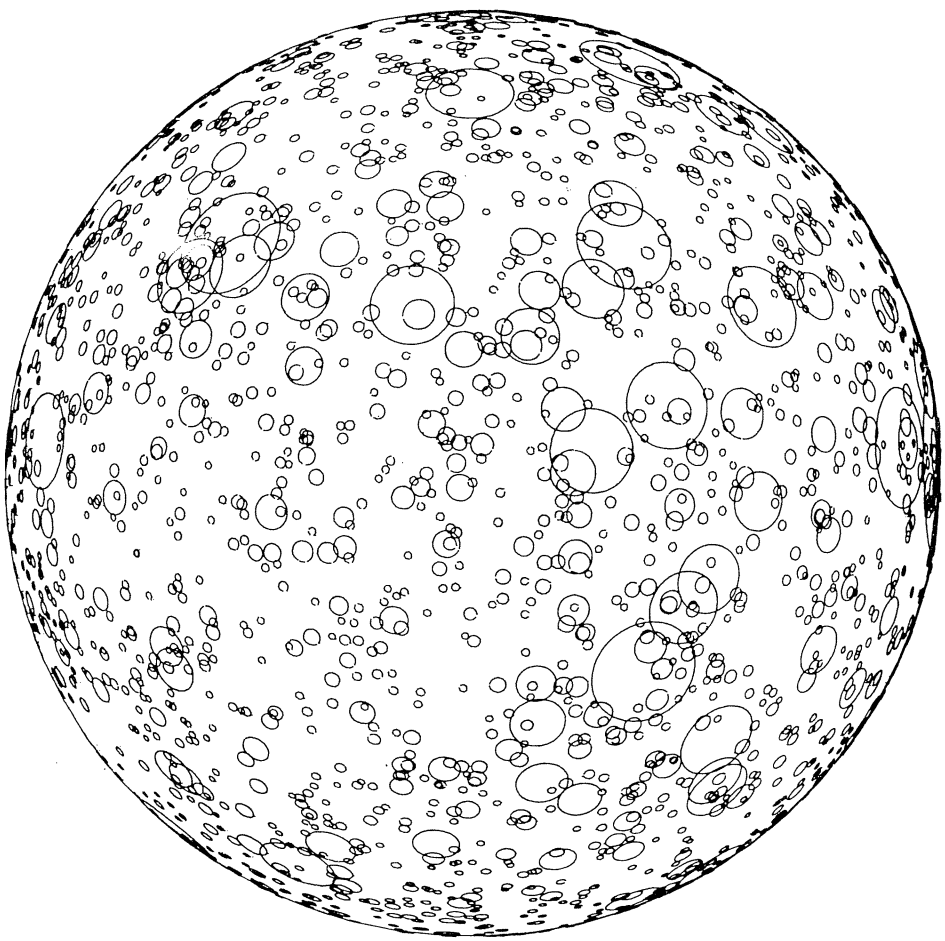
MATHEMATICS MAGAZINE



Vol. 57 No. 3
May 1984

ORDINAL NOTATION • BRIGHTNESS OF VENUS
AUDITING BANKS • CAVALIERI INTEGRATION

MATHEMATICS MAGAZINE

THE MATHEMATICAL ASSOCIATION OF AMERICA

Volume 57, Number 3, May 1984
(ISSN— 0025- 570X)

COVER: Moon, by Robert Dixon. See illustrations, p. 130, and p. 174.

ARTICLES

NOTES

PROBLEMS

REVIEWS

NEWS AND LETTERS

EDITORIAL POLICY

Mathematics Magazine is a journal which aims to provide inviting, informal mathematical exposition. Manuscripts for the *Magazine* should be written in a clear and lively expository style and stocked with appropriate examples and graphics. Our advice to authors is: say something new in an appealing way or say something old in a refreshing way. The *Magazine* is not a research journal and so the style, quality, and level of articles should realistically permit their use to supplement undergraduate courses. The editor invites manuscripts that provide insight into the history and application of mathematics, that point out interrelationships between several branches of mathematics and that illustrate the fun of doing mathematics.

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: Doris Schattschneider, Editor, Mathematics Magazine, Moravian College, Bethlehem, PA 18018. 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

Frederick Gass ("Constructive Ordinal Notation Systems") attended Phillips Academy, received a bachelor's degree from DePauw University with majors in mathematics and Romance languages, and received a Ph.D. in mathematics from Dartmouth College in 1968. He has been on the faculty of Miami University since 1968 except for one year spent as Visiting Scholar at Talladega College. Writing this article has satisfied a long-standing desire to popularize concepts that he has enjoyed privately since graduate school.

ILLUSTRATIONS

The cover, *Moon,* by **Robert Dixon** was computer-generated, and shows an isotropic distribution of circles on a spherical surface, with radii having a hyperbolic distribution. See Mandelbrot, *Fractal Geometry of Nature,* Freeman, 1982. More details appear on p. 174.

The cowering bank manager on p. 144 was illustrated by **Caspert Duk.**

All other illustrations were provided by the authors.

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, five times a year: January, March, May, September, and November.

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. 20036. 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 Membership/Subscriptions Department, Mathematical Association of America, 1529 Eighteenth Street, N.W., Washington, D.C. 20036.

PRINTED IN THE UNITED STATES OF AMERICA

Constructive Ordinal Notation Systems

G: Adam, did you find a good system for naming ordinals?
A: Ordinals? I thought you said "animals."

Frederick Gass
Miami University
Oxford, OH 45056

Cantor's ideas for the development of transfinite numbers can be traced back through a sequence of memoirs to his early research on trigonometric series. In this early research, Cantor included a study of the irrational numbers, using rational Cauchy sequences to define them, and it is in this setting that he seems to have drawn the inspiration for sequences of counting numbers that go beyond the integers.

Recall that if S is an increasing Cauchy sequence of rationals, then S belongs to some equivalence class of Cauchy sequences that by definition constitutes a certain number x, either rational or irrational. Viewing the situation geometrically, we identify x with the point on the real line that is the limit point of the sequence S, and we use this image to guide most of our thinking about x. In fact, prior to the time of Weierstrass it had been presumed that this geometrical point of view was a sufficiently rigorous approach to the theory of irrational numbers.

By analogy with this view of the irrationals, Cantor proposed in 1882 (Introduction to [3], p. 54) that the sequence, $0,1,2\ldots$ be considered to have a least upper bound or limit, now denoted by ω, and that ω be followed in order by numbers $\omega+1$, $\omega+2$, $\omega+3$, and so on. The process of passing from a number x to its immediate successor $x+1$ was Cantor's "first principle of generation", and the process of passing from a denumerable increasing sequence of numbers to its limit was his "second principle of generation." If you begin with 0 and repeatedly apply the first principle, you obtain the natural numbers (the nonnegative integers), which Cantor called the "first number class." By applying the second principle once, you obtain ω; then repeated application of the first principle yields $\omega+1, \omega+2, \omega+3,\ldots$, and then another application of the second principle yields $\omega+\omega$, also called $\omega\cdot 2$. In this way the generating process continues on and on. Incidentally, if you are curious about the use of $\omega\cdot 2$ rather than $2\cdot\omega$, then you might enjoy reading about the noncommutative arithmetic of these transfinite numbers. (Section 21 of [9] is a good reference for this topic. You will learn that $2\cdot\omega$ is equal to ω, as is $1+\omega$.)

The set of all numbers generated from 0 through use of both principles is Cantor's "second (cumulative) number class," and it evidently cannot be denumerable, for otherwise it could be extended by application of the second principle. In order to obtain more numbers, Cantor introduced a third principle that would lead from the second number class to its limit, Ω. From there, one obtains $\Omega+1, \Omega+2,\ldots,\Omega+\omega,\ldots,\Omega+\Omega,\ldots$ through use of all three principles. And of course one need never stop for want of a new principle.

The objects introduced in this way are called **ordinal numbers** because they serve to describe the sequential order of elements in any well-ordering. Each nonzero ordinal is classified as a **successor** if it is generated by Cantor's first principle, and otherwise it is a **limit** ordinal. For ordinals α and β, $\alpha<\beta$ if and only if α precedes β in the sequence of ordinals. This relation has the well-ordering property in that any non-empty set of ordinals contains a least element.

In order to place his ordinal number theory on a mathematically sound basis, Cantor used an equivalence class approach similar to the familiar one for the reals. Thus in [**2**] he formally introduced ordinals as isomorphism classes of well-ordered sets. For a good description of this approach, see [**10**]. A quite different approach due to von Neumann, described in [**7**] and [**9**], is preferred by most modern set-theorists. An interesting brief account of Cantor's life and work is contained in [**4**].

The ordinal numbers play a major role in modern set theory, and they are used occasionally in various branches of mathematics as a vehicle for extended inductive proofs. Applications of ordinals occur in topology, the area in which undergraduates are most likely to meet them. For example, Cantor's second number class with its order topology is a countably compact Hausdorff space that is not compact. The second number class is especially useful in constructing examples because every countable subset of the class has its least upper bound in the class ([**15**], pp. 10, 11, and 117). Another topological use of the ordinals is the formation of a hierarchy for Borel sets: at the 0th level of the hierarchy are the basic open sets. Given the αth level, we define the $\alpha+1$st level to contain all sets that are either a countable union of α-level sets or a countable intersection of α-level sets or the difference of two α-level sets. Finally, if λ is a limit ordinal, then the λth level contains all sets that belong to any previous level, so λ is a sort of gathering-together level. There are other ways of organizing the Borel sets into a hierarchy, but this way is perhaps the simplest to describe. As an exercise you can show that only ordinals of the second number class are needed here, because no new sets are obtained beyond the Ωth level ([**7**], pp. 123, 124).

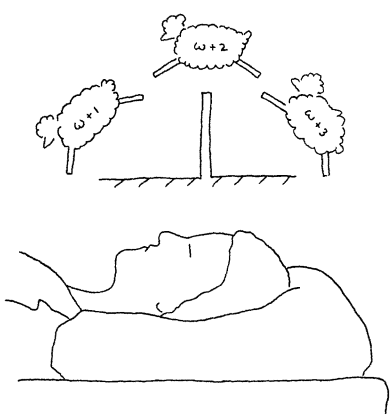
Ordinals are used in a similar fashion to index more extensive hierarchies of sets and functions in modern set theory. In the hierarchy of constructible sets ([**7**], Chapter V), each ordinal belongs to and represents a certain stage at which sets of a particular complexity are formed. Consequently the ordinals are often viewed as the backbone of the hierarchy, with each ordinal a vertebra.

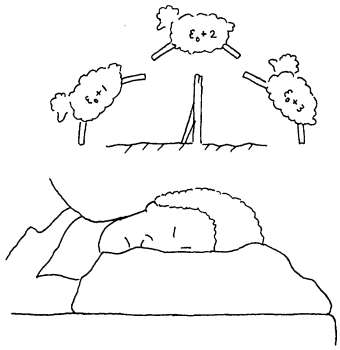
Polynomials in ω

Let us imagine the ordinals being generated by Cantor's first two principles and list several of them at strategically-chosen points so as to keep track of the growth process:

$$0,1,2,\ldots,\omega,\omega+1,\omega+2,\ldots,\omega\cdot 2,\omega\cdot 2+1,\omega\cdot 2+2,\ldots,\omega\cdot 3,$$
$$\omega\cdot 3+1,\omega\cdot 3+2,\ldots,\omega\cdot n,\omega\cdot n+1,\omega\cdot n+2,\ldots,\omega^2,\ldots,$$
$$\omega^3,\ldots,\omega^n,\ldots.$$

In the sequence indicated here, ordinals written in the form $\omega\cdot n$ are followed by a sequence $\omega\cdot n+1,\omega\cdot n+2,\omega\cdot n+3$, and so on, with $\omega\cdot n+m$ understood to mean $(\omega\cdot n)+m$. The limit of this sequence is $\omega\cdot(n+1)$, from whence another sequence leads to $\omega\cdot(n+2)$. The ordinals $\omega,\omega\cdot 2,\omega\cdot 3,\ldots$ form a denumerable increasing sequence whose limit we call $\omega\cdot\omega$ or ω^2.





Repeating the process, beginning with ω^2 rather than with 0, leads up to $\omega^2 \cdot 2$, and similarly we obtain $\omega^2 \cdot 3$, $\omega^2 \cdot 4, \ldots$, a sequence whose limit is called $\omega^2 \cdot \omega$ or ω^3. If you experiment by writing out names for some of the other ordinals, the following ideas will probably occur to you.

First, the notation assigned to the typical ordinal is a kind of polynomial in ω with positive integer coefficients, such as $\omega^2 \cdot 5 + \omega + 2$. (The ordinal mentioned here is quite far down our list. To reach it, go out to the limit ordinal $\omega^2 \cdot 5$ and start counting off numbers from there as if from 0: $1, 2, 3, \ldots, \omega, \omega + 1, \omega + 2$. The last one counted will be $\omega^2 \cdot 5 + \omega + 2$.) Second, this process of generating ordinals can go on indefinitely, but the system of polynomials in ω eventually gives out. For example, the increasing sequence $\omega, \omega^2, \omega^3, \ldots$ has a limit ordinarily called ω^ω, and one can go on to larger ordinals with progressively more complicated exponential notations, such as

$$\omega^{\omega^\omega}, \quad \omega^{\omega^{\omega^\omega}}, \ldots. \tag{*}$$

What would you call the limit of the sequence indicated in (*)? The usual name is ε_0.

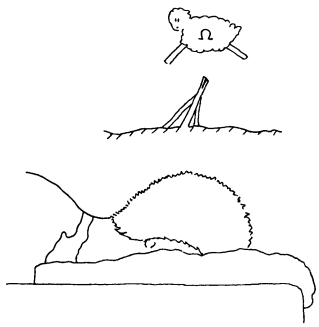
We call any expression of the form $\omega^{m_1} \cdot n_1 + \cdots + \omega^{m_k} \cdot n_k + n_{k+1}$ a (finite) **polynomial in ω**, provided that the coefficients n_i and exponents m_i are natural numbers, and the exponents are arranged in decreasing order. Given a polynomial in ω that denotes an ordinal β, you can quickly determine certain information about β that I shall call "the essential information":

1. You can tell from the polynomial whether β is 0, a successor, or a limit ordinal.
2. If β is a successor, then you can give a polynomial for its immediate predecessor, $\beta - 1$.
3. If β is a limit ordinal, then you can tell how to write polynomials that name a denumerable increasing sequence of ordinals with limit β (called a **fundamental sequence for β**).

To illustrate item 3, suppose that β is $\omega^3 \cdot 5 + \omega^2 \cdot 4$. Then $\omega^3 \cdot 5 + \omega^2 \cdot 3 + \omega \cdot n$, for $n = 1, 2, 3, \ldots$, gives a fundamental sequence for β. Although we can determine other information from the polynomial for β (such as the polynomial notation for β's immediate successor $\beta + 1$), we confine our attention to facts that show the genealogy of β with respect to Cantor's principles.

Since there are only countably many polynomials in ω, this notation system provides names for only a countable subset of Cantor's second number class. We wish to extend the system to a more comprehensive one that still conveys the essential information about each ordinal. The least ordinal for which there is no polynomial in ω is the ordinal ω^ω. One obvious option for extending the polynomial system is to continue with notations like ω^ω that allow the use of exponents greater than or equal to ω. This option works fine up to the number ε_0, as I suggested earlier. In order to extend the notation system beyond that point, one can simply admit ε_0 as a new symbol and allow it to be used in polynomial expressions such as $\varepsilon_0, \varepsilon_0 + 1, \varepsilon_0 + 2, \ldots, \varepsilon_0 + \omega, \ldots, \varepsilon_0 \cdot 2$, and $\varepsilon_0^2 + \omega^\omega \cdot 5 + \omega^3 \cdot 4 + \omega \cdot 14 + 85$.

With more explicit information about the polynomial expressions in this extended system, you could check for yourself that all the essential information is conveyed. But you can readily see that even this system is subject to extension. The major question formalized and answered in the remainder of this paper concerns the *existence of a maximal ordinal notation system*—one that



conveys the essential information and provides notations for the largest possible set of ordinals.

Maximal systems do exist, although at this point your intuition might strongly suggest otherwise. Indeed if λ is the least ordinal not named by some particular maximal system, could not the system be extended by adjoining a special symbol to denote λ? After we achieve the main results of this paper, we shall resolve this apparent problem.

Algorithms

Reviewing the essential information that is to be contained in a notation system, we notice the phrases "you can tell" and "you can give." The evident intent of these phrases is not that one can achieve something by means of luck or ingenuity, but rather that anyone who can follow directions can do it. In other words, there are routine procedures—algorithms, to use the standard modern term—for accomplishing the stated tasks.

To be a bit more specific about the nature of these algorithms, I ask you to pick some general-purpose programming language like BASIC or Pascal—one that you will be able to use or just imagine using for the rest of this article—and assume that all the algorithms are expressed in that language. It may seem rather vague to leave this choice open, but the specifics of the language that you pick will not really matter. Furthermore, it has been established that all standard programming languages have the same theoretical capability when it comes to expressing mathematical procedures ([**13**], p. 114, presents a technical version of this fact. See also Chapters 6–8 of [**1**]).

There is an important question that should be considered, though: whether every algorithm can indeed be expressed in terms of a computer program. (We are speaking of algorithms for manipulating symbols, of course; not algorithms for physical tasks such as tying shoelaces.) An early version of this question was central to the pioneering work of logicians in the 1930's who invented systems of computation equivalent to modern programming languages. Turing machines and recursive functions are perhaps the most famous of those systems. Since that time, every proposed procedure that is evidently algorithmic has been shown to be Turing programmable, so that now virtually all logicians and computer scientists accept the thesis that algorithmic procedures are precisely the programmable ones. The original version of this thesis is credited to Church or to Church and Turing jointly. See Chapter 1 of [**14**] for more discussion of this topic. The branch of mathematics that has grown out of these considerations is called "recursive function theory" or "the theory of effective computability," which I abbreviate as "computability theory."

If F is a computer program and x is an input, then $F(x)$ denotes the output, if any, when F is run with input x. If there is no output, then $F(x)$ is undefined. If F and G are programs, then $F(x) = G(x)$ means that either $F(x)$ and $G(x)$ are both defined and equal, or they are both undefined.

It is traditional and still fairly common practice in computability theory to restrict oneself to natural numbers as the inputs and outputs of algorithms. Natural numbers are also used to identify whole algorithms, in much the way that serial numbers identify appliances; sets of natural

numbers are used to provide ordinal notations. Besides being traditional, this natural number approach ties many ideas together neatly, and so it is the approach we shall follow.

In the first place, then, we shall assume that all programs mentioned in this article are intended for use with natural number inputs and outputs. Any nonnumerical outputs will be considered the same as no output at all. Also, "natural number" will be abbreviated as simply "number."

The decision to use only numbers as ordinal notations may seem too drastic or oversimplifying until you see how systems like our polynomials in ω can be transformed into strictly numerical systems without any loss of information. One way to transform the polynomials is to consider the symbols $0,1,2,\ldots,8,9,+,\cdot$, and ω to be the digits of a base-13 numeration systems. Then, for instance, the polynomial $\omega^2 \cdot 5 + \omega \cdot 15 + 2$ can be rewritten slightly as $\omega 2 \cdot 5 + \omega \cdot 15 + 2$ and identified with the number that it represents as a base-13 numeral. In this example, the number would be

$$12 \times 13^{10} + 2 \times 13^9 + 11 \times 13^8 + 5 \times 13^7 + 10 \times 13^6$$
$$+ 12 \times 13^5 + 11 \times 13^4 + 13^3 + 5 \times 13^2 + 10 \times 13 + 2.$$

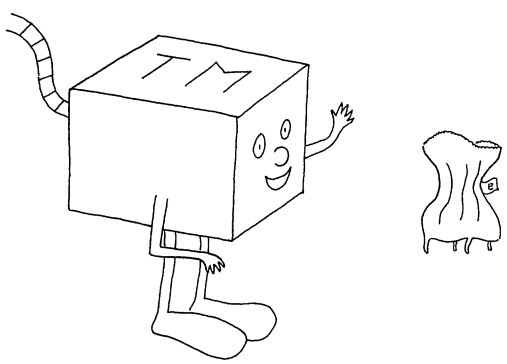
In this way each polynomial is associated with a unique number, and from that number the original polynomial can be recovered via straightforward arithmetic.

In a similar fashion one can associate a unique number with each line of a computer program written in your language. (Let $0,1,2,\ldots,8,9,+,\ldots$ be the symbols of your language and consider them to be the digits of a numeration system. Then each line can be interpreted as a numeral in the system.) If $L_1, L_2, \ldots, L_k$ are the lines of a program and $n_1, n_2, \ldots, n_k$ are the corresponding numbers, then we combine the n_i's into a single number e that represents the entire program. One way to combine them is to let

$$e = 2^{n_1} \cdot 3^{n_2} \cdot 5^{n_3} \cdots p^{n_k},$$

where p is the kth prime number. Given the number e, we can reverse the process by factoring e into distinct prime powers, observing the exponents, then decoding each exponent into the program line it represents.

The idea of associating numbers with programs or formulas or any set of formal expressions is credited to Gödel, who used this technique in his famous paper [**8**] on undecidability. **Gödel number** is now a standard term for the numbers that result. In this article, if e is the Gödel number of a computer program, then I shall write $\{e\}$ to denote that program. If the number e is not the Gödel number of a program, then $\{e\}$ will denote some particular program (for you to choose) that never gives an output, regardless of the input. So now every number e represents some program $\{e\}$, and every program is represented by at least one number. I say "at least one" because the particular program that I invited you to choose will be $\{e\}$ for many numbers e. Furthermore, if you consider a program to be unchanged by certain minor rearrangements, then many programs will be represented by more than one e.



"Gödel number, please?"

Constructive Ordinal Notation Systems

S. Kleene gave the following definition that unites most of the ideas discussed so far, using the term "r-system" ([**11**]). In the definition, L is the set of ordinal notations and f matches the notations with the ordinals they name, while K, P and Q provide the essential information.

DEFINITION. A **constructive ordinal notation system** (CONS) is a pair (L, f) in which L is a set of natural numbers, f is a function from L into the ordinal numbers, and there are programs K, P and Q having the properties listed below.

1. If $f(x) = 0$, then $K(x) = 1$,
   if $f(x)$ is a successor ordinal, then $K(x) = 2$, and
   if $f(x)$ is a limit ordinal, then $K(x) = 3$.
2. If $f(x)$ is a successor ordinal $\beta + 1$, then $P(x)$ is a notation for the immediate predecessor β.
3. If $f(x)$ is a limit ordinal λ, then $Q(x)$ is the Gödel number of a program, and the outputs $\{Q(x)\}(0), \{Q(x)\}(1), \{Q(x)\}(2), \ldots$ are notations for a fundamental sequence for λ.

EXAMPLE. Let L be the set whose elements are all numbers of the form 2^n or $2^n \cdot 3$ for $n \geqslant 0$. We shall take $1, 2, 2^2, 2^3, \ldots$ as notations for the finite ordinals $0, 1, 2, 3, \ldots$, respectively, and $3, 2 \cdot 3, 2^2 \cdot 3, 2^3 \cdot 3, \ldots$ as notations for $\omega, \omega + 1, \omega + 2, \omega + 3, \ldots$, respectively. Consequently the ordinal-assigning function f is given by

$$f(x) = \begin{cases} n & \text{if } x = 2^n \\ \omega + n & \text{if } x = 2^n \cdot 3. \end{cases}$$

The programs K and P should be constructed so as to have the following output features:

$$K(x) = \begin{cases} 1 & \text{if } x = 1 \\ 3 & \text{if } x = 3 \\ 2 & \text{if } x = 2^n \text{ or } 2^n \cdot 3 \text{ for } n \geqslant 1, \end{cases} \qquad P(x) = \begin{cases} 2^n & \text{if } x = 2^{n+1} \\ 2^n \cdot 3 & \text{if } x = 2^{n+1} \cdot 3. \end{cases}$$

We have considerable leeway in constructing K and P, because their behavior on inputs not belonging to L is irrelevant. For instance, we could construct P so that $P(x) = [x/2]$ for all x, using the greatest-integer function to insure integer outputs.

Since 3 is the only notation for a limit ordinal ($f(3) = \omega$), $Q(3)$ is the only output that must be carefully planned when you construct program Q for this CONS. Find a program that prints out the value 2^n for each input number n, and suppose that e is the Gödel number of the program. Then $\{e\}$ generates notations for a fundamental sequence for ω, and therefore Q could be any program constructed so that $Q(3) = e$.

The next example is patterned after the system S_1 defined by Kleene in [**11**]. It will turn out to be a maximal CONS.

EXAMPLE. $\mathscr{S} = (\mathscr{L}, \mathscr{F})$. I shall describe $\mathscr{L}$ and $\mathscr{F}$ by indicating for each ordinal number β which numbers belong to $\mathscr{L}$ as notations for β, that is, by describing the set $\mathscr{F}^{-1}(\beta)$. $\mathscr{L}$ is then the union of all the nonempty sets $\mathscr{F}^{-1}(\beta)$.

(i) 0 is the unique notation for 0.
(ii) 2^x is a notation for $\alpha + 1$ if and only if x is a notation for α.
(iii) 3^e is a notation for the limit ordinal λ if and only if e is a Gödel number such that $\{e\}(0), \{e\}(1), \{e\}(2), \ldots$ are notations for a fundamental sequence for λ.

In $\mathscr{S}$, then, the finite ordinals $0, 1, 2, 3, 4, 5, 6, \ldots$ receive as their unique notations the numbers

$$0, 1, 2, 4, 16, 2^{16}, 2^{(2^{16})}, \ldots. \tag{1}$$

Beginning with ω, however, each ordinal has many notations. The notations for ω are numbers of

the form 3^e, where $\{e\}$ is a program whose successive outputs $\{e\}(0), \{e\}(1), \{e\}(2), \ldots$ are an increasing subsequence of (1). For each of these notations 3^e, the number $2^{(3^e)}$ is a notation for $\omega + 1$, and so on for $\omega + 2, \omega + 3, \ldots$. Likewise, $\omega \cdot 2$ has many notations, all of the form 3^e, and from them we obtain notations for succeeding ordinals.

The following theorem is a companion to the definition of $\mathscr{S}$, for it insures that $\mathscr{L}$ and $\mathscr{F}$ are well defined. The proofs are examples of ordinal induction. They show that the properties ascribed to ordinal β in the theorem are indeed true for all ordinals. The approach is: show that the property is true for 0; show that the property is true for $\alpha + 1$, if it is true for α; show that the property is true for a limit ordinal λ, if it is true for all ordinals less than λ.

THEOREM. *Let β be any ordinal. In the definition of $\mathscr{S}$,*
(a) *the set $\mathscr{F}^{-1}(\beta)$ is defined, and*
(b) *$\mathscr{F}^{-1}(\beta)$ is disjoint from $\mathscr{F}^{-1}(\delta)$ for all $\delta \neq \beta$.*

Proof. (a) By part (i) of the definition of $\mathscr{S}$, $\mathscr{F}^{-1}(0) = \{0\}$. Next, assume $\mathscr{F}^{-1}(\alpha)$ is defined. Then $\mathscr{F}^{-1}(\alpha + 1) = \{2^x : x \in \mathscr{F}^{-1}(\alpha)\}$. Finally, assume that $\mathscr{F}^{-1}(\alpha)$ is defined for each ordinal α less than the limit ordinal λ. Then $\mathscr{F}^{-1}(\lambda) = \{3^e : \{e\}(0), \{e\}(1), \{e\}(2), \ldots$ belong respectively to sets $\mathscr{F}^{-1}(\alpha_0), \mathscr{F}^{-1}(\alpha_1), \mathscr{F}^{-1}(\alpha_2), \ldots$, where $\alpha_0, \alpha_1, \alpha_2, \ldots$ is a fundamental sequence for $\lambda\}$.

(b) Clearly $\mathscr{F}^{-1}(0)$ is disjoint from $\mathscr{F}^{-1}(\delta)$ for all $\delta \neq 0$. Next, assume that $\mathscr{F}^{-1}(\alpha)$ is disjoint from $\mathscr{F}^{-1}(\delta)$ for all $\delta \neq \alpha$; we show that $\mathscr{F}^{-1}(\alpha + 1)$ must also have the disjointness property. Let $\mathscr{F}^{-1}(\alpha + 1)$ and $\mathscr{F}^{-1}(\delta)$ have some notation in common. That notation must be of the form 2^x, where $x \in \mathscr{F}^{-1}(\alpha)$. Furthermore, δ must be a successor ordinal, and $x \in \mathscr{F}^{-1}(\delta - 1)$. Consequently $\delta - 1$ must equal α by our assumption about $\mathscr{F}^{-1}(\alpha)$, and so $\delta = \alpha + 1$.

Finally, assume that each ordinal less than the limit ordinal λ has the disjointness property described in the theorem. To see that λ must also have the property, let $\mathscr{F}^{-1}(\lambda)$ and $\mathscr{F}^{-1}(\delta)$ have some notation in common. That notation must be of the form 3^e, where $\{e\}(0), \{e\}(1), \{e\}(2), \ldots$ are notations for a fundamental sequence for λ and also for δ. By assumption, each of the notations $\{e\}(n)$ denotes a unique ordinal. Furthermore, it is a set-theoretical fact that the limit of a fundamental sequence is unique. Therefore it must be the case that $\delta = \lambda$.

The last thing we do here, and the simplest, is to verify that $\mathscr{S}$ is in fact a CONS. It's really a do-it-yourself verification. Using your chosen programming language, you can construct programs K, P and Q such that

$$K(x) = \begin{cases} 1 & \text{if } x \text{ is } 0 \\ 2 & \text{if } x \text{ is positive and even} \\ 3 & \text{otherwise,} \end{cases}$$

$$P(x) = y \quad \text{if } x = 2^y,$$
$$Q(x) = e \quad \text{if } x = 3^e.$$

Programs with those characteristics can serve as the auxilary programs for the CONS $\mathscr{S}$.

