL(2, Z_p)} \right| = p - 1$$

implies

$$|SL(2, Z_p)| = \frac{|GL(2, Z_p)|}{p-1} = p(p^2 - 1).$$

The importance of $SL(2, Z_p)$ is tied to the idea of a simple group. The simple group is the single most important object of study in finite group theory and can be compared to the importance of prime numbers in number theory. I'd like to show how to construct a simple group from $GL(2, Z_p)$ for $p > 3$.

First, let's find the center C_0 of $SL(2, Z_p)$. Clearly $C_0 = \{M\colon MN = NM \text{ for all } N \in SL(2, Z_p)\}$. M must be a scalar matrix of the form $\begin{bmatrix} k & 0 \\ 0 & k \end{bmatrix} = kI_2$ such that $\det(M) = 1$. This implies $k^2 \equiv 1 (\text{mod } p)$ which means $k = 1$ or -1 (which is equivalent to $p - 1 (\text{mod } p)$). Therefore $|C_0| = 2$. Since C_0 is a normal subgroup of $SL(2, Z_p)$, we can form the factor group $SL(2, Z_p)/C_0$, which has order $[SL(2, Z_p), C_0] = |SL(2, Z_p)|/2 = p(p^2-1)/2 = pq$ where $q = (p^2-1)/2$. The group $SL(2, Z_p)/C_0$ is called the projective special linear group, denoted by $PSL(2, Z_p)$. It is *always* simple for $p > 3$; however, the proof is not easy (see [**5**, p. 162]). The smallest such group is $PSL(2, Z_5)$, which has order 60 and is isomorphic to A_5. If $p = 7$ however, we get a new simple group of order 168, which is not isomorphic to A_n for any n (compare orders).

Here are some of the groups we've discussed along with their orders:

Group	$p=2$	$p=3$	$p=5$	$p=7$
$GL(2, Z_p)$	6	48	480	2016
$SL^*(2, Z_p)$	6	48	240	672
$SL(2, Z_p)$	6	24	120	336
$TL(2, Z_p)$	2	12	80	252
$DL(2, Z_p)$	1	4	16	36
$STL(2, Z_p)$	2	6	10	14
$UL(2, Z_p)$	2	3	5	7
$C(GL(2, Z_p))$	1	2	4	6
$PSL(2, Z_p)$	6	12	60	168

As a concluding remark, one may wonder about $GL(n, Z_p)$ for $n \geqq 3$. In general, $GL(n, Z_p)$ has all the subgroups mentioned previously and $PSL(n, Z_p)$ is simple (if $n \geqq 3$). In addition, the following relationships among the subgroups are always true:

$$SL(n, Z_p) \leqq GL(n, Z_p)$$

$$DL(n, Z_p) < TL(n, Z_p)$$

$$C(GL(n, Z_p)) \leqq DL(n, Z_p) \leqq TL(n, Z_p) < GL(n, Z_p)$$

where $H \leqq G$ indicates H is a subgroup of G and $<$ indicates a proper subgroup.

For an elementary treatment of group theory, an excellent source would be Gallian [**2**]. The other references listed are more advanced. Aschbacher's book [**3**] is the most complete and up-to-date.

REFERENCES

1. Garrett Birkhoff and Saunders Mac Lane, *Algebra*, Macmillan, 1979.
2. Joseph Gallian, *Contemporary Abstract Algebra*, Heath, 1986.
3. Michael Aschbacher, *Finite Group Theory*, Cambridge Univ. Press, 1986.
4. M. I. Kargapolov and Ju. I. Merzljakov, *Fundamentals of the Theory of Groups*, Springer-Verlag, 1979.
5. Joseph J. Rotman, *The Theory of Groups*, Allyn and Bacon, 1978.

PROBLEMS

LOREN C. LARSON, *editor*
St. Olaf College

GEORGE GILBERT, *associate editor*
Texas Christian University

Proposals

To be considered for publication, solutions should be received by March 1, 1991.

1353. *Proposed by Mihály Bencze, Braşov, Romania.*

Prove that

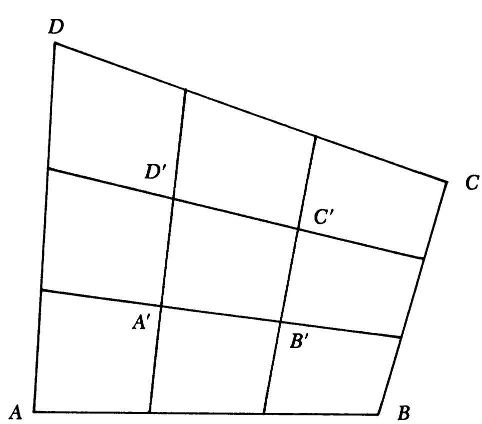
$$(\sqrt{2}-1)\left(\sqrt[3]{6}-\sqrt{2}\right)\ldots\left(\sqrt[n+1]{(n+1)!}-\sqrt[n]{n!}\right)<\frac{n!}{(n+1)^n}.$$

1354. *Proposed by Frank Schmidt, Bryn Mawr College, Bryn Mawr, Pennsylvania, and Rodica Simion, George Washington University, Washington, D.C.*

Let $ABCD$ be a convex quadrilateral in the plane with trisection points joined as in the figure to form nine smaller quadrilaterals.

a. Show that the area of $A'B'C'D'$ is one-ninth the area of $ABCD$.

b. Give necessary and sufficient conditions so that all nine quadrilaterals have equal area.



ASSISTANT EDITORS: CLIFTON CORZATT and THEODORE VESSEY, *St. Olaf College. We invite readers to submit problems believed to be new and appealing to students and teachers of advanced undergraduate mathematics. Proposals should be accompanied by solutions, if at all possible, and by any other information that will assist the editors and referees. A problem submitted as a Quickie should have an unexpected, succinct solution. An asterisk (*) next to a problem number indicates that neither the proposer nor the editors supplied a solution.*

Solutions should be written in a style appropriate for Mathematics Magazine. Each solution should begin on a separate sheet containing the solver's name and full address.

Solutions and new proposals should be mailed in duplicate to Loren Larson, Department of Mathematics, St. Olaf College, Northfield, MN 55057.

1355. *Proposed by Murray S. Klamkin, University of Alberta, Edmonton, Alberta, Canada.*

Determine the extreme values of

$$F \equiv x_1 x_2 \cdots x_n - (x_1 + x_2 + \cdots + x_n),$$

where $b \geqslant x_i \geqslant a \geqslant 0$ for all i.

1356. *Proposed by Hüseyin Demir, Middle East Technical University, Ankara, Turkey.*

Let P, Q be points taken on the side BC of a triangle ABC, in the order B, P, Q, C. Let the circumcircles of PAB, QAC intersect at M ($\neq A$) and those of PAC, QAB at N. Show that A, M, N are collinear if and only if P and Q are symmetric in the midpoint A' of BC.

