MATHEMATICS MAGAZINE

CONTENTS

VOLUME 37 • MARCH 1964 • NUMBER 2

MATHEMATICS MAGAZINE

ROY DUBISCH, *Editor*

ASSOCIATE EDITORS

DAVID B. DEKKER
RAOUL HAILPERN
ROBERT E. HORTON
CALVIN T. LONG
SAM PERLIS
RUTH B. RASMUSEN
H. E. REINHARDT
ROBERT W. RITCHIE
J. M. SACHS
HANS SAGAN
S. T. SANDERS (Emeritus)

EDITORIAL CORRESPONDENCE should be sent to the Editor, ROY DUBISCH, Department of Mathematics, University of Washington, Seattle, Washington, 98105. Articles should be typewritten and double-spaced on $8\frac{1}{2}$ by 11 paper. The greatest possible care should be taken in preparing the manuscript, and authors should keep a complete copy. Figures should be drawn on separate sheets in India ink and of a suitable size for photographing.

NOTICE OF CHANGE OF ADDRESS and other subscription correspondence should be sent to the Executive Director, H. M. GEHMAN, Mathematical Association of America, SUNY at Buffalo (University of Buffalo), Buffalo, New York, 14214.

ADVERTISING CORRESPONDENCE should be addressed to F. R. OLSON, Mathematical Association of America, SUNY at Buffalo (University of Buffalo), Buffalo, New York, 14214.

The MATHEMATICS MAGAZINE is published by the Mathematical Association of America at Buffalo, New York, bi-monthly except July-August. Ordinary subscriptions are: 1 year $3.00; 2 years $5.75; 3 years $8.50; 4 years $11.00; 5 years $13.00. Members of the Mathematical Association of America may subscribe at the special rate of 2 years for $5.00. Single copies are 65¢.

PUBLISHED WITH THE ASSISTANCE OF THE JACOB HOUCK MEMORIAL
FUND OF THE MATHEMATICAL ASSOCIATION OF AMERICA

Second class postage paid at Buffalo, New York and additional mailing offices.

Copyright 1964 by The Mathematical Association of America (Incorporated)

INFORMATION THEORY

HARTLEY ROGERS, JR., Massachusetts Institute of Technology

The phrase "Information Theory" is occasionally used, rather broadly, to include two different, but interrelated areas. The first area concerns *analysis, filtering, and prediction of statistical processes.* (These are technical terms, unimportant for what follows.) Basic work in this area was done by Norbert Wiener at M.I.T. The second area concerns *efficient use of communications channels in the transmission of information.* Basic work in this second area was done by Claude Shannon at Bell Telephone Laboratories and at M.I.T. In recent years, the phrase "Information Theory" has been increasingly used to refer specifically to the second of these areas. We shall use the phrase in this latter sense and shall be considering the second area.

Information theory is customarily included as a part of Communications Engineering. It has, however, considerable mathematical content; indeed, some of its basic results consist of new and rather remarkable mathematical facts. For mathematicians, engineers, and other scientists, these facts, and their consequences, have been a source of continuing stimulation and insight. In Parts I and II below, two of these facts, the so-called *Coding Theorems*, will be described in outline. In Part III, some of their uses and implications will be indicated.

As a partial introduction to what follows, we make a brief historical comment. Prior to the 1940's, problems of design in Communications Engineering were largely formulated in terms of broad, generic specifications (e.g., 'to transmit sound waves', 'to transmit a continuously varying voltage', 'to transmit letters of the English alphabet'). Since the 1940's, and as a result of the work of both Wiener and Shannon, design problems have increasingly taken into account the structure, especially the statistical structure, of the messages to be communicated. In many cases, this has yielded more effective and efficient solutions to engineering problems; in some cases, it has given solutions where otherwise none would have been found. Information Theory illustrates, and helps account for, this recent emphasis.

Part I. Assume that two cities, Boston (B) and San Francisco (S), have the following one-way communications channel from B to S. In B there are two pushbuttons, labelled R and G. These are connected by wire and battery to two lights in S. The lights are red and green; if button R is pressed, the red light flashes, and if button G is pressed, the green light flashes. Messages are sent from B to S by a succession of pushes on one or the other button. Assume, furthermore, that once a message is begun, the buttons must be pressed at a steady rate of one push per second without interruption. Finally, assume that for each push the sender in B must pay a substantial fee; so that he will wish to use as few pushes as possible for his messages—or, to put it another way, so that he will wish to send his messages as *fast* as possible. This rather artificial channel is represented in the following diagram.

$$\text{Pushbuttons}\left\{\begin{tikzcd} R \arrow[r] & \text{red} \\ G \arrow[r] & \text{green} \\ B & S \end{tikzcd}\right\}\text{lights.}$$

For brevity, we agree to write 'red' as '0' and 'green' as '1'. We call the symbols '0' and '1' *binary digits*, or, more succinctly, 'bits'. One of the sender's objectives will be to use as few bits as possible for his messages. Our diagram becomes:

$$\begin{tikzcd} 0 \arrow[r] & 0 \\ 1 \arrow[r] & 1 \\ B & S \end{tikzcd}$$

Clearly, the sender needs a *code*. Ordinary Morse Code comes to mind, but Morse Code appears to use *three* signals, *dot*, *dash*, and *pause* (one pause between letters, two pauses between words). The sender will want a *two* signal code. If his messages are in written English, he needs a code for an alphabet of 27 letters (ignoring punctuation, but counting *space* as a 27th letter).

In order to give ideas simply, we shall assume that messages are given in a language with a four letter alphabet: *A*, *B*, *C*, and *space*. We seek an economical code into bits. The following suggests itself.

$$\begin{array}{rl} A \leftrightarrow & 00 \\ B \leftrightarrow & 01 \\ C \leftrightarrow & 10 \\ \textit{space} \leftrightarrow & 11. \end{array}$$

Thus the message *BACA ABA* would be encoded as 0100100011000100. Note that decoding proceeds without difficulty; (a first step is marking off the coded text in pairs). We can measure the *cost* of our code as 2 bits/letter, or, inversely, we can measure its *speed* as 1/2 letter/bit.

Can we do better than this? Rather surprisingly (at first glance), further economies may be possible, if we take into account the *statistical structure* of the given language. For example, we can look at the average frequencies (i.e., proportions) with which the various letters occur. (In ordinary English, for instance, the letter *e* is much more frequent than, say, the letter *q*; a fact that we use in solving newspaper 'cryptogram' puzzles.) We give two examples to show how further economies might be made.

Example I. Assume that, in our four letter alphabet, the four letters occur, *on the average*, in the following proportions.

A:	1/2
B:	1/4
C:	1/8
space:	1/8.

We try the code,

$$\begin{aligned} A &\leftrightarrow 0 \\ B &\leftrightarrow 10 \\ C &\leftrightarrow 110 \\ \textit{space} &\leftrightarrow 111. \end{aligned}$$

The message *BACA ABA* would now be encoded as 10011001110100. We note that decoding is again unambiguous (once we have located the beginning of the message), although it is a slightly more complex operation than before.

How economical is this code? Occasional individual messages may cost more than before (just as occasional sentences in English can have more *q*'s then *e*'s), but the cost, *on the average*, *over a long run*, can be computed as:

$$1/2(1) + 1/4(2) + 1/8(3) + 1/8(3) = 1.75 \text{ bits/letter,}$$

or the speed as .57 letter/bit. This is an improvement over the previous cost of 2 bits/letter and speed of .5 letter/bit.

Example II. Here we assume that messages are sent in a different language, a language with 2 letters, *A* and *B* (and no space); and we assume that these letters occur in proportions,

A:	2/3
B:	1/3.

At first glance, the only possible codes appear to be

$$\begin{aligned} A &\leftrightarrow 0 \\ B &\leftrightarrow 1, \end{aligned} \qquad \text{or} \qquad \begin{aligned} A &\leftrightarrow 1 \\ B &\leftrightarrow 0, \end{aligned}$$

either of which has a cost of 1 bit/letter.

This cost can be improved on, however, if pairs of letters are used. (We call a pair of letters a *digram*, a triplet of letters a *trigram*.) Assume that digrams occur in the average proportions,

AA:	4/9
AB:	2/9
BA:	2/9
BB:	1/9.

We try the code,

$$\begin{aligned} AA &\leftrightarrow 0 \\ AB &\leftrightarrow 10 \\ BA &\leftrightarrow 110 \\ BB &\leftrightarrow 111. \end{aligned}$$

(Thus *AAABAABABBAA* would be encoded as 01001101110.) The average cost of this code can be computed as:

$$4/9(1) + 2/9(2) + 2/9(3) + 1/9(3) = 17/9 \text{ bits/digram}$$
$$= 17/18 = .94 \text{ bit/letter.}$$

And we have an improvement.

There is a lack of symmetry to this code which leads us to suspect (correctly) that further improvements may be possible. We also note that it can encode only an even number of letters. For an odd length message, special end-of-message provisions would be necessary. With reasonably long messages, or with message traffic high enough so that messages could be joined end to end, the influence of this 'end effect' on our cost figure would be negligible.

Some terminology. A code like that in Example II, which associates blocks of bits with blocks of letters, is called a *block code.* A message language with statistical properties of the kind used in Examples I and II is called a *statistical source*, or, more simply, a *source.* The average proportions of letters, digrams, etc. in a source are called their *probabilities* from that source. A source is called *independent* if, once the letter probabilities are fixed, the remaining statistical structure is the same as would be obtained if successive letters of messages were determined by identical, independent experiments, each producing a single letter according to the fixed single letter probabilities. E.g., the digram frequencies assumed in Example II are those which would hold in an independent source with frequencies A 2/3, B 1/3; for they coincide with results of the usual multiplication law for independent probabilities. Ordinary English is not an independent source: e.g., the probability of *th* is not the product of the probabilities for t and for h, but is instead somewhat larger.

If a source is not independent, substantial economies beyond those indicated by the single letter frequencies *may* be possible. For example, assume an alphabet the same as in Example I, but assume that the only words ever used are *ABABACA* and *BACAABA*, where each word occurs with probability 1/2. Although the single letter frequencies are the same as in Example I, the following block code would clearly suffice,

$$ABABACA \textit{ space} \leftrightarrow 0$$
$$BACAABA \textit{ space} \leftrightarrow 1,$$

and we have a cost of .125 bit/letter, or a speed of 8 letters/bit. This source is far from being independent, since, in an independent source with the same letter frequencies, each of these words would (by the product rule) have probability below .001.

We are now ready to state the main fact of Part I. It is known as the *Source Coding Theorem.*

THEOREM I. *Given a statistical source with an alphabet of k letters, occurring in average proportions $p_1, p_2, \cdots, p_k$, there is a number H such that:*

(1) *no block code has cost below H bits/letter*;

(2) *block codes can be found whose average cost is within any given tolerance of H* (more simply, with mild inaccuracy, "block codes exist costing only H bits/letter");

(3) *H is not greater than* $-(p_1 \log_2 p_1+p_2 \log_2 p_2+ \cdots +p_k \log_2 p_k)$; ($\log_2 x$ is that power of 2 which yields x e.g., $\log_2 (1/8=-3)$. The quantity $-(p_1 \log_2 p_1+ \cdots +p_k \log_2 p_k)$, is positive, since the logarithms are all negative. Of course, $p_1+p_2+ \cdots +p_k=1$);

(4) *if the source is independent, H equals* $-(p_1 \log_2 p_1+ \cdots +p_k \log_2 p_k)$. *$H$ is called the entropy of the given source.*

(This theorem has also been called the *First Coding Theorem* and the *Noiseless Channel Coding Theorem.*)

Applying Theorem I to the preceding examples, we get the following:

Example I. $-(p_1 \log_2 p_1 + \cdots + p_4 \log {}_2p_4)=-(1/2(-1)+1/4(-2)+1/8(-3)+1/8(-3))=1.75$ bits/letter. Hence, if the source in Example I happens to be independent, the code of Example I is best possible.

Example II. $-(p_1 \log_2 p_1 + p_2 \log_2 p_2)=-(2/3 \log_2 2/3+1/3 \log_2 1/3)=.92$ bit/letter.

Here there must, in any case, be a code better than that given.

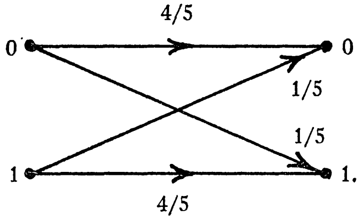
We do not give a proof of the theorem. A rough argument leading to the logarithmic formula is not difficult. A fully general mathematical proof requires satisfactory and precise definitions for the concept of *statistical source* and the notion of *average cost.* This is a matter of some depth and subtlety. Further details can be found in *Mathematical Foundations of Information Theory* by A. I. Khinchin, Dover Publications, 1957 (translated from the Russian).

So far, our attention has been limited to a rather special and artificial channel. We might call it the *bit-channel.* What about a channel of the kind used with Morse Code? Or a channel with 10 buttons instead of 2? Surprisingly, and conveniently, questions of cost and coding for these other channels can be rather easily formulated and answered in terms of our bit-channel. This is done by first introducing further codings between the other given channels and the bit-channel. (For instance, as a first approximation, the *dot*, *dash*, and *pause* of Morse Code might be taken as the code blocks 0, 10, and 11 on the bit-channel.) More detailed investigation (which we omit) shows that the bit-channel and its entropy measure can serve as a useful absolute standard in codability questions. The bit-channel rarely exists in practice; it is a helpful fiction for discussing properties of any given source.

Indeed, it is possible to think of a statistical source as producing a measurable and possibly conservable "substance" which we call *information.* It can be measured in *bits* (the number of them necessary for coding); and, as Shannon has remarked, the entropy of a source, in bits/letter, is then a measure of the *rate* at which that source produces information, much as an amount in tons/day might be a measure of the output of a steel mill.

What about one of the most commonly used sources in daily life: ordinary written English? What is its apparent entropy? Using observed letter proportions, the theorem yields: H no greater than 4.03. As English is not an independent source, H may be considerably below 4.03. In Part III we shall mention experiments that point to an entropy of approximately 1 bit/letter. We shall also, in Part III, briefly comment on application of the entropy concept to "sources" that appear to give a continuously varying signal as output, rather than letters of an alphabet. Spoken English, with its output of sound waves, is an example of such an apparent "continuous source."

Part II. Consider a different and more realistic one-way channel from city B to city S. It has the same pushbuttons at B and the same lights at S. Assume, however, that random disturbances occur along the line in such a way that when a button is pressed, there is probability 4/5 that the correct light will flash and probability 1/5 that the incorrect light will flash. The channel can be diagrammed,



Assume, as before, that buttons are pushed at a rate of one push per second. Assume, further, that occurrence of successive *signal-errors* (i.e., incorrect flashes) follows the probability rules for a sequence of independent events (i.e., probabilities multiply).

A channel with error producing disturbances is called a *noisy channel.* A noisy channel such as the above would be obtained, for instance, if a single wire were laid from B to S, if pressing button 0 or 1 put a positive or negative voltage (with respect to ground) on the wire, and if light 0 or 1 flashed according as a sensitive and synchronized measurement of a positive or negative voltage were made at S. For voltages of ordinary size, losses along the line and variations due to thermal agitation of electrons could introduce errors of the kind we are assuming.

