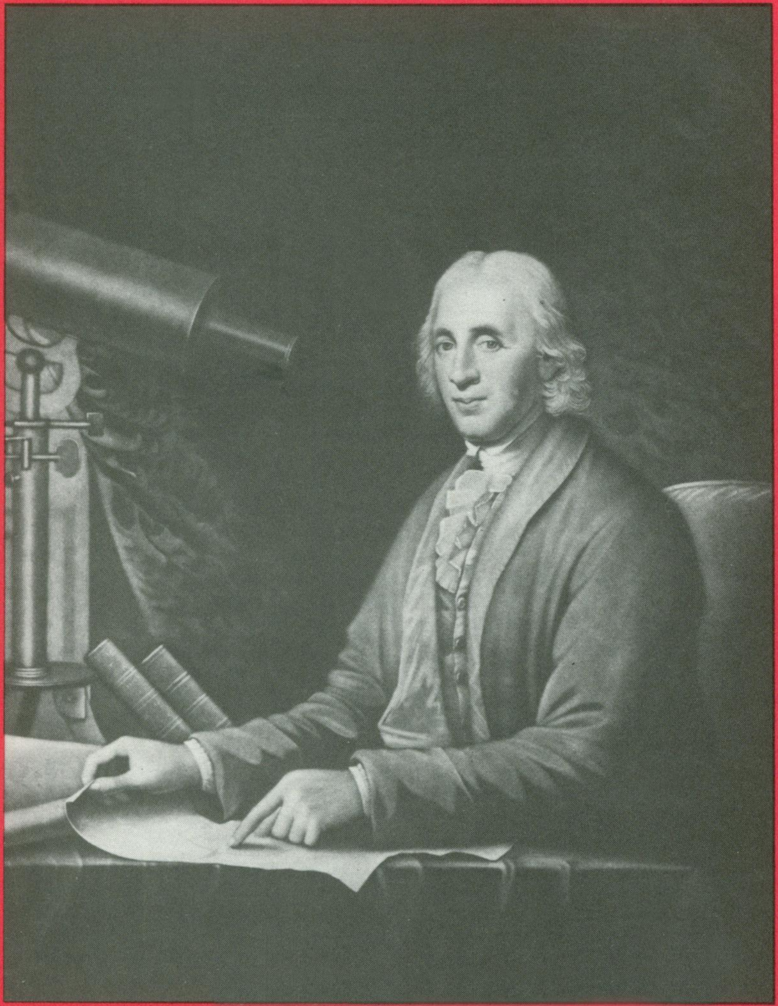
MATHEMATICS MAGAZINE



Vol. 60 No. 1
February 1987

DAVID RITTENHOUSE • MARKOV CHAINS
PERFECT SHUFFLES • ELECTRICAL LEMMA

MATHEMATICS MAGAZINE

Volume 60, Number 1, February, 1987
(ISSN—0025-570X)

COVER: *David Rittenhouse* by Edward Savage (after Charles Willson Peale), from the National Portrait Gallery, Washington, D.C.

ARTICLES

NOTES

PROBLEMS

REVIEWS

NEWS AND LETTERS

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: G. L. Alexanderson, Editor, Mathematics Magazine, Santa Clara University, Santa Clara, CA 95053. 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 one copy and keep one copy. Illustrations should be carefully prepared on separate sheets in black ink, the original without lettering and two copies with lettering added.

AUTHORS

Kent Morrison attended the University of California at Santa Cruz, where he received a doctorate in mathematics. He is on the faculty of California Polytechnic State University in San Luis Obispo and has taught at Utah State, Haverford, and U.C. Santa Cruz. He became interested in shuffle groups upon reading the short article by Gina Kolata in *Science* about the work of Diaconis, Graham, and Kantor. His research interests are in algebraic and differential geometry.

Steve Medvedoff received his B.S. and M.S. degrees from California Polytechnic State University in San Luis Obispo and is currently employed by AT&T Technologies, Inc. He was introduced to perfect shuffle groups by Kent Morrison, who suggested the topic for his senior project and served as his advisor. His other interests include topology, scientific applications programming, and creative writing.

Frederick A. Homann, S.J., received a doctorate in mathematics from the University of Pennsylvania in 1959 where his dissertation director, Hans Rademacher, raised his interest in the history of mathematics. He has written on Christopher Clavius and the 16th century renaissance of Euclidean geometry. The present study of David Rittenhouse comes from a history of mathematics course given at St. Joseph's University in Philadelphia's tricentennial year, 1982.

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 $11 included as part of the annual dues. (Annual dues for regular members, exclusive of annual subscription prices for MAA journals, are $22. Student, unemployed and emeritus members receive a 50% discount; new members receive a 30% dues discount for the first two years of membership.) The non-member/library subscription price is $28 per year. Bulk subscriptions (5 or more copies) are available to colleges and universities for classroom distribution to undergraduate students at a 41% discount ($6.50 per copy—minimum order $32.50).
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. 20077-9564. Back issues may be purchased, when in print, from P. and H. Bliss Company, Middletown, CT 06457. 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

Groups of Perfect Shuffles

Some questions are answered but many remain about the mathematics of card shuffling.

STEVE MEDVEDOFF
KENT MORRISON
California Polytechnic State University
San Luis Obispo, CA 93407

There are two ways to perfectly shuffle an ordinary deck of cards. First divide the deck in half and then interleaf the cards. The top card either remains on top or becomes the second card. A perfect shuffle is difficult but not impossible to perform. There are magicians who can execute a perfect shuffle and there are even a few who can do eight consecutive perfect shuffles—leaving the top card on top—to bring the deck back to its original position.

In 1983 a fascinating paper appeared dealing with the mathematics of perfect shuffles [**4**]. The work of Persi Diaconis, Ron Graham, and William Kantor completely determines the structure of the permutation groups generated by the two perfect shuffles of a deck containing an even number of cards. Incidentally, Diaconis was a professional magician before he became a mathematician-statistician and is able to perform eight perfect shuffles. Graham is also an amateur juggler. An interesting account of their work by Gina Kolata appeared in *Science* [**7**]. In this paper we will describe their results briefly but we will focus on problems that generalize theirs, problems that remain unsolved for the most part and problems that can be attacked in an experimental way by the tools of undergraduate algebra. We offer the subject of shuffle groups as a promising area in which to do exploratory group theory.

Shuffle groups

The mathematics of card shuffles has a long history and has been of most interest to magicians. There are card tricks based on mathematical principles rather than sleight-of-hand, or a combination of the two. One trick that children pass on to each other is the three-pile trick using 27 cards. The paper of Diaconis, Graham, and Kantor has a long section on the history of the mathematics of shuffles and there are two articles by Martin Gardner ([**5**], [**6**]) in his highly readable style.

What we mean by a **perfect shuffle** is a particular way of permuting the cards in a deck. We generalize the usual shuffle, in which the deck is divided into two piles, by allowing the deck to be divided into several equal piles. Then these piles are interleaved perfectly. For example, consider a deck of 33 cards. First divide the deck into three equal piles, the top, middle, and bottom piles each having eleven cards. Put these piles side by side in the order: top, middle, bottom. Next rearrange the piles in any of the six possible ways. Finally, pick up the cards from left to right, one at a time. The resulting arrangement is a perfect 3-shuffle (or ternary shuffle). There are six distinct 3-shuffles. If the cards in our deck are numbered $1,2,\dots,33$, then after dividing into piles, we envision them like this:

1	12	23
2	13	24
.	.	.
.	.	.
.	.	.
11	22	33

Now re-position the piles like this (one of six possibilities):

12	1	23
13	2	24
.	.	.
.	.	.
.	.	.
22	11	33

Next pick up the cards from left to right. The new order is $12, 1, 23, 13, 2, 24, \ldots, 22, 11, 33$. In this way we have six permutations of the numbers $1, \ldots, 33$ and we ask: *what subgroup of the symmetric group S_{33} do they generate*? Actually we have already asked this particular question and managed to answer it, but this is the sort of problem we are interested in. By the way, the answer is that they generate *all* of S_{33}, but it takes quite a bit of work to get the answer.

As you can see from the example, we have an infinite number of groups to study. If we want to divide our deck into k piles before shuffling, then the deck size must be a multiple of k, say kn. For positive integers k and n we generate a subgroup of the permutation group S_{kn}. The generators are the perfect k-shuffles, of which there are $k!$. We call this subgroup $G_{k,kn}$, and we would simply like to know what $G_{k,kn}$ looks like for all possible k and n. In our example with 33 cards we know $G_{3,33} = S_{33}$.

The 2-shuffles, or binary shuffles, are the usual shuffles that we attempt in order to mix up a deck of playing cards. It is the corresponding family of groups $G_{2,2n}$ that have been completely determined by Diaconis, Graham, and Kantor. Although we do not normally shuffle cards by dividing a deck into three or more piles, there are uses of k-shuffles in card tricks. The three pile trick using 27 cards involves the group $G_{3,27}$, which happens to be very much smaller than S_{27}.

Here is a brief summary of what is known about the shuffle groups:

(1) The binary shuffle groups $G_{2,2n}$ are all taken care of. There are five infinite families and two exceptional cases [**4**]. We will describe them later.
(2) We have determined G_{k,k^m}, and we will describe it in this paper.
(3) $G_{3,3n}$ is understood for deck sizes up to 63 (that is, for $n \leq 21$), and we have a solid conjecture for all n, a classification into three families.
(4) $G_{4,4n}$ is understood for decks up to 32 cards, and we have a conjecture for all n, a classification into four families.

We determined the structure of $G_{3,3n}$ and $G_{4,4n}$ for small values of n using the computer system for group theory called CAYLEY. The CAYLEY system was invaluable for concrete knowledge of these groups and gave us the data for the conjectured classification for $k = 3$ and 4. It was also a tremendous amount of fun to use. For access to the program and for help in using it, we would like to thank John Cannon, who has developed CAYLEY over the last twenty years, and Charles Sims and the Rutgers University Mathematics Department whose version of CAYLEY we used. CAYLEY is an immense system of hundreds of algorithms, in 250,000 lines of code, that is designed for the computational algebra of groups by generators and relations, permutation groups, finite fields and their polynomial rings, and linear algebra over finite fields. (CAYLEY is available from John Cannon, University of Sydney, Sydney, Australia, for a modest fee in both VAX and CYBER implementations. It is an expert system that works best with an expert's hand but is used, too, for laboratory work in undergraduate algebra courses.) We would also like to thank William Kantor and Martin Gardner for their interest in this work.

The fundamentals

Now it is time to go into the mathematics of the problem we have outlined. First we develop the notation. Generally, it is more convenient to number the cards beginning with 1 but sometimes it is better to begin with 0. We also number the piles from 1 to k (sometimes 0 to $k-1$). For each permutation σ in S_k, a permutation of the piles, we have the corresponding shuffle that we denote s_σ. The group $G_{k,kn}$ is generated by the elements s_σ for $\sigma \in S_k$. We use the convention that σ is a bijection of the set $\{1,\dots,k\}$ and that $\sigma(2)=3$ means that the second pile is moved to the third position. In cycle notation we would have $(2\ 3\dots)$ somewhere in the expression for σ. The shuffle s_σ can be written as the product of two operations

$$s_\sigma = p_\sigma s_I,$$

first performing the permutation p_σ followed by the shuffle s_I. Here we define p_σ to be the permutation of the deck that is accomplished by dividing it into piles as if to shuffle, permuting the piles according to σ, and then restacking the piles without interleaving with the leftmost pile on top. The shuffle s_I denotes the one in which the piles are not permuted. Thus p_σ permutes the piles and s_I does the interleaving. Notice that we write our operations left to right. Although it was not apparent before, now we see that p_σ is an element of $G_{k,kn}$. Furthermore, we see that we do not need $k!$ generators. All we need are enough elements of the form p_σ, so that the σ's generate S_k, along with the shuffle s_I. It is convenient to call s_I the **standard shuffle** and to denote it by s. We can generate $G_{k,kn}$ with three generators s, p_σ, p_τ, where σ and τ are generators for S_k. (It is easy to see that you can generate S_k with two generators. For example you may use $(1\ 2)$ and $(2\ 3\dots k)$.) When $k=2$, there are only two generators for $G_{2,2n}$.

It is true that $p_\sigma p_\tau = p_{\sigma\tau}$ since successive permutation of the piles is a permutation of the piles, but $s_\sigma s_\tau$ is not $s_{\sigma\tau}$. If we knew how to write $s_\sigma s_\tau$ in a nice way the whole problem would not be hard.

The first step is to determine the parity of our shuffles so that we know when $G_{k,kn}$ is contained in the alternating group A_{kn}.

LEMMA 1. *If n is odd and $\sigma \in S_k$ is an odd permutation, then p_σ is odd; otherwise p_σ is even.*

Proof. A permutation of two piles has the effect of interchanging n pairs of cards. Thus each transposition of σ results in n transpositions for p_σ.

LEMMA 2. *If either k or n is congruent to either* 0 *or* 1 (*mod* 4) *then s is even; otherwise s is odd.*

Proof. We will show that s can be written as $\frac{n(n-1)}{2}\frac{k(k-1)}{2}$ transpositions. Visualize the deck after cutting:

$$\begin{array}{ccccc} 1 & n+1 & 2n+1 & \cdots & (k-1)n+1 \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ n & 2n & 3n & \cdots & kn. \end{array}$$

Card 1 will stay put. Card 2 will have $k-1$ new cards in front of it after shuffling. Card 3 will have $2(k-1)$ new cards in front of it. Thus the cards in the first column will require $(k-1)+2(k-1)+\cdots+(n-1)(k-1)$ transpositions of adjacent cards to put them back on top. This sum is $(n(n-1)/2)(k-1)$. Now analyze the second pile in the same way. The number of adjacent transpositions required is $(n(n-1)/2)(k-2)$. For the rest of the piles we see that

$$\frac{n(n-1)}{2}[(k-1)+(k-2)+\cdots+2+1] = \frac{n(n-1)}{2}\frac{k(k-1)}{2}$$

adjacent transpositions are required to restore the deck to its original order.

Now we can say when the generators of $G_{k,kn}$ are all even permutations. Lemma 1 requires that n be even. Lemma 2 shows that if $n \equiv 0 \pmod 4$ then k can be anything, while if $n \equiv 2 \pmod 4$, k must be congruent to 0 or 1 (mod 4). This proves the following result on parity.

THEOREM 1. *If either of the following conditions holds, then $G_{k,kn}$ is a subgroup of A_{kn}:*
(i) $n \equiv 0 \pmod 4$
(ii) *n is even and $k \equiv 0$ or $1 \pmod 4$.*

Otherwise $G_{k,kn}$ contains an odd permutation.

COROLLARY. *If $n \equiv 0 \pmod 4$, then $G_{3,3n}$ is contained in A_{3n}.*

Lemma 2 and Theorem 1 are contained in [9] with different notation. Generically, that is for almost all k and n, we expect the parity theorem determines the structure of $G_{k,kn}$ as either S_{kn} or A_{kn}. The cases in which the group is not either of these are the cases of most interest.

We can determine the orders of two of the shuffles: s and s_{rev}. Here **rev** denotes the permutation "reverse" that reverses the order of the piles. Unfortunately we do not know what the orders of the other shuffles are.

PROPOSITION 1. *The order of s in $G_{k,kn}$ is the order of $k \pmod{kn-1}$, i.e., the smallest power of k congruent to $1 \pmod{kn-1}$, or equivalently, the order of k in the multiplicative group of units of the ring $\mathbb{Z}_{kn-1}$.*

Proof. Now it is convenient to number the cards from 0 to $kn-1$, because s fixes 0 and $kn-1$ and on the rest it acts by the rule $s\colon i \to ki \pmod{kn-1}$. Card 1 goes to position k, card 2 goes to position $2k$ and so on. This formula works for $i = 0$ but not for $i = kn-1$. Then the order of s is the smallest positive integer e such that $k^e i \equiv i \pmod{kn-1}$ for all i, but this is equivalent to $k^e \equiv 1 \pmod{kn-1}$.

PROPOSITION 2. *The order of s_{rev} is the order of $k \pmod{kn+1}$.*

Proof. This time number the cards from 1 to kn. After cutting, the cards are arranged in the pattern:

$$\begin{array}{cccc} 1 & n+1 & \cdot & \cdot & \cdot \\ 2 & n+2 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ n & 2n & \cdot & \cdot & kn. \end{array}$$

Now card i in row j and column p satisfies $i = (p-1)n + j$. The shuffle s_{rev} picks up the cards from right to left so that s_{rev} picks up all the cards in the rows above and the cards to the right of the same row before picking up a given card. That means the card in row j and column p will have $k(j-1)$ cards in the rows above it and $k-p$ cards to the right so it will be in position $k(j-1)+k-p+1$. A little arithmetic shows that this is congruent to $ki \pmod{kn+1}$. Thus $s_{\text{rev}}\colon i \to ki \pmod{kn+1}$ and it follows that the order of s_{rev} is as claimed.

This proof is not illuminating because we had to know the answer ahead of time. It is a little easier to see when $k = 2$ and we found that result in [4]. We simply tried out the obvious generalization for a couple of cases with $k = 3$ and 4. Convinced that it was true, we constructed a proof.

As we mentioned before, $G_{2,2n}$ has only two generators. Martin Gardner writes in [5] that Alex Elmsley, a magician, coined the terms "out-shuffle" and "in-shuffle." The out-shuffle is s_I and it leaves the first card on top or on the outside. The in-shuffle is $s_{(12)}$ and it puts the first card in the second position or inside. For $k > 2$, we continue to call s_I the out-shuffle and s_{rev} the in-shuffle, though there are other shuffles that leave the first card outside and that put the first card as deeply as possible inside. Elmsley used the letters O and I to stand for the two shuffles. He noticed in 1957 that a sequence of shuffles denoted by a sequence of O's and I's had the effect of bringing the top card down to the position whose binary expansion was the corresponding sequence of zeros and ones. You must number the cards $0,1,2,\dots$. This happy fact generalizes to k-shuffles. In [**6**] there is a description of the ternary representation for a deck of 27 cards. We have not found a published proof for the general case so we give one here.

PROPOSITION 3. *To bring the top card to position r, label the cards from 0 to $kn-1$, label the piles from 0 to $k-1$, and express r in base k as $r = d_m k^m + \cdots + d_1 k + d_0$, with $0 \le d_i < k$. For each i let τ_i be any permutation that transposes 0 and d_i. Then $s_{\tau_m} \cdots s_{\tau_1} s_{\tau_0}$ maps 0 to r.*

Proof. Let τ be any permutation that transposes 0 and d and is otherwise arbitrary. For card i where $0 \le i \le n-1$, we have $s_\tau(i) = ki + d$. Now the result follows by induction on m. Assuming that $s_{\tau_m} \cdots s_{\tau_1}$ has put the top card into position $i = d_m k^{m-1} + \cdots + d_2 k + d_1$, then $s_{\tau_0}(i) = d_m k^m + \cdots + d_1 k + d_0 = r$. Note that $i \le n-1$ since $r \le kn-1$, and so the rule applies to i.

From this it follows that the group $G_{k,kn}$ acts *transitively* on the set of cards. That means for any two cards i and j there is an element of the group that moves i to j.

Card tricks can be based on the algorithm of Proposition 3. A deck of kn cards is used from which the spectator draws a card. The card is replaced in the deck without letting the magician see it. The magician deals out k piles of n cards face up and asks the spectator to identify the pile containing the card. The magician gathers the cards and repeats the process. After several repetitions, the unknown card appears on the top of the deck. Dealing out the cards and stacking up the pile is the inverse of a shuffle, $(s_\sigma)^{-1}$, where σ is the permutation corresponding to the order in which the piles are picked up. The magician picks up the piles so that the pile with the card and the top pile have their positions interchanged. This procedure inverts the algorithm of Proposition 3 and brings the spectator's card to the top. The cards must be dealt out m times where m is the smallest integer such that $kn \le k^m$. For maximum effect for the same amount of work the magician should use a deck with k^m cards.

Deck size a power of k

An ancient trick going back centuries is the "three-pile trick" using 27 cards. In this version the mystery card appears in the middle—the thirteenth card—after three rounds. A generalization of this trick using m^m cards and an analysis of the trick were given by M. Gergonne in 1813. They are discussed in [**1**] and can be done with any values of k and n. However, the 27 card deck does seem to have a special fascination. It turns out that the shuffle groups G_{k,k^m}, where the deck size is a power of k, are quite special. They are quite small compared to the other shuffle groups because there is a lot of rigidity in the deck for these shuffles.

As a senior project at Cal Poly, one of us (Medvedoff) set out to find out anything he could about $G_{k,kn}$ for $k \ge 3$, only having seen [**7**], a brief account of the work of Diaconis, Graham, and Kantor on binary shuffle groups. Soon he noticed that there was something special about $G_{3,9}$ and $G_{3,27}$ and all the groups G_{k,k^m}. The main result in the senior project is the calculation of the order of G_{k,k^m}. It is $m(k!)^m$. After Medvedoff explained the results to his advisor (Morrison), both of us worked on the structure of the group. Without the CAYLEY program or any electronic computer, we found that a deck of cards—a primitive cellulose computer—was essential in understanding the group. In particular, we figured out $G_{3,27}$, whose order is 648, and then abstracted the key features. We strongly advise you to prepare a deck of 27 cards using the ace through nine of three suits. Arrange them from ace to nine in each suit and stack the suits. This is your initial order. Now start shuffling. Divide the deck into suits. Move the suits around and then pick up the cards from left to right. Keep the cards face up and pick up the cards so that the first one picked up is the top card. Doing this you will develop an understanding of $G_{3,27}$ so that you will know what the possible configurations are and how to arrive at them. Then you may finish reading this note.

We label the cards with m-tuples whose entries are the integers $1,2,\dots,k$ and arrange them 'lexicographically' so that the first card is $(1,1,\dots,1)$, the second is $(1,1,\dots,1,2)$, and the last card is $(k,k,\dots,k)$. After executing the shuffle s_σ the cards are arranged in the following order:

$$\begin{gathered}(\sigma^{-1}(1),1,\dots,1)\\ (\sigma^{-1}(2),1,\dots,1)\\ \vdots\\ (\sigma^{-1}(k),1,\dots,1)\\ (\sigma^{-1}(1),1,\dots,1,2)\\ \vdots\\ (\sigma^{-1}(k),1,\dots,1,2)\\ \vdots\\ (\sigma^{-1}(k),k,\dots,k).\end{gathered}$$

The card labeled $(i_1,\dots,i_m)$ is now in position $(i_2,\dots,i_m,\sigma(i_1))$. The sequence of shuffles $s_{\sigma_1}\cdots s_{\sigma_m}$ puts cards $(i_1,\dots,i_m)$ into position $(\sigma_1(i_1),\dots,\sigma_m(i_m))$. The elements of G of the form $s_{\sigma_1}\cdots s_{\sigma_m}$ make up a subgroup N of G that we will show to be normal. With 27 cards, the first digit determines the suit. After any three shuffles the suits are back together, although they may be moved around. To see that N is normal, we conjugate any element of N by any shuffle s_τ. We determine $s_\tau(s_{\sigma_1}s_{\sigma_2}\cdots s_{\sigma_m})s_\tau^{-1}$ by its effect on $(i_1,\dots,i_m)$. First s_τ sends it to $(i_2,\dots,i_m,\tau(i_1))$. Then the sequence of m shuffles sends that to $(\sigma_1(i_2),\sigma_2(i_3),\dots,\sigma_{m-1}(i_m),\sigma_m(\tau(i_1)))$. Then s_τ^{-1} maps to the $(\tau^{-1}(\sigma_m(\tau(i_1))),\sigma_1(i_2),\sigma_2(i_3),\dots,\sigma_{m-1}(i_m))$. Thus,

$$s_\tau(s_{\sigma_1}\cdots s_{\sigma_m})s_\tau^{-1}=s_{\tau\sigma_m\tau^{-1}}s_{\sigma_1}\cdots s_{\sigma_{m-1}}.$$

Notice that we write $\tau^{-1}\circ\sigma_m\circ\tau$ as $\tau\sigma_m\tau^{-1}$ because our convention is to write operations from left to right. This element is in N because it is a product of m shuffles.

Next, $G/N\cong\mathbb{Z}_m$ because any group element can be written as $s_{\sigma_1}\cdots s_{\sigma_m}s_I^e$, uniquely if $0\le e<m$. This is because s_I cyclically permutes $(i_1,\dots,i_m)$ by moving the first component to the end, and any sequence of shuffles maps $(i_1,\dots,i_m)$ to some cyclic permutation of $(\sigma_1(i_1),\dots,\sigma_m(i_m))$. We identify $\mathbb{Z}_m$ with the subgroup of G consisting of the powers of s_I, which has order m. The subgroup N is the product $(S_k)\times\cdots\times(S_k)$, m times. If we multiply $(s_{\sigma_1}\cdots s_{\sigma_m})(s_{\tau_1}\cdots s_{\tau_m})$, we get $s_{\sigma_1\tau_1}\cdots s_{\sigma_m\tau_m}$ by considering what happens to $(i_1,\dots,i_m)$.