1357. *Proposed by George Gilbert, St. Olaf College, Northfield, Minnesota.*

Let $\mathbf{A}_n = (a_{ij})$ be the $n \times n$ (band) matrix with

$$a_{ij} = \begin{cases} 1, & \text{if } |i-j| \leqslant 1; \\ 0, & \text{otherwise.} \end{cases}$$

Define $\mathbf{B}_n$ just as $\mathbf{A}_n$ except that $b_{nn} = 2$, and $\mathbf{C}_n$ just as $\mathbf{A}_n$ except $c_{n,n-1} = 2$.

a. Prove that the characteristic polynomial of $\mathbf{B}_n$ divides that of $\mathbf{A}_{2n}$.
b. Prove that the characteristic polynomial of $\mathbf{C}_n$ divides that of $\mathbf{A}_{2n-1}$.
c. Prove that $\mathbf{C}_n$ is diagonalizable.

Quickies

Answers to the Quickies are on pages 279–80.

Q766. *Proposed by Jeffrey Shallit, Dartmouth College, Hanover, New Hampshire.*

Let $a_1, a_2, \ldots, a_k$ be distinct real numbers, $k \geqslant 1$. Show that the k functions

$$|x - a_1|, \quad |x - a_2|, \ \ldots, |x - a_k|$$

are linearly independent.

Q767. *Proposed by V. F. Lev, Moscow, USSR.*

Consider a sequence of positive integers $a_0, a_1, a_2, \ldots$ where each term is equal to the number of divisors of the previous one. Find all a_0 such that the sequence does not contain any squares.

Q768. *Proposed by Murray S. Klamkin, University of Alberta, Edmonton, Alberta, Canada.*

Show that

$$(a+b+c)^n (a^{2n} + b^{2n} + c^{2n}) \geqslant (a^n + b^n + c^n)(a^2 + b^2 + c^2)^n,$$

where $a, b, c \geqslant 0$ and $n \geqslant 1$.

Solutions

Diagonals of Exscribed Quadrangles **October 1989**

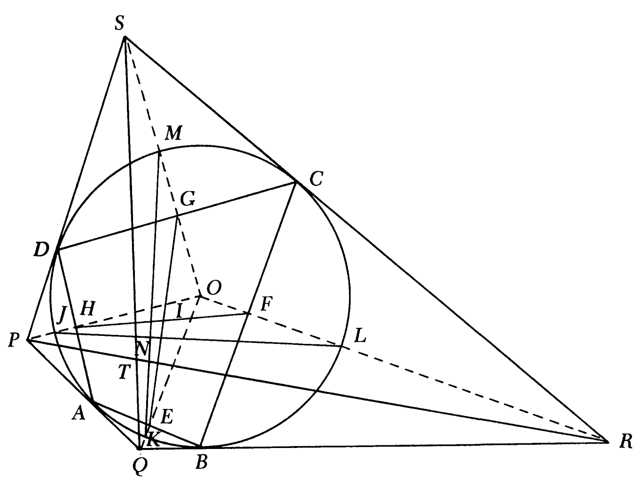
1327. *Proposed by Hüseyin Demir, Middle East Technical University, Ankara, Turkey.*

Let the sides PQ, QR, RS, SP of a convex quadrangle $PQRS$ touch an inscribed circle at A, B, C, D and let the midpoints of the sides AB, BC, CD, DA be E, F, G, H. Show that the angle between the diagonals PR, QS is equal to the angle between the bimedians EG, FH.

I. *Solution by Jordi Dou, Barcelona, Spain.*

Let the inscribed circle have radius r and center O. Let J, K, L, M be the intersection points of the circle with the lines OHP, OEQ, OFR, OGS respectively.

Let $N = JL \cap KM$, $T = PR \cap QS$, $I = EG \cap FH$. Note that JL and KM are perpendicular [because the arcs JAK and LCM together comprise half the perimeter of the circle], so [since $\triangle JOL$ is isosceles] KM is parallel to the angle bisector of $\angle JOL$. Also, note that the lines PR, HF are antiparallel with respect to the sides of $\angle JOL$ [that is, $\angle OHF = \angle ORP$ and $\angle OFH = \angle OPR$], because $OH \cdot OP = OF \cdot OR = r^2$, and so $\triangle OHF$, $\triangle ORP$ are similar, [since $OH/OF = OR/OP$]. This [together with the fact that $\triangle JOL$ is isosceles] implies that the lines PR, HF form equal angles (say α) with JL. Similarly, the lines QS, EG form equal angles (say β) with KM. We then have $\angle GIF = 90° - (\alpha + \beta)$ while $\angle RTS = 90° + (\alpha + \beta)$, and we are done.



II. *Solution by Jiro Fukuta, Motosu-gun, Gifu-ken, Japan.*

Let O be the center of the inscribed circle of the quadrangle $PQRS$ and r be the length of the radius. Let P, Q, R, S be denoted by the complex numbers $\alpha, \beta, \gamma, \delta$, respectively, on the complex plane with the origin at O. Then E, F, G, H correspond to $r^2/\bar{\beta}, r^2/\bar{\gamma}, r^2/\bar{\delta}, r^2/\bar{\alpha}$, respectively.

To obtain the conclusion, it is sufficient to prove that

$$F \equiv \left(\frac{\alpha - \gamma}{\beta - \delta}\right) \div \left(\frac{r^2/\bar{\beta} - r^2/\bar{\delta}}{r^2/\bar{\alpha} - r^2/\bar{\gamma}}\right)$$

is real. We have

$$F=\frac{\alpha-\gamma}{\beta-\delta}\cdot\frac{\dfrac{r^2\bar{\gamma}-r^2\bar{\alpha}}{\bar{\alpha}\bar{\gamma}}}{\dfrac{r^2\bar{\delta}-r^2\bar{\beta}}{\bar{\beta}\bar{\delta}}}$$

$$=\frac{\alpha-\gamma}{\beta-\delta}\cdot\frac{\bar{\alpha}-\bar{\gamma}}{\bar{\beta}-\bar{\delta}}\cdot\frac{\overline{\beta\delta}}{\overline{\alpha\gamma}}$$

$$=\frac{|\alpha-\gamma|^2}{|\beta-\delta|^2}\overline{\left(\frac{\delta}{\alpha}\cdot\frac{\beta}{\gamma}\right)}.$$

But $(\delta/\alpha)(\beta/\gamma)$ is real, because $\arg(\delta/\alpha)+\arg(\beta/\gamma)=\arg POS+\arg ROQ=\pi$. This completes the proof.

Also solved by Duane M. Broline, Timothy V. Craine, John F. Goehl, Jr., Francis M. Henderson, Paul Martin, and the proposer.

Sums of Composite Odd Numbers **October 1989**

1328. *Proposed by Ronald E. Ruemmler, Middlesex County College, Edison, New Jersey.*

Find the largest even integer which cannot be expressed as the sum of two composite odd numbers.

Solution by Garrett R. Vargas, student, Stanford University, Stanford, California.

Let n be an even number. Then each of the following expresses n as the sum of two odd numbers:

$$n=(n-15)+15,$$
$$n=(n-25)+25,$$
$$n=(n-35)+35.$$

Note that at least one of $n-15, n-25, n-35$ must be divisible by three, and hence n can be expressed as the sum of two composite odd numbers if $n>38$. Indeed, it is easily verified that 38 cannot be expressed as the sum of two composite odd numbers.