Three theorems from computability theory will be crucial to the proof of $\mathscr{S}$'s maximality. The programs mentioned in the first two theorems are of fundamental importance, as is evidenced by the fact that [**13**] calls a programming system "acceptable" if and only if it contains them. The third theorem, the Recursion Theorem, is satisfied by an acceptable programming system, and it serves to justify program descriptions in which certain outputs are affected by others (the recursion aspect). To quote [**14**], p. 179, "It is a deep result in the sense that it provides a method for handling, with elegance and intellectual economy, constructions that would otherwise require extensive, complex treatment."

Given any two numbers e and n, and possibly considerable patience, one can determine the program $\{e\}$, apply it with input n, and give the output $\{e\}(n)$ when and if the computation ever halts. As a matter of fact, this whole process can be carried out by a sort of universal program U whose existence is stated below in the Enumeration Theorem. The title of the theorem is prompted by the way U "enumerates" all possible programs $\{e\}$ as e ranges through the natural numbers. The program U acts something like a modern day compiler.

ENUMERATION THEOREM. *There is a program U such that $U(e, n) = \{e\}(n)$ for all numbers e and n.* ([**13**], p. 82; [**14**], p. 22, ϕ_e means $\{e\}$.)

Another task that can be handled by a special program is that of forming compositions of given programs.

COMPOSITION THEOREM. *There is a program C such that for all numbers z and e, $C(z, e)$ is a Gödel number of a program, and $\{C(z, e)\}(n) = \{z\}(\{e\}(n))$ for all numbers n.* ([**13**], p. 83; [**14**], p. 24.)

The fact that the next theorem has many useful applications is indicated by the wide variety of forms in which it has been expressed. The one given here is best suited to the application I shall make of it with respect to $\mathscr{S}$.

RECURSION THEOREM. *If F is any program using two inputs, then there is a number z_0 such that $F(z_0, x) = \{z_0\}(x)$ for all numbers x.* ([**5**], p. 176, Theorem 7.4)

This result can be viewed as a fixed-point theorem, if we let F_z denote the one-input program obtained by fixing z as the first input of F. According to the Recursion Theorem, the mapping $\{z\} \to F_z$ has a fixed point.

The Main Theorem

As an exercise you might prove by ordinal induction that the ordinals named by a CONS form an initial segment of the ordinal numbers. In other words, if β receives a notation from (L, f), then so do all ordinals less than β. The next result shows that $\mathscr{S}$, defined in our previous section, is maximal in that it covers all the ordinals named by any other CONS. It even claims the existence of a program T that transforms the notations of a given CONS into corresponding ones of $\mathscr{S}$.

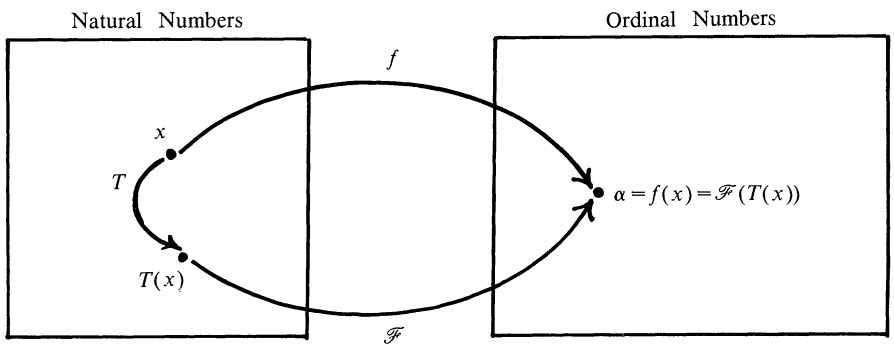
THEOREM. *If (L, f) is a given* CONS, *then there is a program T such that $T(L) \subseteq \mathscr{L}$ and $\mathscr{F}(T(x)) = f(x)$ for each x in L.*

Proof. Let K', P' and Q' be the auxiliary programs of the given CONS. With the definitions of CONS and $\mathscr{S}$ in mind, we will seek a program T with the following features:

$$T(x) = \begin{cases} 0 & \text{if } x \text{ denotes } 0 \\ 2^{T(P'(x))} & \text{if } x \text{ denotes a successor} \\ 3^e & \text{if } x \text{ denotes a limit ordinal,} \end{cases} \tag{2}$$

where e is a Gödel number, and $\{e\}(n) = T(\{Q'(x)\}(n))$ for all numbers n. The proof that such a program T (if it exists) would satisfy this theorem is by induction on the ordinal β denoted by x, and is left as an exercise. Whether or not you perform that exercise, you should inspect the description of T closely enough to see how it provides $\mathscr{S}$-notations for the ordinals named by (L, f). Also notice that recursion is involved in the second and third lines of (2), where $T(x)$ is determined by T values at notations for previous ordinals.

The following construction of T shows a typical application of the Recursion Theorem. In definition (3) below, think of the z's as candidates for being a Gödel number of T. When we get a fixed point z_0 for (3), $\{z_0\}$ will have exactly the properties in (2).



The relationships among f, $\mathscr{F}$ and T

Suppose that F is a program with the following features, where U and C are the programs mentioned in the Enumeration Theorem and the Composition Theorem.

$$F(z,x)=\begin{cases}0 & \text{if } K'(x)=1\\ 2^{U(z,P'(x))} & \text{if } K'(x)=2\\ 3^{C(z,Q'(x))} & \text{if } K'(x)=3.\end{cases} \tag{3}$$

A brief inspection of (3) suggests how programs K', P', Q', U and C are used as subroutines in constructing F.

Now let z_0 be the special number that is predicted in the Recursion Theorem, and let's rewrite the information of (3) in other terms.

$$F(z_0,x)=\{z_0\}(x)=\begin{cases}0 & \text{if } x \text{ denotes } 0\\ 2^{\{z_0\}(P'(x))} & \text{if } x \text{ denotes a successor}\\ 3^{e} & \text{if } x \text{ denotes a limit ordinal,}\end{cases}$$

where $e=C(z_0,Q'(x))$ is a Gödel number of $\{z_0\}$ composed with $\{Q'(x)\}$. This program $\{z_0\}$ has all the features described in (2), and so it is the T we are after.

Let λ be the least ordinal not provided with a notation by $\mathscr{S}$. You can check that λ would have to be a limit ordinal. Now, it would seem easy enough to extend $\mathscr{S}$ by simply adjoining a notation for λ, but since $\mathscr{S}$ is maximal, there is presumably no way to make a CONS that extends it. Suppose we attempt an extension by selecting the number 5 to represent λ. Then $\mathscr{F}(5)=\lambda$, $K(5)=3$, $P(5)$ is irrelevant, and $Q(5)=e$ where $\{e\}$ is a program whose outputs name a fundamental sequence for λ. But in that case 3^e would already belong to $\mathscr{S}$ as a notation for λ, which contradicts our definition of λ. There are other strategies that one might use in attempting an extension (e.g., allow new notations for ordinals less than λ so that $\{e\}$ can approach λ by a new route), but the maximality of $\mathscr{S}$ ensures that all of them will fail. $\mathscr{S}$ reaches as far into the transfinite as is possible for a CONS.

Incidentally, the ordinals for which $\mathscr{S}$ provides notations are called the **constructive ordinals**. Since $\mathscr{S}$ is a CONS, they form a countable initial segment of Cantor's second number class.

Questions of Recursiveness

A final question I want to raise about CONS' in general and $\mathscr{S}$ in particular is a global one about the nature of the whole set of notations. One requirement that we might have considered adding to the essential information is "You can tell whether a given expression is an ordinal notation." The question here is one of recognizing a notation when you see one. For example, a

polynomial in ω is certainly recognizable as such. On the other hand, it isn't clear how easily one could determine whether or not a given number 3^e is an ordinal notation in $\mathcal{S}$.

Similar questions of recognition occur with regard to Gödel numbers for programs. For example, "Can one recognize whether a given number e is the Gödel number of a program?" and "Can one recognize whether e is the Gödel number of a program that gives an output for every input n?"

To put it another way, we are concerned here with the complexity of a set of numbers, whether it be a set of ordinal notations or a set of Gödel numbers. One way to make the question more rigorous is to ask whether the given set is **recursive**, which is to say whether there exists an algorithm for distinguishing the members of that set from all other numbers. It turns out that for $\mathcal{S}$ or any other maximal CONS, the set of ordinal notations is not recursive, and so in this sense one must sacrifice recognition in order to have maximality. (This result is a corollary of facts derived in [**12**] about the complexity of another maximal CONS called $\mathcal{O}$.)

Incidentally, the answer to the first Gödel number question I mentioned above depends partly upon the language (or dialect thereof) that you assume, and partly upon what you require of a bona fide program. In the case of extremely lean and simple programming systems such as Turing machines, the answer is "yes" ([**5**], p. 60, item (11)). The second Gödel number question above is similar to the famous Halting Problem of computability theory and has a negative answer (Section 1.9 of [**14**], especially Theorem VIII). See [**14**], Section 2.2, for notes about similar questions in other branches of mathematics, including **Hilbert's tenth problem**. A good up-to-date exposition of that famous problem is contained in [**6**].

Recursive Ordinals

The two basic ingredients in our approach to the constructive ordinals have been algorithms, as embodied in computer programs, and natural numbers. The former give rise to the term "constructive," while the latter are a convenient though not really essential choice as ordinal notations. Another way to combine these ingredients in the study of ordinals is via the well-ordering concept, as follows. An ordinal is called **recursive** if it is the order type of an algorithmic well-ordering of natural numbers. In other words, the ordinal must be order-isomorphic to some well-ordering W whose field is a set of natural numbers and whose ordered pairs can be distinguished from all others by some algorithm.

For example, to show that $\omega + 2$ is recursive, one might produce the sequence

$$2,3,4,5,\ldots,0,1 \tag{4}$$

in which 0 and 1 follow all the other numbers, then define W to be $\{(x, y)$: x precedes y in (4)$\}$ and show by means of a program that W is algorithmic. More generally, it is not hard to show that the set of all recursive ordinals is a countable (since there are only countably many algorithms) initial segment of Cantor's second number class, and the least nonrecursive ordinal is a limit ordinal.

This recursive ordinal concept is a fairly straightforward constructive analogue of Cantor's well-ordering approach to the ordinals, whereas the main thrust of this article has been to constructivize Cantor's principles of ordinal generation. Remarkably, these two avenues lead to the same destination, for the recursive ordinals turn out to be precisely the constructive ordinals ([**14**], Section 11.8). This result is particularly satisfying because it gives evidence that the set of constructive ordinals is a natural one, being stable under two quite different characterizations.

Acknowledgement

I wish to thank the editor and a referee for valuable suggestions in the preparation of this article. I also wish to thank my friend and former teacher and advisor Don Kreider for introducing me to this subject matter.

References

[1] G. Boolos and R. Jeffrey, Computability and Logic, 2nd ed., Cambridge Univ. Press, 1980.
[2] G. Cantor, Beiträge zur Begründung der Transfiniten Mengenlehre, Math. Ann., 46 (1895) 481–512 and 49 (1897) 207–246.
[3] _____, Contributions to the Founding of the Theory of Transfinite Numbers, Dover, 1947, English translation with introduction and notes by P. E. B. Jourdain.
[4] W. Dauben, George Cantor and the origins of transfinite set theory, Scientific American, 248 (1983) 122–131.
[5] M. Davis, Computability and Unsolvability, McGraw-Hill, 1958.
[6] _____, Hilbert's tenth problem is solvable, Amer. Math. Monthly, 80 (1973) 233–269.
[7] K. Devlin, Fundamentals of Contemporary Set Theory, Springer-Verlag, 1979.
[8] K. Gödel, Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme, I, Monatshefte fur Mathematik und Physik, 38 (1931) 173–198.
[9] P. Halmos, Naive Set Theory, Van Nostrand, 1960.
[10] E. Kamke, Theory of Sets, Dover, 1950.
[11] S. C. Kleene, On notation for ordinal numbers, J. Symbolic Logic, 3 (1938) 150–155.
[12] _____, On the forms of predicates in the theory of constructive ordinals (second paper), Amer. J. Math., 77 (1955) 405–428.
[13] M. Machtey and P. Young, An Introduction to the General Theory of Algorithms, North-Holland, 1978.
[14] H. Rogers, Jr., Theory of Recursive Functions and Effective Computability, McGraw-Hill, 1967.
[15] S. Willard, General Topology, Addison-Wesley, 1968.

Our Mathematical Alphabet

As I recite the alphabet to one who's only three
The world of mathematics is opened up to me.

For a, b, c are constants or parameters assigned
And D is a determinant or distance undefined.

The image which e gives to me is one I can't erase
For it can only mean for me the logarithmic base.

F, G, H are functions with appropriate domain
And i's a unit vector in the Gauss or complex plane.

J's a Bessel function and another kind is K.
L's a linear operator or inductance one could say.

M and N are integers but m could be a mass.
O is the number zero but $\varnothing$'s the empty class.

P and Q give odds that you will win or lose a bet.
R gives correlation of two variables you have met.

I see before me Einstein's world when I hear S and T
For they make me think of space and time and relativity.

At this point I'm so deep in thought of time and space and such
That velocity components u, v, w don't seem much.

Perhaps some day that three-year old may learn when fully grown
Why x, y, z imply for me how much there is unknown.

—M. R. SPIEGEL

References

[1] G. Boolos and R. Jeffrey, Computability and Logic, 2nd ed., Cambridge Univ. Press, 1980.
[2] G. Cantor, Beiträge zur Begründung der Transfiniten Mengenlehre, Math. Ann., 46 (1895) 481–512 and 49 (1897) 207–246.
[3] _____, Contributions to the Founding of the Theory of Transfinite Numbers, Dover, 1947, English translation with introduction and notes by P. E. B. Jourdain.
[4] W. Dauben, George Cantor and the origins of transfinite set theory, Scientific American, 248 (1983) 122–131.
[5] M. Davis, Computability and Unsolvability, McGraw-Hill, 1958.
[6] _____, Hilbert's tenth problem is solvable, Amer. Math. Monthly, 80 (1973) 233–269.
[7] K. Devlin, Fundamentals of Contemporary Set Theory, Springer-Verlag, 1979.
[8] K. Gödel, Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme, I, Monatshefte fur Mathematik und Physik, 38 (1931) 173–198.
[9] P. Halmos, Naive Set Theory, Van Nostrand, 1960.
[10] E. Kamke, Theory of Sets, Dover, 1950.
[11] S. C. Kleene, On notation for ordinal numbers, J. Symbolic Logic, 3 (1938) 150–155.
[12] _____, On the forms of predicates in the theory of constructive ordinals (second paper), Amer. J. Math., 77 (1955) 405–428.
[13] M. Machtey and P. Young, An Introduction to the General Theory of Algorithms, North-Holland, 1978.
[14] H. Rogers, Jr., Theory of Recursive Functions and Effective Computability, McGraw-Hill, 1967.
[15] S. Willard, General Topology, Addison-Wesley, 1968.

Our Mathematical Alphabet

As I recite the alphabet to one who's only three
The world of mathematics is opened up to me.

For a, b, c are constants or parameters assigned
And D is a determinant or distance undefined.

The image which e gives to me is one I can't erase
For it can only mean for me the logarithmic base.

F, G, H are functions with appropriate domain
And i's a unit vector in the Gauss or complex plane.

J's a Bessel function and another kind is K.
L's a linear operator or inductance one could say.

M and N are integers but m could be a mass.
O is the number zero but $\varnothing$'s the empty class.

P and Q give odds that you will win or lose a bet.
R gives correlation of two variables you have met.

I see before me Einstein's world when I hear S and T
For they make me think of space and time and relativity.

At this point I'm so deep in thought of time and space and such
That velocity components u, v, w don't seem much.

Perhaps some day that three-year old may learn when fully grown
Why x, y, z imply for me how much there is unknown.

—M. R. SPIEGEL

NOTES

Designing an Auditing Procedure, or How to Keep Bank Managers on Their Toes

C. L. MALLOWS
N. J. A. SLOANE
Mathematics and Statistics Research Center
Bell Laboratories
Murray Hill, NJ 07974

A bank inspector has N banks under his supervision, and wishes to plan his visits to these banks for many years in advance so that

(i) there is a high probability, or even certainty, that every bank will be audited at least once a year,
(ii) the visits are unexpected, and
(iii) the total number of banks he visits each year is minimized.

Furthermore the plan must be fair: every bank must be treated in the same way. We wish to design a method for selecting M banks to be audited each week so as to satisfy conditions (i)–(iii) as far as possible. We assume the banks will know (or can find out) what algorithm we intend to use.

We first analyze two of the more obvious inspection plans (Plans A and B), and show that neither meets all the requirements. Then three schemes (Plans C, D and E) are described that do satisfy (i)–(iii). Plans C and D both have drawbacks, however, and the final scheme, Plan E, appears to be the best. Some numerical values of the parameters for the case of 2000 banks are given in TABLES I–III.

We are not aware of any previous work on this problem, although [**4**], [**8**], [**9**] and the paradox of the unexpected hanging [**3**] are tangentially related.

The inspection plans

We assume that the diligent inspector spends 50 weeks each year visiting his banks (this parameter can easily be changed.) The other parameters in the analysis are:

N = total number of banks (assumed for simplicity to be a multiple of 50),
P = probability that every bank is visited at least once during the year,
Q = probability that the ith bank is visited more than once during the year (this will be independent of i),
$H_{g,t}$ = probability that the ith bank is visited in week t, conditional on that bank having been last visited g weeks previously (i.e., in week $t-g$), for $t = 1,2,\ldots,50$ and $g = 1,2,\ldots,t-1$,
T = average number of banks visited by the inspector during each year, and
W = average number of weeks between visits to the ith bank.

In practice, a team of inspectors would be used to visit such a large number of banks each week. But for simplicity we speak as if only a single inspector were involved. Also, unlike the notorious traveling salesman problem, we are not concerned with the distance the inspector has to travel each week. (Perhaps the inspections can be carried out by telephone.)

Our first three sampling schemes work on a calendar year basis; everything starts afresh at the beginning of the year.

Plan A: Weekly selection without replacements. At the beginning of each year N balls bearing the names of the banks are placed in an urn and thoroughly shuffled. Each week $M = N/50$ balls are drawn without replacement and these banks are inspected.

This plan certainly satisfies conditions (i) and (iii), since by the end of the year all the banks will have been visited exactly once (in fact $P = 1$, $Q = 0$ and $T = N$). Condition (ii) is partially satisfied, since a bank manager does not know in advance when he will be audited. On the other hand once he has been visited he knows he is safe from further inspection for the rest of the year, so that $H_{g,t} = 0$ for all g, $t \geqslant 1$. With this scheme the average time W between the inspector's visits is one year. Since one of the goals implied by condition (ii) is that the inspections may occur at any time, this plan is not completely satisfactory. Also there is a good chance that the time between visits will be substantially longer than one year, if a visit is made early in one year and the next visit is late in the following year.

Plan B: Weekly selection with replacement between weeks. Each week the names of *all* N banks are mixed in an urn and a random subset of M names is selected (without replacement) for inspection.

This plan certainly satisfies condition (ii), for whether or not a bank is inspected one week is independent of whether it was inspected the previous week. In fact $H_{g,t} = M/N$, independently of g and t. Furthermore it is clear that, if M is sufficiently large, condition (i) is also satisfied. How large must M be? To answer this, we note that this is a version of the classical occupancy problem [**1**], [**2**], [**5**].

In one year there are a total of $\binom{N}{M}^{50}$ possible ways of selecting the banks to be inspected, each having probability $\binom{N}{M}^{-50}$. The number of these selections having the property that a particular set of k banks all fail to be visited during the year is $\binom{N-k}{M}^{50}$. The principle of inclusion and exclusion [**1**, p. 242], [**2**, Section IV.2] shows that the probability that there are exactly k banks not visited during the year is

$$\binom{N}{k} \sum_{i=0}^{N-k-M} (-1)^i \binom{N-k}{i} \binom{N-k-i}{M}^{50} \Big/ \binom{N}{M}^{50}. \tag{1}$$

In particular, the probability that every bank is visited in the year is

$$P = \sum_{i=0}^{N-M} (-1)^i \binom{N}{i} \binom{N-i}{M}^{50} \Big/ \binom{N}{M}^{50}. \tag{2}$$

This expression can be evaluated exactly only for small values of N and M. For larger values the following Poisson approximation is appropriate (compare [**1**, Chapter 14], [**2**, p. 94]). The probability that the ith bank is not visited during the year is

$$\frac{\binom{N-1}{M}^{50}}{\binom{N}{M}^{50}} = \left(1 - \frac{M}{N}\right)^{50},$$

and therefore the average number that are not visited is

$$\lambda = N\left(1 - \frac{M}{N}\right)^{50}. \tag{3}$$

It is straightforward to show from (1) that if N and M increase while λ remains fixed, then the probability that exactly k banks remain unvisited is approximately

$$e^{-\lambda} \frac{\lambda^k}{k!}.$$

In particular, the probability that every bank is visited is

$$P \approx e^{-\lambda}. \tag{4}$$

Thus using Plan B with N banks and a desired probability P, we choose a value of λ from (4), and take

$$M = N\left\{1 - \left(\frac{\lambda}{N}\right)^{1/50}\right\}. \tag{5}$$

The total number of banks visited during the year is then

$$T = 50M \approx 50N\left\{1 - \left(\frac{\ln P^{-1}}{N}\right)^{1/50}\right\}. \tag{6}$$

The probability that the ith bank is visited more than once is

$$\begin{aligned} Q &= 1 - \left(1 - \frac{M}{N}\right)^{50} - \frac{50M}{N}\left(1 - \frac{M}{N}\right)^{49} \\ &\approx 1 - \frac{\lambda}{N} \cdot \frac{N + 49M}{N - M}, \end{aligned} \tag{7}$$

using (3), and the average time between visits is approximately

$$W = \frac{50}{P} = 50e^{\lambda} \text{ weeks}. \tag{8}$$

The difference $N - M$ is $\lambda^{.02}N^{.98}$ (from (5)), which approaches 0 as N increases (for λ fixed). TABLE I gives some typical values of P, λ, M and T for $N = 2000$. It is clear that although conditions (i) and (ii) are satisfied, (iii) is not, as T is unacceptably large.

It is worth remarking that, in the case $M = 1$, (4) also leads to an approximation for the Stirling numbers of the second kind. Although this approximation seems very natural in the present context, it does not appear to have been pointed out before. (It is not included for example among the half-dozen similar approximations analyzed in Chapter 16 of [1].)

Our next scheme is a generalization of Plans A and B, and appears to satisfy all three conditions.

TABLE I. **Typical values of the parameters for Plan B when visiting $N = 2000$ banks. The columns show P = probability that every bank is visited in the year, λ = average number not visited, M = number visited each week, $T = 50M$ = total number of visits during a year.**

P	λ	M	T
.75	.288	324	16220
.80	.223	333	16640
.85	.163	343	17170
.90	.105	358	17880
.95	.051	381	19060

Plan C: A and B combined. As in Plan A, $M = N/50$ names are drawn without replacement for each of the 50 weeks. In addition, for each week we place the $N - M$ names not drawn for that week in an urn, and choose from it a random subset of K additional names. All $M + K$ banks are visited that week.

This plan combines some of the best features of Plans A and B, since each bank will be drawn exactly once during the year in one of the sets of size M, and in addition may also be drawn in any other week in one of the sets of size K. If $K = 0$, it reduces to Plan A, while if we change M from $N/50$ to zero, it reduces to Plan B. For this scheme $P = 1$, $T = 50(M + K)$ and

$$Q = 1 - \left(1 - \frac{K}{49M}\right)^{49}. \tag{9}$$

The parameters $H_{g,t}$ and W are more difficult to calculate, and a formula for W is derived at the end of this paper. We can see, however, that $H_{g,t}$ depends strongly on g and t. Suppose for example that $g \geqslant t$, so that in the current year the ith bank has not yet been visited. Then

$$H_{g,t} = \frac{M}{N - M(t-1)} + \frac{N - Mt}{N - M(t-1)} \cdot \frac{K}{N - M}$$
$$= \frac{1}{51 - t}\left(1 + (50 - t)\frac{K}{49M}\right),$$

which increases from $(M + K)/N$ to 1 as t goes from 1 to 50. If $g < t$, the situation is improved; for these cases we have $K/N \leqslant H_{g,t} \leqslant (K + M)/N$. (We are assuming that when it is visited, a bank does not learn whether the visit is an M-type or a K-type.)

Some typical values of the parameters when $N = 2000$ are shown in TABLE II. Taking K in the range 10 to 40 gives an inspection scheme which meets requirements (i)–(iii), and were it not for the strong dependence of $H_{g,t}$ on g and t in the range $g \geqslant t$, Plan C would be a quite satisfactory solution.

The last two plans to be considered are *stationary*, in the technical sense that no part of the calendar year has any special role. They do not have the drawbacks of Plans A, B and C that the probability that a particular bank is visited in a particular week can vary over a wide range as a function of the calendar date. As a side benefit, this approach will result in our satisfying condition (i) in a stronger sense than before; we shall be able to guarantee that no bank is ever left unvisited for more than twelve months.

TABLE II. **Typical values of the parameters for Plan C when $N = 2000$ and $M = 40$. The columns show K = number of banks chosen randomly (in addition to the M that are chosen without replacement), $1 - Q$ = probability that any given bank is visited more than once during the year, $T = 50(M + K)$ = total number of visits, W = average time between visits.**

K	$1 - Q$	T	W
0	1	2000	50
10	.778	2500	40
20	.605	3000	33.3
30	.470	3500	28.6
40	.364	4000	25
50	.282	4500	22.2
60	.218	5000	20
70	.168	5500	18.2
80	.130	6000	16.7

Plan D: Independent renewals. To implement this scheme, we first specify a probability distribution $\pi = \{\pi_1, \pi_2, \ldots, \pi_{50}\}$ on the integers $1, 2, \ldots, 50$. (We shall soon see what properties this distribution should have.) Then, for each bank separately, the following plan is followed. First, an initial visit is scheduled, for week t_1, say, in a way that will be described shortly. Then an integer g_1 between 1 and 50 is chosen, with $\text{Prob}\{g_1 = j\} = \pi_j$, for $1 \leqslant j \leqslant 50$, and the second visit is scheduled for week $t_1 + g_1$. A second integer g_2 is chosen (with the same probability distribution as g_1), independent of everything else so far, and the third visit is scheduled for $t_1 + g_1 + g_2$, and so on. Once the plan is under way, the average time between visits is $W = \pi_1 + 2\pi_2 + \cdots + 50\pi_{50}$, which is simply the mean of the distribution π.

We have still to specify how t_1 is to be chosen. We do this in such a way as to ensure that the probability that a chosen bank is visited in any particular week is a constant, independent of the week. The constant will turn out to be simply $1/W$. It is easy to see that for this to happen we must make

$$\text{Prob}\{t_1 = k\} = \frac{1}{W}(\pi_k + \pi_{k+1} + \cdots + \pi_{50}), \tag{10}$$

for $k = 1, \ldots, 50$. This much complication seems to be unavoidable if the sampling plan is to be completely independent of the calendar.

Since under Plan D the banks are treated independently, it is straightforward to derive the following formulae for P, Q (and more generally the probability that a particular bank is visited more than once during *any* period of 50 consecutive weeks), $H_{g,t}$, and T (and more generally the average number of banks visited during any period of 50 consecutive weeks). We denote the right-hand side of (10) by p_k. Then

$$P = 1,$$

$$Q = \sum_{k=1}^{49} p_k(\pi_1 + \pi_2 + \cdots + \pi_{50-k})$$

$$= 1 - W\sum_{k=1}^{50} p_k p_{51-k}, \tag{11}$$

$$H_{g,t} = \frac{\pi_g}{\pi_g + \pi_{g+1} + \cdots + \pi_{50}} = \frac{\pi_g}{Wp_g}, \tag{12}$$

$$T = 50N/W. \tag{13}$$

Thus $H_{g,t}$ is independent of t, as desired, and can be written simply as H_g. Note that $H_{50} = 1$. Also M, the number of banks that are visited in any given week, has an average value of N/W and a variance of $N(W-1)/W^2$. These quantities do not depend on the distribution π, except through its mean W, and we choose this so that the resulting value of M has a high probability of being acceptable.

For example, when $N = 2000$, if we wish to make no more than 70 inspections per week, we may take $W = 40$, so that M has a mean of 50 and a standard deviation of 6.98, and so has only a small chance of exceeding 70. (This would happen with probability 0.003, i.e., about once every six years.) The total number T of inspections per 50-week period is 2500 ± 49.4.

Now we consider how the probability distribution π should be chosen, supposing that its mean W is given in advance. Clearly it is undesirable to have $\pi_g = 0$ for any g, since this makes $H_g = 0$ and the bank is certain not to be visited that week. Also H_{50} is constrained to be 1. A reasonable choice for π is to make $H_1 = H_2 = \cdots = H_{49} = H$, say, which we can do by setting

$$\pi_k = H(1-H)^{k-1}, \quad k = 1, \ldots, 49, \tag{14}$$

$$\pi_{50} = (1-H)^{49}, \tag{15}$$

TABLE III. **Typical values of the parameters for Plans D and E when $N = 2000$. For Plan D the columns show W = average time between visits to any bank, T = average number of banks visited in any 50-week period, H = probability that a particular bank is visited in any week, conditioned on the previous visit having been less than 50 weeks earlier, π_{50} = probability that a bank goes 49 weeks without a visit, and M = average number of banks visited per week. These values also apply (approximately) to Plan E, in which case T and M are constants.**

W	T	H	π_{50}	M
44.34	2255	.005	.782	45
39.50	2532	.01	.611	51
35.35	2829	.015	.477	57
31.79	3146	.02	.371	63
28.72	3482	.025	.289	70
26.06	3837	.03	.224	77
23.76	4209	.035	.174	84
21.75	4598	.04	.135	92
20.00	5000	.045	.105	100
18.46	5417	.05	.081	108

where H is determined by the equation

$$\text{mean}\{\pi\} = \frac{1}{H}\left(1-(1-H)^{50}\right) = W. \tag{16}$$

Also

$$p_k = (1-H)^{k-1}/W \text{ for } k = 1,\ldots,50. \tag{17}$$

Some numerical values are given in TABLE III.

The probability distribution π we have arrived at, given by equations (14) and (15), is very nearly the distribution of waiting-time until the appearance of the first head in a sequence of independent coin-tosses with Prob{head} = H at each trial. The only difference is that at trial 50 the outcome "head" is forced. Thus Plan D is only a minor modification of Plan B! This suggests two things: Plan D should be easy to implement, and can be easily modified to make the number of inspections each week a constant.