We will describe the possible arrangements of the 27 card deck whose original order is A to 9 in the suits spades, hearts, clubs. The normal subgroup N is easy to describe, since an element of N leaves all cards in the same suit together. For example, the element $s_{(12)}s_{(23)}s_I$ acts as shown in FIGURE 1. Here, $s_{(12)}$ permutes the suits by interchanging the first—spades—with the second—hearts. Then $s_{(23)}$ permutes the subsuits of the suits, the second subsuit 4,5,6 is

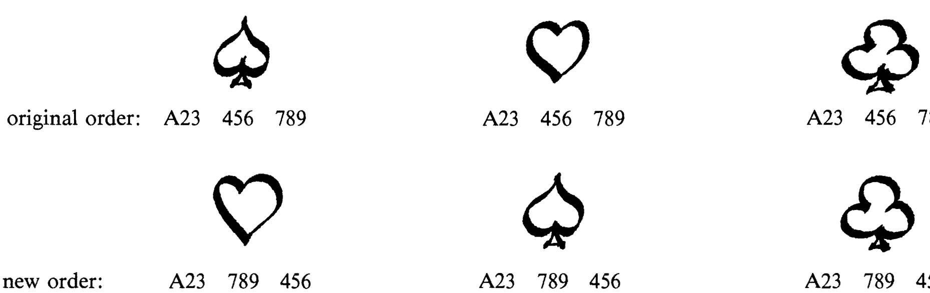


FIGURE 1.

interchanged with the third subsuit 7, 8, 9. This takes place in each of the suits. Finally, the last shuffle s_I leaves the cards in each subsuit alone. The arrangement within each suit is the same. An arbitrary element of the group is something in N followed by s_I^0, s_I or s_I^2. Of course, s_I^0 leaves well enough alone. A single power of s_I leaves the deck with the suits changing every card, while s_I^2 puts three cards of the same suit together. Then s_I^3 puts 9 cards of each suit together, which brings the deck back to N. To determine whether a given arrangement is in $G_{3,27}$ and what it is, first determine the power of s_I that is required. Then shuffle by s_I as many times as needed to reach N. You can recognize the element $s_{\sigma_1}s_{\sigma_2}s_{\sigma_3}$ in N by noting how the suits are permuted by σ_1, the subsuits by σ_2, and the individual cards in the subsuits by σ_3.

Now that we see $G = G_{3,27}$ has a normal subgroup $N = S_3 \times S_3 \times S_3$ and quotient group $K = \mathbb{Z}_3$, we have to put them together to get G. G is not the direct product of N and K, but rather a semidirect product. There are two standard ways to describe semidirect products, both useful. The first is that a group G with normal subgroup N and quotient $K = G/N$ is a **semidirect product** of N by K if there is a homomorphism i: $K \to G$ that identifies K with a subgroup of G and such that the composition $K \to G \to G/N = K$ is the identity map on K. In our example, the quotient G/N can be identified with the cyclic subgroup generated by the standard shuffle s_I. The second definition of a semidirect product begins with groups K and N and builds G. If we have a group homomorphism θ: $K \to \mathrm{Aut}(N)$ where $\mathrm{Aut}(N)$ is the group of automorphisms of N, then we define a group $G = N \times_\theta K$ called the semidirect product of N by K. As a set, G is the Cartesian product $N \times K$. We define the group operation $(n_1, k_1)(n_2, k_2) = (n_1\theta(k_1)(n_2), k_1k_2)$. If the homomorphism θ is not explicitly given, the notation $N \rtimes K$ is also used, emphasizing that N is the normal subgroup. Identify N and K with the subgroups $N \times \{1\}$ and $K \times \{1\}$. Then N will be normal and $\theta(k)$ will correspond to conjugation of $(1, k)$ on $N \times \{1\}$.

Now the action θ of $K = \mathbb{Z}_m$ on $N = S_k \times \cdots \times S_k$ is easy to describe. The generator s_I of $\mathbb{Z}_m$ conjugates an element of N, say $s_{\sigma_1} \cdots s_{\sigma_m}$, by

$$s_I(s_{\sigma_1} \cdots s_{\sigma_m})s_I^{-1} = s_{\sigma_m}s_{\sigma_1} \cdots s_{\sigma_{m-1}}.$$

So we have our theorem.

THEOREM 2. *The shuffle group G_{k,k^m} is the semidirect product of $S_k \times \cdots \times S_k$, m factors, by $\mathbb{Z}_m$ acting by cyclic permutation of the factors. In particular, the generator of $\mathbb{Z}_m$ corresponding to s_I permutes $S_k \times \cdots \times S_k$ by $(\sigma_1, \dots, \sigma_m) \mapsto (\sigma_m, \sigma_1, \dots, \sigma_{m-1})$. The order of the group is $m(k!)^m$.*

Additional comment. This semidirect product is an example of a **wreath product**. The binary shuffle groups are wreath products or very close to wreath products. The symmetry group of Rubik's cube is a wreath product, too. A wreath product is constructed from a group H and a permutation group $P \subset S_m$ by taking the semidirect product of $H \times \cdots \times H$, m factors, with P acting by permuting the factors. Thus we can say G_{k,k^m} is the wreath product of S_k with $\mathbb{Z}_m$.

The binary shuffle groups

To understand $G_{2,2n}$ there is an important symmetry principle called **central symmetry**. Number the cards $1, 2, \dots, n-1, n, n', (n-1)', \dots, 2', 1'$. After shuffling, the order is:

out shuffle: $1, n', 2, (n-1)', \dots, n-1, 2', n, 1'$.
in shuffle: $n', 1, (n-1)', 2, \dots, 2', n-1, 1', n$.

The centrally symmetric pair $\{i, i'\}$ is now in the pair of positions $\{j, j'\}$ for some j. So the cards i and i' that are the same distance from the center remain the same distance from the center. Thus the shuffle group $G_{2,2n}$ is a subgroup of the group $B_n \subset S_{2n}$ consisting of centrally symmetric permutations. B_n is the Weyl group of a simple Lie algebra and has other descriptions. It is the symmetry group of the n-dimensional cube with vertices $\pm e_1, \dots, \pm e_n$. A symmetry must map the pair of vertices $\pm e_i$ to a pair of vertices $\pm e_j$ for some j. Such a linear map on $\mathbb{R}^n$ is represented by a signed permutation matrix: each row and column has only one nonzero entry and that entry is 1 or -1. There is a surjective homomorphism $B_n \twoheadrightarrow S_n$ that forgets the signs in the permutation matrix. With our card deck it means we only keep track of the induced permutation on the set of n symmetric pairs. Therefore, we can consider the parity of the induced permutation in S_n as well as the parity of the permutation in S_{2n}. We have group homomorphisms sgn and $\overline{\text{sgn}}$ from B_n to the group $\{\pm 1\}$:

$$\text{sgn}: B_n \hookrightarrow S_{2n} \to \{\pm 1\} \quad \overline{\text{sgn}}: B_n \twoheadrightarrow S_n \to \{\pm 1\}.$$

We also have the product of these homomorphisms, $\text{sgn}\,\overline{\text{sgn}}$: $B_n \to \{\pm 1\}$, which is a group homomorphism. The binary shuffle groups $G_{2,2n}$ are given by B_n and various kernels of these three homomorphisms. They consist of five families and two special cases.

(0) If $n \equiv 0 \pmod 4$, $n > 12$ and not a power of 2, then $G = \operatorname{Ker} \text{sgn} \cap \operatorname{Ker} \overline{\text{sgn}}$.
(1) If $n \equiv 1 \pmod 4$, then $G = \operatorname{Ker} \overline{\text{sgn}}$.
(2) If $n \equiv 2 \pmod 4$, then $G = B_n$.
(3) If $n \equiv 3 \pmod 4$, then $G = \operatorname{Ker} \text{sgn}\,\overline{\text{sgn}}$.
(4) If $2n = 2^m$ then $G = (\mathbb{Z}_2)^m \times_\theta \mathbb{Z}_m$, where $\mathbb{Z}_m$ acts cyclically on the factors.
(5) The two anomalous cases
  (i) If $n = 6$, then G is the semidirect product $(\mathbb{Z}_2)^6 \rtimes \text{PGL}(2,5)$. See [**4**] for a description of the action.
  (ii) If $n = 12$, then G is the semidirect product of $(\mathbb{Z}_2)^{11}$ by the Mathieu group M_{12}. The group M_{12} is a subgroup of S_{12} and so it acts naturally on $(\mathbb{Z}_2)^{12}$ by permuting the factors. This action restricts to the subspace of vectors whose components sum to zero. The subspace is 11-dimensional, hence isomorphic to $(\mathbb{Z}_2)^{11}$.

For the ordinary deck of 52 cards, n is 26 and $26 \equiv 2 \pmod 4$. The shuffle group is all of B_{26}. The only restriction on possible configurations is that of central symmetry. We can count the order directly. There are 52 places for the first card, but then the last one must go to a determined location. There are 50 remaining places for the second card, and so on. There are $52 \cdot 50 \cdot 48 \cdots 4 \cdot 2$ arrangements possible. This is $2^{26}(26)!$, a large number, but a small fraction of $(52)!$.

Ternary shuffle group

TABLE 1, compiled with the aid of CAYLEY, points to an obvious conjecture about $G_{3,3n}$. The

$3n$	$G_{3,3n}$	$3n$	$G_{3,3n}$
3	S_3	39	S_{33}
6	S_6	42	S_{42}
9	$(S_3)^2 \rtimes \mathbb{Z}_2$	45	S_{45}
12	A_{12}	48	A_{48}
15	S_{15}	51	S_{51}
18	S_{18}	54	S_{54}
21	S_{21}	57	S_{57}
24	A_{24}	60	A_{60}
27	$(S_3)^3 \rtimes \mathbb{Z}_3$	63	S_{63}
30	S_{30}		
33	S_{33}		
36	A_{36}		

TABLE 1

evidence shows that $G_{3,3n}$ is as large as possible subject to the parity theorem as long as n is not a power of 3.

CONJECTURE. *The classification of* $G_{3,3n}$ *is given by three families*:
(1) *If* n *is a multiple of* 4, *then* $G_{3,3n} = A_{3n}$.
(2) *If* n *is not a multiple of* 4 *and not a power of* 3, *then* $G_{3,3n} = S_{3n}$.
(3) *If* $3n = 3^m$, *then* $G_{3,3n} = (S_3)^m \rtimes \mathbb{Z}_3$.

Part (3) has been proved, but we include it for a complete statement.

We have been able to verify the computer results by hand in any of the individual cases, but we have not been able to prove the conjecture. Each case looks a little different but the strategy is the same. We will present a proof that $G_{3,21} = S_{21}$ to illustrate how you can verify the results.

For generators, we use $s = s_I$, p, and q, where p permutes the first and second piles of 7 cards and q cyclically permutes all three piles. Notice that p and q are not shuffles but permutations that we earlier called p_σ. Number the cards from 0 to 20 and recall that s is multiplication by 3 modulo 20 and s fixes 20. The cycle forms of the generators are:

$$s = (1,3,9,7)(2,6,18,14)(4,12,16,8)(5,15)(11,13,19,17)$$

$$p = (7,14)(8,15)(9,16)(10,17)(11,18)(12,19)(13,20)$$

$$q = (0,7,14)(1,8,15)(2,9,16)(3,10,17)(4,11,18)(5,12,19)(6,13,20).$$

Let $G = G_{3,21}$; G is transitive by Proposition 3. Now s and p both fix the top card 0 so they lie in the stabilizer subgroup of 0. The subgroup generated by s and p permutes $1,2,\dots,20$ and by looking at the cycles of s and p we can see that this subgroup is transitive on $1,2,\dots,20$. Thus G is **doubly transitive,** meaning that any pair (i_1, i_2) can be mapped to any other pair (j_1, j_2) by some $g \in G$: $g(i_1) = j_1$ and $g(i_2) = j_2$. A doubly transitive group is **primitive**, meaning that there is no way to partition the deck into subsets of equal size $X_1, \dots, X_r$ (other than the singleton subsets and the whole set) so that the permutations map each X_i to some X_j. (It may help to note that the binary shuffle groups are not primitive because the deck partitions into the central symmetric pairs which have this property.) A classical theorem of Jordan states that a primitive group containing a transposition is the symmetric group and one containing a 3-cycle is at least the alternating group [**11**]. We compute sq and find

$$sq = (0,7,8,11,20,6,4,19,3,16,15,12,2,13,5,1,10,17,18)(9,14).$$

The long cycle has length 19, so that $(sq)^{19} = (9,14)$. Hence G contains a transposition and by Jordan's theorem, $G = S_{21}$.

The strategy that works on these particular groups is to show that G is doubly transitive, hence primitive, and to exhibit a transposition or a 3-cycle, which is found by experimentation. In general, to show that $G_{3,3n}$ is doubly transitive it would suffice to show that the subgroup generated by s and $p_{(2,3)}$ is transitive on $\{1,\dots,3n-1\}$ as both of them fix 0. Here $p_{(2,3)}$ is the permutation of piles 2 and 3, where the piles are numbered $1,2,3$. For $n \le 21$, we know that this subgroup, denoted by $\langle s, p_{(2,3)} \rangle$, is transitive on $\{1,\dots,3n-1\}$, having checked those cases with CAYLEY. In fact, in these cases $\langle s, p_{(2,3)} \rangle$ is either the alternating or the symmetric group, A_{3n-1} or S_{3n-1}, the alternating group occurring when n is a multiple of 4.

There are two ways to proceed to get a full proof of the structure conjecture.

(1) Show $\langle s, p_{(2,3)} \rangle$ is transitive and then show that $G_{3,3n}$ always contains a 3-cycle.

(2) Show $\langle s, p_{(2,3)} \rangle$ is A_{3n-1} or S_{3n-1}. From this it follows that $G_{3,3n}$ is A_{3n} or S_{3n}.

$k = 4$ and beyond

What happens for $k \geq 4$? We believe that, generically, $G_{k,kn}$ is A_{kn} or S_{kn} according to Theorem 1, but not so generically as the case of $G_{3,3n}$ indicates. We found that $G_{4,8}$ is not the expected A_8 and $G_{4,32}$ is not the expected A_{32}. The order of $G_{4,8}$ is 1344, whereas $|A_8| = 20{,}160$, and the group is isomorphic to the semidirect product of $(\mathbb{Z}_2)^3$ by GL(3,2), the group of invertible 3×3 matrices over the field $\mathbb{Z}_2$, which acts linearly on the vector space $(\mathbb{Z}_2)^3$. This is just the affine group of the 3-dimensional vector space V over $\mathbb{Z}_2$ consisting of the maps f: $V \to V$ of the form $f(x) = \varphi(v) + v_0$ for an invertible linear map φ and with $v_0 \in V$. GL(3,2) is also the same as PGL(3,2) or PL(3,2) since the general linear groups and the projective (general) linear groups are the same over fields of characteristic 2. GL(3,2) is a simple group of order 168, the smallest nonabelian simple group larger than A_5 of order 60. Likewise, $G_{4,32}$ is the affine group of the 5-dimensional vector space over $\mathbb{Z}_2$, so $G_{4,32} = (\mathbb{Z}_2)^5 \rtimes \mathrm{GL}(5,2)$. Both of these results come from using CAYLEY and we have only verified $G_{4,8}$ by hand. Our data are summarized in TABLE 2.

$4n$	$G_{4,4n}$
8	$\mathbb{Z}_2^3 \rtimes GL(3,2)$
12	S_{12}
16	$(S_4)^2 \rtimes \mathbb{Z}_2$
20	S_{20}
24	A_{24}
28	S_{28}
32	$\mathbb{Z}_2^5 \rtimes GL(5,2)$

TABLE 2

CONJECTURE. *There are four families classifying* $G_{4,4n}$.

(1) *If n is a power of* 4, *let* $4n = 4^m$. *Then* $G_{4,4n}$ *is* $(S_4)^m \rtimes \mathbb{Z}_m$.

(2) *If n is an odd power of* 2, *let* $4n = 2^{2m+1}$. *Then* $G_{4,4n}$ *is* $(\mathbb{Z}_2)^{2m+1} \rtimes GL(2m+1,2)$.

(3) *If n is even and not a power of* 2, *then* $G_{4,4n} = A_{4n}$.

(4) *If n is odd, then* $G_{4,4n} = S_{4n}$.

Part (1) is proved already. Part (2) is an obvious target to attack, but to understand $G_{4,8}$ using pencil, paper, and cards is tedious. The eight cards correspond to the eight points in the 3-dimensional vector spaces $(\mathbb{Z}_2)^3$, the top card being the origin $(0,0,0)$. We express each of the three generators as an affine map so that we know $G_{4,8}$ is a subgroup of the affine group. This is straightforward but does not suggest a way of generalizing to 2^{2m+1} cards. Finally we check that each of the eight translations is in $G_{4,8}$ and that we can fix the top card and map the three basis cards $(0,0,1)$, $(0,1,0)$, and $(1,0,0)$ to any three linearly independent positions. For the last part, working with a deck of eight cards has been convincing and we have to admit we have not written out the details. There must, however, be a better way to do it. You do not want to go on to $G_{4,32}$ and the larger groups in the same way.

Conjecturing that $G_{9,27}$ might be something along the line of $G_{4,8}$, we checked it with CAYLEY but found that $G_{9,27} = S_{27}$ and was not an affine group over the field $\mathbb{Z}_3$. Some shuffle groups for $k \geq 5$ are in TABLE 3.

k	kn	$G_{k,kn}$
5	10	A_{10}
5	15	S_{15}
5	20	A_{20}
6	12	S_{12}
6	18	S_{18}
7	14	S_{14}
8	16	A_{16}
9	27	S_{27}

TABLE 3

William Kantor reports that he has shown that $G_{k,kn}$ is A_{kn} or S_{kn} when $k \geq 4$ and k does not divide n. We have not seen his argument. This leaves open the case $k = 3$. It also suggests that when k^2 divides the deck size as in $G_{4,8}$ or $G_{9,27}$ you must be more careful.

S. B. Morris, in his 1974 thesis **[8]**, and together with R. E. Hartwig in **[9]**, consider generalized shuffles of decks of size $kn + m$, $0 \leq m < n$. Shuffle by dividing the deck into k piles, m of them having $n + 1$ cards and the rest having n cards. They consider the out-shuffle or "generalized faro shuffle" and the permutation called the "simple cut" that moves the top card to the bottom of the deck. They determine when the group generated by these two permutations is the symmetric group or the alternating group. They also determine the order of the generalized in-shuffle. (We gave a proof in Proposition 2 for $m = 0$.)

M. Davio and C. Ronse, both of the Philips Research Laboratories in Brussels, have generalized the shuffles using a mixed-radix formalism (mixed-base notation). Their generalizations are in a different direction than ours but could fit into a common generalization of the notion of shuffling. They are interested in shuffles for their application to problems in parallel processing and switching network design. We refer the interested reader to **[3]** and **[10]**. Incidentally, the structure of $G_{2,2^m}$ is of interest in the computational scheme of the Fast Fourier Transform. See **[4]** for a description.

One last question

You probably have noticed the big difference between the binary shuffle groups, $k = 2$, and the groups with $k \geq 3$. We cannot resist mentioning that the groups generated by the out-shuffle s_I and the in-shuffle s_{rev} may be more like the binary shuffle groups which have just those generators. Let $H_{k,kn} \subset G_{k,kn}$ be the subgroup generated by s_I and s_{rev}. Our final question is: *What is the structure of $H_{k,kn}$ for all k and n*? Of course, for $k = 2$ we know the answer since $H_{2,2n} = G_{2,2n}$, but for $k \geq 3$ this is a much more difficult question. For $H_{3,3n}$ we have data from CAYLEY up to 69 cards. There is some tantalizing regularity but not enough. The whole situation is much more complicated than the case $k = 2$ or the case $G_{3,3n}$. One intriguing result is that $H_{3,24} = H_{2,24}$, which is a semidirect product of $(\mathbb{Z}_2)^{11}$ by the Mathieu group M_{12}. We have also heard indirectly that $H_{d,24} = H_{2,24}$ for any d a proper divisor of 24. The groups $H_{k,kn}$ have central symmetry, the key feature in the binary shuffles, and so it is not surprising they bear so much resemblance.

References

[1] W. W. Rouse Ball, H. S. M. Coxeter, Mathematical Recreations and Essays, 12th ed., University of Toronto Press, Toronto, 1974.

[2] J. J. Cannon, Description of CAYLEY, preprint, Univ. of Sydney.

[3] M. Davio, Kronecker products and shuffle algebra, IEEE Trans. Computers, C-30 (1981) 116–125.

[4] P. Diaconis, R. L. Graham, W. M. Kantor, The mathematics of perfect shuffles, Adv. Appl. Math., 4 (1983) 175–196.

[5] M. Gardner, Mathematical Carnival, Vintage, New York, 1977.
[6] ______, Mathematics, Magic, and Mystery, Dover, New York, 1956.
[7] G. Kolata, Perfect shuffles and their relation to math, Science, 216 (1982) 505–506.
[8] S. B. Morris, Permutations by cutting and shuffling—a generalization to q-dimensions, Thesis, Duke University, 1974.
[9] S. B. Morris, R. E. Hartwig, The generalized faro shuffle, Discrete Math., 15 (1976) 333–346.
[10] C. Ronse, A generalization of the perfect shuffle, Discrete Math., 47 (1983) 293–306.
[11] H. Wielandt, Finite Permutation Groups, trans. R. Bercov, Academic Press, New York, 1964.

"Such a really remarkable discovery. I wanted your opinion on it. About God. You know the formula: m over nought equals infinity, m being any positive number? Well, why not reduce the equation to a simpler form by multiplying both sides by nought? In which case, you have m equals infinity times nought. That is to say that a positive number is the product of zero and infinity. Doesn't that demonstrate the creation of the universe by an infinite power out of nothing? Doesn't it?"...

"'Well,' began Lord Edward, at the other end of the electrified wire, forty miles away, his brother knew, from the tone in which the single word was spoken, that it was no good. The Absolute's tail was still unsalted."

—ALDOUS HUXLEY, *Point Counter Point,*
Chapter XI

David Rittenhouse: Logarithms and Leisure

FREDERICK A. HOMANN
St. Joseph's University
Philadelphia, PA 19131

The life and times of David Rittenhouse (1732–1796) suggest that leisure and the love of calculation united in early America to stimulate mathematical practice and start, ever so obscurely, our research tradition. Of all the Colonial and Revolutionary natural philosophers, Rittenhouse was the most adept at calculation and mathematics even though his strength and accomplishments lay more in astronomy and the manufacture of precision scientific instruments [**2**], [**3**], [**4**]. Self-taught in mathematics, Rittenhouse read with care Newton's *Principia Mathematica*. "Some intricate calculation or other, always takes up my idle hours," he wrote [**1**, p. 221], [**4**, III]. William Barton, his early biographer, remarked that "he considered all his hours as 'idle' ones which were not occupied in some manual employment" [**1**, p. 221]. François-Jean Marquis de Chastellux (1734–1788), member of the Académie Française and American military ally in the Revolution, observed that Rittenhouse was "not a mathematician of the class of the Eulers, and the D'Alemberts" (almost a self-evident truth, Brooke Hindle notes), yet he had many American admirers, of whom Thomas Jefferson was the foremost [**3**, p. 344]. W. C. Rufus summarizes Rittenhouse's papers in [**9**]. Rooted in problems of astronomy and measurement of time, they reflect the inchoate, isolated state of mathematics in Colonial and Federal America and the work of good, if undeveloped pragmatic mathematical talent. Rittenhouse's paper of 12 August 1795: "Method of raising the common Logarithm of any Number immediately," read when he was president of the American Philosophical Society, is a good case in point (see [**4**], [**6**]). Printed posthumously in 1799, it approximated in continued fraction form the base-10 logarithm of a positive number. Logarithms were of great use in Colonial surveying; Laplace claimed they doubled the life of the astronomer.

Rittenhouse's method in his own spare words is this:

> The logarithm of any number is the index of that power of 10 which is equal to the given number. This index will always be fractional, unless the number be divisible by 10 without any remainder.
>
> If the number be greater than 10, divide it by the highest power of 10 that will leave the quotient not less than 1. The index of that power is the first figure, or index of the logarithm. Divide 10 by the quotient so found raised to the highest power that will leave the new quotient not less than unity. Divide the last divisor by the last quotient raised to its proper power, and proceed in this manner until a sufficient number of divisions are made, which will be when the quotient approaches nearly to unity. Make a compound fraction, taking the successive indexes of the powers you divide by for denominators and unity for numerators. Reduce this compound fraction to a simple one, and that by division to a decimal fraction, which together with the index first found (if any) will be the logarithm required. [**6**].

As an example, Rittenhouse then computed log 99 to nine decimal places.

To unravel his text, we can work with $N = 20 = 2 \cdot 10^1$, and then formalize his algorithm with notation we will use in this example.

