Vol. 62, No. 1 February 1989

MATHEMATICS MAGAZINE

- Brouwerian Counterexamples
- Fields with Simple Binomial Theorem
- The Shape of *e*

An Official Publication of The MATHEMATICAL ASSOCIATION OF AMERICA

EDITORIAL POLICY

The aim of *Mathematics Magazine* is to provide lively and appealing mathematical exposition. This is not a research journal and, in general, the terse style appropriate for such a journal (lemma-theorem-proof-corollary) is not appropriate for an article for the *Magazine*. Articles should include examples, applications, historical background, and illustrations, where appropriate. They should be attractive and accessible to undergraduates and would, ideally, be helpful in supplementing undergraduate courses or in stimulating student investigations. Articles on pedagogy alone, unaccompanied by interesting mathematics, are not suitable. Neither are articles consisting mainly of computer programs unless these are essential to the presentation of some good mathematics. Manuscripts on history are especially welcome, as are those showing relationships between various branches of mathematics and between mathematics and other disciplines.

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

Send new manuscripts to: G. L. Alexanderson, Editor, Mathematics Magazine, Santa Clara University, Santa Clara, CA 95053. Manuscripts should be typewritten and double spaced and prepared in a style consistent with the format of *Mathematics Magazine*. Authors should submit the original and one copy and keep one copy. Illustrations should be carefully prepared on separate sheets in black ink, the original without lettering and two copies with lettering added.

AUTHORS

Mark Mandelkern ("Brouwerian Counterexamples") completed his schooling at Rochester in 1966. His dissertation could not have been written without the kindly encouragement of Leonard Gillman.

His work in constructive mathematics began in 1970. His interest in the subject extends beyond its intrinsic appeal and its significance for the foundations of mathematics. In education mathematics has always played a key role in forming the capacity for clear thinking. He believes that the constructive approach, with its fine distinctions and its particular emphasis on meaning, may serve as a model for the even sharper deductive structures required for dealing with problems in the modern world.

Cover. L. E. J. Brouwer. Photograph on cover and on p. 5 courtesy of the Brouwer Archive.

Acknowledgement. The quotation on p. 57 was kindly provided by Ralph P. Boas.

Vol. 62, No. 1 February 1989

MATHEMATICS MAGAZINE

The *MATHEMATICS MAGAZINE* (ISSN 0025–570X) is published by the Mathematical Association of America at 1529 Eighteenth Street, N.W., Washington, D.C. 20036 and Montpelier, VT, bimonthly except July/August.

The annual subscription price for the MATHEMATICS MAGAZINE to an individual member of the Association is $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 nonmember/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. Microfilmed issues may be obtained from University Microfilms International, Serials Bid Coordinator, 300 North Zeeb Road, Ann Arbor, MI 48106.

Advertising correspondence should be addressed to Ms. Elaine Pedreira, Advertising Manager, The Mathematical Association of America, 1529 Eighteenth Street, N.W., Washington, D.C. 20036.

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

Postmaster: Send address changes to Mathematics Magazine Membership/Subscriptions Department, Mathematical Association of America, 1529 Eighteenth Street, N.W., Washington, D.C. 20077-9564.

PRINTED IN THE UNITED STATES OF AMERICA

ARTICLES

Brouwerian Counterexamples

An 80-year-old but little-known method demonstrates the lack of numerical meaning in many classical theorems.

MARK MANDELKERN
New Mexico State University
Las Cruces, NM 88003

Not only did he launch a controversy which continues to the present day, but in his critique of classical mathematics early in the century, L. E. J. Brouwer also initiated a new mode of reasoning. In contrast to the idealistic thought common since Greek times, characterized by the notion that truth exists independently of humans, Brouwer realistically held as true only what was currently known. Thus truth changes from day to day, and also from person to person. Brouwer developed a type of counterexample which shows when a given statement is not true in this realistic sense. A recent paper [**32**] discussed, with no technical details, the general aspects of the controversy over constructivity which ensued from Brouwer's work. This paper considers the technical details involved in Brouwerian counterexamples; it tries to steer clear of polemics, leaving readers to decide for themselves whether or not the counterexamples indicate a need for constructive considerations in the practice of mathematics.

The term *classical mathematics*, as used here, refers to the sort of mathematics taught in virtually every school and college classroom in the world. Work in *constructive mathematics*, using only constructive methods, is (at least for the present) carried on by only a very small minority of mathematicians.

There is a certain danger in devoting an entire paper to negativistic results, which could give the erroneous impression that the purpose of constructive mathematics is to consider pathological counterexamples (see Appendix). The only reason for these counterexamples is to show that certain theorems in classical mathematics are not constructively valid, and thus to indicate the need for their replacement by positive constructive results. A few of these positive results are indicated here, but for a more complete exposition the reader must consult the literature; for a thorough introduction, see [**3**] or [**6**].

A new kind of counterexample Brouwer's critique of many classical theorems, his claim that they lacked numerical meaning, consisted of demonstrations showing that their truth would imply solutions to problems for which, in fact, no solutions were known. Thus he concluded that certain classical theorems could not be true, for if they were true, many people would be trying to use them to solve the unsolved problems. Such a demonstration is called a *Brouwerian counterexample*; it differs from an ordinary counterexample, which demonstrates that a given statement implies another statement which is known to be false.

Brouwer's examples were typically constructed on an *ad hoc* basis, using a variety of unsolved problems more or less at random—sometimes quite famous problems such as "Fermat's Last Theorem," and sometimes problems remarkable only for their insignificance, such as questions concerning the digits in the decimal expansion of π. More systematically, Errett Bishop [3] formulated several general *omniscience principles*, each of which implies the solution to a vast number of unsolved problems, or at least leads to certain information about such problems which in fact is not available. A Brouwerian counterexample then becomes a demonstration that a certain statement implies one of these omniscience principles. One advantage of this is that it shows more clearly the nonconstructivities in classical mathematics, since there will always be unsolved problems which the omniscience principles would solve. Here we first give a few of Brouwer's original *ad hoc* examples, and then develop the systematic formulation.

The least upper bound principle A characteristic and important theorem of classical mathematics is the *least upper bound principle*. To show its nonconstructive nature (its lack of numerical meaning), we'll show that if it were true, it would provide a finite method leading either to a proof of Fermat's Last "Theorem" or to an explicit counterexample. Since no one knows such a method, no one can claim that the least upper bound principle is true in a numerical sense. Fermat probably had no proof (see [**42**, Ch. 13] and [**44**]), but if he did, then for him it was a "theorem", while we can speak only of "Fermat's Last Problem."

Since Fermat's Last Problem concerns equations in integers, while the least upper bound principle concerns sets of real numbers, we must somehow represent the former by the latter. We'll construct a set S of real numbers by a sequence of steps, as follows. We start with step number 3. Using only positive integers up to and including 3, we look for a solution to the equation $x^n + y^n = z^n$, with $n = 3$. Finding no solution, we put the real number $1/3$ into our set S. (Even when any integers x, y, z are allowed, there is no solution with $n = 3$; see, for example [**7**, Th. 44.9].) If we had found a solution, we would have put the number $1 - 1/3$ into S. This process for constructing S continues; at the kth step we consider all quadruples (x, y, z, n) of positive integers up to k, with the restriction $n \geqslant 3$. When we find no solution to the Fermat equation, we put $1/k$ into S; when we do find a solution, we put $1 - 1/k$ into S. This defines the set S; see FIGURE 2. (There are many ways to form a suitable set; there is nothing special about the method used here.) Notice that the procedure at the kth step is always a finite one, although the whole process is infinite.

Now we apply the *hypothesis* that the least upper bound principle is true. We take this assertion in the numerical sense that we can calculate least upper bounds; that is, we can find explicit, arbitrarily close, rational approximations. Let t be the least upper bound of the set S. For example, someone might specify the number t using a decimal expansion. While such an expansion is infinite, only a finite process is required to calculate any given digit. We need calculate only one digit to tell whether t is less than 0.6 or more than 0.4. In the first case, it is easy to predict with absolute certainty that the Fermat equation will never be solved (with $n \geqslant 3$); you have proved Fermat's Last Theorem. In the second case, by the definition of least upper bound there is a number x in the set S that is more than 0.4. Now this number x must be of the form $1 - 1/k$. Looking at this integer k, you know exactly where to find, using a finite process, a solution to the Fermat equation, a counterexample to Fermat's Last Theorem. In either case you have solved Fermat's Last Problem. It is convenient to restate this counterexample concisely as follows.

Example 1. The *least upper bound principle* is nonconstructive; it would imply a solution to Fermat's Last Problem.

How do constructivists answer the classicist who (omnisciently) looks at Example 1

FIGURE 1
L. E. J. Brouwer (1881–1966). This article commemorates the 80th anniversary of Brouwer's seminal doctoral thesis.

and says, "Well, it's perfectly obvious that the least upper bound of the set S is either $1/3$ or 1." Their answer is, "If you could really tell which one of these alternatives actually holds, you wouldn't be here discussing it with us. You would be at your desk writing either your proof of Fermat's Last Theorem or your proof that you have a finite procedure for finding a counterexample." Saying that a number with certain properties, *if it exists*, must be equal to one or the other of two known numbers, is quite different from saying that such a number actually exists. When a constructivist

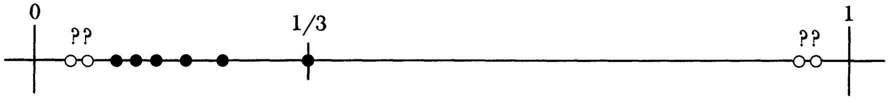


FIGURE 2.
Here are a few of the points, and possible points, of the set S constructed in Example 1 by searching for solutions of the Fermat equation. The points begin at $1/3$ and proceed to the left. However, if a solution is ever found, then points suddenly begin to appear at the far right of the interval. What is the least upper bound of S?

says that a number is either this or that, he is prepared to say which; or at least to give a *finite* procedure by which it may be determined.

Thus a person who really had a constructive proof of the least upper bound principle, a finite procedure for calculating rational approximations to the least upper bound, would immediately begin applying it to the set S to find a solution to Fermat's Last Problem, and similarly for hundreds of other unsolved problems in number theory and analysis.

How does constructive analysis proceed without the important least upper bound principle? By restricting to a certain class of sets, which suffice for all constructive applications, one obtains a constructive substitute; see Theorem 3 below.

Counterexamples such as the above are the basis of Brouwer's critique of classical mathematics, begun in 1907 [12], and the motivation behind modern Bishop-type constructive mathematics [3]. In the following sections we consider a few more informal counterexamples, and then, using these as a guide, adopt a precise formulation. The resulting analysis will eliminate the apparent dependence on specific unsolved problems. Thus, while Example 1 shows that the least upper bound principle is nonconstructive because it implies a solution to Fermat's Last Problem, it remains nonconstructive even if tomorrow somebody solves the problem.

Numerical meaning The general notion of numerical meaning was discussed in [32]; here we try to make this idea more precise. It is the strict constructive notion of numerical meaning which allows us to proceed from the existence of a least upper bound to a solution of Fermat's Last Problem. The least upper bound is a real number; thus we must formulate a precise definition of a constructive real number. This is nothing other than the same real number that is used throughout classical analysis, both elementary and advanced, but with a stricter interpretation. A real number is a Cauchy sequence of rational numbers. Each term of the sequence is a rational number, an explicit quotient of integers. There is no difficulty or controversy about the ordinary integers, their rational quotients, and the finite operations among these. But the idea of an infinite sequence is more difficult. The notion of infinity has had a long and interesting history, and has been discussed by thinkers from the ancient Greeks to modern astronomers. In mathematics there is a sharp contrast between the idea of a potential infinity and that of an actual infinity. This is, perhaps, the crux of the controversy between classical (idealistic) mathematics and constructive (realistic) mathematics. In the latter, only a potential infinity is considered. This means that while the definition of a real number allows you to calculate however many terms of the approximating sequence you want (and have time for), still you can never expect to calculate all the terms.

A real number might be defined by a sequence of approximating rational numbers expressed as finite decimals. For example, $\sqrt{2}$ can be defined by the sequence

$$1, 1.4, 1.41, 1.414, 1.4142, 1.41421, 1.414213, \ldots.$$

This short list of approximations is the result of only the first few applications of the well-known rule which explicitly defines each term of the sequence. A sequence is a function, or rule, rather than an infinite list; the latter has only a potential existence. We will focus attention on such rules. Thus the least upper bound principle (if constructive) must provide a *finite procedure* by which any desired approximation to the least upper bound may be calculated. Example 1 shows that the classical theorem on least upper bounds provides no such procedure. The notion of constructive procedure requires that it be finite only *in principle*. Errett Bishop expressed this as follows. "How do you know whether a proof is constructive? Try to write a computer program. If you can program a computer to do it, it should be constructive. Notice I said write the program. Don't necessarily run it on the computer and wait around for the result." [4]

Trichotomy of real numbers This is a precept of classical mathematics so ingrained in the thought of working mathematicians and students that assaults on its validity are usually met not with rigorous defensive measures, but (even more effectively) with complete disregard. That any given real number is either positive, negative, or zero may seem so familiar an idea as to be intuitively true, but nevertheless it is nonconstructive. To show this, it suffices to show that the trichotomy principle, interpreted so as to have numerical meaning, would imply the solution to an unsolved problem. For the latter we take the *perfect number problem*. A positive integer is *perfect* if it is the sum of its proper divisors, for example $6 = 1 + 2 + 3$, $28 = 1 + 2 + 4 + 7 + 14$, 496, etc.; see [7; 11.25–11.26]. Nobody knows whether or not there exists an odd perfect number. Rather than constructing a set to establish a connection between the principle under test and the unsolved problem, as in the last example, in this case we need a real number. It requires only a finite procedure to test an integer for perfection; the results are used to define a sequence $\{\alpha_k\}$. At the kth step, if no odd perfect number $\leqslant k$ is found, define $\alpha_k = 0$, but if one is found, define $\alpha_k = 1/2^k$. Now define

$$\alpha = \sum_{k=1}^{\infty} \alpha_k.$$

Clearly $\alpha \geqslant 0$; applying the hypothesis that the trichotomy principle has numerical meaning, either $\alpha = 0$ or $\alpha > 0$. In the first case it follows that each α_k is 0; thus we have a theorem, every perfect number is even. In the second case at least one of the terms α_k must be positive, and this leads us straight to an explicit construction of an odd perfect number. (We have used constructive notions of positive real numbers and convergent sequences and series, which are discussed below.) Thus we have proved the following.

Example 2. The *principle of trichotomy of real numbers* is nonconstructive; it would imply a solution to the perfect number problem.

Discontinuous functions It is a common practice in elementary analysis courses to demonstrate the importance of continuity conditions in certain theorems by giving examples of discontinuous functions, showing what can go wrong when continuity is not present. However, there are problems in the construction of such discontinuous functions.

Consider a typical function often mentioned in calculus classes; it is defined on the unit interval [0, 1] by

$$f(0) = 0$$
$$f(x) = 1 \quad \text{whenever } x > 0.$$

This definition presents serious constructive difficulties. As we saw in Example 2, there is no general finite procedure for deciding whether a given real number in the interval is zero or positive. Thus there is no general finite procedure for deciding, for a given point x in the interval, what value f assigns to x. This means that while the definition above does define a function, this function is not defined on the entire interval, but only at those points for which one knows one or the other of the two alternatives, zero or positive. Example 2 constructs a number in the interval for which neither alternative is known.

There is nothing nonconstructive about the definition above, it defines a function which may even be sketched as shown in FIGURE 3, but this function is not defined on the entire unit interval. Thus a Brouwerian counterexample on this topic will concern the following statement.

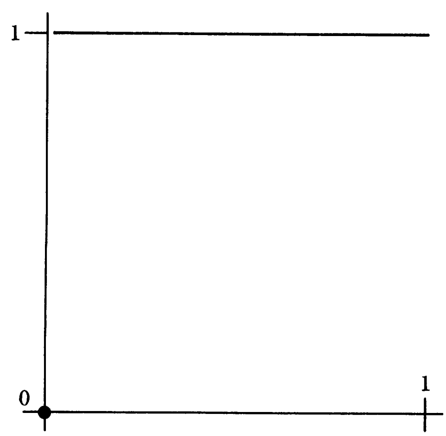


FIGURE 3.
A good try at defining a discontinuous function on the closed unit interval, but it doesn't work. There are points x for which we don't know whether $x = 0$ or $x > 0$; at such points the function is not defined.

DISCONTINUOUS FUNCTION PRINCIPLE (DFP). *There exists a function f defined on the closed unit interval* [0, 1] *such that* $f(0) = 0$, *and* $f(x) = 1$ *whenever* $x > 0$.

This statement, although it uses some of the same phrases, is distinctly weaker than the statement that the definition above defines a function on the entire interval; the latter statement has already been shown to be nonconstructive by Example 2. The statement DFP does not state that the conditions $f(0) = 0$, and $f(x) = 1$ for $x > 0$, define a function on the entire interval, but rather that some function exists, defined somehow on all of [0, 1], which has these values at these points, whatever other values it may have at whatever other points.

For a Brouwerian counterexample to DFP, we will use another unsolved problem from number theory, the Goldbach Conjecture. This conjecture says that every even number greater than 2 can be represented as the sum of two primes, such as $4 = 2 + 2$, $6 = 3 + 3, \ldots,$ $1928 = 1201 + 727$, etc. Goldbach stated this in Moscow, in

1742 ($= 1429 + 313$), but to this day no one knows whether it is true or false; see [**17**, p. 30] or [**41**, p. 47]. Use the method of Example 2; test each even number in succession to obtain a sequence $\{\alpha_k\}$ of nonnegative rational numbers, and a real number $\alpha = \Sigma \alpha_k$, where α_k represents the result of testing the kth even number, beginning with 4. For any given even number, it requires only a finite calculation to test the Goldbach Conjecture in that instance. If for every even number > 2 there do exist two prime summands, then all α_k will be 0 and thus $\alpha = 0$. But if the Goldbach Conjecture ever fails, then some α_k will be $1/2^k$, and $\alpha > 0$. Thus we have constructed a real number α in the unit interval such that the Goldbach Conjecture is true if and only if $\alpha = 0$.

Now apply the hypothesis that DFP is true, and use it to calculate $f(\alpha)$. The resulting information about the Goldbach Conjecture is a bit different from that obtained in the first two examples. This means that the nonconstructivity of DFP is different from that of the least upper bound principle and trichotomy; DFP is a less powerful hypothesis. Calculate a rational approximation to $f(\alpha)$ sufficient to determine whether $f(\alpha) < 1$ or $f(\alpha) > 0$. In the first case α cannot be positive and thus from the elementary constructive properties of real numbers it follows that $\alpha = 0$; the Goldbach Conjecture is true. The second case is different; we can conclude that α cannot be zero, but it does not follow that $\alpha > 0$. (The constructive ordering of the real line is discussed below. The difference arises because $\alpha \leqslant 0$ is an essentially negativistic statement (equivalent to the impossibility of $\alpha > 0$), whereas $\alpha > 0$ is an affirmative statement which requires a construction that is not available in this situation.) So in the second case we can only conclude that the Goldbach Conjecture cannot be true, without, however, finding an explicit counterexample. This strange sort of situation has not been known to arise in number theory or analysis. Nevertheless, such a partial solution to an unsolved problem would certainly be interesting, and of pragmatic significance in focusing further research. Thus our derivation of such unavailable information about the Goldbach Conjecture serves sufficiently well to establish the nonconstructivity of DFP; we have proved the following.

Example 3. The *discontinuous function principle* is nonconstructive; it would imply either a proof of the Goldbach Conjecture or a proof of its falsity (without an explicit counterexample).

The weaker conclusion in this example, compared with the first two, is not due to properties of the Goldbach Conjecture, but rather of continuity. We could equally well have used the Goldbach Conjecture in the first two examples. The systematic formulation of Brouwerian counterexamples will clarify these matters by separating the principles under test from the unsolved problems.

The intermediate value theorem Before systematizing the method of Brouwerian counterexamples, we'll discuss one which subjects the intermediate value theorem to a test for numerical meaning, relating it to the decimal expansion of π. The intermediate value theorem says that a continuous function which is positive at the left end of an interval, and negative at the right, must be zero at some point between. See, for example [**22**, p. 190]. This theorem was first proved rigorously (in the classical sense) by Bernard Bolzano, in Bohemia, in 1817 [**8**]. Bolzano was one of the first modern mathematicians who tried to eliminate geometric intuition from proofs; he might well have appreciated the constructive approach of this century; see [**1**, p. 15], [**21**], and [**26**]. The nonconstructivity of the intermediate value theorem was discussed informally in [**32**].

Since there is no connection between intermediate values and questions about digits in the expansion of π, the need for a systematic analysis becomes even more pressing. Questions about these digits were a favorite of Brouwer's; for example, see [**15**, p. 6]. They are of no significance; their only value is that it is unlikely that anyone will ever try to solve them, and thereby necessitate a revision of all the *ad hoc* examples in which they are used.

Here we'll show that the intermediate value theorem, if constructively valid, would lead to an answer to the question "If the sequence of digits 123456789 ever appears in the decimal expansion of π, and the digit 9 in the first such sequence occurs at the nth place, will n then be even or odd?" Notice that no integer n is actually defined by this question. Only if someday someone finds a sequence 123456789 will n then be defined. An answer to the question would be only a prediction whether, in this event, n will be even or odd. This is a much weaker question than whether or not such a sequence of digits does occur. The use of the weaker question reflects nothing about the number π, but rather about the intermediate value theorem under test, which is not enough to answer the stronger question.

To apply the intermediate value theorem, we must define a continuous function, and to define this function, we first define a real number β. This number is defined by an infinite series, and this in turn is based on the decimal expansion of π. Define

$$\beta = \sum_{k=1}^{\infty} \frac{\alpha_k}{10^k},$$

where the factors α_k are integers to be defined in a moment. This is almost like giving a decimal expansion, but with an important difference: the integers α_k may be positive *or* negative. The rule for α_k is as follows. If, in the decimal expansion of π, the kth digit is at the end of a sequence 123456789, and this is the *first* such sequence, then α_k is 1 when k is even, but -1 when k is odd; otherwise, α_k is 0. Because the sequence $\{\alpha_k\}$ is bounded, the series converges and defines the real number β. Since we know that no sequence 123456789 occurs in at least the first few thousand digits, β is a very small number; but we don't know whether it is positive, negative, or zero.

To define a continuous function f on the closed unit interval $[0,1]$, we first give f the values

$$f(0) = 1$$
$$f(1/3) = f(2/3) = \beta$$
$$f(1) = -1$$

and then complete the definition of f by using straight lines between these points. It must be shown that it is indeed possible to define constructively a continuous function in this way, but we postpone this until a later section. This is no minor point to be glossed over, for some similar constructions, such as the one discussed above in connection with DFP, are *not* constructively valid, and one must distinguish carefully.

FIGURE 4 shows three views of this function, corresponding to three possibilities for the number β, but it is dangerously misleading. We cannot say that one of the curves shown represents f. Example 2 shows that we do *not*, in general, know which possibility holds for a given real number β. In fact, the figure represents precisely those cases in which we are not interested, where there is no unsolved problem. When constructing Brouwerian counterexamples, the figures typically only remind us of what we don't know.

Now we apply the intermediate value theorem to the function f in order to "solve"

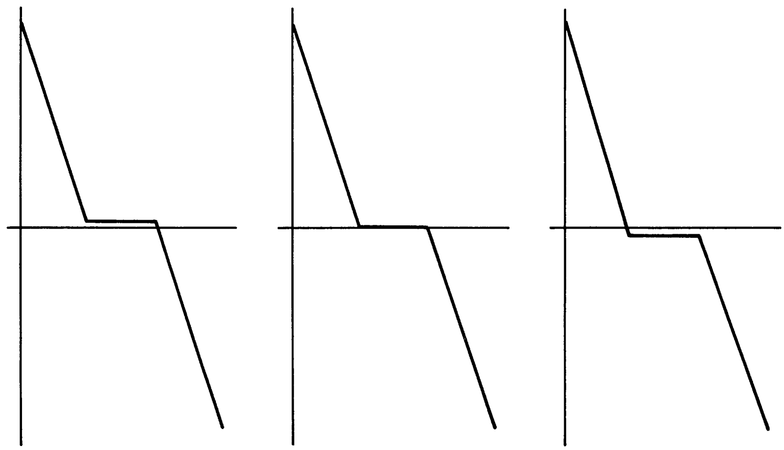


FIGURE 4
Three dangerously misleading views of the function f used in Example 4. Where f crosses the axis, nobody knows.

the "digits of π" problem. Since the function f has values both above and below the axis, the intermediate value theorem claims to provide a specific crossing point, a number x for which $f(x)=0$. To investigate the numerical meaning of the intermediate value theorem, we interpret this claim in a strict constructive sense—we suppose that an explicit rule is given for the construction of x. If x is given by a decimal expansion, we then calculate the first digit. (We do not mean to imply that all real numbers have constructively defined decimal expansions, for there are some problems about this; see [**36**]. Nevertheless, using these expansions serves sufficiently well to demonstrate the method, which is essentially the same no matter how the real number x is approximated.) If the first digit is 5 or less, then it is clear that β cannot be positive, and it follows that the number n in the "digits of π" question (were it to exist) would be odd. On the other hand, if this digit is more than 5, then n would be even. Note that FIGURE 4 is useful in visualizing the proof, but this is only because we are under the spell of the intermediate value hypothesis.

Thus the intermediate value theorem, if constructive, would lead to a solution to the "digits of π" problem. Since, in fact, we have no such solution, we conclude that the intermediate value theorem is constructively invalid. This result is recorded as follows:

Example 4. The *intermediate value theorem* is nonconstructive; it would imply a solution to the "digits of π" problem.

Using sequences to encode unsolved problems To analyze the nature of a Brouwerian counterexample, notice that Example 1 can be broken down into two distinct parts. First we construct a set S which represents the (unsolved) Fermat Problem. Then we use the "theorem" under investigation, in this case the least upper bound principle, to obtain information about this set, thereby "solving" the unsolved problem. The set S plays merely an intermediary role; for other examples other mathematical objects may be used. A set is used in Example 1 because the least upper bound principle is about sets.

It is convenient and helpful to use only a few standard objects for these examples. Surprisingly, they can almost all be handled with a (seemingly) very simple object, an ordinary infinite sequence of integers, and even then using only two integers, 0 and 1. In fact, a sequence is really not so simple an object; the catch is in the (potentially) infinite process which defines the sequence. Since each of us is a merely finite being, no one can actually carry out the infinite process to see what happens. Only occasionally can we predict what would happen; that is the great accomplishment of mathematics, transcending the finiteness of human existence to obtain accurate predictions for infinite processes.

Consider a typical sequence each of whose terms is either 0 or 1. A proof that all the terms are zero constitutes a proof of some theorem, while the existence of a term equal to 1 would be a counterexample (in the usual sense). By mere calculation using the rule defining the sequence, we might find the first million terms to be all 0, but we have no way to tell what might happen in the next million, or further. The essence of a typical theorem in number theory or analysis is *prediction*. No one has ever proved anything by calculating all the terms of an infinite sequence, only by predicting the outcome of such potentially infinite calculations.

The significant fact about these simple sequences is that most of the unsolved problems of number theory and analysis can be encoded using them. To encode Fermat's Last Problem using a sequence, we define the term a_n as 0 in the first case indicated in Example 1, and as 1 in the second case. If you can prove that all the terms of the sequence are 0 (that is, *predict* in a convincing way that each term calculated, no matter how far into the future, will be 0), then you will have proved Fermat's Last Theorem. On the other hand, if someone ever calculates a term that turns out to be 1, then he or she will have a counterexample to Fermat's Last Theorem (in the ordinary sense), and can say that Fermat's Last Theorem is false. It is also conceivable (although less likely) that someone could prove that Fermat's Last Theorem is contradictory; then again it would be false, but in a different sense, as discussed above in connection with Example 3, and below in connection with WLPO.

Decision sequences These sequences of 0's and 1's are so useful for investigating the numerical meaning of mathematical statements, and for analyzing the nature of Brouwerian counterexamples, that it is convenient to adopt a few conventions for their use; in this way we'll see the similarity and relationship between different counterexamples. While it is possible to use sequences with any integers as terms, positive, negative, or zero, we will find it much simpler to use only sequences of 0's and 1's. These suffice for all presently known counterexamples, although one should be prepared for unforeseen complications which might arise in the future. This choice of only two values represents the situation where we use some finite process to look for something (as in the examples above) and either we find it or we don't; the 0 or 1 simply records our results. There are counterexamples in which more values may be convenient; for example, 0, -1, and 1 (which we used in the definition of β in Example 4). However, the systematic use of a single type of sequence has advantages, such as in the comparison of counterexamples, which outweighs this convenience. In situations where a term -1 in a sequence might be handy, we can use instead a factor such as $(-1)^n$ elsewhere (as in Example 4* below). Thus our decision sequences will consist only of 0's and 1's. Similarly, sequences with at most one term equal to 1 have often been useful for counterexamples, but again these are easily converted to examples using decision sequences, by the device used in Example 9 below.