To implement Plan D, once the start-up phase is over and every bank has been visited at least once, each week the inspector must visit

(a) all banks that have gone 49 weeks without a visit, and

(b) a random selection of the remaining banks, with each bank having independently a chance H of being selected.

In the start-up phase, we must implement the probability distribution $(p_1,\ldots,p_{50})$ given in (10), (17). For each bank, we simply toss a coin having Prob{head} = H until the first head appears, and schedule the first visit to occur in the corresponding week, except that if no head appears in the first 50 tosses, we start again at week 1.

The modification of Plan D in which the number of inspections per week is constant is given by our final scheme, Plan E, which combines most of the attractive features of Plans B and D.

Plan E: Truncated geometric renewals, constrained to have fixed sample size. In this scheme, after a start-up phase that is described below, each week we determine which banks have gone 49 weeks without a visit. Suppose there are S of them. Then that week the inspector visits

(a) these S banks, and

(b) a random subset of size $M - S$ of the remaining banks.

Thus every week exactly M banks are visited, and no bank goes for more than 50 weeks between visits; also for each bank there is a constant probability each week that it will be visited (except if it is 50 weeks since the last visit, in which case another visit is sure). It is easy to see that we shall never find $S > M$, since the banks in the S group are a subset of the M banks that were inspected exactly 50 weeks ago.

The number of different ways a schedule for 50 consecutive weeks can be written down (with M visits each week, and so that every bank is visited at least once) is simply $Z = \binom{N}{M}^{50} P$ where P is given in (2) above. It turns out that when Plan E is used, then for any set of 50 consecutive

weeks, each of these Z ways is equally likely. A proof of this result is given below.

Now we must consider how to start up Plan E. If we simply use Plan B (weekly random selections with replacement) for the first 49 weeks, we run the risk that more than M banks will escape visitation, so that in the 50th week we have an impossible task. Better would be Plan C, which does force every bank to be visited at least once in the first year. However, this start-up rule does not exactly achieve the stationarity condition, though the approximation appears to be quite good in the cases (small N and a "year" with only 2 or 3 weeks) that we have been able to work out in detail. To achieve a better approximation, it suffices to generate several years' worth of Plan E (with Plan C start-up) before coming to a part that will actually be used.

Although Plan E is simple to state and to implement, we have not been able to obtain tractable expressions for all of its parameters. As an approximation, however, we can use the values given in TABLE III (for $N = 2000$ banks), interpreting the columns as follows:

M = number of banks visited each week,
W = average time between visits to any bank,
T = $50M$ = total number of visits per 50-week period,
H = approximate probability that a particular bank is visited in any week, except when the previous visit was 49 weeks earlier, and
π_{50} = approximate probability that a bank goes 49 weeks without a visit.

Calculation of the expected time between visits in Plan C

We consider the ith bank, and say that it is in state $(j,0)$, where $1 \leqslant j \leqslant 50$, if at the beginning of the jth week it has not yet been drawn in one of the M-sets, and in state $(j,1)$ if it has. Let $P(j,0)$ and $P(j,1)$ denote the probabilities of being in these states, and let $\pi(j,0)$ [resp. $\pi(j,1)$] be the probability of going from state $(j,0)$ to state $(j+1,0)$ [resp. $(j+1,1)$]. These states and probabilities are related by the Markov chain shown in FIGURE 1. For $1 \leqslant j \leqslant 49$ we have (using $N = 50M$)

$$\pi(j,0) = \frac{\binom{N-jM-1}{M}}{\binom{N-jM}{M}} = \frac{50-j}{51-j},$$

$$\pi(j,1) = \frac{1}{51-j},$$

and therefore $P(j,0) = (51-j)/2500$, $P(j,1) = (j-1)/2500$, for $1 \leqslant j \leqslant 50$. The probability of this bank being drawn during the jth week in the K-set, given that it was not drawn in the M-set,

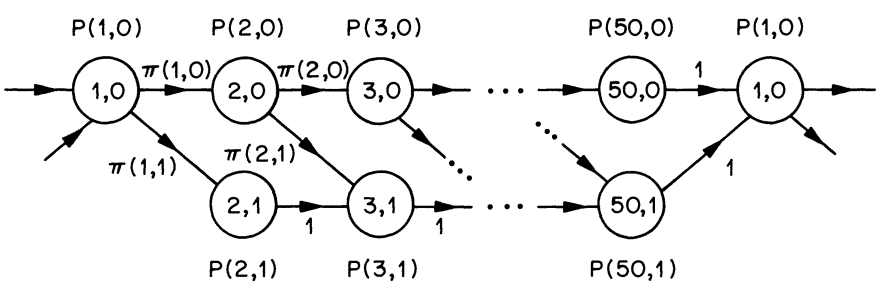


FIGURE 1. **Markov chain describing Plan C.**

is

$$\sigma = 1 - \frac{\binom{N-M-1}{K}}{\binom{N-M}{K}} = \frac{K}{49M}.$$

We wish to determine W, the average number of weeks to the next time this bank is drawn, given that it has just been drawn. There are three possibilities: (a) the bank was in state $(j,0)$ and in the jth week was drawn in the M-set, (b) the bank was in state $(j,1)$ and in the jth week was drawn in the K-set, and (c) the bank was in state $(j,0)$ and in the jth week was not drawn in the M-set but was drawn in the K-set. The expression for W contains terms corresponding to these three possibilities; we omit the details and simply state the result, which is

$$W = \frac{W_a + W_b + W_c}{D_a + D_b + D_c}, \tag{18}$$

where

$$W_a = \sum_{j=1}^{50} P(j,0)\pi(j,1)W_a',$$

$$W_a' = \sum_{r=1}^{50-j} r(1-\sigma)^{r-1}\sigma + \sum_{r=51-j}^{100-j} r(1-\sigma)^{r-1}\left\{\prod_{i=1}^{r+j-51}\pi(i,0)\right\}\{\pi(r+j-50,1)+\pi(r+j-50,0)\sigma\},$$

$$W_b = \sum_{j=1}^{50} P(j,1)\sigma W_a',$$

$$W_c = \sum_{j=1}^{50} P(j,0)\pi(j,0)\sigma W_c',$$

$$W_c' = \sum_{r=1}^{50-j} r\left\{\prod_{i=j+1}^{r+j-1}\pi(i,0)\right\}(1-\sigma)^{r-1}\{\pi(r+j,1)+\pi(r+j,0)\sigma\},$$

and

$$D_a = \sum_{j=1}^{50} P(j,0)\pi(j,1),$$

$$D_b = \sum_{j=1}^{50} P(j,1)\sigma,$$

$$D_c = \sum_{j=1}^{50} P(j,0)\pi(j,0)\sigma.$$

(An empty product is equal to 1 by convention.) Therefore

$$W_a = \sum_{j=1}^{50} \frac{49M+(j-1)K}{2500\times 49M}\cdot\left[\sum_{r=1}^{50-j} r(1-\sigma)^{r-1}\sigma + \sum_{r=51-j}^{100-j}\frac{r(1-\sigma)^{r-1}}{50}\{1+(100-r-j)\sigma\}\right], \tag{19}$$

$$W_c = \sum_{j=1}^{50}\frac{\sigma}{2500}\sum_{r=1}^{50-j} r(1-\sigma)^{r-1}\{1+(50-r-j)\sigma\}, \tag{20}$$

and the denominator is

$$D_a + D_b + D_c = \frac{M+K}{50M}. \tag{21}$$

Equations (18)–(21) were used to calculate the values shown in TABLE II.

Proof of the equidistribution property of Plan E

We first establish some notation. Let $X_1, X_2, \ldots$ be the subsets of $\{1,2,\ldots,N\}$ that indicate which banks are to be visited in weeks $1,2,\ldots$. The number of visits each week is $|X_t| = M$, for all t. We have to show that under Plan E, if

$$\left| \bigcup_{j=t-49}^{t} X_j \right| = N \tag{22}$$

so that all the banks are visited in the 50-week period $(t-49,\ldots,t)$, then

$$\text{Prob}\{X_{t-49},\ldots,X_{t-1},X_t\} = \frac{1}{Z}. \tag{23}$$

The defining property of Plan E is that the conditional probability

$$\text{Prob}\{X_t | X_{t-49},\ldots,X_{t-1}\} \tag{24}$$

is constant over all sets X_t such that (22) holds. Let the *coverage* of $\{X_{t-49},\ldots,X_{t-1}\}$ be

$$C_t = \left| \bigcup_{j=t-49}^{t-1} X_j \right|.$$

Then, at week t, $S = N - C_t$ banks are forced into X_t, leaving $M-(N-C_t)$ to be chosen randomly from the remaining $N-(N-C_t) = C_t$ banks. Thus the probability in (24) equals

$$\frac{1}{\binom{C_t}{M-N+C_t}} = \frac{1}{\binom{C_t}{N-M}}.$$

To prove that (23) is correct, it is sufficient to show that this specification satisfies the recurrence condition

$$\text{Prob}\{X_{t-49},\ldots,X_t\} = \sum_{X_{t-50}} \text{Prob}\{X_{t-50},\ldots,X_{t-1}\}\,\text{Prob}\{X_t | X_{t-49},\ldots,X_{t-1}\}, \tag{25}$$

where the sum is over all sets X_{t-50} such that

$$\left| \bigcup_{j=t-50}^{t-1} X_j \right| = N.$$

By the same argument as before, the number of such sets is just $\binom{C_t}{N-M}$, so the sum in (25) has this many terms, each equal to $1/Z\binom{C_t}{N-M}$; the sum is thus $1/Z$, and the result is established.

Finally, we note that there is no difficulty in implementing any of these schemes. In particular efficient algorithms are readily available for choosing random subsets from a larger set (see [**6**, §3.4.2], [**7**]).

Acknowledgements. The question studied in this paper was raised by Mr. Chungming An of Bell Laboratories, Indian Hill, Illinois, in connection with the auditing of telephone offices. We are grateful to him for telling us about the problem, and to J. Inglis, H. O. Pollak, and L. A. Shepp for helpful discussions.

References

[1] F. N. David and D. E. Barton, Combinatorial Chance, Hafner, New York, 1962.

[2] W. Feller, An Introduction to Probability Theory and Its Applications, 2nd ed., John Wiley, New York, 1957.
[3] M. Gardner, The Unexpected Hanging and Other Mathematical Diversions, Simon and Schuster, 1972, pp. 11–23.
[4] I. D. Hill, The economic incentive provided by sampling inspection, Appl. Stat., 9 (1960) 69–81.
[5] N. L. Johnson and S. Kotz, Urn Models and Their Application, John Wiley, New York, 1977.
[6] D. E. Knuth, The Art of Computer Programming, Volume 2: Seminumerical Algorithms, 2nd ed., Addison-Wesley, Reading, MA, 1981.
[7] A. Nijenhuis and H. S. Wilf, Combinatorial Algorithms for Computers and Calculators, 2nd ed., Academic Press, New York, 1978.
[8] J. N. K. Rao and J. E. Graham, Rotation designs for sampling on repeated occasions, J. Amer. Statist. Assoc., 59 (1964) 492–509.
[9] P. Whittle, Optimum preventative sampling, J. Oper. Res. Soc. Amer., 2 (1954) 197–203.

On the Existence of Group Automorphisms Whose Inverse Is Their Reciprocal

RICHARD E. DOWDS
ALBERT D. POLIMENI
State University College
Fredonia, NY 14063

Students of elementary mathematics frequently confuse $f^{-1}(x)$ with $1/f(x)$ when working with functions of a real variable, but students in junior and senior level mathematics courses have matured sufficiently to understand the difference between the two expressions. Since such students normally take a course in abstract algebra, we propose finding an algebraic setting in which $f^{-1}(x)$ and $(f(x))^{-1}$ actually coincide. We use elementary number theory to determine conditions on a finite group which guarantee the existence of an automorphism f such that $f^{-1}(x) = (f(x))^{-1}$. Our methods also yield the answer to the analogous problem for a finite dimensional vector space over a field. Since the existence of an automorphism of the desired type will require the group to be abelian, we shall adopt additive notation for groups. Our formula then becomes $f^{-1}(x) = -f(x)$.

Let $(G, +)$ be a finite group. A function $f\colon G \to G$ is called **naturally invertible** provided f is one-to-one (hence onto) and $f^{-1}(x) = -f(x)$ for each $x \in G$. We use the terminology **ni-function** to mean a naturally invertible function. Note that if $f\colon G \to G$ is a ni-function, then for each $x \in G$,

$$f^{-1}(f(x)) = x = -f(f(x)) = -f^2(x), \tag{1}$$

hence (replacing x by $-x$ in (1))

$$x = f^2(-x). \tag{2}$$

Immediate consequences of (1) and (2) are the properties

(i) $f^2(x) = -x$, for each $x \in G$,
(ii) $f^{-1}(x) = f(-x)$, for each $x \in G$ and
(iii) $f^4 = i_G$, the identity function on G.

If f is an automorphism of G, it is one-to-one, and $f(-x) = -f(x)$ for all $x \in G$, so f is a ni-function if and only if (i) holds. In this case, we call f a **ni-automorphism**. From (i) and the fact that f^2 is an automorphism, we conclude that a group having a ni-automorphism must be abelian.

Our first theorem characterizes cyclic groups which have a ni-automorphism.

[2] W. Feller, An Introduction to Probability Theory and Its Applications, 2nd ed., John Wiley, New York, 1957.
[3] M. Gardner, The Unexpected Hanging and Other Mathematical Diversions, Simon and Schuster, 1972, pp. 11–23.
[4] I. D. Hill, The economic incentive provided by sampling inspection, Appl. Stat., 9 (1960) 69–81.
[5] N. L. Johnson and S. Kotz, Urn Models and Their Application, John Wiley, New York, 1977.
[6] D. E. Knuth, The Art of Computer Programming, Volume 2: Seminumerical Algorithms, 2nd ed., Addison-Wesley, Reading, MA, 1981.
[7] A. Nijenhuis and H. S. Wilf, Combinatorial Algorithms for Computers and Calculators, 2nd ed., Academic Press, New York, 1978.
[8] J. N. K. Rao and J. E. Graham, Rotation designs for sampling on repeated occasions, J. Amer. Statist. Assoc., 59 (1964) 492–509.
[9] P. Whittle, Optimum preventative sampling, J. Oper. Res. Soc. Amer., 2 (1954) 197–203.

On the Existence of Group Automorphisms Whose Inverse Is Their Reciprocal

RICHARD E. DOWDS
ALBERT D. POLIMENI
State University College
Fredonia, NY 14063

Students of elementary mathematics frequently confuse $f^{-1}(x)$ with $1/f(x)$ when working with functions of a real variable, but students in junior and senior level mathematics courses have matured sufficiently to understand the difference between the two expressions. Since such students normally take a course in abstract algebra, we propose finding an algebraic setting in which $f^{-1}(x)$ and $(f(x))^{-1}$ actually coincide. We use elementary number theory to determine conditions on a finite group which guarantee the existence of an automorphism f such that $f^{-1}(x) = (f(x))^{-1}$. Our methods also yield the answer to the analogous problem for a finite dimensional vector space over a field. Since the existence of an automorphism of the desired type will require the group to be abelian, we shall adopt additive notation for groups. Our formula then becomes $f^{-1}(x) = -f(x)$.

Let $(G, +)$ be a finite group. A function $f\colon G \to G$ is called **naturally invertible** provided f is one-to-one (hence onto) and $f^{-1}(x) = -f(x)$ for each $x \in G$. We use the terminology **ni-function** to mean a naturally invertible function. Note that if $f\colon G \to G$ is a ni-function, then for each $x \in G$,

$$f^{-1}(f(x)) = x = -f(f(x)) = -f^2(x), \tag{1}$$

hence (replacing x by $-x$ in (1))

$$x = f^2(-x). \tag{2}$$

Immediate consequences of (1) and (2) are the properties

(i) $f^2(x) = -x$, for each $x \in G$,
(ii) $f^{-1}(x) = f(-x)$, for each $x \in G$ and
(iii) $f^4 = i_G$, the identity function on G.

If f is an automorphism of G, it is one-to-one, and $f(-x) = -f(x)$ for all $x \in G$, so f is a ni-function if and only if (i) holds. In this case, we call f a **ni-automorphism**. From (i) and the fact that f^2 is an automorphism, we conclude that a group having a ni-automorphism must be abelian.

Our first theorem characterizes cyclic groups which have a ni-automorphism.

THEOREM 1. *Let G be a cyclic group of order $p_1^{\alpha_1} \cdot p_2^{\alpha_2} \cdots p_r^{\alpha_r}$, where $p_1 < p_2 < \cdots < p_r$ are primes. Then G has a ni-automorphism if and only if $p_2, p_3, \ldots, p_r$ are of the form $4k+1$ and either $p_1^{\alpha_1} = 2$ or p_1 is of the form $4k+1$.*

Proof. We first examine the case when G is a cyclic group with generator a and order p^α, p a prime. If f is an automorphism of G, then $f(a) = ta$, where $(p, t) = 1$, and f is a ni-automorphism if and only if $f^2(a) = -a$. But $f^2(a) = -a$ means $t^2 a = -a$, which is equivalent to

$$t^2 + 1 \equiv 0 \bmod p^\alpha. \tag{3}$$

Hence, G has a ni-automorphism if and only if (3) has solutions. When $p = 2$, it is easily seen that (3) has solutions if and only if $\alpha = 1$. If $p > 2$, then (3) has solutions if and only if the congruence $t^2 + 1 \equiv 0 \bmod p$ has solutions (see [3], Theorem 10, p. 149). But by Euler's criterion, $t^2 + 1 \equiv 0 \bmod p$ has solutions (i.e., -1 is a quadratic residue mod p) if and only if $(-1)^{(p-1)/2} \equiv 1 \bmod p$ (see [3], Theorem 1, p. 136). This can happen if and only if p is of the form $4k+1$. We now consider the general case; suppose that G is cyclic with generator a and order $p_1^{\alpha_1} \cdot p_2^{\alpha_2} \cdots p_r^{\alpha_r}$. Then $a = a_1 + a_2 + \cdots + a_r$, where the order of a_i is $p_i^{\alpha_i}$, $1 \leqslant i \leqslant r$ (see [2], Prop. 3, p. 214). As before, f is a ni-automorphism of G if and only if $f(a) = ta$, where $t^2 a = -a$, or equivalently, where $t^2 a_i = -a_i$ for $1 \leqslant i \leqslant r$. Hence we arrive at the system of congruences $t^2 \equiv -1 \bmod p_i^{\alpha_i}$, $1 \leqslant i \leqslant r$, which, by the Chinese Remainder Theorem, has a solution if and only if $p_2, \ldots, p_r$ are primes of the form $4k+1$ and either $p_1^{\alpha_1} = 2$ or p_1 is a prime of the form $4k+1$. This completes the proof.

If p is a prime, then a group G is called an **elementary abelian p-group** (briefly, an ***E(p)*-group**) if it is a direct sum of groups of order p; such a group G may be viewed as a vector space over the field Z_p of p elements. If G is an elementary 2-group of order 2^n, then condition (i) becomes $f^2(x) = -x = x$, or $f^2 = i_G$. Thus for this case, $f = i_G$ is a ni-automorphism of G. There exist many other ni-automorphisms of G. To see this, simply view G as a vector space over Z_2; then obtaining an automorphism f for which $f^2 = i_G$ is equivalent to finding an $n \times n$ matrix (entries in Z_2) whose square is the identity matrix. For example, the matrix with 1's along the "off" diagonal and 0's elsewhere will do. The characterization of $E(p)$-groups which have ni-automorphisms for $p > 2$ is settled by the next theorem.

THEOREM 2. *Let G be an $E(p)$-group of order p^n, where p is an odd prime. Then G has a ni-automorphism if and only if n is even or p is a prime of the form $4k+1$.*

Proof. We view G as an n-dimensional vector space over Z_p and first prove the sufficiency of the stated condition. If $n = 2r$, then the automorphism given by the $n \times n$ matrix

$$A_1 = \begin{bmatrix} B & & 0 \\ & B & \\ & \ddots & \\ 0 & & B \end{bmatrix},$$

where B is the 2×2 matrix

$$B = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix},$$

is a ni-automorphism of G. If p is a prime of the form $4k+1$ and $\pm\alpha$ are solutions to the congruence $t^2 + 1 \equiv 0 \bmod p$, consider an $n \times n$ diagonal matrix of the form diag $(\alpha_1, \alpha_2, \ldots, \alpha_n)$ where $\alpha_i = \alpha$ or $-\alpha, 1 \leqslant i \leqslant n$. There are 2^n such matrices and each of them will determine a ni-automorphism of G.

To show the condition is necessary, suppose that $f\colon G \to G$ is a ni-automorphism of G. Then $f^2(a) = -a$ for each $a \in G$, hence f satisfies the polynomial $t^2 + 1$ over Z_p. If p is not of the form $4k+1$, then $t^2 + 1$ is irreducible over Z_p and $t^2 + 1$ is the minimal polynomial of f. Thus (see [1], p. 307) there is a basis for G in which the matrix of f is of the same form as A_1 above, hence n must be even. This concludes the proof.

The proof of Theorem 2 makes clear the condition for the existence of a ni-automorphism of a finite dimensional vector space over Z_p, since the only property of Z_p necessary for the proof is whether or not t^2+1 has a root in Z_p. Thus we have the following result.

THEOREM 3. *Let F be a field and let V be an n-dimensional vector space over F. Then V has a ni-automorphism if and only if n is even or there is an element a in F satisfying $a^2=-1$.*

If V is an arbitrary (perhaps infinite dimensional) vector space over a field F in which $a^2=-1$ for some a in F, then the linear map f: $V\to V$ defined by $f(v)=av$, for v in V satisfies $f^2(v)=-v$, so f is a ni-automorphism of V.

Finally, we consider the existence of a ni-automorphism for an arbitrary finite abelian group. Recall that every finite abelian group is a direct sum of cyclic p-groups, that is, if A is an abelian group of order $p_1^{\alpha_1}\cdot p_2^{\alpha_2}\cdots p_r^{\alpha_r}$, where $p_1<p_2<\cdots<p_r$ are primes, then

$$A=\langle a_{11}\rangle\oplus\cdots\oplus\langle a_{1m_1}\rangle\oplus\cdots\oplus\langle a_{r1}\rangle\oplus\cdots\oplus\langle a_{rm_r}\rangle,$$

where $|a_{ij}|$ (the order of a_{ij}) is a power of p_i for $i=1,2,\ldots,r$, and $j=1,\ldots,m_i$. If A_i denotes the p_i-Sylow subgroup of A, $i=1,2,\ldots,r$, then

$$A=A_1\oplus A_2\oplus\cdots\oplus A_r$$

and the automorphism group of A, denoted by $\mathrm{Aut}(A)$, is given by

$$\mathrm{Aut}(A)=\mathrm{Aut}(A_1)\oplus\cdots\oplus\mathrm{Aut}(A_r).$$

Thus A will have a ni-automorphism if and only if each A_i has a ni-automorphism, $i=1,2,\ldots,r$. Using techniques similar to those in the proofs of Theorems 1 and 2, it can be shown that A will have a ni-automorphism provided conditions (I) and (II) below are satisfied.

(I) Either:

(i) $|a_{1j}|=2$ for $j=1,2,\ldots,m_1$,
or (ii) p_1 is a prime of the form $4k+1$,
or (iii) m_1 is even and $|a_{1j}|=p_1^{\beta_1}$ for some β_1 and $j=1,2,\ldots,m_1$,

and

(II) For $2\leqslant i\leqslant r$, either:

(i) p_i is a prime of the form $4k+1$,
or (ii) m_i is even and $|a_{ij}|=p_i^{\beta_i}$ for some β_i and $j=1,2,\ldots,m_i$.

The above conditions are sufficient for the existence of a ni-automorphism, but they are not necessary. For example, consider $A=\langle a_{11}\rangle\oplus\langle a_{12}\rangle$, where $|a_{11}|=2$, $|a_{12}|=4$. Denote a_{11} by a, a_{12} by b and define

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

Then

$$f^2(a)=f(a+2b)=f(a)+2f(b)=a=-a$$

and

$$f^2(b)=f(a+b)=f(a)+f(b)=3b=-b.$$

It follows that $f^2(x)=-x$ for every x in A and f is a ni-automorphism. On the other hand, conditions (I) and (II) are not both satisfied.

The authors wish to acknowledge the many helpful suggestions made by the editors and the referees.

References

[1] I. N. Herstein, Topics in Algebra, 2nd ed., Xerox, 1975.
[2] O. F. G. Schilling and W. S. Piper, Basic Abstract Algebra, Allyn and Bacon, 1975.
[3] J. E. Shockley, Introduction to Number Theory, Holt, Rinehart and Winston, 1967.

A Note on Cavalieri Integration

M. A. Malik
Concordia University
Montreal, Canada

In 1635, Bonaventura Cavalieri (1598–1647) published his great treatise "Geometria Indivisibilibus," which is regarded an important milepost in the history of calculus. In this work, he introduced the idea of an "indivisible" of a given planar piece as a chord of that piece and regarded the planar piece as made up of an infinite set of such parallel indivisibles [**4**]. The concept that a geometric solid was the aggregate of parallel cross-sections had been used by Archimedes when he considered a sphere to be composed of circular plane discs in his study on ratio between volumes of sphere and cylinder. However, Cavalieri exploited the idea that two planar pieces, each made up of the same infinite set of parallel indivisibles, have the same area. Using this principle, he established some important results, a prelude to the calculus, that are analogous to the well-known formula

$$\int_0^a x^n\,dx = \frac{a^{n+1}}{n+1}.$$

To describe Cavalieri's method and his results, let $ABCD$ be a parallelogram with diagonal BD (see FIGURE 1). Let c be any chord of the parallelogram which is parallel to AB and let x be the part of c which is a chord of triangle ABD. Then, considering c as an indivisible of the parallelogram and x as an indivisible of the triangle, Cavalieri claimed that

$$\sum c = 2\sum x,$$

where Σ represents the sum being taken over all such chords with endpoints on DA. (This collection of chords c and of chords x are regarded as describing the parallelogram $ABCD$ and the triangle ADB, respectively, and Σc and Σx give their respective areas.) In fact, FIGURE 1 shows that $\Sigma c = \Sigma(x+y) = \Sigma x + \Sigma y = 2\Sigma x$ since for each chord x in triangle ADB there is a chord $y = x$ in triangle CBD. Cavalieri further asserted that the sum of the squares of chords in the parallelogram is three times the sum of the squares of chords in triangle ADB, that is,

$$\sum c^2 = 3\sum x^2. \tag{1}$$

Cavalieri's proof of his claim is given in confusingly verbose geometric terminology [**1**; p. 86]. Translating to algebraic notation, we can follow Cavalieri's method of proof of (1), which begins by noting that

$$\sum c^2 = \sum (x+y)^2 = \sum x^2 + \sum y^2 + 2\sum xy,$$

and since (arguing as before) $\Sigma x^2 = \Sigma y^2$,

$$\sum c^2 = 4\sum x^2 - 4\sum \alpha^2, \tag{2}$$

where $\alpha = \left|\frac{x-y}{2}\right|$. To establish (1) from (2), it is necessary to show that

$$\sum x^2 - 4\sum \alpha^2 = 0. \tag{3}$$

To prove (3), Boyer [**1**; p. 87], following Cavalieri, claims that "in each of the triangles generated by α there are half as many indivisibles as there are in that generated by x. Furthermore the indivisibles of the former are half as long as the corresponding ones in the latter." A different argument is offered by Edwards [**3**; p. 108], who considered the problem for a square instead of a parallelogram and thought of Σx^2 as the volume of a pyramid (see also Bosman [**1**; p. 373]).

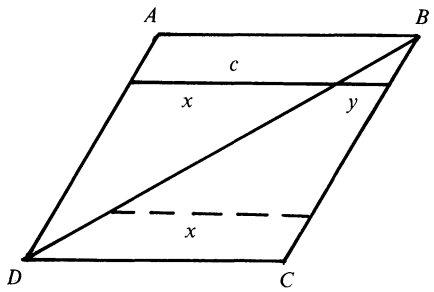


FIGURE 1

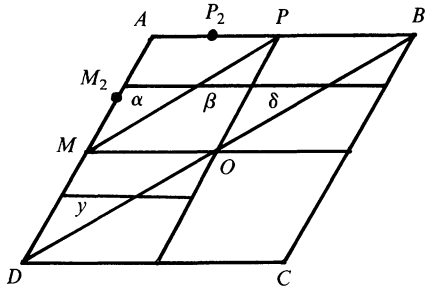


FIGURE 2

We offer a different method to establish (3) which uses Cavalieri's technique of representing sums of indivisibles as sums of other indivisibles, but which results in an equation which requires (in the spirit of Cavalieri's time) a "descent" argument or (in the modern spirit) a limit, to complete the proof. Let O be the center of parallelogram $ABCD$, and let M and P be the midpoints of AD and AB respectively (see FIGURE 2). Consider the trapezoid $AMOB$ as composed of three triangles: AMP, OPM and POB. Let x be an arbitrary chord in $AMOB$ which is parallel to AB and let α, β, δ represent the parts of x in these triangles as shown in FIGURE 2. Note there that α when restricted to triangle AMP is the same as in (2).