Example III. As a first simple problem, assume that we wish to send a single important bit of information with high accuracy from B to S. (For instance, we might wish to communicate, at a set time, news of success or failure of some scientific project in B; where this news might be vital for the immediate choice between two different and expensive lines of investigation at S.) The first signaling scheme that suggests itself is

(a) send 0 for news of failure,
send 1 for news of success.

Unfortunately, the probability of *message-error* is .20, an intolerably high value. This might be improved by physical changes in the channel, such as better wire or higher voltage. After a little thought, however, we see that proper coding can achieve the same effect more cheaply and simply. One possibility for coding would be,

(b) *Encoding*	*Decoding*
failure →000	$\left.\begin{matrix}000\\001\\010\\100\end{matrix}\right\}\rightarrow\text{failure}$
success →111	$\left.\begin{matrix}110\\101\\011\\111\end{matrix}\right\}\rightarrow\text{success.}$

Note that the decoding process is more complex. Because of signal-errors, all possible sequences of bits can occur for decoding. This particular decoding could be called *majority vote decoding*, for obvious reasons. Of course, message-error can still occur. What is its probability? An elementary calculation gives the value .10, a noticeable improvement.

This coding technique could be extended further:

(c) *Encoding*	*Decoding*
failure →string of 16 0's	by majority vote.
success→string of 16 1's	

Here the probability of message error is below .01. This might well be a tolerable value for the purposes of the communicators in *B* and *S*. If not, the strings could be made still longer. We thus find that our problem is solved by suitable coding.

Example IV. Turn now to the problem of sending longer messages. Assume that our messages are already coded into bits as described in Part I. Our problem is to send these bits as accurately as we can over the noisy channel. From Example III, an obvious answer suggests itself: repeat each bit of the message, in turn, sufficiently often. Thus, if we repeat each message bit 16 times, the error probability per message bit drops below .01. In this way, we can reduce the error as much as we like, but we can only do so at the expense of also reducing the *rate* of transmission in *bits* per *signal.* (Where, by *bit* we mean message-bit, and by *signal* we mean channel-signal.) Thus we have,

Probability of error (per bit)	*Rate (bits/signal)*
.20	1
.10	.33
.01	.06

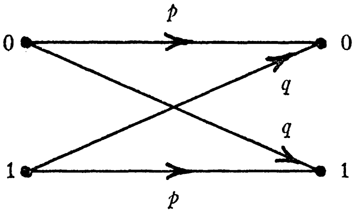
It would appear, from Example IV, that in order to keep pushing the error probability down to zero, we must also keep pushing the rate down to zero. However, and this is perhaps the most remarkable basic result of Information Theory, *this is not true.* We first state this extraordinary fact in rough form for our given noisy channel.

Provided we are dealing with messages involving long sequences of bits, there is a positive value C such that the error probability can be reduced to any desired smallness without our having to lower the average rate in bits/signal, below C. This is accomplished by appropriate block codes from message-bits to channel-signals.

In order to lower the error without lowering the rate below C, we must pay the penalty of having the blocks in the code lengthened—both the blocks of message-bits and the blocks of channel-signals. This may introduce a delay time between sender and receiver. However, we do not pay the penalty of having to lower the average rate of transmission below C.

This fact is included in the theorem to be stated immediately below. The value of C varies, in general, from channel to channel. The theorem will give a formula for finding the value of C in certain special cases.

Some terminology. A *binary symmetric* channel is any channel like that used above, but with possibly different fixed probabilities in place of 4/5 and 1/5. We call these probabilities p and q and require that $p+q=1$. The channel used above is a particular binary symmetric channel. Such channels have the following diagram.



We now state the theorem. It is the main fact of Part II, and is known as the *Noisy Channel Coding Theorem.*

THEOREM II. *Given a noisy channel, there is a number C such that:*

(1) *for any tolerance ϵ and any value R below C, a block code can be found which transmits at average rate R bits/symbol* (*or letter*) *with a maximum error probability smaller than ϵ;*

(2) *for average rates above C, it is not true that for each tolerance ϵ a code can be found with maximum error probability smaller than ϵ;*

(3) *assertions* (1) *and* (2) *are true whether 'error' is interpreted as error per message-bit or as error per code block*;

(4) *for a binary symmetric channel,* $C=1+p\log_2 p+q\log_2 q$.

C is called the capacity *of the channel.*

(This theorem has also been called the *Second Coding Theorem.* Note: the situation when R is exactly equal to C is rather special. If "error" is error per message-bit, then the desired block codes can be found for all tolerances ϵ. If "error" is error per code block, then, in some cases, the desired codes do not exist.)

Applying the theorem to the noisy channel first considered, we obtain:

$$C = 1 + 4/5 \log_2 4/5 + 1/5 \log_2 1/5 = .29 \text{ bit/signal.}$$

Therefore we can get block codes with smaller and smaller error probability without going much beyond a ratio of three signals to one bit.

The following example shows the phenomenon described in Theorem II beginning to occur. We take the given channel with $C=.29$, and use codes with $R=.2$.

Example V. (a) Consider first the block code from single bits to signals,

Encoding	*Decoding*
0→00000	by majority vote.
1→11111	

By elementary calculation, the probability of error per bit is .058. Under this code, pairs would be encoded,

$$\begin{aligned} 00 &\rightarrow 0000000000\\ 01 &\rightarrow 0000011111\\ 10 &\rightarrow 1111100000\\ 11 &\rightarrow 1111111111 \end{aligned}$$

For each pair the probability of error is $2(.058)-(.058)^2=.11+$.

(b) Consider next the following block code from pairs of bits to signals. This is a *basic* block code in that (unlike the coding of pairs in (a)) it cannot be decomposed into smaller blocks.

$$\begin{aligned} 00 &\rightarrow 1111100000\\ 01 &\rightarrow 0101010101\\ 10 &\rightarrow 1000001011\\ 11 &\rightarrow 0010111110 \end{aligned}$$

Under appropriate decoding (which we do not describe) it is possible to show that error probability per pair in code (b) is below .10 and hence, is smaller than that in code (a). (For binary symmetric channels with much smaller q, the error probability in (b) will be about half of that in code (a).) Calculations are elementary, but time consuming, and we omit them. Thus going from a one-to-five block code to a two-to-ten block code has reduced the error without affecting the rate of .2 bits/signal.

Applications of Theorem I and Theorem II can be combined. Thus if we use the source of Example I with the (4/5, 1/5) noisy channel, we have entropy $H=1.75$ bits/letter and capacity $C=.29$ bit/signal. If coding operations are combined, we get $C/H=.17$ letter/signal, or, inversely, $H/C=6.0$ signals/letter, as the limiting value for block codes that give the desired low error transmission of information. Of course, these codes may involve long blocks, so that individual source messages can be sent at close to the limiting rate only if they are sufficiently long.

The known proofs of Theorem II are complex. Details can be found in the Khinchin volume mentioned above. A recent summary proof is given in an expository paper by D. Blackwell: "Information Theory", in *Modern Mathematics for the Engineer, 2nd series*, edited by Beckenbach (McGraw-Hill, 1961). Theorem II is the major basic result of Information Theory. We have given it in a simple and restricted form; it can be extended to other kinds of noisy channels with more general methods and formulas for computing C. This will be commented on in Part III.

As Example V suggests, (with its small error improvement), significant error improvement at a given rate may demand a substantial increase in basic block length. We discuss this further at the end of Part III. Furthermore, not all block codes of the right length give improved error values. The problem of explicitly *finding* the useful codes is difficult. Theorem II tells us they exist, but known proofs of the theorem do not provide efficient ways for finding these codes. Much recent work has gone into the problem of identifying such codes and describing them in simple and useful forms.

The desirable codes whose existence is proved in Theorem II are called *error-correcting codes.*

Part III. Theorems I and II were discovered by Shannon and presented in his papers: "The Mathematical Theory of Communication," *Bell System Technical Journal,* 1948, pp. 379–423 and 623–656; and "Communication in the Presence of Noise," *Bell System Technical Journal,* 1949, pp. 10–21. In these papers, Shannon also formulates results for 'continuous' sources and channels, i.e., sources and channels whose behavior is customarily associated with continuous wave forms. We have neglected these so far.

Since the publication of Shannon's work, researchers in increasing numbers have studied Information Theory and its implications for other parts of engineering and science. They have sought generalizations of the basic mathematical theorems. They have worked on applications to Communications Engineering, and they have studied various applications and possible applications in such other areas of science as genetics, biology, physics, psychology and the study of ordinary language. We comment on a few of these matters below.

Communications Engineering. As Shannon showed in his papers, the concepts of entropy and capacity can be applied to *continuous sources* (such as a voice speaking English) and *continuous channels* (such as sound waves in air) without difficulty. Given both a *source* and a *receiver* (such as a human), a finite entropy (in bits/second) can be meaningfully associated with the source if the receiver used is limited in its ability to distinguish between similar wave forms. This limitation enables us, in effect, theoretically to "quantize" the original messages. With a human receiver, experimental evidence suggests an entropy of from 20 to 50 bits/second for spoken English. The significance of this figure will be discussed further below.

Independently of source and receiver, a finite capacity can be associated with a continuous channel (such as sound waves in air) if disturbances of a random kind (such as background sounds in a crowded noisy room) are present, and if the message signals are limited in frequency and in average amplitude. Such disturbances are usually called *noise*. Background noise is invariably present in continuous channels, if only from thermal agitation in equipment used. Amplitude limitation always holds, and frequency limitation usually holds to a good approximation. Using results of Wiener, Shannon derived the formula $C=W \log_2 (1+P/N)$ bits/second for the case of the most random kinds of noise. Here W is width of frequencies used, P is a measure of average signal amplitude, N is a measure of average noise amplitude. We slight details of this whole important area, except to point out that C is positive even when P is considerably smaller than N (and when the signal would hence appear to be "buried" in the noise). The general result of Theorem II can be shown to hold in this continuous case; and therefore, with appropriate coding and decoding of wave forms, nearly error-free communication is still possible. Estimates of capacity for several continuous channels will be given below.

It should be remarked that binary channels are increasingly used in engineering practice. Digital computers communicate in bits; and even when continuous signals are used, recent devices reduce these signals to bits for certain kinds of processing. Thus the associated coding problems occur more and more often in the direct and simple form of Parts I and II. Communication rates on binary channels are usually quite high—often well into the thousands of bits/second. This has two consequences. First, messages are long enough so that, in principle, quite long basic block length can be used in error-correcting codes; thus low error communication at close to capacity is possible. Second, high speed digital computing equipment is necessary for the encoding and decoding operations, if advantage is to be taken of the higher rates made possible by the error-correcting codes.

Like entropy, the concept of *capacity* has proved to have a natural and absolute physical significance. Just as entropy of a source was likened to the output rate of a steel mill, so, as Shannon has remarked, can capacity of a channel (in bits/signal or bits/second) be likened to the maximum load capacity of a conveyor belt (in tons/day). If entropy (steel output) is below channel capacity (belt capacity) we know that the information (steel) produced can be satisfactorily transported, provided that it is properly coded (cut up, arranged and packed onto the moving belt).

Practical consequences of these ideas, even in rough qualitative form, have been profound. Once the capacity of a channel is calculated, we can see whether or not we are using it efficiently. If not, we can try, by appropriate coding, to improve matters. The idea of improved efficiency by coding is an old one. Bell was working on "multiplexing" of a telegraph channel (to carry several messages at once) when he discovered the telephone. The advantage of the capacity concept is that it shows us how far we are from maximum possible use. Indeed, it is fair to say that the recently successful search for better multiplex techniques on the Atlantic telephone cables was stimulated by knowledge of the small percentage of capacity in use.

Information theory also provides an excellent conceptual framework for discussing questions of channel use when noise is high and capacity is low. Thus, to improve communication in such a situation, we can try to raise capacity (by increasing P or W in the case of a continuous channel), or we can slow the transmission rate below capacity and look for error-correcting codes as suggested by Theorem II. In the latter case, we may have to pay the penalty of using extensive computing machinery to do the encoding and decoding. For a while, the coding approach was studied as a possible means of sending TV signals over the Atlantic telephone cables. In its raw, uncoded form, TV demands a higher capacity than the cables provide. The development of the Telstar and ECHO satellites, however, yields higher capacity channels where it is possible, and economically preferable, to send the uncoded TV signal rather than to use computers for better coding. Coding by digital computers may come into its own in outer space communications. Here, by miniaturization, computers may prove to be more economical of size and weight than are the enlarged sources of power that give increased signal amplitude; thus we may in this case meet the challenge of a low capacity channel (with P well below N) by coding rather than by increasing capacity. Of course, in all such situations, the ultimate engineering decision will depend upon the various costs involved.

In practice, the distinction between 'continuous' and 'discrete' (i.e., alphabetic) processes is somewhat arbitrary. It is more a matter of choosing a convenient and efficient mathematical model than it is of deciding whether a process is inherently discrete or inherently continuous. The latter question is often hazy. A virtue of the concepts of *entropy* and *capacity* is that they can be formulated for either the discrete or the continuous case.

The concepts and theorems of Information Theory play a role in engineering that is similar to the role that has long been played by the basic conservation and irreversibility laws of physical science. They mark off a clear boundary between the possible and the impossible. On one side of this boundary, the engineer can give his ingenuity free play—with hope of success; on the other side, he knows that any search (like the classical search for a perpetual motion energy source) must fail.

One final comment on the engineering implications of Theorem II: Theorem II shows how, by coding, individual signal errors in a channel may be compensated for in a way that yields overall reliable communication. It suggests looking for similar methods and results in the case of a quite different problem, that of building a reliable machine out of many, possibly unreliable, components. (Rocket firing difficulties illustrate the problem.) So far, no results comparable to Theorem II in generality and significance have been obtained. In particular, no counterpart to the concept of *capacity* has been found, although several interesting theorems have been proved.

Languages. The entropy of ordinary written English can be estimated by various means. Using single letter proportions, Theorem I gives 4.03 as a maximum value for H. A detailed proof of Theorem I also furnishes more complex formulas that give improved maximum values for H in terms of the digram proportions, in terms of the trigram proportions, etc. Using observed frequencies for English, we obtain from the digram formula the value 3.32 and from the trigram formula the value 3.1.

A simple experiment due to Shannon gives a still better estimate. Take a fairly long sentence (Shannon often used: "there is no reverse on a motorcycle a friend of mine found this out rather dramatically the other day"), and use as subject for the experiment a person to whom the sentence is unknown. Give the subject a diagram of the sentence where each letter is represented as an empty box. (There are 101 letters, including spaces, in Shannon's sentence.) Ask the subject to guess individual letters (from the 27 letter alphabet) until the first letter of the sentence is correctly guessed, then enter this letter on the diagram and proceed to the second letter, etc. Count the number of guesses required by the subject to determine the full sentence. The subject is encouraged to use his background knowledge of English in order to keep his score as low as possible. (In Shannon's sentence, most subjects require only one guess each for the first three letters, and two or three guesses for the fourth letter. Some subjects, once the '*m*' in '*motorcycle*' is reached, will get the rest of '*motorcycle*' at one guess per letter; but the '*r*' and '*v*' in '*reverse*' may require as many as 15 or 20 guesses each.) An average score for this sentence is about 200 guesses, or on the average, two guesses per letter. Let us assume, for the sake of further argument, that our subject has an identical twin who can be counted on always to make the same guesses under the same circumstances. If we encode right and wrong guesses of the subject as 0 and 1, the guesses made provide a string of bits equal in length to the number of guesses. The identical twin could be used, at some distant point, to decode the string and recover the original sentence, the 0's and 1's being used to tell him when his guesses are right. Thus Shannon's sentence could be coded using no more than 2 bits/letter. If this result is typical, we can, by Theorem I, assert that the entropy of English is no more than 2. The identical twins can be thought of as playing the role of complex computers in their encoding and decoding operations.