Example 1 was typical in that the searches were cumulative; at each step the search included all previous searches. In practice, there would be no need to repeat all the previous work, but it was convenient to express the results that way, because if one search is successful (the result recorded as a 1), then all succeeding searches will be also. Thus the sequence of recorded results consists of an initial segment of 0's, and then (sometimes) a 1 occurs, after which all the remaining terms are also 1's. In Example 1, Fermat's Last Theorem is true if and only if the sequence consists only of 0's, and it is false (in the strong sense of an explicit counterexample) if and only if there exists a 1 in the sequence. In view of these considerations we adopt the following.

Definition. A *decision sequence* is a nondecreasing sequence $\{a_n\}$ of 0's and 1's.

By "nondecreasing" we mean that $a_n \leqslant a_{n+1}$ for all n. The only question of interest for a given decision sequence is whether or not 1's begin to appear somewhere. It is convenient to assume that decision sequences under consideration begin with a 0, for otherwise there is no problem.

Omniscience principles Decision sequences form the connecting link between nonconstructive classical theorems and unsolved problems. This link is made precise by formulating general statements about decision sequences. In Example 1, for example, a solution to an unsolved problem results if we can tell whether or not a certain decision sequence contains a 1. On the other hand, the least upper bound principle, if true in a numerical sense, would provide precisely that information for any decision sequence. Thus we formulate the following:

LIMITED PRINCIPLE OF OMNISCIENCE (LPO). *Given any decision sequence* $\{a_n\}$, *there is a finite procedure which results either in a proof that* $a_n = 0$ *for all* n, *or in the construction of an integer* n *such that* $a_n = 1$.

One usually says only "either all $a_n = 0$ or some $a_n = 1$". The explicit finite procedure, and the proof or construction, are implied. From a classical point of view, it is obvious that either all the terms are zero, or there is a 1. How can anyone imagine a situation in which neither of these alternatives is true? Such a situation would be one instance of what is referred to as the "Middle" in the *Principle of Excluded Middle.* This principle, which we'll refer to as EM, concerns not only decision sequences, but states that any (meaningful) statement is either true or false. Aristotle formulated this principle, but used it only in finite situations, in which it is constructively valid. When it is (inappropriately) applied to the mathematics of the infinite, it leads to results such as the least upper bound principle, the trichotomy principle, the discontinuous function principle, and the intermediate value principle, which, as we have seen, are nonconstructive. The principle of excluded middle is discussed further in [32, pp. 275–276]. Bishop referred to EM (with a reformulation) as the *Principle of Omniscience* ([3, p. 9] or [6, p. 11]); LPO is a special case which applies only to denumerable problems. To interpret LPO as stating that some statements are true or false requires a bit of care. LPO says that for any decision sequence, the statement "some term has value 1" is either true or false. "True" must be taken in the strict sense that an integer n is constructed and a proof is given that $a_n = 1$; then "false" leads to a proof that all $a_n = 0$. On the other hand, the statement "every term has value 0" does not produce the same results, because its falsity only involves a certain contradiction, from which the construction of an integer n, such that $a_n = 1$, does not follow. This is a good example of the difference between the existential and universal quantifiers, when used constructively.

Though it may seem obvious from a classical viewpoint, LPO appears in an entirely different light when viewed constructively. Classically, since according to EM every statement is either true or false, the middle alternative may be described as saying that both possibilities mentioned in LPO are false, and that indeed is unthinkable. But the constructive interpretation of the middle alternative is simply that we do not know, and this is not only possible, but is actually the present situation regarding many unsolved problems.

The term "omniscience" is used in naming these principles to remind us that we are not omniscient! The position of LPO in Example 1 is now clear; we (effectively) showed first that the least upper bound principle implies LPO, and then we showed that LPO implies a solution to Fermat's Last Problem. It is convenient to state and prove explicitly the first part of Example 1 as follows:

Example 1*. The *least upper bound principle* is nonconstructive; it implies LPO.

Proof. Let $\{a_n\}$ be any decision sequence, and let S denote the set of values of its terms. The set S has at least one element, and at most two, but in general we do not know exactly how many; in any event it is a nonvoid bounded set of real numbers. Using the least upper bound principle as an hypothesis, let t denote the least upper bound of S. We need only calculate a rational approximation within $1/6$ to tell whether the real number t is less than $2/3$ or more than $1/3$. (For more details, see the Constructive Dichotomy Lemma below.) In the first case, it is clear that each term in the given decision sequence is zero. (This does not mean that we actually calculate all the infinitely many terms of the sequence and check each one, but rather that we are able to *predict*, with absolute certainty, that no matter how many terms may be calculated, no matter by whom, and no matter how far into the future, each term will turn out to be 0.) In the second case, using the definition of least upper bound, we construct a number x in S that is more than $1/3$. Since this number x must be a term of the given decision sequence, it must be equal to 1. Thus we have arrived at one or the other of the two alternatives stated in LPO.

Similarly, we restate Example 2 as follows; the proof is left as an exercise.

Example 2*. The *principle of trichotomy of real numbers* is nonconstructive; it implies LPO.

The power of LPO It is not the situation that if Fermat's Last Theorem is proved tomorrow, then the least upper bound principle would suddenly be constructive. Example 1* shows not only that the least upper bound principle implies a solution to Fermat's Last Problem, as shown in Example 1, but that it implies LPO, which would yield solutions to *hundreds* of unsolved problems. Here we'll give only a few examples.

Example 5. LPO implies solutions to each of the following problems:
(a) Fermat's Last Problem.
(b) The Perfect Number Problem.
(c) The Goldbach Conjecture.
(d) The Riemann Hypothesis.

Proof. (a) We showed above how Fermat's Last Problem may be encoded as a decision sequence $\{a_n\}$, such that if all $a_n = 0$, then Fermat's Last Theorem is true, but if some term $a_k = 1$ is ever calculated, it will lead to a counterexample. Thus LPO solves the problem.

(b) and (c) are left as exercises.

(d) The Riemann Hypothesis is a long-standing, unsolved problem involving the *Riemann zeta function* $\zeta(s)$ of the complex variable s; this function plays an important role in the theory of prime numbers [39, pp. 424–431], [43]. The hypothesis states that, aside from a sequence of "trivial" roots, each root $s = \sigma + it$ of $\zeta(s)$ lies on the vertical line $\sigma = 1/2$. For each positive integer n, a finite calculation allows one to determine either that $|\sigma - 1/2| < 1/n$ for all nontrivial roots with $|t| < n$, or that

$|\sigma - 1/2| > 0$ for some such root. Define $a_n = 0$ or $a_n = 1$, accordingly. If all $a_n = 0$, then for each nontrivial root $s = \sigma + it$ we have $|\sigma - 1/2| < 1/n$ for all $n > |t|$, and therefore $\sigma = 1/2$; this proves the Riemann Hypothesis! On the other hand, if some $a_n = 1$, then we have a counterexample.

These few examples should suffice. Most problems in number theory and analysis can be encoded as decision sequences to which LPO applies. For some problems it is a bit more difficult. For example, the question of whether or not there are infinitely many twin primes leads us to construct a sequence of 0's and 1's (not nondecreasing) and to ask whether there are infinitely many 1's. LPO answers this question also; see the section below on the Bolzano-Weierstrass Principle.

Discontinuous functions and WLPO A close look at Example 3 leads us to formulate another omniscience principle, and to restate the example.

WEAK LIMITED PRINCIPLE OF OMNISCIENCE (WLPO). Given any decision sequence $\{a_n\}$, there is a finite procedure which produces either a proof that $a_n = 0$ for all n, or a proof that "$a_n = 0$ for all n" is contradictory.

Example 3*. The *discontinuous function principle* is nonconstructive; it implies WLPO.

The power of WLPO is less than that of LPO, although it is enough to establish the nonconstructivity of certain classical theorems. The conjectures listed in Example 5 can each be encoded into a decision sequence $\{a_n\}$ such that the conjecture is true if and only if all $a_n = 0$. For any one of these, WLPO would provide a finite procedure leading either to a proof of the conjecture or to a proof of its falsity, but without a counterexample. Although this would not settle the problem completely, such an application of WLPO would be of great pragmatic value. It would either give you a proof, or show that a proof was impossible, in which case you could give up trying to find a proof, and concentrate further efforts on the search for a counterexample. Thus Brouwerian counterexamples using WLPO are sufficient to indicate the nonconstructivity of certain classical theorems and "constructions."

The intermediate value theorem and LLPO Considering Example 4, one might reasonably ask "What does it matter whether the first sequence 123456789 (if any) in the digits of π ends at an even or an odd place?" This is another good reason to free Brouwerian counterexamples from such *ad hoc* considerations. A close look at Example 4 shows that the intermediate value theorem leads to another omniscience principle.

LESSER LIMITED PRINCIPLE OF OMNISCIENCE (LLPO). Given any decision sequence $\{a_n\}$, there is a finite procedure which predicts whether the first integer k (if any), such that $a_k = 1$, is even or odd.

Thus the first part of Example 4 may be expressed as follows.

Example 4*. The *intermediate value theorem* is nonconstructive; it implies LLPO.

Proof. Let $\{a_n\}$ be a decision sequence. This sequence has no -1's, as did the sequence used in Example 4, and thus we use here a different definition for β:

$$\beta = \sum_{n=1}^{\infty} \frac{(-1)^n a_n}{10^n}.$$

The definition of the function f is the same as before; FIGURE 4 gives us an omniscient view of its properties. The intermediate value theorem, if constructive, would give an explicit procedure for the construction of a point x with $f(x)=0$. It is convenient to use the constructive dichotomy lemma (see the next section below); thus either $x<2/3$ or $x>1/3$. If $x<2/3$, then it is clear that β cannot be positive, and it follows that the first integer k such that $a_k=1$ (if any) would be odd. On the other hand, if $x>1/3$, then k would be even.

The power of LLPO is even less than that of WLPO, but still sufficient to demonstrate the nonconstructivity of certain classical theorems, because there is in fact no such finite procedure available to us. The relevance of LLPO for Brouwerian counterexamples depends on the fact that people agree that the discovery of such a general finite procedure seems extremely unlikely, even impossible.

Constructive constructions It is time to discuss the basic constructive properties of real numbers and functions which were used in the above counterexamples. The title of this section reminds us of the fact that in classical mathematics the term "construction" appears frequently, but rarely in the sense used here.

For a complete description of the construction of the real numbers, one must refer to Chapter 2 of [3] or [6], and for the construction of the extended real numbers, to [29]. Here we consider briefly only a few of the most important concepts. A real number is a Cauchy sequence of rational numbers. Thus, given any real number, arbitrarily close rational approximations are always available. The notion of *positive real number* is crucial, and closely connected with the idea of constructive existence. When we say that a real number x is *positive*, we mean that we have explicitly constructed a positive integer k and a rational approximation q within $1/k$ of x, and have proved that $q>1/k$. (Since q is a quotient of integers, this is a good example of an application of the fundamental constructivist thesis, that all concepts should be reduced to elementary calculations with the integers.) Thus, to prove a real number x is positive requires a concrete construction; a proof that $x\leqslant 0$ is contradictory will not suffice. (On the other hand, the definition of $x\leqslant 0$, while given in an affirmative manner, is equivalent to the statement that $x>0$ is contradictory; see [3, Lm. 5, p. 24] or [6, Ch. 2, 2.18].)

With this definition of *positive*, and the resultant notion of strict inequality, we consider now one of the most frequently used constructive properties of real numbers. It is the main constructive substitute for trichotomy, and reflects the essence of a real number, given only by approximations.

CONSTRUCTIVE DICHOTOMY LEMMA. *If a and b are real numbers with $a<b$, then for any real number x, either $x<b$ or $x>a$.*

Proof. The given condition $a<b$ means that $b-a>0$; thus a positive integer k can be constructed with $b-a>2/k$. It follows that there is a rational number s such that $a+1/k<s<b-1/k$. Choose a rational approximation r to within $1/k$ of the given real number x. Since trichotomy does hold for the rational numbers, we have either $r\leqslant s$, in which case $x<b$, or we have $r>s$, in which case $x>a$. For more details, see [3, Cor., p. 24] or [6, Ch. 2, 2.17]. The lemma is illustrated in FIGURE 5.

Constructive definitions for convergence of sequences and series are also straightforward. One says that a sequence $\{x_n\}$ of real numbers *converges* to a real number x if one has constructed a sequence $\{N_k\}$ of *convergence parameters* with the property that $|x-x_n|<1/k$ whenever $n\geqslant N_k$. This is exactly the same as the classical definition, except that classically one says something like "for all k there exists N_k

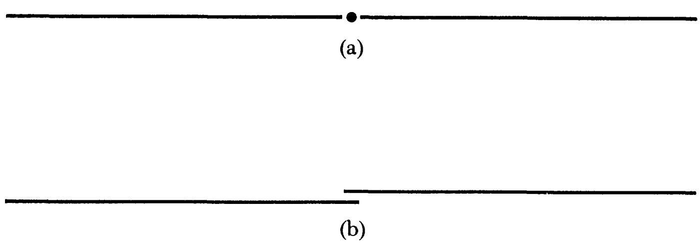


FIGURE 5
Classical trichotomy vs. constructive dichotomy of the real numbers. In the omniscient view (a), the real line is divided precisely into three distinct nonoverlapping parts. The constructive view (b) reveals only two cases. They are not mutually exclusive; however, the region of overlap may be as small as desired, according to the degree of precision required.

such that ..." without being careful to explain in what sense "there exists" is to be understood. Classically one allows convergence to be proved by assuming that such parameters N_k do not exist and deriving a contradiction, while constructively these parameters must be explicitly constructed by means of a finite procedure. The convergence of an infinite series reduces in the usual way to convergence of the sequence of partial sums.

We can now justify the definition of a continuous function by means of straight lines, as in Examples 4 and 4*. We have the function f defined clearly enough on the three subintervals $[0,1/3]$, $[1/3,2/3]$, and $[2/3,1]$. However, these subintervals do not constitute the entire interval $[0,1]$, because we have no finite procedure which determines in which subinterval a given point lies. (A slight modification of Example 2* shows that such a procedure would imply LPO.) We give only a brief sketch of the definition of f. The important conditions, which do hold here, are that uniformly continuous functions are used on each subinterval, and that they connect properly. To define f at an arbitrary point x of $[0,1]$ means to give an approximation to $f(x)$ to within ε, for any $\varepsilon > 0$. We may assume that $x < 2/3$. The other case given by the dichotomy lemma, when $x > 1/3$, is similar. Using the dichotomy lemma again, we have either $x < 1/3$ or $x > 1/3 - \varepsilon/6$. In the first case we have the value of f already defined, while in the second case it suffices to use β as the required approximation to $f(x)$. We leave the remaining details, including the continuity of f, as an exercise for the reader (who may wish to consult Chapter 2 of [3] or [6]). An alternative method for constructing such functions is given in Example 14.28 of [31].

Using the constructive dichotomy lemma, we can replace many nonconstructive classical theorems by constructive substitutes which are fully adequate for the constructive development of analysis. The intermediate value theorem is replaced by a constructive theorem which, given any small positive number ε, constructs a point x at which $|f(x)| < \varepsilon$. This is [**3**, Ch. 2, ex. 11], or [**6**, Ch. 2, 4.8]; we give only a brief sketch of the proof, leaving the reader to fill in the details. Use the uniform continuity of f to construct $\delta > 0$ such that $|f(x) - f(y)| < \varepsilon$ whenever $|x - y| < \delta$, and divide the interval into n subintervals, each of length less than δ. At each subdivision point the dichotomy lemma determines either that f has value less than ε or that f has value more than 0. At the ends of at least one subinterval we must obtain opposite determinations; at the left end of the first such subinterval we find a suitable point x. Thus we have the following.

THEOREM 1. *If f is a uniformly continuous function on a closed bounded interval $[a,b]$, with $f(a) > 0$ and $f(b) < 0$, then for any $\varepsilon > 0$ there exists a point x in the interval such that $|f(x)| < \varepsilon$.*

As for any theorem in constructive mathematics, the phrase "there exists" appearing above is to be understood in the strict constructive sense; in this case, from the definition of f and the other data given, an explicit finite procedure is obtained which constructs the point x.

Sometimes the intermediate value theorem is also expressed in the classically equivalent form of its contrapositive, as, for example, in [**20**, p. 80]. Constructively, the contrapositive of a statement follows from the statement, but is not an equivalent. One form of contrapositive of Theorem 1 states that a function bounded away from zero cannot assume opposite signs at the endpoints. We prove a stronger version of this in the next theorem, which seems not to have appeared previously. A function f is said to be *never zero* on a set S if $|f(x)| > 0$ for every x in S; it is said to *have constant sign* on S if $f(x) \cdot f(y) \geqslant 0$ for all x and y in S.

THEOREM 2. *If a function f is continuous and never* 0 *on an interval* $[a, b]$, *then f has constant sign on* $[a, b]$.

Proof. Since $|f(a)| > 0$, it follows that either $f(a) > 0$ or $f(a) < 0$ (use a rational approximation). We need consider only the case $f(a) > 0$; the other case is similar. Let y be any point of the interval; either $f(y) > 0$ or $f(y) < 0$. Suppose the second case occurs; then the sets $U = \{x : f(x) < 0\}$ and $V = \{x : f(x) > 0\}$ are nonvoid open subsets of $[a, b]$ which cover the entire interval. By the constructive connectivity theorem [**27**, Thm. 2], these sets have a common point, which is absurd. Thus the first case must obtain, and $f(y) > 0$.

Other constructive forms of the intermediate value theorem are also available; some are listed in [**3**, p. 59] and [**6**, p. 63]. This multiplicity of constructive forms is typical. After the fracturing of a classical theorem by a Brouwerian counterexample, constructive workers pick up the pieces and remold them into a number of different constructively valid theorems, each displaying a different aspect of the situation.

The least upper bound principle also has a powerful constructive substitute. We restrict the theorem to totally bounded sets. A set S is said to be *totally bounded* if, given any $\varepsilon > 0$, we can construct a finite subset F such that every point of S lies within ε of some point of F. The maximum of the points in F gives us an approximation to the least upper bound of S to within ε. Since giving arbitrarily close approximations to a real number is equivalent to defining it, we say that the least upper bound of S exists. We state the result as follows.

THEOREM 3. *Every nonvoid totally bounded set of real numbers has a least upper bound and a greatest lower bound.*

The details are in [**3**, Thm. 3, p. 34] or [**6**, Ch. 4, 4.3]. For least upper bounds and greatest lower bounds in the extended real number system, see [**28**] and [**29**]; for alternative constructive notions, see [**31**, sec. 4].

The procedures produced by the above theorems might be quite lengthy; for some comments on this, see [**3**, p. 3] or [**6**, p. 6], and [**5**]. In an important application, rather than giving a solution on a hand calculator in a half a minute, the procedures could lead instead to years of work trying to write programs efficient enough to produce the solution on a large computer in only a month. Nevertheless, we maintain the distinction between an infinite calculation, which we have absolutely no hope of actually carrying out, and a finite process, however long. The important questions on the efficiency of procedures belong to the second phase of the constructivization of mathematics. It is too soon to demand progress on this problem from the very few present-day constructivists. More help is needed—invited!—urgently awaited!—perhaps it will come from among the readers of this MAGAZINE.

Twin primes and the Bolzano-Weierstrass theorem A famous unsolved problem in number theory is whether or not there are infinitely many twin primes of the form p and $p+2$, such as 3 and 5, 5 and 7, 11 and 13,..., 209267 and 209269, etc. See, for example [17, p. 31] and [42, Ch. 1]. Although one might wish to utilize this problem in a Brouwerian counterexample, a difficulty arises. At first glance neither LPO nor the other principles mentioned above seem to imply a solution to this problem. The *Bolzano-Weierstrass principle*, although unrelated to any problem in number theory, presents some similar difficulties; although it clearly implies LPO, the converse is not immediate. The difficulties, which stem from the need to consider arbitrary sequences of positive integers, are resolved by the following.

THEOREM 4. *The following are equivalent.*

(a) LPO. Limited Principle of Omniscience. *For any decision sequence* $\{a_n\}$, *either all* $a_n = 0$ *or some* $a_n = 1$.

(b) BSP. Bounded Sequence Principle. *Any sequence of positive integers is either bounded or unbounded.*

(c) KSP. Constant Subsequence Principle. *Any bounded sequence of positive integers has a constant subsequence.*

(d) BWP. Bolzano-Weierstrass Principle. *Any bounded sequence of real numbers has a convergent subsequence.*

(e) MSP. Monotone Sequence Principle. *Any bounded monotone sequence of real numbers converges.*

We give here only a sketch of the proof that the *Constant Subsequence Principle* implies the *Bolzano-Weierstrass Principle*. The proof of BWP proceeds in essentially the same manner as in most elementary analysis texts, by interval-halving. Using the constructive dichotomy lemma, we may assume that the "halves" of the interval have a small overlap, and thus one can tell in which of these halves each term of the given sequence $\{x_n\}$ lies. Constructive danger is first sighted at the point of deciding which half contains infinitely many terms of the sequence. Here KSP navigates an omniscient course. Define $p_n = 1$ or $p_n = 2$ according as x_n lies in the left or right half. Since the sequence $\{p_n\}$ is bounded, KSP provides a constant subsequence; this corresponds to a subsequence of $\{x_n\}$ which lies wholly in one half of the interval, towards which we should steer. The reader can try the rest of the proof as an exercise, or refer to [33].

We now consider the twin prime problem; extending Example 5.

Example 6. LPO implies a solution to the twin prime problem.

Proof. Construct a sequence $\{p_n\}$ of positive integers as follows; if both n and $n+2$ are prime, define $p_n = n$, and otherwise $p_n = 1$. There are infinitely many twin primes if and only if the sequence $\{p_n\}$ is unbounded. Thus the Bounded Sequence Principle would yield a solution to the twin prime problem.

Because of this example, one might say that the Bolzano-Weierstrass Principle and the Monotone Sequence Principle are nonconstructive because they each imply a solution to the Twin Prime Problem. The Bounded Sequence Principle may also be used in connection with other questions in number theory, such as whether or not there exist infinitely many even numbers which are sums of two primes.

The limited principle of existence Brouwerian counterexamples utilizing the nonconstructive omniscience principles LPO, WLPO, and LLPO quite clearly demonstrate the nonconstructivity of classical theorems, because there are unsolved problems in analysis and number theory for which these principles would yield solutions or information not actually available. There are other classical theorems, however, which imply principles involving decision sequences which certainly seem nonconstructive, and for which no proofs are at hand, and yet for which there are no known unsolved problems whose solution they would provide. Some of these principles relate to very central problems in constructive mathematics. In this section we discuss one of these, LPE, which involves fundamental properties of the real numbers, and in a later section another, WLPE, which is related to the constructivity of continuity theorems.

THE LIMITED PRINCIPLE OF EXISTENCE (LPE). Given any decision sequence $\{a_n\}$, for which it is contradictory that $a_n = 0$ for all n, there is a finite procedure which results in the construction of an integer n such that $a_n = 1$.

The contrast of conditions in LPE evokes the sharp antithesis between classical and constructive mathematics; between pseudoexistence, derived from a proof by contradiction, and constructive existence, derived from an explicit finite process. However, we have no example of an unsolved problem whose solution would be given by LPE. Thus a counterexample involving LPE, such as in the next section below, provides less conclusive evidence of nonconstructivity than one involving the other omniscience principles. Still, a finite procedure as specified in LPE seems, from a constructive viewpoint, very unlikely. In any event, a counterexample involving LPE has pragmatic value in that it tends to limit further efforts to prove the conjecture and intensifies efforts to find a Brouwerian counterexample in the strict sense. LPE is sometimes referred to as Markov's Principle. In recursive function theory, in contrast to the strict Bishop-type constructive mathematics discussed in this paper, arguments are made for the plausibility of LPE, and it is often used as an axiom.

Consider a decision sequence $\{a_n\}$ of the sort considered in LPE: it is contradictory that $a_n = 0$ for all n; yet we have no proof that there exists an integer n such that $a_n = 1$. Using the method of the above counterexamples, we obtain a real number

$$\alpha = \sum_{n=1}^{\infty} \frac{a_n}{10^n},$$

which is clearly $\geqslant 0$ but cannot be 0, because then all the terms a_n in the decision sequence would be necessarily 0. On the other hand, we have no proof that $\alpha > 0$, for this would mean that we had constructed some term $a_n = 1$. Since α bears to 0 a relation not covered by the conventional terminology and symbols, we adopt the following.

Definition. A real number α will be said to be *almost positive* when $\alpha \geqslant 0$ and it is contradictory that $\alpha = 0$. This condition will be written $\alpha \cdot\!> 0$.

With this terminology, LPE has a simple expression: *every almost positive real number is positive.*

Irrational numbers When forming definitions for the constructive development of mathematics, one has a fairly wide choice. This is because the classical definitions typically have a variety of classically equivalent, but constructively quite distinct, formulations, and one must exercise great care in choosing a definition with useful numerical meaning. One example, in a sense the first to arise following the generation of the real number system out of the rationals, is the definition of irrational number. Consider a real number x, and two forms of the classical definition:

(i) For every rational number q, the equality $x = q$ is contradictory.

(ii) For every rational number q, the inequality $|x - q| > 0$ holds.

Although these two conditions are classically equivalent, they are constructively quite different, as the following example shows.

Example 7. The statement "*If x is a real number such that the equality $x = q$ is contradictory for every rational number q, then $|x - q| > 0$ for every rational q*" is nonconstructive; it implies LPE.

Proof. Assuming the statement in quotes, we will obtain a proof of LPE. Let $\{a_n\}$ be a decision sequence such that it is contradictory that all $a_n = 0$, and let x be the real number defined by

$$x = \sum_{n=1}^{\infty} \frac{a_n\sqrt{2}}{2^n}.$$

First we show that x satisfies condition (i). Let q be a rational number and suppose that $x = q$. Since trichotomy holds for the rationals, we have either $x = 0$ or $x > 0$. In the first case it follows that all $a_n = 0$, a contradiction. In the second case, the decision sequence becomes constantly 1 from some point on, and the sum of the series has the form $x = \sqrt{2}/2^k$ where k is the largest integer such that $a_k = 0$. But such a value for x cannot be rational, so again we have a contradiction. Thus condition (i) is satisfied. By hypothesis, x satisfies condition (ii). Since $q = 0$ is rational, it follows that $x > 0$, and there exists an integer n such that $a_n = 1$. This proves LPE.

Nevertheless, even with the strict definition (ii), one has no difficulty finding a plentiful supply of irrational numbers. For example, the usual proof that $\sqrt{2}$ is irrational, while not constructive, becomes so with only a little more care. The familiar classical proof, that for any integer p and any integer $q \neq 0$, the numbers p^2/q^2 and 2 are distinct, is constructively valid; it involves only integers, about which there are no constructive complications. The definition of $\sqrt{2}$ presents no difficulties; for example, decimal approximations suffice, since the precise rule for the determination of each approximation involves only finite decimals. To show that $\sqrt{2}$ is irrational in the strong sense of condition (ii) above, we must show that for any p and q, the inequality $|p/q - \sqrt{2}| > 0$ holds; this means we must construct a positive integer k with $|p/q - \sqrt{2}| > 1/k$. It suffices to consider the case $1 < p/q < 2$, for otherwise one simply takes $k = 3$. What the traditional proof actually shows, using Euclid's *Fundamental Theorem of Arithmetic*, is that p^2 and $2q^2$ are distinct integers (since there are an even number of factors 2 in the unique prime decomposition of p^2, but an odd number in the decomposition of $2q^2$). Thus these integers differ by at least 1, and we have

$$|p/q - \sqrt{2}| \cdot |p/q + \sqrt{2}| \cdot q^2 = |p^2 - 2q^2| \geqslant 1.$$

Since $0 < |p/q + \sqrt{2}| < 4$ it follows that

$$|p/q - \sqrt{2}| > 1/4q^2$$

and we may take $k = 4q^2$. This is typical of much of classical mathematics, which, along with many nonconstructivities, does contain a vast amount of numerical meaning which merely needs to be made explicit, although this usually requires more effort than in this example.