Clearly, the "index of the logarithm" (or characteristic) of 20 is 1. Set the quotient $20/10 = 2 \equiv Q_0 = 10^{1/r_0}$, where we want to write $1/r_0$ (the log of 2) as a continued fraction. Set $10 = Q_{-1} = 2^{r_0}$. Then $r_0 = n_0 + 1/r_1$, where $n_0 = 3$ is the largest integer ("its proper power") such that the new quotient $Q_1 \equiv Q_{-1}/Q_0^{n_0} = 10/2^{n_0} \geqslant 1$ (that is, "not less than unity"). For $n_0 = 3$, $Q_1 = 1.2500$. Set $1.2500 = 2^{1/r_1}$. Then $r_1 = n_1 + 1/r_2$, where $n_1 = 3$ is the largest integer such that the next quotient $Q_2 \equiv Q_0/Q_1^{n_1} = 2/(1.2500)^{n_1} \geqslant 1$. For $n_1 = 3$, $Q_2 = 1.0240$. So we have $1/r_0 = 1/(3 + 1/(3 + 1/r_2))$. A third iteration of the process gives $r_2 = n_2 + 1/r_3$, where $n_2 = 9$ is the largest integer such that $Q_3 \equiv Q_1/Q_2^{n_2} \geqslant 1$. For $n_2 = 9$, $Q_3 = 1.0097$. The quotients Q_0, Q_1, Q_2, Q_3 "approach nearly to unity." Let $E_0 = 1/n_0$, $E_1 = 1/(n_0 + 1/n_1)$, $E_2 = 1/(n_0 + 1/(n_1 + 1/n_2))$. Then $E_0 = 1/3 = .33333\ldots$, $E_1 = 3/10 = .30000$, $E_2 = 28/93 = .30107$. We get $\log 20 \doteq 1.30107$.

So, to approximate $\log N$, where $N = 10^I M$, I a nonnegative integer (the "index of the logarithm" of N), and $1 \leq M < 10$, Rittenhouse set up (in our terms and notation) the sequences $\{Q_k\}_0^\infty, \{n_k\}_0^\infty, \{E_k\}_0^\infty$. He put

$$Q_{-1} = 10, \qquad Q_0 = M \geqslant 1, \qquad Q_{k+1} = Q_{k-1}/Q_k^{n_k} \geqslant 1, \qquad k = 0,1,2,3,\ldots,$$

where n_k is chosen as the largest integer such that $Q_{k+1} \geqslant 1$. He put

$$E_k = \cfrac{1}{n_0 + \cfrac{1}{n_1 + \cfrac{1}{n_2 + \ddots + \cfrac{1}{n_k}}}}.$$

If N is a rational power of 10, $Q_{k+1} = 1$ after a finite number of steps and the process ends, producing the exact value of $\log N$. It is easy to show from the definitions of Q_k and n_k that $1 \leqslant Q_{k+1} < Q_k$, and that $\lim_{k\to\infty} Q_k = 1$. Also, as is characteristic of continued fractions,

$$10^{E_{2q+1}} \leqslant Q_0 = M \leqslant 10^{E_{2q}},$$

that is,

$$E_{2q+1} \leqslant \log M \leqslant E_{2q}. \tag{1}$$

Equality obtains if the process has stopped. We see easily that $\lim_{q\to\infty}(E_{2q} - E_{2q+1}) = 0$, or $\lim_{k\to\infty} E_k = \log M$, and $\log N = I + \log M$.

Accuracy of these approximations E_k comes in terms of the Q_k ("...until a sufficient number of divisions are made, which will be when the quotient $[Q_k]$ approaches nearly to unity"). If we use the recursive definition of Q_k to express $\log M$ in terms of $\log Q_k$ we get, for example,

$$0 \leqslant E_2 - \log M = \log Q_3/(n_0 + n_2 + n_0 n_1 n_2) \leqslant (1/3)\log Q_3,$$
$$0 \leqslant \log M - E_3 = \log Q_4/(1 + n_0 n_1 + n_2 n_3 + n_0 n_1 n_2 n_3)$$
$$\leqslant (1/5)\log Q_4.$$

Rittenhouse's algorithm is tedious in practice without a computer. It differs markedly from earlier (1688) infinite series work of Mercator and others (for the natural logarithm) [**11**, pp. 129–137]. Rittenhouse left no hint how he came upon it, nor did he note the characteristic property (1). His result, however, was both anticipated and repeated. Brook Taylor (1685–1731) had already published a form of this procedure in 1717 [**13**]. Much later, in 1954, Daniel Shanks at the Naval Ordnance Laboratory, independently of both Taylor and Rittenhouse, discovered the algorithm in a general form for any base, and published it "because of its mathematical beauty and its adaptability to high speed computing machines." Shanks showed how it could be programmed, and, with some results of A. Khintchine, proved that its "rate of convergence... is, on the average, about one decimal place per complete cycle" (that is, for each Operation II of his program) [**10**].

An earlier Rittenhouse paper had a similar history. "A Method of Finding the Sum of Several Powers of the Sines" dealt with a pendulum clock problem, and, for positive integral n, correctly evaluated (in our terms and notation) the integrals $\int_0^{\pi/2} \sin^n x\,dx$ [**4**], [**6**]. These values had already been determined by Wallis (1655) and Pascal (1659) [**11**], [**12**]. Although he gave no explicit details of his procedure, Rittenhouse claimed he could prove his results for $n = 1,2$, and asked

N°. IX.

Method of raising the common Logarithm of any Number immediately, by David Rittenhouse, *President of the Society.*

Read Aug. 12, 1795. THE logarithm of any number is the index of that power of 10 which is equal to the given number. This index will always be fractional, unless the number be divisible by 10 without any remainder.

If the number be greater than 10, divide it by the highest power of 10 that will leave the quotient not less than 1. The index of that power is the first figure, or index of the logarithm. Divide 10 by the quotient so found raised to the highest power that will leave the new quotient not less than unity. Divide

* *Note.* When U exceeds 90° take its supplement and in that case deduct the result of the calculation from two right angles, and the remainder will be the true anomaly.

K the

Rittenhouse's paper of 1799.

the laſt diviſor by the laſt quotient raiſed to its proper power, and proceed in this manner until a ſufficient number of diviſions are made, which will be when the quotient approaches nearly to unity. Make a compound fraction, taking the ſucceſſive indexes of the powers you divide by for denominators and unity for numerators. Reduce this compound fraction to a ſimple one, and that by diviſion to a decimal fraction, which together with the index firſt found (if any) will be the logarithm required.

Example of the Calculation.

Required the Logarithm of 99.

Divided by $\frac{99}{10^1} = 9.9$. Here 1 is the index.

Divided by $\frac{10}{9.9^1} = 1.010101 = a$.

a, raiſed to its higheſt power, 228.

$aa = 1.020304$	Firſt quotient, 9.9	$= b = 1.001059$
20406	Divided by $a^{228} = 9.889521$	1001
306	.010479	59
4	9889	$b^2 = 1.002119$
$a^4 = 1.041020$	590	2004
41641	494	119
1041	96	$b^4 = 1.004242$
21	89	4[illegible]27
$a^8 = 1.083723$	7	201
86698		42
3251		$b^8 = 1.008502$
759	× by b	1009
22		59
3		$b^9 = 1.009570$
$a^{16} = 1.174456$	$a = 1.010101$	$= c = 1.000526$
117446	Divided by $b^9 = 1.009570$	526
82212	.000531	$c^2 = 1.001052$
4698	505	
470	26	
59		
7		
$a^{32} = 1.379348$		

Calculation

Calculation Continued.

$a^{32} = 1.379348$
413804
96554
12414
415
55
11

$a^{64} = 1.902600$
1.712340
3805
1141

$a^{128} = 3.619886$
× by a^{64} 3.257897
7240
2172

$a^{192} = 6.887195$
× by a^{32} 2.066158
482104
61985
2066
275
55

$a^{224} = 9.499838$
× by a^{4} 379993
9500
190

$a^{228} = 9.889521$

Divided by $\left.\begin{matrix} b = 1.001059 \\ c^{2} = 1.001052 \end{matrix}\right\} = d = 1.000007$

7

The quotient d, is now ſo ſmall that it is not neceſſary to proceed further in this way, for the decimals of c, divided by the decimals of d will give the power required, viz. 75.

Making a compound fraction, as before directed, with the ſeveral powers of the diviſors in the order they ſtand we have.

$$\cfrac{1}{1+\cfrac{1}{228+\cfrac{1}{9+\cfrac{1}{2+\cfrac{1}{75}}}}}$$

Which reduced 75

Gives the 151 = 75 × 2 + 1

1434 = 151 × 9 + 75

327103 = 1434 × 228 + 151

Simple fraction = $\frac{327103}{328537}$ = 327103 × 1 + 1434

Denominator \ Numerator /
328537) 327103,0 (.995635194.8. The decimal part of the log. of 99. true to the 9th place, and 3 too much in the 10th.

2956833
3141970
2956833
1851870
1642685
2086850
1971222
1156280
985611
170669
164268
6400
3285
3115
2957
158
131
27

K 2 *Experiments*

Robert Patterson, professor at the University of Pennsylvania, to supply proof for all cases. Patterson's response is not known, but in 1803 Nathaniel Bowditch submitted to the American Philosophical Society a paper (unpublished) in response to Rittenhouse, and in 1814 Owen Nulty, a mathematics professor at Dickinson College, gave a complete treatment with standard integral notation in a letter read to the Philosophical Society 15 November 1816, and published in 1818 [7], [8]. Can we see in these events the beginning of the Problem Sections of our contemporary journals?

Rittenhouse the mathematician would not have been surprised by the antecedent work of Taylor, Wallis, and Pascal. He would have no doubt been delighted by the efforts of Bowditch, Nulty, and Shanks. Rittenhouse the instrument maker would certainly have used our computers to increase his "idle" time for his "intricate calculations" that were the start of a prolific American mathematical tradition.

References

[1] William Barton, Memoirs of the Life of David Rittenhouse, Philadelphia, Pa., 1813.

[2] Brooke Hindle, David Rittenhouse, Princeton University Press, Princeton, NJ, 1964.

[3] ______, The Pursuit of Science in Revolutionary America, 1735–1789, Norton, New York, 1974.

[4] ______, (editor), The Scientific Writings of David Rittenhouse, Arno, New York, 1980.

[5] David Rittenhouse, A method of finding the sum of the several powers of the sines, Trans. Am. Phil. Soc., 3 (1793) 155–156.

[6] ______, Method of raising the common logarithm of any number immediately, Trans. Am. Phil. Soc., 4 (1799) 69–71.

[7] Owen Nulty, Complete solution of Rittenhouse's problem, Trans. Am. Phil. Soc., N.S. 1 (1818) 395–400.

[8] Proc. Am. Phil. Soc., 22 (1885) pt. 3, Minutes 1743–1838, p. 330.

[9] W. Carl Rufus, David Rittenhouse as a mathematical disciple of Newton, Scripta Math., 8 (1941) 228–231.

[10] Daniel Shanks, A logarithm algorithm, Math. Tab. Aid. Comp., 8 (1954) 60–64.

[11] J. F. Scott, A History of Mathematics, Barnes and Noble, New York, 1975.

[12] D. J. Struik (ed.), A Source Book in Mathematics, 1200–1800, Harvard Univ. Pr., Cambridge, MA, 1969.

[13] Brook Taylor, A new method of computing logarithms, Philos. Trans. Roy. Soc. London, XXX (1717) 618–622.

Perturbation of Markov Chains

CHRISTINE BURNLEY
95 Horatio Street
New York, NY 10014

A simple fair game can be played by two people who have an equal number of pennies: each flips one coin, and, if the two flipped coins match each other (heads or tails), Player 1 wins both coins; otherwise, Player 2 wins both. The game ends when one of the two players has all the pennies. Clearly each has an equal chance of winning if both start with the same number of pennies, but what if one player has more to begin with? Or what if one player somehow cheats whenever he has only one penny left?

In order to consider these questions, it is useful to model the game with a Markov chain [**2**]. If the total wealth between the two players is n pennies, one can build an $(n+1)\times(n+1)$ matrix P where the rows and columns are numbered 0 through n, and the entries P_{ij} of P are defined as follows: P_{ij} = the probability that Player 1, having i pennies, will have j pennies after the next flip. In this case, P (which is called a **Markov chain** or **transition matrix**), looks like

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ n-1 \\ n \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & 1 & 2 & \cdots & n-1 & n \end{array} \\ \left(\begin{array}{cccccc} 1 & 0 & 0 & \cdots & 0 & 0 \\ 1/2 & 0 & 1/2 & \cdots & 0 & 0 \\ 0 & 1/2 & 0 & \cdots & 0 & 0 \\ & & & \vdots & & \\ 0 & 0 & 0 & \cdots & 0 & 1/2 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{array}\right) \end{array}$$

The row and column numbers of P are called **states**. States 0 and n are **absorbing states** because going to any other state from either has probability 0 (the game is over). An **absorbing Markov chain**, such as our penny game chain, is one that has at least one state from which no other state is accessible and a way to reach one of these states from every other state. Basic calculable quantities for absorbing chains include the *expected number of visits to each state before absorption*, the *mean time to absorption*, and the *probabilities of being absorbed into different absorbing states*. Absorbing Markov chains are typically written with the absorbing rows first. Our chain then looks like

$$P = \begin{array}{c} \\ 0 \\ n \\ 1 \\ 2 \\ \vdots \\ n-1 \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & n & 1 & 2 & \cdots & n-1 \end{array} \\ \left(\begin{array}{cc|cccc} 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 0 & \cdots & 0 \\ \hline 1/2 & 0 & 0 & 1/2 & \cdots & 0 \\ 0 & 0 & 1/2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & 1/2 & 0 & 0 & \cdots & 0 \end{array}\right) \end{array}$$

In general, if P is an $n \times n$ matrix and there are s absorbing states, then P looks like

$$P = \left(\begin{array}{c|c} I & 0 \\ \hline R & Q \end{array}\right)$$

List of Symbols

I	Identity matrix.
P	Transition matrix, unperturbed.
P'	Transition matrix after a perturbation.
$\mathscr{E}$	Perturbation matrix, $P+\mathscr{E}=P'$, $\mathscr{E}=g\delta$.
g	Column vector giving rows to be perturbed.
δ	Row vector of perturbation.
R	Transition matrix from transient to absorbing states in absorbing chains.
Q	Transition matrix from transient to transient states in absorbing chains.
N	Matrix of mean number of visits to transient states before absorption. Also, fundamental matrix for absorbing chains.
c	Column vector of ones.
t	Column vector of mean time to absorption.
B	Matrix of probabilities of being absorbed into particular states.
b	Any vector with nonzero sum.
α	Fixed probability vector for ergodic chains.
M	Matrix of mean first passage times.
Z	Fundamental matrix for ergodic chains.
L	Mean first passage times times fixed probability for destination: $L_{ij}=M_{ij}\alpha_j$.

where I is an $s\times s$ identity matrix, the upper right entries of P are all equal to zero, R is an $(n-s)\times s$ matrix, and Q is an $(n-s)\times(n-s)$ matrix.

It is well known that in order to find the quantities we are interested in, the key quantity to calculate is the **fundamental matrix** N, which equals $(I-Q)^{-1}$, where I is an identity matrix [**2**]. Each entry N_{ij} of N gives the mean number of times starting in state i that the process moves to state j before being absorbed. The mean times to absorption are given by the row sums of N, displayed as a vector t. Thus, $t=Nc$, where c denotes a column vector of all 1's. The probabilities of being absorbed into a particular state are given by the matrix product $B=NR$, where R is the lower left part of the transition matrix. B, therefore, gives us the answer to the question of Player 1's chances of winning when starting wealth is unequal. As the reader may verify, for a game where total wealth is $n=4$,

$$B=NR=\begin{array}{c} \\ 1\\2\\3\end{array}\!\!\begin{array}{c}\begin{array}{ccc}1&2&3\end{array}\\ \begin{pmatrix}3/2&1&1/2\\1&2&1\\1/2&1&3/2\end{pmatrix}\end{array}\begin{array}{c}\begin{array}{cc}0&4\end{array}\\ \begin{pmatrix}1/2&0\\0&0\\0&1/2\end{pmatrix}\end{array}=\begin{array}{c} \\ 1\\2\\3\end{array}\!\!\begin{array}{c}\begin{array}{cc}0&4\end{array}\\ \begin{pmatrix}3/4&1/4\\1/2&1/2\\1/4&3/4\end{pmatrix}\end{array}$$

So, if Player 1 starts with only one penny, his chances of winning are only $1/4$.

What if Player 2 cheats whenever Player 1 has wealth i, so that the probability of Player 1's wealth going to $i+1$ is only $(1/2)-\varepsilon$ and the probability of going to $i-1$ is $(1/2)+\varepsilon$? One could certainly reconstruct P to reflect this inequity and recalculate N. But inverting matrices is tedious, especially if n is large. Is it possible to calculate the new results directly from the old results and knowledge of the perturbation? The answer is yes, as we shall see in the next section.

As we consider the perturbation of Markov chains, we will also take a look at another type of model. For example, picture a circle of n states with probability $1/2$ of taking a step clockwise or counterclockwise each time; there is no state 0. (See FIGURE 1.) This model is an **ergodic Markov**

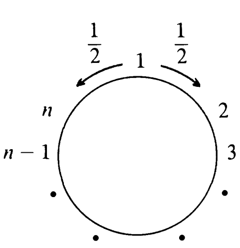


FIGURE 1

chain. An ergodic chain is one in which it is possible from any state to reach any other state and, if the process were to continue forever, the probability of reaching every state is 1. Thus there are no absorbing states. The Markov chain for the process pictured in FIGURE 1 looks like

$$P = \begin{array}{c} \\ 1 \\ 2 \\ \vdots \\ n \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & \cdots & n \end{array} \\ \begin{pmatrix} 0 & 1/2 & 0 & \cdots & 1/2 \\ 1/2 & 0 & 1/2 & \cdots & 0 \\ & & & \vdots & \\ 1/2 & 0 & 0 & \cdots & 0 \end{pmatrix} \end{array}.$$

The interesting values to calculate here are the percentage of time spent in each state in the long run, represented by a vector α, and the average time (number of steps) it takes to go from state i to state j, represented by entry M_{ij} in a matrix M, called the "matrix of mean first passage times." This is a model discussed in detail in [**2**, pp. 159–161]. The model is completely symmetric with

$$\alpha_i = \frac{1}{n} \text{ for all } i$$

$$M_{ij} = \begin{cases} (j-i)(n-j+i) & i<j \\ (i-j)(n-i+j) & i>j \end{cases}.$$

But what if one perturbs state k so that the probability of going clockwise from k is $(1/2)+\varepsilon$ and the probability of going counterclockwise is $(1/2)-\varepsilon$? Intuitively, α should increase for states clockwise from k; is that so? Also, what happens to mean first passage times?

These questions have been addressed by economist John Conlisk [**3**], and we will verify and generalize his results in this note. For another discussion of Markov chain perturbations, see [**4**].

If P is a Markov chain, we define $P' = P + \mathscr{E}$, and call $\mathscr{E}$ the **perturbation matrix**. (In the examples we have discussed, $\mathscr{E}$ is the matrix which has entries $\pm\varepsilon$ in the $i, i-1$ and $i, i+1$ positions and 0's elsewhere.) We will consider forms where $\mathscr{E}$ is factorable into column and row vectors, i.e., $\mathscr{E} = g\delta$, where g is a column vector and δ is a row vector. This type of matrix certainly allows for single row changes by taking g to have a solitary entry of 1 for the desired row and taking δ to be the desired perturbation for that row. One can also use this form for cases in which the perturbations on each row are proportional to one another. However, if changes on different rows are not proportional, then the perturbation matrix $\mathscr{E}$ would have to be considered as a sum of the form $\mathscr{E} = \sum_{i=1}^{k} g_i \delta_i$ and new values would have to be calculated for each step.

To answer our questions about the perturbed Markov chain P', we wish to calculate its fundamental matrix, which is of the form $(I-Y)^{-1}$, where $Y = Q + \mathscr{E}'$ ($\mathscr{E}'$ is a square submatrix of $\mathscr{E}$). Our techniques are not restricted to this particular application to Markov chains. In fact, *any* two $n \times n$ matrices Y_1 and Y_2 satisfy an equation $Y_2 = Y_1 + \mathscr{E}$ for some matrix $\mathscr{E}$, so the perturbation formulas which we will derive could be used to obtain $(I - Y_2)^{-1}$

from $(I - Y_1)^{-1}$ in no more than n applications. (There is no economy in finding $(I - Y_2)^{-1}$ this way if Y_1 and Y_2 differ at every row by disproportionate amounts. It is left to the user to determine when Y_1 and Y_2 are similar enough to warrant the use of our methods in finding $(I - Y_2)^{-1}$. In judging "similarity," it is the number of different entries, not the magnitude of the differences, that determines the efficiency of our method.)

THEOREM 1. *Let Y be an $n \times n$ matrix such that $X = (I - Y)^{-1}$ exists. Let $Y' = Y + g\delta$ where g is a nonzero column vector of length n and δ is a row vector of length n. Then $X' = (I - Y')^{-1}$ exists if and only if $\delta Xg \neq 1$, and if X' exists, then*

$$X' = X + \frac{Xg\delta X}{1 - \delta Xg}.$$

Proof. i) We prove first that if X' exists, then $\delta Xg \neq 1$. If X' exists, then

$$X'(I - Y - g\delta) = 1,$$

so

$$X'(I - Y) = I + X'g\delta.$$

Multiplying on the right by X gives

$$X' = X + X'g\delta X, \tag{1}$$

and multiplying this on the right by g gives

$$X'g = Xg + X'g\delta Xg,$$

so

$$X'g(1 - \delta Xg) = Xg. \tag{2}$$

Now, if $1 - \delta Xg = 0$, then (2) implies that Xg is identically zero, and since X is invertible, this implies that g is the zero vector, a contradiction. Thus, if X' exists, $\delta Xg \neq 1$.

ii) We prove next that if $\delta Xg \neq 1$, then X' exists and

$$X' = X + \frac{Xg\delta X}{1 - \delta Xg}. \tag{3}$$

(This result is suggested by solving (2) for $X'g$ and substituting into (1).) If $\delta Xg \neq 1$, then define X' by equation (3); this is indeed the desired inverse of $(I - Y')$ since $(I - Y')X' = I$, as the reader may verify.

Now, armed with the general perturbed inverse formula (3), we can move to develop formulas to calculate Markov quantities. In Markov perturbation, there are the added restrictions that

1) the matrix P' can have no negative entries and

2) $\mathscr{E}$ must have row sums equal to zero.

The second restriction is a result of the fact that both P and P' have row sums equal to 1. That is, if c is a column vector of all 1's, then

$$\begin{aligned} \mathscr{E}c &= (P' - P)c \\ &= c - c \\ &= \text{column vector of zeros.} \end{aligned}$$

In addition, if $\mathscr{E} = g\delta$, then $\mathscr{E}c = (g\delta)c = g(\delta c) =$ column vector of zeros, so $\delta c = 0$.

Formulas for perturbed Markov chains

Absorbing chains.

We first consider the formulas for absorbing Markov chains, such as the chain that represents our penny game. We will restrict perturbation here to the transient, or nonabsorbing, states since what we seek are ways to express the vector t' and matrix B' in terms of t and B. If one of the absorbing states were perturbed, then s (the number of absorbing states) would be decreased by one, thereby changing the dimensions of N', t' and B'.

Formally, $P' = P + \mathscr{E}$ as before, but $\mathscr{E} = g\delta$ where g is a column vector whose first s entries are 0; we write $g = \begin{pmatrix} 0 \\ g_1 \end{pmatrix}$. Accordingly, we partition the row vector δ, $\delta = (\delta_1, \delta_2)$ where δ_1 is the first s entries of δ, and δ_2 is the remaining n-s entries. Although $\delta c = 0$, either δ_1 or δ_2 could have a nonzero row sum.

With this notation, $Q' = Q + g_1\delta_2$. Using Theorem 1, it is straightforward to see that if $\delta_2 N g_1 \neq 1$, then

$$N' = N + \frac{Ng_1\delta_2 N}{1 - \delta_2 N g_1}, \tag{4}$$

and, since $t' = N'c$,

$$t' = t + \frac{Ng_1\delta_2 t}{1 - \delta_2 N g_1}. \tag{5}$$

Calculating B' is a little trickier, but $B' = N'R'$ works out to

$$B' = B + \frac{Ng_1\delta_2 B}{1 - \delta_2 N g_1} + \frac{Ng_1\delta_1}{1 - \delta_2 N g_1}. \tag{6}$$

It's probably easiest to compute N' first and then get t' and B' directly rather than in terms of t and B.

Ergodic chains

We turn now to the perturbation of ergodic chains and the calculation of the new fixed probability vector, α', and matrix of mean first passage times, M'. There is not a unique fundamental matrix for ergodic chains, but rather a class. The fundamental matrix for an ergodic chain P is any matrix Z which satisfies the equation $Z = (I - P + cb)^{-1}$, where b is any row vector such that the sum of the entries in b is nonzero (i.e., $bc \neq 0$). For a discussion of a yet more general class, see Kemeny's paper [**5**]. The fixed probability vector α is then easily obtained by $\alpha = bZ$. The matrix of mean first passage times M can be computed entry by entry with

$$M_{ij} = \frac{Z_{jj} - Z_{ij}}{\alpha_j}.$$

A more natural quantity suggested by the preceding equation would be a matrix L of values where $L_{ij} = M_{ij}\alpha_j$. Thus $L_{ij} = Z_{jj} - Z_{ij}$. A matrix equation for L in terms of Z can be written if one lets Z_{diag} be a row vector of the diagonal entries in Z. Then $L = cZ_{\text{diag}} - Z$. Assuming that P' is ergodic, Z', α' and L' are all well defined. It will be shown later than $\delta Z g$ cannot equal 1 if P' is ergodic.