Also solved by University of Arizona Undergraduate Problem Solving Lab, F. Balogh (Hungary), Merrill Barnebey, Krystal Bean and Stephen Byers and Shane Hunziker, Duane M. Broline, Patrick Brosnan, Scott Brown, David Callan, Centre College Problem Solving Group, Patrick Costello, William H. Dent, Jr., Clayton W. Dodge, Herbert J. Dulle, Milton P. Eisner, Marjorie A. Fitting, F. J. Flanigan, Kevin Ford (student), Arthur H. Foss, Lorraine L. Foster, Kumar Garimella (student), William Gasarch, S. Gendler, John F. Goehl, Jr., David Graves, Cornelius Groenewoud, Russell Jay Hendel, Francis M. Henderson, Charles V. Heuer, G. A. Heuer, Keith Hirst (England), Jim O. Howard, Ambati Jayakrishna and Ambati Balamurali Krishna (students), H. K. Krishnapriyan and Mark Young (student), John and Libby Krussel, Kee-Wai Lau (Hong Kong), Norman F. Lindquist, Pamela Ann Lipka, David E. Manes, Paul A. Martin, Gonzalo J. Masjuán (Chile), Reiner Martin (student, West Germany), Gary R. Minnich (student), Jean-Marie Monier (France), Robert Patenaude, Allen Pedersen (Denmark), Stephen G. Penrice, Mike Pinter, Vivek Prabhakaran (student), R. Bruce Richter (Canada), Shippensburg University Mathematical Problem Solving Group, Harvey Schmidt, Jr., Nick Singer, Daniel L. Stock, Sahib Singh, Jack Tedeski (student), Sharon Tentarelli (student), Peter and Colleen Vachuska, Jack V. Wales, Jr., Edward T. H. Wang (Canada), William P. Wardlaw, I. A. Webb (student) and J. H. Webb (South Africa), Charles H. Webster, Herbert Wells IV, and the proposer.

Minnich *noted that if $x>68$ then x can be written as two distinct sums of two positive composite odd integers.* The Shippensburg University Mathematical Problem Solving Group *proved that the largest even integer which cannot be written as the sum of $2k$ composite odd integers is* $18k+20$.

Several people noted that this problem was posed, solved, and generalized in two articles: Erwin Just *and* Norman Schaumberger, "*A Curious Property of the Integer* 38," *this* MAGAZINE, 46 (1973), 221, *and*

A. M. Vaidya, *"On Representing Integers as Sums of Odd Composite Integers," this* MAGAZINE, 48 (1975), 221–223.

Inperfect Numbers **October 1989**

1329. *Proposed by Joe Flowers, Northeast Missouri State University, Kirksville, Missouri.*

For real-valued functions f and g defined on the set of positive integers, let $f * g$ denote the Dirichlet product defined by

$$(f * g)(n) = \sum_{d|n} f(d) \cdot g(n/d).$$

Let $\sigma(n)$ denote the sum of all positive factors of n and let σ^{-1} denote the inverse of σ with respect to Dirichlet multiplication. (Note: the identity element is I, where $I(1) = 1$ and $I(n) = 0$ for all $n > 1$.)

a. Evaluate $\sigma^{-1}(p^a)$, where p is prime.

b. Call n an "inperfect" (inverse-perfect) number if $\sigma^{-1}(n) = 2n$. Determine all inperfect numbers.

Solution by David Callan, University of Bridgeport, Bridgeport, Connecticut.

a. The Dirichlet inverse of any arithmetical function f with $f(1) \neq 0$ is given recursively by

$$f^{-1}(1) = \frac{1}{f(1)}, \qquad f^{-1}(n) = -\frac{1}{f(1)} \sum_{\substack{d|n \\ d<n}} f(n/d) f^{-1}(d) \text{ for } n > 1.$$

Recalling $\sigma(p^a) = (p^{a+1} - 1)/(p - 1)$, it is easy to find from this that $\sigma^{-1}(1) = 1$, $\sigma^{-1}(p) = -(p+1)$, $\sigma^{-1}(p^2) = p$ and $\sigma^{-1}(p^a) = 0$ for $a \geqslant 3$.

b. Since σ is multiplicative, so is σ^{-1}. Hence $\sigma^{-1}(6) = (-3)(-4) = 12$ and 6 is "inperfect" (as well as perfect). We show this is the only inperfect number.

Suppose n is inperfect. Thus, $\sigma^{-1}(n) \neq 0$ and letting p denote the largest prime factor of n and a denote the largest power of p that divides n, we have $a = 1$ or 2. Suppose, for a contradiction, that $p > 3$. Then, $p \nmid \sigma^{-1}(q^b)$, $b = 1$ or 2, for any prime $q < p$. So $p^2 \nmid \sigma^{-1}(n)$ and hence $a = 1$. But then $p \nmid \sigma^{-1}(p)$, violating $n | \sigma^{-1}(n)$. Thus, $p \leqslant 3$ and the result follows easily.

Also solved by Seung-Jin Bang (Korea), S. K. Berberian, W. E. Briggs, Duane M. Broline, Chico Problem Group, James T. Cross, Jesse Deutsch, David Doster, Hugh M. Edgar, Lorraine L. Foster, Heinz-Jürgen Seiffert (West Germany), H. K. Krishnapriyan, David E. Manes, Reiner Martin (student, West Germany), Gonzalo J. Masjuán (Chile), Kim McInturff, Maria Luisa Oliver (Argentina), R. Patenaude, Ken Rebman, James Sellers, Nick Singer, Sahib Singh and Stephen I. Gendler, William P. Wardlaw, Western Maryland College Problems Group, Kenneth L. Yocom, and the proposer.

Binomial Identity **October 1989**

1330. *Proposed by K. L. McAvaney, Deakin University, Geelong, Victoria, Australia.*

Show that for all integers $p \geqslant 2$ and $n = 2, 4, \ldots, 2p - 2$,

$$\sum_{k=1}^{p} (-1)^k \binom{2p}{p+k} k^n = 0.$$

Solution by The Chico Problem Group, California State University, Chico, California.

We use the generating function

$$g(x)=(1-e^{x})^{2p}=\sum_{j=0}^{2p}\binom{2p}{j}(-1)^{j}e^{jx}$$
$$=\sum_{k=-p}^{p}\binom{2p}{p+k}(-1)^{k+p}e^{(k+p)x},$$

where we substitute $j=k+p$.

Let

$$F(x)=e^{-px}(1-e^{x})^{2p}=(-1)^{p}\sum_{k=-p}^{p}\binom{2p}{p+k}(-1)^{k}e^{kx}$$

so that

$$F^{(n)}(x)=(-1)^{p}\sum_{k=-p}^{p}\binom{2p}{p+k}(-1)^{k}k^{n}e^{kx}.$$

For $x=0$, $F^{(n)}(0)=(-1)^{p}(S_1+S_2)$, where

$$S_1=\sum_{k=-p}^{-1}\binom{2p}{p+k}(-1)^{k}k^{n},\quad\text{and}\quad S_2=\sum_{k=1}^{p}\binom{2p}{p+k}(-1)^{k}k^{n}.$$

We note that if n is even, $S_1=S_2$. But

$$F(x)=(e^{-x}-1)^{p}(1-e^{x})^{p}$$
$$=(-1)^{p}(1-e^{-x})^{p}(1-e^{x})^{p}$$
$$=(-1)^{p}(2-2\cosh^{p}x)$$
$$=(-1)^{p}4^{p}\sinh^{2p}(x/2).$$

Thus, $F(x)$ has a power series whose lowest power of x is x^{2p}; hence, $F^{(n)}(0)=0$ for all n, $1\leqslant n\leqslant 2p-1$. In particular, if n is even, then $S_2=0$, which was to be proved.