Omniscience principles and real numbers Each of the four omniscience principles so far discussed has an equivalent formulation involving the ordering of the real numbers. It is convenient to have these formulations available, for ease in constructing counterexamples, and for finding relationships between the various omniscience principles.

THEOREM 5. *Each of the omniscience principles listed below has the two equivalent formulations indicated.*

(a) The Limited Principle of Omniscience (LPO).

(i) *For any decision sequence* $\{a_n\}$, *either all* $a_n = 0$ *or some* $a_n = 1$.

(ii) *For any real number* x, *either* $x \leqslant 0$ *or* $x > 0$.

(b) The Weak Limited Principle of Omniscience (WLPO).

(i) *For any decision sequence* $\{a_n\}$, *either all* $a_n = 0$ *or it is contradictory that all* $a_n = 0$.

(ii) *For any real number* x, *either* $x \leqslant 0$ *or* $x \cdot > 0$.

(c) The Lesser Limited Principle of Omniscience (LLPO).

(i) *For any decision sequence* $\{a_n\}$, *either the first integer* k (*if any*), *such that* $a_k = 1$, *is even, or it is odd.*

(ii) *For any real number* x, *either* $x \leqslant 0$ *or* $x \geqslant 0$.

(d) The Limited Principle of Existence (LPE).

(i) *For any decision sequence* $\{a_n\}$, *if it is contradictory that all* $a_n = 0$, *then some* $a_n = 1$.

(ii) *For any real number* x, *if* $x \cdot > 0$, *then* $x > 0$.

From this theorem it follows that LPO implies WLPO, that WLPO implies LLPO, and that LPO is equivalent to WLPO and LPE combined. The theorem is easily verified using the following two lemmata connecting real numbers and decision sequences.

LEMMA 1. *For any real number* x *there exists a corresponding decision sequence* $\{a_n\}$ *such that*

(i) $x \leqslant 0$ *if and only if all* $a_n = 0$

(ii) $x > 0$ *if and only if some* $a_n = 1$.

Conversely, for any decision sequence $\{a_n\}$ *there exists a corresponding real number* x *satisfying these two conditions.*

Proof. Let x be a given real number. For each positive integer n, the constructive dichotomy lemma provides a finite procedure which results in one of two conclusions, either $x < 1/n$, or $x > 0$. Define $a_n = 0$ or $a_n = 1$ accordingly, continuing with the later choice once it occurs; this defines a decision sequence $\{a_n\}$. If all $a_n = 0$, the dichotomy lemma always leads to the first alternative; thus $x < 1/n$ for all n, and it follows that $x \leqslant 0$. The converse is clear. If $x > 0$, then there exists an integer n such that $x > 1/n$. At the nth step in the construction of the decision sequence, the second alternative is necessitated; thus some $a_n = 1$. The converse to this is also clear.

Conversely, given a decision sequence $\{a_n\}$, define

$$x = \sum_{n=1}^{\infty} \frac{a_n}{2^n}.$$

The two conditions are easily verified.

The next lemma is useful in connection with LLPO. The strangely hypothetical condition "the first integer k (if any), such that $a_k = 1$, is even" is more conveniently expressed by the straightforward affirmative condition "$a_n = a_{n+1}$ for all even n".

LEMMA 2. *For any real number* x *there exists a corresponding decision sequence* $\{a_n\}$ *such that*

(i) $x \leqslant 0$ *if and only if* $a_n = a_{n+1}$ *for all even* n
(ii) $x \geqslant 0$ *if and only if* $a_n = a_{n+1}$ *for all odd* n.
Conversely, for any decision sequence $\{a_n\}$ *there exists a corresponding real number* x *satisfying these two conditions.*

Proof. For a given real number x, the construction of a suitable decision sequence is left as an exercise; see [31, sec. 2.6]. Conversely, given a decision sequence $\{a_n\}$, define

$$x = \sum_{n=1}^{\infty} \frac{(-1)^{n+1} a_n}{n}.$$

The elementary topology of the real line A typical theorem found in elementary analysis courses is the following; see, for example [18, 3.19.6].

THE OPEN INTERVAL HYPOTHESIS. Any nonvoid open set on the real line is the union of an at most countable family of disjoint open intervals.

Example 8. The *open interval hypothesis* is nonconstructive, it implies LPO.

Proof. The hypothesis is nonconstructive even if limited to *bounded* sets, and even if *any* family of disjoint open intervals, not necessarily countable, is allowed. We assume the open interval hypothesis and derive a proof of LPO. It is convenient to use Theorem 5; thus for any real number x we must show either $x \leqslant 0$ or $x > 0$. Define $y = |x|$; then $y \geqslant 0$ and it will suffice to show that either $y > 0$ or $y = 0$. Using the dichotomy lemma, we may assume $y < 1$, for if $y > 0$ then there is nothing more to prove. Consider the set

$$U = (-1, y) \cup (0, 1).$$

Clearly U is an open set. (In fact, it is a countable, even finite, union of open intervals, but they are not necessarily disjoint. Disjointness is the crucial part of the statement being tested; it determines the connected components of a set.) By hypothesis, U has a decomposition $U = \bigcup_\alpha I_\alpha$ into disjoint open intervals. Since the point $1/2$ lies in U, it must lie in one of the intervals I_α; this interval has the form (a, b). Applying the constructive dichotomy lemma, either $a < 0$ or $a > -1$. In the first case the point 0 lies in U, and it follows that $y > 0$, while in the second case it follows that $y = 0$. This proves LPO. The set U is illustrated in FIGURE 6.

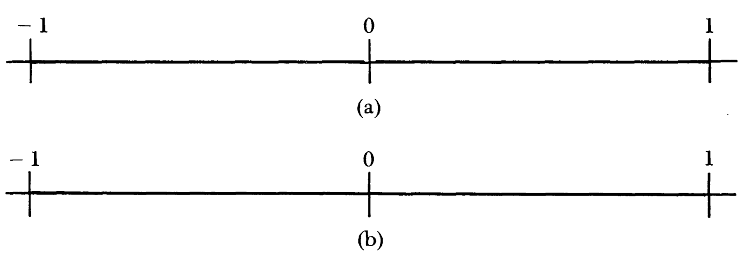


FIGURE 6.
Two views of the open set $U = (-1, y) \cup (0, 1)$ used in Example 8. If the number y is 0, then U consists of two disjoint open intervals, but if y is positive, then U is the entire interval $(-1, 1)$. We have no way to determine that one of these cases applies. What are the connected components of U?

However, most constructively important open sets on the line *can* be resolved into countably many disjoint open intervals; see [**29**] and [**30**]. For other open problems on the topology of the line in need of constructive work, see [**10**] and [**31**].

Counterexamples in algebra A well-known classical theorem in elementary abstract algebra states that every field has a characteristic which is either 0 or a prime p. (The characteristic is the least positive integer m such that $m \cdot e = 0$, where e is the identity of the field, if such an integer exists, and 0 when no such positive integer m exists. See, for example [**2**, pp. 386–392] or [**45**, pp. 91–93].)

Example 9. The statement "*Every field has a characteristic which is either* 0 *or a prime*" is nonconstructive; it implies LPO.

Proof. Let $\{a_n\}$ be a decision sequence, let $\{p_n\}$ be the ordered sequence of all positive primes, and define

$$A = \{0\} \cup \{p_k : a_{k-1} < a_k\}.$$

Note that A has at least one element and at most two, but we do not know which. (The inequality is a device used with decision sequences to indicate conveniently "the first k such that $a_k = 1$ (if any)" and, for the sake of uniformity in counterexamples, to alleviate the need for "sequences of 0's and 1's with at most one term equal to 1".) In the domain $\mathbb{Z}$ of integers, let P be the subring generated by A, and let D be the quotient ring $\mathbb{Z}/P$. Then D is an integral domain; the proof (a rocky shallows) is left for the reader. Thus D has a quotient field F. Apply now the hypothesis that F has a characteristic. If this characteristic is 0, then in the decision sequence all $a_n = 0$. But if this characteristic is a prime, with position k in the sequence $\{p_n\}$, then $a_k = 1$. Thus the hypothesis implies LPO.

Whether or not D is itself a field is also a mystery. For more details, and some related positive constructive results, see [**25**] and [**35**]. Using, in place of $\{p_n\}$, the sequence of primes having residue 1 modulo 4, and considering the polynomial $x^2 + 1$, one may prove the following. (Hint: see [**7**, Thm. 13.2].)

Example 10. The statement "*Every polynomial over a field is either irreducible or can be factored into irreducible polynomials*" is nonconstructive; it implies LPO.

Continuity One of the oldest constructivity problems is whether or not every real-valued function on the closed unit interval is continuous. We formulate this problem as follows.

Continuity Principle (CP). Every real-valued function on the closed unit interval is continuous.

The typical classical counterexample to this principle is constructively invalid, as shown above in the section on discontinuous functions. Brouwer [**13**] proved CP. His proof, however, was not constructive in the strict sense; it used methods of questionable constructivity (which are still used in intuitionistic mathematics). Expositions of Brouwer's proof may be found in [**19**, Ch. 3] and [**23**, Ch. 3].

Since CP is classically false, we consider only the following (classically true) weak mutation of CP:

Limited Continuity Principle (LCP). Every real-valued function on the closed unit interval, which is nondecreasing and approximates intermediate values, is continuous.

By *nondecreasing* we mean that $f(x) \leqslant f(y)$ whenever $x \leqslant y$. *Approximates intermediate values* means that if $f(x) < \lambda < f(y)$ and $\varepsilon > 0$, then there exists a point z in the interval such that $f(z)$ is within ε of λ; (cf. Theorem 1). Classically, for nondecreasing functions, this is equivalent to attaining intermediate values exactly. Thus LCP is a partial converse to the intermediate value theorem; classical proofs are found in some calculus texts, for example [22, p. 192].

A Brouwerian counterexample to LCP is available, but, as for the counterexample above concerning the irrationals, not in the strict sense. It does not relate LCP to an unsolved problem, but to a principle, which, in the fashion of LPE, only "seems" nonconstructive. Thus we enter here into recent results about which there is room for difference of opinion. The principle involved is a weak form of LPE, and a converse to a weak form of constructive dichotomy. An equivalent form of the constructive dichotomy theorem is: if $c > 0$, then for any real number x, either $x > 0$ or $x < c$. A weak form of this is: if $c > 0$, then for any real number x, either $x \cdot > 0$ or $x < \cdot c$ (for this notation see the section on LPE above). The converse to this is the following.

THE WEAK LIMITED PRINCIPLE OF EXISTENCE (WLPE). If c is a real number such that, for any real number x either $x \cdot > 0$ or $x < \cdot c$, then $c > 0$.

Any real number c which satisfies the condition "for any real number x either $x \cdot > 0$ or $x < \cdot c$" is said to be *pseudo-positive*. Thus WLPE has the simple expression "every pseudo-positive real number is positive." Clearly, any pseudo-positive real number is almost positive; thus LPE implies WLPE. For a more complete discussion of LPE and its weaker versions, see [34].

The structure of WLPE strongly suggests that it is nonconstructive; it purports to derive quite affirmative information, the construction of specific integers which demonstrate that c is positive, from a mere dichotomy of negativistic conditions. However, this is somewhat speculative and WLPE will require more time for a definitive evaluation. Nevertheless, taking the notion of *nonconstructive* in the broad sense, that a classical theorem is reduced to an elementary, simply expressed, classical property of the real numbers which, from our extensive experience with similar properties, seems to preclude all possibility of constructive proof, we have the following counterexample to LCP. For the proof, see [31, Thm. 16.5].

Example 11. The *limited continuity principle* is nonconstructive; it is equivalent to WLPE.

Appendix. Negativistic counterexamples vs. positive constructive developments There is a danger in devoting an entire paper to negativistic counterexamples. As Errett Bishop has written, "The counterexamples are deceitful. The reader is asked not to form the impression that the purpose of constructive mathematics is to consider pathological numbers. The only reason for discussing such numbers is to show that certain statements are not constructively valid" [3, p. 60] [6, p. 65]. The danger is less today, however, than in 1967 when the above quote appeared. Prior to that time one had only Brouwer's critique (providing the crucial motivation for the constructivization of mathematics), certain intuitionistic results (often mixed with nonconstructive elements such as free choice sequences), some idealistic logical considerations using formal systems (an approach diametrically opposed to Bishop's strict constructivist thesis), and results in recursive function theory (extensive, but only semi-constructive, because of restricted concepts of number and function, and some use of nonconstructive reasoning). But until 1967 there were few systematic, strictly constructive advances. Thus it was important that Bishop try to correct the prevalent misunderstandings. At this time, however, we have available Bishop's monumental work [3], which constructivizes a large portion of analysis, and indicates the direction for further positive constructive work.

The purpose of this article is to describe Brouwer's critique of 80 years ago, showing the nonconstructivities in classical mathematics. This critique must be evaluated in the crucially different present-day context, in which there are available not only powerful methods for the constructive development of mathematics, but also sufficient examples of their application. The references below include only a few recent constructive advances; their bibliographies provide more extensive references.

REFERENCES

1. G. Birkhoff, ed., *A Source Book in Classical Analysis*, Harvard University Press, Cambridge, 1973.
2. G. Birkhoff and S. MacLane, *A Survey of Modern Algebra*, rev. ed., MacMillan, New York, 1953.
3. E. Bishop, *Foundations of Constructive Analysis*, McGraw-Hill, New York, 1967. (To be reprinted by Springer Verlag, Berlin, with an introduction by G. Stolzenberg.)
4. ______, *Aspects of constructivism*, Notes on lectures given at New Mexico State University, Las Cruces, 1972.
5. ______, Schizophrenia in contemporary mathematics (1973), in *Errett Bishop; Reflections on Him and His Research*, ed. by M. Rosenblatt, Amer. Math. Soc., Providence, 1985.
6. E. Bishop and D. Bridges, *Constructive Analysis*, (Grundlehren der mathematischen Wissenschaften, 279), Springer, Berlin, 1985.
7. E. D. Bolker, *Elementary Number Theory, an algebraic approach*, Benjamin, New York, 1970.
8. B. Bolzano, Rein analytischer Beweis des Lehrsatzes, dass zwischen je zwey Werthen, die ein entgegengesetztes Resultat gewähren, wenigstens eine reelle Wurzel der Gleichung liege, Abh. der Kgl. Ges. der Wiss. Prague, 1817, 1–60.
9. D. Bridges, *Constructive Functional Analysis*, Pitman, London, 1979.
10. ______, Connectivity properties of metric spaces, *Pacific J. Math.* 80 (1979), 325–331.
11. D. Bridges and R. Mines, What is constructive mathematics?, *Math. Intelligencer* 6 (1984), no. 4, 32–38.
12. L. E. J. Brouwer, *Over de grondslagen der wiskunde*, Thesis, Amsterdam, 1907.
13. ______, Beweis, daß jede volle Funktion gleichmäßig ist, *Proc. Akad. Amsterdam* 27 (1924), 189–194.
14. ______, *Collected Works, vol. 1, Philosophy and Foundations of Mathematics*, ed. A. Heyting, North-Holland, Amsterdam, 1975.
15. ______, *Cambridge Lectures on Intuitionism*, Cambridge University Press, 1981.
16. Y. K. Chan, Notes on constructive probability theory, *Ann. Probability* 2 (1974), 51–75.
17. R. Courant and H. Robbins, *What is Mathematics?*, Oxford University Press, London, 1941.
18. J. Dieudonné, *Foundations of Modern Analysis*, Academic Press, New York, 1960.
19. M. Dummett, *Elements of Intuitionism*, Clarendon Press, Oxford, 1977.
20. L. Gillman and R. H. McDowell, *Calculus*, Norton, New York, 1973.
21. I. Grattan-Guinness, *The Development of the Foundations of Mathematical Analysis from Euler to Riemann*, The MIT Press, Cambridge, 1970.
22. G. H. Hardy, *A Course of Pure Mathematics*, 10th ed., University Press, Cambridge, 1958.
23. A. Heyting, *Intuitionism, an introduction*, 3rd ed., North-Holland, Amsterdam, 1971.
24. W. Julian, ε-continuity and monotone operations, *Pacific J. Math.* 115 (1984), 385–390.
25. W. Julian, R. Mines, and F. Richman, Algebraic numbers, a constructive development, *Pacific J. Math.* 74 (1978), 91–102.
26. D. R. Kurepa, Around Bolzano's approach to real numbers, *Czech. Math. J.* 32 (1982), 655–666.
27. M. Mandelkern, Connectivity of an interval, *Proc. Amer. Math. Soc.* 54 (1976), 170–172.
28. ______, Suprema of located sets, *J. London Math. Soc.* (2), 20 (1979), 161–164.
29. ______, Resolutions on the line, *Pacific J. Math.* 88 (1980), 91–100.
30. ______, Components of an open set, *J. Austral. Math. Soc.* (*Series A*) 33 (1982), 249–261.
31. ______, Constructive Continuity, *Memoirs Amer. Math. Soc.* 277, (1983).
32. ______, Constructive Mathematics, this MAGAZINE 58 (1985), 272–280.
33. ______, Limited omniscience and the Bolzano-Weierstrass principle, *Bull. London Math. Soc.*, in press.
34. ______, Constructively complete finite sets, Z. *Math. Logik Grundlag. Math.*, in press.
35. R. Mines, Algebraic number theory, a survey, *The L. E. J. Brouwer Centenary Symposium*, North-Holland, Amsterdam, 1982, pp. 337–358.
36. J. Myhill, What is a real number?, *Amer. Math. Monthly* 79 (1972), 748–754.

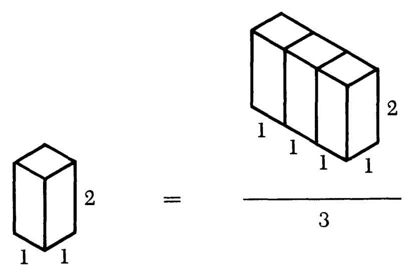
37. F. Richman, Meaning and information in constructive mathematics, *Amer. Math. Monthly* 89 (1982), 385–388.
38. F. Richman, G. Berg, W. Julian, and R. Mines, The constructive Jordan curve theorem, *Rocky Mountain J. Math.* 5 (1975), 215–236.
39. S. Saks and A. Zygmund, *Analytic Functions*, Drukarnia Naukowa, Warsaw, 1952.
40. G. Stolzenberg, Review: Foundations of Constructive Analysis by E. Bishop, *Bull. Amer. Math. Soc.* 76 (1970), 301–323.
41. D. J. Struik, *A Source Book in Mathematics, 1200-1800*, Cambridge, 1969 (reprint, Princeton, 1986).
42. H. Tietze, *Famous Problems of Mathematics*, Graylock Press, Baltimore, 1965.
43. E. C. Titchmarsh, *The Theory of the Riemann zeta-function*, Oxford, 1951.
44. H. S. Vandiver, Fermat's Last Theorem, *Amer. Math. Monthly* 53 (1946), 555–578.
45. B. L. van der Waerden, *Modern Algebra*, vol. 1, Ungar, New York, 1949. (Original; *Moderne Algebra*, Springer, Berlin, 1931.)

Proof without Words

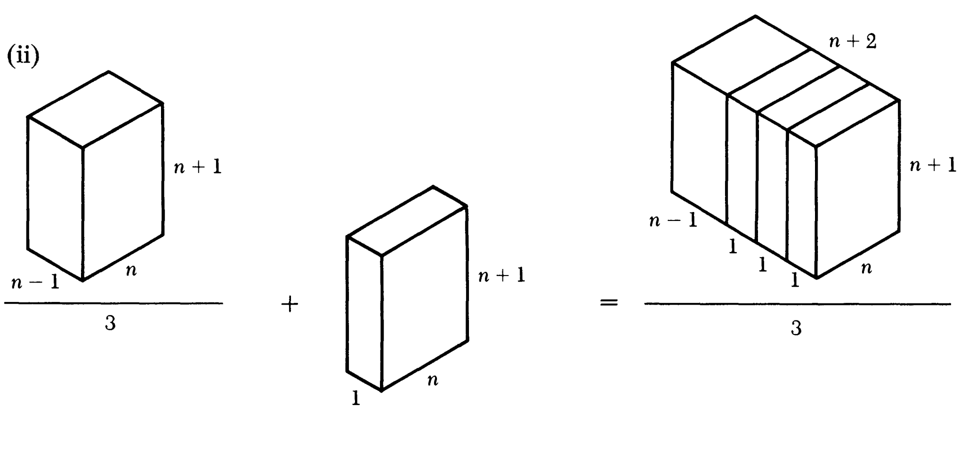
THEOREM. $(1 \times 2) + (2 \times 3) + (3 \times 4) + \cdots + (n-1)n = \dfrac{(n-1)n(n+1)}{3}$, $n = 1, 2, 3, \ldots$.

Proof.

(i)



(ii)



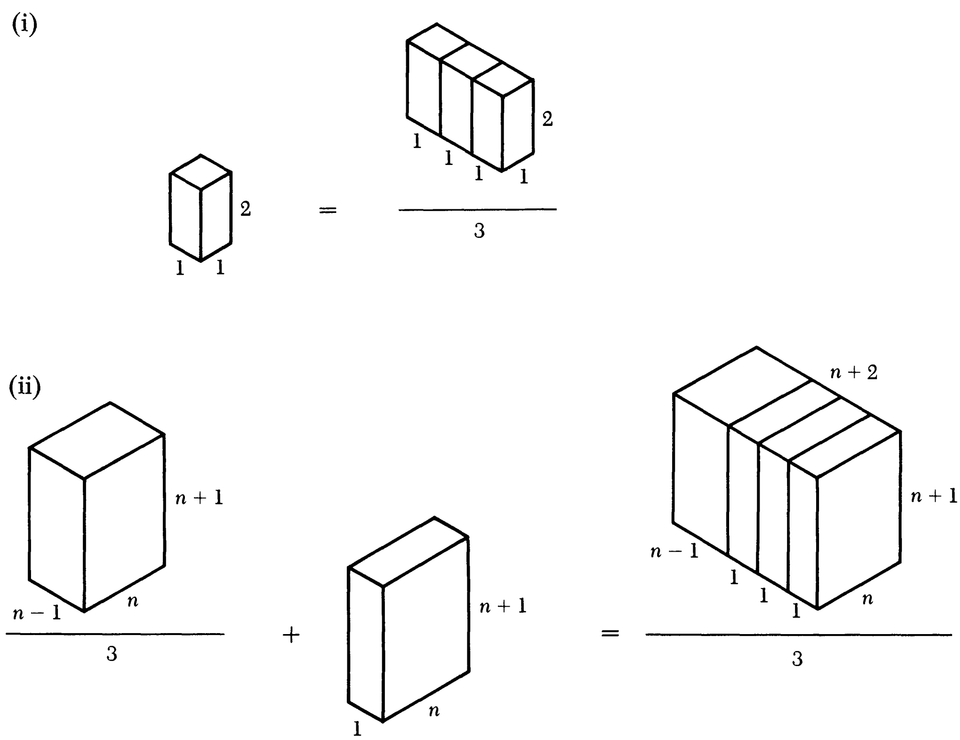
—T. C. WU
York College, CUNY
Jamaica, NY 11451

37. F. Richman, Meaning and information in constructive mathematics, *Amer. Math. Monthly* 89 (1982), 385–388.
38. F. Richman, G. Berg, W. Julian, and R. Mines, The constructive Jordan curve theorem, *Rocky Mountain J. Math.* 5 (1975), 215–236.
39. S. Saks and A. Zygmund, *Analytic Functions*, Drukarnia Naukowa, Warsaw, 1952.
40. G. Stolzenberg, Review: Foundations of Constructive Analysis by E. Bishop, *Bull. Amer. Math. Soc.* 76 (1970), 301–323.
41. D. J. Struik, *A Source Book in Mathematics, 1200-1800*, Cambridge, 1969 (reprint, Princeton, 1986).
42. H. Tietze, *Famous Problems of Mathematics*, Graylock Press, Baltimore, 1965.
43. E. C. Titchmarsh, *The Theory of the Riemann zeta-function*, Oxford, 1951.
44. H. S. Vandiver, Fermat's Last Theorem, *Amer. Math. Monthly* 53 (1946), 555–578.
45. B. L. van der Waerden, *Modern Algebra*, vol. 1, Ungar, New York, 1949. (Original; *Moderne Algebra*, Springer, Berlin, 1931.)

Proof without Words

Theorem. $(1 \times 2) + (2 \times 3) + (3 \times 4) + \cdots + (n-1)n = \dfrac{(n-1)n(n+1)}{3}$, $n = 1, 2, 3, \ldots$.

Proof.



—T. C. Wu
York College, CUNY
Jamaica, NY 11451

NOTES

Differential Invariants of Curves in Projective Space

RICHARD A. SCOTT
Santa Clara University
Santa Clara, CA 95053

Most students first encounter the Euclidean plane as a Cartesian coordinate system with an inner product. This view is particularly suited for the study of angles and lengths. Another, less familiar, approach is to consider the plane as a set of points which can move in certain prescribed ways. These motions are transformations of the plane, and since one motion followed by another is a third permissible motion, this set of transformations forms a group which acts on the plane. This particular type of group is known as a Lie group and provides an ideal structure for the study of curves. The advantage of this view of the Euclidean plane is that it generalizes not only to higher dimensional Euclidean space, but also to different spaces altogether. The purpose of this paper is to exhibit a method of studying curves in two of these spaces known as the projective plane and projective 3-space, where a different Lie group acts on the set of points.

Two curves are equivalent in Euclidean n-space (R^n) if by some rigid motion, one curve can be brought to coincide with the other. These rigid motions, consisting of translations and rotations, form a Lie group which we call the Lie group of Euclidean transformations, and we say two equivalent curves differ by an element of this group. It is reasonable, then, to consider two curves in projective space equivalent if they differ by an element of a Lie group acting on projective space. In order to study these curves in projective space, we first summarize the classification of curves in Euclidean space. Details can be found in [1].

Curves in Euclidean Space

To characterize curves in Euclidean space, we find quantities that are invariant under the action of the group and which are sufficient to classify the curves. Recall from calculus the curvature function, κ, which is determined by the circle that best fits a curve at each point. In the plane, any two equivalent curves, when parameterized by arclength, give rise to the same curvature function. Moreover, any two curves with the same curvature differ by an element of the Euclidean group, so that every curvature function is identified with an equivalence class of curves in the plane. For curves in 3-space, an additional invariant torsion, τ, is necessary to characterize a curve. This quantity measures how much a curve fails to lie in a plane at any given point. In general, a curve in n-space can be classified by $n-1$ invariants.

The mechanism by which invariants determine a curve is a system of differential equations describing how an orthonormal frame moves along the curve. For curves in

3-space, if T, N, and B are the unit tangent, normal, and binormal vectors (as functions of arclength), this system is given by the Serret-Frenet equations

$$\begin{aligned} T' &= \quad -\kappa N \\ N' &= \kappa T \qquad -\tau B \\ B' &= \qquad \tau N. \end{aligned}$$

Spivak [1] gives a generalization to Euclidean n-space.

In order to understand better a typical curve in Euclidean space, we can describe its local behavior by using curves with constant invariants. The circle in the Euclidean plane is an example of one such constant-invariant curve; the regular helix in 3-space is another. For an arbitrary smooth curve in n-space, there is an osculating constant-invariant curve at each point, provided no invariants vanish along the curve. To complete our study of projective curves, then, we will explore the analogous special curves in projective space.

Curves in Projective Space

The elements of projective n-space, P^n, are one dimensional vector subspaces of R^{n+1}, which can be interpreted geometrically as the set of lines through the origin in R^{n+1}. In order to describe the group of transformations acting on P^n, we first introduce the orthogonal group $O(n+1)$ which consists of all rotations and reflections about the origin in R^{n+1}. The group elements are linear transformations which preserve the usual inner product. $O(n+1)$ acts on the set of lines through the origin in R^{n+1}, but ambiguously. That is, if $A \in O(n+1)$, then both A and $-A$ give the same transformation of P^n. A line through the origin can be represented by two unit vectors pointing in opposite directions so the rotation or reflection of a line is determined by either a transformation A applied to one vector or by $-A$ applied to the opposite vector. To avoid this ambiguity, we identify these two transformations by reducing the orthogonal group modulo $C = \{+I, -I\}$ where I is the identity transformation. The transformation group which acts on P^n, then, is $PO(n+1) = O(n+1)/C$. In this paper we restrict our attention to the cases $n = 2$ and $n = 3$: the projective plane and projective space. In each case, we explain the notion of a curve and proceed to find quantities which act as classifying invariants under the actions of $PO(3)$ and $PO(4)$, respectively.