Denote Σ_{MA} as the sum being taken over MA. It is easy to see that

$$\sum_{MA} x^2 = \sum_{MA} (\alpha + \beta + \delta)^2 = 5 \sum_{MA} \alpha^2 + 4 \sum_{MA} \alpha\beta \tag{4}$$

because $\Sigma_{MA}\alpha^2 = \Sigma_{MA}\beta^2 = \Sigma_{MA}\delta^2$ and $\alpha = \delta$. For $\xi = \left|\frac{\alpha - \beta}{2}\right|$ we have

$$4 \sum_{MA} \alpha\beta = 4 \sum_{MA} \alpha^2 - 8 \sum_{MA} \xi^2. \tag{5}$$

Hence (4) together with (5) becomes

$$\sum_{MA} x^2 = 9 \sum_{MA} \alpha^2 - 8 \sum_{MA} \xi^2. \tag{6}$$

Now, denote Σ_{DM} as the sum being taken over DM and note that for each chord y parallel to MO in the triangle DMO there is a chord α parallel to AP in the triangle AMP with $\alpha = y$; also $DM = MA$. Thus

$$\sum_{DM} y^2 = \sum_{MA} \alpha^2, \qquad \sum \alpha^2 = 2 \sum_{MA} \alpha^2 \quad \text{and} \quad \sum x^2 = \sum_{DM} y^2 + \sum_{MA} x^2.$$

With this observation, from (6) one has

$$\begin{aligned} \sum x^2 - 4 \sum \alpha^2 &= \sum_{DM} y^2 + \sum_{MA} x^2 - 8 \sum_{MA} \alpha^2 \\ &= \sum_{MA} \alpha^2 + 9 \sum_{MA} \alpha^2 - 8 \sum_{MA} \xi^2 - 8 \sum_{MA} \alpha^2 \\ &= 2\left(\sum_{MA} \alpha^2 - 4 \sum_{MA} \xi^2 \right). \end{aligned} \tag{7}$$

The left-hand side of (7) is the left-hand side of (3), which is to be shown to be zero. We now have it equal to twice a "similar" quantity in the parallelogram $APOM$ whose sides are half the length of the sides of the original parallelogram $ADCB$.

Now let M_2 and P_2 be the midpoints of AM and AP respectively, and repeat the above reasoning with reference to the parallelogram of sides AM_2 and AP_2. From (7) this gives

$$\sum x^2 - 4 \sum \alpha^2 = 2^2 \left(\sum_{M_2A} \xi^2 - 4 \sum_{M_2A} \chi^2 \right) \tag{8}$$

where Σ_{M_2A} represents the sum being taken over M_2A and $\chi = \left|\frac{\xi - \eta}{2}\right|$, $\eta = \frac{c}{2^2} - \xi$. Repeating the process for each $n = 1, 2, 3, \ldots$, we conclude from (8) that

$$\begin{aligned} 0 \le \left|\sum x^2 - 4\sum \alpha^2\right| &= 2^n \left| \sum_{M_nA} \xi^2 - 4 \sum_{M_nA} \chi^2 \right| \\ &\le 2^n \sum_{M_nA} \xi^2 = 2^n \sum_{M_nA} \left(\frac{c}{2^n}\right)^2 \\ &\le \frac{\sum c^2}{2^n}. \end{aligned} \tag{9}$$

The claim in (3) is verified by making n large enough (i.e., $n \to \infty$).

References

[1] M. H. Bosman, S. J., Un chapitre de l'oeuvre de Cavalieri, Mathesis, 36 (1922) 336–373.
[2] C. Boyer, Cavalieri, Limit and Discarded Infinitesimals, Scripta Mathematica, 8 (1941) 79–91.
[3] C. H. Edwards Jr., The Historical Development of the Calculus, Springer-Verlag, New York, 1979.
[4] H. Eves, Slicing it Thin, The Mathematical Gardner, ed. David A. Klarner, Wadsworth International, Belmont, California, 1981, pp. 100–111, and Great Moments in Mathematics (before 1650), MAA Dolciani Mathematical Expositions no. 5, 1983, pp. 206–214.

Periodic Equilibria Under Periodic Harvesting

A. C. Lazer
University of Cincinnati
Cincinnati, OH 45221

D. A. Sánchez
University of New Mexico
Albuquerque, NM 87131

A simple deterministic model for the growth of one population which is subject to periodic or seasonal harvesting due to fishing, hunting, or disease is given by the ordinary differential equation

$$\dot{x} = g(x) - h(t). \tag{1}$$

Here $x = x(t)$ is the size of the population at time t, $g(x)$ is a smooth function usually of the form $xf(x)$, where $f(x)$ is the per capita rate of growth, and $h(t)$ is a T-periodic function representing the harvesting effect. The case where $h(t)$ is a constant was discussed by the second author in an earlier note [**6**] for the case of a finite difference model, and by F. Brauer and the second author in [**1**] for the differential equation model.

An interesting question is the following: *what is the maximum number of T-periodic solutions equation (1) can have?* If $g(x)$ is a polynomial of degree 2 or 3 then the answer is 2 and 3, respectively. This is a not so well-known result which can be found in [**5**, pp. 102–119], and which is periodically rediscovered. For the more general case of a differential equation of the polynomial form

$$\dot{x} = x^n + a_{n-1}(t)x^{n-1} + \cdots + a_1(t)x + a_0(t), \; n \geq 4,$$

where Σ_{M_2A} represents the sum being taken over M_2A and $\chi = \left|\frac{\xi - \eta}{2}\right|$, $\eta = \frac{c}{2^2} - \xi$. Repeating the process for each $n = 1, 2, 3, \ldots$, we conclude from (8) that

$$\begin{aligned} 0 \le \left|\sum x^2 - 4\sum \alpha^2\right| &= 2^n \left| \sum_{M_nA} \xi^2 - 4 \sum_{M_nA} \chi^2 \right| \\ &\le 2^n \sum_{M_nA} \xi^2 = 2^n \sum_{M_nA} \left(\frac{c}{2^n}\right)^2 \\ &\le \frac{\sum c^2}{2^n}. \end{aligned} \tag{9}$$

The claim in (3) is verified by making n large enough (i.e., $n \to \infty$).

References

[1] M. H. Bosman, S. J., Un chapitre de l'oeuvre de Cavalieri, Mathesis, 36 (1922) 336–373.
[2] C. Boyer, Cavalieri, Limit and Discarded Infinitesimals, Scripta Mathematica, 8 (1941) 79–91.
[3] C. H. Edwards Jr., The Historical Development of the Calculus, Springer-Verlag, New York, 1979.
[4] H. Eves, Slicing it Thin, The Mathematical Gardner, ed. David A. Klarner, Wadsworth International, Belmont, California, 1981, pp. 100–111, and Great Moments in Mathematics (before 1650), MAA Dolciani Mathematical Expositions no. 5, 1983, pp. 206–214.

Periodic Equilibria Under Periodic Harvesting

A. C. Lazer
University of Cincinnati
Cincinnati, OH 45221

D. A. Sánchez
University of New Mexico
Albuquerque, NM 87131

A simple deterministic model for the growth of one population which is subject to periodic or seasonal harvesting due to fishing, hunting, or disease is given by the ordinary differential equation

$$\dot{x} = g(x) - h(t). \tag{1}$$

Here $x = x(t)$ is the size of the population at time t, $g(x)$ is a smooth function usually of the form $xf(x)$, where $f(x)$ is the per capita rate of growth, and $h(t)$ is a T-periodic function representing the harvesting effect. The case where $h(t)$ is a constant was discussed by the second author in an earlier note [**6**] for the case of a finite difference model, and by F. Brauer and the second author in [**1**] for the differential equation model.

An interesting question is the following: *what is the maximum number of T-periodic solutions equation (1) can have?* If $g(x)$ is a polynomial of degree 2 or 3 then the answer is 2 and 3, respectively. This is a not so well-known result which can be found in [**5**, pp. 102–119], and which is periodically rediscovered. For the more general case of a differential equation of the polynomial form

$$\dot{x} = x^n + a_{n-1}(t)x^{n-1} + \cdots + a_1(t)x + a_0(t), \; n \geq 4,$$

where $a_j(t)$, $j = 0, \ldots, n-1$, are smooth T-periodic functions, the determination of an upper bound on the number of T-periodic solutions is an open question. For a recent discussion of this problem see the articles by A. L. Neto [4] and S. Shashahani [8].

Since the positive periodic solutions of (1) represent stable or unstable periodic equilibria sustained by the population under harvesting, an upper bound on the total number of periodic solutions would be useful. Fortunately, such a bound can often be obtained using the equation of variation to study the Poincaré map of (1).

Let $x(t, a)$ be the solution of (1) satisfying the initial conditions $x(0, a) = a$, and let I be the set of values of a for which the solution $x(t, a)$ exists on $0 \leqslant t \leqslant T$. Then I is an open interval, and by uniqueness of solutions, $x(t, a)$ is strictly increasing on I as a function of a for fixed $t \in [0, T]$. Since

$$\frac{d}{dt}x(t, a) = g(x(t, a)) - h(t),$$

it follows that if $\phi(t, a) = \dfrac{\partial}{\partial a}x(t, a)$ then $\phi(t, a)$ satisfies the variational equation

$$\frac{d}{dt}\phi(t, a) = g'(x(t, a))\phi(t, a), \quad \phi(0, a) = 1,$$

where $g' = dg/dx$. Therefore $\phi(t, a)$ satisfies

$$\phi(t, a) = \exp\left\{\int_0^t g'(x(s, a))\, ds\right\}.$$

The Poincaré map associated with (1) is the map $a \to x(T, a)$ and we define

$$H(a) = x(T, a) - a, \quad a \in I.$$

Therefore, a zero of $H(a)$ corresponds to a periodic solution of (1) and conversely, if multiplicities are taken into account. From the previous analysis it follows that

$$H'(a) = \phi(T, a) - 1 = \exp\left\{\int_0^T g'(x(s, a))\, ds\right\} - 1,$$

so if $n - 1$ is an upper bound on the number of zeros of $H'(a)$ then (1) will have at most n periodic solutions. This leads to the following criterion:

> *If $g''(x)$ is either strictly negative or strictly positive for all x, then (1) will have at most two periodic solutions.*

For, if the above holds, then $g'(x)$ is either strictly increasing or strictly decreasing, from which it follows that $H'(a)$ will be also since $x(s, a)$ is strictly increasing in a. Therefore $H(a)$ can have at most two zeros.

A straightforward stability analysis also shows that in the case where a_1 and a_2 are zeros of $H(a)$ with $a_1 < a_2$ then the corresponding periodic solutions will be respectively, stable and unstable if $g'(x) > 0$, whereas the stability will be reversed if $g'(x) < 0$. In any case, once it is determined that periodic solutions may exist, a further analysis of the direction field of (1) or numerical analysis will help locate the periodic solutions, if any.

EXAMPLE 1. As a simple first example we consider the logistic equation with periodic harvesting

$$\dot{x} = rx\left(1 - \frac{x}{K}\right) - (1 + \varepsilon \cos t), \quad r, K > 0, 0 < \varepsilon < 1. \tag{2}$$

Here $g(x) = rx\left(1 - \dfrac{x}{K}\right)$, and $g''(x) = -2r/K < 0$, so there are at most two 2π-periodic solutions. Denoting the right side of equation (2) by $f(x, t)$ we see that $f(0, t) < 0$, $f(K, t) < 0$, and if $rK/4 > 1 + \varepsilon$ then $f(K/2, t) > 0$. Consequently, there is a (stable) periodic solution with initial value $K/2 < a < K$, and an unstable one with initial value $0 < a < K/2$.

A discussion of the above equation with no harvesting where $K = K(t)$ was also periodic (corresponding to a fluctuating environment) was given by B. D. Coleman, Y. Hsieh, and G. P. Knowles in [2]. See also [7] for a more simplified analysis including harvesting.

EXAMPLE 2. The following model for populations of the North American spruce budworm was given by D. Ludwig, D. D. Jones, and C. S. Holling [3]:

$$\dot{x} = rx\left(1 - \frac{x}{K}\right) - \frac{\beta x^2}{\alpha^2 + x^2}, \quad r, K, \alpha, \beta > 0. \tag{3}$$

The second term on the right side of equation (3) models predation by birds and in the absence of predation, the growth is assumed to be logistic. If additional periodic harvesting (say, due to seasonal spraying) were to occur, then the equation would be of the form (1).

In this case

$$g''(x) = -2\left[\frac{r}{K} + \beta\alpha^2 \frac{\alpha^2 - 3x^2}{(\alpha^2 + x^2)^3}\right].$$

For $x \geqslant 0$, $g''(x)$ will be negative if $r/K - \beta/4\alpha^2 > 0$, and there will be at most two periodic solutions. For appropriate values of the constants there will be a stable equilibrium point x_0 satisfying $K/2 < x_0 < K$ when there is no periodic harvesting—this can be easily seen by graphing the two expressions comprising the right hand side of the differential equation. Under small amplitude periodic harvesting, the equilibrium point will become a periodic solution.

References

[1] F. Brauer and D. A. Sánchez, Constant rate population harvesting: equilibrium and stability, Theoret. Population Biol., 8 (1975) 12–30.

[2] B. D. Coleman, Y. Hsieh, and G. P. Knowles, On the optimal choice of r for a population in a periodic environment, Math. Biosci., 46 (1979) 71–85.

[3] D. Ludwig, D. D. Jones, and C. S. Holling, Qualitative analysis of insect outbreak systems: the spruce budworm and forest, J. Animal Ecol., 47 (1978) 315–332.

[4] A. L. Neto, On the number of solutions of $dx/dt = \Sigma_{j=0}^{\infty} a_j(t)x^j, 0 \leqslant t \leqslant 1$, for which $x(0) = x(1)$, Invent. Math., 59 (1980) 69–76.

[5] V. A. Pliss, Nonlocal Problems of the Theory of Oscillations, Academic Press, New York, 1966.

[6] D. A. Sánchez, Populations and harvesting, SIAM Rev., 19 (1977) 551–553.

[7] _____, Periodic environments, harvesting and a Riccati equation, in "Nonlinear Phenomena in Mathematical Sciences," ed. V. Lakshmikantham, Academic Press, New York, 1982, pp. 883–886.

[8] S. Shashahani, Periodic solutions of polynomial first order differential equations, Nonlinear Anal. 5, (1981) 157–165.

The Maximum Brightness of Venus

DENNIS WILDFOGEL
Stockton State College
Pomona, NJ 08240

Even a casual observer may notice that the planet Venus, at times the dominant object in the evening sky, appears noticeably brighter at some times than at others. While reading a book on popular astronomy [4] one day, I came across the statement (with little explanation) that Venus is at its brightest when the illuminated portion of the apparent disk of the planet is 28% of the whole. Could that fact, I wondered, be determined theoretically? If you reflect for a moment, as I did, on the relative positions of the Earth, Venus and the Sun at various times, you should be able to see that, apart from the observer's local conditions, the brightness of Venus depends primarily

A discussion of the above equation with no harvesting where $K = K(t)$ was also periodic (corresponding to a fluctuating environment) was given by B. D. Coleman, Y. Hsieh, and G. P. Knowles in [2]. See also [7] for a more simplified analysis including harvesting.

EXAMPLE 2. The following model for populations of the North American spruce budworm was given by D. Ludwig, D. D. Jones, and C. S. Holling [3]:

$$\dot{x} = rx\left(1 - \frac{x}{K}\right) - \frac{\beta x^2}{\alpha^2 + x^2}, \quad r, K, \alpha, \beta > 0. \tag{3}$$

The second term on the right side of equation (3) models predation by birds and in the absence of predation, the growth is assumed to be logistic. If additional periodic harvesting (say, due to seasonal spraying) were to occur, then the equation would be of the form (1).

In this case

$$g''(x) = -2\left[\frac{r}{K} + \beta\alpha^2 \frac{\alpha^2 - 3x^2}{(\alpha^2 + x^2)^3}\right].$$

For $x \geqslant 0$, $g''(x)$ will be negative if $r/K - \beta/4\alpha^2 > 0$, and there will be at most two periodic solutions. For appropriate values of the constants there will be a stable equilibrium point x_0 satisfying $K/2 < x_0 < K$ when there is no periodic harvesting—this can be easily seen by graphing the two expressions comprising the right hand side of the differential equation. Under small amplitude periodic harvesting, the equilibrium point will become a periodic solution.

References

[1] F. Brauer and D. A. Sánchez, Constant rate population harvesting: equilibrium and stability, Theoret. Population Biol., 8 (1975) 12–30.

[2] B. D. Coleman, Y. Hsieh, and G. P. Knowles, On the optimal choice of r for a population in a periodic environment, Math. Biosci., 46 (1979) 71–85.

[3] D. Ludwig, D. D. Jones, and C. S. Holling, Qualitative analysis of insect outbreak systems: the spruce budworm and forest, J. Animal Ecol., 47 (1978) 315–332.

[4] A. L. Neto, On the number of solutions of $dx/dt = \Sigma_{j=0}^{\infty} a_j(t)x^j, 0 \leqslant t \leqslant 1$, for which $x(0) = x(1)$, Invent. Math., 59 (1980) 69–76.

[5] V. A. Pliss, Nonlocal Problems of the Theory of Oscillations, Academic Press, New York, 1966.

[6] D. A. Sánchez, Populations and harvesting, SIAM Rev., 19 (1977) 551–553.

[7] _____, Periodic environments, harvesting and a Riccati equation, in "Nonlinear Phenomena in Mathematical Sciences," ed. V. Lakshmikantham, Academic Press, New York, 1982, pp. 883–886.

[8] S. Shashahani, Periodic solutions of polynomial first order differential equations, Nonlinear Anal. 5, (1981) 157–165.

The Maximum Brightness of Venus

DENNIS WILDFOGEL
Stockton State College
Pomona, NJ 08240

Even a casual observer may notice that the planet Venus, at times the dominant object in the evening sky, appears noticeably brighter at some times than at others. While reading a book on popular astronomy [4] one day, I came across the statement (with little explanation) that Venus is at its brightest when the illuminated portion of the apparent disk of the planet is 28% of the whole. Could that fact, I wondered, be determined theoretically? If you reflect for a moment, as I did, on the relative positions of the Earth, Venus and the Sun at various times, you should be able to see that, apart from the observer's local conditions, the brightness of Venus depends primarily

on two factors: the apparent size of its disk as seen from Earth, and the fraction of that disk which is illuminated by the Sun's light. Furthermore, those two factors work in opposition: when Venus is closest to Earth, so that its apparent size is greatest, it is then between the Sun and the Earth, so that its illuminated hemisphere is turned away from us; when Venus is fully illuminated from our vantage point, it is on the far side of the Sun and its apparent size is at a minimum (see FIGURE 1). "Aha!" I thought to myself, "a perfect first-year Calculus maximum/minimum problem!" Indeed, I subsequently used the problem as an end-of-the-term group modeling project, as described in [**11**].

Before solving this problem, let me state it in astronomical terminology. Consider the Earthward-facing hemisphere of a planet or moon. As seen from the Earth, that hemisphere appears as a disk. The fraction of the area of that disk which is illuminated by the Sun's light at a given moment is called the **phase**. The **elongation** of a planet viewed from Earth is the size of the angle Sun-Earth-planet. In books and magazines on popular astronomy (e.g., [**9**]), tables giving phases and elongations are common. Thus our problem may be stated as follows:

> *Find the phase and elongation of Venus at the moment it reaches maximum brightness.*

I will begin by assuming that Venus and Earth move in circular orbits with the Sun at the center of each circle. Of course a planet actually moves in what is essentially an ellipse with the Sun at one focus. But the eccentricity of the Earth's orbit is only .017, and that of Venus' is .007. Thus, for example, the difference between the semi-major and semi-minor axes of the Earth's orbit is only about 13,000 miles out of some 93,000,000—an excellent approximation to a circle! The Sun is about one and a half million miles from the center of this near circle, but that is still less than 2% of the semi-major axis.

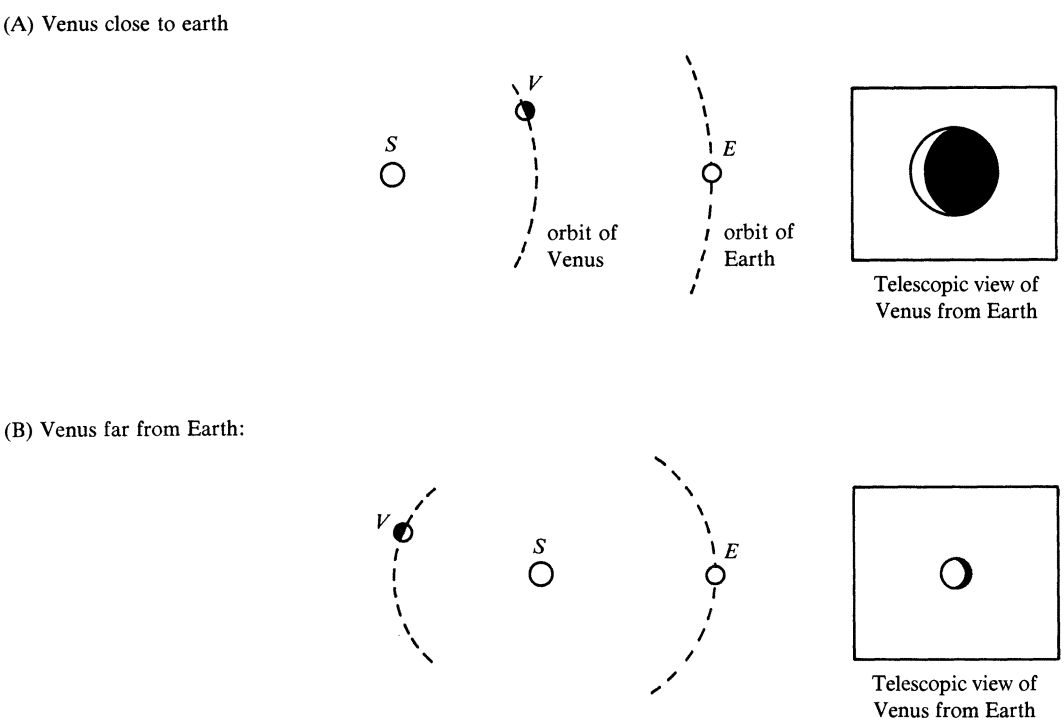


FIGURE 1. **Highly schematic views of the relative positions of Earth (E), Venus (V), and the Sun (S) as seen from above the plane of the Earth's orbit. Representations of the appearance of Venus as viewed from Earth in a telescope appear at the right.**

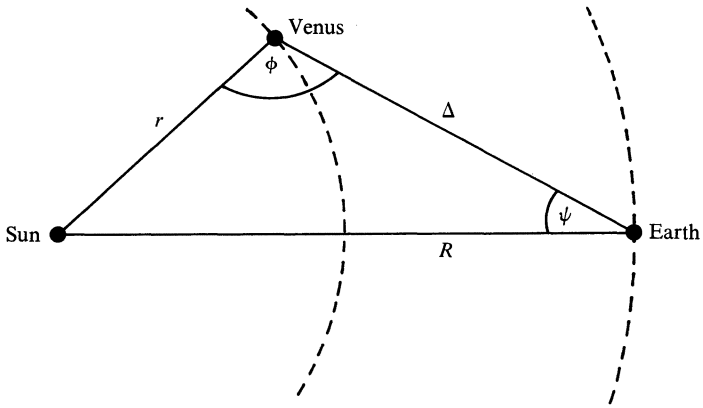


FIGURE 2. **Highly schematic view from above the plane of the solar system.**

I will also assume that the orbits of the planets are coplanar. The orbital plane of Venus is actually inclined by 3° to that of the Earth. One consequence of this is that since the angular size of the Sun as seen from the Earth is $\frac{1}{2}°$, Venus rarely passes directly in front of (a **transit**) or behind (an **occultation**) the Sun. For our purposes, however, the very slight inclination makes extremely little difference in the distance between Earth and Venus. For example, should Venus happen to be at its maximum height above the Earth's orbital plane at the same moment that it is at its closest approach, then the difference between its actual distance and its distance calculated as if it were in the same plane is only about 250,000 miles out of some 24 million, i.e., about 1%. Thus we may safely assume that Venus and Earth orbit in the same plane.

Let R denote the (fixed, by assumption) distance from the Earth to the Sun, r the distance from Venus to the Sun, and Δ the distance from the Earth to Venus (see FIGURE 2). Since the apparent diameter of Venus is essentially inversely proportional to Δ, the apparent area of Venus, and hence its brightness (or luminosity), is inversely proportional to Δ^2. Thus if L denotes the luminosity of Venus and p that planet's phase, then the remarks in the first paragraph may be expressed as

$$L = k\frac{p}{\Delta^2},$$

where k is a constant of proportionality.

If we could now obtain Δ as a function of p, we could then differentiate L with respect to p and obtain the desired result, namely, the value of p which maximizes L. However, Δ turns out to be a rather ugly function of p, so instead we will obtain p as a function of Δ and first maximize L with respect to Δ.

The curve which marks the boundary of the lit and unlit portions of Venus (or of any planet or moon) is called the **terminator**. The terminator is always a great circle. (By the way, Venus is virtually a perfect sphere, as are Mercury and the Moon.) As viewed from Earth, though, the terminator is seen projected onto a plane perpendicular to the line from Earth to Venus and thus appears as half of an ellipse. In FIGURE 3, this projection is represented by the arc NAS. The angle Sun-Venus-Earth, which is denoted by ϕ in FIGURE 2, is the angle FCC' in FIGURE 4. Then the angle $A'CA$ in FIGURE 4 is $\pi - \phi$ and so $AC = \rho\cos(\pi - \phi)$, where ρ is the radius of Venus.

Now the area of the half-ellipse $NASCN$ (FIGURE 3) is

$$\frac{1}{2}\pi\rho(AC) = \frac{1}{2}\pi\rho^2\cos(\pi - \phi),$$

so the area of the crescent $NDSAN$ is

$$\frac{1}{2}\pi\rho^2 - \frac{1}{2}\pi\rho^2\cos(\pi - \phi) = \frac{1}{2}\pi\rho^2(1 + \cos\phi).$$

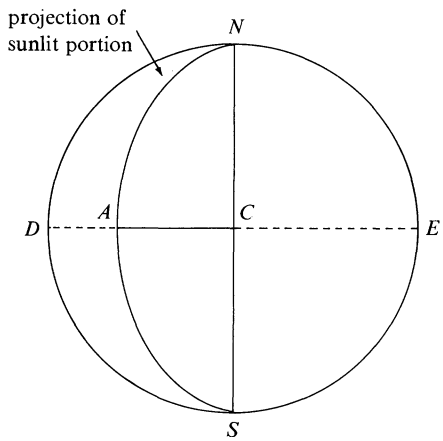


FIGURE 3. **"View" from Earth. Cross-section of Venus outlined by the great circle perpendicular to the line from Earth to Venus. The arc *NAS* is the projection of the terminator on this cross-section. The points *A* and *C* are in the interior of Venus and not actually visible from Earth. The points *D*, *A*, *C*, and *E* appear again in FIGURE 4.**

Thus the phase of Venus is given by

$$p = \frac{\frac{1}{2}\pi\rho^2(1+\cos\phi)}{\pi\rho^2} = \frac{1}{2}(1+\cos\phi).$$

I have derived the expression for p for a "crescent" Venus ($\phi > \pi/2$). The reader should check that the expression is still valid for a "gibbous" Venus ($\phi < \pi/2$) even though the derivation has to be modified slightly.

To get p as a function of Δ, apply the law of cosines to the triangle Sun-Venus-Earth (see FIGURE 2) to get

$$R^2 = \Delta^2 + r^2 - 2r\Delta\cos\phi,$$

so

$$\cos\phi = \frac{\Delta^2 + r^2 - R^2}{2r\Delta}$$

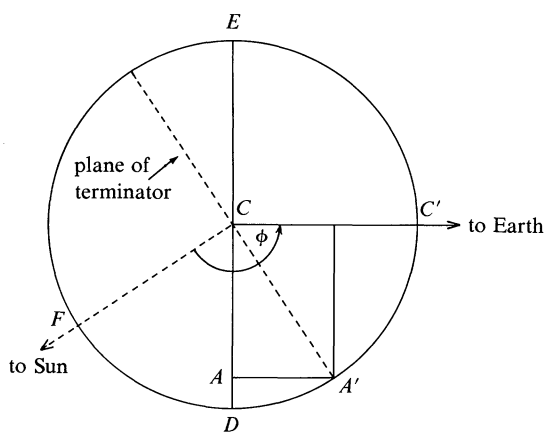


FIGURE 4. **"View" from above. Cross-section of Venus outlined by the great circle parallel to the plane of the Earth's orbit. The line *DCE* (which is interior to the planet and not actually visible to an observer) is the intersection of the plane of FIGURE 3 with this cross-section.**

and

$$p = \frac{1}{2}(1 + \cos\phi) = \frac{2r\Delta + \Delta^2 + r^2 - R^2}{4r\Delta}.$$

We can now express the luminosity of Venus as a function of Δ as follows:

$$L = \frac{kp}{\Delta^2} = K\frac{2r\Delta + \Delta^2 + r^2 - R^2}{\Delta^3},$$

where K is the constant $k/4r$. Then

$$\frac{dL}{d\Delta} = -\frac{K}{\Delta^4}(4r\Delta + \Delta^2 + 3(r^2 - R^2))$$

and so $dL/d\Delta = 0$ if and only if $\Delta^2 + 4r\Delta + 3(r^2 - R^2) = 0$. The roots of the quadratic are

$$\Delta = -2r \pm \sqrt{r^2 + 3R^2}.$$

The negative value for Δ is clearly physically unacceptable. Thus the value $\Delta_M = -2r + \sqrt{r^2 + 3R^2}$ is the only candidate for a relative extreme for L as a function of Δ. It is not hard to see (by, for example, factoring the quadratic) that the sign of $dL/d\Delta$ is opposite to that of the factor $\Delta - \sqrt{r^2 + 3R^2} + 2r$. Thus L is an increasing function for $\Delta < \Delta_M$ and decreasing for $\Delta > \Delta_M$, and so Δ_M is indeed a relative maximum.

Note that the only physically acceptable values for Δ are those in the interval $[|R - r|, R + r]$. Does Δ_M fall in this interval? It is a straightforward exercise in inequalities to show that if $R > r$, then $\Delta_M > R - r$, but that $\Delta_M < R + r$ if and only if $r/R > 1/4$. If $r/R > 1$ (i.e., $r > R$), then $\Delta_M < 0$. For the case of observing Venus from Earth, r/R does satisfy $1/4 < r/R < 1$, so Δ_M is in $[R - r, R + r]$.

We should also like the model developed here to be applicable to other planets, e.g., Venus viewed from Jupiter, Mars viewed from Earth, etc. In general, then, R would represent the distance from the Sun to the observer's planet and r the distance from the Sun to the planet under observation. In some cases, e.g., observing Venus from Jupiter, $r/R < 1/4$. In such a case $\Delta_M > R + r$, so L increases throughout the interval $[R - r, R + r]$. The extreme values of L then occur, of course, at the endpoints of the interval. On the other hand, observing Mars from Earth gives $r > R$. In this case $\Delta_M < 0 < r - R$, so L decreases steadily throughout the interval $[r - R, r + R]$.