Also solved by J. C. Binz (Switzerland), Paul Bracken (Canada), David Callan, David Doster, Robert Doucette, H. K. Krishnapriyan, Allen Pedersen (Denmark), Heinz-Jürgen Seiffert (West Germany), William F. Trench, Luis Verde-Star (Mexico), Robert J. Wagner, Western Maryland College Problems Group, and the proposer.

Many authors based their proofs on the differences of a polynomial. Callan, Doucette, *and* Seiffert *noted that for $n=2p$, the value is $(-1)^{p}\frac{1}{2}(2p)!$, and* Callan *noted the value is $(-1/2)\binom{2p}{p}$ for $n=0$.*

Commuting Matrices and Matrix Polynomials **October 1989**

1331. *Proposed by Daniel Shapiro, The Ohio State University, Columbus, Ohio.*

Suppose that $\mathbf{A}$ and $\mathbf{B}$ are $n\times n$ matrices (with complex entries, say). If there exists another matrix $\mathbf{P}$ with the property that $\mathbf{A}=f(\mathbf{P})$ and $\mathbf{B}=g(\mathbf{P})$ for some polynomials $f(x)$ and $g(x)$, then clearly $\mathbf{AB}=\mathbf{BA}$. Is the converse true? That is, if $\mathbf{AB}=\mathbf{BA}$ for two matrices $\mathbf{A},\mathbf{B}$, does it follow that $\mathbf{A}$ and $\mathbf{B}$ are expressible as polynomials in some matrix $\mathbf{P}$?

Solution by Harvey Schmidt, Jr., Lewis and Clark College, Portland, Oregon.

The answer is no. Let

$$\mathbf{A}=\begin{pmatrix}0&1&0\\0&0&0\\0&0&0\end{pmatrix} \quad \text{and} \quad \mathbf{B}=\begin{pmatrix}0&0&0\\0&0&0\\0&1&0\end{pmatrix}.$$

Then certainly **A** and **B** commute. Just suppose there exists a matrix **P** and polynomials $f(x)$ and $g(x)$ with the property that $\mathbf{A}=f(\mathbf{P})$ and $\mathbf{B}=g(\mathbf{P})$. Then $\mathbf{P}=(p_{ij})$ commutes with both **A** and **B**, so it is easy to check that

$$P=\begin{pmatrix}p_{11}&p_{12}&0\\0&p_{11}&0\\0&p_{32}&p_{11}\end{pmatrix}.$$

Now, if $\mathbf{A}=\alpha_0\mathbf{I}+\alpha_1\mathbf{P}+\alpha_2\mathbf{P}^2=f(\mathbf{P})$ and $\mathbf{B}=\beta_0\mathbf{I}+\beta_1\mathbf{P}+\beta_2\mathbf{P}^2=g(\mathbf{P})$, then it is straightforward to check that

$$\begin{aligned}1&=\alpha_1p_{12}+2\alpha_2p_{11}p_{12}\\0&=\alpha_1p_{32}+2\alpha_2p_{11}p_{32}\\0&=\beta_1p_{32}+2\beta_2p_{11}p_{32}\\1&=\beta_1p_{32}+2\beta_2p_{11}p_{32}.\end{aligned}$$

Notice that the first equation implies that $p_{12}\neq 0$ and the last equation implies that $p_{32}\neq 0$. Dividing the first equation by p_{12} and the second by p_{32}, it follows that the first two equations are inconsistent, a contradiction.

Also solved by Irl C. Bivens, Duane M. Broline, F. J. Flanigan, David E. Manes, Marvin Marcus, Jean-Marie Monier (France), Pei Yuan Wu (Republic of China), and the proposer.

Shapiro *pointed out two cases where the result is true*: (*i*) *A and B are diagonalizable*; Flanagan *attributes the special case of Hermitian matrices to Perlis* (1952), (*ii*) *A is cyclic* (*its minimal polynomial is equal to its characteristic polynomial*). *In this case, B is a polynomial in A*. Wu *and* Flanigan *referred to* Gerstenhaber ("*On dominance and varieties of commuting matrices*", *Annals of Mathematics* 73 (1961), 324–348) *which proved the more general theorem that the dimension of the algebra generated by commuting matrices A and B is at most n. From these two cases, the result follows for all* 2×2 *matrices that commute.*

Comments

1311. Murray Klamkin notes that the result of this problem is equivalent to a generalization he gave to a USA Olympiad Problem (M. S. Klamkin, *USA Mathematical Olympiads, 1972–1976*, MAA, Washington, D.C., 1988, p. 84). Moreover, the result in this reference is valid for both the odd and even cases.

Answers

Solutions to the Quickies on p. 274.

A766. Suppose $b_1|x-a_1|+b_2|x-a_2|+\cdots+b_k|x-a_k|=0$. If, say, $b_1\neq 0$, then

$$b_1|x-a_1|=-b_2|x-a_2|-\cdots-b_k|x-a_k|. \qquad (*)$$

The right-hand side of $(*)$ is differentiable at a_1, but the left-hand side is not. Contradiction.

A767. The required set is exactly the set of all primes. If a_0 is a prime, then $a_1 = a_2 = \cdots = 2$, and the sequence does not contain squares. Conversely, let $a_0 > 1$ be a composite number. Then our sequence decreases until a term equal to 2 occurs. Let i be the first index satisfying $a_i = 2$. Clearly, $i \geqslant 2$ and a_{i-1} is an odd prime. Then a_{i-2} is a square.

A768. By the power mean inequality,

$$\frac{\sum a^{2n}}{\sum a^n} = \frac{\sum a^n \cdot a^n}{\sum a^n} \geqslant \left(\frac{\sum a^n \cdot a}{\sum a^n} \right)^n,$$

where the sums here and subsequently are symmetric over a, b, c. Then

$$\frac{\sum a^{n+1}}{\sum a^n} \geqslant \frac{\sum a^2}{\sum a}$$

since it is equivalent to

$$\sum ab(a^{n-1} - b^{n-1})(a-b) \geqslant 0.$$

The inequality can be easily extended for a more general combination of exponents and to any number of variables.

Blake and Fractals

William Blake said he could see
Vistas of infinity
In the smallest speck of sand
Held in the hollow of his hand.

Models for this claim we've got
In the work of Mandelbrot:
Fractal diagrams partake
Of the essence sensed by Blake.

Basic forms will still prevail
Independent of the scale;
Viewed from far or viewed from near
Special signatures are clear.