For the projective plane, we see that there is a projective curvature invariant, and we will investigate the curves with constant curvature. Our choice of coordinates for "points" (lines) in P^2 will affect the descriptions of these special curves. These curves will be described easily with points on S^2, the sphere in 3-space. However, in accordance with classical projective geometry, we will also describe these curves using points in the plane R^2 together with points at infinity. Both of these surfaces are locally diffeomorphic to the Euclidean plane so we are justified in our choice of coordinates. A useful way to visualize the interaction of these two coordinate choices is to picture a plane tangent to the top of the sphere. A line through the center of the sphere intersects the sphere at two points and the plane at exactly one point. Identifying these antipodal points of the sphere gives a natural map between the two surfaces. To conclude the paper, we describe the "projective helices" in projective space by using coordinates in R^3, again with points at infinity. The more imaginative reader may wish to visualize these curves on the 3-sphere in R^4.

1. Coordinates and Projective Curves

Consider first the set of lines through the origin in R^3. One of these lines can be described as the set of vectors $\{(ax_0, ay_0, az_0) \in R^3 | a \in R\}$, where (x_0, y_0, z_0) is a point on the line other than the origin. To give coordinates to this set of lines, we pick exactly one point along each line and identify this point with its corresponding line. If we restrict the points to be one unit from the origin, we have the coordinates $(x, y, z) \in R^3$ where $x^2 + y^2 + z^2 = 1$. This, however, gives two coordinates for each line since every line intersects the sphere at two antipodal points. We identify these two points and refer to points on the hemisphere where antipodal points on the equator are considered as the same point.

Another classical way to assign coordinates to the projective plane is to choose points in the plane $z = 1$. For each line $\{(ax, ay, az)|a \in R\}$ we pick the point where $a = 1/z$. This time, however, our selection is too restrictive; we have no coordinates for lines parallel to the plane $z = 1$. To include these lines, we refer to them as points at infinity. Together they form the line at infinity L_∞ which by any non-identity projective transformation becomes a line with coordinates in R^2.

A curve in P^2 can now be defined in terms of curves on the 2-sphere. A curve on S^2 is a function γ defined on a closed interval of R; that is, $\gamma:[a, b] \mapsto R^3$ such that $\gamma(t) = (x(t), y(t), z(t))$ and $x^2 + y^2 + z^2 = 1$. We identify all curves which are reparametrizations of other curves. That is, $\gamma(t)$ is equivalent to $\alpha(t)$ if and only if there is a differentiable function $s(t):[a, b] \mapsto [c, d]$ such that $ds/dt > 0$, $\gamma(a) = \alpha(c)$, $\gamma(b) = \alpha(d)$, and $\gamma(t) = \alpha(s(t))$ for $t \in [a, b]$. A curve in P^2 is defined by identifying "antipodal curves" on the 2-sphere. For the remainder of this paper, we will be working with smooth curves, those with continuously differentiable components.

The generalization to P^3 follows naturally. Coordinates can be assigned on the 3-sphere by restricting to $(x, y, z, w) \in R^4$ where $x^2 + y^2 + z^2 + w^2 = 1$ and antipodal points are identified. They can also be assigned by restricting to the hyperplane $w = 1$ and adjoining the plane at infinity (the set of lines parallel to the hyperplane). A curve in projective space is defined as an equivalence class of P^3-valued functions defined on a closed interval of R. In fact it matters not whether this curve is viewed in terms of the homogeneous coordinates R^3 or the coordinates on S^3.

It should be noted at this point that although the choice of coordinates makes no difference locally (which is what we need for this paper), there is a slight difference globally. Because every projective curve is identified with exactly two curves on the corresponding sphere, the group of transformations acting on P^n is not the same as the group acting on S^n. In fact, the group acting on the n-sphere is $O(n+1)$. We use coordinates on the sphere in order to apply our classification mechanism to curves in projective space.

2. The Curve Apparatus in the Projective Plane

Let $R(t) = (x(t), y(t), z(t))$ with $x^2 + y^2 + z^2 = 1$ give a parametrization of a smooth curve in P^2. That an element $A \in PO(3)$ maps this curve to another in the projective plane is verified by noting that $\langle R, R \rangle = 1$ implies $\langle AR, AR \rangle = 1$ where $\langle\ ,\ \rangle$ is the usual inner product. To represent this equivalence class of curves, we find the reparametrization $t(s)$ such that $\langle dR/ds, dR/ds \rangle = 1$. In the case that dR/dt vanishes at a point or at a finite number of points, the curve can be analyzed piecewise. Therefore, we will assume that dR/dt is non-zero along the curve.

Differentiating R, we have

$$\frac{dR}{ds} = \frac{dR}{dt} \cdot \frac{dt}{ds}$$

so that when $t(s)$ satisfies $|dt/ds|^2 = 1/\langle dR/dt, dR/dt \rangle$, $R(s)$ is correctly parametrized. This parametrization is analogous to the arclength parameter in Euclidean geometry since the condition $\langle dR/dt, dR/dt \rangle = 1$ is preserved by the action of the group. Assuming that $T = R' = dR/dt$ is correctly parametrized, we now have two elements of an orthonormal frame, which follows from differentiating $\langle R, R \rangle = 1$ to obtain $\langle R', R \rangle = \langle T, R \rangle = 0$.

To complete the frame, we differentiate $\langle T, T \rangle = 1$, and find that the new vector, T', is perpendicular to T. The equation $\langle T, R \rangle' = \langle T', R \rangle + \langle T, R' \rangle = 0$ shows that T' has a component in the R direction. There is one degenerate case we wish to avoid. If $T' = -R$ at a finite number of points then the curve can be analyzed piecewise to avoid these singularities. Letting $\kappa N = T' + R$ where κ is the real valued function chosen so that $\langle N, N \rangle = 1$, we now have a complete orthonormal frame $\{R, T, N\}$. Note that if $T' + R = 0$ in some neighborhood, we can complete the frame by letting N be one of two unit vectors orthogonal to R and T, in which case $\kappa = 0$ in that neighborhood. Differentiation of $\langle N, N \rangle = 1$, $\langle N, T \rangle = 0$, and $\langle N, R \rangle = 0$ yields the differential equation $N' = -\kappa T$. In order to characterize the rate of change of the frame as it moves along the curve, we combine the equations from the last two paragraphs to give the system

$$\begin{aligned} R' &= \quad\quad T \\ T' &= -R \quad\quad + \kappa N \\ N' &= \quad\quad -\kappa T. \end{aligned}$$

If Γ is the 3 by 3 matrix formed by $[RTN]$ where R, T, and N are column vectors, the following differential equation results

$$\Gamma' = \Gamma \cdot K \tag{2.1}$$

with

$$K = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & -\kappa \\ 0 & \kappa & 0 \end{bmatrix}.$$

A solution to (2.1) is a matrix valued function which represents a path through the Lie group $PO(3)$. Using this equation, we are able to map (lift) curves in the projective plane to curves in the group. We do most of our work in the group where the curve has a particularly nice form, realizing that curves in the plane can be recovered easily from curves in the group by means of this natural identification.

Just as the analogous scalar equation $y' = a \cdot y$ has an exponential solution, existence of solutions to differential equations tells us that (2.1) has a matrix valued solution $\Gamma(t)$. As with the scalar equation, uniqueness of such a solution is established by a choice of initial conditions $\Gamma(0)$ (see [2]). The columns of $\Gamma(0)$ are the initial values of the curve R, its unit tangent vector T, and so on. A projective transformation of the curve in the plane merely amounts to a different choice of the initial condition $\Gamma(0)$ for the differential equation on the group.

The quantity κ is the differential invariant of curves in the projective plane. Existence and uniqueness of solutions to (2.1) guarantees that with every function $\kappa(t)$ is associated the equivalence class of solutions to the differential equation. These equivalence classes in the group are identified with equivalence classes in the plane by the mapping from Γ to its first column vector. This shows that κ does indeed classify curves in the projective plane.

It should be noted that $O(3)$ consists of matrices with determinant either $+1$ or -1. Because the determinant of the solution to (2.1) is a continuous function, the determinant along the curve takes on only one of these values depending on the choice of initial conditions. Classifying up to the action of $O(3)$, then, makes no distinction between a curve which "turns" in one direction with given invariants and a curve which turns in the opposite direction with the same invariants. This is intuitively consistent with the fact that our Lie group allows reflections. In order to distinguish between two such curves, we could insist that κ be chosen so that the determinant of Γ is $+1$. In this case the group acting is the "special orthogonal group" $SO(3)$, the subgroup of $O(3)$ with determinant $+1$. In fact, $SO(3)$ and $PO(3)$ are isomorphic indicating that the group that preserves orientation on the 2-sphere is the same as the group that acts on the projective plane.

Having defined the projective curvature invariant, we now consider the curves with constant curvature. For an arbitrary curve Γ in $O(3)$ and some fixed point $\Gamma_0 = \Gamma(t_0)$ along the curve, there is a constant-invariant curve which has the same tangent matrix at that point. This curve is given exponentially as

$$A \cdot e^{K_0 t} \quad \text{where } K_0 = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & -\kappa_0 \\ 0 & \kappa_0 & 0 \end{bmatrix}$$

and $\kappa_0 = \kappa(t_0)$. Exponentiation of matrices is defined by the power series

$$e^{Bt} = I + B \cdot t + B^2 \cdot \frac{t^2}{2!} + B^3 \cdot \frac{t^3}{3!} + \cdots .$$

A is the element of $O(3)$ that transfers the curve $e^{K_0 t}$ to a point of tangency. Because different choices of A give equivalent curves, it will be enough to find any one curve with curvature κ_0. Such a curve can be found by exponentiating the matrix K_0 or by noting that the desired curve is a circle on the 2-sphere, given by $R_p(t) = (\cos\theta\cos pt, \cos\theta\sin pt, \sin\theta)$ where θ is determined by the equation $p^2\cos^2\theta = 1$ in order that the curve is correctly parametrized. Computation shows the curvature of R_p to be $\kappa = \sqrt{p^2 - 1}$. Note that choosing $p = 1$ forces θ and κ to be zero, giving a great circle on the sphere (in fact, since we are considering only continuously differentiable curves, $\kappa = 0$ in any neighborhood implies that the curve is a great circle). Moreover, for any choice of curvature, there is a choice of p which gives R_p the desired curvature. In other words, we have not only found one curve with constant curvature, but by taking projective images of the R_p's we can obtain every curve of constant curvature.

To see what these constant curvature functions look like in homogeneous coordinates, we divide the triple R_p by its z component to obtain $R_p(t) = (\cot\theta\cos pt, \cot\theta\sin pt)$. This curve is just a circle centered at the origin with radius $\cot\theta$. To get curves away from the origin, however, we need to rotate the curve on S^2 and then switch coordinates. The resulting curve will be a parabola, a hyperbola, or an ellipse, since the rotated cone intersects the plane in one of these non-circular conic sections. This is verified pictorially by intersecting the cone determined by the circle on the sphere and the center of the sphere with the plane $z = 1$. Examples of curves with constant curvature are shown in FIGURE 2.1.

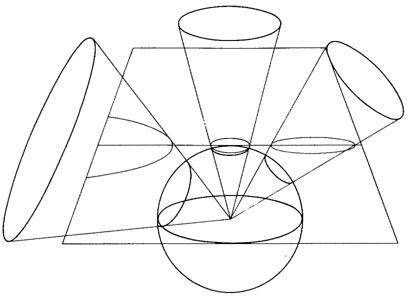


FIGURE 2.1
Curves with constant projective curvature.

3. Projective Space and the Projective Helix

The process of developing an orthonormal frame and computing the rate of change of the frame works just as well for curves in P^3. For a curve R in projective space, we use coordinates on S^3 so that $\langle R, R\rangle = 1$. Differentiating, we obtain the vector $T = R'$ which by a suitable parameter choice is a unit vector perpendicular to R. We compute the "normal" vector N as we did in the projective plane as $\kappa N = T' + R$ and assume N does not vanish at any point along the curve. Because we are now looking for a basis of R^4, we need a fourth vector. Differentiating $\langle N, N\rangle = 1$ shows that the new vector N' is perpendicular to N. The reader can verify that N' has a component $-\kappa$ in the T direction, but no component in the R direction. We choose as our next basis vector $\tau B = N' + \kappa T$ where τ is the real valued function that guarantees $\langle B, B\rangle = 1$.

More differentiation gives the equation $B' = -\tau N$ and we now have a complete set of formulas describing the rate of change of the frame as it moves along the curve. This set is summarized as

$$\begin{aligned} R' &= \quad T \\ T' &= -R \quad + \kappa N \\ N' &= \quad -\kappa T \quad + \tau B \\ B' &= \quad -\tau N \quad . \end{aligned}$$

Letting Γ be the matrix of column vectors $[RTNB]$, this system is reformulated as

$$\Gamma' = \Gamma \cdot K, \tag{3.1}$$

where

$$K = \begin{bmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & -\kappa & 0 \\ 0 & \kappa & 0 & -\tau \\ 0 & 0 & \tau & 0 \end{bmatrix}.$$

Again, existence and uniqueness of solutions to differential equations guarantees that the differential invariants κ and τ are sufficient to classify curves in P^3.

Finding curves with constant invariants is more difficult in 3-space, because we no longer have the visual aid of the sphere S^2. As a generalization of the curve on the 2-sphere, however, we offer a curve with constant projective curvature and torsion,

$$R = (\cos\theta\cos pt, \cos\theta\sin pt, \sin\theta\sin qt, \sin\theta\cos qt) \tag{3.2}$$

where p and q are arbitrary constants subject to the restrictions $|p| > 1$, $|q| < 1$. (R

can also be obtained by exponentiating the matrix in (3.1).) The angle θ is determined by the parametrization of R. In particular, $p^2\cos^2\theta + q^2\sin^2\theta = 1$ implies

$$\theta = \cos^{-1}\sqrt{\frac{1-q^2}{p^2-q^2}}\,.$$

For these curves, formulas for κ and τ are $\kappa = \sqrt{(p^2-1)(1-q^2)}$ and $\tau = p\cdot q$. Additional calculations will show that for any choice of κ and τ, there are suitable choices for p and q.

Before attempting to describe these curves in full, we look at the special case when $\tau = 0$. Substituting $q=0$ into (3.2) gives the family of curves $(\cos\theta\cos pt,$ $\cos\theta\sin pt, 0, \sin\theta)$ in the plane $z=0$. As one would hope, these are the same curves found for the projective plane, and the curvature invariants are the same.

Even if the curvature invariant is not constant, an analytic curve with zero torsion will lie in a plane. To prove this, we construct a frame for such a curve R and realize that R and all of its derivatives can be written as linear combinations of the vectors R, T, and N. The three initial vectors R_0, T_0, and N_0 span a three dimensional subspace of R^4. Letting B_0 be the constant unit vector in the orthogonal complement of this subspace, we consider the Taylor series expansion $\langle R, B_0\rangle = \langle R_0, B_0\rangle + \langle R, B_0\rangle'_{t=0}\cdot t + \langle R, B_0\rangle''_{t=0}\cdot(t^2/2!)\cdots$. With B_0 a constant vector, the differentiation moves to the R inside the inner product. Rewriting each of these R derivatives as linear combinations of R, T, and N and evaluating at $t=0$, we see that because B_0 is perpendicular to the initial R, T, and N vectors, the Taylor series expansion is zero. Because $\langle R, B_0\rangle$ is an analytic function, it must be identically zero. This shows that the position vector of the curve is always perpendicular to B_0, and, consequently, the curve lies in a 3-plane in R^4.

In order to visualize the projective helices, we use the coordinates in R^3 with points at infinity. Dividing by the w (fourth) component gives

$$R = (\cot\theta\sec qt\cos pt, \cot\theta\sec qt\sin pt, \tan qt)$$

where p, q, and θ are as before. This curve in R^3 satisfies the equation

$$\frac{x^2+y^2}{\cot^2\theta} - z^2 = 1,$$

so that R lies on a hyperboloid of one sheet. An example of such a curve is shown in FIGURE 3.1. It can be shown that certain projective transformations map a hyperboloid to other hyperboloids and, consequently, a curve on this hyperboloid to curves on the other hyperboloids. Even more important, though, is the fact that this family of curves includes a representative from every equivalence class of constant-invariant curves in projective space.

Further study of these helices shows them to be periodic if and only if p/q is rational. If $p/q = m/n$ with $m, n \in \mathrm{Z}$ and $(m,n)=1$, then the period of R is $2\pi n/q$. These particular helices have only a finite number of points in the plane at infinity. Because $z(t) = \tan qt$ has period π/q, these helices pass through the plane at infinity $2n$ times before repeating. For the curve with $p=2$ and $q=1/2$, the period is 4π and this curve passes through the plane at infinity twice. This can be seen in FIGURE 3.2. A helix with p/q irrational will have an infinite number of points at infinity and can be shown to be a dense subset of the hyperboloid on which it lies.

The method of classifying curves which we have demonstrated in this paper uses facts about Lie groups to understand the geometric properties of curves. In particular, we knew that the tangent vector to a curve in the group would be a matrix of a given form. A similar analysis can be performed on curves in other spaces provided the Lie

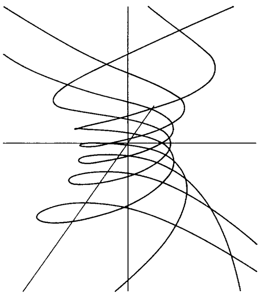


FIGURE 3.1
A projective helix lying on a hyperboloid of one sheet.

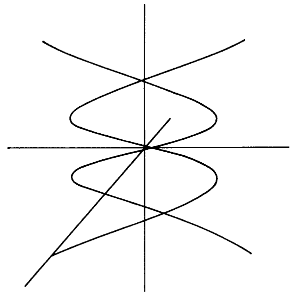


FIGURE 3.2
A projective helix with $p = 2$ and $q = \frac{1}{2}$.

group action is well defined on the space. Another example of such a space is relativistic space-time where the group acting is the Lorentz group. For a discussion of curves in this space, see [3].

The author wishes to acknowledge the faculty guidance of Frank A. Farris and the financial support of the Science Advisory Committee at Santa Clara University.

REFERENCES

1. Michael Spivak, *A Comprehensive Introduction to Differential Geometry*, vol. 2, Publish or Perish Publishing Company, Berkeley, CA. 1979.
2. D. A. Sanchez, R. C. Allen, and W. T. Kyner, *Differential Equations, An Introduction*, Addison-Wesley, Reading, MA. 1983.
3. Frank A. Farris, Classifying curves in space and space-time, *Exposition. Math.* 6 (1988), 81–89.

A Geometrical Approach to Cramer's Rule

J. W. ORR
Crookstown
County Cork, Ireland

Gabriel Cramer (1704–52) published in 1750 [1] his celebrated Rule (already known to Maclaurin [2]) for solving simultaneous equations by means of determinants, which he had obtained by "the science of algebra." It may be of interest to consider how, upon certain assumptions, it may be obtained geometrically.

Let us start with the two simultaneous equations

$$ax + by = k$$
$$cx + dy = l$$

and write these in matrix form

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} k \\ l \end{bmatrix}.$$

Now consider the matrix as representing, not the usual *point* transformation, by

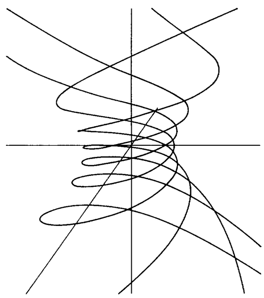


FIGURE 3.1
A projective helix lying on a hyperboloid of one sheet.

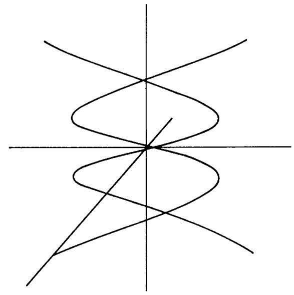


FIGURE 3.2
A projective helix with $p = 2$ and $q = \frac{1}{2}$.

group action is well defined on the space. Another example of such a space is relativistic space-time where the group acting is the Lorentz group. For a discussion of curves in this space, see [3].

The author wishes to acknowledge the faculty guidance of Frank A. Farris and the financial support of the Science Advisory Committee at Santa Clara University.

REFERENCES

1. Michael Spivak, *A Comprehensive Introduction to Differential Geometry*, vol. 2, Publish or Perish Publishing Company, Berkeley, CA. 1979.
2. D. A. Sanchez, R. C. Allen, and W. T. Kyner, *Differential Equations, An Introduction*, Addison-Wesley, Reading, MA. 1983.
3. Frank A. Farris, Classifying curves in space and space-time, *Exposition. Math.* 6 (1988), 81–89.

A Geometrical Approach to Cramer's Rule

J. W. ORR
Crookstown
County Cork, Ireland

Gabriel Cramer (1704–52) published in 1750 [1] his celebrated Rule (already known to Maclaurin [2]) for solving simultaneous equations by means of determinants, which he had obtained by "the science of algebra." It may be of interest to consider how, upon certain assumptions, it may be obtained geometrically.

Let us start with the two simultaneous equations

$$ax + by = k$$
$$cx + dy = l$$

and write these in matrix form

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} k \\ l \end{bmatrix}.$$

Now consider the matrix as representing, not the usual *point* transformation, by which the point at the position (x, y) moves to, or maps into, the position (k, l), all coordinates referring to the same fixed axes, but a *coordinate* or *axis* transformation, by which the fixed point (k, l) is relabelled (x, y) in consequence of a change in the axes and so in the whole coordinate system. (The two transformations are mutually inverse.) The equation now states that the point with coordinates (k, l) relative to the original axes has coordinates (x, y) relative to the transformed axes with unit (basis) vectors (a, c) and (b, d).

(Note: As the basis vectors of a coordinate or axis transformation have unit length in that transformation, they are here referred to as "the unit vectors" of the transformation. Similarly the area or volume spanned by the unit vectors is referred to as "the unit area" or "the unit volume" of the transformation. The word "transformation" refers both to the process of obtaining one set of coordinates from another, and to the coordinate system resulting from that process.)

Now call the original axes $K'OK$ and $L'OL$, and call the points (k, l), (a, c) and (b, d) P, Q and R, respectively. Join OP, OQ and OR and complete the parallelograms $OQSR$ and $OPTR$. Extend TP to meet RS in M and OQ in N.

We now have P as the point with coordinates x and y relative to the transformed axes along OQ and OR, and since NP is parallel to OR,

$$x = \frac{\text{length of } ON}{\text{length of } OQ}$$

$$= \frac{\text{area of parallelogram ONMR}}{\text{area of parallelogram OQSR}}$$

(because the areas of parallelograms of the same height are proportional to the lengths of their bases)

$$= \frac{\text{area of parallelogram } OPTR}{\text{area of parallelogram } OQSR}$$

(because $ONMR$ and $OPTR$ have the same height and the same base OR)

$$= \frac{\begin{vmatrix} k & b \\ l & d \end{vmatrix}}{\begin{vmatrix} a & b \\ c & d \end{vmatrix}}$$

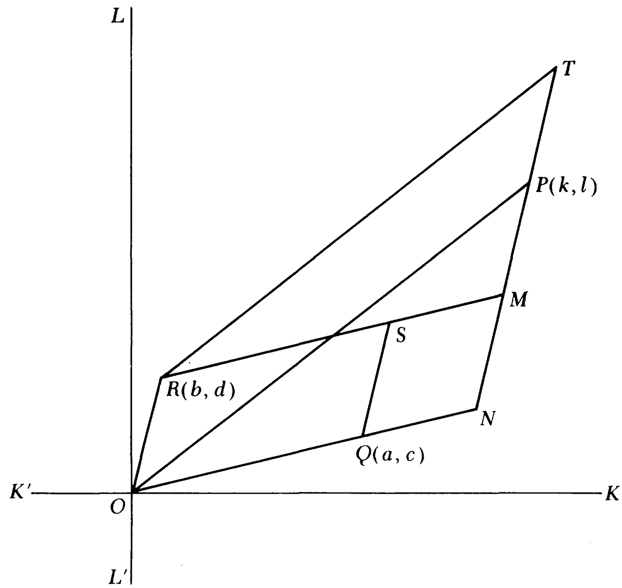
by the interpretation of a determinant as the area of the parallelogram spanned by its column vectors (the unit area of the corresponding transformation).

Similarly by completing the parallelogram on OP and OQ we find that

$$y = \frac{\begin{vmatrix} a & k \\ c & l \end{vmatrix}}{\begin{vmatrix} a & b \\ c & d \end{vmatrix}}.$$

We thus obtain Cramer's Rule for two equations in two unknowns.

The argument holds for all positions of P, Q, and R, provided account is taken of the sense of lines and the sense of description of areas. If OQ, OP and OR are simply rotated from the position in the diagram without transposition, the reasoning is unaffected. If two of them are transposed there is a reversal of (a) the sense of one or other of them, (b) the sense of description of the area spanned by them, and (c) the sign of that area. There are two special cases. If the columns of coefficients are dependent, so that O, Q and R are collinear, but (k, l) is independent, TP and OQ fail to meet, and there is no solution (the determinant of coefficients is zero). If (k, l)



is also dependent, so that P is also collinear, then TP and OQ coincide, and the position of N (determining the value of x) is arbitrary (both determinants are zero). The position of N also determines the value of y, which is given by the quotient of the lengths of NP and OR.

The extension of the argument to three dimensions seems fairly clear, given a suitable convention for determining sign. If we take the volume spanned by the original unit vectors to be positive, then the unit volume of the transformation could be considered positive or negative according to whether the number of pairs of unit vectors transposed by the transformation was even or odd. An equivalent criterion would be the number of unit vectors whose sense was reversed by the transformation. The special cases carry over with the substitution of "coplanar" for "collinear."

(Note: The result for two or three dimensions may be obtained very neatly by writing the original equations in column vector form, representing a coordinate or axis transformation, and taking vector products or scalar triple products. The procedure, however, is possibly more algebraic than geometrical.)

The further extension of the argument to higher dimensions is perhaps allowable by analogy. In each case the coordinate required is given by the quotient of two determinants, that in the denominator being the determinant of coefficients, and that in the numerator being the same except that the constant terms replace the coefficients of the required coordinate. In these determinants

(1) the columns of coefficients of coordinates other than the one required represent the vectors spanning the common base;

(2) the column of coefficients of the required coordinate represents the unit vector along the axis of that coordinate; and

(3) the column of constant terms represents the vector whose projection parallel to the base on the axis of the required coordinate gives the value of that coordinate.

REFERENCES

1. Gabriel Cramer, *Introduction à l'Analyse des Lignes Courbes Algébriques*, Geneva, 1750. See Henrietta Midonick (ed.), *The Treasury of Mathematics*, Philosophical Library, N.Y., 1965.
2. Carl B. Boyer, *A History of Mathematics*, Wiley, N.Y. 1968.

The Metamorphosis of a Manuscript

The following verse was submitted for publication in the MAGAZINE.

Discoveries

The theorem found by our toiler,
Is hot stuff, a genuine "boiler".
But, o what a blow
When learned to know:
The thing was discovered by Euler.

Marta Sved
University of Adelaide
Adelaide S.A. 5001, Australia

The referee responded with the following report with suggested revision:

I like this poem;
It strikes home.
I'm another for whom Euler
Was an antecedent spoiler.

I sent your ditty to G. L. A.,
Along with a little note to say
(And I hope he has it with him)
That you might improve your rhythm.

I thought the line, "When learned to know,"
Was not really what you yearned to show.
And so, not wanting your rhyme to be discrediting,
I nevertheless suggested a little bit of editing.

And the author responded:

Dear Referee,
Unknown to me,
Though "learnt to know"
The fine rhythmic flow
Of your verse in your letter.
I admit, it is better
Than Australian slang
With Hungarian twang.

However,
I think, we well may
Let G. L. A.
Make the decision
About the revision.
For printing the verse as I read it,
The coauthor must be given the credit.
So without stealing the thunder
Of the voice from Down Under

It's my obligation
To make the great opus
A joint publication.

This is the result:

Discoveries

The theorem found by our toiler
Is hot stuff, a genuine "boiler".
But, wouldn't you know,
Alas, what a blow,
The thing was discovered by Euler.

'Twas ever so!