To use Theorem 1, consider $Y = P + cb$ so that X will correspond to Z. The same b can be used for both Z and Z', so the theorem still holds to obtain

$$Z' = Z + \frac{Zg\delta Z}{1 - \delta Z g}. \tag{7}$$

Multiplying on the left by b gives

$$\alpha' = \alpha + \frac{\alpha g \delta Z}{1 - \delta Z g}.$$

This equation would be more useful if it used L instead of Z, since Z is not unique. Substituting L for Z gives

$$\alpha' = \alpha + \frac{\alpha g \delta (cZ_{diag} - L)}{1 - \delta (cZ_{diag} - L) g},$$

and since $\delta c = 0$,

$$\alpha' = \alpha - \frac{\alpha g \delta L}{1 + \delta L g}.$$

Unfortunately, it does not seem possible to get a general form perturbation result for M free of Z.

However, let us consider the more specific perturbation in which only two entries of P are disturbed, that is, only the kth row of P is altered. Here, $\mathscr{E}$ is all zeros except for $\mathscr{E}_{kl} = +\varepsilon$ and $\mathscr{E}_{km} = -\varepsilon$. The matrix $\mathscr{E}$ then factors as $g\delta$ where g is all zeros except $g_k = 1$, and δ is all zeros, except $\delta_l = +\varepsilon$, $\delta_m = -\varepsilon$.

For these vectors g and δ, we have

$$[Zg\delta Z]_{ij} = \varepsilon Z_{ik}(Z_{lj} - Z_{mj})$$

and

$$\delta Z g = \varepsilon(Z_{lk} - Z_{mk}).$$

Using the equation for Z' gives

$$Z'_{ij} = Z_{ij} + \frac{\varepsilon Z_{ik}(Z_{lj} - Z_{mj})}{1 - \varepsilon(Z_{lk} - Z_{mk})},$$

which translates to

$$L'_{ij} = L_{ij} + \frac{\varepsilon(L_{ik} - L_{jk})(L_{mj} - L_{lj})}{1 - \varepsilon(L_{mk} - L_{lk})}. \tag{8}$$

Entry-by-entry formulas for α would make a useful companion to (8) and they can be found without using Theorem 1. First observe that for fundamental matrices Z and Z',

$$Z'(Z')^{-1} = ZZ^{-1}$$

so

$$Z'Z^{-1} - Z'Z'^{-1} = Z'Z^{-1} - ZZ^{-1}.$$

Recalling that

$$Z^{-1} - Z'^{-1} = (I - P + cb) - (I - (P + \mathscr{E}) + cb) = \mathscr{E}$$

gives

$$Z'\mathscr{E} = (Z' - Z)Z^{-1}.$$

Multiplying by b on the left and Z on the right yields

$$\alpha' \mathscr{E} Z = \alpha' - \alpha.$$

Solving the last equation for particular components shows that

$$\alpha'_i = \alpha_i + \alpha'_k \varepsilon(L_{mi} - L_{li}),$$

so

$$\alpha'_k = \frac{\alpha_k}{1 - \varepsilon(L_{mk} - L_{lk})}. \tag{9}$$

Using (9) and (8) we can now compute any entry of α', as well as any entry of the matrix L', for the specialized 2-entry perturbation. Clearly, these two items give all the information needed to obtain new mean first passage times, or M'.

The penny game with cheating

Turning back to our penny game, we can now answer what happens if Player 2 cheats whenever Player 1 has wealth k so that his chance of winning a penny is only $1/2 - \varepsilon$. Kemeny and Snell [1, p. 151] provide the fact that, in an unperturbed absorbing random walk with $p = 1/2$ (our penny game is an example),

$$N_{ij} = \begin{cases} \frac{2}{n} j(n-i) & \text{if } j \le i \\ \frac{2}{n} i(n-j) & \text{if } j \ge i \end{cases} \tag{10}$$

$$t_i = i(n-i).$$

Perturb point k so that $P_{k,k-1} = (1/2) + \varepsilon$ and $P_{k,k+1} = (1/2) - \varepsilon$. By the perturbation formula (4),

$$N'_{ij} = N_{ij} + \frac{\varepsilon N_{ik}(N_{k-1,j} - N_{k+1,j})}{1 - \varepsilon(N_{k-1,k} - N_{k+1,k})} = N_{ij} + \frac{\varepsilon N_i^* N_j^*}{1 - \varepsilon N_k^*},$$

where N_i^* is the part dependent on i, N_j^* depends on j, and N_k^* depends on k. Using (10), we obtain

$$N_{k-1,k} = \frac{2}{n}(k-1)(n-k) = \frac{2}{n}(kn - k^2 - n + k)$$

$$N_{k+1,k} = \frac{2}{n}k(n-k-1) = \frac{2}{n}(kn - k^2 - k)$$

so

$$N_k^* = \frac{2}{n}(2k-n).$$

Straightforwardly,

$$N_i^* = \begin{cases} \frac{2}{n} i(n-k) & i < k \\ \frac{2}{n} k(n-i) & i \ge k. \end{cases}$$

$N_j^* = N_{k-1,j} - N_{k+1,j}$ will depend on j:

$$N_j^* = \begin{cases} \frac{2}{n}(2j) & j < k \\ \frac{2}{n}(2j-n) & j = k \\ \frac{2}{n}(2j-2n) & j > k. \end{cases}$$

So

$$\sum_{j=1}^{n-1} N_j^* = \frac{2}{n}(2kn - n^2),$$

as the reader may verify.

We can now compute new values for time to absorption (end of game) and Player 1's probability of winning. Using (5) and (10), calculation of t' is reasonably simple:

$$t'_i = t_i + \begin{cases} \dfrac{\varepsilon \frac{2}{n} i(n-k)2(2k-n)}{1 - \varepsilon \frac{2}{n}(2k-n)} & \text{if } i < k \\ \dfrac{\varepsilon \frac{2}{n} k(n-i)2(2k-n)}{1 - \varepsilon \frac{2}{n}(2k-n)} & \text{if } i \ge k. \end{cases}$$

Player 1's probability of winning is shown by the second column of B, since that gives the probabilities of being absorbed into state n. The entries of the second column of B are obtained by multiplying the $(n-1)$ column of N by $1/2$.

$$B'_{in} = B_{in} + \frac{1}{2}\frac{N_i^* N^*_{(j=n-1)}}{1-\varepsilon N_k^*}$$
$$= B_{in} - \frac{\frac{2}{n}\varepsilon}{1-\frac{2}{n}\varepsilon(2k-n)} \cdot \begin{cases} \frac{2}{n} i(n-k) & \text{if } i<k \\ \frac{2}{n} k(n-i) & \text{if } i \geq k. \end{cases}$$

As expected, Player 1's probability of winning has decreased for every value of i, since the numerator of the additional part is always positive, as is the denominator.

$$1-\varepsilon\frac{2}{n}(2k-n) > 1-\frac{1}{2}\cdot\frac{2}{n}(2k-n)$$
$$> 1-\frac{1}{n}(n)$$
$$> 0.$$

An ergodic perturbation

Now we return to our circular random walk. Because of the circular symmetry, perturbing state 1 is as illuminating as perturbing any other, so let

$$P'_{12} = \frac{1}{2}+\varepsilon, \qquad P'_{1n} = \frac{1}{2}-\varepsilon.$$

Since $\alpha'_1 = \alpha_1 + 0 = 1/n$ and $\alpha'_i = \alpha_i(1+\varepsilon\alpha'_k(M_{ni}-M_{2i}))$, we have

$$\alpha'_i = \begin{cases} \alpha_i, & i=1 \\ \alpha_i + \frac{\varepsilon(n-4)}{n^2}, & i=2 \\ \alpha_i + \frac{\varepsilon(2n-4i+4)}{n^2}, & 2<i<n \\ \alpha_i + \frac{\varepsilon(4-n)}{n^2}, & i=n. \end{cases}$$

One would imagine that increasing the clockwise flow out of state 1 would increase the fixed probability measures for points to the right of state 1 and decrease them to the left of state 1. Indeed $\alpha'_i - \alpha_i$ is positive for $1<i<(n+2)/2$, negative for $i>(n+2)/2$, and zero for $i=1$ and $i=(n+2)/2$, if there are an even number of states on the circle.

The formulas for L'_{ij} here are

$$L'_{ij} = L_{ij} + \frac{\varepsilon(L_{i1}-L_{j1})(L_{nj}-L_{2j})}{1-\varepsilon(L_{n1}-L_{21})}$$
$$= L_{ij} + \varepsilon\left(\frac{1}{n}\right)^2[(i-1)(n-i+1)-(j-1)(n-j+1)](M_{ij}-M_{2j}),$$

$$L'_{ij} = \begin{cases} L_{ij}, & j=1 \\ L_{ij} + \varepsilon\left(\frac{1}{n}\right)^2((i-1)(n-i+1)-(n-1))(2n-4), & j=2 \\ L_{ij} + \varepsilon\left(\frac{1}{n}\right)^2((i-1)(n-i+1)-(j-1)(n-j+1))(2n-4i+4), & 2<j<n \\ L_{ij} + \varepsilon\left(\frac{1}{n}\right)^2((i-1)(n-i+1)-(n-1))(4-2n), & j=n. \end{cases}$$

(Details of the calculations to obtain these formulas are contained in [**1**].) These formulas give a path to finding M'_{ij}'s by using them in connection with formulas for α'_j. In general, they show that even a small perturbation upsets and complicates the computation of Markov values for otherwise symmetric chains. To an even greater degree than in the absorbing random walk model, chains where p is the probability of going clockwise and $p \neq 1/2$ have prohibitively complicated original formulas. Calculations would be straightforward but tedious since the formulas do not simplify, and multiple cases for values of i and j must be considered when computing L'_{ij}.

John Conlisk posed the question that if an ergodic Markov chain P is perturbed so that P_{kl} is increased and P_{km} is decreased, intuition suggests that the fixed probability vector would increase at l and decrease at m, but can this be proven? He answered his own specific question, with attention to the rates of change of basic quantities in [**3**].

Formulas developed here confirm the derivative formulas found in Conlisk's paper. Assume the basic perturbation of $+\varepsilon$ at P_{kl} and $-\varepsilon$ at P_{km}. Then

$$\begin{aligned}\frac{\partial \alpha_l}{\partial \varepsilon} &= \lim_{\varepsilon \to 0} \frac{\alpha'_l - \alpha_l}{\varepsilon} = \lim_{\varepsilon \to 0} \frac{\alpha_k M_{ml} \alpha_l}{1 - \varepsilon \alpha_k (M_{mk} - M_{lk})} \\ &= \alpha_k M_{ml} \alpha_l .\end{aligned}$$

Summarizing similar results for the rest of α,

$$\frac{\partial \alpha_i}{\partial \varepsilon} = \begin{cases} \alpha_k M_{ml} \alpha_l & \text{for } i = l \\ -\alpha_k M_{lm} \alpha_m & \text{for } i = m \\ \alpha_k (M_{mi} - M_{li}) \alpha_i & \text{otherwise.} \end{cases}$$

Since α and M are always nonnegative, this establishes that for simple, two-place perturbations, the fixed probability vector always increases in the state where the transition probability towards it was increased, and decreases where the transition probability towards it was decreased. The direction of the change for other entries in α' depends on certain entries in M. Inspection of the formula for L' reveals that six values from L would have to be known in order to predict the direction of change from entries of M to entries of M', and that again assumes that only two entries of P have been altered. Formulas treating absorbing chains yield equally complicated results when answering questions of simple increase/decrease. (These questions are considered in more detail in [**1**].)

Degenerate cases

Both formulas (4) for N' and (7) for Z' have expressions in the denominator of the form $1 - \delta X g$, where X is the original matrix. The perturbation formulas will therefore "blow up" if $\delta X g = 1$. Let's examine the nature of the perturbations for which $\delta X g = 1$.

We first consider absorbing chains. One would expect the formulas to indicate something if a perturbation creates a new absorbing state within the transient part of an absorbing Markov chain. The situation is easy to check; consider a single-row change in the ith row of Q that causes the ith row of Q' to be absorbing, that is, it is identical to the ith row of I. Letting e_i be the basis vector with the ith component equal to one and components elsewhere equal to zero, $g = e_i^T$ and $\delta = e_i(I - Q)$. Then

$$\begin{aligned}\delta N g &= [e_i(I - Q)](I - Q)^{-1}[e_i^T] \\ &= e_i e_i^T \\ &= 1.\end{aligned}$$

Is this the only case in which $\delta N g$ can be 1? Not quite. The same problem will occur if a new ergodic set is formed within the chain. An **ergodic set** is a closed set of states in which every state is accessible from every other state and from which no other states can be reached. A state in an ergodic set is an ergodic state. Each absorbing state by itself constitutes an ergodic set.

THEOREM 2. *In an absorbing Markov chain*, $\delta N g = 1$ *if and only if a new ergodic set is formed.*

Proof. For absorbing chains, $(I - Q')^{-1}$ exists if and only if there is no ergodic class within Q'. This is because if there were an ergodic set E contained in Q', then Q' could be rewritten listing states in E first. If w were a fixed vector for E and $\bar{0}$ a vector of zeros then $(w, \bar{0})$ would be a fixed vector for Q', thus giving $(I - Q')$ a nontrivial nullspace. Equivalently, then, $(I - Q')^{-1}$ does not exist, so, by Theorem 1, $\delta N g = 1$.

The results for ergodic chains are somewhat surprising. We assume, as above, that only perturbations resulting in new transition matrices with nonnegative entries are under consideration.

THEOREM 3. *If* P *and* P' *are ergodic, then* $\delta Z g$ *cannot equal* 1.

Proof. By specializing a result in a forthcoming paper by John Kemeny, we can show that $(I - P' + cb)^{-1}$ exists if and only if there is a single ergodic set in P'. This condition occurs only if P' is ergodic or if P' has a single absorbing state. Again by Theorem 1, the existence of $(I - P' + cb)^{-1}$ means $\delta Z g \neq 1$.

By considering single-row perturbations, an upper bound for $\delta Z g$ can be established. Perturb row k so that

$$\begin{aligned}\delta Z g &= \sum_i \delta_i Z_{ik} \\ &= \sum_i \delta_i (Z_{kk} - (Z_{kk} - Z_{ik})) \\ &= \sum_i \delta_i Z_{kk} - \sum_i \delta_i L_{ik} \\ &= -\sum_i \delta_i L_{ik},\end{aligned}$$

since $\Sigma_i \delta_i = 0$. To maximize $-\Sigma_i \delta_i L_{ik}$, assume that all the positive entries of δ fall where $L_{ik} = 0$, namely at $i = k$. In order to keep the entries of P' nonnegative, the inequality $-\delta_i \leqslant P_{ki}$ holds for each entry. So

$$-\sum_i \delta_i L_{ik} \leqslant \sum_i P_{ki} L_{ik} = \alpha_k \left(\frac{1}{\alpha_k} - 1 \right) = 1 - \alpha_k$$

(see Kemeny, Snell, and Thompson [6]). Summarizing, $\delta Z g \leqslant 1 - \alpha_k$. Note that $\delta Z_g = 1 - \alpha_k$ if and only if all the positive entries of δ fall on k and the negative entries cancel out the rest of row k. That condition is equivalent to making an absorbing state at row k. In this case, there is still a unique fixed vector α', but since it is not strictly positive (in fact, $\alpha' = e_k$, the kth row of I), some mean first passage times are infinite.

The author thanks Professors J. Laurie Snell and John G. Kemeny for their help and suggestions on this paper.

References

[1] Christine Burnley, Perturbation of Markov Chains, Senior Honors Thesis, Dartmouth College, 1984, unpublished.

[2] John G. Kemeny and J. Laurie Snell, Finite Markov Chains, Van Nostrand, 1960; Springer-Verlag, 1975.

[3] John Conlisk, Comparative statics for Markov chains, J. of Economic Dynamics and Control, 9(1985) 139–51.

[4] Carl D. Meyer, Jr. and James N. Shoaf, Updating finite Markov chains by using techniques of group matrix inversion, J. Statistical Comput. Simul., 11(1980) 163–181.

[5] John G. Kemeny, Generalization of a fundamental matrix, Linear Algebra and Its Applications, 38(1981) 193–206.

[6] John G. Kemeny, J. Laurie Snell, and Gerald L. Thompson, Introduction to Finite Mathematics, 3rd ed., Prentice Hall, Inc., 1974.

Coefficients of the Characteristic Polynomial

Louis L. Pennisi
University of Illinois at Chicago
Box 4348, Chicago, Il 60680

This note derives the determinantal formulas for the coefficients in the characteristic polynomial of a matrix. In some areas of mathematics and other fields of science one has to determine the characteristic polynomial

$$P(\lambda) = \det(\underline{A} - \lambda \underline{I}) = b_0\lambda^n + b_1\lambda^{n-1} + b_2\lambda^{n-2} + \cdots + b_{n-1}\lambda + b_n, \tag{1}$$

of a matrix $\underline{A}$, where $\underline{A} = (a_{ij})$ is an $n \times n$ matrix over the field of real numbers (or complex numbers) and $\underline{I}$ is the $n \times n$ identity matrix. Once $P(\lambda)$ is determined, the sought for eigenvalues λ_i, $i = 1, 2, \ldots, n$, of $\underline{A}$ may be obtained by solving the characteristic equation $P(\lambda) = 0$. When $n \leqslant 3$, usually one obtains $P(\lambda)$ by expanding the $\det(\underline{A} - \lambda \underline{I})$ by minors. When $n \geqslant 3$, this task may become somewhat laborious. For a method of determining the b_i's in (1), see [**1**], where the following recursive formula is found:

$$\begin{aligned} &b_0 = (-1)^n, \qquad b_1 = -(-1)^n T_1, \qquad b_2 = -\frac{1}{2}\left[b_1T_1 + (-1)^n T_2\right], \\ &b_3 = -\frac{1}{3}\left[b_2T_1 + b_1T_2 + (-1)^n T_3\right], \ldots, \\ &b_n = -\frac{1}{n}\left[b_{n-1}T_1 + b_{n-2}T_2 + \cdots + b_1T_{n-1} + (-1)^n T_n\right], \end{aligned} \tag{2}$$

where $T_1, T_2, \ldots, T_n$ denote the traces of the matrices $\underline{A}, \underline{A}^2, \ldots, \underline{A}^n$, respectively. (The *trace* of an $n \times n$ matrix $\underline{B}$ is the sum of the elements on its main diagonal.)

We shall now present an alternative method for determining the coefficients of the characteristic polynomial. This method will involve the expansion of determinants of order 1 through n rather than computing the traces of the matrices $\underline{A}, \underline{A}^2, \ldots, \underline{A}^n$.

For our purpose, we shall consider the polynomial

$$g(t) = \det(\underline{K} + t\underline{I}) \quad (\text{the } + \text{ sign intended}), \tag{3}$$

where $\underline{K} = (k_{ij})$ is an $n \times n$ matrix over the field of real numbers (or complex numbers). We now ask, what is the coefficient C_r of t^r in (3)? To separate the occurrences of t, let

$$f(t_1, t_2, \ldots, t_n) = \det(\underline{K} + \operatorname{diag}(t_1, t_2, \ldots, t_n)).$$

Then $g(t) = f(t, t, \ldots, t)$, and C_r is the sum of the coefficients of the terms of total degree r in $f(t_1, t_2, \ldots, t_n)$. The fact that $f(t_1, t_2, \ldots, t_n)$ is of degree 1 in each t_i separately simplifies the accounting:

$$C_r = \sum_{i_1 < i_2 < \cdots < i_r} \frac{\partial^r}{\partial t_{i_1} \partial t_{i_2} \cdots \partial t_{i_r}} f(t_1, t_2, \ldots, t_n) \Bigg|_{\substack{t_1 = 0 \\ t_2 = 0 \\ \vdots \\ t_n = 0}}$$

where $1 \leqslant i_1$, $i_r \leqslant n$, and $0 \leqslant r \leqslant n$.

Thus, the C_r's are now expressed in terms of the elements of the matrix $\underline{K}$. Let D be the determinant as a function of the n^2 entries:

$$D(k_{11}, k_{12}, \ldots, k_{nn}) = \begin{vmatrix} k_{11} & \cdots & k_{1n} \\ \cdot & \cdots & \cdot \\ \cdot & \cdots & \cdot \\ \cdot & \cdots & \cdot \\ k_{n1} & \cdots & k_{nn} \end{vmatrix}.$$

Then

$$C_r = \sum_{i_1 < i_2 < \cdots < i_r} \frac{\partial^r}{\partial k_{i_1 i_1} \partial k_{i_2 i_2} \cdots \partial k_{i_r i_r}} D. \tag{4}$$

Because of the cluster of n^2 variables, it may be wise to spell out the transition to formula (4). We can reduce this cluster by suppressing the dependence of D on its off-diagonal variables k_{ij} where $i \neq j$. This suppression is feasible because these off-diagonal variables are parameters that remain fixed throughout this whole discussion. So let us write

$$f(t_1, t_2, \ldots, t_n) = D^*(k_{11} + t_1, k_{22} + t_2, \ldots, k_{nn} + t_n)$$

where D^* is a function of only n variables $k_{11}, k_{22}, \ldots, k_{nn}$ defined by

$$D^*(k_{11}, k_{22}, \ldots, k_{nn}) = D(k_{11}, k_{12}, \ldots, k_{nn}).$$

In this notation, it is manifest that

$$\frac{\partial^r f}{\partial t_{i_1} \partial t_{i_2} \cdots \partial t_{i_r}} = D^*_{k_{i_1 i_1} k_{i_2 i_2} \cdots k_{i_r i_r}}(k_{11} + t_1, k_{22} + t_2, \cdots, k_{nn} + t_n),$$

where the subscripts denote partial derivatives of D^*.

The rule for expansion of a determinant by a row (or column) tells us that $\partial D / \partial k_{ij}$ is the (i, j) cofactor of $\underline{K}$; when $i = j$ (on the diagonal), this is the (i, i)-minor. Thus, the partial derivative in (4) is just the subdeterminant of $\underline{K}$ resulting from crossing out the rows and columns numbered $i_1, i_2, \ldots, i_r$.

EXAMPLE. Let us determine $g(t) = \det(\underline{K} + t\underline{I})$, where

$$\underline{K} = \begin{pmatrix} k_{11} & k_{12} & k_{13} \\ k_{21} & k_{22} & k_{23} \\ k_{31} & k_{32} & k_{33} \end{pmatrix}.$$

Solution.

$$D(k_{11}, k_{12}, \ldots, k_{33}) = \begin{vmatrix} k_{11} & k_{12} & k_{13} \\ k_{21} & k_{22} & k_{23} \\ k_{31} & k_{32} & k_{33} \end{vmatrix}. \tag{5}$$

By definition:

$$C_0 = D(k_{11}, k_{12}, \ldots, k_{33}) \tag{6}$$

$$C_1 = \sum_{i_1} \frac{\partial}{\partial k_{i_1 i_1}} D = \frac{\partial D}{\partial k_{11}} + \frac{\partial D}{\partial k_{22}} + \frac{\partial D}{\partial k_{33}} \tag{7}$$

$$= \begin{vmatrix} 1 & k_{12} & k_{13} \\ 0 & k_{22} & k_{23} \\ 0 & k_{32} & k_{33} \end{vmatrix} + \begin{vmatrix} k_{11} & 0 & k_{13} \\ k_{21} & 1 & k_{23} \\ k_{31} & 0 & k_{33} \end{vmatrix} + \begin{vmatrix} k_{11} & k_{12} & 0 \\ k_{21} & k_{22} & 0 \\ k_{31} & k_{32} & 1 \end{vmatrix}$$

$$= \begin{vmatrix} k_{22} & k_{23} \\ k_{32} & k_{33} \end{vmatrix} + \begin{vmatrix} k_{11} & k_{13} \\ k_{31} & k_{33} \end{vmatrix} + \begin{vmatrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{vmatrix}$$

$$C_2 = \sum_{i_1 < i_2} \frac{\partial^2}{\partial k_{i_1 i_1} \partial k_{i_2 i_2}} D = \frac{\partial^2 D}{\partial k_{11} \partial k_{22}} + \frac{\partial^2 D}{\partial k_{11} \partial k_{33}} + \frac{\partial^2 D}{\partial k_{22} \partial k_{33}} \tag{8}$$

$$= \begin{vmatrix} 1 & 0 & k_{13} \\ 0 & 1 & k_{23} \\ 0 & 0 & k_{33} \end{vmatrix} + \begin{vmatrix} 1 & k_{12} & 0 \\ 0 & k_{22} & 0 \\ 0 & k_{32} & 1 \end{vmatrix} + \begin{vmatrix} k_{11} & 0 & 0 \\ k_{21} & 1 & 0 \\ k_{31} & 0 & 1 \end{vmatrix}$$

$$= k_{33} + k_{22} + k_{11}$$

$$C_3 = \sum_{i_1 < i_2 < i_3} \frac{\partial^3}{\partial k_{i_1 i_1} \partial k_{i_2 i_2} \partial k_{i_3 i_3}} D = \frac{\partial^3 D}{\partial k_{11} \partial k_{22} \partial k_{33}} = \begin{vmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{vmatrix} = 1. \tag{9}$$

Thus, our desired polynomial $g(t)$ is given by

$$g(t) = C_0 + C_1 t + C_2 t^2 + C_3 t^3, \tag{10}$$

where C_0, C_1, C_2, and C_3 are given by (6)–(9), respectively.