I will summarize below the outcome of our model, making use of some further astronomical terminology. A planet is called **inferior** (**superior**) if it is closer to (further from) the Sun than the observer's planet. When a superior planet is, from the observer's perspective, 180° away from the Sun it is said to be at **opposition** (see FIGURE 5). When it is again lined up with the Sun and the observer but beyond the Sun, it is in **conjunction**. An inferior planet will also line up with the Sun and the observer in two different configurations. When it is on the near side of the Sun it is in **inferior conjunction**; when it is beyond the Sun it is in **superior conjunction**.

We may then summarize our results as follows:

If $R > r$ and $r/R > 1/4$, then the maximum brightness of the observed planet occurs at $\Delta_M = -2r + \sqrt{r^2 + 3R^2}$. Minimum values of brightness occur at the endpoints of the interval $[R - r, R + r]$, i.e., at inferior and superior conjunction.

If $r/R < 1/4$, then the brightness of the observed planet increases steadily from a minimum at $\Delta = R - r$ (inferior conjunction) to a maximum at $\Delta = R + r$ (superior conjunction).

If $r > R$, then the brightness of the observed planet decreases steadily from a maximum at $\Delta = r - R$ (opposition) to a minimum at $\Delta = r + R$ (conjunction).

The summary statements above refer to the time interval from inferior to superior conjunction for an inferior planet or from opposition to conjunction for a superior planet. The variation in

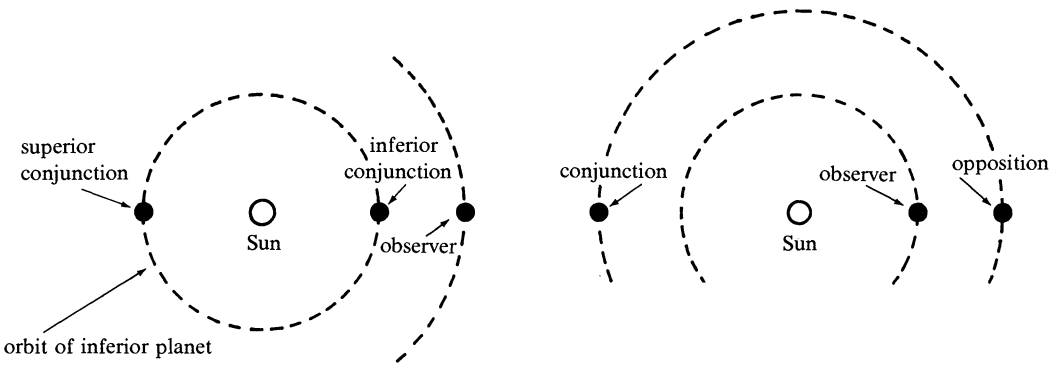


FIGURE 5. **At left, an inferior planet is shown in its two possible alignments (*inferior* and *superior conjunction*) with the line from the observer through the sun. The corresponding alignments (*opposition* and *conjunction*) for a superior planet are shown at right.**

brightness is then of course reversed as the planets shift until they are once again in the original configuration.

How well does this model correspond to observation? For the case of observing Venus ($r = 6.72 \times 10^7$ miles) from Earth ($R = 9.29 \times 10^7$ miles), certainly $1/4 < r/R < 1$, so the calculated value $\Delta_M = 4.00 \times 10^7$ is a relative maximum for L. This value for Δ_M yields $p = 26.6\%$, which is in very good agreement with the figure of 28% quoted at the beginning of this article.

The elongation of Venus (ψ in FIGURE 2) can be calculated from Δ by again using the law of cosines:

$$\cos\psi = \frac{\Delta^2 + R^2 - r^2}{2\Delta R}.$$

For $\Delta = \Delta_M = 4.00 \times 10^7$, the angle ψ is 39.7°. Also, by using the law of sines, ψ can be calculated directly from the phase p as

$$\psi = \arcsin\left(\frac{r}{R}\sin(\arccos(2p-1))\right). \tag{1}$$

To test this model further, I have used data from the *American Ephemeris* [**5**, pp. 4 and 369] for the phase of Venus on some selected dates to compute ψ according to equation (1). TABLE 1 shows the excellent agreement between these computed values and the actual values listed in the *Ephemeris*.

Note that phase and elongation are purely geometric concepts, i.e., they have nothing to do with brightness. The excellent agreement shown in TABLE 1 between the actual and the calculated elongation suggests that even the small difference between our calculated value of 26.6% for the phase of Venus at maximum brightness and the quoted value of 28% is *not* accounted for by our

1977 Date	Phase	ψ actual	ψ calculated from (1)
Jan. 17	.550	47	46.03
Feb. 1	.471	47	46.23
Feb. 16	.376	45	44.49
Feb. 26	.301	42	41.57
Mar. 3	.259	39	39.33
Mar. 8	.214	36	36.39
Mar. 18	.121	28	28.15
Apr. 2	.016	10	10.46

TABLE 1

simplifying geometric assumptions. Have we made some other assumption about brightness that could account for the difference? Yes, in fact, we made an unstated assumption about the *uniformity* of the brightness of the disk of Venus.

When we claim that the luminosity L is proportional to the phase p, we are assuming that if a particular portion of Venus is fully lit during two different phases then it appears equally as bright at each of those times. This is, unfortunately, not the case. Actually, over-all brightness varies non-linearly with the phase (even assuming that distance remains fixed) in a complicated way that depends markedly on the reflective properties of the planet or moon under observation. (By the way, making this non-linear relation explicit for the case of an ideal sphere might be a good project for students in a multivariable calculus class.) For example, when the Earth's moon appears half lit (at First or Last Quarter) it is not half as bright as when it is full but rather only one-eleventh as bright! [**8**, p. 36]. However, the non-linearity of the effect is substantially mitigated on Venus because it is the thick cloud cover and not that planet's surface which reflects sunlight. Thus when the phase is 26.6%, Venus is not quite as bright as our model indicates, but the actual maximum brightness, which must be determined empirically, occurs when the phase is only very slightly greater.

Interestingly, astronomers *define* [**6**, p. 209] the **brilliancy** of Venus as the quantity

$$\frac{2r\Delta + \Delta^2 + r^2 - R^2}{r^3\Delta^3}.$$

This is, of course, essentially what I used for the brightness L. Even though this is a purely theoretical quantity, tables of celestial phenomena actually list the moment of greatest brilliancy, not of greatest brightness (see [**7**, p. 29], [**9**, p. 39]).

I have shown that our model fits the data very well for the case of observing Venus from Earth, two planets with low orbital eccentricity. But what about Mercury, with an eccentricity of .206, or even Mars with an eccentricity of .098? (The figure .098 may seem low, but even before the advent of the telescope Kepler was able to deduce that planets move in ellipses rather than circles by studying the orbit of Mars.) The expressions derived here for phase, elongation and brilliancy will still be accurate if, instead of using values for r and R representing the *mean* distance of those planets from the Sun, values are used which more closely approximate the true values for those planets in the configurations under consideration. Using simple analytic geometry and some knowledge of astronomical coordinate systems, it is not too hard (and it would be a good project for students) to make such approximations for r and R (see for example [**1**]; for a shortcut, see [**3**, p. 297ff]). Even without such refinements, the model still gives a good qualitative account of the variation in brightness of the planet under observation (except that in the case of Mercury, the non-linear variation of brightness with respect to phase has a significant effect).

It did not surprise me that astronomers have long known about the simple model I had worked out on my own to find out whether the phase of Venus at maximum brightness could be calculated theoretically. What did surprise me was that this simple model is actually in very good agreement with observation. I would imagine there are other astronomical phenomena that are amenable to analysis with the level of mathematics used in this note. Indeed, elementary calculus is used at several junctures in texts on positional astronomy [**2**], [**10**].

Once, astronomy and mathematics were practiced by the same people. One sees little evidence of this in today's calculus texts. The instructor who is willing to look at the astronomical literature, or better yet, work out some models for him or herself, most likely will be rewarded with interesting material for use in introductory calculus courses.

References

[1] P. Duffet-Smith, Practical Astronomy with Your Calculator, Cambridge Univ. Press, 1979.
[2] D. McNally, Positional Astronomy, Wiley, 1974.
[3] D. H. Menzel, Field Guide to the Stars and Planets, Houghton-Mifflin, 1964.
[4] J. Muirden, Astronomy with Binoculars, T. Y. Crowell, 1979.

[5] Nautical Almanac Office of the U.S.A., American Ephemeris and Nautical Almanac for the year 1977, U.S. Government Printing Office, 1975.
[6] Nautical Almanac Offices of the U.K. and the U.S.A., Explanatory Supplement to the Astronomical Ephemeris and the American Ephemeris and Nautical Almanac, Her Majesty's Stationery Office, 1961.
[7] G. Ottewell, The Astronomical Calendar for the year 1981, Author, 1980.
[8] _____, The Astronomical Companion, Author, 1979.
[9] J. R. Pearcy, ed., The Observer's Handbook 1980, Royal Astronomical Society of Canada, 1979.
[10] W. Smart, Textbook on Spherical Astronomy, Cambridge Univ. Press, 1931, 4th ed., 1956.
[11] D. Wildfogel, A mock symposium for your calculus class, Amer. Math. Monthly, 90 (1983) 52–53.

The Circumdisk and its Relation to a Theorem of Kirszbraun and Valentine

RALPH ALEXANDER
University of Illinois
Urbana, IL 61801

In this note we consider the following problem: *Given a finite set of points in euclidean m-space, characterize the radius R of the smallest disk (closed solid sphere) which contains those points.* We believe that our solution to this problem is new, characterizing R in terms of a well-known quadratic form. Moreover, it provides a new proof of the very appealing theorem of Kirszbraun-Valentine: *If a collection of disks (of varying radii) in E^m having nonempty intersection are rearranged so that corresponding distances between centers do not increase, then the rearranged collection also has nonempty intersection.* Whether or not the volume of the intersection can decrease remains a problem which baffles mathematicians.

We first need to review some of the basic ideas that will be required. Let $\mathbf{p}_0, \mathbf{p}_1, \ldots, \mathbf{p}_n$, $n \geqslant 1$, be a collection of distinct points in euclidean space E^m. Among all those disks (closed solid spheres) which contain these points, there is a unique disk of smallest radius, called the **circumdisk**. In FIGURE 1 we give a simple but very useful example. Here $m = 2$, $n = 2$, and the unbroken circle with diameter $|\mathbf{p}_1 - \mathbf{p}_2|$ bounds the circumdisk.

The existence of a minimal containing disk follows from the Blaschke selection theorem [**12**, p. 37]. However, those readers familiar with sequential compactness in E^m can easily concoct a proof of existence. Such a disk is unique, for if there were two distinct minimal disks, centered at $\mathbf{u}$ and $\mathbf{u}'$, respectively, and having radius R, then a disk of radius $\sqrt{R^2 - \frac{1}{4}|\mathbf{u} - \mathbf{u}'|^2}$ centered at $\frac{1}{2}(\mathbf{u} + \mathbf{u}')$ would also contain all the points $\mathbf{p}_i$.

Some writers use the word circumsphere instead of circumdisk. However, we wish to reserve the former as a term for a generalization of the notion **circumcircle**. Thus, if $\mathbf{p}_0, \mathbf{p}_1, \ldots, \mathbf{p}_n$ are points in E^m, we define the **circumsphere** to be the sphere of least radius on which all the points lie, provided such a sphere exists. For example, three collinear points have no circumsphere. In FIGURE 1, observe that $\mathbf{p}_0, \mathbf{p}_1, \mathbf{p}_2$ lie on a circumsphere (the circle determined by the dashed arc)

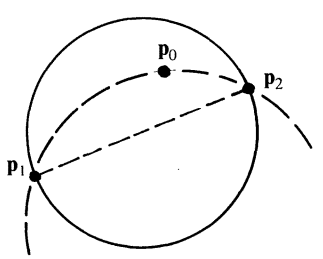


FIGURE 1

[5] Nautical Almanac Office of the U.S.A., American Ephemeris and Nautical Almanac for the year 1977, U.S. Government Printing Office, 1975.
[6] Nautical Almanac Offices of the U.K. and the U.S.A., Explanatory Supplement to the Astronomical Ephemeris and the American Ephemeris and Nautical Almanac, Her Majesty's Stationery Office, 1961.
[7] G. Ottewell, The Astronomical Calendar for the year 1981, Author, 1980.
[8] _____, The Astronomical Companion, Author, 1979.
[9] J. R. Pearcy, ed., The Observer's Handbook 1980, Royal Astronomical Society of Canada, 1979.
[10] W. Smart, Textbook on Spherical Astronomy, Cambridge Univ. Press, 1931, 4th ed., 1956.
[11] D. Wildfogel, A mock symposium for your calculus class, Amer. Math. Monthly, 90 (1983) 52–53.

The Circumdisk and its Relation to a Theorem of Kirszbraun and Valentine

RALPH ALEXANDER
University of Illinois
Urbana, IL 61801

In this note we consider the following problem: *Given a finite set of points in euclidean m-space, characterize the radius R of the smallest disk (closed solid sphere) which contains those points.* We believe that our solution to this problem is new, characterizing R in terms of a well-known quadratic form. Moreover, it provides a new proof of the very appealing theorem of Kirszbraun-Valentine: *If a collection of disks (of varying radii) in E^m having nonempty intersection are rearranged so that corresponding distances between centers do not increase, then the rearranged collection also has nonempty intersection.* Whether or not the volume of the intersection can decrease remains a problem which baffles mathematicians.

We first need to review some of the basic ideas that will be required. Let $\mathbf{p}_0, \mathbf{p}_1, \ldots, \mathbf{p}_n$, $n \geqslant 1$, be a collection of distinct points in euclidean space E^m. Among all those disks (closed solid spheres) which contain these points, there is a unique disk of smallest radius, called the **circumdisk**. In FIGURE 1 we give a simple but very useful example. Here $m = 2$, $n = 2$, and the unbroken circle with diameter $|\mathbf{p}_1 - \mathbf{p}_2|$ bounds the circumdisk.

The existence of a minimal containing disk follows from the Blaschke selection theorem [**12**, p. 37]. However, those readers familiar with sequential compactness in E^m can easily concoct a proof of existence. Such a disk is unique, for if there were two distinct minimal disks, centered at $\mathbf{u}$ and $\mathbf{u}'$, respectively, and having radius R, then a disk of radius $\sqrt{R^2 - \frac{1}{4}|\mathbf{u} - \mathbf{u}'|^2}$ centered at $\frac{1}{2}(\mathbf{u} + \mathbf{u}')$ would also contain all the points $\mathbf{p}_i$.

Some writers use the word circumsphere instead of circumdisk. However, we wish to reserve the former as a term for a generalization of the notion **circumcircle**. Thus, if $\mathbf{p}_0, \mathbf{p}_1, \ldots, \mathbf{p}_n$ are points in E^m, we define the **circumsphere** to be the sphere of least radius on which all the points lie, provided such a sphere exists. For example, three collinear points have no circumsphere. In FIGURE 1, observe that $\mathbf{p}_0, \mathbf{p}_1, \mathbf{p}_2$ lie on a circumsphere (the circle determined by the dashed arc)

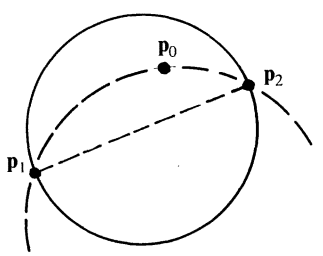


FIGURE 1

which is not the boundary of the circumdisk. If a circumsphere exists, its uniqueness is established in the same manner as for the circumdisk.

Recall that a set is **convex** if for any two points in the set, the line segment joining them also lies in the set. While our claims concerning convexity should be clear at least for the plane ($m = 2$), the excellent references [2] and [12] may be consulted for the general theory of convex sets in E^m. The **convex hull** of the points $\mathbf{p}_0, \mathbf{p}_1, \ldots, \mathbf{p}_n$ is the intersection of all convex sets containing them. It is easily checked that the convex hull consists of all points $\mathbf{p} = x_0\mathbf{p}_0 + \cdots + x_n\mathbf{p}_n$, where the x_i are nonnegative real numbers and $x_0 + x_1 + \cdots + x_n = 1$. It is clear that the circumdisk contains the convex hull, and that at least two of the points $\mathbf{p}_0, \mathbf{p}_1, \ldots, \mathbf{p}_n$ lie on the boundary of the circumdisk (FIGURE 1 illustrates this). Those points which lie on this boundary will be called ***c*-extreme points.**

Our first theorem is standard, but a discussion is included for completeness and for future reference. Lemma 1, which actually gives a metric characterization of the convex hull, has various applications in geometric extremal problems. It says that if the point $\mathbf{u}$ lies in the convex hull of a finite set of points $\{\mathbf{q}_i\}$, then given any other point $\mathbf{u}'$, some one of the $\mathbf{q}_i$ is closer to $\mathbf{u}$ than to $\mathbf{u}'$.

LEMMA 1. *Let the point* $\mathbf{u}$ *lie in the convex hull of the points* $\mathbf{q}_0, \mathbf{q}_1, \ldots, \mathbf{q}_s$. *If* $\mathbf{u}'$ *is distinct from* $\mathbf{u}$, *then for some* i, $|\mathbf{u} - \mathbf{q}_i| < |\mathbf{u}' - \mathbf{q}_i|$.

To prove this, choose H to be the $(m-1)$ flat (or hyperplane) through $\mathbf{u}$ which is perpendicular to the segment $\overline{\mathbf{u}\mathbf{u}'}$. Then for at least one value of i, $\mathbf{q}_i$ must lie in the closed halfspace of H which does not contain $\mathbf{u}'$; this choice of i works.

THEOREM 1. *Let* $\mathbf{p}_0, \mathbf{p}_1, \ldots, \mathbf{p}_n$ *be points in* E^m. *Then*

(a) *the center of the circumdisk is contained in the convex hull of the c-extreme points,*
(b) *the circumdisk of the c-extreme points is the circumdisk of the set of points* $\mathbf{p}_i$, *and*
(c) *the boundary of this disk is the circumsphere of the c-extreme points.*

All of the statements in Theorem 1 are illustrated in FIGURE 1.

Proof. (a). Let $\mathbf{u}$ be the center of the circumdisk which has radius R, and suppose that $\mathbf{u}$ does not lie in the convex hull K of the c-extreme points. Let $\delta_1 > 0$ be the distance from $\mathbf{u}$ to K, and let $\delta_2 > 0$ be the least of all distances from $\mathbf{u}$ to those $\mathbf{p}_i$ which are not c-extreme points. Since K is convex and closed, there will be exactly one point $\mathbf{u}_1$ in K such that $|\mathbf{u} - \mathbf{u}_1| = \delta_1$. Choose the point $\mathbf{u}_2$ on the segment $\overline{\mathbf{u}\mathbf{u}_1}$ so that $|\mathbf{u} - \mathbf{u}_2| = \delta$, where $\delta = \frac{1}{2}\min(\delta_1, \delta_2)$. A contradiction is obtained by observing that all points $\mathbf{p}_i$ are contained in a disk centered at $\mathbf{u}_2$ of radius $\sqrt{R^2 - \delta^2}$.

(b) and (c). Lemma 1, applied to the center $\mathbf{u}$, may now be used to deduce that no disk, centered at $\mathbf{u}'$, of smaller radius than R can contain the c-extreme points. Lemma 1 also shows that the c-extreme points can lie on no smaller sphere than the boundary of the circumdisk.

Theorem 1 can be made sharper; a theorem of Carathéodory can be invoked to show that the center of the circumdisk is contained in the convex hull of $m+1$ or fewer c-extreme points. This strengthened version is not needed for the present work, but a discussion of Carathéodory's theorem may be found in Eggleston's book [2].

We next introduce the quadratic form

$$Q = \sum_{i,j=0}^{n} |\mathbf{p}_i - \mathbf{p}_j|^2 x_i x_j. \tag{1}$$

For many purposes it is convenient to rewrite Q as

$$Q = \sum_{i,j=0}^{n} \left(|\mathbf{p}_i|^2 + |\mathbf{p}_j|^2\right) x_i x_j - 2\left|\sum_{i=0}^{n} x_i \mathbf{p}_i\right|^2. \tag{2}$$

The transition from (1) to (2) is accomplished by using the identity

$$|\mathbf{p}_i - \mathbf{p}_j|^2 = |\mathbf{p}_i|^2 + |\mathbf{p}_j|^2 - 2\mathbf{p}_i \cdot \mathbf{p}_j$$

and the fact that the dot product is linear. It is clear from (1) that the value of Q at $(x_0, x_1, \ldots, x_n)$ is independent of the choice of coordinate system in E^m.

LEMMA 2. *Suppose* $\mathbf{p}_0, \mathbf{p}_1, \ldots, \mathbf{p}_n$ *lie in a disk of radius* R, *and the nonnegative numbers* x_i *satisfy* $x_0 + x_1 + \cdots + x_n = 1$. *Then the quadratic form* Q *in* (1) *satisfies*

$$\text{Max } Q \leqslant 2R^2. \tag{3}$$

To prove this, choose the center of the disk as the origin and note from (2) that

$$Q \leqslant \sum \left(|\mathbf{p}_i|^2 + |\mathbf{p}_j|^2\right) x_i x_j \leqslant 2R^2 \left(\sum x_i\right)^2 = 2R^2.$$

LEMMA 3. *Suppose the points* $\mathbf{q}_0, \mathbf{q}_1, \ldots, \mathbf{q}_s$ *possess a circumsphere of radius* R. *Then subject to* $x_0 + x_1 + \cdots + x_s = 1$, *the quadratic form* Q *in* (1) *satisfies*

$$\text{Max } Q = 2R^2.$$

Moreover, a nonnegative vector ($x_i \geqslant 0$ *for each* i) *which maximizes* Q *exists when the center of the circumsphere lies in the convex hull of the* $\mathbf{q}_i$.

Proof. If the center of the circumsphere is the origin, then (2) implies

$$Q = 2R^2 - 2\left|\sum_i x_i \mathbf{q}_i\right|^2. \tag{4}$$

The center of the circumsphere must lie in the flat spanned by the $\mathbf{q}_i$, or else a reflection through this flat would contradict the uniqueness of the circumsphere. Thus, there is a suitable vector $(x_0, x_1, \ldots, x_s)$ such that $\Sigma x_i \mathbf{q}_i = 0$. If the center lies in the convex hull of the $\mathbf{q}_i$, then the x_i may be chosen nonnegative.

Using these results, we can now easily characterize the radius of the circumdisk of a pointset in terms of Q.

THEOREM 2. *Suppose that the points* $\mathbf{p}_0, \mathbf{p}_1, \ldots, \mathbf{p}_n$ *in* E^m *possess a circumdisk of radius* R. *Then subject to* $x_i \geqslant 0$ *for each* i, *and* $x_0 + x_1 + \cdots + x_n = 1$, *the quadratic form* Q *in* (1) *satisfies*

$$\text{Max } Q = 2R^2.$$

Lemma 2 says that Max $Q \leqslant 2R^2$. However, if we set $x_i = 0$ whenever $\mathbf{p}_i$ is not c-extreme, then Theorem 1 and Lemma 3, applied to the c-extreme points, imply that there actually is a suitable vector $(x_0, x_1, \ldots, x_n)$ at which $Q = 2R^2$.

COROLLARY. *Let* $\mathbf{p}_0, \mathbf{p}_1, \ldots, \mathbf{p}_n$ *and* $\mathbf{p}'_0, \mathbf{p}'_1, \ldots, \mathbf{p}'_n$ *be points in* E^m *such that*

$$|\mathbf{p}'_i - \mathbf{p}'_j| \leqslant |\mathbf{p}_i - \mathbf{p}_j| \text{ for all } i, j. \tag{5}$$

Then the circumdisk of the $\mathbf{p}'_i$ is no larger than that of the $\mathbf{p}_i$.

Proof. Let $y_0, y_1, \ldots, y_n$, $y_i \geqslant 0$, $y_0 + y_1 + \cdots + y_n = 1$, be chosen to maximize the form Q' associated with the points $\mathbf{p}'_i$. Then

$$2R'^2 = Q'(y_0, y_1, \ldots, y_n) \leqslant Q(y_0, y_1, \ldots, y_n) \leqslant 2R^2.$$

It may be seen that, subject to (5), $R = R'$ if and only if there is a subset of the c-extreme points among the $\mathbf{p}_i$, containing the center of the circumdisk in its convex hull, which is congruent to a corresponding subset of the $\mathbf{p}'_i$.

There are several well-known results due to Kirszbraun [5] and Valentine [11] on intersections of disks. The following theorem (or its equivalent), described in our introduction, is a consequence

of the Corollary.

THEOREM 3. *Suppose the disks $D_0, D_1, \ldots, D_n$ in E^m with respective radii $R_0, R_1, \ldots, R_n$ and centers $\mathbf{p}_0, \mathbf{p}_1, \ldots, \mathbf{p}_n$ have nonempty intersection. If the disks are rearranged to be centered at $\mathbf{p}'_0, \mathbf{p}'_1, \ldots, \mathbf{p}'_n$ in such a manner that $|\mathbf{p}'_i - \mathbf{p}'_j| \leqslant |\mathbf{p}_i - \mathbf{p}_j|$ for all i, j, then the intersection of the rearranged disks remains nonempty.*

Proof. If all the disks have the same radius R, the theorem follows immediately from the Corollary, since a collection of disks of radius R has nonempty intersection if and only if the circumdisk of their centers has radius not exceeding R. (Incidentally, this shows that the Corollary itself follows at once from Theorem 3.) We now utilize a simple "Chinese checkers" geometric construction to deduce the theorem from this special case.

Choose R greater than the maximum of the R_i. Place $2n$ disks of radius R in E^{m+1}, centered at the $2n$ points $(\mathbf{p}_i, \pm\sqrt{R^2 - R_i^2})$. Note that the intersection of these disks of radius R is a convex set which is symmetric with respect to the hyperplane E^m (or $x_{m+1} = 0$), and which is nonempty since its intersection with E^m is the intersection of the original disks in E^m of varying radii. (See FIGURE 2.) Likewise place $2n$ disks of radius R centered at the $2n$ points $(\mathbf{p}'_i, \pm\sqrt{R^2 - R_i^2})$. Observe that the two sets of centers of disks of radius R satisfy the distance inequality, thus the rearranged disks have nonempty intersection. Since this intersection is a convex set which is symmetric with respect to the hyperplane E^m, it follows that the disks in E^m centered at the $\mathbf{p}'_i$ also have nonempty intersection.

Some versions of Theorem 3 state (at least implicitly) that if a collection of disks has nonempty intersection, then the intersection contains a point in the convex hull of the centers. Theorem 1 guarantees this to be true when all radii are equal. The geometric construction described above quickly extends the validity to the situation of varying radii.

A number of proofs and generalizations of the theorems of Kirszbraun and Valentine have appeared; of these, the note of Schoenberg [**10**] is probably the most closely related to the present work. The articles of Mickle [**7**] and Grünbaum [**3**], [**4**] should be consulted for elegant extensions of the Kirszbraun-Valentine theory. A very pretty generalization of spheres having nonempty intersection, due to G. Minty, is also described in [**4**].

The relation of the form Q to the theory of metric geometry was investigated by Schoenberg in a series of very interesting papers; [**9**] provided an early elegant example. More recently the author [**1**] has been attempting to find practical methods of employing the form Q in the solution of geometric problems in E^m. Problem 7 in the article by Klee [**6**] in this *Magazine* provided impetus for the present note.

We close with some questions. Suppose $D_0, D_1, \ldots, D_n$ are unit disks in the plane. If the centers are moved closer together, can the perimeter of the intersection decrease? Consideration of the evolute of the boundary of the intersection leads to further problems of interest. Are there generalizations of Helly's theorem which might be helpful? The theorem of Sallee [**8**] could be a step in the right direction.

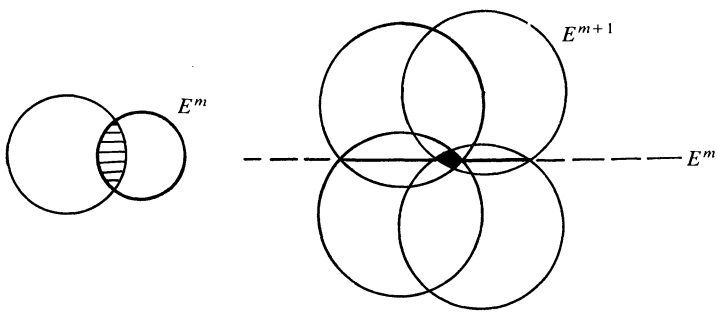


FIGURE 2

We wish to thank the referee and Branko Grünbaum for many helpful suggestions about the exposition.

References

[1] R. Alexander, Metric averaging in euclidean and Hilbert spaces, Pacific J. Math, 85 (1979) 1–9.
[2] H. G. Eggleston, Convexity, Cambridge University Press, Cambridge, 1958.
[3] B. Grünbaum, On a theorem of Kirszbraun, Bull. Res. Counc. of Israel, 7F (1958) 129–132.
[4] _____, A generalization of theorems of Kirszbraun and Minty, Proc. Amer. Math. Soc., 13 (1962) 812–814.
[5] M. Kirszbraun, Uber die Zusammenziehenden und Lipschitzschen Transformationen, Fund. Math., 22 (1934) 77–108.
[6] V. Klee, Some unsolved problems in plane geometry, this MAGAZINE, 52 (1979) 131–145.
[7] E. J. Mickle, On the extension of a transformation, Bull. Amer. Math. Soc., 55 (1949) 160–164.
[8] G. T. Sallee, A Helly-type theorem for widths, The Geometry of Metric and Linear Spaces, ed. L. M. Kelly, Springer-Verlag, 1975, pp. 227–232.
[9] I. J. Schoenberg, Metric spaces and positive definite functions, Trans. Amer. Math. Soc., 44 (1938) 522–536.
[10] _____, On a theorem of Kirszbraun and Valentine, Amer. Math. Monthly, 60 (1953) 620–622.
[11] F. A. Valentine, A Lipschitz condition preserving extension for a vector function, Amer. J. Math., 67 (1945) 83–93.
[12] _____, Convex Sets, McGraw-Hill, New York, 1964.