When you magnify a spot,
What you had before, you've got.
Smaller, smaller, smaller yet,
Still the same details are set;

Finer than the finest hair
Blake's infinity is there,
Rich in structure all the way—
Just as the mystic poets say.

—J. D. MEMORY
North Carolina State University

A767. The required set is exactly the set of all primes. If a_0 is a prime, then $a_1 = a_2 = \cdots = 2$, and the sequence does not contain squares. Conversely, let $a_0 > 1$ be a composite number. Then our sequence decreases until a term equal to 2 occurs. Let i be the first index satisfying $a_i = 2$. Clearly, $i \geqslant 2$ and a_{i-1} is an odd prime. Then a_{i-2} is a square.

A768. By the power mean inequality,

$$\frac{\sum a^{2n}}{\sum a^n} = \frac{\sum a^n \cdot a^n}{\sum a^n} \geqslant \left(\frac{\sum a^n \cdot a}{\sum a^n} \right)^n,$$

where the sums here and subsequently are symmetric over a, b, c. Then

$$\frac{\sum a^{n+1}}{\sum a^n} \geqslant \frac{\sum a^2}{\sum a}$$

since it is equivalent to

$$\sum ab(a^{n-1} - b^{n-1})(a - b) \geqslant 0.$$

The inequality can be easily extended for a more general combination of exponents and to any number of variables.

Blake and Fractals

William Blake said he could see
Vistas of infinity
In the smallest speck of sand
Held in the hollow of his hand.

Models for this claim we've got
In the work of Mandelbrot:
Fractal diagrams partake
Of the essence sensed by Blake.

Basic forms will still prevail
Independent of the scale;
Viewed from far or viewed from near
Special signatures are clear.

When you magnify a spot,
What you had before, you've got.
Smaller, smaller, smaller yet,
Still the same details are set;

Finer than the finest hair
Blake's infinity is there,
Rich in structure all the way—
Just as the mystic poets say.

—J. D. MEMORY
North Carolina State University

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 the mathematics literature. Readers are invited to suggest items for review to the editors.

Peterson, Ivars, Curves for a tighter fit: Number theory provides a novel strategy for packing spheres efficiently, *Science News* 137 (19 May 1990), 316-317.

Elliptic curves, which served in the last few years as the font for new attacks on Fermat's Last Theorem and new algorithms for factoring, now have yielded new strategies for finding dense packings of spheres in higher dimensions.

Cipra, Barry A., How to catch a cheating computer, *Science* 248 (1 June 1990), 1079-1080; Cipra, Barry A., Computational complexity theorists tackle the cheating computer conundrum, *SIAM News* 23:4 (July 1990), 1, 18-20.

For a number of years, mathematicians have used proofs of primality that are probabilistic in nature. Whether an integer is prime can be "proved" by subjecting it to testing; if any test is negative, the integer is not prime, but a positive result only adds to the likelihood of primality. Passing a sequence of, say, 30 such independent tests, each with probability one-half of a false positive, provides very strong evidence that the number is prime. A similar approach, "interactive proof," allows efficient (polynomial-time but still probabilistic) verification of solutions to a wide class of problems. Adi Shamir (Weizmann Institute) has shown that the problems is PSPACE (a class that includes NP) are exactly those for which there are interactive proofs.

Kolata, Gina, Giant leap in math: 155 divided to 0, *New York Times* (20 June 1990) A8 (National Edition); Number field sieve produces factoring breakthrough, *SIAM News* 23:4 (July 1990), 1, 11.

Several hundred researchers, using their computers in parallel and communicating results by electronic mail, have factored a 155-digit number on the "10 Most Wanted Numbers" list. The achievement, which took only one month, lengthens the shadow over cryptographic schemes that are based on the difficulty of factoring. A 7-digit factor of the number ($2^{512} + 1$—the ninth Fermat number) had already been known, so the new achievement was the factoring of a 148-digit number (into 99-digit and 49-digit factors). The method uses a new algorithm, a generalization by Hendrik Lenstra (UC-Berkeley) of an approach of John Pollard (Reading, England) that is based on factorization of integers in algebraic number fields; the method applies to integers of the form $a^k + b$ for a and b very small and k very large. For the problem just solved, the results done in parallel were assembled to construct a sparse linear system with more than 200,000 equations, which was then reduced to a dense system of 72,000 equations by a stuctured Gaussian elimination. The latter system was solved in three hours.

Peterson, Ivars, *Islands of Truth: A Mathematical Mystery Cruise*, Freeman, 1990; xvii + 325 pp + 16 color plates, $19.95. ISBN 0-7167-2113-9

The author of *The Mathematical Tourist* (Freeman, 1989) has set off on another tour, and you and your students are all invited to enjoy splendidly lucid explanations of recent research developments at the frontiers of mathematics. Much of the material here appeared in somewhat different form in *Science News* over the past eight years.

Tufte, Edward R., *Envisioning Information*, Graphics Press (Box 430, Cheshire, CT 06410), 1990; 126 pp, $48. ISBN none

A sequel to the author's *The Visual Display of Quantitative Information* (1983), this masterful and beautiful book "celebrates escapes from flatland," displaying "design strategies for enhancing the dimensionality and density of portrayals of information." From the many examples, Tufte distills general principles. Users of so-called "business graphics" software (especially in color) badly need this book, and students of statistics would benefit from its examples and observations.

Starfield, Anthony M., et al., *How to Model It: Problem Solving for the Computer Age*, McGraw-Hill, 1990; xii + 206 pp (P). ISBN 0-07-005897-0

A refreshing change from modeling books that are really taxonomies, this book concentrates on fostering an active learning environment. The book arose from, and is designed for, a course whose main classroom activity is modeling in groups of two to four. More than other books, it encourages the participants to ask themselves (and each other) reflective questions about what they are doing and what they are learning.

Gardner, Martin, *The New Ambidextrous Universe: Symmetry and Asymmetry from Mirror Reflections to Superstrings*, 3rd rev. ed., W.H. Freeman, 1990; xiv + 392 pp, $19.95. ISBN 0-7167-2092-2

A substantial revision of Gardner's well-known and justly famous work on mirror symmetry and asymmetry. Five chapters (on antimatter, monopoles, theories of matter, spin, and superstrings) are entirely new.

Andrews, Edmund L., Patents: "Breeding" computer programs, *New York Times* (30 June 1990), 18 (National Edition).

John R. Koza (Stanford University) has received a patent for a particular approach to generating "genetic" algorithms. Koza's method has the computer start with a problem and relevant variables and possible operations on them, generate programs at random, then "mate" those programs that show signs of effectiveness in solving the problem.

Samuelson, Pamela, Legally speaking: Should program algorithms be patented?, *Communications of the Association for Computing Machinery* 33:8 (August 1990), 23-27.

A professor of law summarizes relevant decisions of the US Supreme Court, points out the need for a standard of patentability, and notes a forthcoming study by the Office of Technology Assessment of the US Congress.

Gleick, James, The census: Why we can't count, *New York Times Magazine* (15 July 1990), 22-26, 54.

Cites the difficulties, failures, and expense of the US decennial census, particularly the argument over whether the raw data should be adjusted for undercounting.