Taking

$$\underline{K} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 2 & 2 \\ 3 & 2 & 3 \end{pmatrix},$$

we have:

$$C_0 = \begin{vmatrix} 1 & 2 & 3 \\ 4 & 2 & 2 \\ 3 & 2 & 3 \end{vmatrix} = 4, \qquad C_1 = \begin{vmatrix} 2 & 2 \\ 2 & 3 \end{vmatrix} + \begin{vmatrix} 1 & 3 \\ 3 & 3 \end{vmatrix} + \begin{vmatrix} 1 & 2 \\ 4 & 2 \end{vmatrix} = 2 - 6 - 6 = -10,$$

$$C_2 = 3 + 2 + 1 = 6, \qquad C_3 = 1.$$

Substituting these values in (10), we obtain

$$g(t) = -4 - 10t + 6t^2 + t^3.$$

REMARK. The relationship between $g(t)$ of (3) and $P(\lambda)$ of (1) is given by $g(-\lambda) = P(\lambda)$, where going from (1) to (3), we have put $\underline{K} = \underline{A}$ and $t = -\lambda$.

Reference

[1] L. A. Zadeh and C. A. Desoer, Linear System Theory: The State Approach, McGraw-Hill, New York, 1963, pp. 303–305.

Matrices as Sums of Invertible Matrices

N. J. LORD

Tonbridge School

Kent TN9 1JP England

While it is a trivial truism that not every matrix is invertible, it does not seem to be well known that every matrix can be expressed as the sum of two invertible matrices. The proof of this makes a good exercise in elementary linear algebra and, although a direct proof is short, we have expanded the discussion to indicate several contrasting lines of attack.

For convenience we shall adopt the following notation:

$\mathbb{F}$ will denote the field under consideration;

q will denote the number of elements of $\mathbb{F}$ if $\mathbb{F}$ is finite;

$M(n,\mathbb{F})$ will denote the ring of $n \times n$ matrices with entries in $\mathbb{F}$, where $n > 1$;

$G(n,\mathbb{F})$ will denote the group of invertible matrices in $M(n,\mathbb{F})$;

I (or I_n, for emphasis) will denote the $n \times n$ identity matrix.

The theorem that we are going to prove then is as follows:

$$= k_{33} + k_{22} + k_{11}$$

$$C_3 = \sum_{i_1 < i_2 < i_3} \frac{\partial^3}{\partial k_{i_1 i_1} \partial k_{i_2 i_2} \partial k_{i_3 i_3}} D = \frac{\partial^3 D}{\partial k_{11} \partial k_{22} \partial k_{33}} = \begin{vmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{vmatrix} = 1. \tag{9}$$

Thus, our desired polynomial $g(t)$ is given by

$$g(t) = C_0 + C_1 t + C_2 t^2 + C_3 t^3, \tag{10}$$

where C_0, C_1, C_2, and C_3 are given by (6)–(9), respectively.

Taking

$$\underline{K} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 2 & 2 \\ 3 & 2 & 3 \end{pmatrix},$$

we have:

$$C_0 = \begin{vmatrix} 1 & 2 & 3 \\ 4 & 2 & 2 \\ 3 & 2 & 3 \end{vmatrix} = 4, \qquad C_1 = \begin{vmatrix} 2 & 2 \\ 2 & 3 \end{vmatrix} + \begin{vmatrix} 1 & 3 \\ 3 & 3 \end{vmatrix} + \begin{vmatrix} 1 & 2 \\ 4 & 2 \end{vmatrix} = 2 - 6 - 6 = -10,$$

$$C_2 = 3 + 2 + 1 = 6, \qquad C_3 = 1.$$

Substituting these values in (10), we obtain

$$g(t) = -4 - 10t + 6t^2 + t^3.$$

REMARK. The relationship between $g(t)$ of (3) and $P(\lambda)$ of (1) is given by $g(-\lambda) = P(\lambda)$, where going from (1) to (3), we have put $\underline{K} = \underline{A}$ and $t = -\lambda$.

Reference

[1] L. A. Zadeh and C. A. Desoer, Linear System Theory: The State Approach, McGraw-Hill, New York, 1963, pp. 303–305.

Matrices as Sums of Invertible Matrices

N. J. Lord
Tonbridge School
Kent TN9 1JP England

While it is a trivial truism that not every matrix is invertible, it does not seem to be well known that every matrix can be expressed as the sum of two invertible matrices. The proof of this makes a good exercise in elementary linear algebra and, although a direct proof is short, we have expanded the discussion to indicate several contrasting lines of attack.

For convenience we shall adopt the following notation:

$\mathbb{F}$ will denote the field under consideration;

q will denote the number of elements of $\mathbb{F}$ if $\mathbb{F}$ is finite;

$M(n, \mathbb{F})$ will denote the ring of $n \times n$ matrices with entries in $\mathbb{F}$, where $n > 1$;

$G(n, \mathbb{F})$ will denote the group of invertible matrices in $M(n, \mathbb{F})$;

I (or I_n, for emphasis) will denote the $n \times n$ identity matrix.

The theorem that we are going to prove then is as follows:

THEOREM. *Let A be in $M(n,\mathbb{F})$. Then A can be expressed as the sum of two elements of $G(n,\mathbb{F})$ where*

(i) *the summands may be taken as distinct unless A is the zero matrix and $\mathbb{F}$ has characteristic* 2;

(ii) *the decomposition is unique only if A is a nonzero 2×2 matrix with entries in the field with two elements.*

We deal first with the easiest case: $\mathbb{F}$ is infinite. For x in $\mathbb{F}$, let $p(x)$ denote the nonzero polynomial function $\det(A-xI)$. Since p has degree n, $p(x)$ vanishes for at most n values of x, so we can certainly find x_0 in $\mathbb{F}$ with $x_0\neq 0$, $p(x_0)\neq 0$ and $A\neq 2x_0I$.

Then $x_0I\in G(n,\mathbb{F})$, because $x_0\neq 0$; $A-x_0I\in G(n,\mathbb{F})$, because $p(x_0)\neq 0$; and $x_0I\neq A-x_0I$, because $A\neq 2x_0I$. Thus $A=(A-x_0I)+x_0I$ provides a splitting of A as the sum of two distinct elements of $G(n,\mathbb{F})$. (This argument will also carry through for a finite field provided that $n+3\leqslant q$.)

For $\mathbb{F}$ finite and $q>2$, we can use a counting argument based on a comparison of the sizes of $M(n,\mathbb{F})$ and $G(n,\mathbb{F})$. First note that an element of $M(n,\mathbb{F})$ can be obtained by putting any of the q elements of $\mathbb{F}$ into n^2 'slots' so $|M(n,\mathbb{F})|$, the cardinality of $M(n,\mathbb{F})$, is q^{n^2}.

Next, an element of $G(n,\mathbb{F})$ is characterised by the fact that its columns form an ordered linearly independent set. The first column is subject only to the restriction that it is nonzero: q^n-1 choices. The second column is subject only to the restriction that it is not linearly dependent on the first column: q^n-q choices. Continuing in this manner we see that

$$|G(n,\mathbb{F})|=(q^n-1)(q^n-q)(q^n-q^2)\cdots(q^n-q^{n-1})$$
$$=q^{\frac{1}{2}n(n-1)}(q-1)(q^2-1)\cdots(q^n-1).$$

Further progress hinges on establishing the inequality:

$$2|G(n,\mathbb{F})|>|M(n,\mathbb{F})|+1.$$

In view of our formulae, this is equivalent to showing

$$2(q-1)(q^2-1)\cdots(q^n-1)>q^{\frac{1}{2}n(n+1)}+q^{-\frac{1}{2}n(n-1)}$$

or

$$(1-q^{-1})(1-q^{-2})\cdots(1-q^{-n})>\tfrac{1}{2}(1+q^{-n^2}).$$

Since $q\geqslant 3$,

$$(1-q^{-1})(1-q^{-2})\cdots(1-q^{-n})\geqslant(1-3^{-1})(1-3^{-2})\cdots(1-3^{-n})$$

and

$$\tfrac{1}{2}(1+q^{-n^2})\leqslant\tfrac{1}{2}(1+3^{-n^2})<\tfrac{1}{2}(1+3^{-n}).$$

So it is enough to show that $(1-3^{-1})(1-3^{-2})\cdots(1-3^{-n})>\frac{1}{2}(1+3^{-n})$, which is easily done by induction on n ($n\geqslant 2$).

Next, define a map T_A from $G(n,\mathbb{F})$ to $M(n,\mathbb{F})$ by $T_A(X)=A-X$ and let imT_A denote its image. T_A is clearly injective, so $|imT_A|=|G(n,\mathbb{F})|$. Moreover, our estimate above shows that

$$|imT_A\cap G(n,\mathbb{F})|=|imT_A|+|G(n,\mathbb{F})|-|imT_A\cup G(n,\mathbb{F})|$$
$$\geqslant 2|G(n,\mathbb{F})|-|M(n,\mathbb{F})|$$
$$>1.$$

So, there are at least two elements in $imT_A\cap G(n,\mathbb{F})$, say Y_1,Y_2 with

$$T_A(X_1)=Y_1\quad\text{and}\quad T_A(X_2)=Y_2.$$

Both of these provide decompositions of $A=X_1+Y_1=X_2+Y_2$ and one provides a splitting into distinct elements of $G(n,\mathbb{F})$ unless $\mathbb{F}$ has characteristic 2. For if $X_1\neq Y_1$ we are done. Otherwise

$A = 2X_1 = 2Y_1$ and if also $X_2 = Y_2$ we deduce that $A = 2Y_2$. But $A = 2Y_1 = 2Y_2$ yields the contradiction $Y_1 = Y_2$ unless $\mathbb{F}$ has characteristic 2, in which case lack of distinctness forces A to be the zero matrix.

(It is worth noting that something of the flavour of this proof can be recaptured over the infinite fields $\mathbb{R}$ (and $\mathbb{C}$) by the following topological approach. $G(n,\mathbb{R})$ is an open, dense subset of $M(n,\mathbb{R})$ and imT_A, being essentially just a translate of $G(n,\mathbb{R})$, is also open and dense. Thus $G(n,\mathbb{R}) \cap imT_A$ shares these properties and supplies an open, dense set of candidates with which to split A in the required manner.)

We are however still left with the stubborn case $q = 2$. The necessity to adopt a fresh approach here will in fact provide an alternative proof for the previous cases as well. Our motivation comes from facing the question: do we need to split every matrix in $M(n,\mathbb{F})$ or can we manage with just splitting some suitably representative (or canonical) matrices? At this point recall (for example, from [**1**, p. 167]) that there are matrices P, Q in $G(n,\mathbb{F})$ such that $PAQ = \begin{pmatrix} I_k & 0 \\ 0 & 0 \end{pmatrix}$. Thus it is enough to effect a splitting of the matrices $\begin{pmatrix} I_k & 0 \\ 0 & 0 \end{pmatrix}$ for $k = 1,2,\dots,n$. (The zero matrix has already been dealt with.) If $q > 2$ (or if $\mathbb{F}$ is infinite) we can choose any a in $\mathbb{F}$ with $a \neq 0,1$ to give:

$$\begin{pmatrix} I_k & 0 \\ 0 & 0 \end{pmatrix} = \left(\begin{array}{c|c} (1-a)I_k & 0 \\ \hline 0 & -aI_{n-k} \end{array}\right) + aI_n$$

with both matrices easily seen to be invertible and distinct. If $q = 2$ we consider the cases of even k and odd k separately. If $k = 2r$ is even then:

$$\begin{pmatrix} I_k & 0 \\ 0 & 0 \end{pmatrix} = \left(\begin{array}{cc|c} I_r & I_r & 0 \\ I_r & 0 & \\ \hline \multicolumn{2}{c|}{0} & I_{n-k} \end{array}\right) + \left(\begin{array}{cc|c} 0 & I_r & 0 \\ I_r & I_r & \\ \hline \multicolumn{2}{c|}{0} & I_{n-k} \end{array}\right)$$

is a suitable splitting. If k is odd, it is enough to provide a splitting of I_3 since we can then abut the splitting of I_{k-3} given above. But

$$I_3 = \begin{pmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix}$$

clinches the matter. Finally, we consider uniqueness of the splitting. From our previous discussion, for there to be any chance of uniqueness q must be 2 and A must be nonzero. Uniqueness then forces the number of nonzero elements of $M(n,2)$ to be the same as the number of unordered pairs of distinct elements of $G(n,2)$. That is, $|M(n,2)| - 1 = \frac{1}{2}|G(n,2)|(|G(n,2)| - 1)$. The last equation can be quickly shown to hold only when $n = 2$.

Reference

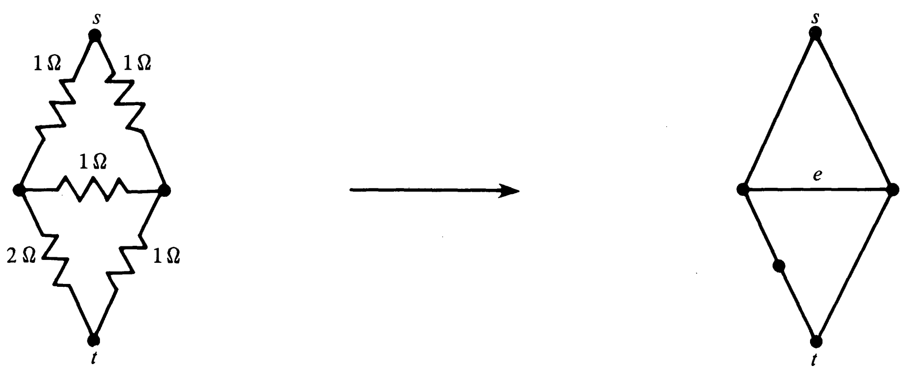
[1] S. Lipschutz, Linear Algebra, McGraw-Hill, 1974.

An Electrical Lemma

LOUIS W. SHAPIRO
Howard University
Washington, D.C. 20059

Let s and t be distinct vertices of a connected graph G. The notation and definitions used here follow the excellent text of Bollabás [1]. Next let N be the number of spanning trees of G while $F(s,t)=F$ is the number of spanning forests with two components, one containing s and one containing t. Call such a forest a *thicket*. The lemma in question states that $R_{s,t}=F/N$ where $R_{s,t}$ is the resistance of G between s and t if each edge of G represents a one ohm resistor. Since the right-hand side is a strictly graphical quantity this is a fascinating result which deserves to be better known.

By way of illustration, look at the smallest example not immediately solved by the series and parallel laws.



Note the adjustment for the two-ohm resistor. Some counting (consider separately the cases where the edge e is and is not present) yields 11 spanning trees and 13 thickets and thus

$$R_{s,t}=13/11 \text{ ohms}.$$

Given an edge ab in G let $\mathscr{N}(s,a\to b,t)$ be the set of spanning trees such that the tree has a path from s to t of the form $s,v_2,\ldots,a,b,\ldots,v_{m-1},t$ where $a=s$ or $b=t$ is allowed. $\mathscr{N}(s,b\to a,t)$ is defined similarly. Let $N(s,a\to b,t)=|\mathscr{N}(s,a\to b,t)|$ be the cardinality of $\mathscr{N}(s,a\to b,t)$. If every edge of G represents a 1-ohm resistor then we have the following result:

LEMMA 1. *Put a 1-ampere current between s and t and let* $w_{ab}=[N(s,a\to b,t)-N(s,b\to a,t)]/N$. *Then* w_{ab} *is the current in the edge ab oriented from a to b.*

Before the proof, we recall some facts about circuits. A simple electrical circuit can be regarded as a graph where each edge ab has a nonnegative number associated with it called its resistance, r_{ab}. We have current entering the circuit at s and leaving at t. This gives a voltage (or potential) at each vertex a denoted v_a. For an edge ab the voltage (difference) is defined to be $v_{ab}=v_a-v_b$. There will also be a current, w_{ab}, flowing through this edge and Ohm's law specifies that

$$v_{ab}=w_{ab}r_{ab}.$$

Orient the edges of G in any way. Then note that $v_{ab} = -v_{ba}$ and that $w_{ab} = -w_{ba}$.

Voltages and currents must also satisfy Kirkhoff's laws.

KIRKHOFF'S CURRENT LAW states that the sum of the currents out of any vertex a is zero. If $a_1, a_2, \ldots, a_n$ are the edges incident with a then,

$$w_{a_1} + w_{a_2} + \cdots + w_{a_n} = 0. \tag{C}$$

KIRKHOFF'S VOLTAGE LAW states that the sum of the voltages around any cycle $x_1x_2, x_2x_3, \ldots, x_nx_1$ must be zero. Then,

$$v_{x_1x_2} + v_{x_2x_3} + \cdots + v_{x_nx_1} = 0.$$

Note that if the resistance in each edge is 1 ohm, then we obtain

$$w_{x_1x_2} + w_{x_2x_3} + \cdots + w_{x_nx_1} = 0 \tag{V}$$

for the voltage law, and in this case both of Kirkhoff's laws can be stated in terms of currents. We will show that defining w_{ab} as $(1/N)(N(s, a \to b, t) - N(s, b \to a, t))$ implies that (V) and (C) are satisfied. A lifetime of turning on light bulbs or an induction proof, say, on the number of edges, shows the uniqueness of the currents in the circuit.

Proof of Lemma 1. For each thicket, S, define a function f_S on the edges of G as follows:

$$f_S(ab) = \begin{cases} 1 & \text{if } a \text{ is in } S_s,\ b \text{ is in } S_t \\ -1 & \text{if } a \text{ is in } S_t,\ b \text{ is in } S_s \\ 0 & \text{otherwise,} \end{cases}$$

where $S = S_s \cup S_t$ is a thicket. Let $\mathscr{S} = \{S | S \text{ is a thicket}\}$. Note that $f_S(ab) = 1$ if and only if $S \cup (ab)$ is a spanning tree in the set $\mathscr{N}(s, a \to b, t)$ while -1 indicates an element of $\mathscr{N}(s, b \to a, t)$. Let C be a circuit in G, and notice that as you traverse this circuit, you pass from S_s to S_t just as often as you pass from S_t to S_s. This yields the first equality in

$$\begin{aligned} 0 = \sum_{s \in \mathscr{S}} \sum_{ab \in C} f_S(ab) &= \sum_{ab \in C} \sum_{S \in \mathscr{S}} f_S(ab) \\ &= \sum_{ab \in C} N(s, a \to b, t) - N(s, b \to a, t) \\ &= \frac{1}{N} \sum_{ab \in C} N(s, a \to b, t) - N(s, b \to a, t) \\ &= \sum_{ab \in C} w_{ab}. \end{aligned}$$

To show that Kirkhoff's current law holds, let x be any vertex of G other than s or t. Let S be any spanning tree of G such that x is on the path in S between s and t. If this path consists of $s, \ldots, q, x, r, \ldots, t$ then we obtain two thickets from S by removing either the edge qx or the edge xr. Call the thickets obtained T and $\overline{T}$, respectively. Then $f_T(qx) = f_{\overline{T}}(xr)$ or $f_T(xq) + f_{\overline{T}}(xr) = 0$. Summing over all spanning trees with x on the path between s and t yields

$$w_{x1} + w_{x2} + \cdots + w_{xn} = 0,$$

where the vertices adjacent to x are labeled $x1, x2, \ldots, xn$.

It remains only to verify the current law at s and t. Let $A(s)$ be all the vertices adjacent to s. Then on the one hand $N = \Sigma_{b \in A(s)} N(s, s \to b, t)$, while on the other hand $0 = N(s, b \to s, t)$ for any $b \in A(s)$. This can be rewritten as:

$$1 = (1/N) \sum_{b \in A(s)} N(s, s \to b, t) - N(s, b \to s, t) = \sum_{b \in A(s)} w_{sb}.$$

But the external power source is bringing in 1-ampere of current (or starting at s the current flow to the external power source is -1). Thus, we have

$$0 = \sum w_{sb} - 1 \quad \text{verifying the current law at } s.$$

The proof at t is very similar and this finishes the proof of the lemma.

Now, for the proof of the main theorem. Let P be a path from s to t and note that for any thicket S, P goes from S_s to S_t one more time than it goes from S_t to S_s. This justifies the second equality in

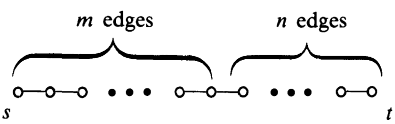
$$\begin{aligned} F &= \sum_{S \in \mathcal{S}} 1 = \sum_{S \in \mathcal{S}} \sum_{ab \in P} f_s(ab) \\ &= \sum_{ab \in P} \sum_{S \in \mathcal{S}} f_s(ab) = \sum_{ab \in P} N(s, a \to b, t) - N(s, b \to a, t) \\ &= \sum_{ab \in P} N w_{ab} = N \sum_{ab \in P} v_{ab}. \end{aligned}$$

The fifth equality follows from the definition of w_{ab}. The last equality follows from the lemma and the use of one ohm resistors.

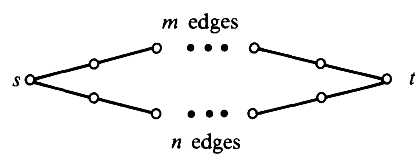
Thus $F/N = \Sigma_{ab \in P} v_{ab} = R_{st}$, since a rephrasing of Kirkhoff's voltage law says that R_{st} is independent of path.

The following graphs give interesting results when the Electrical Lemma is applied.

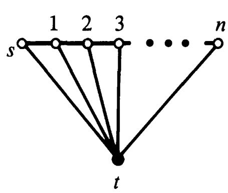
G_1:



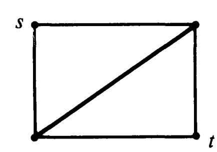
G_2:



G_3:



G_4:



In fact, G_1 and G_2 give the entering wedge towards proving the series and parallel laws. Suitable combinations of G_1 and G_2 can be used to prove the series and parallel laws for resistors whose value is rational. Continuity could then be used to remove this rationality restriction.

We would also like to remove the restriction on 1-ohm resistors. This can be done by associating the tree product $\Sigma_{ab \in T} r_{ab}$ with any tree. Any thicket, S, has associated with it $\Sigma_{ab \in S} r_{ab} \cdot \Pi_{xy \in C} r_{xy}$ with C the set of all edges connecting S_s to S_t.

Then N is replaced by $\Sigma_{\text{all trees}}$ (tree products) and F is replaced by $\Sigma_{\text{all thickets}}$ (thicket products).

Reference

[1] Bela Bollabás, Graph Theory, Springer, New York, 1979, p. 31–33.

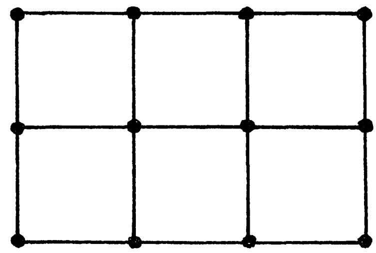
LOREN C. LARSON, Editor
BRUCE HANSON, Associate Editor
St. Olaf College

Proposals

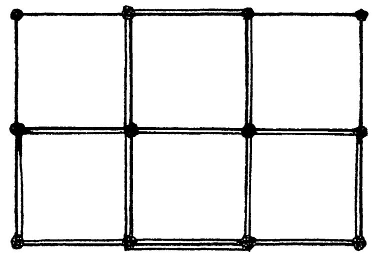
To be considered for publication, solutions should be received by July 2, 1987.

1257. *Proposed by Bruce Reznick, University of Illinois, Urbana-Champaign.*

Let $G(r,s)$ denote the grid of rs points arranged in r rows and s columns with adjacent points connected, r, $s \geqslant 2$. (See illustration of $G(3,4)$.)



Let $R(r,s)$ denote the minimum number of rectangles needed to cover the edges and vertices of $G(r,s)$. The accompanying illustration shows that $R(3,4) \leqslant 3$.



It is easy to check that $G(3,4)$ cannot be covered by two rectangles, so $R(3,4)=3$.
Find a simple expression for $R(r,s)$.

1258. *Proposed by Jerrold W. Grossman, Oakland University, Michigan.*
Let $s_1, s_2, s_3, \dots$ be the sequence of positive integers (in increasing order) whose binary representation contains no two consecutive ones. The sequence begins (in binary notation) 1, 10, 100, 101,... . How can one efficiently determine s_n from n and vice versa?

1259. *Proposed by John P. Hoyt, Lancaster, Pennsylvania.*
Given two nonzero integers a and b with $|a| \neq |b|$, define a sequence $(b_n)_{n=1}^{\infty}$ by the recurrence

$$b_1 = b, \qquad b_2 = b^2 - a^2, \quad \text{and} \quad b_{k+2} = \frac{b_{k+1}^2 - a_{k+1}^2}{b_k} \text{ for } k \geqslant 1.$$

Show that b_n is an integer for all $n \geqslant 1$.

1260. *Proposed by William P. Wardlaw, U.S. Naval Academy, Annapolis, Maryland.*
Among all ways of writing a given permutation in S_n as a product of cycles, does the disjoint cycle representation use the fewest number of nontrivial cycles? If so, give a proof; if not, describe a correct minimal representation.