Marta Sved
University of Adelaide
Adelaide S.A. 5001, Australia

Dave Logothetti
Santa Clara University
Santa Clara, CA 95053

The Shape of *e*

MARY COUGHLIN
University of Toledo
Toledo, OH 43606

For the ancient Greeks, mathematics was geometry and, therefore, numbers in particular were conceived as geometric entities. Even today, it is difficult for us to think of a number like π apart from its relation to the circle. Modern mathematics, on the other hand, is more analytic in nature and therefore e, the base of the natural logarithm and a 17th-century invention, is not usually thought of as a geometric entity. Nevertheless, the question of what e might "look like" had the Greeks invented it is an interesting one. While e, like any transcendental number, is not constructible as a length in the classical sense, perhaps it does exist as the area of some "natural," aesthetically pleasing, region. Is there, for example, an elliptic or hyperbolic region which can be identified with e as the circle is identified with π? The fact that e is defined through the use of the hyperbolic functions suggests that our search for a geometric interpretation of e begin with the hyperbola. This note is the result of following that suggestion.

Our search ended with a figure which we call a *tetragram*.[1] (See FIGURE 1.) The convex hull [3] of this tetragram is a square region of area e. A tetragram is a symmetric star-shaped figure obtained from two conjugate rectangular hyperbolas [1]

[1]The tetragram is a four-pointed star. We chose the name on a purely linguistic basis because of the analogy between it and the name of the five-pointed star called the Pythagorean pentagram or triple triangle [2].

It's my obligation
To make the great opus
A joint publication.

This is the result:

Discoveries

The theorem found by our toiler
Is hot stuff, a genuine "boiler".
But, wouldn't you know,
Alas, what a blow,
The thing was discovered by Euler.

'Twas ever so!

Marta Sved
University of Adelaide
Adelaide S.A. 5001, Australia

Dave Logothetti
Santa Clara University
Santa Clara, CA 95053

The Shape of *e*

MARY COUGHLIN
University of Toledo
Toledo, OH 43606

For the ancient Greeks, mathematics was geometry and, therefore, numbers in particular were conceived as geometric entities. Even today, it is difficult for us to think of a number like π apart from its relation to the circle. Modern mathematics, on the other hand, is more analytic in nature and therefore e, the base of the natural logarithm and a 17th-century invention, is not usually thought of as a geometric entity. Nevertheless, the question of what e might "look like" had the Greeks invented it is an interesting one. While e, like any transcendental number, is not constructible as a length in the classical sense, perhaps it does exist as the area of some "natural," aesthetically pleasing, region. Is there, for example, an elliptic or hyperbolic region which can be identified with e as the circle is identified with π? The fact that e is defined through the use of the hyperbolic functions suggests that our search for a geometric interpretation of e begin with the hyperbola. This note is the result of following that suggestion.

Our search ended with a figure which we call a *tetragram*.[1] (See FIGURE 1.) The convex hull [3] of this tetragram is a square region of area e. A tetragram is a symmetric star-shaped figure obtained from two conjugate rectangular hyperbolas [1]

[1]The tetragram is a four-pointed star. We chose the name on a purely linguistic basis because of the analogy between it and the name of the five-pointed star called the Pythagorean pentagram or triple triangle [2].

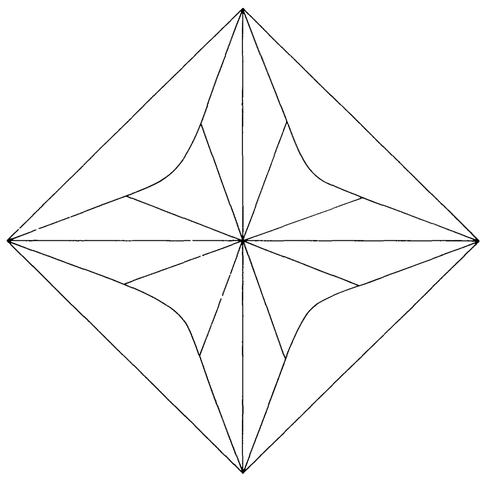


FIGURE 1

with transverse axes of unit length. Since any tangent to a hyperbola intersects an asymptote, tangents that are drawn symmetrically to the two hyperbolas will meet on the asymptotes and form a star. Hence, the region interior to the star is bounded by the two hyperbolas and the line segments which join the hyperbolas to the asymptotes. We choose the points of tangency so that those rays from the center that are either asymptotic to the curves or pass through the points of tangency divide the region into subregions of equal areas. We shall establish the following theorem.

THEOREM. *The total area inside the tetragram is* $3/2$ *and the area of the convex hull of the tetragram is* e.

This entire note is essentially a proof of this theorem.

Although the search for a "picture" of e might be thought of as an enterprise suitable to the ancient Greeks, we have not hesitated to use modern methods in our search. It is the picture we are interested in obtaining, not a construction by straightedge and compass or by more general synthetic methods. We begin with a coordinate-free approach which, although it is not constructive, is more in the "spirit of the Greeks" than the detailed analytic treatment we give later.

For convenience, we begin with a rectangular hyperbola with transverse axis, i.e., the axis which joins the vertices of the hyperbola, of length $2\sqrt{2}$ and center 0. (Later we will reduce this hyperbola to one with transverse axis of length 1.) The tetragram is obtained as follows. First, choose one branch H of the hyperbola and call its vertex V. (See FIGURE 2.) Next, choose a point I on one asymptote so that OI measures one unit. Then choose a point P on H, the branch of the hyperbola, with the following two properties:

1) $\overline{OP}$ and $\overline{VI}$ intersect in a point Q.

2) The area of triangle OQI is equal to the area of the region bounded by $\overline{VQ}$, $\overline{PQ}$, and the hyperbola. We call these regions A_2 and A_1, respectively. The existence of such a point P, i.e., a point such that the regions A_1 and A_2 have equal areas, follows from continuity considerations.

Now, extend $\overline{OI}$ to $\overline{OC}$ so that $\overline{PC} \perp \overline{OC}$, and let A be the point at which lines through $\overline{OV}$ and $\overline{PC}$ intersect. Finally, let B be the point of intersection of the asymptote through segment OI with the tangent to the hyperbola at P. We denote the area of $\triangle OAB$ by α and we will show that $\alpha = e$. Indeed, $\triangle OAB$ provides us with a "picture" of e, but we reject it in favor of a more attractive one.

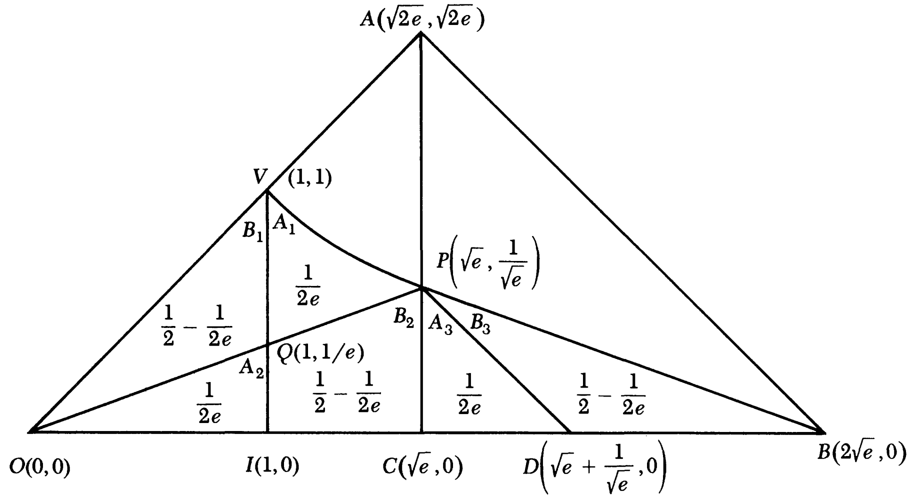


FIGURE 2

In order to facilitate the computation we turn now to the analytic description of the tetragram. We begin with the hyperbola whose cartesian equation is $xy = 1$. H is the branch of the hyperbola in the first quadrant and the corresponding vertex V is $(1,1)$.

At this point we state and prove a lemma which we will use to prove that $\triangle OAB$ has area e. The proof is a simple application of elementary calculus.

LEMMA. *Let P be any point on H and let the tangent at P meet an asymptote at B. Then $\triangle OPB$ is an isosceles triangle with $|\overline{OP}| = |\overline{PB}|$.*

Proof. The slope m of line OP is

$$m = \frac{1/x}{x} = \frac{1}{x^2},$$

and the slope of line PB, tangent to the curve $y = 1/x$, is $-1/x^2$. Therefore, the conclusion follows.

In order to show that α is the number e, we use the fact that the area of $\triangle OPC$ is $1/2$. This is true because any point P on the curve $xy = 1$ has coordinates $(x, 1/x)$; therefore, the area of any triangle OPC equals

$$\frac{1}{2}xy = \frac{1}{2}\left(x \cdot \frac{1}{x}\right) = \frac{1}{2}.$$

For the same reason the area of $\triangle OVI$ is also $1/2$. In addition, since regions A_1 and A_2 were chosen to have equal areas, it follows that the curved regions OVP and $IVPC$ each has area $1/2$. Next, we use the fact that $\triangle OAB$ is an isosceles right triangle to obtain coordinates for C. That the triangle is isosceles follows from the lemma; $\angle OAB$ is a right angle because $\angle AOB \cong \angle ABO$ and each of these angles measures $\pi/4$. Now, let the length of one side of $\angle OAB$ be x. Then the area of $\triangle OAB$ is $(1/2)x^2$; since we denote the area of $\triangle OAB$ by α, we have $\alpha = (1/2)x^2$. It follows that

$$x = \sqrt{2\alpha} \quad \text{and} \quad |\overline{OC}| = x\cos\frac{\pi}{4} = \sqrt{c}\,.$$

Thus, C has coordinates $(\sqrt{\alpha}, 0)$ and the area of $IVPC$ is given by

$$\int_1^{\sqrt{\alpha}} \frac{1}{t}\,dt = \ln\sqrt{\alpha} = \frac{1}{2}\ln\alpha.$$

But the area of $IVPC$ is $1/2$ and, therefore, $(1/2)\ln\alpha = 1/2$. This implies $\ln\alpha = 1$ and therefore, $\alpha = e$.

Thus, as FIGURE 2 shows, we have constructed an isosceles right triangle, $\triangle OAB$, which has area e. The tetragram is obtained by reflecting $\triangle OAB$ in the lines $x = 0$, $y = 0$, $y = \pm x$ and obtaining eight triangles similar to $\triangle OAB$ about the origin. (See FIGURE 3.) These triangles which we obtain from the hyperbolas $xy = \pm 1$ together with appropriate tangents enclose a square region of area $8e$. Because we are looking for a representation of e, we reduce the figure by $1/8$. Thus, the equations of the hyperbolas which underlie this last figure are $xy = \pm 1/8$ and the square which is the convex hull of the tetragram has area e. The reader may wish to show that the regions in FIGURE 3 with the same shading have the same area. In fact, any pair of adjacent regions with one black and one lined vertically or one white and one lined horizontally has area $1/16$. Consequently, in each quadrant the tetragram has three subregions each of area $1/8$. Thus the region bounded by the tetragram has total area $3/2$ and therefore, the region bounded by the square and the tetragram has area $e - 3/2$.

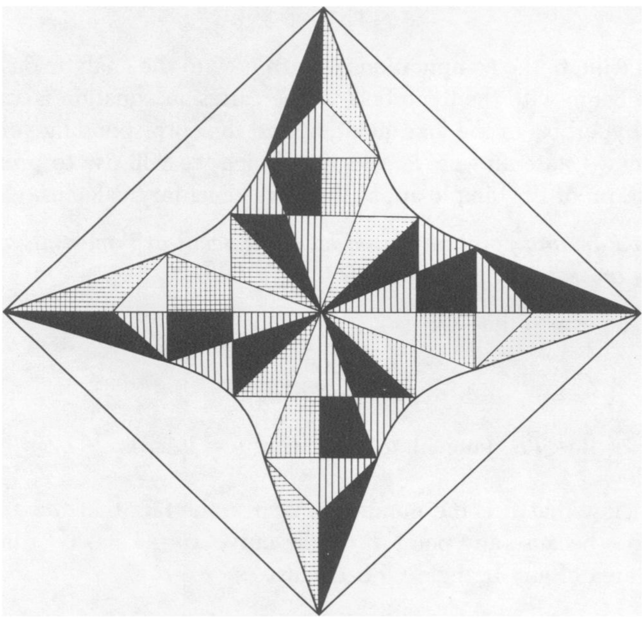


FIGURE 3

Thus, through the use of analytic methods, we have obtained a geometric interpretation of e as the area of the convex hull of the figure we called a tetragram. It is easy to see that this figure cannot be constructed in the classical sense; otherwise, the transcendental number e would be constructible as a length. The tetragram is admittedly more complex than the circle; nevertheless, it is a fairly simple figure with considerable aesthetic appeal. Therefore we may, with some justification, call it the shape of e.

REFERENCES

1. T. M. Apostol, *Calculus*, vol. I, Blaisdell, New York, 1961.
2. H. E. Huntley, *The Divine Proportion*, Dover Publications, Inc., New York, 1979.
3. J. R. Smart, *Modern Geometries*, 2nd edition, Brooks Cole, Monterey, 1978.

Advanced Advanced Calculus: Counting the Discontinuities of a Real-Valued Function with Interval Domain

JOHN KLIPPERT
James Madison University
Harrisonburg, VA 22807

1. Introduction Many of the current textbooks in advanced calculus cover Lebesgue's Criterion for the Riemann integrability of a real-valued function f defined and bounded on the compact interval $[a, b]$ (see e.g. [1], [2]) and, hence, the notions of "countability" and "measure zero." Since the student has learned that a countable set of real numbers has measure zero, by reflecting upon Lebesgue's condition that the set of discontinuities of f must have measure zero, an inquisitive student might ask: "Can we 'count' the various types of discontinuities of f?" Answering this question is a worthwhile project for the student, which involves only a little extra effort (and stays within the realm of advanced calculus). Of course, the question can be dealt with long before a presentation of Lebesgue's Theorem, provided the student is familiar with the concepts of continuity and countability.

2. Preliminaries The student who has studied the concept of countability should be well acquainted with the following definitions and results.

Definition 2.1. Let S be any set and let N be the set of natural numbers. We say that S is *denumerable* $\Leftrightarrow$ there exists a one-to-one correspondence between N and S. Thus S is denumerable if there exists a 1–1 function whose domain is N and whose range is S.

Definition 2.2. If a set S is either finite or denumerable, we say that S is *countable* (otherwise, *uncountable*).

Result 2.1. Between any two distinct real numbers lies a rational number.

Result 2.2. Q, the set of rational numbers, is countable (in fact, denumerable).

Result 2.3. A subset of a countable set is countable.

Result 2.4. If $f\colon A \to B$ is $1-1$ and B is countable, then so is A.

Result 2.5. No interval is countable (if we agree that the empty set and singletons are not considered intervals).

Result 2.6. The union of a countable collection of countable sets is countable.

3. Counting the discontinuities of a real-valued function defined on an interval Suppose f is a real-valued function defined (but not necessarily bounded) on an interval I. Let

$$R = \{x_0 \in I | f \text{ has a } \textit{removable} \text{ discontinuity at } x_0\}.$$

Thus, $x_0 \in R \Leftrightarrow$ there exists a real number L such that

$$\lim_{x \to x_0^+} f(x) = L = \lim_{x \to x_0^-} f(x), \text{ but } L \neq f(x_0).$$

Let

$$J = \{ x_0 \in I | f \text{ has a } \textit{jump} \text{ discontinuity at } x_0 \}.$$

So $x_0 \in J \Leftrightarrow$ there exist real numbers L_1, L_2 such that $\lim_{x \to x_0^+} f(x) = L_1$, $\lim_{x \to x_0^-} f(x) = L_2$, but $L_1 \neq L_2$. Let

$$E = \{ x_0 \in I | f \text{ has an } \textit{essential} \text{ discontinuity at } x_0 \}.$$

In other words, $x_0 \in E \Leftrightarrow \lim_{x \to x_0^+} f(x)$ does not exist or $\lim_{x \to x_0^-} f(x)$ does not exist, where the phrase "does not exist" includes the possibilities $+\infty$ and $-\infty$. Putting $f(x_0^-) = \lim_{x \to x_0^-} f(x)$ and $f(x_0^+) = \lim_{x \to x_0^+} f(x)$, we observe that $E = E_1 \cup E_2 \cup E_3$ where

$$E_1 = \{ x_0 \in I | f(x_0^-) \text{ does not exist and } f(x_0^+) \text{ does not exist} \},$$
$$E_2 = \{ x_0 \in I | f(x_0^-) \text{ exists and } f(x_0^+) \text{ does not exist} \},$$
$$E_3 = \{ x_0 \in I | f(x_0^-) \text{ does not exist and } f(x_0^+) \text{ exists} \}.$$

We call $x_0 \in E_1$ an *essential discontinuity of the first kind* and $x_0 \in E_2 \cup E_3$ an *essential discontinuity of the second kind.*

REMARK. Observe that any endpoint which is contained in I is included in E.

REMARK. If D is the set of discontinuities of f on I, we have $D = R \cup J \cup E_1 \cup E_2 \cup E_3$ where R, J, E_1, E_2 and E_3 are pairwise disjoint.

EXAMPLE. By defining f on [0, 1] by

$$f(x) = \begin{cases} 1, & x \text{ rational} \\ 0, & x \text{ irrational}, \end{cases}$$

and noting that $E = E_1 = [0, 1]$, we have (Result 2.5) that E_1 is uncountable.

This example shows that, in general, we can't "count" the essential discontinuities of the first kind of a real-valued function defined on an interval. We can, however, establish the following:

THEOREM 3.1. *Suppose f is a real-valued function defined (but not necessarily bounded) on an interval I. Then $R \cup J \cup E_2 \cup E_3$ is countable.*

REMARK. If f is monotonic on I, then $D = J$, and we obtain the familiar result that a monotonic function has countably many discontinuities.

4. Theorem 3.1: A discussion and proof An argument for proving that $R \cup J$ is countable is suggested in [3] (p. 100, #17) and in [4] (p. 131, #3). This argument, which uses the density of the set of rational numbers, Q, in the set of real numbers, $\mathbb{R}$ (Result 2.1), involves constructing a vector-valued function from J into a subset of $Q^3 = Q \times Q \times Q$ and showing that this function is 1–1. A similar argument is then used for R. The technique relies upon the existence of *both* one-sided limits for points in $R \cup J$ and, thus, is not applicable to $E_2 \cup E_3$. By introducing the notion of "pointwise oscillation of a function" we can, however, use a different technique that covers $E_2 \cup E_3$ as well as $R \cup J$. Although this notion is typically introduced at the time of the proof of Lebesgue's Theorem, it is certainly appropriate to introduce it earlier as an adjunct to the concept of continuity. During the course of the discussion of "oscillation," the student covers the following definitions and results.

DEFINITION 4.1. If x_0 is in the interior of I, f is bounded on some neighborhood of x_0, and $h > 0$ is small enough so that $[x_0 - h, x_0 + h]$ is contained in this neighborhood, we call

$$W_f[x_0 - h, x_0 + h] = \sup_{x \in [x_0 - h, x_0 + h]} f(x) - \inf_{x \in [x_0 - h, x_0 + h]} f(x)$$

the *oscillation of f on* $[x_0 - h, x_0 + h]$ and $\omega_f(x_0) = \lim_{h \to 0^+} W_f[x_0 - h, x_0 + h]$ the *oscillation of f at* x_0. If f is unbounded on each neighborhood of x_0, we put $\omega_f(x_0) = +\infty$.

REMARK. The obvious modifications are made if x_0 is an endpoint of I.

REMARK. Since (for f bounded near x_0)$W_f[x_0 - h, x_0 + h]$ is a nonnegative, nondecreasing function of h, $\omega_f(x_0)$ is well defined and is nonnegative.

RESULT 4.1. If f is bounded on $[a, b]$, then

$$W_f[a, b] = \sup_{x, y \in [a, b]} |f(x) - f(y)|.$$

RESULT 4.2. Suppose f is defined (but not necessarily bounded) on an interval I and $x_0 \in I$. Then f is continuous at $x_0 \Leftrightarrow \omega_f(x_0) = 0$.

The proof of Theorem 3.1 is a modification of an approach suggested by L. M. Levine in [4] (p.280, #4). We begin with

THEOREM 4.1. *Suppose f is a real-valued function defined (but not necessarily bounded) on an interval I.*

Let $S = R \cup J \cup E_2 = \{x_0 \in I | f(x_0^-)$ exists and f is discontinuous at $x_0\}$. Then S is countable.

We argue Theorem 4.1 as follows. For each natural number n, let $S_n = \{x_0 \in S | \omega_f(x_0) > 1/n\}$. If $x_0 \in S$, then (by Result 4.2) $\omega_f(x_0) > 0$. If $\omega_f(x_0)$ is finite, according to the Archimedean property of the reals, there is a natural number k such that $k \cdot \omega_f(x_0) > 1$. If $\omega_f(x_0) = +\infty$, this same inequality holds for *any* natural number k. Thus $x_0 \in S_k$ for some k and, hence, $x_0 \in \bigcup_{n=1}^{\infty} S_n$. So $S \subseteq \bigcup_{n=1}^{\infty} S_n$, and since clearly $\bigcup_{n=1}^{\infty} S_n \subseteq S$, we have $S = \bigcup_{n=1}^{\infty} S_n$. If we can show that each S_n is countable, then S will also be countable (Result 2.6). To that end, let n be fixed and let $x_0 \in S_n$. We claim that x_0 is the right endpoint of an interval disjoint from S_n and use this observation to argue the countability of S_n. Since $x_0 \in S_n$, $f(x_0^-)$ exists and so there exists a positive number $\delta_n(x_0)$ such that $|f(x) - f(x_0^-)| < 1/2n$ provided $x_0 - \delta_n(x_0) < x < x_0$. We observe that f is then necessarily bounded on $(x_0 - \delta_n(x_0), x_0)$ and that if $x, y \in (x_0 - \delta_n(x_0), x_0)$, then

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

From this it follows that if $a, b \in (x_0 - \delta_n(x_0), x_0)$ and $a < b$, then (by Result 4.1) $W_f[a, b] \leqslant 1/n$. So if $z \in (x_0 - \delta_n(x_0), x_0)$ and h is small enough so that $[z - h, z + h] \subseteq (x_0 - \delta_n(x_0), x_0)$, then $W_f[z - h, z + h] \leqslant 1/n$ and, hence, $\omega_f(z) = \lim_{h \to 0+} W_f[z - h, z + h] \leqslant 1/n$. Letting $I_n(x_0) = (x_0 - \delta_n(x_0), x_0)$, we have thus shown that $I_n(x_0) \cap S_n = \varnothing$ and so *each* $x_0 \in S_n$ is the right endpoint of an interval disjoint from S_n. Finally, if $x_0 \in S_n$, let $q_n(x_0)$ be a rational number such that $q_n(x_0) \in I_n(x_0)$ (Result 2.1) and define a function $F: S_n \to Q$ by $F(x_0) = q_n(x_0)$. To see that F is 1–1, let $x_0, z_0 \in S_n$ and without any loss of generality suppose $x_0 < z_0$. Then $I_n(x_0) \cap I_n(z_0) = \varnothing$. If not, we would have $x_0 \in I_n(z_0)$ (since x_0 is the right endpoint of $I_n(x_0)$) and thus $I_n(z_0) \cap S_n \neq \varnothing$, which contradicts the construction of $I_n(z_0)$. Since $I_n(x_0) \cap I_n(z_0) = \varnothing$, $q_n(x_0) \neq q_n(z_0)$ and F is 1–1. By Results 2.2 and 2.4, S_n is countable.

We finish this section by offering the following corollaries.

COROLLARY 4.1. *Suppose f is a real-valued function defined (but not necessarily bounded) on an interval I. Then E_3 is countable.*

COROLLARY 4.2. (THEOREM 3.1). *Suppose f is a real-valued function defined (but not necessarily bounded) on an interval I. Then $R \cup J \cup E_2 \cup E_3$ is countable.*

The second corollary (Theorem 3.1) is an immediate consequence of Theorem 4.1, the first corollary, and Result 2.6. To prove the first corollary, recall that $E_3 = \{x_0 \in I \mid f(x_0^+)$ exists and $f(x_0^-)$ does not exist$\}$ and put $\mathscr{I} = \{x \mid -x \in I\}$. On $\mathscr{I}$, let $g(x) = f(-x)$. Then $E_3 = E_2(g)$. Applying Theorem 4.1 to g on $\mathscr{I}$ and then using Result 2.3 gives us that $E_2(g)$ is countable and hence so is E_3.

Finally we observe that Theorem 3.1 tells us that the only discontinuities which are not *necessarily* countable are the essential discontinuities of the first kind.

5. A "distillation" of Lebesgue's theorem By the time a student has finished studying the proof of Lebesgue's Theorem, he or she should be well acquainted with the following definition and results.

DEFINITION 5.1. Let $\mathbb{R}$ be the set of real numbers and suppose $S \subseteq \mathbb{R}$. We say that S is of *Lebesgue measure zero* $\Leftrightarrow$ for every $\varepsilon > 0$, there exists a countable collection of open intervals $I_1, I_2, \ldots$, each of finite length such that $S \subseteq \bigcup_1^\infty I_i$ and $\sum_1^\infty |I_i| < \varepsilon$ where $|I_i|$ = length of I_i. If S has Lebesgue measure zero, we write $m(S) = 0$.

RESULT 5.1. If $A \subseteq B \subseteq \mathbb{R}$ and $m(B) = 0$, then $m(A) = 0$.

RESULT 5.2. Suppose $S_i \subseteq \mathbb{R}$ and $m(S_i) = 0$, for each $i \in N$. If $S = \bigcup_1^\infty S_i$, then $m(S) = 0$.

RESULT 5.3. If S is a countable subset of $\mathbb{R}$, then $m(S) = 0$.

REMARK. Although a discussion of the Cantor set is outside the realm of most advanced calculus courses, the student can at least be made aware that the converse in Result 5.3 is false.

RESULT 5.4. No interval has Lebesgue measure zero (if we agree that $\varnothing$ and singletons are not considered intervals).

RESULT 5.5. (LEBESGUE'S THEOREM) Suppose f is a real-valued function defined and bounded on $[a, b]$. Let $D = \{x \in [a, b] \mid f$ is discontinuous at $x\}$. Then

$$f \text{ is integrable on } [a, b]$$

$$\Leftrightarrow$$

$$m(D) = 0.$$

Now let's suppose f is a real-valued function defined and bounded on $[a, b]$. As previously observed (Section 3), $D = R \cup J \cup E_1 \cup E_2 \cup E_3$ where R, J, E_1, E_2 and E_3 are pairwise disjoint. According to Lebesgue's Theorem, if f is integrable on $[a, b]$, then $m(D) = 0$ and since $E_1 \subseteq D$, we'll have (by Result 5.1) that

$$f \text{ is integrable on } [a, b] \\ \Rightarrow \\ m(E_1) = 0.$$

On the other hand, by Theorem 3.1, $R \cup J \cup E_2 \cup E_3$ is countable and, hence, (Result 5.3) $m(R \cup J \cup E_2 \cup E_3) = 0$. So *if* $m(E_1) = 0$, we have (Result 5.2) that $m(D) = m(R \cup J \cup E_1 \cup E_2 \cup E_3) = 0$. Hence,

$$m(E_1) = 0 \\ \Rightarrow \\ f \text{ is integrable on } [a, b],$$

and we have established our "distilled" Lebesgue's Theorem:

RESULT 5.6. (LEBESGUE'S THEOREM, second version) Suppose f is a real-valued function defined and bounded on $[a, b]$. Then

$$f \text{ is integrable on } [a, b] \\ \Leftrightarrow \\ m(E_1) = 0.$$

REMARK. The crux of the second version is that if f is *any* real-valued function defined and bounded on $[a, b]$, then, a priori, $m(R) = m(J) = m(E_2) = m(E_3) = 0$. Of course we need not have $m(E_1) = 0$. Again, let f be defined on $[0, 1]$ by

$$f(x) = \begin{cases} 1, & x \text{ rational} \\ 0, & x \text{ irrational.} \end{cases}$$

Then $E_1 = [0, 1]$ and so (by Result 5.4) $m(E_1) \neq 0$.