Peterson, Ivars, Rock and roll bridge: A new analysis challenges the common explanation for a famous collapse, *Science News* 137 (2 June 1990), 344-346.

The traditional explanation for the dramatic collapse of the Tacoma Narrows suspension bridge in 1940 is simple harmonic oscillator resonance, resulting from the applied force of wind vortices. But such an explanation—even if it makes for a great example in a differential equations course—depends critically on linearity of response of the bridge. P. Joseph McKenna (U. of Connecticut) and Alan C. Lazer (U. of Miami) claim that in fact suspension bridges are characterized by nonlinearity, because of the inelasticity of the vertical cables. Nonlinear theory predicts that a suspension bridge can respond nonlinearly to a whole range of forcing frequencies, not just to one that matches the bridge's natural (linear) oscillation. For future bridges, McKenna and Lazer recommend under-cables to counterbalance the nonlinear effects of the upper cables.

Reingold, Edwin M., Mathematics made easy: A Japanese teaching method adds up in the U.S., *Time* (4 June 1990), 83-84, (9 July 1990), 6.

In times of crisis, nostrums sell well. The Kumon method from Japan, for increasing speed and accuracy in computation, has been adopted by a number of US elementary schools. The method covers arithmetic to calculus; it features self-paced mastery learning and emphasizes mechanics over theory. (Does the system allow or encourage calculators? I'm doubtful.) This article reports better self-confidence, self-image, and motivation for students, which seem to be the major results that teachers and parents are looking for. (Apart from this laudable result of maintaining the children's psychological well-being, what happened to learning as a goal?) Most of all, I hope that a cycle of fresh instances of the Hawthorne effect ("any change produces transient improvement") does not turn out to be our best approach to the problems of mathematics learning.

Olsen, L.F., and W.M. Schaffer, Chaos versus noisy periodicity: Alternative hypotheses for childhood epidemics, *Science* 249 (3 August 1990), 499-504.

"Especially in the physical sciences, the ubiquity of chaotic fluctuations is now well established." This article on models for chickenpox and measles begins with a sentence I found striking because the establishment referred to has taken only five years or so. Meanwhile, the prevailing position of calculus in the first-year college curriculum is a consequence of the same realization being made about calculus—but many decades ago. The modeling in this article makes use of a whole panoply of more-recent mathematics (time series, power spectra, nonlinear DEs, phase portraits, autocorrelation, multiplicative confidence intervals, chaos, return map, fractals, trajectorial divergence, Lyapunov characteristic exponent, correlation dimension, Monte Carlo simulations, noise-induced bifurcations). If this is the kind of mathematics our students going into other fields will use, and those becoming mathematicians will need to know in order to be successful modelers, where and when are they going to learn it? And when will the mathematics curriculum begin to resemble in flexibility the curricula in other sciences, in which contemporary discoveries enter freshman textbooks within a year or two?

Nahin, Paul J., Oliver Heaviside, *Scientific American* 262:6 (June 1990), 122-129.

What we now call Maxwell's equations for electromagnetism in fact are the distillation by Oliver Heaviside of Maxwell's original 20 equations in 20 variables. Heaviside, together with Gibbs, was responsible for physicists adopting vectors. This article only briefly alludes to the operational calculus invented by Heaviside, but gives some details of his life and his role in making long-distance telephony feasible.

NEWS AND LETTERS

LETTERS TO THE EDITOR

Editor:

The lovely article *A Fibonacci-like Sequence of Composite Numbers* by Donald E. Knuth [63 (1990), 21-25] shows that a certain pair (a, b) of relatively prime 17 digit numbers begins a 'Fibonacci-like' sequence all of whose members are composite integers, thereby improving an earlier result of Ron Graham that exhibited such a pair in which a and b each had around 33 digits. We remark here that if we have any such pair in which $a < b$ and $b - a$ is a composite then the smaller pair $(b-a, a)$ still has the required property.

Now the smallest pair found by Knuth has $a > b$, so the remark does not apply to it. However one of the pairs that he found but did not use because it was larger than the chosen one *does* satisfy $a < b$, and so it can be improved. Precisely, the third pair from the bottom on page 24 is

$$a = 52{,}615{,}644{,}495{,}813{,}559$$
$$b = 80{,}820{,}849{,}424{,}504{,}095,$$

and we can indeed project it backwards as remarked above. It turns out that this pair is actually (A_4, A_5) in a composite sequence that begins with

$$a = A_0 = 20{,}615{,}674{,}205{,}555{,}510$$
$$b = A_1 = 3{,}794{,}765{,}361{,}567{,}513.$$

These are smaller than his final selection roughly by factors of 3 and 16, respectively.

Herbert S. Wilf
University of Pennsylvania
Philadelphia, PA 19104-6395

Editor:

The April issue of *Mathematics Magazine* contains an article by Roger Herz-Fischler on "Dürer's Paradox or Why an Ellipse Is Not Egg-Shaped". The author refers to a note I wrote for *The Australian Mathematics Teacher* in 1974 on "Albrecht Dürer and the Ellipse". This note is a shortened form of one of the talks I gave on Dürer while spending part of a sabbatical in Australia.

Herz-Fischler seems to question my assertion in this note that the equations of the conic sections follow from Dürer's methods if one uses a theorem on similar triangles. He remarks that I give no further details, and that the theorem on similar triangles in "non-specified"!

On p. 79 Herz-Fischler says that he will prove that the ellipse he is considering has the same width at points symmetrically located with respect to the centre O. He devotes rather a lot of space and much irrelevant trigonometry to the proof. But using my "non-specified" theorem on similar triangles, that corresponding sides are proportional, this can be done in a few lines.

Dürer made his offending drawing of an ellipse on woodblocks, but with modern instruments and good paper Herz-Fischler is unable to make any of his triangles isosceles, and as a result his w_1 and w_2 in Figure 5 are visibly unequal. He might have modified his article, and perhaps not published it, if he had seen my book *Geometry and the Visual Arts*, currently a Dover Publication. This book contains a fair amount of material on Dürer and his magnificent work in geometry education. Originally published in 1976 by Penguin Books as *Geometry and the Liberal Arts*, it has been translated, with sometimes essential title modifications, into Russian, Japanese, Spanish and Dutch. It is now a Dover book, but it seems to be unknown to the many who advised Herz-Fischler on his paper.

After finding the equation of an elliptic section of the cone, Herz-Fischler invites his readers to find the equation of a parabolic section, by presumably imitating his methods, which he states on p. 80 to be the simplest approach. I find the equation of a parabolic section in my book, rather simply. The present revival of interest in Euclidean geometry is to be welcomed, and may give some relief from the strident and well-funded emphasis that geometry is only what you do with supercomputers, but the pleasure derived from geometry, as with music, depends very much on elegant performance.