In the preceding example, we have limited the subject to guessing letters. We could do things differently and allow him to ask any yes-or-no questions he pleased. (E.g., "Are the first three letters '*the*'?", "Do the next ten letters form the word 'motorcycle'?") With this latter method, the results depend more strongly on the amount of practice the subject has had. An intelligent subject, experienced with other sentences, can do Shannon's sentence in about 150 questions. This gives a still better estimate of 1.5 as a maximum value for the entropy of English. Minimum values can also be calculated from letter guessing experiments or from other facts (such as the existence of crossword puzzles, see below). Present evidence indicates that the entropy of written English, in sentences of average length, lies between .5 and 1.5 bits/letter.

For a source with a k letter alphabet, let I stand for $-(p_1 \log_2 p_1 + \cdots + p_k \log_2 p_k)$, and let H be the true entropy. In Information Theory, the quantity $(I-H)/I$ is called the *redundance* of the source. One hundred per cent redundance means zero entropy and, by Theorem I, no output of information; there can be no significant variety in sequences of letters put out by the source. Zero redundance, on the other hand, as can be shown from consideration of the digram and trigram formulas, etc., means that almost any sequence of letters can occur in a legitimate message. This implies that in a language with low redundance, two and three dimensional crossword puzzles should be easy to devise. Now in English, two dimensional puzzles are fairly easy to invent, while three dimensional puzzles are extremely difficult. As mentioned above, this fact can be used to give estimates of redundance and hence entropy. For more detailed discussion of these matters, see Shannon's paper: "Prediction and Entropy of Printed English," *Bell System Technical Journal*, 1951, pp. 50–64.

The redundance of English and other languages is evident in our ability to reconstruct damaged texts and to solve cryptograms. Clearly redundancy can be useful. Indeed, if we turn back to Theorem II, and think of English sentences not as strings of letters but as abstract single units (like the units we select when we choose a pre-written birthday telegram), then a sentence written out in ordinary form can be viewed as a rather effective block coding of the original unit. When the amount of noise (e.g., typographical errors) is not too great, we have highly reliable communication.

Spoken language gives an even more direct illustration of Theorem II. We first note that speech is intelligible to the ear in the presence of relatively high background noise. It is a not uncommon experience to have the noise level in a room reach a rather distinct threshold where ordinary conversation becomes unintelligible. At this threshold, the high noise level has lowered channel capacity below the necessary minimum. When this happens, we do one or more of three things to continue communication, and all three accord with our theory. We speak more loudly (raise channel capacity), use a simpler vocabulary (lower entropy and therefore lower transmission rate), or speak more slowly (lower transmission rate). The last is made qualitatively plausible by noting that when we use a slower transmitted signal, we cause a variation in the noisy received signal which is less probable as noise, and therefore more detectable, than before. Even more convincingly in accord with the theory is the following not uncommon experience. Someone is speaking to us in the presence of considerable noise. He utters a long phrase. The phrase sounds unintelligible up to the last word or two. We hear the last words and suddenly the entire phrase takes form in our mind. The phrase acts like a block of code, and we do not decode until the entire block is received. It is tempting to speculate that the sound forms of speech are (in part) an error-correcting code developed in the course of evolutionary growth.

We conclude with several further comments on spoken English. These comments are in some measure speculative, and concern an area where both the choice of concepts and the interpretation of experimental evidence can be uncertain.

1) *Entropy.* To define entropy for a discrete source, we need only take account of the source itself. To define entropy for a continuous source, we need also take account of characteristics of the receiver. (This was indicated early in Part III above.) With a human receiver, the entropy of spoken English was estimated at from 20 to 50 bits/second. That the figure is as low as 20 to 50 bits/second appears to be due, not to inability of the receiving ear to discriminate, but rather to limited ability of the receiving central nervous system to process and analyse the signal received. (Here, definitions and experiments must be framed with care and subtlety. We do not discuss the matter in detail.) A variety of studies have suggested that 20 to 50 bits/second is a maximum rate at which the human organism can "genuinely" process information. If there are ways, involving higher rates, in which the human receiver responds to voice signals, then, when such ways are discovered, the entropy figure for spoken English can be revised upward.

Twenty bits/second, for ordinary spoken English, is equivalent to about 2 bits/letter of the corresponding written words, a figure that is higher than the figure of 1 bit/second for written English. The additional information includes what we are accustomed to calling "tone of voice," "emotion," etc.; it enables us, for instance, to distinguish one person's voice from another's.

In discussing *entropy* of spoken and of written English, we have been assuming that these sources are, in some sense, statistical. To what extent does language in fact possess the properties of a true, mathematically defined, statistical source? Considerable recent attention has been given to this question. It is not yet fully settled.

2) *Capacity.* Characteristics of the receiver are not taken into account when capacity of a continuous channel is defined. At the frequencies and amplitude of ordinary voice signals, the capacity of the acoustic (sound waves in air) channel in a relatively quiet room is from 200,000 to 300,000 bits/second. The capacity of the usual telephone channel is on the order of 20,000 bits/second. Several thousand bits/second appears to be a minimum figure for transmitting normal voice signals at normal speed over an acoustic or telephone-like channel without further coding. Below this figure, the received signals will not be understood by a human receiver. (When the continuous channel capacity formula is applied, this minimum involves a value of P/N, the *signal to noise ratio*, that is well below 1.)

3) *Coding.* In what ways can a voice signal be coded? As suggested before, one way of making a voice message more noise resistant is to transmit it more slowly. This is much like using the simple repetition code discussed at the beginning of Part II. The contrast between the entropy figure of 20 to 50 bits/second and the minimum voice channel capacity figure of several thousand bits/second suggests the study of more complex codings. Practical progress in such study has been slow. Promising recent work has involved a search for simple discrete ways of describing the physical production of voice signals in a human source (e.g., simple and useful ways of describing alterations in the shape of the human voice chamber). Simple discrete ways of describing the initial physical reception of voice signals in a human receiver are also being studied.

In principle, Theorem II tells us that, with sufficiently elaborate coding, understandable voice signals could be transmitted at normal speed over any channel (discrete or continuous) of capacity greater than 50 bits/second. Indeed, if our entropy figure is correct and if we have full knowledge of characteristics of the particular human receiver, we should be able to transmit over such a channel with very high apparent fidelity. Of course, problems of technology put such extreme coding achievements into a distant and speculative future; and striving for such goals might not ever be economically reasonable.

Final Comments on Theorem II. Theorem II is, in a sense, the reverse of Theorem I. Theorem I considered the extent to which messages can be compactly coded into bits, with a consequent elimination of redundancy and statistical structure in the coded message. Theorem II shows how statistical structure and redundance can be *introduced* to provide efficient error-correction.

Of course, the most remarkable feature of Theorem II is that it shows how, with high probability, errors in a block can be corrected *in full detail* without the necessity of dropping the transmission rate below a certain level. The use of longer and longer block length is correlated with the fact that it is easier *fully* to reconstruct a damaged text when a larger quantity of the damaged text is available for reconstruction. In cryptographic work, encipherment can be compared with damage or noise. Shannon began to suspect the truth of Theorem II from statistics accumulated in the course of cryptographic work. These statistics pointed to the existence of a critical length of enciphered text (in traditional secret codes) beyond which the deciphered 'solution' was unique.

In conclusion, we note that the relationship between error probability and block length among the good codes of Theorem II can be given as follows.

ϵ is approximately proportional to $e^{-\gamma N}$, where ϵ is maximum error probability per block, and N is block length. e is $2.718 \cdots$, the base of the natural logarithms, and γ depends on channel capacity C and transmission rate R. For R close to C, γ is approximately proportional to $(C-R)^2$.

Symbols, Signals and Noise, by J. R. Pierce (Harper, 1961) gives a popular exposition of Information Theory. It discusses a variety of applications and includes many interesting examples. A somewhat more technical survey ot the basic theory and its engineering implications will be found in *Transmission of Information*, by R. M. Fano (M.I.T., 1961). A recent bibliography of work done in Information Theory can be obtained from Bell Telephone Laboratories Publications Department, 463 West Street, New York City 14.

MATRICES IN TEACHING TRIGONOMETRY

A. R. AMIR-MOÉZ, University of Florida

In the September-October 1958 issue of this MAGAZINE in a brief article we described a use of vectors in teaching trigonometry [1]. In this note we would like to add to that a few ideas concerning the use of matrices in obtaining

In principle, Theorem II tells us that, with sufficiently elaborate coding, understandable voice signals could be transmitted at normal speed over any channel (discrete or continuous) of capacity greater than 50 bits/second. Indeed, if our entropy figure is correct and if we have full knowledge of characteristics of the particular human receiver, we should be able to transmit over such a channel with very high apparent fidelity. Of course, problems of technology put such extreme coding achievements into a distant and speculative future; and striving for such goals might not ever be economically reasonable.

Final Comments on Theorem II. Theorem II is, in a sense, the reverse of Theorem I. Theorem I considered the extent to which messages can be compactly coded into bits, with a consequent elimination of redundancy and statistical structure in the coded message. Theorem II shows how statistical structure and redundance can be *introduced* to provide efficient error-correction.

Of course, the most remarkable feature of Theorem II is that it shows how, with high probability, errors in a block can be corrected *in full detail* without the necessity of dropping the transmission rate below a certain level. The use of longer and longer block length is correlated with the fact that it is easier *fully* to reconstruct a damaged text when a larger quantity of the damaged text is available for reconstruction. In cryptographic work, encipherment can be compared with damage or noise. Shannon began to suspect the truth of Theorem II from statistics accumulated in the course of cryptographic work. These statistics pointed to the existence of a critical length of enciphered text (in traditional secret codes) beyond which the deciphered 'solution' was unique.

In conclusion, we note that the relationship between error probability and block length among the good codes of Theorem II can be given as follows.

ϵ is approximately proportional to $e^{-\gamma N}$, where ϵ is maximum error probability per block, and N is block length. e is $2.718 \cdots$, the base of the natural logarithms, and γ depends on channel capacity C and transmission rate R. For R close to C, γ is approximately proportional to $(C-R)^2$.

Symbols, Signals and Noise, by J. R. Pierce (Harper, 1961) gives a popular exposition of Information Theory. It discusses a variety of applications and includes many interesting examples. A somewhat more technical survey ot the basic theory and its engineering implications will be found in *Transmission of Information*, by R. M. Fano (M.I.T., 1961). A recent bibliography of work done in Information Theory can be obtained from Bell Telephone Laboratories Publications Department, 463 West Street, New York City 14.

MATRICES IN TEACHING TRIGONOMETRY

A. R. AMIR-MOÉZ, University of Florida

In the September-October 1958 issue of this MAGAZINE in a brief article we described a use of vectors in teaching trigonometry [1]. In this note we would like to add to that a few ideas concerning the use of matrices in obtaining

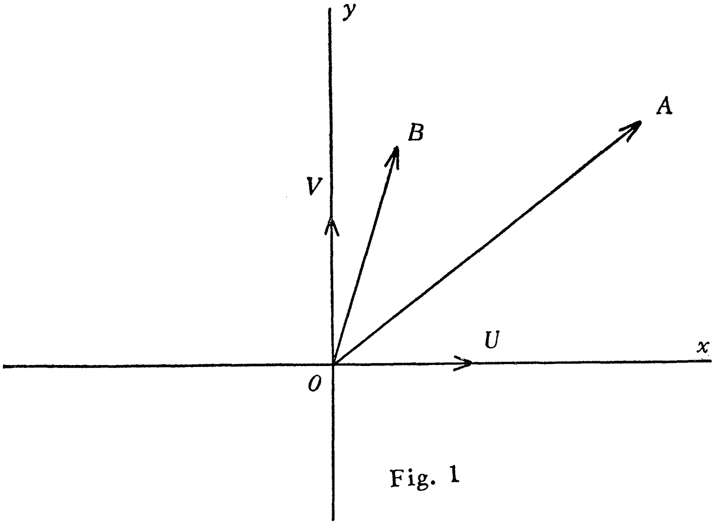


Fig. 1

trigonometric functions of $t+s$ in terms of the ones of t and s. In [1] a formula for cos $(t-s)$ was obtained using the inner product of vectors. Here we would like to obtain formulas for sin $(t+s)$ and cos $(t+s)$. The advantage of the use of matrices is that we can obtain the formulas in general without referring to complicated diagrams.

1. Linear transformations. A function f on the set of vectors of the plane into the set of vectors of the plane is called a linear transformation if:

$$\text{I.}\quad f(\mathrm{A}+\mathrm{B}) = f(\mathrm{A}) + f(\mathrm{B}),$$

$$\text{II.}\quad f(a\mathrm{A}) = af(\mathrm{A}),$$

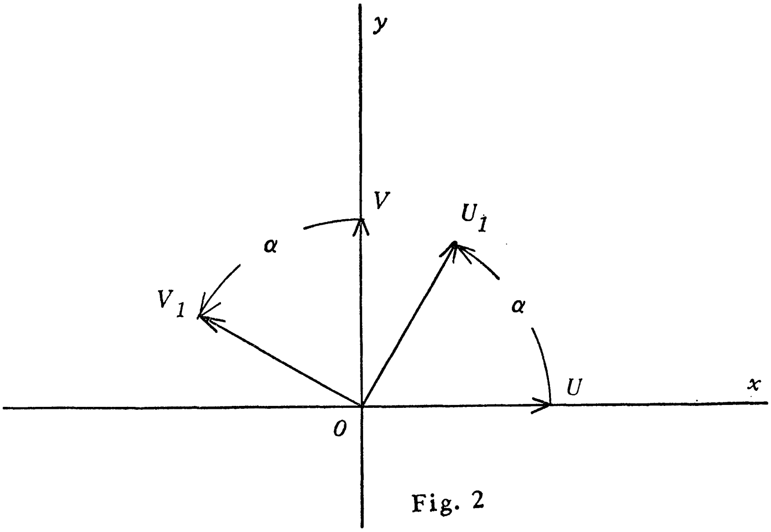


Fig. 2

where a is a real number and **A** and **B** are any two vectors. Indeed this idea may be studied in any book in linear algebra, for example [2].

2. The matrix of a linear transformation. Consider the unit vectors **U** on the x-axis and **V** on the y-axis (Fig. 1). Suppose $f(\mathbf{U})=\mathbf{A}$, $f(\mathbf{V})=\mathbf{B}$. If (a_{11}, a_{12}) is the set of coordinates of the point A and (a_{21}, a_{22}) is the set of coordinates of the point B, then the matrix of f is defined to be the array of numbers

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}.$$

3. Matrix of a rotation. For any vector **A** we define $\mathbf{B}=f(\mathbf{A})$ in such a way that $|\mathbf{A}|=|\mathbf{B}|$ and directed angle from **A** to **B** is α. Then we call f the rotation of the plane through an angle α.