1261. *Proposed by Stanley Rabinowitz, Digital Equipment Corporation, Nashua, New Hampshire.*
a) What is the area of the smallest triangle with integral sides and integral area?
*b) What is the volume of the smallest tetrahedron with integral sides and integral volume?

Quickies

Solutions to the Quickies appear on p. 50.

Q717. *Submitted by M. S. Klamkin, University of Alberta.*
Are there any integral solutions to the Diophantine equation

$$x_1^{1987} + 2x_2^{1987} + 4x_3^{1987} + \cdots + 2^{1986}x_{1987}^{1987} = 1986x_1x_2 \cdots x_{1987}?$$

(Note: This problem is an extension of a problem from the Wisconsin Talent Search.)

Q718. *Submitted by Bjorn Poonen, (student), Harvard College.*
Suppose $f(x)$ and $g(x)$ are nonzero real polynomials satisfying $f(x^2+x+1) = g(x)f(x)$. Show that $f(x)$ has even degree.

Solutions

Augmented Nonsingular Matrix **February 1986**

1231. *Proposed by Martin Feuerman, New Jersey Medical College, Newark.*

Let A be a $t \times t$ real symmetric matrix of rank $t-1$ such that $A\mathbf{1} = 0$, where $\mathbf{1}$ is the $t \times 1$ vector with each element equal to 1, and let

$$A^* = \begin{pmatrix} A & \mathbf{1} \\ \mathbf{1}' & 0 \end{pmatrix}.$$

(The prime denotes transpose.) Prove that A^* is nonsingular.

I. *Solution by Luz M. DeAlba, Drake University.*

Assume that

$$\begin{pmatrix} A & \mathbf{1} \\ \mathbf{1}' & 0 \end{pmatrix}\begin{pmatrix} \mathbf{x} \\ k \end{pmatrix} = 0$$

for some $t-1$ column vector $\mathbf{x}$, and some real number k. Then matrix multiplication shows that

$$A\mathbf{x} + k\mathbf{1} = 0,$$
$$\mathbf{1}'\mathbf{x} = 0.$$

Multiply the first of these equations on the left by $\mathbf{1}'$ to get

$$\mathbf{1}'A\mathbf{x} + \mathbf{1}'k\mathbf{1} = 0.$$

Since A is symmetric, $\mathbf{1}'A = (A\mathbf{1})' = 0$, and, therefore, $k\mathbf{1}'\mathbf{1} = 0$. Thus $k = 0$, and it follows that $A\mathbf{x} = 0$. Since the rank of A is $t-1$, the nullity of A is 1. Therefore, $A\mathbf{x} = 0$ is satisfied only for $\mathbf{x} = a\mathbf{1}$ for some constant a. But then from $\mathbf{1}'\mathbf{x} = 0$ we get $\mathbf{1}'a\mathbf{1} = a\mathbf{1}'\mathbf{1} = 0$ and this implies that $a = 0$. Thus the nullity of A^* is 0, and A^* is nonsingular.

II. *Solution by Mike Bolla, Torrance, California, and Dan Kalman, Rancho Palos Verdes, California.*

In general, if $\mathbf{x}$ is a nonzero $t \times 1$ vector such that $A\mathbf{x} = 0$, the matrix $\begin{pmatrix} A \\ \mathbf{x}' \end{pmatrix}$ has rank 1 greater than A, for $\mathbf{x}'$ is orthogonal to all rows of A. In the case at hand, $\begin{pmatrix} A \\ \mathbf{1}' \end{pmatrix}$ has rank t, whence its transpose $(A \quad \mathbf{1})$ has rank t also. Now $\begin{pmatrix} \mathbf{1} \\ 0 \end{pmatrix}$ is evidently a null vector of $(A \quad \mathbf{1})$. By the preceding general remarks, the rank of $\begin{pmatrix} A & \mathbf{1} \\ \mathbf{1}' & 0 \end{pmatrix}$ is $t+1$, so this last matrix is invertible. Note that $\mathbf{1}$ may be replaced by any nonzero null vector $\mathbf{v}$ of the symmetric matrix A, and $\begin{pmatrix} A & \mathbf{v} \\ \mathbf{v}' & 0 \end{pmatrix}$ will have rank 2 more than A.

As a generalization, if A is symmetric and $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$ is an orthogonal basis for its null space, then the matrix

$$\begin{pmatrix} A & \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_k \\ \mathbf{v}_1' & 0 & 0 & \cdots & 0 \\ \mathbf{v}_2' & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ \mathbf{v}_k' & 0 & 0 & & 0 \end{pmatrix}$$

is invertible.

III. *Solution by John O. Bennett, Linton, Indiana.*

Add the first $t-1$ rows of A^* to the t-th row and then add the first $t-1$ columns to the t-th column, to get the matrix

$$B^* = \left(\begin{array}{cccc|cc} & & & & 0 & 1 \\ & A_{t-1} & & & 0 & 1 \\ & & & & \vdots & \vdots \\ & & & & 0 & 1 \\ \hline 0 & 0 & \cdots & 0 & 0 & t \\ 1 & 1 & \cdots & 1 & t & 0 \end{array}\right),$$

where A_{t-1} is the $(t-1)\times(t-1)$ submatrix of A obtained by leaving out row t and column t. Since A has rank $t-1$, A_{t-1} is nonsingular. Now $\det A^* = \det B^*$, and by cofactor expansion, $\det B^* = -t^2 \det A_{t-1} \neq 0$.

IV. *Solution by Harvey Schmidt, Jr., Lewis and Clark College.*

Since A is symmetric and $A\mathbf{1}=0$, it is immediate that

$$(A^*)^2 = \begin{pmatrix} A^2+E & 0 \\ 0 & t \end{pmatrix},$$

where E is the $t\times t$ matrix with each entry equal to 1. Clearly it suffices to show that $\operatorname{rank}(A^2+E)=t$, for then $\operatorname{rank} A^* \geqslant \operatorname{rank}(A^*)^2 = t+1$.

If $\mathbf{v}_1,\dots,\mathbf{v}_t$ is a basis consisting of eigenvectors of A with associated eigenvalues $\lambda_1=0$, $\lambda_2,\dots,\lambda_t$ and $\mathbf{v}_1=\mathbf{1}$, let β_i be the sum of the coordinates of $\mathbf{v}_i$. Then $E\mathbf{v}_i=\beta_i\mathbf{1}=\beta_i\mathbf{v}_i$. Hence, with respect to the basis $\mathbf{v}_1,\dots,\mathbf{v}_t$, the matrix for A^2+E is

$$\begin{pmatrix} \beta_1 & \beta_2 & \beta_3\cdots & \beta_t \\ 0 & \lambda_2^2 & 0\cdots & 0 \\ 0 & 0 & \lambda_3^2\cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0\cdots & \lambda_t^2 \end{pmatrix}$$

which clearly has rank t since $\beta_1 = t \neq 0$.

V. *Solution by J. S. Frame, Michigan State University.*

The symmetric idempotent matrix $E=\mathbf{1}\cdot\mathbf{1}'/t$ is orthogonal to A, so $B=A+E$ has rank t, is invertible, and satisfies

$$BE=E^2=E=EB=B^{-1}E, \quad B\cdot\mathbf{1}=\mathbf{1}=B^{-1}\cdot\mathbf{1}.$$

Hence

$$\begin{pmatrix} B^{-1}-E & \mathbf{1}/t \\ \mathbf{1}'/t & 0 \end{pmatrix}\begin{pmatrix} A & \mathbf{1} \\ \mathbf{1}' & 0 \end{pmatrix} = \begin{pmatrix} (B^{-1}-E)(B-E)+E & 0 \\ 0 & \mathbf{1}'\cdot\mathbf{1}/t \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & 1 \end{pmatrix}.$$

Thus A^* has an inverse and is nonsingular.

Also solved by Mike Bolla and Dan Kalman (two additional solutions), François Dubeau (Canada), Thomas E. Elsner, J. S. Frame (an additional solution based on considering the Moore-Penrose pseudoinverse of A), Stanley Gudder, David Hill, María Ascension López Chamorro (Spain), Kathleen Lewis, J. C. Linders (The Netherlands), N. J. Lord (England), Marvin Marcus, William A. Newcomb, Donald H. Pelletier, Bjorn Poonen (student), João Filipe Queiró (Portugal), Bruce Richter, Allen J. Schwenk, Daniel B. Shapiro, Michiel Smid (The Netherlands), Leo Thurston, William P. Wardlaw, Western Maryland College Problem Group, Edward T. Wong, Yan-loi Wong (student; two solutions), Ken Yanosko, and the proposer.

Several noted that A need not be symmetric for the result to hold. Elsner, for instance, showed that $\begin{pmatrix} A & \mathbf{u} \\ \mathbf{v}' & k \end{pmatrix}$, where **u** and **v** are $t \times 1$ vectors and k is a scalar, is nonsingular if and only if (i) **u** is not in the range of A, and (ii) the null space of A and $\mathbf{v}'$ intersect trivially. Queiró showed that if s and t are natural numbers, $s \le t$, and A is a $t \times t$ complex hermitian matrix of rank $t-s$ such that $AB = 0$, where B is a $t \times s$ full complex matrix (i.e., rank $B = s$), then $\begin{pmatrix} A & B \\ B' & 0 \end{pmatrix}$ is nonsingular.

Shapiro provided the following extension: Let F be a field and A an $n \times n$ matrix over F of rank $n-1$. Let $\mathbf{v}_0, \mathbf{w}_0$ be nonzero column vectors in F^n satisfying $A\mathbf{v}_0 = 0$ and $\mathbf{w}_0'A = 0$. Let $\tilde{A} = \begin{pmatrix} A & \mathbf{x} \\ \mathbf{y}' & \delta \end{pmatrix}$ for some $\mathbf{x}, \mathbf{y}$ in F^n and some δ in F. Then $\tilde{A}$ is nonsingular if and only if $\mathbf{y}'\mathbf{v}_0 \neq 0$ and $\mathbf{w}_0'\mathbf{x} \neq 0$. He generalized this in the following way. Let V, W be vector spaces over F. A linear map T: $V \oplus W \to V \oplus W$ can be represented as a matrix $T = \begin{pmatrix} f & g \\ h & l \end{pmatrix}$ where f: $V \to V$, g: $W \to V$, h: $V \to W$, and l: $W \to W$ are linear maps. Theorem: Let V, W be vector spaces over a field F and let $T = \begin{pmatrix} f & g \\ h & l \end{pmatrix}$ as above. If T is nonsingular, then $\dim(\ker f) \le \dim W$. Moreover, T is nonsingular and $\dim(\ker f) = \dim W$ if and only if $\operatorname{im} f \cap \operatorname{im} g = 0$, $\ker g = 0$, and $\ker f \cap \ker h = 0$.

The Euler Line and an Angle Bisector — February 1986

1232. *Proposed by J. T. Groenman, Arnhem, and D. J. Smeenk, The Netherlands.*

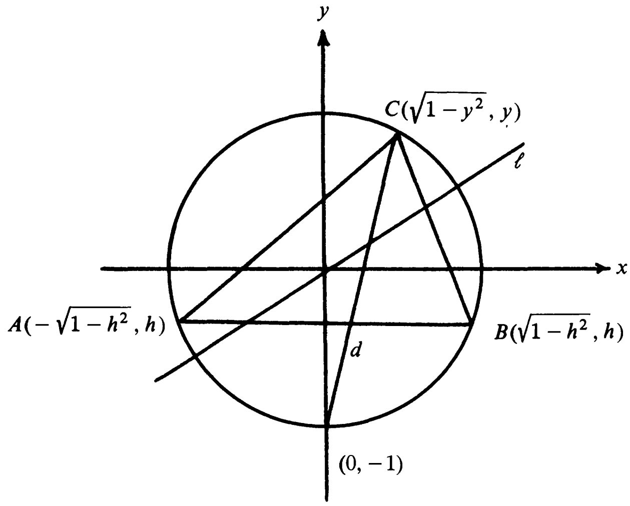
Let l be the Euler line of the nonisosceles triangle ABC (with sides a, b, c and angles α, β, γ), and let d be the internal angle bisector of γ. (The Euler line of a triangle contains the centroid, circumcenter, and orthocenter.) Prove that:

(a) l is perpendicular to d if and only if $\gamma = \pi/3$; and

(b) l is parallel to d if and only if $\gamma = 2\pi/3$.

I. *Solution by Richard E. Pfiefer, San Jose State University.*

Without loss of generality, we may assume that the circumradius $R = 1$, the circumcenter is at the origin $(0,0)$, and we may assign coordinates as shown in the diagram. Here $-1 < h < 1$ and $h < y < 1$.



The centroid has coordinates $(1/3)(\sqrt{1-y^2}, 2h+y) \neq (0,0)$, so the Euler line has slope $(2h+y)/\sqrt{1-y^2}$ and line d has slope $(y+1)/\sqrt{1-y^2}$.

(a) l is perpendicular to d if and only if

$$\begin{aligned} (2h+y)(y+1) + (1-y^2) &= 0, \\ 2h(y+1) + (y+1) &= 0, \\ h &= -1/2, \\ \gamma &= \pi/3. \end{aligned}$$

(b) l is parallel to d if and only if

$$2h + y = y + 1,$$
$$h = 1/2,$$
$$\gamma = 2\pi/3.$$

II. *Solution by William A. Newcomb, Lawrence Livermore National Laboratory.*

View the problem in the Gaussian complex plane, and let the Gaussian coordinate of an arbitrary point P be written as $z(P)$. Without loss of generality, we can choose the unit circle $|z| = 1$ for the circumcircle and orient the axes so that $z(C) = -1$. We then have

$$z(A) = e^{i\theta} \quad \text{and} \quad z(B) = e^{i\phi},$$

say, where $-\pi < \theta, \phi < \pi$. Since the angle γ is inscribed in the arc joining A and B (and not containing C), we see that $\gamma = (1/2)|\theta - \phi|$. If D is the midpoint of this arc, then $z(D) = e^{i\xi}$, where $\xi = (\theta + \phi)/2$, and the line CD is the internal bisector d of angle γ.

The circumcenter is at the origin O, and the Euler line l is the line joining O to M, where $z(M) = z(A) + z(B) + z(C) = e^{i\theta} + e^{i\phi} - 1 = 2\cos\gamma e^{i\xi} - 1$. (The centroid of the triangle is 1/3 of the way from O to M.) The criterion of orthogonality of d and l is $\mathrm{Re}(W) = 0$, and the criterion of parallelism is $\mathrm{Im}(W) = 0$, where

$$\begin{aligned} W &= (z(D) - z(C))\bar{z}(M) \\ &= (e^{i\xi} + 1)(2\cos\gamma e^{-i\xi} - 1) \\ &= (1 + \cos\xi)(2\cos\gamma - 1) - i\sin\xi(2\cos\gamma + 1). \end{aligned}$$

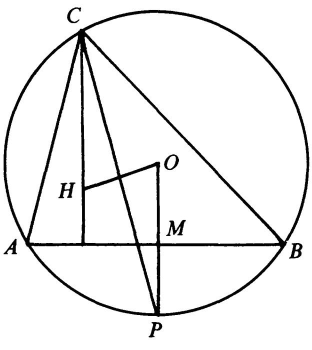
Since $-\pi < \xi < \pi$ and $\xi \neq 0$, neither $\sin\xi$ nor $1 + \cos\xi$ can vanish. Both statements now follow directly.

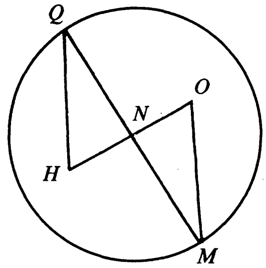
III. *Solution by Bjorn Poonen, student, Harvard College.*

Let H be the orthocenter and O the circumcenter of triangle ABC. Let d intersect the circumcircle at P. Finally let M be the midpoint of AB and let Q be the midpoint of CH.

First we prove the following claim: $\mathbf{CH} = 2\mathbf{OM}$. ($\mathbf{CH}$ and $\mathbf{OM}$ are directed line segments.)

Proof. Let N be the center of the nine-point circle, which passes through M and Q. Then $HN = NO$, and $HQ \| MO$, since both segments are perpendicular to AB. Hence triangle QHN is similar to triangle MON, so $\mathbf{CH} = 2\mathbf{QH} = 2\mathbf{OM}$, as described. (Lines QH and OM do not coincide since triangle ABC is nonisosceles.)





Now $\angle AOP = 2\angle ACP = \gamma$, so $\mathbf{OM} = (\cos\gamma)\mathbf{OP}$. Thus from our first claim, $\gamma = \pi/3$ if and only if $\mathbf{CH} = \mathbf{OP}$, and $\gamma = 2\pi/3$ if and only if $\mathbf{CH} = -\mathbf{OP}$.

If $\mathbf{CH} = \mathbf{OP}$, then $CHPO$ is a parallelogram with $CO = OP$, so $CHPO$ is a rhombus, and its diagonals (l and d) must be perpendicular. Conversely, if l is perpendicular to d, then $\mathbf{CH} = \mathbf{OP}$, since there is only one point on line CH which lies on the line perpendicular to CP

passing through O. This proves (a).

Also $\mathbf{CH} = -\mathbf{OP}$ if and only if $CHOP$ is a parallelogram, which is true if and only if l is parallel to d, so we are done.

Also solved by Francisco Bellot (Spain), J. Heuver (Canada), Tan Shiu Kin (Hong Kong), L. Kuipers (Switzerland), J. C. Linders (The Netherlands), N. J. Lord (England; solution with vectors methods), Helen M. Marston, David Morin (two solutions), J. M. Stark, Michael Vowe (Switzerland), and the proposers.

A Rational Approximation of $\log^2(1+x)$ — February 1986

1233. *Proposed by Robert E. Shafer, Berkeley, California.*

Prove that if $x > -1$ and $x \neq 0$, then

$$\frac{x^2}{1+x+\dfrac{x^2}{12}-\dfrac{\dfrac{x^4}{120}}{1+x+\dfrac{31}{252}x^2}} < \log^2(1+x) < \frac{x^2}{1+x+\dfrac{x^2}{12}-\dfrac{\dfrac{x^4}{240}}{1+x+\dfrac{1}{20}x^2}}.$$

Solution by Paul Bracken, Toronto, Canada.

Write the right-most inequality in the form

$$\log^2(1+x) < \left(\frac{x^2}{1+x}\right)\left(\frac{1+x+\dfrac{x^2}{20}}{1+x+\dfrac{2x^2}{15}}\right), \qquad x > -1.$$

Introduce the variable u through the transformation $1 + x = e^u$, and note that $x = 0$ corresponds to $u = 0$. The proposed inequality is

$$u^2 < (e^u + e^{-u} - 2)\left(\frac{1+\frac{1}{20}(e^u+e^{-u}-2)}{1+\frac{2}{15}(e^u+e^{-u}-2)}\right). \tag{1}$$

Now using the series expansions for the exponential functions we get

$$e^u + e^{-u} - 2 = 2\sum_{k=1}^{\infty}\frac{u^{2k}}{(2k)!},$$

$$(e^u + e^{-u} - 2)^2 = 2\sum_{k=1}^{\infty}\frac{(2u)^{2k}}{(2k)!} - 8\sum_{k=1}^{\infty}\frac{u^{2k}}{(2k)!}.$$

Using these, we see that (1) takes the form

$$u^2\left(1+\frac{2}{15}\left(2\sum_{k=1}^{\infty}\frac{u^{2k}}{(2k)!}\right)\right) < 2\sum_{k=1}^{\infty}\frac{u^{2k}}{(2k)!} + \frac{1}{20}\left(2\sum_{k=1}^{\infty}\frac{(2u)^{2k}}{(2k)!} - 8\sum_{k=1}^{\infty}\frac{u^{2k}}{(2k)!}\right)$$

and this is equivalent to

$$\sum_{k=1}^{\infty}\frac{2}{(2k)!}\left(\frac{1}{(2k+2)(2k+1)} + \frac{1}{5}\left(\frac{2^{2k}-1}{(2k+2)(2k+1)} - \frac{2}{3}\right)\right)u^{2k+2} \equiv \sum_{k=1}^{\infty}\frac{2}{(2k)!}a_k u^{2k+2} > 0.$$

For all $u \neq 0$, $u^{2k+2} > 0$, so the above inequality will be valid if for all k, $a_k \geqslant 0$ with at least one $a_k > 0$. By putting $k = 1, 2, 3$, we find that $a_1 = a_2 = a_3 = 0$ and for all $k \geqslant 3$,

$$2^{2k} - 1 > (2k+2)(2k+1).$$

This implies that $a_k > 0$ for all $k > 3$.

For the leftmost inequality we again set $e^u = 1 + x$ and put the proposed inequality into the form

$$(e^u+e^{-u}-2)\left(1+\frac{31}{252}(e^u+e^{-u}-2)\right)<u^2\left(1+\frac{13}{63}(e^u+e^{-u}-2)+\frac{23}{60\cdot 63}(e^u+e^{-u}-2)^2\right).$$

Using the series expansions of $e^u+e^{-u}-2$ and $(e^u+e^{-u}-2)^2$ we get

$$\sum_{k=1}^{\infty}\left[-\frac{2}{(2k+2)!}-\frac{31}{126}\left(\frac{2^{2k+2}}{(2k+2)!}-\frac{4}{(2k+2)!}\right)\right.$$
$$\left.+\frac{26}{63}\frac{1}{(2k)!}+\frac{23}{30\cdot 63}\left(\frac{2^{2k+2}}{4(2k)!}-\frac{4}{(2k)!}\right)\right]u^{2k+2}>0,$$

$$\sum_{k=1}^{\infty}\frac{2}{(2k)!}\left[\left(-\frac{1}{(2k+2)(2k+1)}+\frac{13}{63}\right)\right.$$
$$\left.+\frac{1}{63}\left(\left(\frac{23}{15\cdot 4}-\frac{31}{(2k+2)(2k+1)}\right)2^{2k}-\left(\frac{23}{15}-\frac{31}{(2k+2)(2k+1)}\right)\right)\right]u^{2k+2}$$
$$\equiv\sum_{k=1}^{\infty}\frac{2}{(2k)!}b_k u^{2k+2}>0.$$

Substitute $k=1,2,3$ into b_k and find that $b_1=b_2=b_3=0$. For $k=4$,

$$\frac{13}{63}>\frac{1}{(2k+2)(2k+1)},$$

and for $k\geqslant 4$,

$$\left(\frac{23}{15\cdot 4}-\frac{31}{(2k+2)(2k+1)}\right)2^{2k}>2>\frac{23}{15}.$$

These equations imply that $b_k>0$ for all $k\geqslant 4$, and the proof is complete.

Also solved by Kee-wai Lau (Hong Kong), and the proposer.

Sorted Integers **February 1986**

1234. *Proposed by the Computer Science Problem Seminar, Stanford University.*

A positive integer is said to be "sorted" if the digits in its decimal notation are nondecreasing from left to right.

(a) Let x be any integer whose decimal notation consists of an arbitrary number of 3's followed by an arbitrary number of 6's followed by a single 7. Prove that x^2 is sorted. For example, $33366667^2=1113334466688889$.

(b)* Which positive integers x are such that both x and x^2 are sorted?

(a) *Solution by Roger B. Nelsen, Lewis and Clark College.*

We shall use the following notation (borrowed from the chemists) in what follows: denote repeated digits by subscripts, i.e., the example in this problem has $x=3_3 6_4 7_1$ with $x^2=1_3 3_3 4_2 6_3 8_4 9_1$.

Suppose $x=3_k 6_n 7_1$ where $k\geqslant 0$ and $n\geqslant 0$. We will prove that

$$x^2=\begin{cases}1_k 3_k 4_{n-k+1} 6_k 8_n 9_1, & \text{if } n+1\geqslant k;\\ 1_k 3_{n+1} 5_{k-n-1} 6_{n+1} 8_n 9_1, & \text{if } n+1<k;\end{cases}$$

and thus x^2 is sorted.

Proof. If $x=3_k 6_n 7_1$, then

$$x=3\cdot 10^{n+1}\left(\frac{10^k-1}{9}\right)+60\left(\frac{10^n-1}{9}\right)+7,$$

so that

$$3x=10^{n+k+1}+10^{n+1}+1.$$

Then

$$9x^2=10^{2n+2k+2}+2\cdot10^{2n+k+2}+10^{2n+2}+2\cdot10^{n+k+1}+2\cdot10^{n+1}+1,$$

and, hence,

$$\begin{aligned} x^2 &= \left(\frac{10^{2n+2k+2}}{9}\right)+2\left(\frac{10^{2n+k+2}-1}{9}\right)+\left(\frac{10^{2n+2}-1}{9}\right) \\ &\quad +2\left(\frac{10^{n+k+1}-1}{9}\right)+2\left(\frac{10^{n+1}-1}{9}\right)+1, \\ &= 1_{2n+2k+2}+2_{2n+k+2}+1_{2n+2}+2_{n+k+1}+2_{n+1}+1, \end{aligned}$$

from which the desired conclusion follows.

(b) *Solution by R. Glenn Powers, Western Kentucky University.*

If we assume that both x and x^2 are sorted, the values of x are given by the following results (using the same notation as that introduced in part (a).