6. An alternate approach to counting R ∪ J An approach for dealing solely with $R \cup J$ is mentioned at the beginning of Section 4 and is suggested in [3] and [4]. As a challenge, the student might consider the following alternate approach to this problem:

1) Prove:

LEMMA 1. *Any collection of disjoint open intervals is countable.*

2) Use Lemma 1 to prove:

LEMMA 2. *Let* $\mathbb{R}$ *be the set of real numbers and suppose* $S \subseteq \mathbb{R}$. *Let* $S' = \{x_0 \in \mathbb{R} \mid x_0$ *is an accumulation point of* $S\}$. *Suppose* $S \cap S' = \varnothing$. *Then* S *is countable.*

3) Let I be any interval and put $S_n = R_n \cup J_n$, $n = 1, 2, \ldots,$ where

$$R_n = \left\{ x_0 \in I \,\middle|\, |f(x_0) - \lim_{x \to x_0} f(x)| > \frac{1}{n} \right\},$$

and $J_n = \{x_0 \in I \mid |f(x_0^+) - f(x_0^-)| > 1/n\}$. Show that $R \cup J = \bigcup_{n=1}^{\infty} S_n$ and use Lemma 2 to prove that each S_n is countable. Now use the fact that the union of a countable collection of countable sets is countable to establish the following:

THEOREM. *Let f be a real-valued function (not necessarily bounded) defined on any interval. Then $R \cup J$ is countable.*

REFERENCES

1. Nicholas J. DeLillo, *Advanced Calculus with Applications*, Macmillan, New York, 1982.
2. William F. Trench, *Advanced Calculus*, Harper and Row, New York, 1978.
3. Walter Rudin, *Principles of Mathematical Analysis*, third edition, McGraw-Hill, New York, 1976, p. 100, #17.
4. Karl R. Stromberg, *An Introduction to Classical Real Analysis*, Wadsworth, Belmont, California, 1981, p. 131, #3, p. 280, #24.

Rings on R^2

H. HOLDEN
Gonzaga University
Spokane, WA 99258

A common example of a multiplication of ordered pairs is afforded by the complex numbers when we identify the ordered pair of real numbers (a, b) with $a + bi$. The multiplication is $(a, b) * (c, d) = (ac - bd, ad + bc)$. Another example is provided by the rational numbers extended by the square root of 2. If we identify the ordered pair of rational numbers (a, b) with $a + b\sqrt{2}$ we have $(a, b) * (c, d) = (ac + 2bd, ad + bc)$.

How many ways can we define a multiplication of ordered pairs which will result in a ring and which of these rings will be fields? Can we generalize our results to analyze the multiplication of ordered triples?

The intent of this note is to address these questions using the basic concepts of linear algebra. We will construct all the multiplications for ordered pairs of real numbers and give an elementary proof of the fact that there is no three-dimensional analogue for the complex numbers.

Notation: R^2 and R^3 will denote the vector spaces on $R \times R$ and $R \times R \times R$ over the field R of *real numbers*.

Let $N[R^2]$ be any ring defined on $R \times R$ with identity $(1,0)$. Let $X, Y, Z \in N[R^2]$ and $r \in R$. We will assume $N[R^2]$ satisfies the following conditions in addition to the axioms for a ring.

a) Addition in $N[R^2]$ is vector addition in R^2.

b) The real numbers are imbedded as ordered pairs of the form $(r,0)$ and $(r,0) * X = X * (r,0)$ for all $X \in N[R^2]$ and $r \in R$.

c) Scalar multiplication in the vector space R^2 is identified with multiplication in and $J_n = \{x_0 \in I \mid |f(x_0^+) - f(x_0^-)| > 1/n\}$. Show that $R \cup J = \bigcup_{n=1}^{\infty} S_n$ and use Lemma 2 to prove that each S_n is countable. Now use the fact that the union of a countable collection of countable sets is countable to establish the following:

THEOREM. *Let f be a real-valued function (not necessarily bounded) defined on any interval. Then $R \cup J$ is countable.*

REFERENCES

1. Nicholas J. DeLillo, *Advanced Calculus with Applications*, Macmillan, New York, 1982.
2. William F. Trench, *Advanced Calculus*, Harper and Row, New York, 1978.
3. Walter Rudin, *Principles of Mathematical Analysis*, third edition, McGraw-Hill, New York, 1976, p. 100, #17.
4. Karl R. Stromberg, *An Introduction to Classical Real Analysis*, Wadsworth, Belmont, California, 1981, p. 131, #3, p. 280, #24.

Rings on R^2

H. HOLDEN
Gonzaga University
Spokane, WA 99258

A common example of a multiplication of ordered pairs is afforded by the complex numbers when we identify the ordered pair of real numbers (a, b) with $a + bi$. The multiplication is $(a, b)*(c, d) = (ac - bd, ad + bc)$. Another example is provided by the rational numbers extended by the square root of 2. If we identify the ordered pair of rational numbers (a, b) with $a + b\sqrt{2}$ we have $(a, b)*(c, d) = (ac + 2bd, ad + bc)$.

How many ways can we define a multiplication of ordered pairs which will result in a ring and which of these rings will be fields? Can we generalize our results to analyze the multiplication of ordered triples?

The intent of this note is to address these questions using the basic concepts of linear algebra. We will construct all the multiplications for ordered pairs of real numbers and give an elementary proof of the fact that there is no three-dimensional analogue for the complex numbers.

Notation: R^2 and R^3 will denote the vector spaces on $R \times R$ and $R \times R \times R$ over the field R of *real numbers*.

Let $N[R^2]$ be any ring defined on $R \times R$ with identity $(1,0)$. Let $X, Y, Z \in N[R^2]$ and $r \in R$. We will assume $N[R^2]$ satisfies the following conditions in addition to the axioms for a ring.

a) Addition in $N[R^2]$ is vector addition in R^2.

b) The real numbers are imbedded as ordered pairs of the form $(r,0)$ and $(r,0)*X = X*(r,0)$ for all $X \in N[R^2]$ and $r \in R$.

c) Scalar multiplication in the vector space R^2 is identified with multiplication in

$N[R^2]$ as follows: $rX = (r,0) * X$. Since $(a,b) = a(1,0) + b(0,1)$ and $(c,d) = c(1,0) + d(0,1)$, multiplication in $N[R^2]$ is uniquely determined by the product of $(0,1)$ with itself. If $(0,1)*(0,1) = (p,q)$ then the multiplication is given by

$$(a,b)*(c,d) = (ac + pbd, ad + bc + qbd). \tag{1}$$

What choices of p and q will satisfy the ring axioms for multiplication? It is not immediately clear that they can be chosen independently. There may be some constraints imposed by associativity, however, we shall soon show that any choice of p and q produces an associative multiplication. We will denote the ring $N[R^2]$ with this multiplication by $N[R^2, p, q]$. The ring of complex numbers is $N[R^2, -1, 0]$.

If $X \in N[R^2, p, q]$ does X have a multiplicative inverse? Is X a divisor of zero? If each X in $N[R^2, p, q]$ could be represented by a matrix these would be easy questions to answer. So, we will proceed to associate a 2×2 matrix with each element of $N[R^2, p, q]$ by observing the following:

$$X*((r,0)*Y) = (r,0)*(X*Y) \quad \text{and} \quad X*(Y+Z) = X*Y + X*Z. \tag{2}$$

The first equation follows from the associativity of $*$ and the commutativity of scalars. The second is the distributive law of the ring. Next we recall that a mapping F from R^2 to R^2 is linear if

$$F(rY) = rF(Y) \quad \text{and} \quad F(Y+Z) = F(Y) + F(Z). \tag{3}$$

Comparing (2) and (3) we see that left multiplication by any X in the ring $N[R^2, p, q]$ is a linear transformation of the vector space R^2! Thus we can associate a 2×2 matrix with each element of $N[R^2, p, q]$. We will denote this matrix by $M(X)$. If we treat ordered pairs as column vectors then $X * Y$ will be computed as $M(X)Y^T$ and the columns of $M(X)$ will be the images of $(1,0)$ and $(0,1)$, namely, $X*(1,0)$ and $X*(0,1)$. The matrices which correspond to $(1,0)$ and $(0,1)$ are given by:

$$M(1,0) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad M(0,1) = \begin{pmatrix} 0 & p \\ 1 & q \end{pmatrix}.$$

It is easy to verify that $M(X+Y) = M(X) + M(Y)$ and $M(rX) = rM(X)$ by examining these matrix equations column by column. The first column of $M(X+Y)$ is $(X+Y)*(1,0)$ and the first column of $M(X) + M(Y)$ is $X*(1,0) + Y*(1,0)$. The verification is easily completed in this manner and thus:

$$M(X) = M(a(1,0) + b(0,1)) = \begin{pmatrix} a & pb \\ b & a+qb \end{pmatrix}.$$

If $X = (a,b)$ and $Y = (c,d)$ then $X * Y$ is computed as follows.

$$X*Y = \begin{pmatrix} a & pb \\ b & a+qb \end{pmatrix}\begin{pmatrix} c \\ d \end{pmatrix} = \begin{pmatrix} ac+pbd \\ ad+bc+qbd \end{pmatrix}. \tag{4}$$

Thus, we see that X is a left zero divisor iff the kernel of $M(X)$ is nontrivial and X has a right inverse iff $M(X)$ is invertible.

Let $S = \{M(X) | X \in N[R^2, p, q]\}$. It is easy to see that S is closed under matrix addition. It is somewhat surprising that S is also closed under matrix multiplication for any choice of p and q *since a direct computation shows that* $M(X)M(Y) = M(X*Y)$. Hence, S forms a subring of the ring of 2×2 matrices. We will denote this subring by $M[R^2, p, q]$. *Next we establish an isomorphism between* $N[R^2, p, q]$ *and the ring of* *matrices* $M[R^2, p, q]$. *Since matrix multiplication is associative this isomorphism will induce associativity in* $N[R^2, p, q]$ *for all* p *and* q.

THEOREM 1. $M[R^2, p, q]$ *is isomorphic to* $N[R^2, p, q]$.

Proof. Define $M: N[R^2, p, q] \to M[R^2, p, q]$ by the matrix $M(X)$ defined as above. Since the first column of $M(X)$ is X itself, it is easy to see that M is 1:1 and onto. *Note that the second column is uniquely determined by the values of* a *and* b *in the first column and the choice of* p *and* q.

$M(X+Y) = M(X) + M(Y)$ and $M(X * Y) = M(X) * M(Y)$ are verified directly by matrix arithmetic.

The multiplication in $N[R^2, p, q]$ given by (1) is commutative by symmetry and thus $M[R^2, p, q]$ is also commutative. We can confirm this by verifying that, for fixed p and q, all the matrices in $M[R^2, p, q]$ have the same eigenvectors, namely,

$$\left(p, \left(q + \sqrt{q^2+4p}\right)/2\right) \quad \text{and} \quad \left(p, \left(q - \sqrt{q^2+4p}\right)/2\right). \tag{5}$$

THEOREM 2. $N[R^2, p, q]$ *is a field iff* $q^2 + 4p < 0$.

Proof. Let $X = (a, b)$ and $X \neq (0,0)$. $\text{Det}(M(X)) = a^2 + qab - pb^2$. $\text{Det}(M(X)) \neq 0$ iff $q^2 + 4p < 0$.

If X has an inverse then $M(X^{-1})$ must be $M(X)^{-1}$. If $q^2 + 4p \geqslant 0$ and p and q are not both zero then there exists an eigenvector Y (see (5)) such that $M(X)Y = kY = (kI)Y$ and even right cancellation fails.

When $\text{Det}(M(X))$ is a quadratic form which is never zero we are able to define a field for ordered pairs of reals. If $\text{Det}(M(X))$ were a cubic then there would always be at least one real zero and hence a zero divisor. With this in mind, we proceed to construct a subring of the ring of 3×3 real matrices for any ring of ordered triples of reals.

Let $N[R^3]$ be any ring defined on $R \times R \times R$ with identity $(1,0,0)$. We will assume the following.

a) Addition in N is the addition of the vector space R^3.

b) The real numbers are imbedded as ordered triples of the form $(r,0,0)$ and $(r,0,0) * X = X * (r,0,0)$ for all $X \in R^3$ and $r \in R^3$.

c) Scalar multiplication in the vector space R^3 is identified with multiplication in N as follows: $rX = (r,0,0) * X$.

The multiplication in $N[R^3]$ is completely determined by the products of the elements in the canonical basis. We will assume the following multiplication table.

$*$	$(1,0,0)$	$(0,1,0)$	$(0,0,1)$
$(1,0,0)$	$(1,0,0)$	$(0,1,0)$	$(0,0,1)$
$(0,1,0)$	$(0,1,0)$	(p_1, p_2, p_3)	(q_1, q_2, q_3)
$(0,0,1)$	$(0,0,1)$	(r_1, r_2, r_3)	(s_1, s_2, s_3)

We note that it is no longer possible to make arbitrary choices for all the above column vectors and still have a multiplication which satisfies the associative law. *If associativity holds then it imposes restrictions on the products of* $(0,1,0)$ *and* $(0,0,1)$. *For example, if* p_3 *and* s_2 *are nonzero then* $r_1 = q_1$, $r_2 = q_2$, *and* $r_3 = q_3$ *as the reader can verify by examining the cubes of* $(0,1,0)$ *and* $(0,0,1)$.

We can again interpret left multiplication as a linear transformation and we can associate a matrix $M(X)$ with each ordered triple X.

We will treat ordered triples as column vectors and compute $X * Y$ as $M(X)Y$. The matrices of the canonical basis elements are shown below.

$$\begin{array}{cccc} X: & (1,\ 0,\ 0) & (0,\ 1,\ 0) & (0,\ 0,\ 1) \\ M(X): & \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} & \begin{pmatrix} 0 & p_1 & q_1 \\ 1 & p_2 & q_2 \\ 0 & p_3 & q_3 \end{pmatrix} & \begin{pmatrix} 0 & r_1 & s_1 \\ 0 & r_2 & s_2 \\ 1 & r_3 & s_3 \end{pmatrix}. \end{array}$$

We can still show $M(X+Y) = M(X) + M(Y)$ and $M(rX) = rM(X)$ for any choice of the p, q, r, and s vectors as we did earlier with $N[R^2]$.

If $X = (a, b, c)$ then

$$M(X) = \begin{pmatrix} a & bp_1 + cr_1 & bq_1 + cs_1 \\ b & a + bp_2 + cr_2 & bq_2 + cs_2 \\ c & bp_3 + cr_3 & a + bq_3 + cs_3 \end{pmatrix}$$

and

$$\begin{aligned} \mathrm{Det}(M(X)) = a^3 &+ (p_2 + q_3)a^2b + (r_2 + s_3)a^2c + (p_2q_3 - p_3q_2 - p_1)ab^2 \\ &+ (r_2s_3 - r_3s_2 - s_1)ac^2 + (p_2s_3 - p_3s_2 - q_1 - r_1)abc \\ &+ (r_3s_1 - r_1s_3 + p_1s_2 - p_2s_1)bc^2 + (p_3s_1 - p_1s_3 + p_1q_2 - p_2q_1)b^2c \\ &+ (p_3q_1 - p_1q_3)b^3 + (r_1s_2 - r_2s_1)c^3. \end{aligned}$$

Using theorem 1 we took advantage of associativity in $M[R^2]$ to establish associativity in $N[R^2]$. This time choose the products of the basis elements in $N[R^3]$ so that multiplication is associative in $N[R^3]$ and construct the isomorphic subring of 3×3 matrices. Now the value of the isomorphism comes from the determinant in $M[R^3]$ which can be used to detect the presence of zero divisors back in $N[R^3]$.

If $\mathrm{Det}(M(X)) = 0$ then there exists a nonzero Y such that $M(X)Y = (0,0,0)$ and thus $X * Y = 0$. If we choose $X = (x,1,0)$ then $\mathrm{Det}(M(X))$ is a monic cubic in x which has at least one real root and hence there is always a zero divisor of the form $(x,1,0)$ which establishes the fact that there are no fields on R^3.

REFERENCES

The problem of constructing a multiplication for a vector space is discussed in the literature under the subject heading of hypercomplex numbers or linear associative algebra. A good narrative description of the development of this topic can be found in M. Kline, *Mathematical Thought from Ancient to Modern Times*, Oxford University Press, 1972, Chapter 32, and E. Kramer, *The Nature and Growth of Modern Mathematics*, Hawthorn Books, 1970, Chapter 28.

The search initiated by Hamilton and Grassmann for a three-dimensional analogue of the complex numbers was proven to be futile by Frobenius in 1878. A nice presentation of this theorem of Frobenius appears in I. Herstein, *Topics in Algebra*, Blaisdell, 1964, Chapter 7.

A brief introduction to hypercomplex numbers may be found in B. van der Waerden, *Modern Algebra*, Vol. I, Frederick Ungar, 1953. References to a more complete treatment of the subject can be found in the bibliography at the end of chapter 32 in the text by Kline.

Further material of historical and technical interest appears in M. Crowe, *A History of Vector Analysis*, University of Notre Dame Press, 1967, which has recently become available as a Dover reprint.

Fields with the Simple Binomial Theorem

DAVID E. DOBBS
University of Tennessee
Knoxville, Tennessee 37996

1. Introduction

Our starting point is a famous mistake often made by students of elementary algebra. Widely known as the "freshman exponentiation law" (FEL), it would simplify $(a+b)^n$ to a^n+b^n. Of course, this is wrong: consider the real numbers $a=b=1$ and $n=2$! What is *right* is the binomial theorem: if R is a commutative ring, a and b elements of R, and n any positive integer, then

$$(a+b)^n = a^n + \sum_{i=1}^{n-1} \binom{n}{i} a^{n-i}b^i + b^n.$$

Notice that this assertion *sometimes* reduces to FEL, making FEL right! For instance, it is well known (cf. [1, Theorem 1, p. 177]) that if R is a (commutative integral) domain of characteristic $p>0$, then $(a+b)^p = a^p+b^p$ for all $a, b \in R$.

The following definition will help to focus matters. If $n \geqslant 2$ is a positive integer and R is a domain (resp., field), then R is called an **n-domain** (resp., **n-field**) if $(a+b)^n = a^n+b^n$ for all $a, b \in R$. *Our basic question is*: *what are the n-domains*? Surely the field of real numbers isn't an n-field but $\mathbb{F}_2$ *is*, for each $n \geqslant 2$. (In general, $\mathbb{F}_m$ denotes the field of cardinality m.) By the above remark, each domain of positive characteristic p is a p-domain. As we noted in [2, Exercise 6(c), p. 10], the 2-domains are precisely the domains of characteristic 2, but the list of 3-domains includes $\mathbb{F}_2$ as well as the domains of characteristic 3. It is possible, by lengthy *ad hoc* considerations, to complete the following table of all (up to isomorphism) n-domains for $2 \leqslant n \leqslant 10$. ($Ch(R)$ will denote characteristic of R.)

n	n-domains R
2	$ch(R)=2$
3	$\mathbb{F}_2$; $ch(R)=3$
4	$ch(R)=2$
5	$\mathbb{F}_2$; $\mathbb{F}_3$; $\mathbb{F}_4$; $ch(R)=5$
6	$\mathbb{F}_2$
7	$\mathbb{F}_2$; $\mathbb{F}_3$; $\mathbb{F}_4$; $ch(R)=7$
8	$ch(R)=2$
9	$\mathbb{F}_2$; $\mathbb{F}_5$; $\mathbb{F}_8$; $ch(R)=3$
10	$\mathbb{F}_2$; $\mathbb{F}_4$

One purpose of this article is to develop enough theory to justify speedily the above table with a minimum of *ad hoc* calculations.

It is shown in Corollary 2.4 that, apart from the types of n-domains suggested above, each positive integer $n \geqslant 2$ admits only finitely many "exceptional" n-domains (up to isomorphism), and each of these is a finite field of cardinality less than n. The "rarity" theme is pursued in Corollary 2.6: up to isomorphism, $\mathbb{F}_2$ is the only domain that is an n-domain for three consecutive values of n. But "rarity" is relative, for Section 3 develops two infinite classes of "exceptional" n-domains. Nevertheless, Sections 3 and 4 complete the justification of the above table, using in particular some congruences derived via elementary group theory in Theorem 4.2. For readers interested in the (avoidable) *ad hoc* reasoning by which the above table might be built from scratch, an invitation is provided in Remark 4.3.

Although this article has serious goals, it intentionally has modest prerequisites. To increase accessibility, we have chosen as references two concrete introductions to abstract algebra. Beyond the rudiments, the reader should know Lagrange's theorem in group theory and the structure of finite fields, both of which are referenced below as needed.

2. Exceptions are Rare

To indicate some of the special flavor of the subject, we begin by showing that each n-domain R has positive characteristic. To prove (the contrapositive of) this, suppose that $\mathrm{ch}(R)=0$. Then R contains an isomorphic copy of $\mathbb{Z}$, in which $2^n \neq 2$; thus $(1+1)^n \neq 1^n + 1^n$ in R.

We next give some additional useful results and pertinent examples.

PROPOSITION 2.1. (a) *A domain R is an n-domain if and only if the quotient field of R is an n-field.*

(b) *Let $n \geqslant 2$ be a positive integer. Then a field F is an n-field if and only if $(a+1)^n = a^n + 1$ for each $a \in F$.*

(c) *$\mathbb{F}_2$ is an n-field for each positive integer $n \geqslant 2$.*

(d) *If R is a domain such that $p = ch(R) > 0$ and if $n = p^k$ for some positive integer k, then R is an n-domain.*

Proof. (a) Just the "only if" assertion needs attention, since each subring of an n-field is an n-domain. Let F be the quotient field of an n-domain R. Consider $\alpha, \beta \in F$; choose $a, b, c \neq 0$ in R so that $\alpha = ac^{-1}$ and $\beta = bc^{-1}$. Then $(\alpha+\beta)^n = ((a+b)c^{-1})^n = (a+b)^n c^{-n} = (a^n + b^n)c^{-n} = (ac^{-1})^n + (bc^{-1})^n = \alpha^n + \beta^n$, as desired.

(b) Again, just the "only if" part needs a proof. To see that $(d+e)^n = d^n + e^n$ for each $d, e \in F$, note first that $e \neq 0$ without loss of generality. Then repeat the calculation in the proof of (a), using $(a+1)^n = a^n + 1$, with $R = F$, $\alpha = d$, $\beta = e$, $a = de^{-1}$, $b = 1$ and $c = e^{-1}$.

(c) This may be checked easily for all choices of parameters. Alternately, one verifies the criterion from (b). If $a = 0$, this assertion is plain. In the remaining case, $a = 1$, it suffices to remark that $1 + 1 = 0$ in $\mathbb{F}_2$.

(d) By (a), we need only show $(a+b)^{p^k} = a^{p^k} + b^{p^k}$ for each a, b in the quotient field of R. By induction on k, only the case $k = 1$ is needed. But this case is a well-known consequence of the fact that the combinatorial symbol $\binom{p}{i}$ is a multiple of p for $1 \leqslant i \leqslant p-1$ (cf. [**1**, Theorem 1, p. 177]).

Before giving our first "rarity" result (Corollary 2.4), we shall need another well-known fact about combinatorial symbols. Its proof, included next for the sake of completeness, was shown to the author by Wilbur Jonsson in 1964 as a tool in proving Sylow's Theorems.

LEMMA 2.2. *If p is prime and m, k are positive integers, then*

$$\binom{p^m k}{p^m} \equiv k (\mathrm{mod}\, p).$$

Proof. Set $n = p^m k$. Expanding $\alpha = (X+1)^n$ via the binomial theorem in the polynomial ring $\mathbb{F}_p[X]$, we find that the numerical coefficient of the term in X^{p^m} is $\binom{n}{p^m}$, viewed in $\mathbb{F}_p$. But Proposition 2.1(d) leads to $\alpha = (X^{p^m}+1)^k$. Thus, by another application of the binomial theorem, the canonical image of $\binom{k}{1} = k$ in $\mathbb{F}_p$ is the coefficient of the term in X^{p^m}. Hence, $\binom{n}{p^m} = k$, viewed in $\mathbb{F}_p$.

THEOREM 2.3. *Let R be a domain, X an indeterminate over R, and $n \geqslant 2$ a positive integer. Then the following conditions are equivalent:*
(1) *$R[X]$ is an n-domain;*
(2) $(X+1)^n = X^n + 1$ in $R[X]$;
(3) $p = \mathrm{ch}(R) > 0$ *and* $n = p^k$ *for some positive integer* k.

Proof. (3) $\Rightarrow$ (1) by Proposition 2.1(d), and (1) $\Rightarrow$ (2) trivially. We next show that (2) $\Rightarrow$ (3). Assume (2). By comparing coefficients of X^{n-1} in $(X+1)^n$ and X^n+1, $n=0$ in R. Hence $p = \mathrm{ch}(R) > 0$ and $p|n$. If (3) fails, the fundamental theorem of arithmetic gives positive integers m and $k \geqslant 2$ so that $n = p^m k$ and $p \nmid k$. By the binomial theorem, the coefficient of X^{p^m} in $(X+1)^n$ is $\binom{n}{p^m}$. Thus, by (2) and Lemma 2.2, $k = 0$ in R; and so $p|k$, the desired contradiction.

In case n is prime and $k=1$, the equivalence (2) $\Leftrightarrow$ (3) in Theorem 2.3 was noted in [**2**, Exercise 8, p. 30].

It will be convenient to introduce the following terminology. A domain R is called an *n*-**exception** in case R is an n-domain, R is not isomorphic to $\mathbb{F}_2$, and n is not an integral power of $\mathrm{ch}(R)$. This section's title is now justified by the statement of

COROLLARY 2.4. *Let $n \geqslant 2$ be a positive integer. Then each n-exception is a finite field of cardinality less than n. Thus, up to isomorphism, there are (for given n) only finitely many n-exceptions.*

Proof. Let R be an n-exception for which the first assertion fails. Since each finite domain is a field, $|R| \geqslant n$. As R is an n-domain, each element of R is a root of the polynomial $g = (X+1)^n - (X^n+1)$. Since $\deg(g) < n$, it follows that $g = 0 \in R[X]$ (cf. [**1**, Corollary 1, p. 131], [**2**, p. 40]). By Theorem 2.3 [(2) $\Rightarrow$ (3)], we have $p = \mathrm{ch}(R) > 0$ and $n = p^k$ for some positive integer k. This contradicts the n-exceptionality of R. Thus the first assertion holds. The final assertion also holds, since any two finite fields of the same cardinality are isomorphic (cf. [**1**, Theorem, p. 280], [**2**, pp. 92–93]).

The table exhibited in Section 1 (whose entries will be justified in Sections 3 and 4) reveals that Corollary 2.4 is best-possible. For instance, $\mathbb{F}_4$ is a 5-exception (and $\mathbb{F}_8$ is a 9-exception).

The proof of Corollary 2.4 reveals that for each positive integer $n \geqslant 2$, a domain R is an n-domain if (and only if) $(a+1)^n = a^n + 1$ for each $a \in R$. This generalizes Proposition 2.1(b).

Another immediate, and amusing, consequence of Corollary 2.4 is a new proof that each n-domain has positive characteristic.

The above table also motivates the next "rarity" result. Indeed, observe from the table that $\mathbb{F}_4$ is an n-field for consecutive values of n (namely 4, 5); so is $\mathbb{F}_8$ (with $n = 8, 9$); and $\mathrm{ch}(\mathbb{F}_4) = 2 = \mathrm{ch}(\mathbb{F}_8)$.

THEOREM 2.5. *Let $n \geqslant 2$ be a positive integer and R an n-domain. Then the following conditions are equivalent:*

(1) *R is an $(n+1)$-domain;*
(2) *R is a finite field, $ch(R)=2$, and $a^n = a$ for each $a \in R$.*
Moreover, if these conditions hold, the cardinality of R is at most n.

Proof. For $a \in R$,

$$(a+1)^{n+1} = (a+1)^n(a+1) = (a^n+1)(a+1) = a^{n+1} + a^n + a + 1.$$

$(1) \Rightarrow (2)$: Assume (1). Thus, for $a \in R$, $a^{n+1}+1 = a^{n+1} + a^n + a + 1$. Hence $a^n + a = 0$ for each $a \in R$. Choosing $a = 1$ reveals that $\mathrm{ch}(R) = 2$. Hence $a^n = a$ for each $a \in R$. Finally, R is finite (and hence a field) of cardinality at most n, since $X^n - X$ has at most n roots in the quotient field of R.

$(2) \Rightarrow (1)$: Assume (2). By Proposition 2.1(b), it is enough to show that $(a+1)^{n+1} = a^{n+1} + 1$ for each $a \in R$. By the equation displayed above, this amounts to showing $a^n + a = 0$. But this follows directly from (2).

COROLLARY 2.6. *Up to isomorphism, $\mathbb{F}_2$ is the only domain which is an n-domain for three consecutive values of n.*

Proof. As noted in Proposition 2.1(c), $\mathbb{F}_2$ is an n-domain for each $n \geqslant 2$. Conversely, if R is an i-domain for all $i \in \{n, n+1, n+2\}$, then Theorem 2.5 gives $a^n = a = a^{n+1}$ for each $a \in R$. Thus, if $0 \neq a \in R$, cancellation of a^n yields $1 = a$. Then $R = \{0,1\} \cong \mathbb{F}_2$.