To obtain the matrix of this rotation we again consider the unit vectors **U** and **V** respectively on the x and y axes (Fig. 2). Let $f(\mathbf{U})=\mathbf{U}_1$ and $f(\mathbf{V})=\mathbf{V}_1$. Then we observe that the set of coordinates of U_1 is $(\cos\alpha, \sin\alpha)$ and the set of coordinates of V_1 is $(-\sin\alpha, \cos\alpha)$. Thus the matrix of the rotation through an angle α is

$$\begin{pmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{pmatrix}.$$

4. Product of linear transformations. For two linear transformations f and g we define $f\cdot g$ as follows. Let $f(\mathbf{A})=\mathbf{B}$ and $g(\mathbf{B})=\mathbf{C}$. Then $(f\cdot g)(\mathbf{A})=g[f(\mathbf{A})]=\mathbf{C}$ (note the order).

Now for the matrix of $f\cdot g$ we suppose that the matrices of f and g are respectively

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \text{ and } \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}.$$

This means that

$$f(\mathbf{U})=a_{11}\mathbf{U}+a_{12}\mathbf{V},\quad f(\mathbf{V})=a_{21}\mathbf{U}+a_{22}\mathbf{V},\quad g(\mathbf{U})=b_{11}\mathbf{U}+b_{12}\mathbf{V},\quad g(\mathbf{V})=b_{21}\mathbf{U}+b_{22}\mathbf{V}.$$

Since the transformations f and g are linear we can write

$$\begin{aligned}(f\cdot g)(\mathbf{U}) &= g[f(\mathbf{U})] = a_{11}g(\mathbf{U}) + a_{12}g(\mathbf{V}) \\ &= a_{11}(b_{11}\mathbf{U} + b_{12}\mathbf{V}) + a_{12}(b_{21}\mathbf{U} + b_{22}\mathbf{V}) \\ &= (a_{11}b_{11} + a_{12}b_{21})\mathbf{U} + (a_{11}b_{12} + a_{12}b_{22})\mathbf{V}.\end{aligned}$$

That is, the set of coordinates of $(f\cdot g)$ (**U**) is

$$(a_{11}b_{11} + a_{12}b_{21},\ a_{11}b_{12} + a_{12}b_{22}).$$

Similarly the set of coordinates of $(f\cdot g)$ (**V**) can be obtained as

$$(a_{21}b_{11} + a_{22}b_{21},\ a_{21}b_{12} + a_{22}b_{22}).$$

Thus the matrix of $f\cdot g$ is

$$\begin{pmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{pmatrix}.$$

Thus we define this matrix to be the product of

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \text{ and } \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}.$$

5. Trigonometric functions of $t+s$. Consider the rotation f of the plane through an angle t. The matrix of this rotation is

$$\begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix}.$$

The matrix of the rotation g of the plane through an angle s will be

$$\begin{pmatrix} \cos s & \sin s \\ -\sin s & \cos s \end{pmatrix}.$$

The transformation $f \cdot g$ means the rotation of the plane through an angle t first and then the rotation of the plane through an angle s, i.e., the rotation of the plane through an angle $t+s$ whose matrix is

$$\begin{pmatrix} \cos (t+s) & \sin (t+s) \\ -\sin (t+s) & \cos (t+s) \end{pmatrix}.$$

On the other hand the matrix of $f \cdot g$ is

$$\begin{pmatrix} \cos t \cos s - \sin t \sin s & \cos t \sin s + \sin t \cos s \\ -\sin t \cos s - \cos t \sin s & -\sin t \sin s + \cos t \cos s \end{pmatrix}.$$

This implies that

$$\cos (t+s) = \cos t \cos s - \sin t \sin s$$

and

$$\sin (t+s) = \cos t \sin s + \sin t \cos s.$$

Other formulas can be obtained from these.

References

1. Ali R. Amir-Moéz, Teaching trigonometry through vectors, this MAGAZINE, 31 (1958) 19–23.

2. A. R. Amir-Moéz, and A. L. Fass, Elements of linear spaces, Pergamon Press, 1962, pp. 16, 67, 123.

CLASSROOM NOTE ON $e^{i\alpha}$

J. PRITCHETT, Rutherford College of Technology

An introductory lesson on the relationship between $\cos\alpha+i\sin\alpha$ and $e^{i\alpha}$ to a class who have just been shown De Moivre's Theorem:

"We have already shown that

$$(\cos\alpha + i\sin\alpha)(\cos\beta + i\sin\beta) = \cos(\alpha+\beta) + i\sin(\alpha+\beta)$$

and

$$(\cos\alpha + i\sin\alpha)^n = \cos n\alpha + i\sin n\alpha.$$

You will notice that if we write $\cos\alpha+i\sin\alpha=f(\alpha)$ that $\cos\alpha+i\sin\alpha$ has the property

$$(f(\alpha))^n = f(n\alpha) \quad \text{and} \quad f(\alpha)\times f(\beta) = f(\alpha+\beta).$$

Can anyone think of any other function which obeys this law?"

With a little bit of luck, someone mentions indices.

"So that we have, if

$$F(\alpha) = a^\alpha$$

then $F(\alpha)^n = F(n\alpha)$ and $F(\alpha)\times F(\beta) = F(\alpha+\beta)$.

Is it a coincidence that $\cos\alpha+i\sin\alpha$ and a^α obey the same laws, or are they the same function?

Let us examine $\cos\alpha+i\sin\alpha$ a little further.

If we differentiate it we get $i(\cos\alpha+i\sin\alpha)$, i.e., apart from the factor i, it appears to be unchanged by differentiation.

Can anyone remember any other function with this property?"

Again someone will remember e^x or even e^{ax}.

"If we differentiate $e^{i\alpha}$ we get $ie^{i\alpha}$.

It appears very likely in fact that $\cos\alpha+i\sin\alpha$ and $e^{i\alpha}$ are the same function."

SOLUTIONS

S58. $m=n=1$ and $m=n=3$ are obvious solutions of both equations. For any value of n other than 1 or 3 the terminal digit of $\sum$ is clearly 3, whence $\sum$ cannot be a square.

S59. $\phi(x, y) = |x-1| + |x+1| + |y-1| + |y+1| - 4 = 0$. This set consists of all points in and on the square with vertices at $(\pm 1, \pm 1)$.

S60. In the United States, he should walk in a direction such that the uptown tracks are kept on his left. Presumably, in London, it would be in the opposite direction. That is, if the trains run the same way as the automobile.

(Trickies on page 126)

ANSWERS

A331. If $y=D-1$ we may write $x^2+(x+1)^2=D^2$, a familiar Pythagorean relation where D is restricted to the odd denominators of the convergents to $\sqrt{2}$. Then $x=(\pm\sqrt{\{2D^2-1\}}-1)/2$.

Then the first five solutions in positive integers are x, $y=3$, 4; 20, 28; 119, 168; 696, 984; 4059, 5740; $\cdots$.

A332.

$$\begin{array}{cccccc}
x^5 & -\ x^4 & -\ \ x^3 & & & \\
 & +\ x^4 & -\ \ x^3 & -\ \ x^2 & & \\
 & & +\ 2x^3 & -\ 2x^2 & -\ 2x & \\
 & & & -\ 2x^2 & +\ 2x & +\ 2 \\
\hline
x^5 & & & -\ 5x^2 & & +\ 2
\end{array}$$

so $x^5-5x^2+2=(x^2-x-1)(x^3+x^2+2x-2)$.

A333. All Pythagorean triplets can be formed from $a=m^2-n^2$, $b=2mn$, $c=m^2+n^2$ for $m>n\geq 1$. Since the differences of consecutive squares exhaust all odd numbers ≥ 3, and since, for $n=1$, $2m$ exhaust all even numbers ≥ 4, any integer ≥ 3 can be part of a triplet. The values 1 and 2 obviously fail. In particular given a odd, $b=(a^2-1)/2$, $c=(a^2+1)/2$; given a even, $b=a^2/4-1$, $c=a^2/4+1$.

A334. The ratio $n/2$ follows immediately from

$$\sum_{r=0}^{n} r\binom{n}{r}^p = \sum_{r=0}^{n} (n-r)\binom{n}{r}^p \text{ or from}$$

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

A335. Since $\phi^n(x)=0$, this implies that $\phi^{n-1}(x)=1$, 2. (Then $n=1$ implies $x=1$ or 2.) But $\phi^{n-1}(x)=1$, 2 implies that $\phi^{n-2}(x)=3$, 4, or 6, but only 3 is acceptable. (Then $n=2$ implies $x=3$.) Finally $\phi^{n-2}(x)=3$ has no solution for $\phi^{n-3}(x)$; then $x=1$, 2, or 3.

(Questions on page 126)

AN APPLICATION OF CONTINUANTS

S. L. BASIN, Sylvania Electronic Systems, Mountain View, California

1. Introduction. The following note on continuants (defined below) is intended to demonstrate their usefulness in studying recurrent sequences. In particular, the correspondence between a class of *simple continuants* and the well known Fibonacci sequence is demonstrated. Several continuant identities, discovered between the years 1853 and 1880, listed by T. Muir [1] are given together with their corresponding Fibonacci identities.

2. Definition. A continuant of order n is defined as an nth order determinant all of whose elements are zero except those along the principal diagonal and minor diagonals, i.e., along the two adjacent diagonal lines parallel to and on either side of the principal diagonal.

The nth order continuant,

$$(1)\qquad \begin{vmatrix} a_1 & b_1 & & & & \\ c_1 & a_2 & b_2 & & 0 & \\ & c_2 & a_3 & b_3 & & \\ & & c_3 & a_4 & b_4 & \\ & & & & \ddots & \\ & 0 & & & & \cdot\, b_{n-1} \\ & & & & c_{n-1} & a_n \end{vmatrix}$$

is usually denoted by,

$$(2)\qquad K\begin{bmatrix} & b_1 & b_2 \cdots b_{n-1} & \\ a_1 & a_2 & a_3 \cdots a_n & \\ & c_1 & c_2 \cdots c_{n-1} & \end{bmatrix} \quad\text{or}\quad K(1,\, n).$$

If in $K(1, n)$, $b_1=b_2=b_3= \cdots =b_{n-1}=1$ and $c_1=c_2=c_3= \cdots =c_{n-1}=-1$ the determinant is called a *simple continuant* which we denote by $K(1, n)_1$, or $K(a_1a_2a_3 \cdots a_n)$ when the latter notation is not ambiguous.

3. Recurrent Sequences. Given any second order homogeneous difference equation with constant coefficients,

$$(3)\qquad u_n = a_{n-1}u_{n-1} + b_{n-2}u_{n-2}$$

we may easily obtain an expression for u_n in terms of the initial values u_0 and u_1 by use of continuants. As an example, consider $n=6$, for which we obtain,

$$(4)\qquad \begin{aligned} -u_6 + a_5u_5 + b_4u_4 \qquad\qquad\qquad\qquad\qquad\qquad &= 0\\ -u_5 + a_4u_4 + b_3u_3 \qquad\qquad\qquad\qquad &= 0\\ -u_4 + a_3u_3 + b_2u_2 \qquad\qquad &= 0\\ -u_3 + a_2u_2 + b_1u_1 \qquad &= 0\\ -u_2 + a_1u_1 + b_0u_0 &= 0. \end{aligned}$$

It follows directly from (4) that u_6 is expressed by the continuant,

$$(5)\qquad u_6 = \begin{vmatrix} a_5 & b_4 & & & \\ -1 & a_4 & b_3 & & \\ & -1 & a_3 & b_2 & \\ & & -1 & a_2 & b_1u_1 \\ & & & -1 & (a_1u_1 + b_0u_0) \end{vmatrix}.$$

The desired result follows from (5) and from Theorem 1.

THEOREM 1 ([1]). *If in any continuant $K\ (1, n)$ the element $a_j = a_j' + a_j''$ then $K(1, n)$ may be written as*

$$K\left\{\begin{matrix} & b_1 & \cdots & b_{j-1} & \\ a_1 & a_2\cdots & a_{j-1} & a_j' & \\ & c_1 & \cdots & c_{j-1} & \end{matrix}\right\}\cdot K(j+1, n) + K(1, j-1)\cdot K\left\{\begin{matrix} & b_j \cdots b_{n-1} & \\ a_j'' & \cdots & a_n \\ & c_j \cdots c_{n-1} & \end{matrix}\right\}.$$

Theorem 1 allows (5) to be rewritten as,

$$u_6 = K\left\{\begin{matrix} & b_4 & & b_3 & & b_2 & & b_1 & \\ a_5 & & a_4 & & a_3 & & a_2 & & a_1 \\ & -1 & & -1 & & -1 & & -1 & \end{matrix}\right\} u_1 + K\left\{\begin{matrix} & b_4 & & b_3 & & b_2 & \\ a_5 & & a_4 & & a_3 & & a_2 \\ & -1 & & -1 & & -1 & \end{matrix}\right\} u_0$$

which is the desired result. If the continuant,

$$K_n = \left|\begin{matrix} a & c & \cdot & \cdot & & & \\ b & a & c & \cdot & & & \\ \cdot & b & a & c & & & \\ \cdot & \cdot & \cdot & \cdot & \cdot & & \\ & & & \cdot & \cdot & \cdot & \\ & & & & & & \\ & & & \cdot & \cdot & \cdot & \end{matrix}\right|_n$$

is expanded down column 1 and the minor of b (row 2, column 1) is again expanded down column 1, the continuant may be expressed by the second order difference equation,

$$K_n = aK_{n-1} - bcK_{n-2}. \tag{6}$$

Since the characteristic equation of (6) is $x^2 - ax^2 + bc = 0$ with roots α and β we have,

$$K_n = \frac{\alpha^{n+1} - \beta^{n+1}}{2^{n+1}(\alpha - \beta)} = \frac{(a + \sqrt{\{a^2 - 4bc\}})^{n+1} - (a - \sqrt{\{a^2 - 4bc\}})^{n+1}}{2^{n+1}\sqrt{\{a^2 - 4bc\}}}. \tag{7}$$

If we let $a = 1$ and $b = -c = \pm 1$ then (7) clearly establishes,

THEOREM 2. *The simple continuant,*

$$K(1, n)_1 = K\left\{\begin{matrix} & -1 & -1 \cdots & -1 & \\ 1, & 1, & 1, & \cdots & 1 \\ & 1 & 1 \cdots & 1 & \end{matrix}\right\} = K\left\{\begin{matrix} & i & i \cdots i & \\ 1, & 1, & 1, \cdots & 1 \\ & i & i \cdots i & \end{matrix}\right\} = F_{n+1}$$

where F_{n+1} is the $(n+1)$th Fibonacci number, and $i = \sqrt{-1}$.