(i) If the unit digit of x is one, then $x=1$.
(ii) If the unit digit of x is two, then $x=2$ or $x=12$.
(iii) If the unit digit of x is three, then $x=3$ or $x=13$.
(iv) If the unit digit of x is four, then $x=3_n4_1$ for some $n\geqslant 0$.
(v) If the unit digit of x is five, then $x=15$ or $x=3_n5_1$ for some $n\geqslant 0$.
(vi) If the unit digit of x is six, then $x=6$, 16, or 116.
(vii) If the unit digit of x is seven, then $x=7$, 17, 117, 3_n7_1, 6_n7_1, $1_16_n7_1$, or $3_n6_m7_1$ for some $n\geqslant 1$, $m\geqslant 1$.
(viii) If the unit digit of x is eight, then $x=38$.
(ix) It is impossible for the unit digit of x to be either zero or nine.

In summary, the positive integers x for which both x and x^2 are sorted are 1, 2, 3, 6, 12, 13, 15, 16, 38, 116, 117, and all numbers of the form $1_16_n7_1$, 3_n4_1, 3_n5_1, and $3_n6_m7_1$ for some $n\geqslant 0$, $m\geqslant 0$.

The following proof illustrates the techniques needed to prove the above results.

Proof of (v). Let $x=a_n10^n+\cdots+a_110+5$. Since x is sorted, each digit of x is at most 5. In addition, $x^2=\cdots+(a_1^2+a_1)10^2+2\cdot10+5$. Since x^2 is sorted and a_1 is at most 5, it follows that $a_1=0$, 1, or 3.

If $a_1=0$, then

$$x=5 \quad\text{and}\quad x^2=25.$$

If $a_1=1$, then

$$x=a_n10^n+\cdots+a_210^2+10+5$$

and $a_2,\dots,a_n$ are at most 1. In addition,

$$x^2=\cdots+3a_210^3+2\cdot10^2+2\cdot10+5.$$

So $a_2=0$. Thus $x=15$ and $x^2=225$.

If $a_1=3$, then

$$x=a_n10^n+\cdots+a_210^2+3\cdot10+5$$

and $a_2,\dots,a_n$ are at most 3. If $a_2=\cdots=a_n=3$, then

$$x=3\cdot10^n+\cdots+3\cdot10+5$$

and

$$x^2 = 10^{2n+1} + \cdots + 10^{n+2} + 2 \cdot 10^{n+1} + \cdots + 2 \cdot 10 + 5.$$

Otherwise, there exists $m \geqslant 1$ such that

$$x = a_n 10^n + \cdots + a_{m+1} 10^{m+1} + 3 \cdot 10^m + \cdots + 3 \cdot 10 + 5$$

with $a_{m+1} \neq 3$. In addition,

$$x^2 = \cdots + (7a_{m+1} + 1)10^{m+2} + 2 \cdot 10^{m+2} + \cdots + 2 \cdot 10 + 5.$$

So $a_{m+1} = 0$. Thus $x = 3 \cdot 10^m + \cdots + 3 \cdot 10 + 5$ for some $m \geqslant 1$.

Part (a) was also solved by Robert Bernstein, Daniel Finkel, Gymnasium Bern-Kirchenfeld Problem Solving Group (Switzerland), Stanley Gudder, Humboldt State University '86 Proof Group, Bill Mixon, Al Nicholson, Maroof A. Quidwai (Pakistan), Michael Vowe (Switzerland), Westmont College Problem Solving Group, Samuel Yates, and the proposers.

Part (b) was also solved by James Grochocinski. Partial solutions to (b) were given by Daniel Finkel, Gymnasium Bern-Kirchenfeld Problem Solving Group (Switzerland), Al Nicholson, Bill Mixon, and Roger B. Nelsen.

Gudden showed that if x and y are integers of the form of part (a) then the product xy is sorted. Mixon discovered infinite families of sorted integers with sorted squares in bases 5 and 9, but found no such infinite families in bases 6, 7, or 8.

Two Mountains Without a Valley — February 1986

1235. *Proposed by Ira Rosenholtz, The University of Wyoming.*

The book *Calculus in Vector Spaces* by Lawrence J. Corwin and Robert H. Szczarba contains the following in its discussion of extrema for functions of several variables.

"Suppose f has local maxima at v_1 and v_2. Then f must have another critical point, v_3, because it is impossible to have two mountains without some sort of valley in between. The other critical point can be a saddle point (a pass between the mountains) or a local minimum (a true valley)."

(a) Show that the impossible is possible.

(b) Is the impossible possible for polynomials?

Solution by Bjorn Poonen, student, Harvard College.

(a) We can let $f(x, y) = -(x^2 - 1)^2 - (x^2 - e^y)^2$. Clearly f has absolute maxima at $(1,0)$ and $(-1,0)$. At any critical point (x, y),

$$\frac{\partial f}{\partial y} = 2e^y(x^2 - e^y) = 0, \quad \text{so } x^2 - e^y = 0.$$

Also,

$$\frac{\partial f}{\partial x} = -4x(x^2 - 1) - 4x(x^2 - e^y) = -4x(x^2 - 1) = 0,$$

so $x = 0$, 1, or -1. Then using $x^2 - e^y = 0$ again yields only the solutions $(1,0)$ and $(-1,0)$.

Part (a) was also solved by the proposer, who used the function $f(x, y) = 4x^2e^y - 2x^4 - e^{4y}$. For related material, see Ira Rosenholtz and Lowell Smylie, "'The Only Critical Point in Town' Test", MATHEMATICS MAGAZINE, May 1985, pp. 149–150. No solutions or comments were received for Part (b).

Functional Bounds on Triangle Ratio — February 1986

1236. *Proposed by Mihaly Bencze, Sacele, Romania.*

Let the functions f and g be defined by

$$f(x) = \frac{\pi^2 x}{2\pi^2 + 8x^2} \quad \text{and} \quad g(x) = \frac{8x}{4\pi + \pi x^2} \quad \text{for all real } x.$$

(a) Prove that if A, B, and C are the angles of an acute-angled triangle, and R is its circumradius then

$$f(A)+f(B)+f(C)<\frac{a+b+c}{4R}<g(A)+g(B)+g(C). \tag{1}$$

(b)* Determine functions f and g, where $f(x)$ and $g(x)$ have the form $x/(u+vx^2)$, with u and v real constants, for which the inequalities in (1) are best possible.

Solution by Bjorn Poonen, student, Harvard College.

It is not clear what is meant by the "best possible" inequalities. (For example, an increase in u may permit a decrease in v.) Hence we provide two functions f and two functions g, each of which is good for some types of triangles. From

$$\frac{1}{2R}=\frac{\sin A}{a}=\frac{\sin B}{b}=\frac{\sin C}{c}$$

we find that

$$\frac{a+b+c}{4R}=\frac{\sin A+\sin B+\sin C}{2}.$$

Now $-\sin x$ is a strictly convex function for $x\in[0,\pi/2]$, and the vector (A,B,C) majorizes $(\pi/3,\pi/3,\pi/3)$ and is majorized by $(\pi/2,\pi/2,0)$. (For definitions, see for example, *Inequalities: Theory of Majorization and Its Applications*, Albert W. Marshall and Ingram Olkin, Academic Press, 1979.) Hence by the Hardy-Littlewood-Pólya Majorization Inequality for Convex Functions,

$$\frac{\sin(\pi/2)+\sin(\pi/2)+\sin 0}{2}<\frac{\sin A+\sin B+\sin C}{2}\leqslant\frac{3\sin(\pi/3)}{2}.$$

Also $A+B+C=\pi$, so we can let $f_1(x)=x/\pi$ and $g_1(x)=x/u$, for u slightly less than $4\pi/3\sqrt{3}$. (These are of the desired form, since v can be zero.) These yield strong inequalities for triangles with two nearly right angles, and equilateral triangles, respectively.

From the power series of the sine function, we have

$$(\sin x)/x=1-x^2/6+x^4/120-x^6/5040+\cdots$$

For $x\in[0,\pi/2]$, the ratio of consecutive terms, $-x^2/(2k)(2k+1)$, has absolute value less than 1 for $k\geqslant 1$, so the sum is alternating. Hence,

$$1-x^2/6<(\sin x)/x<1-x^2/6+x^4/120.$$

Then

$$\begin{aligned}(2+x^2/3)(1/2)(\sin x)/x &< (1+x^2/6)(1-x^2/6+x^4/120)\\ &=1-x^4/36+(x^4/120)(1+x^2/6)\\ &<1-x^4\Big(1/36-1/120\big(1+(\pi/2)^2/6\big)\Big)\\ &<1,\end{aligned}$$

so the inequality of the problem holds with $g_2(x)=x/(2+x^2/3)$. This is better than the g given in part (a), since $2+x^2/3>4\pi/8+(\pi/8)x^2$ for all $x\in[0,\pi/2]$.

We can also let $f_2(x)=x/(2+vx^2)$ for any v such that $(2+vx^2)(1/2)(1-x^2/6)>1$ for all $x\in(0,\pi/2)$. This requires that

$$(v/2-1/6)x^2>vx^4/12$$

$$6v-2>vx^2.$$

Since $0<x<\pi/2$, it suffices to have $6v-2\geqslant v(\pi/2)^2$, which is satisfied for all $v\geqslant 8/(24-\pi^2)$. Using $f_2(x)=x/(2+(8/(24-\pi^2))x^2)$ gives a stronger inequality than given by the function f

in part (a) since $8/(24-\pi^2) < 8/\pi^2$.

Part (a) was also solved by Ragnar Dybvik (Norway), L. Kuipers (Switzerland), and the proposer.

Answers

Solutions to the Quickies which appear on p. 40

A717. The only solution is the trivial one $x_1 = x_2 = \cdots = x_{1987} = 0$. This follows by induction and infinite descent. Note that x_1 must be even. Thus $x_1 = 2y_1$ which gives the identical equation in the variables $x_2, x_3, \ldots, x_{1987}, y_1$.

A718. If $f(x)$ had odd degree, $f(x)$ would have at least one real zero, so we could let r be the greatest real zero. But then $f(r^2+r+1) = g(r)f(r) = 0$, and $r^2+r+1 > r$, contradicting the choice of r.

Comments

Q700. (September 1985). Evaluate

$$(n+1)\int_0^x \frac{(x-t)^n}{(1-t)^{n+2}}\,dt \quad \text{for } x<1 \text{ and } n=0,1,2\ldots.$$

Murray Klamkin, University of Alberta, offered the following solution. Denote the expression by I_n and integrate by parts to obtain

$$I_n - I_{n-1} = -x^n, \qquad I_0 = \frac{x}{1-x}.$$

The latter is a telescoping difference equation whose solution is immediate, i.e.,

$$I_n - I_0 = -(x + x^2 + \cdots + x^n),$$

or

$$I_n = \frac{x^{n+1}}{1-x}.$$

H. M. Srivastava, University of Victoria, noted that the integral can be done directly by setting $t = x\sin^2\theta$ followed by binomial expansion. Klamkin generalized the result to $\int_0^x (x-t)^m/(1-t)^n$; let $1-t=p$, expand, and integrate.

1219. (*Proposed May* 1985, *Solution June* 1986)

H. M. Srivastava, University of Victoria, noted that the sums S_1 and S_2 are found in formulas (B) and (C) in D. P. Verma and A. Kaur, "Summation of some series involving Riemann zeta function," *Indian J. Math.*, 25 (1983) pp. 181–184.

REVIEWS

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.

Cartwright, Mark, Crackly phones and the schoolgirl problem, *New Scientist* (3 July 1986) 36-40.

How to transmit data accurately and efficiently? Mathematics offers an answer in the form of error-correcting codes. This article gives a detailed example (the Hamming code of dimension 4), cites the application of communication about Halley's comet from the spacecraft Giotto (a Reed-Solomon code to correct for bursts of errors), and traces the relevant mathematics to balanced incomplete block designs and their origin in Kirkman's schoolgirl problem. All told, a fine example of pure mathematics being done first and applied much later.

Gould, Stephen Jay, Entropic homogeneity isn't why no one hits .400 any more, *Discover* 7:8 (August 1986) 60-66.

Were oldtime baseball players, a number of whom hit .400, far better than the best players of today, who don't even come close? No, says Gould, the "myth of ancient heroes" in fact records the improvement of play over time. The standard deviation of batting averages has regularly declined over the past hundred years, thereby marking a "maturation" of the sport of baseball. "As variation shrinks around a constant mean batting average, .400 hitting disappears. It is, I think, as simple as that." The decline in variation produces a decrease in the difference between average and star performance.

Kenner, Hugh, The several sorts of sorts, *Discover* 7:5 (May 1986) 78-83.

Popular exposition about the various sorting algorithms, focussing on bubblesort and quicksort. The "grabber" that will make people read this article is comical context (sorting dogs at a dog show) plus superb watercolor illustrations (of the dogs and their owners).

Burn, R.P., *Groups: a Path to Geometry*, Cambridge U Pr, 1985; xii + 242 pp, $49.50.

An excellent approach to group theory through organized problem sets, à la Pólya and Szegö's *Problems and Theorems in Analysis*. (Since answers are supplied, the book is suitable for a Moore-method course only if the students don't have copies.) The focus is concrete and geometric: groups of transformations in two and three dimensions. More than 800 exercises are organized into 23 chapters. "The mathematical knowledge assumed for this course is a confident familiarity with high school mathematics," asserts the author (correctly but narrowly). Also necessary are a strong motivation toward, sympathy for, and ability in reasoning.

Fjelstad, Paul, Extending special relativity via the perplex numbers, *American Journal of Physics* 54:5 (May 1986) 416-422.

In analogy to the complex numbers a system of "perplex" numbers $z = x + hy$ is introduced, where the "hallucinatory" h is such that $|h| = -1$. This system, invented by four freshmen in an experimental subcollege of St. Olaf College, appears to have relevance to physics: it provides a natural way to extend the formalism of relativity for superluminal objects (ones travelling faster than the speed of light).

Williams, Ruth, Building blocks for space and time, *New Scientist* (12 June 1986) 48-51.

Einstein's equations for general relativity are not easy to solve; a technique known as the Regge calculus provides a discretiation and hence an approximation scheme for solving these equations. In 1961 T. Regge introduced physicists to the study of simplicial spacetimes, a geodesic-dome-like way of approximating curved spaces from flat pieces. Now some physicists think that the discrete theory may be more fundamental and the continuous theory only an approximation: spacetime could have a foam-like structure with very large fluctuations both in curvature and in topology.

Gauss, Carl Friedrich, *Disquisitiones Arithmeticae*, English ed., translated by Arthur A. Clarke and revised by William C. Waterhouse, Springer-Verlag, 1986; xx + 472 pp, $58.

Reproduction of the 1966 Yale University Press edition, with small changes to improve the rendering.

Hildebrandt, Stefan, and Anthony Tromba, *Mathematics and Optimal Form*, Freeman, 1985; xiii + 215 pp, $29.95.

Lavishly illustrated popular account of the calculus of variations and its manifestations in nature. "If the reader finally appreciates mathematics as an integral part of our culture, this book will have achieved its aim."

Shanks, Daniel, *Solved and Unsolved Problems in Number Theory*, 3rd ed., Chelsea, 1985; xiv + 304 pp, $18.95.

The original text is unchanged but the chapter entitled "Progress" is almost three times as long as in the 1978 edition. Most of the added topics concern class numbers and class groups. The book remains one of the finest introductions to number theory.

Stigler, Stephen M., *The History of Statistics: The Measurement of Uncertainty before 1900*, Harvard U Pr, 1986; xvi + 410 pp, $25.

Ten years in the making, this book has turned out to be a masterpiece, showing how statistics arose from the interplay of mathematical ideas and the needs of various applied sciences.

Peschel, Manfred, and Mende Werner, *The Predator-Prey Model: Do We Live in a Volterra World?*, Springer-Verlag, 1986; xi + 251 pp, $22.

This book is a unique amalgam of philosophy about growth of systems with mathematical analysis of the resulting differential equations. The focus is on laws of growth, particularly in biology and evolution; the authors' answer to the title question is "yes."

Slocum, Jerry, and Jack Botermans, *Puzzles Old & New: How to Make and Solve Them*, University of Washington Press, 1986; 160 pp, $19.95.

Profusely illustrated history of puzzles of all kinds. Too bad more of them weren't revived during the short Rubik-cube craze.

Chambers, Donald L. (ed.), *A Guide to Curriculum Planning in Mathematics*, Wisconsin Dept. of Public Instruction (125 South Webster St., Madison, WI 53707-7841), 1986; viii + 252 pp, $7 (P).

Although directed to Wisconsin school districts as a model and resource, this guide will interest people elsewhere as well. Starting from national reports of the last few years, it devises instructional objectives for K-12 mathematics. Its foremost recommendation is that schools offer a richer curriculum for the non-college-intending high-school students, who--being in the majority--are to be regarded as the "normal" students. It proposes a "Wisconsin Plan" of three years of mathematics for all, in two curricula: the "informal" one for non-college-bound and the "formal" one (similar to present college-prep math). Both treat algebra, geometry, and statistics in an integrated fashion.

Porter, Theodore M., *The Rise of Statistical Thinking*, Princeton U Pr, 1986; xii + 333 pp, $35.

Non-technical history of nineteenth-century statistics, emphasizing the development of statistics from twin roots in the natural and the social sciences. The author also dwells on the significant philosophical issues raised by the birth of statistical ideas.

Hexyflex and Octyflex. Kinetic sculpture and table jewelry. Flowerday Designs (21 West St., New York, NY 10006), $9.95 each.

Oldtimers will remember Martin Gardner's popularization of A. H. Stone's hexaflexagons, and younger mathematicians may have enjoyed the *M. C. Escher Kaleidocycles* of Doris Schattschneider and Wallace Walker. Here are two new toys, loops of stainless steel rods and springs, that can be turned and twisted through their centers. Moreover, they can be disassembled easily and reassembled in other patterns stimulated by the user's creativity.

Morris, Robert (ed.), *Studies in Mathematics Education*, vol. 4: *The Education of Secondary School Teachers of Mathematics*, UNESCO, 1985; xii + 175 pp (also available in French and Spanish).

The single biggest obstacle to progress in mathematics in most countries of the world is weakness in teacher education. This volume, the outgrowth of a meeting to examine throughout the world whether the teaching of mathematics corresponds to the needs of the majority of pupils and of their society, presents suggestions for both teacher education and support for teachers. Two case studies examine Zimbabwe and China.

Kolata, Gina, Factoring on microcomputers, *Science* (23 May 1986) 935.

Progress in the quadratic sieve method now allows a factoring problem to be broken down into independent subproblems, which can be solved concurrently on microcomputers. Even though a factoring problem may take 100 times as much microcomputer time as time on a Cray XMPO, the cost ratio runs at least as great in the opposite direction; so factoring is in fact migrating to micros.

Kingsland, Sharon E., *Modeling Nature: Episodes in the History of Population Ecology*, U Chicago Pr, 1985; ix + 267 pp, $27.50.

Stresses the dialectic between biological and mathematical approaches, between historical and ahistorical approaches, to the study of population ecology. Chapters center on leading figures, such as Lotka, Volterra, and Gauss. This book is a very important contribution to the humanizing of mathematical modeling.

Brook, Richard, et al. (eds.), *The Fascination of Statistics*, Dekker, 1986; xi + 436 pp, $24.75.

Collection of readable articles on applications of statistics, suitable as a supplement for an elementary statistics course.

Berloquin, Pierre, *The Garden of the Sphinx: 150 Challenging and Instructive Puzzles*, Scribners, 1985; xvii + 186 pp, $13.95.

Mathematical puzzles, requiring nothing beyond high-school mathematics, that have appeared in recent years in *Le Monde*.

Lemonick, Michael D., Physics/digest: A theory of everything; the entire universe could be explained in terms of superstrings, *Science Digest* 98:2 (February 1986) 20

Another theory of extra dimensions in the universe says there are nine space dimensions altogether, with six of them string-like thin (10^{-33}cm across). The math involved includes algebraic geometry and Lie algebras.

Osserman, Robert, *A Survey of Minimal Surfaces*, 2nd ed., Dover, 1986; 207 pp, $8.

Newly-updated survey, including developments through 1985. Although the blurb claims (correctly) that the work is accessible to readers without previous knowledge of differential geometry, a good knowledge of vector calculus and linear algebra is essential. Appendix 3, which relates developments since 1969, together with the corresponding references, takes up one-fourth of the book.

Wheeler, John Archibald, Hermann Weyl and the unity of knowledge, *American Scientist* 74:4 (July-August 1986) 366-375.

Splendid thought-provoking essay occasioned by the hundredth anniversary of Weyl's birth and by Weyl's own essays on remarkable issues connected with the puzzle of existence. Central to one's view of the "how come" of existence has to be quantum theory, "the summit of the exact natural science in our day." Why the quantum? Why time? What about the continuum? (though readers will find Wheeler's discussion of his misnamed "continuum of natural numbers" cryptic indeed).

Nelson, David R., Quasicrystals, *Scientific American* 255:2 (August 1986) 40-51, 120.

Newly-discovered metallic glasses exhibit a crystalline structure with five-fold symmetry. But the theory of crystals, based on the mathematical theory of tilings, has as a prime tenet Barlow's Law, which says that five-fold symmetry is impossible in a crystal (periodic tiling)! In fact, the new materials exhibit the features of the aperiodic Penrose tilings discovered a decade ago; despite being aperiodic, Penrose tilings have both short-range pentagonal symmetry and the long-range orientational order that is usually associated with conventional crystals. Once again, mathematics invented for the pure pleasure of it turns out to have unforeseen applications, proving once again the old saying: "All mathematics is applied mathematics, even if the applications haven't been discovered yet."

Davis, Morton, *The Art of Decision-Making*, Springer-Verlag, 1986; viii + 92 pp, $19.95.

"Common sense is an unreliable guide. To prove it, we pose a number of 'easy' questions with 'obvious' answers, obvious answers that turn out to be wrong." This delightful collection of paradoxes covers general business problems, game theory, voting problems and paradoxes, and logic and probability. The examples should wake up even the least-interested student.

'Buckyball' II: The game continues, *Science News* 128 (21 & 28 December 1985) 396.

Account of the newly-discovered molecule C_{60}--buckminsterfullerene, or "buckyball" for short--that has the shape of a semiregular solid "like a pattern on a soccer ball." Speculation: C_{120} may also be stable, in a symmetric pattern of squares, hexagons, and decagons.

Davis, Robert B., *Learning Mathematics: The Cognitive Science Approach to Mathematics Education*, Ablex, 1984; viii + 392 pp.

Why do students make the errors they do in mathematical work? Cognitive science, which models human learning and cognitive behavior on computer analogues, has been the most successful of educational theories in explaining why and how students make errors. This book provides a highly readable introduction to thinking about learning and teaching mathematics from this perspective; no teacher should be ignorant of the ideas presented here.

Kozlov, Alex, Mathematics/digest: But does it really work? After more than a year, Karmarkar's algorithm remains unverified, *Science Digest* 94:2 (February 1986) 23.

Debate over the details and value of Karmarkar's algorithm--not to mention AT&T's handling of his discovery--continues. Stay tuned.

Kaufmann, Arnold, and Gupta, Madan, *Introduction to Fuzzy Arithmetic: Theory and Applications*, Van Nostrand Reinhold, 1985; xvii + 351 pp, $44.95.

A fuzzy number is a particular kind of fuzzy subset and can be considered a kind of confidence interval. This volume develops arithmetical operations and functions on such entities; the enterprise is refreshing and elementary.

Martin, James, *System Design from Provably Correct Constructs*, Prentice-Hall, 1985; sviii + 392 pp, $40.

"At last we have the ability to create specifications without internal errors and inconsistencies and to automatically produce bug-free code from these ... without manual program coding." The author, the country's most prominent lecturer on software engineering, touts a particular mathematically-based system of program specification and code generation. "The jacket cover is a painting by Pieter Bruegel the elder (1525-1569) representing typical computer specifications and the difficulties of debugging them, prior to the methodology described in this book."

Franksen, Ole Immanuel, *Mr. Babbage's Secret: the Tale of a Cypher--and APL*, Prentice-Hall, 1984; 319 pp, $32.

Charles Babbage delved into cryptography, and this book chases down the letters and documents that display his involvement. The author exhibits how the programming language APL can solve handily certain kinds of ciphers. The book wanders off on many tangents; their appeal will depend on the reader's taste and interests.

McKelvey, Robert W. (ed.), *Environmental and Natural Resource Mathematics*, AMS, 1985; xii + 143 pp, $34.

Revised lecture notes from the AMS 1984 Summer Short Course. Topics include pest management, pollution control, depletion of resources, capitalizing renewable resources, control theory, and international trade; there is also the transcript of the panel discussion. Notable is G. Chichilnisky's (Columbia) article on trade, which takes up more than one-third of the book; it demonstrates a "transfer paradox," whereby an international loan can benefit the lender far beyond interest received and simultaneously further impoverish the recipient.

Crypton, Dr., Is democracy mathematically unsound?: Nobel prize-winning work seems to rule out a perfect democracy, *Science Digest* 93:11 (November 1985) 56-58, 92-93.

The usual voting paradoxes are described.

Rubinstein, Moshe F., *Tools for Thinking and Problem Solving*, Prentice-Hall, 1986; xv + 330 pp, $32.95.

Rubinstein is concerned mainly with decision-making, not mathematical problem-solving; but he brings in elementary mathematical tools (digraphs, matrices, probability tree diagrams, Bayes's theorem, the normal distribution), in a way that makes this book suitable for general readers.