3. Exceptions are Common

By Proposition 2.1(d), each finite field $\mathbb{F}_m$, $m \geqslant 3$, is an n-field for infinitely many $n \geqslant 2$. We shall show next that $\mathbb{F}_m$ is actually an n-exception for infinitely many of these n. In view of Corollary 2.6, this behavior is the "most" that one could hope for.

THEOREM 3.1. *Let F be a finite field, isomorphic to $\mathbb{F}_m$, with $m \geqslant 3$. Then F is an n-exception for infinitely many n, in particular for n any integral power of $2m-1$.*

Proof. It is easy to prove by induction on k that if R is an n-exception, then R is an n^k-exception for each positive integer k. Hence it remains only to show that F is a $(2m-1)$-exception. Now, $p = \mathrm{ch}(F) > 0$ and $m = p^s$ for some positive integer s (cf. [**2**, p. 92]). Thus $2m-1$ is not of the form p^t, lest $1 = 2p^s - p^t$ be divisible by p, a contradiction. Hence it just remains to show that F is a $(2m-1)$-field. To this end, it clearly suffices to show $\alpha^{2m-1} = \alpha$ for each $\alpha \in F$. However, $\alpha^m = \alpha$ by the structure of finite fields (cf. [**2**, p. 92]), whence $\alpha^{2m-1} = \alpha^m\alpha^m\alpha^{-1} = \alpha\alpha\alpha^{-1} = \alpha$.

It follows from Theorem 3.1 that if p is an odd prime, k and t are positive integers and $n = (2p^t - 1)^k$, then there exists an n-exception, namely $\mathbb{F}_{p^t}$. We do not know whether, for *each* odd $n \geqslant 3$, there exists an n-exception.

The next result will help to verify some of the entries in the table given in Section 1.

PROPOSITION 3.2. *If R is both a k-domain and a t-domain, then R is a kt-domain.*

Proof. $(a+1)^{kt} = ((a+1)^k)^t = (a^k+1)^t = (a^k)^t + 1 = a^{kt} + 1$ for each $a \in R$.

We shall soon verify the entries in the table given in Section 1. By marshalling the above work (especially Proposition 2.1(c), (d); Theorem 2.5; and Theorem 3.1), one verifies that the exhibited $\mathbb{F}_m$'s *are* n-fields, except possibly for the pair $(n,m) = (7,3)$. (Note that Proposition 3.2 explains the presence of the pair $(n,m) = (10,4)$, given the "earlier" pairs $(2,4)$ and $(5,4)$. It also permits identification of pairs beyond the capacity of the table, such as (6125,4).) As for the pair (7,3), it suffices to show that $\alpha^7 = \alpha$ for each $\alpha \in \mathbb{F}_3$. Since $\alpha^3 = \alpha$, a calculation in the spirit of the proof of Theorem 3.1 is available: $\alpha^7 = \alpha^3\alpha^3\alpha = \alpha\alpha\alpha = \alpha^3 = \alpha$. Not only does this analysis verify that each entry in Section 1's table deserves to be listed, it generalizes easily to yield

THEOREM 3.1 (bis). *If $F \cong \mathbb{F}_m$ with $m \geqslant 3$ and $n = m^2 - m + 1$, then F is an n^k-exception for each positive integer k.*

4. Table-building

We have just seen how the results obtained in Sections 2 and 3 help one to verify each entry in Section 1's table. *Why are there no other entries*? A systematic explanation depends on results delimiting n-exceptions. We shall next give two such table-building theorems, and then show that the table (for $2 \leqslant n \leqslant 10$) is complete.

THEOREM 4.1. *Let R be an n-domain of characteristic p. Assume that $n = p^m k$ for positive integers m and k such that $k \geqslant 2$ and $p \nmid k$. Then R is a finite field of cardinality less than k.*

Proof. For each $a \in R$,

$$(a^{p^m} + 1)^k = \left((a+1)^{p^m}\right)^k = (a+1)^n = a^n + 1 = \left(a^{p^m}\right)^k + 1.$$

Thus each element of $S = \{a^{p^m} : a \in R\}$ is a root of the polynomial $g = (X+1)^k - (X^k + 1)$. Note that $g \neq 0$ by virtue of Theorem 2.3. Thus $1 \leqslant \deg(g) < k$. It follows that the cardinality of S is at most $k - 1$. However S has the same cardinality as R, since Proposition 2.1(d) easily implies that the function $R \to S$, $a \mapsto a^{p^m}$, is one-to-one. It remains only to note that R, being a finite domain, is a field.

THEOREM 4.2. *Let F be an n-exception of characteristic $p \geqslant 3$. Let k be the order of 2 in the multiplicative group $F^* = F \setminus \{0\}$. Then $p \equiv n \equiv 1 \pmod{k}$.*

Proof. By Corollary 2.4, F is finite. Then k is finite, dividing the cardinality of F^* (cf. [1, Proposition 1, p. 96]).

By hypothesis, $(a+1)^n = a^n + 1$ for each $a \in F$. Choosing $a = 1$ leads to $2^n = 2$; that is, $2^{n-1} = 1$ in F^*. By elementary group theory (cf. [1, Proposition 2, p. 96]), $k | (n-1)$; that is, $n \equiv 1 \pmod{k}$.

It will suffice to prove that $k | (p - n)$; that is, by elementary group theory, that $2^p = 2^n$ in F^*. To this end, express n as $qp + r$ by the division algorithm, for suitable nonnegative integers q, r. Then for each $a \in F$,

$$(a^q + 1)^p a^r = (a^{qp} + 1)a^r = a^{qp+r} + a^r = a^n + a^r.$$

Choosing $a = 1$ leads to

$$2^p = 1^n + 1^r = 1^n + 1 = (1+1)^n = 2^n.$$

The careful reader will note that Theorem 4.2 carries over to any "nonexceptional" finite field F of characteristic $p \geqslant 3$, but this extra fact has no role in table-building.

In order to "build" the table given in Section 1, the interested reader may proceed *ad hoc* (as the author initially did). However, the comments in Section 3 showed, by virtue of the then-accumulated theory, that for each n, $2 \leqslant n \leqslant 10$, each listed domain *is* an n-domain. It remains now to eliminate any other putative tabular entries (for $2 \leqslant n \leqslant 10$). We shall do this by exploiting the above theory, with a minimum of calculation. Many of the verifications result from combining Corollary 2.4 and Theorem 4.2. Consider, for instance, the case $n = 7$: By Corollary 2.4, we need only dispatch $\mathbb{F}_5$. For this, apply Theorem 4.2. The point is to show that the order, k, of 2 in $(\mathbb{F}_5)^*$ does *not* satisfy $5 \equiv 7 \equiv 1 \pmod k$. As $(\mathbb{F}_5)^*$ has cardinality 4, elementary group theory shows $k|4$. But $k \neq 1, 2$ since $1 \neq 2, 4$ in $\mathbb{F}_5$. Thus $k = 4$ and the congruences at issue do *not* hold. Corollary 2.4 and Theorem 4.2 combine similarly, to handle the cases $n = 3, 4, 5, 8$.

For $n = 2$, a direct approach is best. If $(a+1)^2 = a^2 + 1$ for each $a \in R$, the binomial theorem leads to $2a = 0$, whence $\mathrm{ch}(R) = 2$.

Theorem 2.5 is needed in verifying that $\mathbb{F}_4$ is not a 9-domain: just note that $a^8 \neq a$ for each $a \in \mathbb{F}_4 \setminus \{0, 1\}$. Theorem 4.1 guarantees that neither $\mathbb{F}_3$ nor $\mathbb{F}_4$ is a 6-domain. The interested reader will easily complete the other verifications. Perhaps the most interesting remaining case is $n = 10$: By Corollary 2.4, it remains to eliminate $\mathbb{F}_3$, $\mathbb{F}_5, \mathbb{F}_7, \mathbb{F}_8$, and $\mathbb{F}_9$. For all these except $\mathbb{F}_7$ and $\mathbb{F}_8$, Theorem 2.5 and Theorem 4.2 *each* suffice. (Moreover, Theorem 4.1 eliminates $\mathbb{F}_5$ and $\mathbb{F}_8$; and Corollary 2.6 eliminates $\mathbb{F}_8$.) What eliminates $\mathbb{F}_7$? A timely calculation, in the spirit of Section 3! Since $\alpha^7 = \alpha$ for each $\alpha \in \mathbb{F}_7$, $\alpha^{10} = \alpha^4$, and so it suffices to find $a \in \mathbb{F}_7$ such that $(a+1)^4 \neq a^4 + 1$. In fact, each $a \in \mathbb{F}_7 \setminus \{0, 1\}$ works. For instance, you may choose $a = 2$ because $81 \not\equiv 17 \pmod 7$.

The table has been built!

REMARK 4.3. The interested reader should have little trouble in using this article's theory to catalogue the n-domains for $n = 11, 12, \ldots$. By way of contrast to the above appeal to Corollary 2.4 and Theorem 4.2, we close by suggesting that the reader discover how an *ad hoc* search for the 7-exceptions would have settled on $\mathbb{F}_3$ and $\mathbb{F}_4$.

REFERENCES

1. L. Childs, *A Concrete Introduction to Higher Algebra*, Springer-Verlag, New York/Heidelberg/Berlin, 1979.
2. D. E. Dobbs and R. Hanks, *A Modern Course on the Theory of Equations*, Polygonal Publ. House, Passaic, NJ, 1980.

I would urge that an essential element in the equipment of an investigator is a literary education, or, if you prefer it, a training in the means of expression. It is necessary to be articulate, but more than this is desirable. It is desirable to be mathematically articulate, to be able to express mathematical ideas in such a way that they can be comprehended easily by those who have the requisite training.... There is such a thing as style in mathematics, and it is worth cultivating.

—A. E. H. Love, Presidential Address to the London Mathematical Society, *Proc. London Math. Soc.* (2) 14 (1915), 178–188.

PROBLEMS

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

BRUCE HANSON, *associate editor*
St. Olaf College

Proposals

To be considered for publication, solutions should be received by July 1, 1989.

1312. *Proposed by Daniel Ullman, The George Washington University, Washington, DC.*

For $n \geq 1$, let $S_n = \sum \frac{1}{a_1 a_2 \cdots a_m}$, where the sum is taken over all m and all finite sequences of positive integers $a_1, a_2, \ldots, a_m$ such that $a_1 = n$ and $a_{i+1} \leq a_i - 2$ for $1 \leq i \leq m-1$. For example,

$$S_6 = \frac{1}{6} + \frac{1}{6\cdot 4} + \frac{1}{6\cdot 3} + \frac{1}{6\cdot 2} + \frac{1}{6\cdot 1} + \frac{1}{6\cdot 4\cdot 2} + \frac{1}{6\cdot 4\cdot 1} + \frac{1}{6\cdot 3\cdot 1}.$$

Show that S_n is a convergent sequence and find its limit.

1313. *Proposed by Václav Konečný, Ferris State University, Big Rapids, Michigan.*

Let $f(x) = a^x$. For what values of a, $0 < a < 1$, if any, are there distinct points P and Q on the graph of $y = f(x)$ that are symmetrical about the line $y = x$.

1314. *Jany C. Binz, University of Bern, Switzerland.*

Let $t_n = \frac{n(n+1)}{2}$ be the nth triangular number. Find all positive integers m and n such that

$$t_n + t_{n+1} + t_{n+2} = t_m.$$

1315. *Proposed by M. S. Klamkin and A. Liu, University of Alberta, Canada.*

Determine all real values of λ such that the roots of

$$P(x) \equiv x^n + \lambda \sum_{r=1}^{n} (-1)^r x^{n-r} = 0, \qquad (n > 2)$$

are *all* real.

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

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

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

1316. *Proposed by K. R. S. Sastry, Addis Ababa, Ethiopia.*

Characterize the Heronian triangles (a, b, c) in which the Eulerian segment OH between the circumcenter O and the orthocenter H subtends a right angle at the vertex A. (A *Heronian* triangle is one with integral sides and integral area.)

Quickies

Answers to the Quickies are on page 66.

Q742. *Proposed by Murray S. Klamkin, University of Alberta, Canada.*

Let

$$S_n = \frac{\sin^{2n+2}\theta}{\sin^{2n}\alpha} + \frac{\cos^{2n+2}\theta}{\cos^{2n}\alpha}.$$

If $S_k = 1$ for some positive integer k, show that $S_n = 1$ for $n = 1, 2, \dots$.

Q743. *Proposed by David M. Bloom, Brooklyn College of CUNY, New York.*

Evaluate the sum of the series

$$\frac{1}{1\cdot 2\cdot 2!} + \frac{1}{2\cdot 3\cdot 3!} + \frac{1}{3\cdot 4\cdot 4!} + \cdots = \sum_{n=2}^{\infty} \frac{1}{(n-1)n(n!)}.$$

Q744. *Proposed by David Callan, University of Bridgeport, Connecticut.*

Let S be an arithmetic progression of (at least two) positive integers whose common difference is odd. Show that $\sum_{i\in S} 1/i$ is not an integer.

Solutions

Exscribed Quadilateral **February 1988**

1287. *Proposed by C. S. Gardner, Austin, Texas.*

Let P be an interior point of the rectangle $ABCD$. Draw lines through A, B, C, D perpendicular to PA, PB, PC, PD respectively. Show that the area of the convex quadrilateral enclosed by these four lines is equal to or greater than twice the area of the rectangle. When do we have equality?

Solution by Con Amore Problem Group, Royal Danish School of Educational Studies, Copenhagen, Denmark.

We place rectangle $ABCD$ in a coordinate system as shown in FIGURE 1. We use the point names of the figure, and assign coordinates (a, b) to A and (u, v) to P. We now justify the coordinates shown for Q, R, S.

The circle with diameter PH passes through A and D, so its center is on the perpendicular bisector of AD, which is the x-axis, and so H has ordinate $-v$. Similarly, F has ordinate $-v$, and each of F and G has abscissa $-u$. It follows that the segments FH and EG are parallel to the axes, and their point of intersection is $(-u, -v)$.

We now turn to FIGURE 2 which shows a rectangle inscribed in a right-angled triangle. Using the names of the figure, we have $y/q = p/x$, which implies

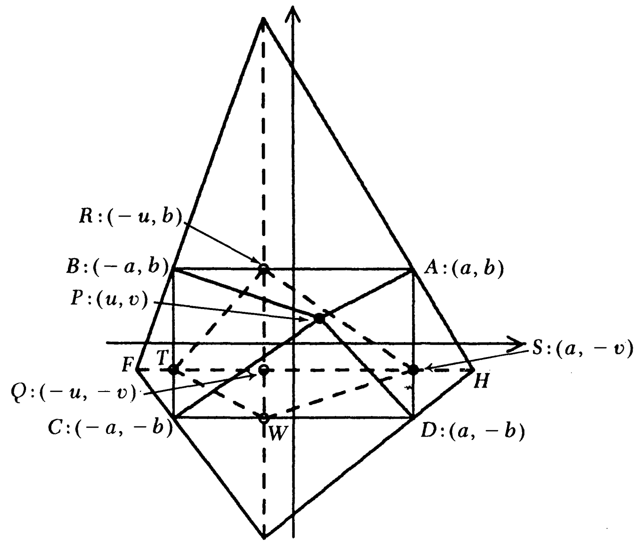


FIGURE 1

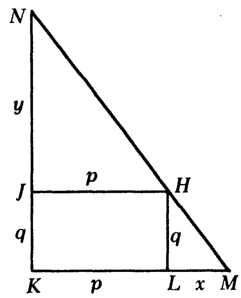


FIGURE 2

$$\frac{1}{2}(p+x)(q+y) = \frac{1}{2}(p+x)\left(q+\frac{pq}{x}\right) = \frac{q}{2x}(p+x)^2$$
$$= \frac{q}{2x}\left((p-x)^2+4px\right)$$
$$\geqslant 2pq,$$

with equality if and only if $x = p$ and $y = q$.

Returning to FIGURE 1, this means that Area $EQH \geqslant 2 \cdot$ Area $ARQS$, with equality if and only if

$$EH \| RS, \tag{1}$$

and similarly with areas EQF and $BTQR$, FQG and $CWQT$, GQH and $DSQW$. It follows that

$$\text{Area } EFGH \geqslant 2 \cdot \text{Area } ABCD, \tag{2}$$

with equality if and only if we have (1) *and*

$$EF \| RT, \quad FG \| TW, \quad \text{and} \quad GH \| WS. \tag{3}$$

But (3) follows from (1). For (1) $\Rightarrow ER = RQ \Rightarrow EF \| RT$, and similarly $EF \| RT \Rightarrow FG \| TW \Rightarrow GH \| WS$.

So there is equality in (2) if and only if (1) is satisfied. Now (1) $\Leftrightarrow \overrightarrow{PA} \perp \overrightarrow{RS} \Leftrightarrow (a-u)(a+u)+(b-v)(-v-b)=0 \Leftrightarrow u^2-v^2=a^2-b^2$. This means that there is equality in (2) if and only if P is on the curve

$$x^2-y^2=a^2-b^2, \qquad |x| \leqslant a, \qquad |y| \leqslant b, \tag{4}$$

which is the union of diagonals AC and BD when $a=b$ (i.e. $ABCD$ is a square), and otherwise the part inside $ABCD$, of an equilateral hyperbola with asymptotes $y = \pm x$, and passing through A, B, C, and D.

When $a \geqslant b$ (as in FIGURE 1) the curve (4) meets the x-axis in $z_{\pm} = (\pm\sqrt{a^2-b^2}, 0)$ for which a construction is shown in FIGURE 3. The figure also shows a construction of the tangents to (4) at A, B, C, D. This construction is based on the following argument. The tangent at A has equation

$$ax - by = a^2 - b^2,$$

hence slope a/b, and so it is orthogonal to BD (with slope $-b/a$); similarly with the tangents at B, C, D.

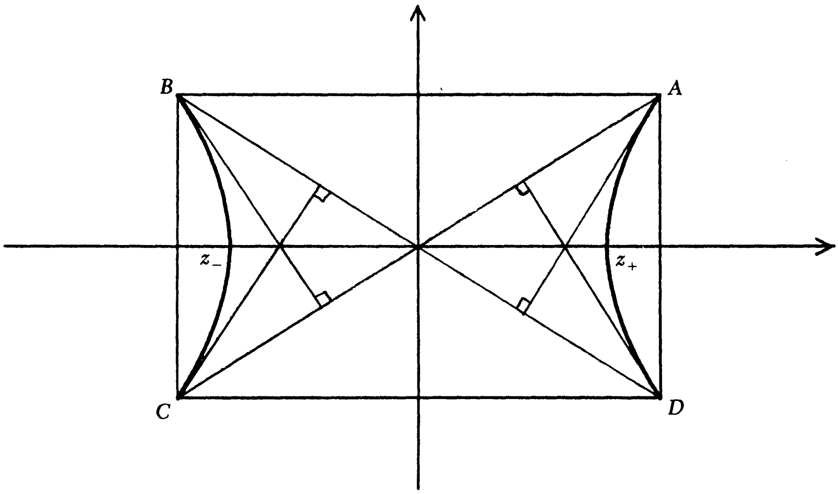


FIGURE 3

Also solved by Nicolas K. Artemiadis (Greece), Michael Bertrand, J. C. Binz (Switzerland), Nirdosh Bhatnagar, Onn Chan (student), Chico Problem Group, Gill Y. Choi, C. Patrick Collier, Hüseyin Demir (Turkey), Terence J. DeSousa, David Earnshaw (Canada), Martin Feuerman, Nick Franceschine, J. M. Stark, John F. Goehl, Jr., Francis M. Henderson, Chuck Hixon, Thomas Jager, Hans Kappus (Switzerland), Farhood Pouryoussefi Kermany (student, Iran), Václav Konečný, L. Kuipers (Switzerland), N. J. Lord (England), Kim McInturff, Herby McKaig (student, Canada), Aron Pinker, Harry D. Ruderman, Volkard Schindler (East Germany), David Schumann, Seshadri Sivakumar (Canada), Adam Stinchcomke, Michael Vowe (Switzerland), John T. Ward, Harry Weingarten, Robert L. Young, and the proposer.

Quickie Functional Equation **February 1988**

1288. *Proposed by Leroy F. Meyers, The Ohio State University, Columbus.*

Find all differentiable real-valued functions f defined on the entire real line such that $f(x)f'(x) = 0$ for all real x.

Solution by Donald F. Bailey, Trinity University, San Antonio, Texas.

If f is differentiable and real-valued on the real line and $f(x)f'(x)=0$ for all real x then clearly

$$\frac{d}{dx}f(x)^2=0,$$

which implies $f(x)^2=K$ for some constant K. Then the continuity of f implies that f is a constant function.

Also solved by Nicolas K. Artemiadis (Greece), Bruce W. Atkinson, Seung-Jin Bang (Korea), S. F. Barger, Jim Bolen, William W. Bosch, Nicholas Buck (Canada), Onn Chan (student), Chico Problem Group, Con Amore Problem Group (Denmark), Camilla A. Daggett, Charles R. Diminnie, Fred Dodd, Michael W. Ecker, Russel Euler, Joe Flowers, Bill Frederick, Kay Griswold and Edward C. Nichols, Lee O. Hagglund, Joe Howard, Thomas Jager, Farhood Pouryoussefi Kermany (student, Iran), Robert Kowalski, Don Krug, Kathleen E. Lewis, N. J. Lord (England), Patrick Dale McCray, Kim McInturff, David E. Manes, Andreas Müller (Switzerland), Robert L. Raymond, Mohammad Riazi, Adam Riese, Randy K. Ross, Harry D. Ruderman, H.-J. Seiffert (West Germany), Zun Shan and Edward T. H. Wang (Canada), Daniel B. Shapiro, Michiel Smid (The Netherlands), Thomas J. Smith, Stephen Spielberg, Jan Söderkvist (Sweden), Gerald Thompson, Michael Vowe (Switzerland), William P. Wardlaw, William Watkins, William V. Webb, John C. Wenger, C. Wildhagen (The Netherlands), Veronique Ziegler (student), and the proposer.

Jager also proved that linear functions are the only twice-differentiable functions that satisfy $f(x)f''(x)=0$ for all real x.

Sliding Beads **February 1988**

1289. *Proposed by M. S. Klamkin, University of Alberta, Canada.*

Two identical beads slide on two straight wires intersecting at right angles. If the beads start from rest in any position other than the intersection point of the wires and attract each other in an arbitrary mutual fashion but also subject to a drag proportional to the speed, show that the beads will arrive at the intersection simultaneously.

Solution by Yan-loi Wong (student), University of California, Berkeley.

Resolving the attractive force $f(x,y)$ along the axes (the wires), the equations of motion for the two beads are given by

$$x''=\frac{-xf(x,y)}{\sqrt{x^2+y^2}}+kx',$$

$$y''=\frac{-yf(x,y)}{\sqrt{x^2+y^2}}+ky',$$

where k is a positive constant. Eliminating $f(x,y)$ we find that

$$yx''-xy''=k(yx'-xy'),$$
$$(yx'-xy')'=k(yx'-xy').$$

It follows that

$$yx'-xy'=Ce^{kt}$$

for some constant C. Since $x'(0)=y'(0)=0$, we have $yx'-xy'=0$. Hence, until

$$y(t)=0,\qquad \frac{d}{dt}\left(\frac{x}{y}\right)=0,$$

so that $x(t)=Ay(t)$ for some nonzero constant A, and this gives the desired result.

Also solved by The Chico Problem Group, The Con Amore Problem Group (Denmark), John F. Goehl, Jr., Thomas Jager, Farhood Pouryoussefi Kermany (student, Iran), Václav Konečný, L. Kuipers (Switzerland) and M. Kuipers (The Netherlands), N. J. Lord (England), Kim McInturff, Andreas Müller (Switzerland), M. Riazi-Kermani, Volkhard Schindler, and the proposer.

The proposer notes that this problem without drag occurs in C. J. Coe, *Theoretical Mechanics*.

A Combinatorial Sum February 1988

1290. *Proposed by Edward T. H. Wang, Wilfrid Laurier University, Canada.*

For positive integers n and r, let $\left\langle {n \atop r} \right\rangle = \binom{n+r-1}{r}$. Find a closed form expression for

$$\sum_{r=1}^{k} r\left\langle {n \atop r} \right\rangle,$$

where k denotes a positive integer.

I. *Solution by Barry Brunson, Western Kentucky University, Bowling Green.*

Note that $r\left\langle {n \atop r} \right\rangle = n\left\langle {r \atop n} \right\rangle$, and therefore

$$\sum_{r=1}^{k} r\left\langle {n \atop r} \right\rangle = \sum_{r=1}^{k} n\binom{n+r-1}{n} = n\sum_{r=1}^{k}\left(\binom{n+r}{n+1} - \binom{n+r-1}{n+1}\right) = n\binom{n+k}{n+1},$$

since the last sum is telescoping.

II. *Solution by David Graves, Elmira College, New York.*

A closed form expression for the summation is

$$n\binom{n+k}{k-1}.$$

This is easily proven by induction, with induction step as follows:

$$\begin{aligned}\sum_{r=1}^{k+1} r\left\langle {n \atop r} \right\rangle &= n\binom{n+k}{k-1} + (k+1)\left\langle {n \atop k-1} \right\rangle\\ &= n\binom{n+k}{k-1} + (k+1)\binom{n+k}{k+1}\\ &= n\binom{n+k}{k-1} + n\binom{n+k}{k} = n\binom{n+k+1}{k}.\end{aligned}$$

III. *Solution by N. J. Lord, Tonbridge School, Kent, England.*

Start by observing that

$$(1-x)^{-n} = \sum_{r=0}^{\infty}\left\langle {n \atop r} \right\rangle x^r. \qquad (*)$$

Differentiation yields

$$n(1-x)^{-n-1} = \sum_{r=1}^{\infty} r\left\langle {n \atop r} \right\rangle x^{r-1}.$$

$\sum_{r=1}^{k} r\left\langle {n \atop r} \right\rangle$ will then be the coefficient of x^{k-1} in the expansion of

$$(1-x)^{-1}n(1-x)^{-n-1}=n(1-x)^{-n-2}$$

(consider $(1+x+x^2+\cdots)\sum_{r=1}^{\infty}r\left\langle {n \atop r} \right\rangle$). By $(*)$, this is

$$n\left\langle {n+2 \atop k-1} \right\rangle = n\binom{n+k}{n+1}.$$

IV. *Solution by Y. H. Kwong, SUNY College of Fredonia, New York.*

Let $[m]$ denote the set of integers $\{1,2,\ldots,m\}$. There are $\left\langle {n \atop r} \right\rangle$ ways to select r elements from $[n]$, with repetition allowed. Given any such selection S, arrange its elements in ascending order: $a_1 \leqslant a_2 \leqslant \cdots \leqslant a_r$. For each i where $1 \leqslant i \leqslant r$, construct a new selection S_i by removing a_i and replacing all a_j with $n+1$ where $j>i$ and $a_j = a_i$. So S_i is a selection of $r-1$ elements from $[n+1]$, with repetition allowed. Observe that this construction gives an n-to-1 correspondence. To standardize the sizes of these selections, add $k-r$ copies of $n+2$ to each S_i, so that we have selections of $k-1$ elements, with repetition allowed, chosen from $[n+2]$. Repeat the same construction on all other selections of r elements, $1 \leqslant r \leqslant k$. Then we can conclude that

$$\sum_{r=1}^{k} r\left\langle {n \atop r} \right\rangle = n\left\langle {n+2 \atop k-1} \right\rangle = n\binom{n+k}{k-1}.$$

V. *Solution by James C. Hickman, University of Wisconsin, Madison.*

$$\begin{aligned}\sum_{r=1}^{k} r\left\langle {n \atop r} \right\rangle &= \frac{1}{(n-1)!}\sum_{r=1}^{k}(n+r-1)^{(n)}\\ &= \frac{1}{(n+1)(n-1)!}\sum_{r=1}^{k}\Delta(n+r-1)^{(n+1)}\\ &= \frac{(n+r-1)^{(n+1)}}{(n+1)(n-1)!}\bigg|_{1}^{k+1}\\ &= \frac{(n+k)^{(n+1)}}{(n+1)(n-1)!}.\end{aligned}$$

Here, $n^{(r)} = n(n-1)\cdots(n-r+1)$ and Δ is the difference operator.