Proof. The proof follows from (7) since (7) becomes the Binet formula when $a = 1$ and $b = -c = \pm 1$ or $b = c = i$ in the characteristic equation of (6). An alternate proof follows from

Lemma 1. *The number of terms in the expansion of a continuant is given by the three equivalent expressions,*

(i) $$\frac{1}{2^{n+1}\sqrt{5}}\left[(1+\sqrt{5})^{n+1}-(1-\sqrt{5})^{n+1}\right]$$

(ii) $$\frac{1}{2^n}\left[\binom{n+1}{1}+5\binom{n+1}{3}+5^2\binom{n+1}{5}+\cdots\right]$$

(iii) $$\left[\binom{n}{0}+\binom{n-1}{1}+\binom{n-2}{2}+\cdots+\binom{n-j}{j}+\cdots\right].$$

Proof. Each term in the expansion of $K(1, n)$ consists of n factors. The first term consists of the product of all n principal diagonal elements. The next terms consist of all possible permutations obtained by replacing two successive principal diagonal elements by a bc pair, then replacing four successive principal diagonal elements by two bc pairs, . . . , etc. The process is continued until r bc pairs replace $n-2r$ successive principal diagonal elements; this process yields all the permutations of $n-r$ objects, $n-2r$ of which are alike and r are alike since we consider each bc pair as a single entity, i.e.,

$$\frac{(n-r)!}{(n-2r)!\,r!}=\binom{n-r}{r}.$$

Since the expansion process is started by replacing zero principal diagonal elements, then one pair of successive principal diagonal elements, then two pair, three pair, . . . etc. the total number of terms in the expansion is given by

$$\left[\binom{n}{0}+\binom{n-1}{1}+\binom{n-2}{2}+\cdots+\binom{n-j}{j}+\cdots\right]$$

which is equivalent to (i) and (ii) in Lemma 1. The above expansion process is probably better understood by considering the expansion of

(8) $$\begin{vmatrix} a & b & 0 & 0 & 0 \\ c & a & b & 0 & 0 \\ 0 & c & a & b & 0 \\ 0 & 0 & c & a & b \\ 0 & 0 & 0 & c & a \end{vmatrix} \begin{aligned} &= a\cdot a\cdot a\cdot a\cdot a - a\cdot a\cdot a\cdot \underline{bc} - a\cdot a\cdot \underline{bc}\cdot a - a\cdot \underline{bc}\cdot a\cdot a \\ &\quad - \underline{bc}\cdot a\cdot a\cdot a + a\cdot \underline{bc}\cdot \underline{bc} + \underline{bc}\cdot a\cdot \underline{bc} + \underline{bc}\cdot \underline{bc}\cdot a. \end{aligned}$$

If each term is forced to be unity by requiring $a=1$, $b=-c=\pm 1$, or $b=c=i$, then the value of $K(1, n)_1$ is clearly F_{n+1}.

4. Identities. The following Fibonacci identities result from well-known continuant identities [1] and Theorem 2:

(i) $K(1,n)_1 = K(1,r)_1\cdot K(r+1,n)_1 + K(1,r-1)_1\cdot K(r+2,n)_1$

(Continuant identity)

$F_{n+1}=F_{r+1}\cdot F_{n-r+1}+F_r\cdot F_{n-r}$ (Fibonacci identity).

(ii) $$K(1,n)_1 = \{K(1,h)_1\cdot K(h+1,r)_1 + K(1,h-1)_1\cdot K(h+2,r)_1\}\cdot K(r+1,n)_1 + \{K(1,h)_1\cdot K(h+1,r-1)_1 + K(1,h-1)_1\cdot K(h+2,r-1)_1\}\cdot K(r+2,n)_1.$$

$$F_{n+1} = \{F_{h+1}\cdot F_{r-h+1} + F_hF_{r-h}\}F_{n-r+1} + \{F_{h+1}\cdot F_{r-h} + F_hF_{r-h-1}\}F_{n-r}.$$

(iii) $$K(1,n)_1 = a_1\cdot K(2,n) + K(3,n)_1 \text{ or}$$
$$K(1,n)_1 = a_nK(1,n-1)_1 + K(1,n-2)_1$$
$$F_{n+1} = F_n + F_{n-1}.$$

(iv) $$K(1,n)_1\cdot K(h,r)_1 = K(1,r)_1\cdot K(h,n)_1 + (-1)^{r-h+1}K(1,h-2)_1\cdot K(r+2,n)_1$$
$$F_{n+1}F_{r-h+2} = F_{r+1}\cdot F_{n-h+2} + (-1)^{r-h+1}F_{h-1}\cdot F_{n-r}.$$

(v) $$K(1,n)_1\cdot K(2,n-1)_1 - K(1,n-1)_1\cdot K(2,n)_1 + (-1)^{n-1} = 0$$
$$F_{n+1}\cdot F_{n-1} - F_n\cdot F_n + (-1)^{n-1} = 0 \Rightarrow F_{n+1}F_{n-1} - F_n^2 = (-1)^n.$$

(vi) $$K(1,n)_1\cdot K(2,n-2)_1 - K(1,n-2)_1\cdot K(2,n)_1 + (-1)^n = 0$$
$$F_{n+1}\cdot F_{n-2} - F_{n-1}\cdot F_n + (-1)^n = 0.$$

(vii) If $K(1,n)$ is factored we have,

$$K(1,n) = a_1\left(a_2 - \frac{b_2c_2}{a_1}\right)\left(a_3 - \frac{b_3c_3a_1}{K(1,2)}\right)\left(a_4 - \frac{b_4c_4K(1,2)}{K(1,3)}\right)\cdots \left(a_n - \frac{b_nc_nK(1,n-2)}{K(1,n-1)}\right).$$

This factorization results in the identity, $F_{n+1} = \prod_{i=1}^{n}\left(1 + \frac{F_{i-1}}{F_i}\right)$.

(viii) The continuant identity in (v) may be rewritten as

$$\left\{\frac{K(1,n)_1}{K(2,n)_1} - \frac{K(1,n-1)_1}{K(2,n-1)_1}\right\} = \frac{(-1)^n}{K(2,n)_1\cdot K(2,n-1)_1}.$$

Substituting this into the identity,

$$\frac{K(1,n)_1}{K(2,n)_1} = \frac{K(1,1)}{K(2,1)} + \left\{\frac{K(1,2)_1}{K(2,2)_1} - \frac{K(1,1)_1}{K(2,1)_1}\right\} + \left\{\frac{K(1,3)_1}{K(2,3)_1} - \frac{K(1,2)_1}{K(2,2)_1}\right\} + \left\{\frac{K(1,4)_1}{K(2,4)_1} - \frac{K(1,3)_1}{K(2,3)_1}\right\} + \cdots + \left\{\frac{K(1,n)_1}{K(2,n)_1} - \frac{K(1,n-1)_1}{K(2,n-1)_1}\right\}$$

results in

$$\frac{K(1,n)_1}{K(2,n)_1} = a_1 + \frac{1}{K(2,2)_1K(2,1)_1} - \frac{1}{K(2,3)_1K(2,2)_1} + \frac{1}{K(2,4)_1\cdot K(2,3)_1} - \frac{1}{K(2,5)_1\cdot K(2,4)_1} + \frac{1}{K(2,6)_1K(2,5)_1} - \cdots + \frac{(-1)^n}{K(2,n)_1K(2,n-1)_1}.$$

The corresponding Fibonacci identity is

$$\frac{F_{n+1}}{F_n} = 1 + \sum_{i=1}^{n} \frac{(-1)^i}{F_i F_{i-1}} \quad \text{and} \quad \frac{1+\sqrt{5}}{2} = 1 + \sum_{i=1}^{\infty} \frac{(-1)^i}{F_i F_{i-1}} \cdot$$

(ix) The simple continuants $K(1, n)_1$ and $K(2, n)_1$ are prime to each other, provided $a_1, a_2, \cdots, a_n$ are integers, which establishes that two successive Fibonacci numbers are relatively prime, i.e. $(F_n, F_{n+1}) = 1$.

(x) The continuant identity, (iii) $K(1, n)_1 = a_1 K(2, n)_1 + K(3, n)_1$ may be rewritten as the continued fraction,

$$\frac{K(1,n)_1}{K(2,n)_1} = a_1 + \cfrac{1}{\cfrac{K(2,n)_1}{K(3,n)_1}} = a_1 + \cfrac{1}{a_2 + \cfrac{K(4,n)_1}{K(3,n)_1}} = a_1 + \cfrac{1}{a_2 + \cfrac{1}{\cfrac{K(3,n)_1}{K(4,n)_1}}}$$

$$= a_1 + \frac{1}{a_2+}\;\frac{1}{a_3+}\;\cdots\;\frac{1}{+\,a_n},$$

from which

$$\frac{F_{n+1}}{F_n} = 1 + \frac{1}{1+}\;\frac{1}{1+}\;\cdots\;\frac{1}{+1} \cdot$$

Two additional identities are given when considering odd order centrosymmetric continuants, i.e.,

$$\text{(xi)}\quad a_m K \left\{ \begin{matrix} b_1\, b_2 \cdots b_{m-1}\, b_{m-1} \cdots b_1 \\ a_1\;\, a_2 \cdots a_m \quad a_{m-1} \cdots a_1 \\ c_1\, c_2 \cdots c_{m-1}\; c_{m-1} \cdots c_1 \end{matrix} \right\}_{2m-1} = [K(1,m)]^2 - [b_{m-1}c_{m-1}K(1, m-2)]^2 \cdot$$

The corresponding Fibonacci identity is $F_{2m} = F_{m+1}^2 - F_{m-1}^2$, and

$$\text{(xii)}\quad K \left\{ \begin{matrix} b_1 \cdots b_{m-1},\, c_{m-1} \cdots c_1 \\ a_1 \cdots \qquad a_m \cdots \qquad a_1 \\ c_1 \cdots c_{m-1},\, b_{m-1} \cdots b_1 \end{matrix} \right\}_{2m-1} = K(1, m-1)[K(1,m) - b_{m-1}c_{m-1}K(1, m-2)].$$

The corresponding identity is given by,

$$F_{2m} = F_m[F_{m+1} + F_{m-1}] = F_m L_m,$$

where L_m is the mth Lucas number.

5. Generating Functions and Chebyshev Polynomials.

LEMMA 2. *If* $1/(acx^2 - bx + 1) = K_0 + K_1 x + K_2 x^2 + K_3 x^3 + \cdots$ *then*

$$K_n = K \left\{ \begin{matrix} a,\; a, \cdots a \\ b,\; b,\; b, \cdots, b \\ c,\; c, \cdots c \end{matrix} \right\} .$$

Proof. Direct expansion of the generating function gives the desired result. Since the generating function for the Chebyshev polynomials of the second kind, $U_n\ (x)$ is

$$\frac{1}{u^2 - 2xu + 1} = U_0 + U_1(x)u + U_2(x)u^2 + \cdots,$$

let $ac=1$ and $b=2x$ in Lemma 2; e.g., we may take $a=c=1$,′ therefore,

$$\tag{9} \begin{vmatrix} 2x & 1 & & 0 \\ 1 & 2x & 1 & \\ & 1 & 2x & 1 \\ 0 & & & \ddots \end{vmatrix}_n = U_n(x).$$

The Chebyshev polynomials $S_n(x)$ may be generated by the successive principal coaxal determinants of

$$\tag{10} \begin{vmatrix} x & 1 & & 0 \\ 1 & x & 1 & \\ & 1 & x & 1 \\ 0 & & & \ddots \end{vmatrix}_n = S_n(x)\ \cdot$$

The closed form of these Chebyshev polynomials is easily derived from (7), i.e.,

$$\tag{11} U_n(x) = \frac{1}{2\sqrt{\{x^2-1\}}}\left\{(x+\sqrt{\{x^2-1\}})^{n+1} - (x-\sqrt{\{x^2-1\}})^{n+1}\right\}$$

and

$$\tag{12} S_n(x) = \frac{1}{\sqrt{\{x^2-4\}}}\left\{\left(\frac{x+\sqrt{\{x^2-4\}}}{2}\right)^{n+1} - \left(\frac{x-\sqrt{\{x^2-4\}}}{2}\right)^{n+1}\right\}.$$

The relationship between Fibonacci numbers, Chebyshev polynomials, and hyperbolic functions is clearly derived from (11), since

$$U_n\left(\frac{1}{2i}\right) = (i)^n \frac{1}{\sqrt{5}}\left\{\left(\frac{1+\sqrt{5}}{2}\right)^{n+1} - \left(\frac{1-\sqrt{5}}{2}\right)^{n+1}\right\}.$$

Therefore,

$$\tag{13} F_{n+1} = (i)^{-n}U_n\left(\frac{1}{2i}\right) \quad \text{where} \quad i = \sqrt{-1}.$$

Computing $U_n(\frac{3}{2})$ from (11) we have,

$$U_n\left(\frac{3}{2}\right) = \frac{1}{\sqrt{5}}\left\{\left(\frac{3+\sqrt{5}}{2}\right)^{n+1} - \left(\frac{3-\sqrt{5}}{2}\right)^{n+1}\right\}.$$

Observing that

$$\left(\frac{3+\sqrt{5}}{2}\right)=\left(\frac{1+\sqrt{5}}{2}\right)^2, \qquad \left(\frac{3-\sqrt{5}}{2}\right)=\left(\frac{1-\sqrt{5}}{2}\right)^2 \text{and}$$

$$U_n\left(\frac{3}{2}\right)=\frac{1}{\sqrt{5}}\left\{\left(\frac{1+\sqrt{5}}{2}\right)^{2(n+1)}-\left(\frac{1-\sqrt{5}}{2}\right)^{2(n+1)}\right\}$$

we obtain,

$$F_{2n+2}=U_n\left(\frac{3}{2}\right).$$

From (12) we observe that,

$$F_{2n+2}=S_n(3)$$

and

$$F_{n+1}=(i)^{-n}S_n(-i), \quad \text{where} \quad i=\sqrt{-1}.$$

From (13) we derive

$$F_{n+1}=(-1)^{n/2}\left\{\frac{\sinh\left[\ (n+1)\ \text{arc}\cosh\left(\frac{1}{2i}\right)\right]}{\sinh\left[\text{arc}\cosh\left(\frac{1}{2i}\right)\right]}\right\}.$$

6. Fibonacci Polynomials. The polynomials $f_n(x)=xf_{n-1}(x)+f_{n-2}(x)$ where $f_0(x)=0; f_1(x)=1$, investigated by Jacobsthal [2] and Basin [3] are generated by the successive principal coaxal determinants of

$$\text{(13)} \qquad \begin{vmatrix} x & 1 & & 0 & \\ -1 & x & 1 & & \\ & -1 & x & 1 & \\ 0 & & & \ddots & \end{vmatrix}_n .$$

The closed form given by (7) is

$$\text{(14)} \quad f_n(x)=\frac{1}{\sqrt{\{x^2+4\}}}\left\{\left(\frac{x+\sqrt{\{x^2+4\}}}{2}\right)^{n+1}-\left(\frac{x-\sqrt{\{x^2+4\}}}{2}\right)^{n+1}\right\}.$$

The polynomials defined in [3] by,

$$\text{(15)} \qquad \begin{aligned} b_0(x)&=1, \qquad B_0(x)=1\\ b_n(x)&=xB_{n-1}(x)+b_{n-1}(x) \qquad (n\geqq 1)\\ B_n(x)&=(x+1)B_{n+1}(x)+b_{n-1}(x) \qquad (n\geqq 1)\end{aligned}$$

may be generated by the successive principal coaxal determinants of,

$$\mathcal{B}_n = \begin{vmatrix} 1 & 1 & & & \\ -1 & x & 1 & & 0 \\ & -1 & 1 & 1 & \\ & & -1 & x & 1 \\ 0 & & & & \ddots \end{vmatrix}_n . \tag{16}$$

Expanding (16) as in the proof of Lemma 1 we obtain

$$\mathcal{B}_n = \sum_{j=0}^{[n/2]} \binom{n-j}{j} x^{[n/2]-j} \tag{17}$$

where $[n/2]$ is the greatest integer less than or equal to $n/2$. The closed form of $b_n(x)$ and $B_n(x)$ may be obtained by observing that, $b_n(x) = \mathcal{B}_{2n}$ and $B_n(x) = \mathcal{B}_{2n+1}$. We hope that the above examples have demonstrated some of the useful applications of continuants.