Is Every Continuous Function Uniformly Continuous?

RAY F. SNIPES
Bowling Green State University
Bowling Green, OH 43403

In any elementary treatment of uniform continuity, examples are given to show that the answer to our title question is *no*. The most often cited functions are probably

$$f\colon (0,1] \to R$$
$$x \to 1/x$$

and

$$g\colon R \to R$$
$$x \to x^2.$$

To show that these functions are continuous but not uniformly continuous it is assumed that the metric is the usual Euclidean or standard one, i.e., the distance between two real numbers x and y is $d_s(x, y) = |x - y|$. In this note we are concerned with finding different metrics on R so that not only are continuous functions (like the above) also uniformly continuous, but so that the new metrics are *equivalent* to d_s.

First, we recall the definitions of continuity and uniform continuity. Let (X, d_1) and (Y, d_2) be metric spaces (see [**10**]). A function $f\colon D \to Y$, where $D \subseteq X$, is (d_1, d_2)-**continuous at a point** a in D if: for each $\varepsilon > 0$, there exists a $\delta_a > 0$ such that

$$x \in D \quad \text{and} \quad d_1(x, a) < \delta_a \text{ imply } d_2(f(x), f(a)) < \varepsilon.$$

The function f is (d_1, d_2)-**continuous** if it is (d_1, d_2)-continuous at every point of its domain D. In contrast, a function $f\colon D \to Y$ is (d_1, d_2)-**uniformly continuous** if: for each $\varepsilon > 0$, there exists a $\delta > 0$ such that

$$x, y \in D \quad \text{and} \quad d_1(x, y) < \delta \text{ imply } d_2(f(x), f(y)) < \varepsilon.$$

We wish to thank the referee and Branko Grünbaum for many helpful suggestions about the exposition.

References

[1] R. Alexander, Metric averaging in euclidean and Hilbert spaces, Pacific J. Math, 85 (1979) 1–9.
[2] H. G. Eggleston, Convexity, Cambridge University Press, Cambridge, 1958.
[3] B. Grünbaum, On a theorem of Kirszbraun, Bull. Res. Counc. of Israel, 7F (1958) 129–132.
[4] _____, A generalization of theorems of Kirszbraun and Minty, Proc. Amer. Math. Soc., 13 (1962) 812–814.
[5] M. Kirszbraun, Uber die Zusammenziehenden und Lipschitzschen Transformationen, Fund. Math., 22 (1934) 77–108.
[6] V. Klee, Some unsolved problems in plane geometry, this MAGAZINE, 52 (1979) 131–145.
[7] E. J. Mickle, On the extension of a transformation, Bull. Amer. Math. Soc., 55 (1949) 160–164.
[8] G. T. Sallee, A Helly-type theorem for widths, The Geometry of Metric and Linear Spaces, ed. L. M. Kelly, Springer-Verlag, 1975, pp. 227–232.
[9] I. J. Schoenberg, Metric spaces and positive definite functions, Trans. Amer. Math. Soc., 44 (1938) 522–536.
[10] _____, On a theorem of Kirszbraun and Valentine, Amer. Math. Monthly, 60 (1953) 620–622.
[11] F. A. Valentine, A Lipschitz condition preserving extension for a vector function, Amer. J. Math., 67 (1945) 83–93.
[12] _____, Convex Sets, McGraw-Hill, New York, 1964.

Is Every Continuous Function Uniformly Continuous?

RAY F. SNIPES
Bowling Green State University
Bowling Green, OH 43403

In any elementary treatment of uniform continuity, examples are given to show that the answer to our title question is *no*. The most often cited functions are probably

$$f\colon (0,1] \to R$$
$$x \to 1/x$$

and

$$g\colon R \to R$$
$$x \to x^2.$$

To show that these functions are continuous but not uniformly continuous it is assumed that the metric is the usual Euclidean or standard one, i.e., the distance between two real numbers x and y is $d_s(x, y) = |x - y|$. In this note we are concerned with finding different metrics on R so that not only are continuous functions (like the above) also uniformly continuous, but so that the new metrics are *equivalent* to d_s.

First, we recall the definitions of continuity and uniform continuity. Let (X, d_1) and (Y, d_2) be metric spaces (see [**10**]). A function $f\colon D \to Y$, where $D \subseteq X$, is (d_1, d_2)-**continuous at a point** a in D if: for each $\varepsilon > 0$, there exists a $\delta_a > 0$ such that

$$x \in D \quad \text{and} \quad d_1(x, a) < \delta_a \text{ imply } d_2(f(x), f(a)) < \varepsilon.$$

The function f is (d_1, d_2)-**continuous** if it is (d_1, d_2)-continuous at every point of its domain D. In contrast, a function $f\colon D \to Y$ is (d_1, d_2)-**uniformly continuous** if: for each $\varepsilon > 0$, there exists a $\delta > 0$ such that

$$x, y \in D \quad \text{and} \quad d_1(x, y) < \delta \text{ imply } d_2(f(x), f(y)) < \varepsilon.$$

Clearly every (d_1, d_2)-uniformly continuous function is (d_1, d_2)-continuous.

Our concern is to find metrics d_1 and d_2 on R so that (d_1, d_2)-continuous functions $f\colon D \to R$, where $D \subseteq R$, are also (d_1, d_2)-uniformly continuous. Note that if the metric on R is the trivial or discrete metric d_t, defined by $d_t(x, y) = 0$ if $x = y$ and by $d_t(x, y) = 1$ if $x \neq y$, then every function $f\colon D \to R$ is (d_t, d_s)-continuous and (d_t, d_s)-uniformly continuous. In order to avoid such trivialities, we shall require the metrics d_1 and d_2 each to be *equivalent* to the standard metric d_s on R. A metric d on R is equivalent to the metric d_s on R if they each generate the same topology on R or, more simply, if they each generate the same convergent sequences in R (and corresponding points to which those sequences converge). With this requirement, the (d_1, d_2)-continuous functions (from subsets of R into R) are precisely the (d_s, d_s)-continuous functions. Our title question has now become:

> *Is it possible to find two metrics d_1 and d_2 on R, each equivalent to d_s, such that every (d_s, d_s)-continuous function $f\colon D \to R$ is also (d_1, d_2)-uniformly continuous?*

The following functions, defined for all x and y in R, are metrics on R each equivalent to the standard metric d_s:

(a) $$d_b(x, y) = \frac{|x - y|}{1 + |x - y|}$$ (the **bounded metric** for R),

(b) $$d_\infty(x, y) = \left| \frac{x}{1 + |x|} - \frac{y}{1 + |y|} \right|$$ ("**vanishing at infinity**" metric for R).

For discussions of these metrics, see [**9**], pp. 151–152, and [**3**], pp. 135–137. The metric d_b is called the bounded metric for R since every subset of R is bounded with this metric. With the standard metric, d_s, the distance between two consecutive integers n and $n + 1$ is always 1; with the bounded metric, d_b, this distance is always $1/2$. With the metric d_∞, we have

$$d_\infty(n, n+1) = \begin{cases} \dfrac{1}{(n+1)(n+2)} & \text{if } n \geqslant 0 \\ \dfrac{1}{(-n)(1-n)} & \text{if } n < 0. \end{cases}$$

Clearly, $\lim_{n \to +\infty} d_\infty(n, n+1) = 0$ and $\lim_{n \to -\infty} d_\infty(n, n+1) = 0$, hence the name "vanishing at infinity" metric for d_∞. Note that $\lim_{x \to +\infty} d_\infty(x, -x) = 2$.

Although the (d_s, d_s)-continuous function $f\colon (0,1] \to R$, defined by $f(x) = 1/x$, is not (d_s, d_s)-uniformly continuous, f is (d_s, d_∞)-uniformly continuous. For if x, y are in $(0,1]$, we have

$$d_\infty(f(x), f(y)) = d_\infty(1/x, 1/y) = \left| \frac{1/x}{1 + |1/x|} - \frac{1/y}{1 + |1/y|} \right|$$

$$= \left| \frac{1}{1+x} - \frac{1}{1+y} \right| = \left| \frac{x - y}{(1+x)(1+y)} \right| \leqslant |x - y| = d_s(x, y).$$

Clearly, then, given $\varepsilon > 0$, we can choose $\delta = \varepsilon$ and have

$$x, y \in (0,1] \quad \text{and} \quad d_s(x, y) < \delta \text{ imply } d_\infty(f(x), f(y)) < \varepsilon.$$

Likewise, the (d_s, d_s)-continuous but not (d_s, d_s)-uniformly continuous function $g\colon R \to R$, defined by $g(x) = x^2$, is (d_∞, d_∞)-uniformly continuous. In general, every function $f\colon D \to R$, where $D \subseteq R$ and $f(x)$ is a *polynomial* in x, is (d_∞, d_∞)-uniformly continuous. A proof of this is suggested later in this note.

N. Levine [**6**] has shown that given *any* (d_s, d_s)-continuous function $f\colon R \to R$, there exists a metric d_f on R, defined by

$$d_f(x, y) = d_s(x, y) + d_s(f(x), f(y)),$$

which is equivalent to d_s and is such that f is (d_f, d_s)-uniformly continuous. This last fact follows

from the inequality $d_s(f(x), f(y)) \leqslant d_f(x, y)$. He also obtained the following two theorems. Given a *finite collection* of (d_s, d_s)-continuous functions $\{f_1, f_2, \ldots, f_n\}$ from R into R, there exists a metric d on R, defined by

$$d(x, y) = d_s(x, y) + \sum_{j=1}^{n} d_s\big(f_j(x), f_j(y)\big)$$

and equivalent to d_s, such that all the functions $\{f_1, f_2, \ldots, f_n\}$ are (d, d_s)-uniformly continuous. Lastly, given a *countable collection* of (d_s, d_s)-continuous functions $\{f_n\colon n \in N\}$ from R into R, there exists a metric d on R, defined by

$$d(x, y) = d_s(x, y) + \sum_{j=1}^{\infty} \frac{1}{2^j} d_b\big(f_j(x), f_j(y)\big)$$

and equivalent to d_s, such that all the functions $\{f_n\colon n \in N\}$ are (d, d_b)-uniformly continuous, where d_b is the bounded metric for R.

These examples certainly show that metrics, different from but equivalent to d_s, can be found to turn sets of continuous functions that are not necessarily uniformly continuous into functions that are. In fact this can be done simultaneously for any countable collection of continuous functions or it can be done simultaneously for all polynomial functions. On the other hand, it *cannot* be done simultaneously for *all* continuous functions; the answer to our title question is *no*. We prove this in the next section.

Proofs

The most important theorem relating continuity and uniform continuity is the following (see [**10**], p. 87).

THEOREM 1. *Let (X, d_1) and (Y, d_2) be metric spaces, and let $M \subseteq X$. If $f\colon M \to Y$ is (d_1, d_2)-continuous and M is compact, then f is (d_1, d_2)-uniformly continuous.*

Thus restricting the domain (so it is compact) can make every continuous function with that domain uniformly continuous.

Theorem 1 can be used in a rather interesting way to obtain our earlier results for the functions $f(x) = 1/x$ and $g(x) = x^2$. Let R^* be the extended real number system, i.e., let $R^* = R \cup \{-\infty, +\infty\}$ where $-\infty$ and $+\infty$ are two distinct objects not in R. The metric d_∞^* on R^* is defined by

$$d_\infty^*(x, y) = |D(x) - D(y)|$$

where

$$D(x) = \begin{cases} \dfrac{x}{1+|x|} & \text{if } x \in R \\ -1 & \text{if } x = -\infty \\ 1 & \text{if } x = +\infty. \end{cases}$$

Clearly d_∞^* is an extension of the metric d_∞ on R to the set R^*. We extend the function $f(x) = 1/x$ to the interval $[0,1]$ by defining f^* as follows:

$$f^*\colon [0,1] \to R^*$$
$$f^*(x) = f(x) = 1/x, \text{ if } x \in (0,1]$$
$$f^*(0) = +\infty.$$

Here $[0,1]$ is considered a subset of R with its usual metric d_s. Since f^* is (d_s, d_∞^*)-continuous and the set $[0,1]$ is compact, Theorem 1 implies that the function f^* is (d_s, d_∞^*)-uniformly continuous. Consequently, the restriction of f^* to $(0,1]$, namely f, is (d_s, d_∞^*)-uniformly continuous and hence (since the range of f is a subset of R) f is (d_s, d_∞)-uniformly continuous.

Similarly, we can extend $g(x) = x^2$ to the function g^*:

$$g^*: R^* \to R^*$$
$$g^*(x) = g(x) = x^2, \text{ if } x \in R$$
$$g^*(-\infty) = +\infty$$
$$g^*(+\infty) = +\infty.$$

Since g^* is (d_∞^*, d_∞^*)-continuous and the set R^* is compact, the function g^* is (d_∞^*, d_∞^*)-uniformly continuous hence g: $R \to R$ is (d_∞, d_∞)-uniformly continuous. The same argument can be used to prove that every polynomial function f: $R \to R$ is (d_∞, d_∞)-uniformly continuous. In fact, Theorem 1 can be used in this manner to show that every continuous function f: $D \to R$, where $D \subseteq R$, which has an extension to a continuous function f^*: $D^* \to R^*$ where D^* is a closed subset of R^*, is (d_∞, d_∞)-uniformly continuous. Some such functions are the logarithm function $\ln x$, the exponential function e^x, and the rational function defined by $f(x) = 1/x^2$. Moreover, every rational function with vertical asymptotes only at points a where $\lim_{x \to a^+} f(x) = \lim_{x \to a^-} f(x)$ is (d_∞, d_∞)-uniformly continuous. (Incidentally, the function f: $(0,1] \to R$ defined by $f(x) = 1/x$ is also (d_∞, d_∞)-uniformly continuous. Of course, an example of a rational function which is not (d_∞, d_∞)-uniformly continuous is f: $R \setminus \{0\} \to R$ defined by $f(x) = 1/x$.)

H. Hueber has characterized metric spaces (E, d) which have the property that all continuous real-valued functions on E are uniformly continuous *and* has used this to give a new necessary and sufficient condition for compactness [**4**]. Generalizing his result (and its proof), we have the following theorem.

THEOREM 2. *Let (E, d) be a metric space and let d_2 be a metric on R equivalent to the standard metric d_s on R. Then E is compact if and only if both of the following conditions hold*:

(1) *Every (d, d_2)-continuous function f: $E \to R$ is (d, d_2)-uniformly continuous.*
(2) *For every $\varepsilon > 0$, the set $\{x \in E | d(x, E \setminus \{x\}) > \varepsilon\}$ is finite.*

In this theorem, for each point x in E and each subset D of E, the symbol $d(x, D)$ denotes the distance from x to D, i.e., $d(x, D) = \inf\{d(x, y): y \in D\}$. Notice that the set of limit points of E is the set $\{x \in E | d(x, E \setminus \{x\}) = 0\}$.

Theorem 2 can be used to show that it is *not* possible to find two metrics d_1 and d_2 on R, each equivalent to d_s, such that *every* (d_1, d_2)-continuous function f: $R \to R$ is (d_1, d_2)-uniformly continuous. We simply let (E, d) of Theorem 2 be the metric space (R, d_1). Since the open sets of (R, d_1) are the same as the open sets of (R, d_s), the set R is not d_1-compact. Further, the set of d_1-limit points of R is the set of d_s-limit points of R, i.e., $\{x \in R | d_1(x, R \setminus \{x\}) = 0\} = R$. Clearly (2) holds, i.e., for every $\varepsilon > 0$, we have $\{x \in R | d_1(x, R \setminus \{x\}) > \varepsilon\} = \varnothing$ which is a finite set. Thus Theorem 2 implies that statement (1) is false, i.e., it is not the case that every (d_1, d_2)-continuous function f: $R \to R$ is (d_1, d_2)-uniformly continuous. Since d_1 and d_2 were arbitrary metrics on R, each equivalent to d_s, our proof is complete.

A number of papers have been concerned with finding (or showing the existence or nonexistence of) a metric d on a set E such that, with appropriate restrictions on E, every real-valued continuous function on E is uniformly continuous. See the papers by N. Levine [**5**], M. Atsuji [**1**], N. Levine and W. G. Saunders [**7**], W. C. Waterhouse [**11**], and S. G. Mrowka [**8**]. Results from these papers appear in the book by A. Wilansky [**12**]; see p. 59, Problems 113 and 115. Several theorems in these references can be used, as we used Theorem 2, to prove that the answer to our title question is *no*.

Of course the simplest way to show that the answer to our title question is *no* is to find an example, for metrics d_1 and d_2 on R equivalent to d_s, of a continuous function f: $R \to R$ which is not (d_1, d_2)-uniformly continuous. Let d_1 and d_2 be metrics on R with d_1 equivalent to d_s. For each natural number $n \geq 2$, n is a d_s-limit point (and hence a d_1-limit point) of the interval

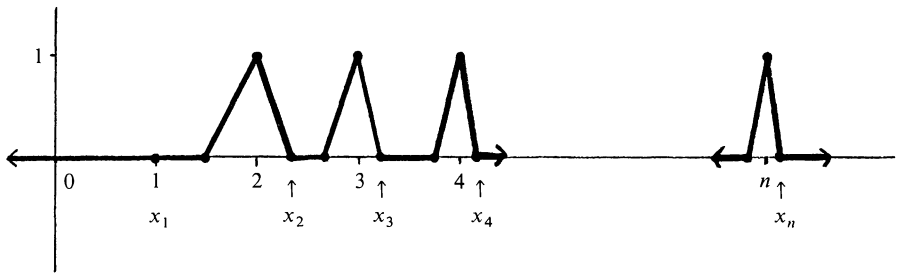


FIGURE 1

$[n, n+1/n)$. Thus there exists a real number x_n in $(n, n+1/n)$ such that $d_1(n, x_n) < 1/n$. Choose $x_1 = 1$. Then $\{x_n\}$ is a sequence in R satisfying

$$\begin{cases} n < x_n < n + 1/n & \text{for } n = 2,3,4,\ldots \\ d_1(n, x_n) \to 0 & \text{as } n \to +\infty. \end{cases}$$

Define a continuous function f: $R \to R$ by

$$f(x) = \begin{cases} nx + (1 - n^2) & \text{for } n - 1/n \leqslant x \leqslant n,\ n \geqslant 2 \\ \dfrac{1}{n - x_n}x - \dfrac{1}{n - x_n}x_n & \text{for } n < x \leqslant x_n,\ n \geqslant 2 \\ 0 & \text{otherwise.} \end{cases}$$

The graph of f is an "infinite saw-blade" with the "teeth" all the same height but getting "sharper and sharper" (see FIGURE 1). Note that for $n \geqslant 2$, we have $d_2(f(n), f(x_n)) = d_2(1,0)$ which is a nonzero constant. Clearly, f is not (d_1, d_2)-uniformly continuous: there exists a real number $\varepsilon = \frac{1}{2}d_2(1,0) > 0$ such that for every $\delta > 0$ there exist real numbers n and x_n such that $d_1(n, x_n) < \delta$ and $d_2(f(n), f(x_n)) = d_2(1,0) \geqslant \varepsilon$. Even though earlier examples in this note may have suggested that the problem of lack of uniform continuity has something to do with functions being unbounded, this last example should dispel that notion since (with $d_2 = d_s$) f is a bounded function which is not (d_1, d_2)-uniformly continuous.

Our examples in this note confirm that uniform continuity is a metric concept, not a topological concept: a continuous function may be uniformly continuous if one pair of metrics is used but not uniformly continuous when another pair of equivalent metrics is used. Other examples which illustrate this can be readily found in textbooks on topology (see [**10**], p. 91; and [**3**], p. 155) and analysis (see [**2**], pp. 51 and 53).

References

[1] M. Atsuji, Uniform continuity of continuous functions of metric spaces, Pacific J. Math., 8 (1958) 11–16.
[2] J. Dieudonné, Foundations of Modern Analysis, Academic Press, New York, 1960.
[3] W. W. Fairchild and C. I. Tulcea, Topology, W. B. Saunders Co., Philadelphia, 1971.
[4] H. Hueber, On uniform continuity and compactness in metric spaces, Amer. Math. Monthly, 88 (1981) 204–205.
[5] N. Levine, Uniformly continuous linear sets, Amer. Math. Monthly, 62 (1955) 579–580.
[6] ———, Remarks on uniform continuity in metric spaces, Amer. Math. Monthly, 67 (1960) 562–563.
[7] N. Levine and W. G. Saunders, Uniformly continuous sets in metric spaces, Amer. Math. Monthly, 67 (1960) 153–156.
[8] S. G. Mrowka, On normal metrics, Amer. Math. Monthly, 72 (1965) 998–1001.
[9] L. A. Steen and J. A. Seebach, Jr., Counterexamples in Topology, 2nd ed., Springer-Verlag, New York, 1978.
[10] W. A. Sutherland, Introduction to Metric and Topological Spaces, Clarendon Press, Oxford, 1975.
[11] W. C. Waterhouse, On *uc* spaces, Amer. Math. Monthly, 72 (1965) 634–635.
[12] A. Wilansky, Topology for Analysis, Ginn and Co., Waltham, Mass., 1970.

Comments on the Cover Illustration, *Moon*

Robert Dixon
Royal College of Art
London S.W.7., England

The computer drawing *Moon* is based on Mandelbrot's discussion [**1**] of lunar cratering as an example of fractal form. Circles lying on the surface of a sphere represent craters in a perspective drawing, complete with a hidden-line algorithm to ensure that no craters or parts of craters will appear which lie beyond the viewer's lunar horizon. I used a microcomputer to drive a Calcomp A3 pen plotter from a very short BASIC program.

Two patterns of randomness figure in this lunar model: first, the crater sizes (radii) are given a **hyperbolic** distribution; and, second, the crater centers are scattered in **isotropic** distribution. Each of these distributions is obtained by forming a simple function of the random variable v, defined on the interval $[0,1]$, by Probability$(v \leqslant n) = n$. In BASIC, RND(1) simulates v.

Few large, more medium, and many more small—this is the shape of a hyperbolic distribution. Mandelbrot points us to many natural phenomena which exhibit such a pattern, from astronomical bodies to human incomes. Formally, $\Pr(r \geqslant R) = R^{-p}$, with positive p and $R \geqslant 1$, defines a hyperbolic distribution of r on the interval $[1, \infty]$. So what transformation of v gives the required r?

$$\Pr(v \leqslant n) = n \Leftrightarrow \Pr(v^{-1/p} \geqslant n^{-1/p}) = n \Leftrightarrow \Pr(v^{-1/p} \geqslant R) = R^{-p}.$$

Hence, for the crater radius r, we put $r = v^{-1/p}$.

Note that the procedure for forming an appropriate function of v leads to an inverse function. Mandelbrot suggests that $p = 2$ gives a good approximation to the actual lunar surface, and to those of other known planetary satellites. It is also the value of p which gives statistical self-similarity.

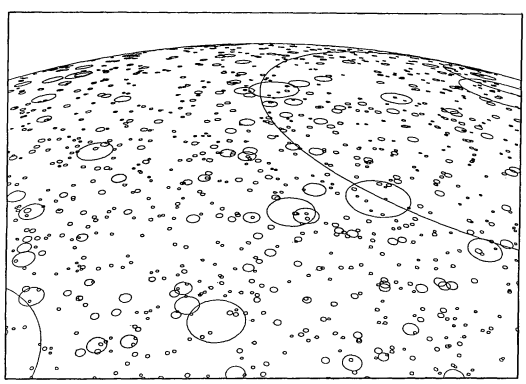
To achieve randomly even scattering of points (x, y) within a plane rectangle is simply a matter of obtaining x and y as independent and linear functions of v. To achieve the analogous scattering on a spherical surface—isotropic distribution—requires a little more subtlety. A solution is obtained by using (longitude, latitude) coordinates for the location of each center, with

$$\text{longitude} = 2\pi v, \qquad \text{latitude} = \arccos(1 - 2v).$$

Making one of the coordinates a linear function of v forces the other coordinate to be a nonlinear function of v. The reasoning behind the above solution is left to the reader to figure out. (Clue: what is the area of the polar cap bounded by latitude θ on a globe of unit radius?)

Reference

[1] Benoit B. Mandelbrot, The Fractal Geometry of Nature, Freeman, 1982.



Lunar Landscape
—R. Dixon

PROBLEMS

LEROY F. MEYERS, Editor
G. A. EDGAR, Associate Editor
The Ohio State University

Proposals

To be considered for publication, solutions should be mailed before October 1, 1984.

1191. Let $P_1, P_2, \ldots, P_n$ be the vertices of a regular polygon inscribed in the unit circle. Let A_n denote the sum, and B_n the arithmetic mean, of the areas of all triangles $P_1P_jP_k$ for $1 < j < k \leqslant n$. Evaluate A_n, show that A_n decreases as $n \to \infty$, and evaluate $\lim_{n\to\infty} B_n$. [*L. Kuipers, Sierre, Switzerland.*]

1192. Let M be an $n \times n$ matrix with complex elements such that $\operatorname{tr} M = 0$. Prove:

(i) $x^*Mx = 0$ for some nonzero vector x;

(ii) there are unitary matrices U such that U^*MU has all its diagonal elements 0. (M^* denotes the conjugate transpose of M.) [*The late H. Kestelman, University College London, England.*]

1193. Let $a_1, a_2, \ldots, a_m$ be fixed nonnegative integers, and let n be a fixed integer with $n \geqslant \sum_{i=1}^m a_i$. Show that

$$\sum_{*} \prod_{i=1}^{m} \binom{k_i}{a_i} = \binom{n+m-1}{n-\sum_{i=1}^m a_i},$$

where the *-sum is over all ordered m-tuples $(k_1, k_2, \ldots, k_m)$ of integers that sum to n. [*Allen J. Schwenk, U.S. Naval Academy.*]

1194. Find a simple smooth function f of the two variables x and y such that:

(1) the differential equation $y' = f(x, y)$ is solvable in closed form, and

(2) the domain of definition of f is connected but not simply connected (and cannot be made so by continuous extension of f). [*J. Walter, RWTH Aachen, West Germany.*]

1195. Phone books, n in number, are kept in a stack. The probability that the book numbered i (where $1 \leqslant i \leqslant n$) is consulted for a given phone call is $p_i > 0$, where $\sum_{i=1}^n p_i = 1$. After a book is used, it is placed at the top of the stack. Assume that the calls are independent and evenly spaced, and that the system has been employed indefinitely far into the past. Let d_i be the average depth of book i in the stack. Show that $d_i \leqslant d_j$ whenever $p_i \leqslant p_j$. Thus, on the average, the more popular books have a tendency to be closer to the top of the stack. (Compare problem 1159, vol. 57 (1984), p. 50.) [*Boris Pittel, The Ohio State University.*]

ASSISTANT EDITORS: DANIEL B. SHAPIRO and WILLIAM A. MCWORTER, JR., *The Ohio State University.*

We invite readers to submit problems believed to be new. Proposals should be accompanied by solutions, if at all possible, and by any other information that will assist the editors. A problem submitted as a Quickie should have an unexpected, succinct solution. An asterisk (*) will be placed next to a problem number to indicate that the proposer did not supply a solution.

Solutions should be written in a style appropriate for Mathematics Magazine. Each solution should begin on a separate sheet containing the solver's name and full address. It is not necessary to submit duplicate copies.

Send all communications to the problems department to Leroy F. Meyers, Mathematics Department, The Ohio State University, 231 W. 18th Ave., Columbus, Ohio 43210.

Quickies

Solutions to Quickies appear at the conclusion of the Problems section.

Q689. Let X and Y be identically distributed random variables, and let c be a constant. Set $a \dot{-} b = \max(a - b, 0)$. Show that

$$E[(X+Y) \dot{-} c] \leqslant E[2X \dot{-} c].$$

[*Eric Bach, Mike Luby, and J. O. Shallit, University of California.*]

Q690. A farmer has a long straight wall along one side of his property. He intends to build a pen in the shape of a right triangle with one leg along the existing wall. He wants to know how to allocate a given amount of fencing between the other leg and the hypotenuse of the pen so as to maximize the area. No computations, please! [*Chester Palmer, Auburn University at Montgomery.*]

Solutions

Quotients of Sums of Two Squares **March 1983**

1167. Let S be the set of all integers of the form $x^2 + y^2$ where x and y are integers. It is readily seen from Fermat's identity that if $u \in S$ and $v \in S$, then $uv \in S$. Prove that if $m \in S$ and $n \in S$ and $m/n = l$ is an integer, then $l \in S$, *by *elementary means*, i.e., without using the well-known characterization of members of S in terms of their factorization into primes. [*Edward T. H. Wang, Wilfrid Laurier University.*]

Solution I (edited): Let l, m, and n be positive integers such that $m \in S$, $n \in S$, and $m/n = l$. Let n' be the largest divisor of l which belongs to S. Then $m = nl = NL$, where $N = nn' \in S$ and $L = l/n'$ is a positive integer. The only divisor of L which belongs to S is 1.

Let t be the smallest positive integer such that $tL \in S$. (Obviously there is such a t, since $LL \in S$, and so $1 \leqq t \leqq L$.) Let $x^2 + y^2$ be a representation of tL, and let ξ and η be the absolutely least remainders of x and y, respectively, modulo t. Thus

$$\xi \equiv x \pmod{t}, \quad \eta \equiv y \pmod{t}, \quad |\xi| \leqq \frac{1}{2}t, \quad \text{and} \quad |\eta| \leqq \frac{1}{2}t,$$

so that $\xi^2 + \eta^2 \equiv x^2 + y^2 \equiv 0 \pmod{t}$, and so $\mu t = \xi^2 + \eta^2 \in S$ for some integer μ. Then

$$\mu t^2 L = (\mu t)(tL) = (\xi^2 + \eta^2)(x^2 + y^2) = (\xi x + \eta y)^2 + (\xi y - \eta x)^2.$$

But $\xi x + \eta y \equiv x^2 + y^2 \equiv 0 \pmod{t}$ and $\xi y - \eta x \equiv xy - yx \equiv 0 \pmod{t}$. Hence $\xi x + \eta y = ut$ and $\xi y - \eta x = vt$ for some integers u and v, so that $\mu t^2 L = (u^2 + v^2)t^2$, or $\mu L = u^2 + v^2 \in S$. But $0 \leqq \mu t = \xi^2 + \eta^2 \leqq \frac{1}{2}t^2$, or $0 \leqq \mu \leqq \frac{1}{2}t < t$, so that $\mu = 0$ by the minimality of t.