VI. *Solution by William F. Trench, Trinity University, San Antonio, Texas.*

By the residue theorem,

$$\left\langle {n \atop r} \right\rangle = \frac{1}{2\pi i}\oint_C \frac{(1+z)^{n+r-1}}{z^{r+1}}\,dz,$$

where C is any circle centered at the origin. Let S_k be the sum in question; then

$$S_k = \frac{1}{2\pi i}\oint_C \frac{(1+z)^n}{z^2}\left(\sum_{r=1}^{k} r\left(\frac{1+z}{z}\right)^{r-1}\right)dz.$$

Differentiating the identity

$$\sum_{r=0}^{k} w^r = (1-w^{k+1})/(1-w)$$

shows that

$$\sum_{r=1}^{k} rw^{r-1} = \frac{(1-w^{k+1})-(k+1)(1-w)w^k}{(1-w)^2}$$

if $w \neq 1$. Now let $w = (1+z)/z$ here to see that (1) can be rewritten as

$$S_k = \frac{1}{2\pi i}\oint_C \left((1+z)^n + \frac{k(1+z)^{n+k+1}}{z^{k+1}} - \frac{(k+1)(1+z)^{n+k}}{z^k} \right) dz.$$

This and the residue theorem imply that

$$S_k = k\binom{n+k+1}{k} - (k+1)\binom{n+k}{k-1} = n\binom{n+k}{k-1}.$$

Also solved by A. H. Alfar (student), Seung-Jin Bang (Korea), Richard Bedient, Nirdosh Bhatnagar, J. C. Binz (Switzerland), Duane M. Broline, Stephen R. Cavior, Onn Chan (student), Chico Problem Group, Mark Collins and Ashley Plank (Australia), Con Amore Problem Group (Denmark), Jim Delany, Frank Dezzi, Fred Dodd, David Doster, Ragnar Dybvik (Norway), Michael J. Fitzgerald, Joe Flowers, Tom Fowler (student), Nick Franceschine, Ralph Garfield, Samuel Gebre-Egziabher (student), Richard A. Gibbs, Zeather Gladney, Leslie V. Glickman (England), Ralph P. Grimaldi, Thomas Jager, Hans Kappus (Switzerland), L. Kuipers (Switzerland), Y. H. Harris Kwong (second solution), Kee-Wai Lau (Hong Kong), Carl Libis, Wilen Lim (student), Pamela A. Lipka, Graham Lord, Ray McClanahan, Kim McInturff, David E. Manes, Armel Mercier (Canada), William Moser (Canada), Roger B. Nelsen, Gillian Nonay, Northern Kentucky University Problem Group, Kennard Reed, Jr., Mohammad Riazi, Harry D. Ruderman, Volkhard Schmidler (East Germany), Edward Schmeichel, Jens Schwaiger (Austria), H. -J. Seiffert (West Germany), Paul Shimura and Veronique Ziegler (students), Michiel Smid (The Netherlands), Robert S. Stacy, Jan Söderkvist (Sweden), T. A. Tarbox, R. S. Tiberio, Michael Vowe (Switzerland), John T. Ward, William P. Wardlaw, William V. Webb, Gregory P. Wene, Harry Weingarten, C. Wildhagen (The Netherlands), Yan-loi Wong (student), Matt Wyneken, and the proposer.

Mercier proved more generally that for any positive integer m,

$$T_k^{(m)}(n) := \sum_{r=1}^{k} r^m\left\langle {n \atop r} \right\rangle = \sum_{j=1}^{m} (n)_j S(m,j)\left\langle {n+1+j \atop k-j} \right\rangle,$$

where $(n)_j = n(n+1)\cdots(n+j-1)$, $S(m,j)$ are Stirling numbers of the second kind and k denotes a positive integer.

A Determinant Inequality — February 1988

1291. *Proposed by Mihály Bencze, Braşov, Romania.*

Let n be a positive integer, and let $\mathbf{A}$ be an $m \times m$ matrix of real numbers such that $\mathbf{A}^{2n+1} = \mathbf{A} - \mathbf{I}$, where $\mathbf{I}$ is the identity matrix. Prove that $(-1)^m(\det \mathbf{A}) > 0$.

Solution by William Watkins, California State University, Northridge.

Each eigenvalue of $\mathbf{A}$ satisfies the equation $x^{2n+1} = x - 1$, and it follows that $\mathbf{A}$ has no nonnegative eigenvalues. Since $\mathbf{A}$ is real, the nonreal eigenvalues occur in pairs, as complex conjugates. Thus there are an even number of nonreal eigenvalues, and their product is positive. The remaining eigenvalues are negative. Since $\det A$ equals the product of all the eigenvalues, the sign of $\det A$ equals the number of negative eigenvalues. But this number has the same parity as m, and the result follows.

Also solved by David Callan, Seung-Jin Bang (Korea), Onn Chan (student), Con Amore Problem Group, Jim Delany, Bill Frederick, Weimin Han, Thomas Jager, Hans Kappus (Switzerland), N. J. Lord (England), Kim McInturff, Colm Mulcahy, Andreas Müller (Switzerland), Zun Shan and Edward T. H. Wang (Canada), Daniel B. Shapiro, Jan Söderkvist (Sweden), William P. Wardlaw, C. Wildhagen (The Netherlands), Yan-loi Wong (student), and the proposer.

Answers

Solutions to the Quickies on p. 59.

A742. More generally, let

$$S_n = \sum_{i=1}^{n} \frac{x_i^{2n+2}}{y_i^{2n}},$$

where

$$x_1^2 + x_2^2 + \cdots + x_n^2 = 1 = y_1^2 + y_2^2 + \cdots + y_n^2.$$

Then if $S_k = 1$ for some positive integer k, $S_n = 1$ for $n = 1, 2, \ldots$.
By Hölder's Inequality

$$\left(\sum_{i=1}^{n} \frac{x_i^{2k+2}}{y_i^{2k}}\right)^{1/(k+1)} \left(\sum_{i=1}^{n} y_i^2\right)^{k/(k+1)} \geq \sum_{i=1}^{n} x_i^2.$$

Since we have the equality case by the hypotheses, we must have $x_i^{2k+2}/y_i^{2k} = \lambda y_i^2$, or, $x_i^2 = c y_i^2$ for all i. It then follows that $c = 1$ and $S_n = 1$ for $n = 1, 2, \ldots$. (Comment: The initial problem for the special case $k = 1$ appears in E. W. Hobson, *A Treatise on Plane and Advanced Trigonometry*, Dover, New York, 1957, p. 96.)

A743. It is easily verified that

$$\frac{1}{(n-1)n(n!)} + \frac{1}{n!} = \frac{1}{(n-1)(n-1)!} - \frac{1}{n(n!)}.$$

Summing this over all $n \geq 2$, the right side telescopes and the sum equals 1. Thus the sum is

$$1 - \sum_{n=2}^{\infty} \frac{1}{n!} = 1 - (e-2) = 3 - e.$$

A744. Strike out the odd terms in S (every second one) and divide each of the remaining terms by two to produce a shorter progression S′ with the same common difference. Repeat the process until ultimately a single number remains. Clearly this number comes from the *unique* term of S divisible by the maximal power of two. Consequently when $\Sigma_{i \in S} 1/i$ is brought to a least common denominator, two divides the denominator but not the numerator. Hence the sum is not an integer.

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.

Lederer, Eric (ed.), Calculus I Exam File, 1986, 250 pp (P). Lederer, Eric M. (ed.), *Linear Algebra Exam File*, 1989, 442 pp (P). Ward, Thomas L., *Probability and Statistics Exam File*, 1985, 346 pp (P). Engineering Press, Inc. (P.O. Box 1, San Jose, CA 95103-0001).

Copies of past examinations from the courses indicated (the probability and statistics is the calculus-based course), together with the professors' handwritten solutions. Problems have been sorted by topic; the professors included all seem to have good handwriting!

Gastwirth, Joseph L., *Statistical Reasoning in Law and Public Policy, Vol.* 1: *Statistical Concepts and Issues of Fairness*, Vol. 2: *Tort Law, Evidence and Health*, Academic Press, 1988; xxii + 925 pp + 49 pp of indices, $84.50.

"Designed as a textbook for students of law and public policy . . . this book introduces the reader to the concepts of statistics and applies them to a wide variety of data sets which arise in the context of legal and public policy decisionmaking." The author, a statistical consultant to the U.S. Office of Management and Budget, includes data from cases in which he served as an expert witness or in which he served as a consultant. The volumes are interesting and informative, and a far cry from the made-up problems that pass for "applications" in most texts.

Traub, Joseph F., et al. (eds.), *Annual Review of Computer Science*, Vol. 3, Annual Reviews Inc., 1988; 423 pp, $45.

Survey papers on topics in computer science, including database security, expert systems, protein folding, geometric reasoning, image analysis, theorem proving, algebraic complexity theory, applications in manufacturing, and computational geometry.

Chaitin, G. J., *Algorithmic Information Theory*, Cambridge U Pr, 1987; xi + 177 pp.

An involved and technical proof of "the strongest possible version of Goedel's incompleteness theorem, using an information-theoretic approach based on the size of computer programs." The conclusion: "[W]e see that proving whether particular exponential diophantine equations have finitely or infinitely many solutions, is absolutely intractable . . . Such questions escape the power of mathematical reasoning. This is a region in which mathematical truth has no discernible structure or pattern and appears to be completely random. These questions are completely beyond the power of human reasoning. Mathematics cannot deal with them."

MacLane, Saunders, and Garrett Birkhoff, *Algebra*, 3rd ed., Chelsea, 1988; xv + 626 pp.

Republication of a classic (1967), with restoration of the chapter on affine and projective geometry (which was omitted from the 2nd ed.). Every graduate student in mathematics should have this book.

Duren, Peter (ed.), *A Century of Mathematics in America*, Part I, AMS, 1988; viii + 477 pp, $57.

Thirty-three essays, a third of them reprinted and the rest original with this volume, ranging through reminiscences, individual biographies, tales of European mathematicians' migrations to America, and accounts of mathematicians in the World Wars and in the McCarthy era.

Richards, Joan L., *Mathematical Visions: The Pursuit of Geometry in Victorian England*, Academic, 1988; xi + 266 pp, $34.95.

Mathematical news did not travel very fast in the 19th century. Yet I am still surprised to consider that the discovery of non-Euclidean geometry did not "reach" England for 30 years (Morris Kline, in *Mathematical Thought from Ancient to Modern Times* (p. 879), says that the published work of Bolyai and Lobachevsky was simply "ignored," because "mathematicians generally exhibited their usual reluctance to entertain radical ideas.") Joan Hutchinson's book traces not only the effect of serious consideration of non-Euclidean geometry upon the English view of geometry, but also details the philosophical fallout and the pedagogical aftershocks. "This book originated as a study of the nineteenth-century reception of non-Euclidean geometry in England. . . . [It] has become a broader study of mathematics in culture, focusing on the pursuit of geometry in Victorian England."

Gardner, Martin, *Penrose Tiles to Trapdoor Ciphers . . . and the Return of Dr. Matrix*, Freeman, 1989; ix + 311 pp, $13.95(P).

This is the 13th collection from Gardner's Mathematical Games column in *Scientific American*. The columns on Penrose tilings, on public key cryptosystems, and on the French Oulipo group have been supplanted by brand new chapters on those topics, because so much has happened since the columns first appeared. Miracle of miracles, Dr. Matrix turns out not to have been killed by the KGB but has been looking into magic squares of Smith numbers.

Gardner, Martin, *Hexaflexagons and Other Mathematical Diversions: The First Scientific American Book of Mathematical Puzzles and Games*, University of Chicago Pr, 1988; ix + 200 pp, $10.95.

Second edition of the very first collection of Gardner's popular *Scientific American* columns. A new afterword and accompanying reference bring the reader up to date on the developments since 1959.

Washburn, Dorothy K., and Donald W. Crowe, *Symmetries of Culture: Theory and Practice of Plane Pattern Analysis*, University of Washington Pr, 1988; x + 299 pp, $40.

This joint work of a mathematician and an anthropologist shows how to use symmetry analysis as a powerful tool to analyze patterned design in material culture. No previous background in mathematics is presumed. Put this book on your coffee table; its hundreds of exquisitely-rendered photographs and drawings will gratify the esthetic senses of your friends and open the door for you to discuss with them the underlying mathematics.

Dossey, John A., et al., *The Mathematics Report Card: Are We Measuring Up? Trends and Achievement Based on the* 1986 *National Assessment*, Educational Testing Service, 1988; 143 pp (P).

Don't look for much good news here. Some ethnic and regional subpopulations show gains in student achievement, but such improvements are mainly confined to lower-order mathematical skills. A new feature of the presentation of the results in this study is the reporting of jackknifed standard errors.

Berggren, J. L., *Episodes in the Mathematics of Medieval Islam*, Springer-Verlag, 1986; xiv + 197 pp, $25.

"I have not attempted to write 'The History of Mathematics in Medieval Islam.' Such a book could not be written yet, for so much material remains unstudied . . . . My aim, rather, has been to exhibit some ways in which writers of the Islamic world contributed to the development of the mathematics one learns in high school." The six chapters—introduction, arithmetic, geometry, algebra, trigonometry, and spherical geometry/trigonometry—all include exercises, as the author intends the book as a textbook "accessible to anyone who has learned high school mathematics."

Pool, Robert, The Allais paradox, *Science* 242 (28 October 1988) 512.

Account of a paradoxical experiment, invented by the 1988 Nobel economics winner, which shows that a "natural" axiom for decision-making is false.

Cohen, Joel, The counterintuitive in conflict and cooperation, *American Scientist* 76: (November-December 1988) 577-583.

Simple network models can reveal surprisingly counterintuitive results: Increasing redundancy in a network can decrease reliability; adding a road to a congested network can make things worse. "It might be inferred from these examples that real conflict and cooperation should be left to the management of experts. Such an inference is not justified. On the contrary, . . . conflict and cooperation require the fresh and unprejudiced analysis of thinkers who have no commitment, historical or bureaucratic, to established rules of thumb."

Ascher, Marcia, Graphs in cultures: A study in ethnomathematics, *Historia Mathematica* 15 (1988) 201-227; Graphs in cultures (II): A study in ethnomathematics, *Archive for History of Exact Sciences* 39:1 (1988) 75-95.

These marvelous articles deal with continuous figure-tracing as it occurs in Oceania and Africa among peoples who do not read or write. The complex of mathematical ideas involved is presented with details of the contexts of the ideas and cultures in which they arose.

Kolata, Gina, Theorist applies computer power to uncertainty in statistics, *New York Times* (8 November 1988) 19, 21.

Heralds the bootstrap method, invented in 1977 by Bradley Efron (Stanford). The method involves taking an original data set, considering it as a population, and taking samples from it (with replacement). From these latter samples, one can estimate the standard deviation of *any* statistic based on the original sample. The bootstrap can require a thousand times as much calculation as traditional methods. In *The Jackknife, the Bootstrap and Other Resampling Plans* (SIAM, 1982), Efron described it as a "prodigious computational spendthrift . . . [a] substitution of computational power for theoretical analysis."

Cipra, Barry A., Computer-drawn pictures stalk the wild trajectory, *Science* 241 (2 September 1988) 1162-1163; Peterson, Ivars, In the shadows of chaos, *Science News* 134 (3 December 1988) 360-361.

The sequences of points generated by a chaotic function are particularly sensitive to the initial starting point, so there is some skepticism whether the computer-drawn graphs of these chaotic trajectories are really accurate. New methods have the computer itself verify that the computed trajectory stays close to a true trajectory.

Cipra, Barry A., Mathematicians reach factoring milestone, *Science* 242 (21 October 1988) 374-375; Peterson, Ivars, Cracking the 100-digit factoring barrier, *Science News* 134 (22 October 1988) 263.

Mathematicians, with their computers working together in parallel, have succeeded in factoring a 100-digit number. The number factored was $11^{104} + 1$, and the method used was the quadratic sieve.

Hoffman, Paul, Fermat still has the last laugh, *Discover* 10.1 (January 1989) 48-50.

Recalls the excitement in February, when Y. Miyaoka (Tokyo Metropolitan) thought he had a proof of Fermat's Last Theorem. The article includes a splendid full-page color reproduction of a little-known portrait of Fermat from the French Academy of Sciences.

Brams, Steven J., and Kilgour, D. Marc, *Game Theory and National Security*, Blackwell, 1988; xiii + 199 pp, $49.95, $19.95 (P).

Game-theoretic analyses of various games (arms races, deterrence, disarmament, Star Wars, threatening, verification) associated with national foreign policies. Most mathematical details are confined to appendices to the chapters.

Committee on the Mathematical Education of Teachers (COMET), *Guidelines for the Continuing Mathematical Education of Teachers*, MAA, 1988; ix + 83 pp.

Offers sample curricula (with syllabi) for master's degree programs for teachers and supervisors of high-school mathematics and for a mathematics concentration for elementary-school teachers.

Minsky, Marvin, *The Society of Mind*, Simon & Schuster, 1986; 339 pp, $9.95 (P).

Minsky asks, "How does the mind work?" His answer: "[Y]ou can build a mind from many little parts, each mindless by itself." The book itself is organized on much the same basis that Minsky attributes to the mind: Each subsection starts on a new page and occupies two-thirds to all of it.

Cipra, Barry A., To have and have knot: When are knots alike?, *Science* (9 September 1988) 1291-1292; Tying up a knotty loose end, *Science News* (29 October 1988) 283.

C. Gordon (Texas) and J. Luecke (Courant Institute) have settled an 80-year-old conjecture in knot theory: Two knots are the same if and only if the space around them is the same, i.e., the knots are homeomorphic if and only if their complements are homeomorphic. In conjunction with a theorem of W. Whitten (Southwestern Louisiana), we now know that two prime knots are homeomorphic if and only if they have the same fundamental group.

Publications by the

Mathematical Association of America

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

New Mathematical Library #30 . . .

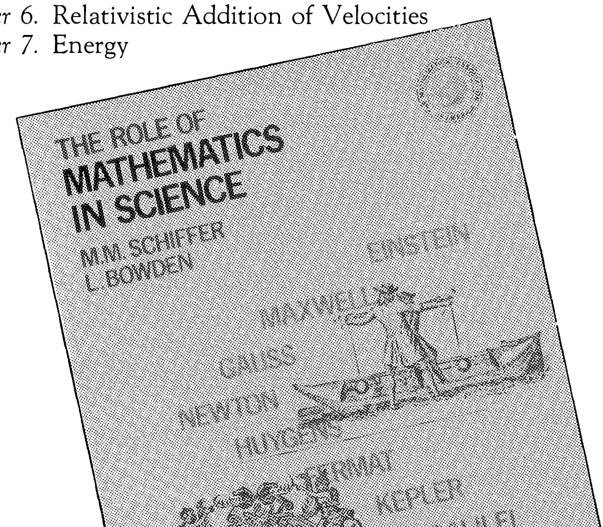
The Role of Mathematics in Science
by Leon Bowden and M.M. Schiffer
220 pp. Paperbound
List: $16.00 MAA Member: $12.50

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

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

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

Table of Contents



U.S. Postal Service

STATEMENT OF OWNERSHIP, MANAGEMENT AND CIRCULATION

Required by 39 U.S.C. 3685)

1A. TITLE OF PUBLICATION: MATHEMATICS MAGAZINE

1B. PUBLICATION NO.: 0 0 2 5 5 7 0 X

2. DATE OF FILING: Sept. 27, 1988

3. FREQUENCY OF ISSUE: bi-monthly except July and August

3A. NO. OF ISSUES PUBLISHED ANNUALLY: Five

3B. ANNUAL SUBSCRIPTION PRICE: Library $45; Indiv.Member $13

4. COMPLETE MAILING ADDRESS OF KNOWN OFFICE OF PUBLICATION *(Street, City, County, State and ZIP+4 Code) (Not printers)*
1529 Eighteenth St.,N.W., Washington, D.C. 20036

5. COMPLETE MAILING ADDRESS OF THE HEADQUARTERS OF GENERAL BUSINESS OFFICES OF THE PUBLISHER *(Not printer)*
1529 Eighteenth St.,N.W., Washington, D.C. 20036

6. FULL NAMES AND COMPLETE MAILING ADDRESS OF PUBLISHER, EDITOR, AND MANAGING EDITOR *(This item MUST NOT be blank)*

PUBLISHER *(Name and Complete Mailing Address)*
The Mathematical Association of America, 1529 Eighteenth St.,N.W.,Washington,D.C. 20036

EDITOR *(Name and Complete Mailing Address)*
Gerald L. Alexanderson, Santa Clara Univ., Dept. of Math., Santa Clara, CA 95053

MANAGING EDITOR *(Name and Complete Mailing Address)*
Harry Waldman, MAA, 1529 Eighteenth St.,N.W., Washington, D.C. 20036

7. OWNER *(If owned by a corporation, its name and address must be stated and also immediately thereunder the names and addresses of stockholders owning or holding 1 percent or more of total amount of stock. If not owned by a corporation, the names and addresses of the individual owners must be given. If owned by a partnership or other unincorporated firm, its name and address, as well as that of each individual must be given. If the publication is published by a nonprofit organization, its name and address must be stated.) (Item must be completed.)*

FULL NAME	COMPLETE MAILING ADDRESS
The Mathematical Association of America, Inc.	1529 Eighteenth St.,N.W. Washington, D.C. 20036

8. KNOWN BONDHOLDERS, MORTGAGEES, AND OTHER SECURITY HOLDERS OWNING OR HOLDING 1 PERCENT OR MORE OF TOTAL AMOUNT OF BONDS, MORTGAGES OR OTHER SECURITIES *(If there are none, so state)*

FULL NAME	COMPLETE MAILING ADDRESS
None	

9. FOR COMPLETION BY NONPROFIT ORGANIZATIONS AUTHORIZED TO MAIL AT SPECIAL RATES *(Section 423.12 DMM only)*
The purpose, function, and nonprofit status of this organization and the exempt status for Federal income tax purposes *(Check one)*

(1) [X] HAS NOT CHANGED DURING PRECEDING 12 MONTHS
(2) [ ] HAS CHANGED DURING PRECEDING 12 MONTHS
(If changed, publisher must submit explanation of change with this statement.)

10. EXTENT AND NATURE OF CIRCULATION *(See instructions on reverse side)*	AVERAGE NO. COPIES EACH ISSUE DURING PRECEDING 12 MONTHS	ACTUAL NO. COPIES OF SINGLE ISSUE PUBLISHED NEAREST TO FILING DATE
A. TOTAL NO. COPIES *(Net Press Run)*	16,457	15,636
B. PAID AND/OR REQUESTED CIRCULATION 1. Sales through dealers and carriers, street vendors and counter sales	0	0
2. Mail Subscription *(Paid and/or requested)*	13,847	12,991
C. TOTAL PAID AND/OR REQUESTED CIRCULATION *(Sum of 10B1 and 10B2)*	13,847	12,991
D. FREE DISTRIBUTION BY MAIL, CARRIER OR OTHER MEANS SAMPLES, COMPLIMENTARY, AND OTHER FREE COPIES	433	328
E. TOTAL DISTRIBUTION *(Sum of C and D)*	14,280	13,319
F. COPIES NOT DISTRIBUTED 1. Office use, left over, unaccounted, spoiled after printing	2,177	2,317
2. Return from News Agents	0	0
G. TOTAL *(Sum of E, F1 and 2—should equal net press run shown in A)*	16,457	15,636

11. I certify that the statements made by me above are correct and complete

SIGNATURE AND TITLE OF EDITOR, PUBLISHER, BUSINESS MANAGER, OR OWNER: Executive Director, MAA

PS Form 3526, Dec. 1985 *(See instruction on reverse)*

TWO BOOKS ON MATHEMATICS EDUCATION

TOWARD A LEAN AND LIVELY CALCULUS
Report of the Conference/Workshop to Develop Curriculum and Teaching Methods for Calculus at the College Level

Ronald G. Douglas, Editor
MAA Notes #6
249 pp., Paperbound, 1987,
ISBN-0-88385-056-7
List: $12.50
Catalog Number NTE-06

Should calculus be taught differently? Can it? Common wisdom says "no"—which topics are taught, and when, are dictated by the logic of the subject and by client departments. The surprising answer from a four-day Sloan Foundation-sponsored conference on calculus instruction, chaired by Ronald Douglas, is that significant change is possible, desirable and necessary. Meeting at Tulane University in New Orleans in January, 1986, a diverse and sometimes contentious group of twenty-five faculty, university, and foundation administrators, and scientists from client departments put aside their differences to call for a leaner, livelier, more contemporary course, more sharply focused on calculus's central ideas and on its role as the language of science.

This volume contains the results of that conference and the papers presented to the conferees. These are certain to be the point of departure and basis for efforts to strengthen and reshape calculus in the next decade.

CALCULUS FOR A NEW CENTURY: A PUMP NOT A FILTER

L.A. Steen, Editor
MAA Notes #8
272 pp., Paperbound, 1987,
ISBN-0-88385-058-3
List: $12.50
Catalog Number NTE-08

Proceedings of the Colloquium held in October 1987 in Washington, D.C. to discuss calculus reform. The mathematical community is challenged to make the introductory calculus course into a pump that feeds more students into science and engineering, not a filter that cuts down the flow. These proceedings, with contributions from over eighty authors, show the full sweep of concerns and approaches of all the groups involved in calculus reform, including those currently teaching traditional and innovative courses, those whose students or employees need to use calculus as a tool, and the department chairs, deans, and others who must mobilize the resources needed for this reform.

THE MATHEMATICAL ASSOCIATION OF AMERICA

ORDER FORM

Please send the following books:

Quantity	Catalogue Number	Unit Price	Price
____	NTE-06	$12.50	____________
____	NTE-08	$12.50	____________
		Total $	____________

Name: ______________________________

Address: ______________________________

______________________________ ☐ Payment enclosed (sent postpaid)

______________________________ ☐ Please bill me (postage and handling extra)

Mail to: Mathematical Association of America, 1529 Eighteenth St., N.W., Washington, DC 20036

Studies in Mathematical Economics

Volume 25 in the MAA Studies in Mathematics
Edited by Stanley Reiter

420 pp. Hardbound
ISBN-0-88385-027-X

List: $42.00
MAA Member: $31.00

> *"For the mathematician desiring to become familiar with modern mathematical, microeconomic theory, this volume is indispensable."*
>
> Robert Rosenthal
> SUNY, Stony Brook
> Department of Economics

Stanley Reiter, as editor, has brought together a distinguished group of contributors in this volume, in order to give mathematicians and their students a clear understanding of the issues, methods, and results of mathematical economics. The range of material is wide: game theory; optimization; effective computation of equilibria; analysis of conditions under which economies will move to the greatest possible efficiency under various forces, and the requirements for the flow of information needed to achieve efficient markets.

The material is interesting at all mathematical levels. For example, the initial article shows how even mathematically simple, concrete, two-person, nonzero sum games present us with the complexities and dilemmas of choices in real life. At the other extreme, the final article, by Debreu, begins by using the power of Kakutani's fixed point theorem to prove the existence of economic equilibria. In between, the reader will find beautiful uses of calculus, topology, combinatorial topology, and other topics.

The chapters of this volume can be read independently, although they are related. The book begins with Meyerson's chapter on game theory and its theoretic foundations. The second chapter, by Simon, starts with the familiar criteria for maxima from calculus and goes on to develop more general tools of mathematical economics, including the Kuhn-Tucker and related conditions. The third contribution, by Mas-Collell, uses the tools of differential topology, including Sard's theorem, to study the competitive equilibria of whole families of economies using a differentiable point of view. Next Kuhn, building on the work of Scarf, shows how methods based on Sperner's lemma can be used to compute equilibria.

The next two chapters by Reiter and Hurwicz explore the properties of systems that are not purely competitive. They bring analytical and topological tools to bear to determine what conditions on the exchange of information are needed to allow such markets to become optimally efficient.

Radner addresses one consequence of what Herbert Simon calls "bounded rationality." Managers neither know all the facts nor do they have unlimited ability to calculate. How should they allocate their time? The tools used to answer this question are fittingly probabilistic.

In the final chapter, Debreu gives four examples of mathematical methods in economics. These four examples alone give a sense of the breadth and nature of the field.

In this study, Reiter and his other contributors show the reader the subtlety and complexity of the subject along with the precision and clarity that mathematics bring to it.

ORDER FROM

The Mathematical Association of America
1529 Eighteenth Street, NW
Washington, DC 20036

CONTENTS

ARTICLES

NOTES

PROBLEMS

REVIEWS

THE MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth Street, NW
Washington D.C. 20036