References

1. Thomas Muir, A Treatise on the theory of determinants, Dover, 1960 edition.

2. E. Jacobsthal, Fibonaccische Polynome und Kreisteilungsgleichungen. Sitzungsberichte der Berliner Math. Gesellschaft, **17** (1919–20) pp. 43–57.

3. S. L. Basin, The Appearance of Fibonacci numbers and the Q matrix in electrical network theory, this MAGAZINE, **36**, No. 2, March, 1963.

BRACKET FUNCTION CONGRUENCES FOR BINOMIAL COEFFICIENTS

L. CARLITZ, Duke University and
H. W. GOULD, West Virginia University

Among the many congruences involving binomial coefficients are some which involve also the bracket function, sometimes called the "greatest integer function."

One of the more interesting of these was the subject of problem 4704 in the *American Mathematical Monthly* [1] where it was shown that a necessary and sufficient condition for p to be prime is that for every integer n

$$\binom{n}{p} \equiv \left[\frac{n}{p}\right] \pmod{p}. \tag{1}$$

The fact that the congruence is true when p is a prime was pointed out in problem 4322 in the *Monthly* [3] and, in fact, the theorem traces to perhaps the time of Wolstenholme.

may be generated by the successive principal coaxal determinants of,

$$\mathcal{B}_n = \begin{vmatrix} 1 & 1 & & & 0 \\ -1 & x & 1 & & \\ & -1 & 1 & 1 & \\ & & -1 & x & 1 \\ 0 & & & & \ddots \end{vmatrix}_n . \tag{16}$$

Expanding (16) as in the proof of Lemma 1 we obtain

$$\mathcal{B}_n = \sum_{j=0}^{[n/2]} \binom{n-j}{j} x^{[n/2]-j} \tag{17}$$

where $[n/2]$ is the greatest integer less than or equal to $n/2$. The closed form of $b_n(x)$ and $B_n(x)$ may be obtained by observing that, $b_n(x) = \mathcal{B}_{2n}$ and $B_n(x) = \mathcal{B}_{2n+1}$. We hope that the above examples have demonstrated some of the useful applications of continuants.

References

1. Thomas Muir, A Treatise on the theory of determinants, Dover, 1960 edition.
2. E. Jacobsthal, Fibonaccische Polynome und Kreisteilungsgleichungen. Sitzungsberichte der Berliner Math. Gesellschaft, **17** (1919–20) pp. 43–57.
3. S. L. Basin, The Appearance of Fibonacci numbers and the Q matrix in electrical network theory, this MAGAZINE, **36**, No. 2, March, 1963.

BRACKET FUNCTION CONGRUENCES FOR BINOMIAL COEFFICIENTS

L. CARLITZ, Duke University and
H. W. GOULD, West Virginia University

Among the many congruences involving binomial coefficients are some which involve also the bracket function, sometimes called the "greatest integer function."

One of the more interesting of these was the subject of problem 4704 in the *American Mathematical Monthly* [1] where it was shown that a necessary and sufficient condition for p to be prime is that for every integer n

$$\binom{n}{p} \equiv \left[\frac{n}{p}\right] \pmod{p}. \tag{1}$$

The fact that the congruence is true when p is a prime was pointed out in problem 4322 in the *Monthly* [3] and, in fact, the theorem traces to perhaps the time of Wolstenholme.

In problem 4322, P. A. Piza raised the question about (1) when p is replaced by an arbitrary power of a prime. The congruence (1) follows from an elegant theorem of Lucas as was pointed out by Fine in his solution to 4322.

We should like to point out the following congruence: Put

$$k = b_0 + 2b_1 + 4b_2 + \cdots + 2^{m-1}b_{m-1} \qquad (0 \leqq b_s \leqq 1).$$

Then

$$\binom{n}{2^m - k} \equiv \sum_j \left[\frac{n+j}{2^m}\right] \pmod 2, \tag{2}$$

where j ranges over the set of nonnegative integers defined by $j = c_0 + 2c_1 + \cdots + 2^{m-1}c_{m-1}$ with $0 \leqq c_s \leqq b_s$. This follows by an induction on k.

An examination of Dickson's History of the Theory of Numbers and a number of other texts failed to turn up this congruence. It may be of interest to list a few special cases of the congruence (all below are modulo 2 of course):

$$\binom{n}{1} \equiv \left[\frac{n}{1}\right], \quad \binom{n}{2} \equiv \left[\frac{n}{2}\right],$$

$$\binom{n}{3} \equiv \left[\frac{n}{4}\right] + \left[\frac{n+1}{4}\right], \quad \binom{n}{4} \equiv \left[\frac{n}{4}\right],$$

$$\binom{n}{5} \equiv \left[\frac{n}{8}\right] + \left[\frac{n+1}{8}\right] + \left[\frac{n+2}{8}\right] + \left[\frac{n+3}{8}\right],$$

$$\binom{n}{6} \equiv \left[\frac{n}{8}\right] + \left[\frac{n+2}{8}\right], \quad \binom{n}{7} \equiv \left[\frac{n}{8}\right] + \left[\frac{n+1}{8}\right],$$

$$\binom{n}{8} \equiv \left[\frac{n}{8}\right], \quad \binom{n}{9} \equiv \left[\frac{n}{16}\right] + \sum_{j=1}^{7} \left[\frac{n+j}{16}\right].$$

For an arbitrary prime p it is easy to show that

$$\binom{n}{p^m - k} \equiv \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} \left[\frac{n+j}{p^m}\right] \pmod p \quad (0 \leqq k \leqq p^m). \tag{3}$$

Indeed this can be proved by induction on k as follows:

$$\begin{aligned}
\binom{n}{p^m - k - 1} &= \binom{n+1}{p^m - k} - \binom{n}{p^m - k} \\
&\equiv \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} \left[\frac{n+j+1}{p^m}\right] - \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} \left[\frac{n+j}{p^m}\right] \\
&\equiv \sum_j (-1)^{k+1-j} \binom{k}{j-1} \left[\frac{n+j}{p^m}\right] + \sum_j (-1)^{k+1-j} \binom{k}{j} \left[\frac{n+j}{p^m}\right] \\
&\equiv \sum_j (-1)^{k+1-j} \binom{k+1}{j} \left[\frac{n+j}{p^m}\right].
\end{aligned}$$

If we put $k = b_0 + pb_1 + \cdots + p^{m-1}b_{m-1} (0 \leqq b_s < p)$ and

$$j = c_0 + pc_1 + \cdots + p^{m-1}c_{m-1} \qquad (0 \leqq c_s < p)$$

then by a familiar theorem of Lucas [2]

$$\binom{k}{j} \equiv \binom{b_0}{c_0}\binom{b_1}{c_1} \cdots \binom{b_{m-1}}{c_{m-1}} \pmod{p}.$$

Thus it is only necessary to retain those j for which $0 \leqq c_s \leqq b_s$. Consequently (3) becomes

$$\binom{n}{p^m - k} \equiv \sum_{j=0}^{k} (-1)^{k-j} \binom{b_0}{c_0}\binom{b_1}{c_1} \cdots \binom{b_{m-1}}{c_{m-1}} \left[\frac{n+j}{p^m}\right] \pmod{p} \tag{4}$$

and for the case $p = 2$ this reduces to (2).

We remark that the very interesting theorem of Lucas has been used by Roberts [4] to obtain some rather curious congruences for binomial coefficients.

References

1. L. E. Clarke, Problem 4704, Amer. Math. Monthly, **63** (1956) 584. Solution: **64** (1957) 597–598.

2. L. E. Dickson, History of the theory of numbers, vol. 1, Chapter 9, Washington, 1919, p. 271.

3. P. A. Piza, Problem 4322, Amer. Math. Monthly, **55** (1948) 642. Solution: **57** (1950) 347–348.

4. J. B. Roberts, On binomial residues, Canad. J. Math., **9** (1957) 363–370.

INNER PRODUCTS OF MULTILINEAR VECTORS

H. RANDOLPH PYLE, Whittier College

Introduction. When a student has learned to use vectors in dealing with the geometry of 3-space, he is ready to use them in the geometry of n-space if he can do so by generalizing the concepts with which he is familiar. This can be done by resorting to tensor analysis, but the intricacies of that notation are not easy to master. There are many advantages in using the vocabulary and the geometrical approach of elementary vector analysis. The algebra of multilinear vectors as developed by Grassmann, Cartan, Bourbaki and other writers does this.

A *bivector* (*area vector*) or 2-vector has two independent constituent line vectors and is interpreted as the parallelogram of which the line vectors are sides. The *magnitude* of the bivector is the area of the parallelogram, and its *orientation* is the orientation of the plane in which the parallelogram lies. The bivector is called the *outer product* of the constituent vectors. Two bivectors are equal when they represent parallelograms with the same area and orientation, and the directions of the constituent vectors are in the same sense.

If we put $k=b_0+pb_1+\cdots+p^{m-1}b_{m-1}(0\leqq b_s<p)$ and

$$j = c_0 + pc_1 + \cdots + p^{m-1}c_{m-1} \qquad (0 \leqq c_s < p)$$

then by a familiar theorem of Lucas [2]

$$\binom{k}{j} \equiv \binom{b_0}{c_0}\binom{b_1}{c_1}\cdots\binom{b_{m-1}}{c_{m-1}} \pmod{p}.$$

Thus it is only necessary to retain those j for which $0\leqq c_s \leqq b_s$. Consequently (3) becomes

$$(4) \qquad \binom{n}{p^m - k} \equiv \sum_{j=0}^{k} (-1)^{k-j}\binom{b_0}{c_0}\binom{b_1}{c_1}\cdots\binom{b_{m-1}}{c_{m-1}}\left[\frac{n+j}{p^m}\right] \pmod{p}$$

and for the case $p=2$ this reduces to (2).

We remark that the very interesting theorem of Lucas has been used by Roberts [4] to obtain some rather curious congruences for binomial coefficients.

References

1. L. E. Clarke, Problem 4704, Amer. Math. Monthly, **63** (1956) 584. Solution: **64** (1957) 597–598.
2. L. E. Dickson, History of the theory of numbers, vol. 1, Chapter 9, Washington, 1919, p. 271.
3. P. A. Piza, Problem 4322, Amer. Math. Monthly, **55** (1948) 642. Solution: **57** (1950) 347–348.
4. J. B. Roberts, On binomial residues, Canad. J. Math., **9** (1957) 363–370.

INNER PRODUCTS OF MULTILINEAR VECTORS

H. RANDOLPH PYLE, Whittier College

Introduction. When a student has learned to use vectors in dealing with the geometry of 3-space, he is ready to use them in the geometry of n-space if he can do so by generalizing the concepts with which he is familiar. This can be done by resorting to tensor analysis, but the intricacies of that notation are not easy to master. There are many advantages in using the vocabulary and the geometrical approach of elementary vector analysis. The algebra of multilinear vectors as developed by Grassmann, Cartan, Bourbaki and other writers does this.

A *bivector* (*area vector*) or 2-vector has two independent constituent line vectors and is interpreted as the parallelogram of which the line vectors are sides. The *magnitude* of the bivector is the area of the parallelogram, and its *orientation* is the orientation of the plane in which the parallelogram lies. The bivector is called the *outer product* of the constituent vectors. Two bivectors are equal when they represent parallelograms with the same area and orientation, and the directions of the constituent vectors are in the same sense.

In 3-space the cross product provides the information we want concerning magnitude and orientation by making use of the normal to the plane, but in spaces of higher dimension a plane has no unique normal and it is necessary to focus attention on the parallelogram itself.

In like manner a trivector (volume vector) or 3-vector is represented by a parallelepiped in n-space, and an r-vector by a parallelotope of corresponding dimensions.

We shall define and develop formulas for the projection of an s-vector on the space spanned by an r-vector $(s \leqq r)$, and for the inner product of two such vectors. This product is a scalar when $s=r$, and a vector of dimension $r-s$ when $s<r$. The vector is in the r-space and it is completely orthogonal to the s-vector. Its magnitude is the product of the magnitudes of the two vectors and the cosine of the angle between them.

Notation. Our discussion will be limited to vectors in a real vector space $V_n(R)$ with an orthogonal coordinate system. An r-vector whose independent constituent line vectors are $A_1, A_2, \cdots, A_r$ will be designated by $A_{(r)}$ (the subscript in parentheses indicates the order of the multiple vector). The outer product is written with brackets, $A_{(r)}=[A_1A_2\cdots A_r]$. An alternative notation particularly useful for two constituent vectors is $A_1\Box A_2$, where the symbol is called "box" and the product is named the "box product." Geometrically this is a most appropriate name. In matrix form

$$A_{(r)} = \begin{pmatrix} A_1 \\ A_2 \\ \cdot \\ \cdot \\ \cdot \\ A_r \end{pmatrix}$$

and its magnitude is found by taking the determinant of the matrix product $A_{(r)}A_{(r)}{}^t$. If we call the magnitude of $A_{(r)}$, D, we have

$$D^2 = \left| A_{(r)}A^t_{(r)} \right| = \begin{vmatrix} A_1\cdot A_1 & A_1\cdot A_2 \cdots A_1\cdot A_r \\ A_2\cdot A_1 & A_2\cdot A_2 \cdots A_2\cdot A_r \\ \cdot\ \cdot\ \cdot\ \cdot & \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \\ A_r\cdot A_1 & A_r\cdot A_2 \cdots A_r\cdot A_r \end{vmatrix}.$$

If $A_{(s)}$, $(s\leq r)$, contains $A_{i_1}, A_{i_2}, \cdots, A_{i_s}$ (constituent vectors of $A_{(r)}$) we shall call it a *subvector* of $A_{(r)}$, and write it $A_{(s)}=[A_{i_1}A_{i_2}\cdots A_{i_s}]=A_{i_1i_2\cdots i_s}$. The subvector of dimension $r-s$ containing the constituent vectors of $A_{(r)}$ not in $A_{(s)}$, with the proper sign is called the *vector complement* of $A_{i_1i_2\cdots i_s}$, and is designated by $\overline{A}_{i_1i_2\cdots i_s}$. The sign is such that $A_{i_1\cdots i_s}\Box\overline{A}_{i_1\cdots i_s}=A_{(r)}$. It can be determined by the number of permutations required to put the subscripts of the vectors in natural order. For instance,

$$\overline{A}_i = (-1)^{i+1}[A_1\cdots A_{i-1}A_{i+1}\cdots A_r].$$

Let A_i^*, $(i=1,\cdots,r)$ be the set of line vectors reciprocal to the set A_i with respect to the r-space spanned by $A_{(r)}$. We have shown in a previous paper [3] that, if α_{ij} is the reduced cofactor of $A_i\cdot A_j$ in the determinant D^2, then $A_i^* = \sum_{j=1}^r \alpha_{ij}A_j$. Now $A_{(r)}^* = [A_1^*A_2^* \cdots A_r^*]$ is the r-vector reciprocal to $A_{(r)}$.