But if $\mu = 0$, then $\xi = \eta = 0$, so that $x = x't$ and $y = y't$ for some integers x' and y', not both 0. Then $tL = (x'^2 + y'^2)t^2$, or $L = kt$, where $k = x'^2 + y'^2 \in S$. The definition of L now implies that $k = 1$, and so $t = L$. Hence the smallest positive multiple of L which belongs to S is LL.

Now let s be the smallest positive element of S such that $sL \in S$. (Obviously there is such an s, since $NL = m \in S$ with $N \in S$. Hence $L = t \leqq s \leqq N$.) Let $a^2 + b^2$ be a representation of sL, and let α and β be the absolutely least remainders of a and b, respectively, modulo L. Thus

$$\alpha \equiv a \pmod{L}, \quad \beta \equiv b \pmod{L}, \quad |\alpha| \leqq \frac{1}{2}L, \quad \text{and} \quad |\beta| \leqq \frac{1}{2}L.$$

Hence $\alpha^2+\beta^2 \equiv a^2+b^2 \equiv 0 \pmod{L}$, so that $hL=\alpha^2+\beta^2 \in S$ for some integer h. But $0 \leqq hL = \alpha^2+\beta^2 \leqq \frac{1}{2}L^2$, so that $0 \leqq h \leqq \frac{1}{2}L < L = t$. Hence $h=0$ by the minimality of t. But if $h=0$, then $\alpha=\beta=0$, so that $a=a'L$ and $b=b'L$ for some integers a' and b', not both 0. Hence $sL=(a'^2+b'^2)L^2$, so that $(a'^2+b'^2)L=s \in S$. However, $a'^2+b'^2 \in S$ and $s \leqq sL$. Hence $a'^2+b'^2=s$, by the minimality of s. Thus $L=1$, and so $l=n' \in S$.

DENNIS HAMLIN, student
University of Minnesota

Solution II: First, let us note that we may restrict our attention to the case where n is a square member of S. For if u and v are in S, with u/v integral, then $(uv)/v^2$ is also integral, with numerator uv and denominator v^2 both in S. Furthermore, we may limit our attention to denominators which are squares of primes, checking to see that the quotient is in S at each step.

Hence we are left with the problem: Given that p^2, where p is prime, is a divisor of a sum x^2+y^2, show that x^2+y^2 can be replaced by an equal sum h^2+k^2, with both h and k multiples of p. .

We factor x^2+y^2 in the ring $Z[i]$ of complex integers, which is a unique factorization domain. [Elementary proofs of the results needed on the Gaussian integers can be found in Niven and Zuckerman, *An Introduction to the Theory of Numbers*, 4th ed., ch 9.] Because $x^2+y^2=(x+yi)(x-yi)$, each prime factor (in $Z[i]$) of $x+yi$ is the conjugate of a prime factor of $x-yi$, and vice versa.

If the prime p (as an element of Z) remains a prime in $Z[i]$, then it will divide $x+yi$ or $x-yi$, and so will divide both the real part x and the imaginary part $\pm y$, as desired.

Otherwise, $p=\pi\bar{\pi}$ for some Gaussian prime π, and one will find either (a) $\pi\bar{\pi}$ as a factor of both $x+yi$ and $x-yi$, in which case p divides both x and y, or (b) $\pi\pi$ as a factor of (say) $x+yi$ and $\bar{\pi}\bar{\pi}$ as a factor of $x-yi$, in which case $x+yi=\pi^2\xi$ and $x-yi=\bar{\pi}^2\bar{\xi}$ for some Gaussian integer ξ, so that $x^2+y^2=h^2+k^2$, where $h+ki=\pi\bar{\pi}\xi=p\xi$ and $h-ki=\pi\bar{\pi}\bar{\xi}=p\bar{\xi}$ and the integers h and k (in Z) are divisible by p.

RICHARD PARRIS
Phillips Exeter Academy

L. Kuipers (Switzerland) and the proposer used the well-known characterization of the positive elements of S as the integers whose prime factorizations contain primes congruent to 3 modulo 4 only with an even exponent. There were two incorrect solutions.

Solution I is similar to the proof of the Norm Theorem in Heinrich Dörrie, *100 Great Problems of Elementary Mathematics*, Dover, New York, 1965, pp. 83–84. *Enzo R. Gentile (Argentina)* remarked that the problem is equivalent to proving that for $l \in Z$, if X^2+Y^2 represents l over Q, then it represents l over Z. This is contained in the Davenport-Cassels theorem, which is found in T. Y. Lam, *The Algebraic Theory of Quadratic Forms*, pp. 273–274, ex. 2, and in F. Lorenz, *Quadratische Formen über Körpern*, Lecture Notes in Math. 130, pp. 18–20. The proof does not use the characterization of elements of S.

An Angle Bisector Triangle's Angle — May 1983

1170. In triangle ABC, the bisectors of the angles A, B, and C are the segments AP, BQ, and CR, respectively. If $AB=4$, $AC=5$, and $BC=6$, find the size of $\angle QPR$. [*John P. Hoyt, Lancaster, Pennsylvania.*]

Solution I: Use of the fact that $BP:PC=AB:AC$, etc., followed by seven applications of the law of cosines, on triangles ABC (three times), AQR, BRP, CPQ, and PQR, to determine $\cos A$, $\cos B$, and $\cos C$, distances QR, RP, and PQ, and finally $\cos\angle QPR$, yields

$$\angle QPR = \arccos\frac{10}{\sqrt{667}} \approx 67°13'10.26''.$$

COMPOSITE of solutions

Solution II: Use the law of cosines to find $\cos A$, and then use the formula

$$\tan\angle QPR = \frac{\sin B + \sin C}{\frac{1}{2} + \cos A} \quad \text{if } A \neq 120°, \tag{1}$$

noting that $\sin B + \sin C = \dfrac{b+c}{a} \sin A$, to obtain

$$\angle QPR = \operatorname{Arctan} \frac{9\sqrt{7}}{10}.$$

A proof of (1) may be found in C. F. Parry, "A variation on the Steiner-Lehmus theme", *Math. Gazette*, v. 62 (1978), pp. 89–94, esp. 93–94. An alternative solution uses vector algebra to compute $\overrightarrow{PQ} \cdot \overrightarrow{PR}$.

JOHN P. HOYT
Lancaster, Pennsylvania

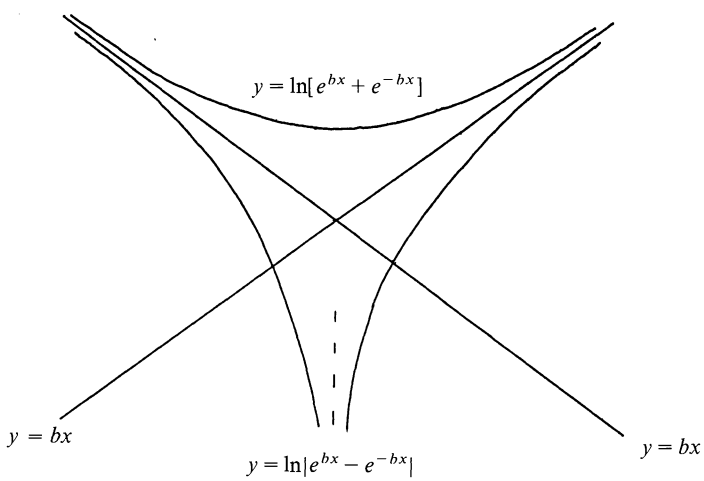
Solved as in Solution I, or with similar use of the law of sines, by Raul Arias (student), Leon Bankoff, Arvin Ben-Zvi (Israel), Walter Bluger (Canada), David Boduch (student), Eric Dew, Ragnar Dybvik (Norway), Milton P. Eisner, Herta Freitag, Ralph Garfield, Lisa M. Graber, G. A. Heuer & C. V. Heuer, Geoffrey A. Kandall, Mark Kantrowitz (student), L. Kuipers (Switzerland), Kurtis H. Lemmert (student), Henry S. Lieberman, David Lindsay, Todd Listerman (student), John J. Martinez, Glen E. Mills, Roger B. Nelsen, Hong Nguyen & Bich Nguyen (students), David F. Paget (Australia), Richard Parris, Lawrence A. Ringenberg, James S. Robertson, Simon W. Strauss, Michael Vowe (Switzerland), Edward T. H. Wang (Canada), and Harry Zaremba. Solved using analytic geometry by Bill Olk (student), Robert S. Stacy (West Germany), Katherine Woerner, and the following students of Kenneth A. Brown, Jr. (jointly): Paul Barkett, Ian Berkowitz, Jodi Cohen, Beth Ehrlich, Gary Goldstein, Jeff Hoffman, Billy Mulligan, Silke Parl, Mike Rosen, Douglas Ross, Eric Rubin, Stephanie Rubin, Greg Simon, Ken Strick, and Lance Washington. There were three incorrect solutions.

Arias, Bankoff, Dew, Robertson, Vowe, and the proposer noted that $A = 2C$. Hence $AP = PC$ and $\Delta APB \sim \Delta ACB$.

An Elementary Differential Equation — May 1983

1171. Find all functions f such that $f''(x) + (f'(x))^2 = b^2$, where b is a given constant. Which of these solutions are polynomials? [*Mark Kantrowitz, student, Brookline, Massachusetts.*]

Solution I: The linear functions given by $f(x) = C \pm bx$ (C an arbitrary constant) are the only polynomial solutions, and they are also exceptional in that they serve as asymptotes for the other solutions, which all satisfy either $(f'(x))^2 > b^2$ or $(f'(x))^2 < b^2$ (if $b \neq 0$). In what follows, we assume that $(f'(x))^2 \neq b^2$ for all x in some interval. Then the differential equation is equivalent to $f''(x)/(b^2 - (f'(x))^2) = 1$. If $b \neq 0$, we can expand by partial fractions and then antidifferentiate:



$$\ln\left|\frac{b+f'(x)}{b-f'(x)}\right| = 2b(x-h),$$

or

$$\frac{b+f'(x)}{b-f'(x)} = \pm e^{2b(x-h)},$$

or

$$f'(x) = b\cdot\frac{\pm e^{2b(x-h)}-1}{\pm e^{2b(x-h)}+1} = b\cdot\frac{e^{b(x-h)}\mp e^{-b(x-h)}}{e^{b(x-h)}\pm e^{-b(x-h)}},$$

so that

$$f(x) = \ln|e^{b(x-h)}\pm e^{-b(x-h)}| + C$$

describes the solutions with $f'(x)\neq\pm b$. The arbitrary constants h and C serve to translate the basic configuration shown in the FIGURE.

If $b=0$, there are the logarithmic solutions

$$f(x) = C + \ln|x-h|,$$

in addition to the linear solutions $f(x)=C$ already noted.

RICHARD PARRIS
Phillips Exeter Academy

Solution II: Let f be a function defined on some interval I such that $f''(x)+(f'(x))^2=b^2$ for all x in I. Put $g=e^f$. Then

$$g'=f'e^f \quad\text{and}\quad g''=f''e^f+(f')^2e^f=e^f\left(f''+(f')^2\right)=b^2e^f=b^2g.$$

Thus, g is a solution of the elementary second order linear differential equation $g''=b^2g$. If $b\neq 0$, then there are constants A and B such that $g(x)=Ae^{bx}+Be^{-bx}$. If $b=0$, then there are constants A and B such that $g(x)=Ax+B$. Consequently,

$$f(x)=\begin{cases}\ln|Ax+B| & \text{if } b=0,\\ \ln|Ae^{bx}+Be^{-bx}| & \text{if } b\neq 0.\end{cases}$$

Suppose f is a polynomial solution which is of degree $n\geq 2$. Then the degree of f'' is $n-2$ and the degree of $(f')^2$ is $2n-2$. Consequently, the sum of f'' and $(f')^2$ cannot be constant. Therefore, $n\leq 1$, and so $f(x)=mx+c$ for some constants m and c. Substitution into the differential equation yields $m^2=b^2$, or $m=\pm b$. Thus, the only polynomial solutions are $f(x)=\pm bx+c$.

MICHAEL B. GREGORY
University of North Dakota

Also solved completely by Robert A. Close, Beatriz Margolis (France), Roger Nelsen & Harvey Schmidt, Jr., Bill Olk (student), and W. R. Utz. Twenty-seven incomplete solutions were submitted. In many of these, it was forgotten that: b could be 0; $(f'(x))^2$ could equal b^2; from $\ln|u|=2bx+c$ it follows that $u=\pm e^{2bx+c}$; or different hyperbolic substitutions for evaluating $\int du/(b^2-u^2)$ are needed according as $b^2>u^2$ or $b^2<u^2$ (if $b\neq\pm u$). An elementary problem need not be solved mechanically.

A Unique Solution — May 1983

1172. Determine the number of real solutions x $(0\leq x\leq 1)$ of the equation

$$(x^{m+1}-a^{m+1})(1-a)^m = \left\{(1-a)^{m+1}-(1-x)^{m+1}\right\}a^m,$$

where $0\leq a\leq 1$ and m is a positive integer. [*M. S. Klamkin, University of Alberta, Canada.*]

Solution: Let $f(x) = \text{l.h.s.} - \text{r.h.s.}$ and note that

$$\begin{aligned}\operatorname{sgn} f'(x) &= \operatorname{sgn}\big((m+1)\big(x^m(1-a)^m - (1-x)^m a^m\big)\big)\\ &= \operatorname{sgn}(x(1-a) - (1-x)a) = \operatorname{sgn}(x-a).\end{aligned}$$

Thus $f(x)$ is minimal when $x = a$. Since $f(a) = 0$, $x = a$ is the only solution to the given equation.

VANIA D. MASCIONI, student
ETH Zürich, Switzerland

Also solved by J. H. Abbott & P. Puri (two solutions), Curtis Cooper, Ragnar Dybvik (Norway), Alberto Facchini (Italy), Gordon Fisher, Chico Problem Group, Chris Hackman (student), Mark Kantrowitz (student), H. Kappus (Switzerland), L. Kuipers (Switzerland), David Lindsay, Beatriz Margolis (France), H. G. Mushenheim, David F. Paget (Australia), Richard Parris, Robert S. Stacy (West Germany; two solutions), and the proposer. There were two incorrect solutions.

Abbott & Puri, *Cooper*, *Parris*, and *Stacy* algebraically reduced the problem to that of solving $(x-a)f(x/a) = (x-a)f((1-x)/(1-a))$, where f is the strictly increasing function defined for $y \in [0,1]$ by $f(y) = 1 + y + \cdots + y^m$. *Abbott & Puri* and the *proposer* used convexity arguments on the mth-power function. *Margolis* noted that m may be any real number not less than 1. *Abbott & Puri*, *Hackman*, and *Parris* noted that there is only one solution for x on the real line if m is an odd integer, but there is a negative solution if m is even.

A Know/Don't Know Game — May 1983

1173. (a) Two positive integers are chosen. The sum is revealed to logician A, and the sum of squares is revealed to logician B. Both A and B are given this information and the information contained in this sentence. The conversation between A and B goes as follows: B starts.

B: "I can't tell what the two numbers are."
A: "I can't tell what the two numbers are."
B: "I can't tell what the two numbers are."
A: "I can't tell what the two numbers are."
B: "I can't tell what the two numbers are."
A: "I can't tell what the two numbers are."
B: "Now I can tell what the two numbers are."

What are the two numbers?

(b) When B first says he cannot tell what the two numbers are, A receives a large amount of information. But when A first says that he cannot tell what the two numbers are, B already knows that A cannot tell what the two numbers are. What good does it do B to listen to A? [*Thomas S. Ferguson, University of California, Los Angeles.*]

Solution: After B's first announcement, the possibilities are reduced to a table which begins as follows:

SUM	SUM OF SQUARES
8	50
9	65
10	50
11	65, 85
13	85, 125, 145
14	130, 170
15	125
16	130, 200
17	145, 185, 205
...	...

The only appearances of 50, 65, 85, 125, and 145 as a sum of squares in this table are already

displayed. A's first announcement rules out the sums 8, 9, 10, and 15. B's second announcement rules out the 65 from the row of sum 11 and the 125 from the row of sum 13. A's second announcement rules out the sum 11. B's third announcement rules out the 85 from the row of sum 13. A's third announcement rules out the sum 13. B's fourth announcement shows one possible answer to be that the sum is 17 and the sum of squares 145, giving 8 and 9 as the numbers.

To show that this is the only possible answer, it is sufficient to show that each subsequent sum in this table from 18 on has at least two entries. By considering the identity $(a+2b)^2+(2a-b)^2 = (a-2b)^2+(2a+b)^2$, pointed out to me by Robert Steinberg, for $b=1,2,3$ and $a \geqq 7$, we may conclude that every row from 22 on contains at least two entries; that rows 18 through 21 also have at least two entries follows by considering in addition the special cases of the identity with (a, b) equal to (5,3), (6,1), (6,2), and (6,5), respectively.

(b) The secondary problem is much easier to answer if A's second announcement is that he *can* tell what the two numbers are. The unique solution is then: sum 11, sum of squares 85, numbers 2 and 9. In this case, also, B knows that A's first announcement is true before A makes it. However, A doesn't know that B knows this, so he announces he cannot tell, in the hope that B has 65 and can immediately say he knows the numbers. This does convey information to B, who is not able to use it yet. B admits this in his second announcement, whereupon A can tell the two numbers. The complete answer to the secondary problem is much more complex, and can be found by moving the announcement of finding the numbers down successively to B's fourth announcement. In each case, there is a unique solution.

THOMAS S. FERGUSON
University of California, Los Angeles

Solved partially (no uniqueness proof) by Anon (Erewhon), Robert Galia, Matthew A. Jaffey, David Lindsay, Richard Parris, James S. Propp (student, (b) only), and Eric S. Rosenthal. There was one incorrect solution.

Related material may be found in A. K. Austin, "A calculus for know/don't know problems", this MAGAZINE, v. 49 (1976), pp. 12–14.

Answers

Solutions to the Quickies which appear near the beginning of the Problems section.

Q689. $a \dot{-} b = (a-b)^+$, the positive part of $a-b$. Now $(a+b)^+ \leqslant a^+ + b^+$. By subtracting $\frac{1}{2}c$ from X and Y we may assume $c=0$. Hence

$$E[(X+Y)^+] \leqslant E[X^+]+E[Y^+] = E[2X^+].$$

Q690. Imagine his neighbor working on the same problem on the other side of the wall. Together, they are using a fixed amount of fencing to build a triangle. But for a given perimeter, the triangle of largest area is equilateral. Hence, fencing should be allocated to the leg and the hypotenuse of the original triangle in the ratio 1 : 2.

REVIEWS

PAUL J. CAMPBELL, Editor
Beloit College

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

Peterson, I., *Faster factoring for cracking computer security*, Science News 125 (14 January 1984) 20.

The 67-digit number $11^{64}+1$ has been factored by mathematicians at Sandia National Labs, using an improved quadratic sieve. They expect to crack a 71-digit number soon. Only eight years ago, the best anyone could do with a day of computer time was to factor a 40-digit number.

Kolata, Gina, *Factoring gets easier; mathematicians are exploiting computer designs to factor large numbers in times that, as recently as 1 year ago, seemed inconceivable*, Science 222 (2 December 1983) 999-1001; reply, *ibid.* 223 (6 January 1984) 8.

"A few years ago, an interest in factoring was the hallmark of a proven eccentric. Now it suddenly relates not only to the transfer of funds between banks but also to the national security." There's a certain amount of irony in this, says John Brillhart (Arizona). The new developments in factoring are a consequence of mathematicians' newfound interest in exploiting computer architecture. In a reply, Gustavus Simmons (Sandia) notes that advances in architecture have been accompanied by corresponding advances in algorithms.

Kolata, Gina, *Another promising code falls: a code that looked too good to be true has a fatal weakness and now can be broken in a few seconds*, Science 222 (16 December 1983) 1224.

Donald Coppersmith (IBM) has shown how to break the public-key cryptosystem called "discrete exponentials." Its premise is that raising to powers is easy but taking logs is hard; but the coding takes place in a Galois field of size 2^n, so one can exploit the fact that in such fields $(x+y)^2 = x^2+y^2$.

Sloane, N.J.A., *The packing of spheres*, Scientific American 250:1 (January 1984) 116-125, 146.

What is the densest way to arrange identical spheres in space? Even for 3-space, the answer is unknown; but dense configurations (particularly in 24 dimensions!) offer applications in fields as diverse as digital signalling and the theory of simple groups.

Dembart, Lee, *Quick fixes with error-correcting codes*, High Technology 4:1 (January 1984) 18-21.

Mentions uses, and gives a simple example, of error-correcting codes. Both block and convolutional codes are described and compared.

Interview: Benoit B. Mandelbrot, Omni (February 1984) 64-66, 102-107.

The inventor of fractals describes how his visual orientation steered his discovery toward success. The interview ends with a discussion of esthetics in architecture, furniture, biology, and fractal art ("symmetrical to the most extreme degree").

Peterson, I., *Ants in labyrinths and other fractal excursions*, Science News 125 (21 January 1984) cover, 42-43.

"Trees are one of our biggest failures." So says B. Mandelbrot (IBM), who creates fractal objects that look like natural patterns. But fractals are also rapidly becoming an important scientific tool; almost every issue of Physical Review Letters has an article using fractals and phenomena with fractal dimension. Why fractals work is still an unanswered question.

McDermott, Jeanne, *Geometrical forms known as fractals find sense in chaos*, Smithsonian 14:9 (December 1983) cover, 110-117.

Skip most of the text; you already knew this much about fractals. It's the illustrations in this article that are worth paying attention to: a fractal dragon resembling ginger root, a mountainscape, fractals as art.

Brams, Steven J., and Fishburn, Peter C., *America's unfair elections: "one voter, one vote" sounds democratic, but mathematicians know otherwise*, The Sciences 23:6 (November-December 1983) 28-34.

It's time again in this U.S. presidential election year to reconsider our voting methods. Brams and Fishburn, authors of Approval Voting, here again plump for approval voting, claiming that under more democratic than "one voter, one vote" is "one voter, n votes."

Bentley, Jon, *Programming pearls*, Communications of the ACM, every issue.

This splendid regular column began in August 1983, a belated component of the earlier revamping of the Communications of the ACM. The column should be regular reading for students in computer science courses beyond the first course in structured programming. Whether it's showing how to write a binary search correctly or advising when to use a data structure, Bentley teaches the elegance necessary in computer science--his column is amptly-named! Exercises are offered, too, with answers deferred to the next issue.

Hayes, Brian, *Computer recreations: on the ups and downs of hailstone numbers*, Scientific American 250:1 (January 1984) 10-16, 146.

Hayes reflects on a problem sometimes called the Collatz problem: what is the behavior of the replacement heuristic $n \to 3n + 1$ if n is odd, $n \to n/2$ if n is even? "The path the series follows [from a given starting n] is rather like the trajectory of a hailstone through a storm cloud, rising in updrafts and then falling under its own weight." The unproved conjecture is that every such sequence terminates in 1.

Olson, Steve, *Sage of software*, Science 84 (January-February 1984) 74-80.

Portrait of Edsger Dijkstra, the dean of present-day computer scientists, who begins by stating "most of NASA's software is full of bugs." The inventor of structured programming, Dijkstra has tried for 20 years to develop the mathematical science of programming. "In programming," he asserts, "elegance is not a dispensable luxury. It is a matter of life and death."

Wilkie, Tom, and Lamb, John, *Maths scoop of the century is "nonsense,"* New Scientist (19 January 1984) 9; reply, *ibid.* (2 February 1984) 46.

According to the Guardian of Jan. 12, a certain Arnold Arnold has developed "computing techniques" that render current NATO ciphers vulnerable and--almost as a byproduct--"prove" Fermat's last theorem. New Scientist attempts to discredit Arnold, but the result is gibberish, exceeded only in its degree of gobbledygook by Arnold's responding letter. The usually highly reliable magazine has revealed more talent for "scoop" than for maths."

Gardner, Martin, Order and Surprise, Prometheus, 1983; 396 pp.

Science: Good, Bad and Bogus, reprinted Gardner's writings about pseudoscience; this sequel comprises essays and reviews mostly on other topics. Notable are "Mathematics and the folkways," a reflection on L.A. White's "The locus of mathematical reality," and "How not to talk about mathematics," a review of P. Davis and R. Hersh's The Mathematical Experience (1980), which Gardner roundly castigates for the same extreme anthropocentric point of view.

Jones, David E.H., The Inventions of Daedalus: A Compendium of Plausible Schemes, Freeman, 1982; 204 pp, (P).

What can you do in the desert with a 720-meter high column of syrup? What are the hazards of microwave-induced mammoth-resuscitation? What is the curious algorithm you'd use to build a nest if you were a termite? Radioastrology, digging for electricity, radioactive levitation, thermal glidoons, tired light, amplified smell, milk of amnesia, and non-Newtonian trousers--they're all here, a hilarious array of technical proposals that harness accepted scientific proposals and ride them relentlessly. Mathematics plays a strong supporting role (sad to say) in the calculations that buttress the feasibility studies for these schemes. Positively the funniest book with serious science in it!

Tufte, Edward R., The Visual Display of Quantitative Information, Graphics Press (Box 430, Chesire, CT 06410), 1983; 197 pp, $34.

A landmark book, *the* successor to Huff's How to Lie with Statistics. A beautiful compendium of the best and the worst in data graphics, together with an authoritative enunciation of principles for display of quantitative information. Statistics students should read and enjoy it. Your library should have it.

Chambers, John M., *et al.*, Graphical Methods for Data Analysis, Wadsworth, 1983; xiv + 395 pp, $24.95, $17.95 (P).

The methodology discussed in this book is oriented not toward communicating information or storing data compactly, but toward analyzing the data to discover its structure. As a result, pie charts and pictograms don't appear; instead, sets of data reveal their character through successive plots, the result of each plot suggesting the next. The authors have chosen to present methods useful in their work (at Bell Labs), ranging from box plots, variations on scatter, plots, and "draftsman's displays," to quantile-quantile plots and data transformations. Elementary statistics suffices as background.

Melzak, Z.A., Bypasses: A Simple Approach to Complexity, Wiley, 1983; xvi + 245 pp, $37.50.

A familiar technique in solving a problem is to transform the problem, solve the transformed problem, then transform back--in symbols, TST^{-1}. Melzak calls this the conjugacy principle, or the *bypass*. Examples and applications vary from theology to humor, from communications to transport, but the bulk of the book is devoted to illustrations in mathematics. Even if the reader occasionally finds the bypass concept has been stretched a bit far, this is a mind-expanding book, dazzling in its attempt to comprehend complexity.

Koblitz, Ann Hibner, A Convergence of Lives: Sofia Kovalevskaia: Scientist, Writer, Revolutionary, Birkhauser, 1983; xx + 305 pp, $19.95.

Splendid biography of the 19th century's preeminent woman scientist. The author has consulted the full range of available sources in Russia and Sweden, and her account steers clear of temptations to speculate or enlarge.

Halmos, P.R., Selecta: Expository Writing, edited by Donald E. Sarason and Leonard Gillman, Springer-Verlag, 1983; xix + 304 pp, $19.80.

A collection of some of the finest writing about mathematics, by a master of the art, 1983 winner of a Pólya Award and the Steele Prize. You'll find some of your favorites here, from "What does the spectral theorem say?" to "American mathematics from 1940 to the day before yesterday" and "How to write mathematics."

Wenninger, Magnus J., Dual Models, Cambridge U Pr, 1983; xii + 156 pp, $19.95.

"The priest with a passion for polyhedra" has brought us a sequel to Polyhedral Models and Spherical Models. Models are presented in photographs, along with line drawings, diagrams, and commentary. Even if you can't pronounce it, you can now make and enjoy a great disdyakistriacontahedron or a rhombicosacron!

Curves, Leapfrogs (Tarquin Publns., Stradbroke, Diss, Norfolk 1P21 5JP, England, 1982; 80 pp, £8.00; (P) £4.95.

A feast of geometric curves: cycloidal, trochoidal, conics, envelopes, and pedal curves. Intended for schoolchildren, this beautiful book concentrates on construction rules for loci and omits algebraic descriptions of the curves.

Steinhaus, Hugo, Mathematical Snapshots, 3rd ed., Oxford U Pr, 1983; 311 pp, $7.95 (P).

This reprint in paperback of the 3rd revised edition (1969) of a 1939 work has actually led to some new mathematics! Figure 2 illustrates a decomposition of an equilateral triangle into four pieces which can be reassembled into a square. Isaac Schoenberg discusses the problem, and others of Steinhaus's "snapshots," in Mathematical Time Exposures (MAA, 1983). It was Don Crowe (UW-Madison), however, who noticed that there is an error in the figure--and thereby hangs a tale.

Steen, Lynn Arthur (ed.), Undergraduate Mathematics Education in the People's Republic of China: Report of a 1983 North American Delegation, MAA, 1984; x + 99 pp, $5 (P).

Report on a June 1983 visit by 58 North American mathematics educators to Chinese universities. Included are copies of curricula and of the entrance examinations for admission to high school, university, and graduate school, which determine whether a student will get into a "key" institution. University teaching loads are 3-4 hours, and student-faculty ratios range from 1 to 5. It would be interesting to read a similar report by a delegation of Chinese (or even Russian) mathematics educators on undergraduate mathematics education here.

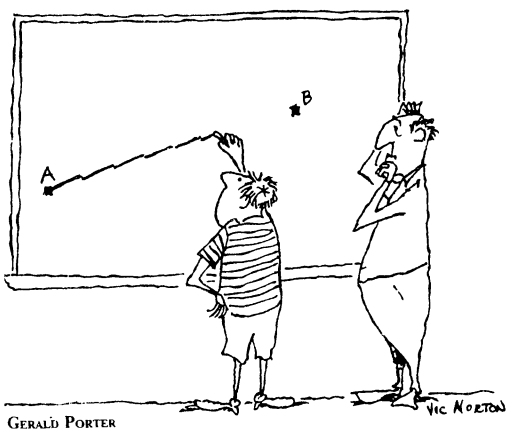
CUPM Panel on Teacher Training, Recommendations on the Mathematical Preparation of Teachers, MAA, 1983; ii + 76 pp, (P).