Radetsky, Peter, The man who mastered motion, *Science 86* (May 1986) 52-60.

Thomas Kane (Stanford) has produced a simpler and more efficient method of solving problems in dynamics, by replacing some of the Lagrange-D'Alembert methodology with what he calls "generalized speeds." His methods have been applied particularly to articulated spacecraft and reconstructing falls from such heights as highway overpasses and hotel windows.

Peterson, Ivars, Computing art: can a computer be taught to take a painting's measure?, *Science News* 129 (1 March 1986) 138-140.

Russell and Joan Kirsch have written a "grammar" for the "Ocean Park" canvases of Richard Diebenkorn, and they have tested it by generating others like them. The Kirsches chose Diebenkorn because of the geometric nature of his work.

Samuelson, Paul A., The 1985 Nobel Prize in Economics, *Science* 231 (21 March 1986) 1399-1401.

Modigliani is not only the name of a famous modernist painter and sculptor--it is also the name of the winner of the 1985 Nobel Prize for Economics. Like most of his predecessor winners, Franco Modigliani (MIT) has used notable mathematics; in particular, he used Brouwer's fixed point theorem to resolve a matter from rational expectationism.

Orbach, R., Dynamics of fractal networks, *Science* 231 (21 February 1986) cover, 777, 814-819.

Percolation networks exhibit fractal geometry in the small, Euclidean geometry at larger scales. There are nice color pictures of the fractals, but the physics is heavy going.

Schattschneider, Doris, In black and white: how to create perfectly colored symmetric patterns, *Computers and Mathematics with Applications* 12B (1986) 673-695.

"The use of isometries to create and 'perfectly color' symmetric tilings and patterns is explained. The reader is not presumed to have knowledge of isometries, group theory, or computer science. A student, designer, teacher, or any other person interested in the interplay of geometry and art (particularly geometric symmetry), and the possibility of implementation using computer graphics, can learn from this paper."

White, Arthur T., *Graphs, Groups, and Surfaces*, revised and enlarged edition, North-Holland, 1984; xiii + 314 pp, $30.

Assuming only an introductory knowledge of group theory and point-set topology, this book investigates the mappings and interactions among graphs, groups, and surfaces. This second edition has six new chapters on voltage-graphs, non-orientable embeddings, block designs, hypergraph embeddings, map automorphism groups, and change ringing.

Parent, D.P., *Exercises in Number Theory*, Springer-Verlag, 1984; x + 541 pp.

Collection of problems and solutions, most from French university examinations. Certain important areas (e.g., diophantine equations) are not treated.

NEWS & LETTERS

27th INTERNATIONAL MATH OLYMPIAD SOLUTIONS

The solutions that follow have been especially prepared for publication in this MAGAZINE *by Loren C. Larson and Bruce Hanson, St. Olaf College.*

1. Let d be any positive integer not equal to 2, 5 or 13. Show that one can find distinct a,b in the set $\{2,5,13,d\}$ such that $ab-1$ is not a perfect square.

Sol. Assume the contrary--that

$$2d-1 = x^2 ,$$

$$5d-1 = y^2 ,$$

$$13d-1 = z^2 ,$$

for some integers x,y,z. From $2d-1 = x^2 \pmod 4$ we conclude that d must be odd. From $(z-y)(z+y) = z^2-y^2 = 8d$ we conclude that either $z+y$ or $z-y$ is divisible by 4. But y and z are both even since d is odd, and therefore both $z+y$ and $z-y$ are divisible by 4. Thus $8d$ is divisible by 16, so d is even, a contradiction.

2. A triangle $A_1A_2A_3$ and a point P_0 are given in the plane. We define $A_s = A_{s-3}$ for all $s \geq 4$. We construct a sequence of points $P_1, P_2, P_3, \ldots$ such that P_{k+1} is the image of P_k under rotation with center A_{k+1} through angle 120° clockwise (for $k = 0,1,2,\ldots$). Prove that if $P_{1986} = P_0$ then the triangle $A_1A_2A_3$ is equilateral.

Sol. View the problem in the complex plane. Let $\eta = e^{\pi i/3} = \frac{1 + i\sqrt{3}}{2}$. Then $P_k-A_k = \eta(A_k-P_{k-1})$ for $k \geq 1$, so

(1) $P_1 = (1+\eta)A_1 - \eta P_0$,

(2) $P_2 = (1+\eta)A_2 - \eta P_1$,

(3) $P_3 = (1+\eta)A_3 - \eta P_2$.

Multiply (1) by η^2 and (2) by $-\eta$, and add the three equations, using the fact that $\eta^3 = -1$, to find that

$$P_3 = P_0 + (1+\eta)(\eta^2 A_1 - \eta A_2 + A_3) .$$

Since $A_s = A_{s-3}$ for $s \geq 4$, we see that in general

$$P_{3k} = P_0 + k(1+\eta)(\eta^2 A_1 - \eta A_2 + A_3) .$$

In particular,

$$P_{1986} = P_0 + 662(1+\eta)(\eta^2 A_1 - \eta A_2 + A_3).$$

Since $P_{1986} = P_0$ and $1 + \eta \neq 0$, it follows that $\eta^2 A_1 - \eta A_2 + A_3 = 0$. Since $\eta^2 = \eta-1$, we find that A_1, A_2, A_3 satisfy

$$A_3 - A_1 = \eta(A_2 - A_1),$$

i.e. A_3 is the image of A_2 under a counterclockwise rotation of 60° about A_1. It follows that $A_1A_2A_3$ is equilateral.

3. To each vertex of a regular pentagon an integer is assigned in such a way that the sum of all the five numbers is positive. If three consecutive vertices are assigned the numbers x,y,z respectively and $y < 0$ then the following operation is allowed: the numbers x,y,z are replaced by $x+y$, $-y$, $z+y$ respectively. Such an operation is performed repeatedly as long as at least one of the five numbers is negative. Determine whether this procedure *necessarily* comes to an end after a finite number of steps.

Sol. 1. This solution, due to U.S. team member Joseph Keane, won the only special prize of the Olympiad.

The key to the solution is to find an integer-valued, non-negative function

$f(x_1,x_2,x_3,x_4,x_5)$ whose value decreases when the given operation is performed. One might conjecture f to be the sum of the absolute values of the five numbers. When the operation is performed, this particular function decreases by $|x| + |z| - |x+y| - |y+z|$. This expression, however, is *not* always positive, but it suggests a possible modification. The desired function might involve the absolute value of pairwise sums as well as the five numbers themselves. Upon testing the new function and continuing this line of attack, we discover in turn that the desired function should include absolute values of sums of triples and foursomes. At last, with a pentagon numbered x_1,x_2,x_3,x_4,x_5, define f by

$$f(x_1,x_2,x_3,x_4,x_5) = \sum_{i \neq j} |s(i,j)|$$

where $s(i,j) = x_i + x_{i+1} + \ldots + x_{j-1}$ for $i \neq j$ and where subscripts run modulo 5.

If y corresponds to x_4, we find that the operation reduces the value of f by $|x_5 + x_1 + x_2 + x_3| - |x_5 + x_1 + x_2 + x_3 + 2x_4| = |s - x_4| - |s + x_4|$, where $s = x_1 + x_2 + x_3 + x_4 + x_5$. Since $s > 0$ and $x_4 < 0$, we see that $|s - x_4| - |s + x_4| > 0$, so f has the property required to prove that the operation can be performed only a finite number of times.

Sol. 2. A recent article in FOCUS about this year's Olympiad described this problem and invited readers to try to discover Keane's elegant solution. Robert W. Floyd, Ramsey Haddad, and Donald E. Knuth, all of Stanford University, correctly surmised that an elegant solution could be based on finding an integer-valued, nonnegative function that decreases whenever the operation is performed. With this beginning they discovered the function used by Keane, and another one as well. Namely, using the notation of the preceding solution, define

$$f(x_1,x_2,x_3,x_4,x_5) = \sum_{i \neq j} s(i,j)^2 .$$

Then the operation reduces the value of f by $4sx_4$, where as before, $y = x_4$, $s = x_1 + x_2 + x_3 + x_4 + x_5$.

In either of these solutions, the parameter "5" can be replaced by any integer $n \geq 3$.

Sol. 3. The proposer.

Begin with the general quadratic form

$$f = \sum_{i,j=1}^{5} Q_{ij} x_i x_j$$

and on the basis of symmetry assume that

$$Q_{ij} = \begin{cases} a & \text{if } i = j \\ b & \text{if } i \text{ and } j \text{ are adjacent} \\ c & \text{otherwise} . \end{cases}$$

Then with $x_4 = y$, (x_1,x_2,x_3,x_4,x_5) is replaced by $(x_1,x_2,x_3+x_4,-x_4,x_4+x_5)$ and f is reduced by $x_4[(c-b)(x_1 + x_2) + (2b - 2a - c)(x_3 + x_4 + x_5)]$. Since $x_4 < 0$ and $s = x_1 + x_2 + x_3 + x_4 + x_5 > 0$, we can insure that f has the desired property by choosing integers a,b,c so that $c-b = 2a-2b-c < 0$. For example, the choice $a=1$, $b=0$, $c=-1$ yields

$$f = \tfrac{1}{2}[(x_1 - x_3)^2 + (x_2 - x_4)^2 + (x_3 - x_5)^2 + (x_4 - x_1)^2 + (x_5 - x_1)^2]$$

which decreases by $-sx_4$ at each step.

Comment. (Don Knuth, Stanford University).

"None of these methods prove termination when the x's aren't integers. But there's an elegant way to do this, based on an idea due to Bernard Chazelle of Princeton. Namely, consider the infinite multiset S of all sums $s(i,j)$ [see Sol. 2] where $1 \leq i \leq 5$ and $j > i$. This set remains almost untouched by the stated operation; many of the $s(i,j)$ switch places with others, while all but one of the others stay unchanged. The only change in S is that, for example, $s(4,5) = x_4$ changes to $-s(4,5)$, when y corresponds to x_4. Thus, exactly one negative element of S changes to positive at each step! There are finitely many negative elements, since $s > 0$. The number of iterations until termination is exactly the number of negative elements of the initial set S, regardless of the order in which the operations are carried out."

4. Let A,B be adjacent vertices of a regular n-gon ($n \geq 5$) in the plane having center at 0. A triangle XYZ, which is congruent to and initially coincides with $0AB$, moves in the plane in such a way that Y and Z each trace out the whole boundary of the polygon, X remaining inside the polygon. Find the locus of X.

Sol. Since $n \geq 5$, Y and Z are always on adjacent edges of the polygon. Assume Y and Z lie on edges AB and BC respectively (see fig.). Since angles YBZ and YXZ are supplementary, $YBZX$ is a cyclic quadrilateral. It follows that angles YBX and YZX are equal and therefore that X lies on the line $B0$. If we let the sides of the polygon have unit length, then $B0 = \frac{1}{2}\csc\left(\frac{\pi}{n}\right)$ and the diameter of the circle circumscribing triangle XYZ is equal to $\frac{YZ}{\sin YXZ} = \csc\left(\frac{2\pi}{n}\right)$. It follows that as Z moves along edge BC from B to C, X traces a line segment of length $\csc\left(\frac{2\pi}{n}\right) - \frac{1}{2}\csc\left(\frac{\pi}{n}\right)$ emanating from 0 and pointing away from B. Thus, the complete locus will be an "asterisk" consisting of n such segments, one for each vertex of the polygon.

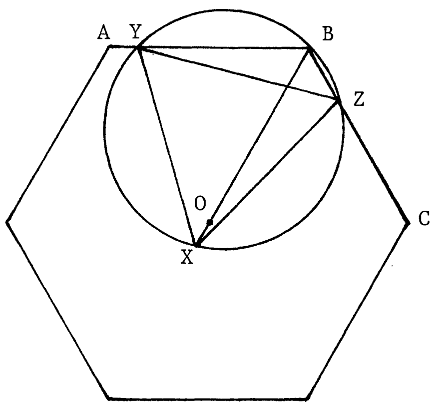


Illustration for $n = 6$

5. Find all functions f, defined on the non-negative real numbers and taking non-negative real values, such that:

(i) $f[xf(y)]f(y) = f(x+y)$ for all $x,y \geq 0$.

(ii) $f(2) = 0$.

(iii) $f(x) \neq 0$ for $0 \leq x < 2$.

Sol. Letting $y = 2$ in (i) and using (ii), we see that $f(z) = 0$ if $x \geq 2$. Now suppose that $0 \leq z < 2$. Letting $y = z$, $x = 2-z$ in (i) we get

$$f((2-z)f(z))f(z) = f(2) = 0 .$$

Since $f(z) \neq 0$, it follows that

$$(2-z)f(z) \geq 2$$

so

$$f(z) \geq \frac{2}{2-z} .$$

Next setting $x = \frac{2}{f(z)}$, $y = z$ in (i) we have

$$0 = f(2)f(z) = f\left(\frac{2}{f(z)} + z\right) .$$

Again, since $f(z) \neq 0$, it follows that

$$\frac{2}{f(z)} + z \geq 2$$

so

$$f(z) \geq \frac{2}{2-z} .$$

Thus the only possible solution is

$$f(z) = \begin{cases} \frac{2}{2-z} & \text{if } 0 \leq z < 2 \\ 0 & \text{if } z \geq 2 . \end{cases}$$

Obviously f satisfies conditions (ii) and (iii), and condition (i) certainly holds for $y \geq 2$. But if $0 \leq y < 2$, then $f(y) = \frac{2}{2-y}$ so $x + y \geq 2$ if and only if $xf(y) \geq 2$. Using this fact, it is straightforward to show that (i) is satisfied in this case also and hence f satisfies all three conditions.

6. One is given a finite set of points in the plane, each point having integer coordinates. Is it always possible to color some of the points in the set red and the remaining points white in such a way that for any straight line L parallel to either one of the coordinate axes the difference

(in absolute value) between the numbers of white points and red points on L is not greater than 1? Justify your answer.

Sol. Such a coloring is always possible. Let S denote the given set of points. For each horizontal or vertical line L which meets S, define a pairing of the points of S along L by joining paired points by an edge. One way to do this is to order the points by the free coordinate $P_1 < P_2 < P_3 < \ldots$ and then to join P_1 to P_2, P_3 to P_4, etc. If the number of points on L is k, then $\lfloor k/2 \rfloor$ edges are thus introduced and the number of unpaired points is zero or one depending on whether k is even or odd. We thus define a graph G whose vertices are the points of S and whose edges are the matching edges so introduced. Note that each component of G is either an isolated vertex, a path, or an even cycle. This follows from the fact that each vertex is incident with at most one horizontal edge and with at most one vertical edge. In particular, each cycle is even since it consists of alternating horizontal and vertical edges.

Color the vertices of G by coloring the vertices of each path or cycle alternately red-white and coloring any isolated vertices red or white arbitrarily. Since each cycle of G is even, any two vertices which are joined by an edge receive opposite colors. This coloring fulfills our requirement. To see this, consider an arbitrary horizontal or vertical line L which meets S. Points of S which are paired receive opposite colors. There is at most one unpaired point on L, so there is a discrepancy of at most one (in absolute value) between red and white points.

ESCHER EXHIBIT

There will be a show of works by M.C. Escher March 10 through April 11 at the Payne Gallery of Moravian College, Bethlehem, PA. The show is co-curated by Doris Schattschneider and is entitled "A Mathematician Views Escher."

HISTORY OF ENGLISH MATHEMATICS

This summer, the course "On the Shoulders of Giants: A History of English Mathematics" will be offered at Oxford University. Lectures, aimed at a general audience, will discuss the lives and contributions of some giants of English mathematics, taking a broad view of the scientific and cultural contexts of their achievements. To supplement the lectures, visits are planned to some places where these men and women lived and worked. Three undergraduate or graduate credits are available for completing this three week course, August 4 - 24. Mathematics instructors and other professionals taking the course for credit may find that the costs of the program qualify for tax purposes as a professional expense.

This course is offered as part of the 1987 Summer Program at Oxford University sponsored by the Pennsylvania Consortium for International Education. The program is designed to provide an opportunity to study at England's oldest university, founded in 1167. Classes as well as room and board will be scheduled at Mansfield, one of Oxford's colleges. Room, board, and one-day trips in conjunction with the course are all covered by the program fee of $750. Tuition ($204 for 3 undergraduate credits and $267 for 3 graduate credits), airfare, and weekend meals are not included. For further information, write to the instructor: Professor Paul Wolfson, Department of Mathematical Sciences, West Chester University, West Chester, PA 19383.

SIXTH INTERNATIONAL CONGRESS ON MATHEMATICAL EDUCATION

The Sixth International Congress on Mathematical Education will be held July 23 - August 3, 1988 in Budapest, Hungary.

The United States Commission on Mathematical Instruction seeks to encourage American participation in ICME 6. The presentations to be made at ICME 6 are organized into Action Groups and Theme Groups as follows:

Action Groups: Early Childhood Years (ages 4-8), Elementary School (ages 7-12), Junior Secondary School (ages 11-16), Senior Secondary School (ages 15-19), Tertiary/Post-Secondary/Academic Institutions (age 18+), Pre-Service Teacher Education; Adult, Technical and Vocational Education.

Theme Groups: The Profession of Teaching, Computers and the Teaching of Mathematics, Problem Solving, Modelling and Applications, Evaluation and Assessment, the Practice of Teaching and Research in Didactics, Mathematics and other subjects, Curriculum towards the Year 2000.

For more information about submitting abstracts for consideration, send your name, mailing address, affiliation, and name of the group to which you wish to make a contribution to:

Professor Eileen L. Poiani
Chairperson, USCMI
Department of Mathematics
Saint Peter's College
Jersey City, NJ 07306

* * * * * * * * * * * * * * * * *

LETTERS TO THE EDITOR

Dear Editor:

It has been brought to our attention that the theorem on page 224 of our recent publication, "Fields for Which the Principal Axis Theorem is Valid" ([this *MAGAZINE*], 59 (1986) pp. 222-225), is a known result--at least if the characteristic of the field is not 2. It was proved by Professor William C. Waterhouse in his article, "Self-adjoint Operators and Formally Real Fields" (*Duke Mathematical Journal*, 43 (1976) p. 241: Proposition 7).

A. Charnow and E. Charnow
California State University
Hayward, CA 94542

Dear Editor:

In the October issue of [this *MAGAZINE*], the article "On the Power Sums of the Roots of a Polynomial" gives a reference to B.L. van der Waerden. In van der Waerden's text, Newton's formulae for the power sums of the roots of a polynomial are given as an exercise for the reader to prove. Chrystal's *Textbook of Algebra*, page 436, gives a proof of Newton's formulae using the barest minimum of symmetric function theory. The formulae are:

Let s_r be the sum of the rth (integral) powers of the roots of the algebraic equation:

$$x^n + p_1x^{n-1} + p_2x^{n-2} + \dots + p_n = 0.$$

Then

$$s_r + p_1s_{r-1} + p_2s_{r-2} + \dots + p_{r-1}s_1 + rp_r = 0, \quad 0 < r \le n .$$

$$s_r + p_1s_{r-1} + p_2s_{r-2} + \dots + p_ns_{r-n} = 0, \quad r > n .$$

I do not think these old formulae are well known. They are very easy to use.

John P. Hoyt
1115 Marietta Ave., 32
Lancaster, PA 17603

* * * * * * * * * * * * * * * * * * * *

CORRECTION

In the News and Letters Section of the October 1986 issue of this *MAGAZINE*, the name of C. Rousseau was inadvertently omitted from the list of those who made up the 15th USAMO.

Action Groups: Early Childhood Years (ages 4-8), Elementary School (ages 7-12), Junior Secondary School (ages 11-16), Senior Secondary School (ages 15-19), Tertiary/Post-Secondary/ Academic Institutions (age 18+), Pre-Service Teacher Education; Adult, Technical and Vocational Education.

Theme Groups: The Profession of Teaching, Computers and the Teaching of Mathematics, Problem Solving, Modelling and Applications, Evaluation and Assessment, the Practice of Teaching and Research in Didactics, Mathematics and other subjects, Curriculum towards the Year 2000.

For more information about submitting abstracts for consideration, send your name, mailing address, affiliation, and name of the group to which you wish to make a contribution to:

Professor Eileen L. Poiani
Chairperson, USCMI
Department of Mathematics
Saint Peter's College
Jersey City, NJ 07306

* * * * * * * * * * * * * * * * *

LETTERS TO THE EDITOR

Dear Editor:

It has been brought to our attention that the theorem on page 224 of our recent publication, "Fields for Which the Principal Axis Theorem is Valid" ([this *MAGAZINE*], 59 (1986) pp. 222-225), is a known result--at least if the characteristic of the field is not 2. It was proved by Professor William C. Waterhouse in his article, "Self-adjoint Operators and Formally Real Fields" (*Duke Mathematical Journal*, 43 (1976) p. 241: Proposition 7).

A. Charnow and E. Charnow
California State University
Hayward, CA 94542

Dear Editor:

In the October issue of [this *MAGAZINE*], the article "On the Power Sums of the Roots of a Polynomial" gives a reference to B.L. van der Waerden. In van der Waerden's text, Newton's formulae for the power sums of the roots of a polynomial are given as an exercise for the reader to prove. Chrystal's *Textbook of Algebra*, page 436, gives a proof of Newton's formulae using the barest minimum of symmetric function theory. The formulae are:

Let s_r be the sum of the rth (integral) powers of the roots of the algebraic equation:

$$x^n + p_1x^{n-1} + p_2x^{n-2} + \dots + p_n = 0.$$

Then

$$s_r + p_1s_{r-1} + p_2s_{r-2} + \dots + p_{r-1}s_1 + rp_r = 0, \quad 0 < r \leq n .$$

$$s_r + p_1s_{r-1} + p_2s_{r-2} + \dots + p_ns_{r-n} = 0, \quad r > n .$$

I do not think these old formulae are well known. They are very easy to use.

John P. Hoyt
1115 Marietta Ave., 32
Lancaster, PA 17603

* *

CORRECTION

In the News and Letters Section of the October 1986 issue of this *MAGAZINE*, the name of C. Rousseau was inadvertently omitted from the list of those who made up the 15th USAMO.

JOIN MAA TODAY!

Send this application to: MAA Membership Department
1529 18th Street, N.W.
Washington, D.C. 20036

Name ____________________

Mailing Address ____________________

____________________ Zip ____________

Employer/School ____________________

Position (Rank) ____________________

Employer's City/State ____________________

Highest Earned Degree ____________ Year Degree Earned ____________

Institution Awarding Degree ____________________

Month/Year of Birth ____________________

Have you been a member of the MAA before? ☐ **Yes** ☐ **No**

Payment Enclosed $____________ *U.S. Funds Only* (see below)

Your membership may begin on July 1 or January 1. You may elect to receive any of the listed combinations of the three MAA journals:

THE AMERICAN MATHEMATICAL MONTHLY **(M)**
MATHEMATICS MAGAZINE **(G)**
THE COLLEGE MATHEMATICS JOURNAL **(J)**

Please circle the appropriate box corresponding to your initial membership period and the selection of journals you wish to receive. (Rates guaranteed for the indicated periods only.) Subscription prices are included with dues (see **).

	M	**G**	**J**	**M + G**	**M + J**	**G + J**	**M + G + J**
Student Membership							
1 year (Jan.-Dec. 1986)	$24	$18	$20	$30	$32	$26	$38
1½ year (July 1986-Dec. 1987)	$36	$27	$30	$45	$48	$39	$57
Regular Membership							
1 year (Jan.-Dec. 1986)	$40	$28	$31	$52	$55	$43	$67
1½ year (July 1986-Dec. 1987)	$60	$42	$46.50	$78	$82.50	$64.50	$100.50

*Student Membership available to high school and undergraduate students and to students regularly enrolled in graduate study at least half time. Enclose letter from school or department official confirming status. Student rates apply to unemployed persons under certain circumstances.

**Annual dues include subscription prices as follows: Reg. Memb. $24 (M), $15 (J), $12 (G), $1 (Focus). Std. Memb. $12 (M), $8 (J), $6 (G), $1 (Focus).

METHOD OF PAYMENT ☐ Check enclosed ☐ VISA ☐ MasterCard ☐ Bill me

Card Number |_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_| Expiration Date |_|_|_|_| *(month/year)*

Interbank Number |_|_|_|_| *(MasterCard only—located above name on card)*

SIGNATURE ____________________

Publications by the

Mathematical Association of America

1529 Eighteenth St., NW
Washington, DC 20036
202 387-5200

New Mathematical Library #30 . . .

The Role of Mathematics in Science
by Leon Bowden and M.M. Schiffer
220 pp. Paperbound
List: $16.00 MAA Member: $12.50

This book is based on a series of lectures given over a period of years to high school mathematics teachers. The lectures focused on topics in mechanics, questions of population growth, probability and other uses of exponential functions, optics and application of matrices to relativity theory.

The aim of the book is to illustrate the power and elegance of mathematical reasoning in science with some examples ranging from the work of Archimedes to that of Einstein. The book starts with problems of the lever, the mirror and the growth of populations and ends up with problems of space travel and atomic energy.

This excellent book will appeal to the advanced high school student, to the undergraduate and to their teachers. A sampling of the Table of Contents should inspire you to order your copy now.

Table of Contents