Inner product of two vectors of the same order. Let $A_{(r)} = [A_1A_2 \cdots A_r]$, $B_{(r)} = [B_1B_2 \cdots B_r]$ be two r-vectors. In each the constituent vectors are independent, so that it spans an r-space. The inner product of $A_{(r)}$ and $B_{(r)}$ is defined as

$$A_{(r)}\cdot B_{(r)} = \left| A_{(r)} \right| \left| B_{(r)} \right| \cos\theta = \begin{vmatrix} A_1\cdot B_1 & A_1\cdot B_2 \cdots A_1\cdot B_r \\ A_2\cdot B_1 & A_2\cdot B_2 \cdots A_2\cdot B_r \\ \cdot\;\cdot\;\cdot\;\cdot & \cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot \\ A_r\cdot B_1 & A_r\cdot B_2 \cdots A_r\cdot B_r \end{vmatrix}.$$

This relation serves to define cosine θ when θ is the angle between the two r-flats.

Now we can write $D^2 = |A_{(r)}|^2 = A_{(r)}\cdot A_{(r)}$, and $A_{(r)}\cdot A_{(r)}^* = 1$.

Projection of an s-vector on an r-flat. In a previous paper [3] we have shown that the projection of a line vector B_i on the r-flat spanned by $A_{(r)}$ is $P_i = \sum_{j=1}^r (B_i\cdot A_j)A_j^*$. Let $B_{(s)} = [B_1B_2 \cdots B_s]$, $(s \leqq r)$, be the vector to be projected on $A_{(r)}$.

DEFINITION. *The projection $P_{(s)}$ of $B_{(s)}$ on the r-flat is the s-vector whose constituent vectors are the projections of the constituent vectors of $B_{(s)}$ on the flat. The angle between $B_{(s)}$ and $A_{(r)}$ is the angle between $B_{(s)}$ and $P_{(s)}$.*

Matrix form of the projection. The projection can be written in matrix form as

$$P_{(s)} = \begin{pmatrix} P_1 \\ P_2 \\ \cdot \\ \cdot \\ P_s \end{pmatrix} = \begin{pmatrix} B_1\cdot A_1 & B_1\cdot A_2 \cdots B_1\cdot A_r \\ B_2\cdot A_1 & B_2\cdot A_2 \cdots B_2\cdot A_r \\ \cdot\;\cdot\;\cdot\;\cdot & \cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot \\ B_s\cdot A_1 & B_s\cdot A_2 \cdots B_s\cdot A_r \end{pmatrix} \begin{pmatrix} A_1^* \\ A_2^* \\ \cdot \\ \cdot \\ A_r^* \end{pmatrix}.$$

When $s=1$, $r=1$,

$$P = (B\cdot A)A^* = \frac{(B\cdot A)A}{|A|^2},$$

which is the usual form of the projection of the line vector B on the line determined by the vector A. When $s=r$, $P_{(r)} = (B_{(r)}\cdot A_{(r)})A_{(r)}$ and $P_{(r)}\cdot A_{(r)}^* = B_{(r)}\cdot A_{(r)}$.

Outer product form of the projection. An alternative form of $P_{(s)}$ is obtained by taking the outer product $P_{(s)} = [P_1P_2 \cdots P_s]$. Since

$$P_k = \sum_{i_1\cdots i_s=1}^{r} (B_k\cdot A_{i_k})A_{i_k}^*,$$

this product becomes

$$P_{(s)} = \sum_{i_1\cdots i_s=1}^{r} \{(B_1\cdot A_{i_1})(B_2\cdot A_{i_2}) \cdots (B_s\cdot A_{i_s})\}[A_{i_1}^*A_{i_2}^* \cdots A_{i_s}^*].$$

As $i_1, \cdots, i_s$ range over the set of values $1, \cdots, r$ there will be $s!$ arrangements for any given set of numbers $j_1, \cdots, j_s$. If we permute $[A^*_{i_1}A^*_{i_2} \cdots A^*_{i_s}]$ so as to get $[A^*_{j_1}A^*_{j_2} \cdots A^*_{j_s}]$, $(j_1 < j_2 < \cdots < j_s)$, there will be changes of sign as the permutations are made. If we affix these signs to the products $(B_1 \cdot A_{i_1})(B_2 \cdot A_{i_2}) \cdots (B_s \cdot A_{i_s})$ we obtain the determinant

$$B_{(s)} \cdot A_{j_1 \cdots j_s} = \begin{vmatrix} B_1 \cdot A_{j_1} & B_1 \cdot A_{j_2} \cdots B_1 \cdot A_{j_s} \\ B_2 \cdot A_{j_1} & B_2 \cdot A_{j_2} \cdots B_2 \cdot A_{j_s} \\ \cdot \; \cdot \; \cdot \; \cdot \; \cdot & \cdot \; \cdot \; \cdot \; \cdot \; \cdot \; \cdot \; \cdot \; \cdot \\ B_s \cdot A_{j_1} & B_s \cdot A_{j_2} \cdots B_s \cdot A_{j_s} \end{vmatrix},$$

so that the outer product form is

$$P_{(s)} = \sum_{j_1 \cdots j_s = 1}^{r} (B_{(s)} \cdot A_{j_1 \cdots j_s}) A^*_{j_1 \cdots j_s}, \qquad (j_1 < j_2 < \cdots < j_s).$$

From this form we have

$$B_{(s)} \cdot P_{(s)} = \sum_{j_1 \cdots j_s = 1}^{r} (B_{(s)} \cdot A_{j_1 \cdots j_s})(B_{(s)} \cdot A^*_{j_1 \cdots j_s}).$$

Let

$$M = \begin{bmatrix} B_1 \cdot A_1 & B_1 \cdot A_2 \cdots B_1 \cdot A_r \\ B_2 \cdot A_1 & B_2 \cdot A_2 \cdots B_2 \cdot A_r \\ \cdot \; \cdot \; \cdot \; \cdot \; \cdot & \cdot \; \cdot \; \cdot \; \cdot \; \cdot \; \cdot \; \cdot \\ B_s \cdot A_1 & B_s \cdot A_2 \cdots B_s \cdot A_r \end{bmatrix}$$

and $(M^*)^t$ be the transpose of the corresponding matrix when A^*_i is substituted for A_i in M. The summation of $B_{(s)} \cdot P_{(s)}$ is equal to the determinant of the matrix product of M and $(M^*)^t$ [5], i.e., $B_{(s)} \cdot P_{(s)} = |M(M^*)^t|$. When the A_i are orthonormal vectors, $M = M^*$. Then $B_{(s)} \cdot P_{(s)} = |MM^t|$.

Since

$$P_i = \sum_{=1}^{r} (B_i \cdot A_j) A^*_j, \qquad P_i \cdot A_k = \sum_{=1}^{r} (B_i \cdot A_j)(A^*_j \cdot A_k).$$

But $A^*_j \cdot A_k = \delta_{jk}$ (Kronecker delta), so that $P_i \cdot A_k = B_i \cdot A_k$. Substituting in M and M^* shows that $B_{(s)} \cdot P_{(s)} = P_{(s)} \cdot P_{(s)}$.

Angle between $B_{(s)}$ and $A_{(r)}$. Cartan [1] has shown that multiple vectors can be added and subtracted, so that from $B_{(s)} \cdot P_{(s)} = P_{(s)} \cdot P_{(s)}$ we have $(B_{(s)} - P_{(s)}) \cdot P_{(s)} = 0$. This means that $Q_{(s)} = B_{(s)} - P_{(s)}$ is perpendicular to $P_{(s)}$. If we designate $Q_{(s)} \cdot Q_{(s)}$ by $Q^2_{(s)}$, multiplication gives $Q^2_{(s)} = B^2_{(s)} - 2(B_{(s)} \cdot P_{(s)}) + P^2_{(s)} = B^2_{(s)} - P^2_{(s)}$, and $B^2_{(s)} = P^2_{(s)} + Q^2_{(s)}$. Consequently $|P_{(s)}| \leqq |B_{(s)}|$. We have defined the angle θ between $B_{(s)}$ and $A_{(r)}$ as the angle between $B_{(s)}$ and $P_{(s)}$, so that $B_{(s)} \cdot P_{(s)} = |B_{(s)}| \, |P_{(s)}| \cos\theta$. Since $B_{(s)} \cdot P_{(s)} = |P_{(s)}|^2$, $|P_{(s)}| = |B_{(s)}| \cos\theta$, and $\cos\theta = |M(M^*)^t|^{1/2}/|B_{(s)}|$.

The inner product of $B_{(s)}$ and $A_{(r)}$. The inverse of the determinant for D^2 is

$$(D^2)^{-1} = \begin{vmatrix} \alpha_{11} & \alpha_{21} & \cdots & \alpha_{r1} \\ \alpha_{12} & \alpha_{22} & \cdots & \alpha_{r2} \\ \cdot & \cdot & \cdot & \cdot \\ \alpha_{1r} & \alpha_{2r} & \cdots & \alpha_{rr} \end{vmatrix}$$

where the α_{ij} are the reduced cofactors of $A_i \cdot A_j$ in D^2. Since

$$A_i^* = \sum_{j=1}^{r} \alpha_{ij} A_j,$$

we have

$$\overset{*}{A}_{i_1 \cdots i_s} = [A_{i_1}^* A_{i_2}^* \cdots A_{i_s}^*] = \sum_{j_1 \cdots j_s = 1}^{r} (\alpha_{i_1 j_1} \alpha_{i_2 j_2} \cdots \alpha_{i_s j_s})[A_{j_1} A_{j_2} \cdots A_{j_s}].$$

If we select a given set of numbers $k_1, k_2, \cdots, k_s$ and consider all the combinations of $j_1, j_2, \cdots, j_s$ in the summation and repeat the argument already used, we have

$$\overset{*}{A}_{i_1 \cdots i_s} = \sum_{k_1 \cdots k_s = 1}^{r} \beta_{i_1 \cdots i_s, k_1 \cdots k_s} A_{k_1 \cdots k_s},$$

where $\beta_{i_1 \cdots i_s, k_1 \cdots k_s}$ is the minor of $(D^2)^{-1}$ involving rows $i_1, \cdots, i_s$ and columns $k_1, \cdots, k_s$. By [4] this minor in $(D^2)^{-1}$ equals the reduced cofactor of the minor in D^2 with corresponding rows and columns. If we refer to our definition of vector complements, we recall that the complement of $A_{i_1 \cdots i_s}$ is $\overline{A}_{i_1 \cdots i_s}$ (an $(r-s)$-vector containing the constituent vectors of $A_{(r)}$ not in $A_{i_1 \cdots i_s}$ with the proper sign attached). The reduced cofactor of D^2 in question is $\alpha_{i_1 \cdots i_s, k_1 \cdots k_s}$ which can be written in the form

$$\alpha_{i_1 \cdots i_s, k_1 \cdots k_s} = \frac{\overline{A}_{i_1 \cdots i_s} \cdot \overline{A}_{k_1 \cdots k_s}}{D^2}.$$

Now

$$\overset{*}{A}_{i_1 \cdots i_s} = \frac{1}{D^2} \sum_{k_1 \cdots k_s = 1}^{r} (\overline{A}_{i_1 \cdots i_s} \cdot A_{k_1 \cdots k_s}) A_{k_1 \cdots k_s}.$$

This reduces to the form for A_i^* when $s = 1$.

Now

$$P_{(s)} = \sum_{i_1 \cdots i_s = 1}^{r} (B_{(s)} \cdot A_{i_1 \cdots i_s}) \overset{*}{A}_{i_1 \cdots i_s}$$

becomes

$$P_{(s)} = \frac{1}{D^2} \sum_{i_1 \cdots i_s} (B_{(s)} \cdot A_{i_1 \cdots i_s})(\overline{A}_{i_1 \cdots i_s} \cdot \overline{A}_{k_1 \cdots k_s}) A_{k_1 \cdots k}$$

and

$$B_{(s)}\cdot P_{(s)} = \frac{1}{D^2}\sum_{k_1\cdots k_s}\sum_{i_1\cdots i_s}(B_{(s)}\cdot A_{i_1\cdots i_s})(B_{(s)}\cdot A_{k_1\cdots k_s})(\overline{A}_{i_1\cdots i_s}\cdot\overline{A}_{k_1\cdots k_s})$$
$$= \left\{\frac{1}{D}\sum_{i_1\cdots i_s=1}^{r}(B_{(s)}\cdot A_{i_1\cdots i_s})\overline{A}_{i_1\cdots i_s}\right\}^2.$$

The quantity inside the braces is an $(r-s)$-vector in the r-flat. We shall call it $N_{(r-s)}/D$. Now $|N_{(r-s)}/D|^2=|P_{(s)}|^2=|B_{(s)}|^2\cos^2\theta$ and since $D=|A_{(r)}|$, $|N_{(r-s)}|=|A_{(r)}|\ |B_{(s)}|\cos\theta$.

DEFINITION. *The inner product of the multiple vectors $B_{(s)}$ and $A_{(r)}$ $(s\leq r)$ is an $(r-s)$-vector in $A_{(r)}$ completely orthogonal to $B_{(s)}$, whose magnitude is $|B_{(s)}|\ |A_{(r)}|$ $\cos\theta$. This vector is*

$$B_{(s)}\cdot A_{(r)} = N_{(r-s)} = \sum_{i_1\cdots i_s=1}^{r}(B_{(s)}\cdot A_{i_1\cdots i_s})\overline{A}_{i_1\cdots i_s}.$$

This reduces to the vector given by Cartan [1] when $s=1$, $r=2$.

A convenient way to write the inner product is

$$B_{(s)}\cdot A_{(r)} = \left|\begin{array}{cccc} B_1\cdot A_1 & B_1\cdot A_2 & \cdots & B_1\cdot A_r \\ B_2\cdot A_1 & B_2\cdot A_2 & \cdots & B_2\cdot A_r \\ \cdot & \cdot & \cdots & \cdot \\ B_s\cdot A_1 & B_s\cdot A_2 & \cdots & B_s\cdot A_r \\ \hline A_1 & A_2 & \cdots & A_r \end{array}\right|$$

which is to be interpreted as meaning that with each s-rowed minor above the line is to be associated the $(r-s)$-vector whose constituent vectors are those in the last line which are in the columns not involved in the s-rowed minor. The sign is determined as one would determine the sign of the cofactor of the s-rowed minor in an r-rowed determinant.

When $r=s+1$, we can write the product as an orthodox determinant with the last row consisting of vectors. For instance when $s=1$, $r=2$,

$$B\cdot A_{12} = \begin{vmatrix} B\cdot A_1 & B\cdot A_2 \\ A_1 & A_2 \end{vmatrix} = (B\cdot A_1)A_2 - (B\cdot A_2)A_1.$$

Thus $B\cdot A_{12}$ is a vector in the plane spanned by A_{12} perpendicular to B.