Dan Pedoe
Professor Emeritus
University of Minnesota

Editor:

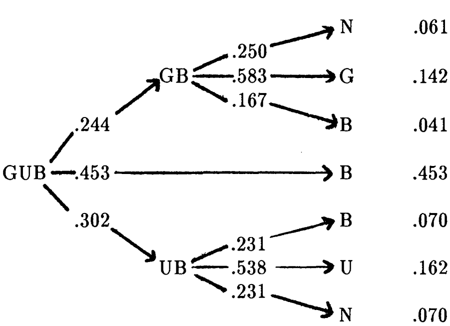
My professor, Dr. Rick Luttmann, assigned our mathematics class the problem of analyzing a truel. His source for the problem was "The Gunfight at the OK Corral" by J. Sandefur (this MAGAZINE 62 (1989), 119-124) in which the problem is solved using Markov chains. Not having been introduced to Markov chains, I found an alternate solution, which I thought might be of interest.

Recall that the problem is a gunfight in which the three participants, the Good, the Bad, and the Ugly draw and fire simultaneously, each combatant firing at the better shot of his opponents. The survivors of each round of the truel fight another round until there is no more than one survivor. Our problem is to find the probabilities of each possible final outcome of the gunfight. We assume that Good hits his target 60% of the time, Ugly hits his 50% of the time, and Bad is on target 30% of the time. We denote by G, U, B, and N the outcomes Good survivies, Ugly survives, Bad survives, and Nobody survives, respectively.

Rather than consider the gunfight round by round as is done using Markov chains, we need consider only the round(s) in which someone is eliminated. We may do this because the probability that the first lethal round will begin with all three combatants still alive is 1, no matter how many nonlethal rounds have occurred. If two or three combatants are killed in the first lethal round, the fight is over. Otherwise, there will be a second lethal round, and the probability that it will involve both survivors of the first lethal round is also 1. What we must do for each lethal round is compute the conditional probabilities of each possible outcome, given that the round was lethal.

We note that: $P(GUB \geq GB) = (.6)(.5)(.7) = .21$ and $P(\text{lethal round}) = 1 - (.4)(.5)(.7) = .86$. Therefore: $P(GUB \geq GB : \text{lethal}) = .21/.86 = .244$.

With similar computations, the other conditional probabilities may be found. Now we are ready for a simple tree diagram:



The probabilities of the four possible outcomes may now be easily determined.

Keith Krieger, student
Sonoma State University
Rohnert Park, CA 94928

Editor:

In the note "Configurations Arising from the Three Circle Theorem" (this MAGAZINE 63 (1990), 116-121) the footnote on p. 116 should appear on p. 121. This note is the one for reference [8] Ushijima, Zoku Sangaku Shōsen.

Hiroshi Okumura
Annaka High School
Annaka Gunma 379-01, Japan

Editor:

In the June 1990 issue of this MAGAZINE, Russell Jay Hendel presents a nice "proof without words" of the familiar formula for the area of a disk, in which he illustrates how a disk of radius R can be transformed into a triangle with height R and base $2\pi R$ [2].

However, this approach to the area of a disk is not new. Virtually the same demonstration appeared in this journal in 1977, together with some historical background for this method of demonstrating the area formula [1]. I have used this approach in several of my elementary classes; students find it quite convincing.

References

1. S. Epstein and M. Hochberg, A Talmudic Approach to the Area of a Circle, this MAGAZINE 50 (1977), 210.
2. R. J. Hendel, Proof without Words: Area of a Disk is πR^2, this MAGAZINE 63 (1990), 188.

Roger B. Nelsen
Lewis & Clark College
Portland, OR 97219

Other readers observed this earlier appearance in the MAGAZINE. It was also pointed out that "proof without words" is an overstatement and something like "heuristic observation without words" might have been more suitable.

—Editor

19th USA MATHEMATICAL OLYMPIAD

The 1990 USAMO was prepared by Douglas Hensley, Gerald Heuer, Gregg Patruno, Bjorn Poonen, Ian Richards (chair), Leo Schneider, and Daniel Ullman.

The publication *Mathematical Olympiads* 1990 will present comprehensive solutions to these problems, along with additional comments and the problems, solutions, and results of the 1990 International Mathematical Olympiad. This helpful booklet will be available from:

Dr. Walter Mientka
Department of Mathematics
University of Nebraska
Lincoln, NE 68588-0322

Copies are \$2.00 for each year, 1976-1990, with a minimum order of \$6.00, payable in US funds.

1. A certain state issues license plates consisting of six digits (from 0 through 9). The state requires that any two plates differ in at least two places. (Thus the plates 027592 and 020592 cannot both be used.) Determine, with proof, the maximum number of distinct license plates that the state can issue.

Sol. The state can issue 10^5 license plates. One method for doing so is to use a "check digit", as follows. For each of the 10^5 five-digit strings $x_1x_2x_3x_4x_5$, define x_6 so that

$$x_6 \equiv x_1 + x_2 + x_3 + x_4 + x_5 \pmod{10}.$$

For any pair of distinct strings $a_1a_2a_3a_4a_5a_6$ and $b_1b_2b_3b_4b_5b_6$ constructed in this way, there must be at least one position j with $1 \leq j \leq 5$ such that $a_j \neq b_j$. But if there is only one such j, then

$$\begin{aligned} a_6 - b_6 &\equiv (a_1 + a_2 + a_3 + a_4 + a_5) \\ &\qquad - (b_1 + b_2 + b_3 + b_4 + b_5) \\ &= a_j - b_j \\ &\not\equiv 0 \pmod{10}, \end{aligned}$$

so $a_6 \neq b_6$. Hence $a_1a_2a_3a_4a_5a_6$ and $b_1b_2b_3b_4b_5b_6$ must differ in at least two places.

To show that no method can produce a greater number of acceptable license plates, observe that among $10^5 + 1$ distinct license plates, two would have to agree in their first 5 places, where only 10^5 distinct combinations are possible. These two license plates would differ in only one place.

2. A sequence of functions $\{f_n(x)\}$ is defined recursively as follows:

$$f_1(x) = \sqrt{x^2 + 48}\ , \quad \text{and}$$

$$f_{n+1}(x) = \sqrt{x^2 + 6f_n(x)} \quad \text{for} \quad n \geq 1.$$

(Recall that $\sqrt{\ }$ is understood to represent the positive square root.) For each positive integer n, find all real solutions of the equation $f_n(x) = 2x$.

Sol. Observe first that $f_n(x)$ is positive for every n and every x, so $f_n(x) = 2x$ admits only positive solutions.

We show that, for each n, the unique solution of $f_n(x) = 2x$ is $x = 4$. To that end, we first prove by induction on n that $x = 4$ is a solution. For $n = 1$ we have $f_1(4) = \sqrt{16 + 48} = 2\cdot 4$. Suppose that $f_k(4) = 2\cdot 4$. Then $f_{k+1}(4) = \sqrt{16 + 6f_k(4)} = \sqrt{16 + 48} = 2\cdot 4$, completing the induction.

To see that there are no other solutions, we show by induction that

(*) for each n, $\frac{f_n(x)}{x}$ decreases as x increases in $(0, \infty)$,

and therefore cannot take on the value 2 more than once. With $n = 1$ we have $\frac{f_1(x)}{x} = \sqrt{1 + \frac{48}{x^2}}$, which decreases as x increases. Now suppose that $\frac{f_k(x)}{x}$ decreases as x increases. Then $\frac{f_{k+1}(x)}{x} =$

$$\sqrt{1 + \frac{6}{x} \cdot \frac{f_k(x)}{x}} \text{ also decreases as } x \text{ increases, and}$$

the proof of (*) is complete.