New sets of course recommendations for elementary, junior high, and high school mathematics teachers. The minimum recommended for high school teachers includes 9 specified courses apart from calculus and linear algebra, for a total of more than 40 semester hours. This amounts to one more course than in the 1971 recommendations, and two more than the 1961 recommendations.

NEWS & LETTERS

THE WINNERS!

The five best (in the editor's opinion) captions to the cartoon without caption in this *Magazine*, January 1984, p. 40, sent in by readers, are published below. Competition was keen--thanks for your entries!



GERALD PORTER

Forget about computer illiteracy! What about math ignorance?
-- David A. Rose

The boy has a computer but never heard of a straightedge.. -- Jim Madison

...and this would be the shortest distance between two pixels.
-- Howard Anton

The human element tries to duplicate computer technology. -- Mark Anderson

I saw it before -- on ancient clay tablets. -- faculty of Drake Univ.

$100,000 CHALLENGE

The Fredkin Foundation of Boston has issued a challenge: it will award $100,000 to the first person who can program a computer to make a major mathematical discovery. For details, contact Woodrow W. Bledsoe (Texas at Austin).

MATHEMATICIANS HONOR WOMEN

At the annual meeting in Louisville in January, 1984, the Mathematical Association of America announced a special Citation honoring those who have furthered the progress of mathematics by enhancing significantly the status of women in mathematics. The full text of the Citation appears in the March-April issue of FOCUS, the MAA newsletter.

SHORT COURSE IN COMPUTATIONAL COMPLEXITY

The sixth Annual Short Course, sponsored by the Northeast Section of the MAA and the University of Maine, will be held June 11-15, 1984 at the University of Maine, Orono, Maine. The week-long course, with principal lecturer Victor Klee, will consist of ten lectures on some of the combinatorial and geometric aspects of optimization. Cost (including course fee, room, and board): $175.00.

For more information contact Grattan Murphy, Dept of Math, University of Maine at Orono, Orono, Maine 04469 or Don Small, Dept of Math, Colby College, Waterville, Maine 04901.

CONFERENCE ON MATH CURRICULA

The Twelfth Annual Mathematics and Statistics Conference at Miami University, Oxford, Ohio, will be held September 28-29, 1984. The conference title is "Mathematics Curricula: Crisis Intervention". Speakers will include Peter Lax, Courant Institute; Arthur Coxford, University of Michigan; Anthony Ralston, SUNY at Buffalo; and John Saxon, Rose State College (Oklahoma). Contributed papers relating to the general theme and appropriate for a general audience of mathematicians are welcome. Topics might include: curricular innovations, articulation between secondary schools and colleges, teacher training and certification. Send abstracts by June 1, 1984 to:

David Kullman, Department of Mathematics and Statistics, Miami University, Oxford, Ohio 45056. Information on preregistration, housing, etc., will be available after July 15.

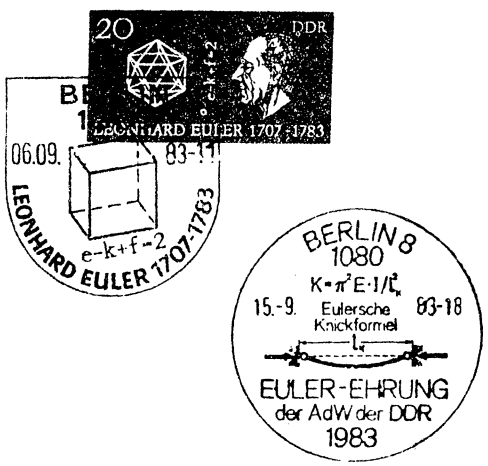
The Ohio Delta Chapter of Pi Mu Epsilon will also hold its Eleventh Annual Student Conference September 28-29, 1984. Undergraduate and graduate students are invited to contribute 15-minute papers, and should send abstracts to Milton Cox, Department of Mathematics and Statistics, Miami University, Oxford, Ohio 45056, by September 20. Lodging at no cost is available for students.

PHILATELISTS TAKE NOTE

To mark the passing of 200 years since the death of Leonhard Euler, and to celebrate Euler's life in Berlin, an East German (DDR) postage stamp and special postmarks for Berlin were issued in 1983. The postmarks display two of Euler's famous formulas (see this *Magazine*, November 1983, pages given):

$e + k + f = 2$ (p. 324) and

$K = \pi^2 EI/l^2$ (p. 321, note missing =).



1983 PUTNAM: SOLUTIONS

The following solutions to the 1983 W.L. Putnam competition problems were prepared for publication in this Magazine *by Loren Larson and Bruce Hanson, with the assistance of the St. Olaf College Problem Solving group.*

A-1. How many positive integers n are there such that n is an exact divisor of at least one of the numbers

$$10^{40},\ 20^{30}\ ?$$

Sol. Let A denote the set of positive integer divisors of 10^{40} $(= 2^{40}\ 5^{40})$ and B the set of positive integer divisors of 20^{30} $(= 2^{60}\ 5^{30})$. By unique factorization of integers, we can compute,

$$|A \cup B| = |A| + |B| - |A \cap B|$$
$$= 41^2 + 61\times 31 - 41\times 31 = 2301.$$

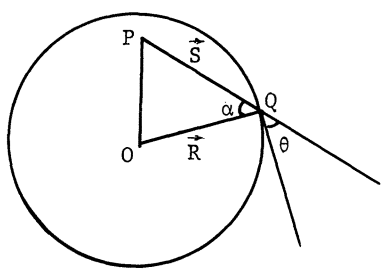
A-2. The hands of an accurate clock have lengths 3 and 4. Find the distance between the tips of the hands when that distance is increasing most rapidly.

Sol. I. The square of the distance between the tips of the hands can be expressed in terms of the cosine of the angle between the hands (using the law of cosines). It is a straightforward calculus problem to find the rate of increase of this distance (first derivative), and to maximize it by setting its derivative (the second derivative) to zero. It turns out that the second derivative is zero when the cosine of the angle between the hands is 3/4, and at this angle the distance between the tips is $\sqrt{7}$.

Sol. II: Based on vector calculus. Assume the clock is centered at O and the small hand is fixed in a vertical position. Let P and Q denote the tip of the short and long hands respectively; let

$$\vec{S} = \overrightarrow{PQ} \text{ and } \vec{R} = \overrightarrow{OQ}.$$

We wish to maximize $\frac{d|\vec{S}|}{dt}$. Let θ be the angle between $\vec{S}$ and $\frac{d\vec{S}}{dt}$ and $\alpha = \angle PQO$.



We know that $\frac{d|\vec{S}|}{dt} = \left|\frac{d\vec{S}}{dt}\right| \cos\theta$.

But $\left|\frac{d\vec{S}}{dt}\right| = \left|\frac{d\vec{R}}{dt}\right| = K$, a constant, and

$\theta = \pi - \alpha - \pi/2 = \pi/2 - \alpha$,
and therefore

$$\frac{d|\vec{S}|}{dt} = K\cos(\pi/2 - \alpha) = K \sin\alpha.$$

We now have a geometry problem: Where should Q be located (on the circle) to maximize α? By the law of sines,

$$\sin\alpha = \left|\frac{OP}{OQ}\right| \sin \angle OPQ,$$

and this is maximized when $\angle OPQ = \pi/2$. At this position,

$$|PQ| = \sqrt{16 - 9} = \sqrt{7}.$$

A-3. Let p be in the set {3, 5, 7, 11, ...} of odd primes and let

$$F(n) = 1 + 2n + 3n^2 + \ldots + (p-1)n^{p-2}.$$

Prove that if a and b are distinct integers in {0, 1, 2, ... , p - 1} then $F(a)$ and $F(b)$ are not congruent modulo p, that is $F(a) - F(b)$ is not exactly divisible by p.

Sol. Throughout the solution, the $\equiv$ symbol means "congruent modulo p". First note that $F(0) \equiv 1$ and $F(1) \equiv 0$. Suppose that $1 < a < p$. We will show that $(1-a)F(a) \equiv 1$. This will imply that $F(a)$ is not equal to $F(0)$ or $F(1)$.

We begin by finding a closed formula for $F(a)$. Differentiate each side of the identity

$$x + x^2 + \ldots + x^{p-1} = \frac{x^p - x}{x-1}, \quad x \neq 1,$$

to get

$$1 + 2x + \ldots + (p-1)x^{p-2} = \frac{(p-1)x^p - px^{p-1} + 1}{(x-1)^2}.$$

Now cross-multiply and replace x by a to get

$$\begin{aligned}(1-a)^2F(a) &\equiv (p-1)a^p - pa^{p-1} + 1\\ &\equiv (p-1)a + 1\\ &\equiv 1 - a,\end{aligned}$$

where we have used the fact that $a^p \equiv a$. Since $1-a$ is relatively prime to p, the last equation implies that $(1-a)F(a) \equiv 1$, as claimed.

Now suppose that $a \neq b$, $1 < a, b < p$. We have just seen that $(1-a)F(a) \equiv 1$ and $(1-b)F(b) \equiv 1$. If $F(a) \equiv F(b)$, the last equations would imply that $1-a \equiv 1-b$, or equivalently, that $a \equiv b$. This contradiction implies that $F(a) \not\equiv F(b)$ and the solution is complete.

A-4. Let k be a positive integer and let $m = 6k - 1$. Let

$$S(m) = \sum_{j=1}^{2k-1} (-1)^{j+1} \binom{m}{3j-1}.$$

For example with $k = 3$,

$$S(17) = \binom{17}{2} - \binom{17}{5} + \binom{17}{8} - \binom{17}{11} + \binom{17}{14}.$$

Prove that $S(m)$ is never zero.

As usual, $\binom{m}{r} = \frac{m}{r!(m-r)!}$.

Sol.
$$S(m) - 1 = \sum_{j=1}^{2k} (-1)^{j+1}\binom{m}{3j-1}$$

$$= \sum_{j=1}^{k} (-1)^{j+1}\left[\binom{m}{3j-1} - \binom{m}{m-3(j-1)}\right]$$

$$= \sum_{j=1}^{k} (-1)^{j+1}\left[\binom{m}{3j-1} - \binom{m}{3j-3}\right].$$

Now

$$\binom{m}{3j-1} = \binom{m}{3j-3}\frac{(m-3j+2)(m-3j+3)}{(3j-1)(3j-2)}$$

$$= \binom{m}{3j-3}\frac{(6k-3j+1)(6k-3j+2)}{(3j-1)(3j-2)},$$

and it follows that

$$\binom{m}{3j-1} \equiv \binom{m}{3j-3} \bmod 3.$$

Therefore $S(m) - 1 \equiv 0 \bmod 3$, and the result follows.

A-5. Prove or disprove that there exists a positive real number u such that $[u^n] - n$ is an even integer for all positive integers n.

Here $[x]$ denotes the greatest integer less than or equal to x.

Sol. Let $a_1 = 3$. Define a_n recursively as follows:

if $n+1$ is odd, a_{n+1} is the smallest odd integer $\geq a_n^{(n+1)/n}$;

if $n+1$ is even, a_{n+1} is the smallest even integer $\geq a_n^{(n+1)/n}$.

Then $\{a_n\}$ is an increasing sequence of positive integers such that

$$a_n^{(n+1)/n} \leq a_{n+1} \leq a_n^{(n+1)/n} + 2. \quad (*)$$

An easy induction shows that $a_n \geq 3^n$ for all n. Let

$$b_n = a_n^{1/n}, \quad c_n = (a_n+1)^{1/n},$$

and let I_n denote the interval $[b_n, c_n)$. Notice that for each $x \in I_n$, $[x^n] - n = a_n - n$ is even. We claim that $\overline{I}_{n+1} \subseteq I_n$ for all n. It is clear from (*) that $b_n \leq b_{n+1}$. To show that $c_n > c_{n+1}$, we must show that $(a_n+1)^{1/n} > (a_{n+1}+1)^{1/(n+1)}$, or equivalently, that $(a_n+1)^{(n+1)/n} > a_{n+1}+1$. By (*), it suffices to show that

$$(a_n+1)^{(n+1)/n} - a_n^{(n+1)/n} > 3.$$

By the mean value theorem,

$$(a_n+1)^{(n+1)/n} - a_n^{(n+1)/n} = \left(\frac{n+1}{n}\right) c^{1/n}$$

for some c, $a_n < c < a_n + 1$. Hence

$$(a_n+1)^{(n+1)/n} - a_n^{(n+1)/n} = \left(\frac{n+1}{n}\right)c^{1/n} > \left(\frac{n+1}{n}\right)a_n^{1/n} > 3$$

(the last inequality by an earlier observation). This completes the proof of the claim.

By the nested interval property and the claim of the last paragraph,

$$\bigcap_{n=1}^{\infty} I_n \neq \emptyset .$$

Let $u \in \bigcap_{n=1}^{\infty} I_n$.

By an earlier observation, $[u^n] - n$ is even for all n.

A-6. Let $\exp(t)$ denote e^t and

$$F(x) = \frac{x^4}{\exp(x^3)} \int_0^x \int_0^{x-u} \exp(u^3+v^3)\, dv\, du.$$

Find $\lim_{x\to\infty} F(x)$ or prove that it does not exist.

Sol. Let $u = t$ and $v = s-t$. Then

$$F(x) \lim_{x\to\infty} \frac{\int_0^x \int_0^s \exp(t^3 + (s-t)^3)\, dt\, ds.}{\exp(x^3)/x^4}$$

By L'Hopital's rule, this is

$$= \lim_{x\to\infty} \frac{x^5 \int_0^x \exp(t^3 + (x-t)^3)\, dt}{(3x^3-4)\exp(x^3)} .$$

Now let $w = t - x/2$ and write this in the form

$$= \lim_{x\to\infty} \frac{x^5 \int_{-x/2}^{x/2} \exp[(x/2+w)^3+(x/2-w)^3]\,dw}{(3x^3-4)\exp(x^3)}$$

$$= \lim_{x\to\infty} \frac{2x^5 \int_0^{x/2} \exp(3xw^2)\, dw}{(3x^3-4)\exp(3x^3/4)} .$$

Let $y = \sqrt{x}w$, and this is

$$= \lim_{x\to\infty} \frac{2\int_0^{x^{3/2}/2} \exp(3y^2)\, dy}{(3x^3-4)\exp(3x^3/4)/x^{9/2}} .$$

Another application of L'Hopital's rule shows this is

$$= \lim_{x\to\infty} \frac{\frac{3}{2}x^{19/2}}{\frac{27}{4}x^{19/2} - \frac{27}{2}x^{13/2} + 18x^{7/2}}$$

$$= 2/9.$$

B-1. Let v be a vertex (corner) of a cube C with edges of length 4. Let S be the largest sphere that can be inscribed in C. Let R be the region consisting of all points p between S and C such that p is closer to v than to any other vertex of the cube. Find the volume of R.

Sol. The desired volume is one-eighth of the volume bounded between the sphere and the cube, namely

$$\frac{1}{8}\left(4^3 - \frac{4}{3}\pi 2^3\right) = 8 - 4\pi/3.$$

B-2. For positive integers n, let $C(n)$ be the number of representations of n as a sum of nonincreasing powers of 2, where no power can be used more than three times. For example, $C(8) = 5$ since 8 is represented:

8, 4+4, 4+2+2, 4+2+1+1, and 2+2+2+1+1.

Prove or disprove that there is a polynomial $P(x)$ such that $C(n) = [P(n)]$ for all positive integers n; here $[u]$ denotes the greatest integer less than or equal to u.

Sol. An obvious correspondence shows that $C(2n+1) = C(2n)$ for each positive integer n (add "1" to each representation of $2n$).

Let $A(n)$ and $B(n)$ denote the number of representations of n, of the desired type, which contain no 1's, or exactly two 1's, respectively. We will use induction to show that

$$A(4n)=n+1, \quad B(4n)=n \qquad n=1,2,3, \ldots$$

$$A(4n+2)=n+1, \quad B(4n+2)=n+1, \quad n=0,1,2, \ldots$$

It is easy to check that $A(2) = B(2) = 1$, $A(4) = 2$, $B(4) = 1$, $A(6) = 2$, $B(6) = 2$. So assume we have proved the formulas for all positive integers less than n, and that $n > 1$.

Notice that $A(4n) = C(2n)$ (multiply the terms of each representation of $2n$ by 2). We claim that $C(2n) = n + 1$. There are two cases to consider. If n is even, say $n = 2m$, we have $C(2n) = C(4m) = A(4m) + B(4m) = (m + 1) + m$ (by the inductive assumption) $= 2m + 1 = n + 1$. If n is odd, say $n = 2m + 1$, $C(2n) = C(4m + 2) = A(4m + 2) + B(4m + 2) = (m + 1) + (m + 1) = n + 1$. It follows that $A(4n) = C(2n) = n + 1$.

Also, $B(4n) = A(4n - 2)$ (add two 1's to those representations of $4n - 2$ which contain no 1's), and by induction, $A(4n - 2) = A(4(n - 1)+2) = n$. Similarly, $A(4n + 2) = C(2n + 1) = C(2n) = n + 1$, and $B(4n + 2) = A(4n) = n + 1$. This completes the induction.

The preceding work shows that $C(n) = [n/2 + 1]$ for each positive integer n, so there is such a polynomial, namely $P(x) = x/2 + 1$.

B-3. Assume that the differential equation

$$y''' + p(x)y'' + q(x)y' + r(x)y = 0$$

has solutions $y_1(x)$, $y_2(x)$, and $y_3(x)$ on the whole real line such that

$$y_1^2(x) + y_2^2(x) + y_3^2(x) = 1$$

for all real x. Let

$$f(x) = (y_1'(x))^2 + (y_2'(x))^2 + (y_3'(x))^2.$$

Find constants A and B such that $f(x)$ is a solution to the differential equation

$$y' + Ap(x)y = Br(x).$$

Sol. The following three identities result from taking the first, second, and third derivative, respectively, of $y_1^2 + y_2^2 + y_3^2 = 1$:

$$y_1y_1' + y_2y_2' + y_3y_3' = 0, \tag{1}$$

$$(y_1')^2 + (y_2')^2 + (y_3')^2 = -(y_1y_1'' + y_2y_2'' + y_3y_3''), \tag{2}$$

$$3(y_1'y_1'' + y_2'y_2'' + y_3'y_3'') = -(y_1y_1''' + y_2y_2''' + y_3y_3'''). \tag{3}$$

We wish to find constants A and B such that

$$[(y_1')^2 + (y_2')^2 + (y_3')^2]' + Ap\,[(y_1')^2 + (y_2')^2 + (y_3')^2] = Br,$$

or equivalently,

$$2(y_1'y_1'' + y_2'y_2'' + y_3'y_3'') + Ap\,[(y_1')^2 + (y_2')^2 + (y_3')^2] = Br.$$

Using equations (2) and (3), this equation is

$$-\frac{2}{3}(y_1y_1''' + y_2y_2''' + y_3y_3''') + Ap\,[-(y_1y_1'' + y_2y_2'' + y_3y_3'')] = Br,$$

or equivalently,

$$-y_1\left(\frac{2}{3}y_1''' + Apy_1''\right) - y_2\left(\frac{2}{3}y_2''' + Apy_2''\right) - y_3\left(\frac{2}{3}y_3''' + Apy_3''\right) = Br.$$

Now use the fact that y_1, y_2, and y_3 satisfy $y''' = -ry - qy' - py''$. These substitutions yield

$$\frac{2}{3}r(y_1^2 + y_2^2 + y_3^2) + \frac{2}{3}q(y_1y_1' + y_2y_2' + y_3y_3') + \frac{2}{3}p(y_1y_1'' + y_2y_2'' + y_3y_3'') - Ap(y_1y_1'' + y_2y_2'' + y_3y_3'') = Br.$$

By (1) and our condition, this is

$$\frac{2}{3}r + p(\frac{2}{3} - A)(y_1y_1'' + y_2y_2'' + y_3y_3'') = Br,$$

and from this we see that $f(x)$ will satisfy the equation if $A = 2/3$ and $B = 2/3$.

B-4. Let $f(n) = n + [\sqrt{n}]$ where $[x]$ is the largest integer less than or equal to x. Prove that, for every positive integer m, the sequence

$$m,\ f(m),\ f(f(m)),\ f(f(f(m))),\ \ldots$$

contains at least one square of an integer.

Sol. If m is not already a perfect square, it will fall between two perfect squares, say

$$n^2 < m < (n+1)^2 .$$

Suppose that $m = n^2 + k$ where $0 < k < n+1$. Then $f(m) = n^2 + n + k$ and $f^{(2)} = f(f(m)) = n^2 + n + k + n = (n+1)^2 + (k-1)$. If $(k-1) \neq 0$, a similar argument shows that $f^{(4)}(m) = (n+2)^2 + (k-2)$. By repeating this reasoning k times, we find that $f^{(2k)}(m) = (n+k)^2$, so the result holds.

Suppose that $m = n^2 + k$ where $n+1 \leq k < 2n+1$. Then $f(m) = n^2 + k + n$. But $(n+1)^2 = n^2 + 2n + 1 = (n^2 + n) + (n + 1) \leq n^2 + n + k = f(m) < n^2 + n + 2n+1 = (n+1)^2 + n$. Thus $f(m)$ has the form $f(m) = (n+1)^2 + s$, where $0 \leq s < n+2$. But now the argument of the preceding paragraph applies and we see that $f^{(2s+1)}(m) = (n+1+s)^2$. This completes the proof.

B-5. Let $||u||$ denote the distance from the real number u to the nearest integer. For positive integers n, let

$$a_n = \frac{1}{n}\int_1^n \left|\left|\frac{n}{x}\right|\right| dx .$$

Determine $\lim_{n\to\infty} a_n$. You may assume the identity

$$\frac{2}{1}\cdot\frac{2}{3}\cdot\frac{4}{3}\cdot\frac{4}{5}\cdot\frac{6}{5}\cdot\frac{6}{7}\cdot\frac{8}{7}\cdot\frac{8}{9} \cdots = \pi/2.$$

Sol.

$$\int_1^n \left|\left|\frac{n}{x}\right|\right| dx = \sum_{k=2}^{n} \int_{n/k}^{n/(k-1)} \left|\left|\frac{n}{x}\right|\right| dx$$

$$= \sum_{k=2}^{n} \left[\int_{n/k}^{2n/(2k-1)} (k - \frac{n}{x})\, dx + \int_{2n/(2k-1)}^{n/(k-1)} (\frac{n}{x} - (k-1))\, dx\right]$$

$$= n \sum_{k=2}^{n} \ln \frac{(2k-1)(2k-1)}{(2k-2)(2k)}$$

$$= n \ln \frac{3}{2}\frac{3}{4}\frac{5}{4}\frac{5}{6}\cdots\frac{(2n-1)(2n-1)}{(2n-2)(2n)} .$$

It follows that $\lim_{n\to\infty} a_n$

$$= \ln\left[2 \lim_{n\to\infty} \frac{1}{2}\frac{3}{2}\frac{3}{4}\frac{5}{4}\cdots\left(\frac{2n-1}{2n-2}\right)\left(\frac{2n-1}{2n}\right)\right]$$

$$= \ln (4/\pi).$$

B-6. Let k be a positive integer, let $m = 2^k+1$, and let $r \neq 1$ be a complex root of $z^m - 1 = 0$. Prove that there exist polynomials $P(z)$ and $Q(z)$ with integer coefficients such that

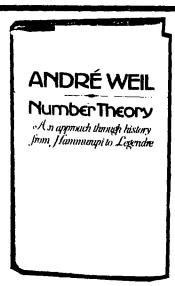
$$(P(r))^2 + (Q(r))^2 = -1.$$

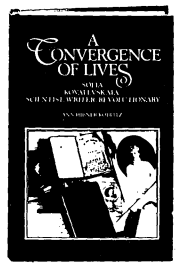
Sol. From the assumption that r is a complex root of $z^m - 1$ $(= (z-1)(z^{m-1} + z^{m-2} + \ldots + z + 1))$, it follows that

$$\begin{aligned}
-1 &= (r+r^2) + (r^3+r^4) + \ldots + (r^{m-2}+r^{m-1}) \\
&= (r^{m+1}+r^2) + (r^{m+3}+r^4) + \ldots \\
&\quad + (r^{2m-2}+r^{m-1}) \\
&= (r^2(1+r^{m-1}) + r^4(1+r^{m-1}) + \ldots \\
&\quad + r^{m-1}(1+r^{m-1}) \\
&= (r^2+r^4+\ldots+r^{m-3}+r^{m-1})(1+r^{m-1}) \\
&= r^2+r^4+\ldots+r^{2^n-2}+r^{2^n})(1+r^{2^n}) \\
&= r^2(1+r^2)(1+r^4)\ldots(1+r^{2^n-1})(1+r^{2^n}).
\end{aligned}$$

The result follows after repeated use of the fact that the product of two sums of two squares is again a sum of two squares: $(f^2 + g^2)(h^2 + k^2) = (fh + gk)^2 + (gh - fk)^2$.

MATHEMATICAL PERSONALITIES





Number Theory
An approach through history from Hammurapi to Legendre
André Weil

Written by one of the foremost mathematicians of this century, this is a history of the oldest, yet most exciting and topical branch of modern mathematics. Andre Weil examines thirty-six centuries of number theory from ancient Babylonia to the workshops of Fermat, Euler, Lagrange, and Legendre.

Jan. 1984 / 384 pp./ Hardcover / $24.95 / O-8176-3141-O

A Convergence of Lives
Sofia Kovalevskaia, Scientist, Writer, Revolutionary
Ann Hibner Koblitz

This fascinating biography of pioneering female scientist Sofia Kovalevskaia examines her contributions to her field and her reception by prominent scientists, writers and political figures of her day — Kropotkin, Dostoevsky, Turgenev, Chekhov, Darwin and others.

Dec. 1983 / 304 pp. / Hardcover / $19.95 / O-8176-3162-3

Mathematical People
Profiles and Interviews
edited by Donald J. Albers and G.L. Alexanderson

Entertaining and in-depth interviews with extraordinary people, including Paul Erdős, Ronald Graham, Donald Knuth and Olga Taussky-Todd. This book uncovers the motivations and personal decisions that led to the great mathematical advances in this century. Written in collaboration with The Mathematical Association of America.

July 1984 / ca. 400 pages / Hardcover / $24.95 / O-8176-3191-7

The Thread
A Mathematical Yarn
Phil Davis

"Philip Davis has raised Digression to a literary form. It is high time, for this book shows Digression is the thread of history." —Gerard Piel, Scientific American.

Second printing

1983 / 112 pp. / Hardcover / $12.95 / 3-7643-3097-X

ORDER FORM

Yes, I would like to read about these fascinating mathematical people. Please send me the following book(s):

copies	title	ISBN	price
_____	Convergence of Lives	3162-3	$19.95
_____	Number Theory	3141-O	$24.95
_____	Mathematical People	3191-7	$24.95 (Coming in July)
_____	The Thread	3097-X	$12.95

☐ bill me ☐ payment enclosed $ _____
☐ VISA ☐ MasterCard
card #: _____ exp. date: _____

Please include $1 for shipping and handling for one book, $2 for shipping and handling for two or more books.
Massachusetts residents pleaase add 5% sales tax.

BIRKHÄUSER BOSTON, INC.

signature: _____

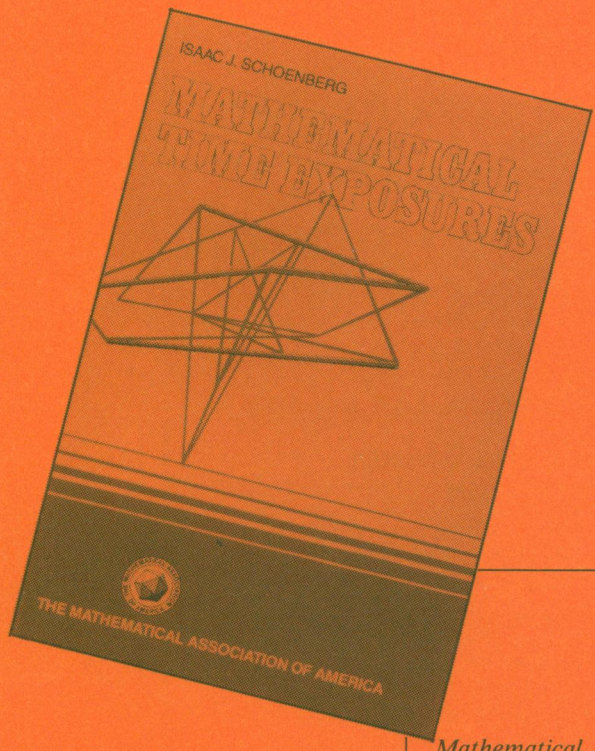
(please print)
bill to: _____

address: _____

city/state/zip: _____

Send to:
BIRKHÄUSER BOSTON, INC.
P.O. Box 3005 Cambridge, MA 02139

New . . .



Available in Hardcover and Paper Editions

279 pages
Hardcover edition:
List: $30.00
MAA Member $22.50

Paper Edition
List: $18.00
MAA Member: $13.50

Mathematical Time Exposures was inspired by Hugo Steinhaus' admirable book, *Mathematical Snapshots,* published in 1938. The title, *Mathematical Time Exposures* was also suggested by photography, but Schoenberg's pace is much more leisurely than Steinhaus'. Schoenberg spends more time on fewer subjects—the "snapshots" become "time exposures." The subject of at least two of the chapters actually antedate the invention of the daguerreotype.

The author manages to bring together concepts from geometry, number theory, algebra, and analysis, frequently mixing them together in the same chapter. The arts are not neglected. Discussions on the tuning of keyboard instruments, the guitar, and the vibrations of strings are discussed, as well as the suggestion of rectilinear models for outdoor sculpture.

Many of the chapters in this book may be read independently, and at least half of them with only a knowledge of precalculus mathematics.

Order From:
The Mathematical Association of America
1529 Eighteenth Street, N.W.
Washington, D.C. 20036

THE MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth Street, N.W.
Washington, DC 20036

MATHEMATICS MAGAZINE VOL. 57, NO. 3, MAY 1984