3. Suppose that necklace A has 14 beads and necklace B has 19. Prove that, for every odd integer $n \geq 1$, there is a way to number each of the 33 beads with an integer from the sequence

$$\{n,\ n+1,\ n+2,\ \ldots,\ n+32\}$$

so that each integer is used once, and adjacent beads correspond to relatively prime integers. (Here a "necklace" is viewed as a circle in which each bead is adjacent to two other beads.)

Sol. We take an integer m, $1 \leq m \leq 18$, and number the beads of necklace A with the consecutive integers $n+m$, $n+m+1$, ..., $n+m+13$. We have to join the ends of this chain together, which is allowed so long as $n+m$ and $n+m+13$ are relatively prime; i.e.,

$$\gcd(n+m,\ n+m+13) = 1.$$

(Since pairs of consecutive positive integers are *always* relatively prime, no other conditions on necklace A are necessary.)

Next, we number the beads of necklace B with the consecutive integers from $n+m+14$ to $n+32$ (note that $m+14 \leq 32$), and then follow $n+32$ with n and continue up to $n+m$-1. Again, all pairs of consecutive positive integers are relatively prime, so the numbering will succeed provided that the following two constraints are satisfied:

$$\gcd(n,\ n+32) = 1, \text{ and}$$
$$\gcd(n+m-1,\ n+m+14) = 1.$$

Thus, since $\gcd(a,b) = \gcd(a,\ b-a)$, the numbering method is valid if

$$\gcd(n,32) = \gcd(n+m-1,15) = \gcd(n+m,13) = 1.$$

The first of these conditions holds automatically, since n is given as odd. The second condition (modulo 15) is equivalent to corresponding conditions modulo 3 and modulo 5. Therefore, the numbering succeeds if m can be chosen such that

$$m \not\equiv 1-n \pmod 3,$$
$$m \not\equiv 1-n \pmod 5,$$
$$m \not\equiv \ -n \pmod{13}.$$

Out of the 18 consecutive possible values for m, only 6 will have $m \equiv 1 - n \pmod 3$, at most 4 will have $m \equiv 1-n \pmod 5$, and at most 2 will have $m \equiv -n \pmod{13}$. This leaves at least $18-(6+4+2) = 6$ values of m satisfying all the requirements for the numbering system to work.

Note. The interval $1 \leq m \leq 18$ can be replaced by certain shorter intervals; e.g., $1 \leq m \leq 5$.

4. Find, with proof, the number of positive integers whose base-n representation consists of distinct digits with the property that, except for the leftmost digit, every digit differs by ± 1 from some digit further to the left. (Your answer should be an explicit function of n in simplest form.)

Sol. Since only the digit 0 is not permitted to begin a number, let us temporarily remove this asymmetry by defining $F(n)$ as the number of suitable base-n integers *plus* the number of base-n digit-strings which begin with 0 but are otherwise suitable. Thus

$F(1) = 1$ digit string [0];
$F(2) = 4$ digit-strings [0, 01, 1, 10];
$F(3) = 11$ digit strings [0, 01, 012, 1, 10, 102, 12, 120, 2, 21, 210];
etc.

Now, to establish a recursive expression for $F(n+1)$, observe that the suitable digit-strings in base $n+1$ fall into three disjoint classes:

(1) a single digit $0,1,2,\ldots,n$;

(2) a suitable digit-string in base n followed by the next-largest unused digit;

(3) a suitable digit-string in base n, with each digit increased by 1, followed by the next-smallest unused digit.

It follows that

$$F(n+1) = n + 1 + 2\cdot F(n).$$

With base case $F(1) = 1$, straightforward induction shows that $F(n) = 2^{n+1}-n-2$.

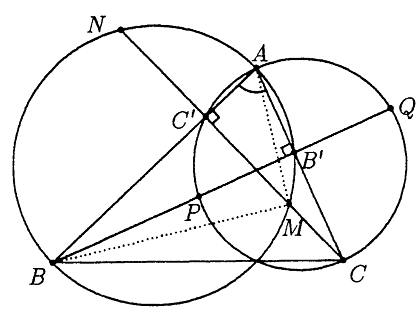
The only suitable digit-strings in base n which are *not* counted as suitable integers in the original problem are

$$0,\ 01,\ \ldots,\ 012\ \cdots(n-1),$$

so there remain $2^{n+1}-2n-2$ suitable integers.

5. An acute-angled triangle ABC is given in the plane. The circle with diameter AB intersects altitude CC' and its extension at points M and N, and the circle with diameter AC intersects altitude BB' and its extension at P and Q. Prove that the points M, N, P, Q lie on a common circle.

Sol.



The perpendicular bisectors of MN and PQ are AB and AC, respectively, so if M,N,P,Q do lie on a circle, the center can only be point A. Since we already have $AM = AN$ and $AP = AQ$, we have only to prove $AM = AP$.

Point C' is the foot of the altitude to the hypotenuse of right triangle ABM; therefore,

$$AM^2 = AB\cdot AC' = AB\cdot AC \cos \angle BAC.$$

Similarly,

$$AP^2 = AC\cdot AB' = AC\cdot AB \cos \angle BAC.$$

Thus $AM = AP$ and the problem is solved.

Help your students discover more meaningful relationships.

Free classroom display device with purchase of 30 calculators.

Showing is much more powerful than telling. So we've developed special classroom displays for our two most advanced calculators.

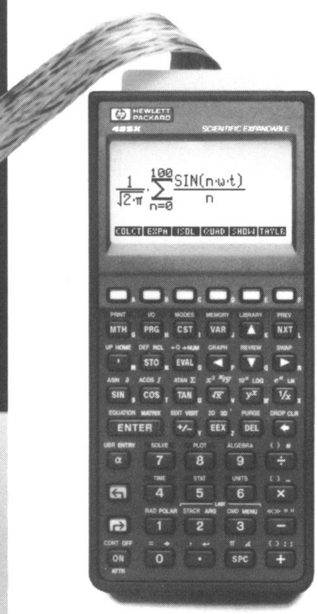
The new scientific expandable HP 48SX calculator and the cost-effective HP 28S are designed to put your students on the cutting edge of calculus and engineering. With more built-in functions and graphics solutions than any other calculator.

If your department or students purchase 30 HP 48SX or HP 28S calculators (or a mix of both), we'll give you a free HP 48SX and plug-in classroom display (an $850 retail value). Or a classroom display version of the HP 28S (a $500 retail value). And free teaching materials in the bargain.

Call us at **(503) 757-2004** between 8am and 3pm PDT for more details. Or write: Calculator Support, Hewlett-Packard, 1000 NE Circle Blvd., Corvallis, OR 97330. Offer ends October 31, 1990, and applies to instructors at college and high school levels only.

There is a better way.

HEWLETT PACKARD



CONTENTS

ARTICLES

NOTES

PROBLEMS

REVIEWS

NEWS AND LETTERS

THE MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth Street, NW
Washington, D.C. 20036