When $s=2$, $r=3$, we may choose $A_{(3)}=[ijk]$, where i, j, k are the unit vectors on the coordinate axes of the 3-space. Now

$$B_{(2)}\cdot A_{(3)} = \begin{vmatrix} B_1\cdot i & B_1\cdot j & B_1\cdot k \\ B_2\cdot i & B_2\cdot j & B_2\cdot k \\ i & j & k \end{vmatrix}.$$

This is the usual form of the cross product, $B_1\times B_2$, so that the classical cross product is the inner product of an area vector and a volume vector in 3-space.

An orthogonal set of vectors. This process can be used to find a set of mutually orthogonal vectors spanning an r-space which is determined by the vectors $A_1, A_2, \cdots, A_r$. When we take the inner product of the s-vector $A_{(s)} = [A_1 \cdots A_s]$, and the $(s+1)$-vector $A_{(s+1)} = [A_1 \cdots A_s A_{s+1}]$, we have a vector in the $(s+1)$-space spanned by $A_{(s+1)}$, which is completely orthogonal to $A_{(s)}$. (The orthogonality is easily verified when we use the determinant form of the product.) Beginning with $s=1$, we can build up the required set of vectors one by one.

Let

$$V_1 = A_1, \qquad V_2 = A_1 \cdot A_{12} = \begin{vmatrix} A_1 \cdot A_1 & A_1 \cdot A_2 \\ A_1 & A_2 \end{vmatrix}$$

$$V_3 = A_{12} \cdot A_{123} = \begin{vmatrix} A_1 \cdot A_1 & A_1 \cdot A_2 & A_1 \cdot A_3 \\ A_2 \cdot A_1 & A_2 \cdot A_2 & A_2 \cdot A_3 \\ A_1 & A_2 & A_3 \end{vmatrix}, \text{ and so on.}$$

$V_1, V_2, V_3, \cdots$ are mutually orthogonal line vectors. This method has the advantage in numerical calculation of requiring no division, with its attendant fractions. The Gram-Schmidt process involves division.

Proof that $B_{s+1} \cdot (B_{(s)} \cdot A_{(r)}) = B_{(s+1)} \cdot A_{(r)}$. Since $N_{(r-s)} = B_{(s)} \cdot A_{(r)}$, we have

$$B_{s+1} \cdot (B_{(s)} \cdot A_{(r)}) = B_{s+1} \cdot N_{(r-s)} = \sum_{i_1 \cdots i_s = 1}^{r} (B_{(s)} \cdot A_{i_1 \cdots i_s})(B_{s+1} \cdot \overline{A}_{i_1 \cdots i_s})$$

But $\overline{A}_{i_1 \cdots i_s}$ is an $(r-s)$-vector, so that $B_{s+1} \cdot \overline{A}_{i_1 \cdots i_s}$ is an $(r-s-1)$-vector. Let $\overline{A}_{i_1 \cdots i_s}|_{i_{s+1}}$ represent the $(r-s-1)$-vector complementary to $A_{i_{s+1}}$ in $\overline{A}_{i_1 \cdots i_s}$. The sign of $\overline{A}_{i_1 \cdots i_s}|_{i_{s+1}}$ is obtained by removing $A_{i_1 \cdots i_s}$ from $A_{(r)}$ first and then removing $A_{i_{s+1}}$ from the subvector $\overline{A}_{i_1 \cdots i_s}$, with the proper sign attached at each step. Now

$$B_{s+1} \cdot N_{(r-s)} = \sum_{i_1 \cdots i_s} (B_{(s)} \cdot A_{i_1 \cdots i_s}) \sum_{i_{s+1}} (B_{s+1} \cdot A_{i_{s+1}}) \overline{A}_{i_1 \cdots i_s}|_{i_{s+1}}.$$

As $i_1, \cdots, i_s, i_{s+1}$ vary over a given set of $s+1$ integers such as $1, 2, \cdots, s+1$, $\overline{A}_{i_1 \cdots i_s}|_{i_{s+1}}$ changes only in sign. When we make use of this sign,

$$\sum_{i_1 \cdots i_s} \sum_{i_{s+1}} (B_{(s)} \cdot A_{i_1 \cdots i_s})(B_{s+1} \cdot A_{i_{s+1}}) = \begin{vmatrix} B_1 \cdot A_{i_1} & B_1 \cdot A_{i_2} & \cdots & B_1 \cdot A_{i_{s+1}} \\ B_2 \cdot A_{i_1} & B_2 \cdot A_{i_2} & \cdots & B_2 \cdot A_{i_{s+1}} \\ \cdot & \cdot & \cdots & \cdot \\ B_s \cdot A_{i_1} & B_s \cdot A_{i_2} & \cdots & B_s \cdot A_{i_{s+1}} \\ B_{s+1} \cdot A_{i_1} & B_{s+1} \cdot A_{i_2} & \cdots & B_{s+1} \cdot A_{i_{s+1}} \end{vmatrix}.$$

But this determinant is $B_{(s+1)} \cdot A_{i_1 \cdots i_s, i_{s+1}}$ so that

$$B_{s+1} \cdot N_{(r-s)} = \sum_{i_1 \cdots i_{s+1}} (B_{(s+1)} \cdot A_{i_1 \cdots i_{s+1}}) \overline{A}_{i_1 \cdots i_{s+1}} = B_{(s+1)} \cdot A_{(r)}$$

as was to be proved.

Proof that $N_{(r-s)}$ is completely orthogonal to $B_{(s)}$. If $B_{s+1}=B_i$, $(i=1,\cdots,s)$, $B_{(s+1)}\cdot A_{i_1\cdots i_{s+1}}=0$ because two rows of the determinant are identical. This shows that $N_{(r-s)}$ is orthogonal to every constituent vector in $B_{(s)}$, and so is completely orthogonal to $B_{(s)}$. Since $P_i\cdot A_k=B_i\cdot A_k$, we get the same result when we take $P_i\cdot N_{(r-s)}$. $B_{s+1}\cdot N_{(r-s)}$ is a vector in $N_{(r-s)}$ perpendicular to B_{s+1}, and $B_{(s+1)}\cdot A_{(r)}$ is a vector in $A_{(r)}$ completely orthogonal to $B_{(s+1)}$. The fact that these expressions are equivalent shows that the two vectors are identical.

$B_{(r-s)}\cdot N_{(r-s)}=B_{(r)}\cdot A_{(r)}$. If we take $r-s$ arbitrary vectors $B_{s+1}, B_{s+2},\cdots, B_r$ and call $B_{(r-s)}=[B_{s+1}B_{s+2}\cdots B_r]$ so that $B_{(s)}\square B_{(r-s)}=B_{(r)}$, we have

$$B_{(r-s)}\cdot N_{(r-s)} = \sum_{i_1\cdots i_s} (B_{(s)}\cdot A_{i_1\cdots i_s})(B_{(r-s)}\cdot \bar{A}_{i_1\cdots i_s}) = B_{(r)}\cdot A_{(r)}.$$

When $s=1$, $r=2$, this reduces to a result obtained by Cartan.

Since $B_{(r-s)}$ is arbitrary and $B_{(r)}\cdot A_{(r)}$ is independent of the particular vectors used to determine the r-vectors, $N_{(r-s)}$ must be independent of the vectors used.

References

1. E. Cartan, Leçons sur la Géométrie des Espaces de Riemann, Gauthier-Villars et Cie, Paris, 1928.

2. Bourbaki, Éléments de Mathématique, Algèbre Multilinéaire, Hermann, Paris, 1958.

3. H. R. Pyle, The projection of a vector on a plane, this MAGAZINE, March-April, 1961.

4. Muir and Metzler, Theory of determinants. Privately published, Albany, N. Y., 1930, p. 168.

5. Thrall and Tornheim, Vector spaces and matrices, Wiley, New York, 1957. p. 127.

TWO PARAMETRIC REPRESENTATIONS OF A CIRCLE IN 3-DIMENSIONAL EUCLIDEAN SPACE

LEO UNGER, Litton Industries, Inc.

Part 1. A curve in Euclidean space is representable, at least theoretically, by a single parameter. Elementary curves, such as conic sections, are usually defined in three dimensional space as the intersection of a plane with a right circular cone whose implicit equations are given. Some of the following paragraphs utilize a similar definition to derive an explicit equation of a circle in 3-dimensional space in terms of a single parameter. The derivation is somewhat specialized in the sense that a unit sphere with its center at the origin is assumed. This restricts the radius and center of the circle. The first restriction is trivial and the second can easily be removed. To do so now would only obscure some aspects of the equations.

Two points on the sphere determine a family of circles. In order to single out a particular one a third point, or an additional equivalent restraint, must be given since three points determine a circle uniquely. For example, a great circle through $\mathbf{r}$ and $\mathbf{r}'$ is uniquely determined since its center is coincident with the center of the sphere. The circle is also uniquely defined when one of its points

Proof that $N_{(r-s)}$ is completely orthogonal to $B_{(s)}$. If $B_{s+1}=B_i$, $(i=1, \cdots, s)$, $B_{(s+1)}\cdot A_{i_1\cdots i_{s+1}}=0$ because two rows of the determinant are identical. This shows that $N_{(r-s)}$ is orthogonal to every constituent vector in $B_{(s)}$, and so is completely orthogonal to $B_{(s)}$. Since $P_i\cdot A_k=B_i\cdot A_k$, we get the same result when we take $P_i\cdot N_{(r-s)}$. $B_{s+1}\cdot N_{(r-s)}$ is a vector in $N_{(r-s)}$ perpendicular to B_{s+1}, and $B_{(s+1)}\cdot A_{(r)}$ is a vector in $A_{(r)}$ completely orthogonal to $B_{(s+1)}$. The fact that these expressions are equivalent shows that the two vectors are identical.

$B_{(r-s)}\cdot N_{(r-s)}=B_{(r)}\cdot A_{(r)}$. If we take $r-s$ arbitrary vectors $B_{s+1}, B_{s+2}, \cdots, B_r$ and call $B_{(r-s)}=[B_{s+1}B_{s+2}\cdots B_r]$ so that $B_{(s)}\Box B_{(r-s)}=B_{(r)}$, we have

$$B_{(r-s)}\cdot N_{(r-s)} = \sum_{i_1\cdots i_s} (B_{(s)}\cdot A_{i_1\cdots i_s})(B_{(r-s)}\cdot \overline{A}_{i_1\cdots i_s}) = B_{(r)}\cdot A_{(r)}.$$

When $s=1$, $r=2$, this reduces to a result obtained by Cartan.

Since $B_{(r-s)}$ is arbitrary and $B_{(r)}\cdot A_{(r)}$ is independent of the particular vectors used to determine the r-vectors, $N_{(r-s)}$ must be independent of the vectors used.

References

1. E. Cartan, Leçons sur la Géométrie des Espaces de Riemann, Gauthier-Villars et Cie, Paris, 1928.
2. Bourbaki, Éléments de Mathématique, Algèbre Multilinéaire, Hermann, Paris, 1958.
3. H. R. Pyle, The projection of a vector on a plane, this MAGAZINE, March-April, 1961.
4. Muir and Metzler, Theory of determinants. Privately published, Albany, N. Y., 1930, p. 168.
5. Thrall and Tornheim, Vector spaces and matrices, Wiley, New York, 1957. p. 127.

TWO PARAMETRIC REPRESENTATIONS OF A CIRCLE IN 3-DIMENSIONAL EUCLIDEAN SPACE

LEO UNGER, Litton Industries, Inc.

Part 1. A curve in Euclidean space is representable, at least theoretically, by a single parameter. Elementary curves, such as conic sections, are usually defined in three dimensional space as the intersection of a plane with a right circular cone whose implicit equations are given. Some of the following paragraphs utilize a similar definition to derive an explicit equation of a circle in 3-dimensional space in terms of a single parameter. The derivation is somewhat specialized in the sense that a unit sphere with its center at the origin is assumed. This restricts the radius and center of the circle. The first restriction is trivial and the second can easily be removed. To do so now would only obscure some aspects of the equations.

Two points on the sphere determine a family of circles. In order to single out a particular one a third point, or an additional equivalent restraint, must be given since three points determine a circle uniquely. For example, a great circle through $\mathbf{r}$ and $\mathbf{r}'$ is uniquely determined since its center is coincident with the center of the sphere. The circle is also uniquely defined when one of its points and the normal of the plane in which it lies are specified. It is in terms of these restraints that the derivation will be made.

The sphere, being a surface, can be described by two independent parameters, say ϕ and λ. The same is true of the plane. A simultaneous solution of the equations of these surfaces expresses one of the parameters in terms of the other. But this is equivalent to describing the intersection of the two surfaces—the circle, which becomes then the locus of a single parameter.

Let $\boldsymbol{r}' = x'\boldsymbol{i} + y'\boldsymbol{j} + z'\boldsymbol{k}$ be the position vector of a fixed point on the sphere and $\boldsymbol{n} = n_1\boldsymbol{i} + n_2\boldsymbol{j} + n_3\boldsymbol{k}$ the unit normal of a plane which contains $\boldsymbol{r}'$. If $\boldsymbol{r} = x\boldsymbol{i} + y\boldsymbol{i} + z\boldsymbol{k}$ is a generic point of the plane its equation is $\boldsymbol{n}\cdot(\boldsymbol{r}-\boldsymbol{r}') = 0$ and in expanded form it becomes

$$(1)\qquad n_1x + n_2y + n_3z = n_1x' + n_2y' + n_3z' = K$$

The spherical coordinates $x = \cos\lambda\cos\phi$, $y = \sin\lambda\cos\phi$, and $z = \sin\phi$, define the unit sphere

$$(2)\qquad \boldsymbol{r} = \cos\lambda\cos\phi\boldsymbol{i} + \sin\lambda\cos\phi\boldsymbol{j} + \sin\phi\boldsymbol{k}$$

and if the circle (great or small) is to lie on the sphere and in the plane it must satisfy the equations of both. Thus

$$(3)\qquad n_1\cos\lambda\cos\phi + n_2\sin\lambda\cos\phi + n_3\sin\phi = K$$

$$-\frac{\pi}{2} \leqq \phi_0 \leqq \phi \leqq \phi_1 < \frac{\pi}{2}, \quad\text{or}\quad -\frac{\pi}{2} < \phi_0 \leqq \phi \leqq \phi_1 \leqq \frac{\pi}{2} \qquad \lambda \neq \text{constant}$$

$$-\frac{\pi}{2} \leqq \phi \leqq \frac{\pi}{2} \qquad \lambda = \text{constant}.$$

The restrictions on equation (3) are necessary to avoid contradictions. Since K is a constant and generally not zero its sign must be fixed. As the domain of ϕ corresponding to the definition of the sphere by (2) is $[-\pi/2, \pi/2]$ without this restriction one would have

$$K = \begin{cases} +n_3, & \phi = +\pi/2 \\ -n_3, & \phi = -\pi/2. \end{cases}$$

This situation stems from the singularities of the sphere at the poles. (See for example, L. Brand, *Vector and Tensor Analysis*, p. 203.) These however are not intrinsic properties of these particular points. They are "victims" of the coordinates chosen to represent the surface. Such singularities are removable by *ad hoc* definitions chosen to suit one's convenience. For example the normal $\partial\boldsymbol{r}/\partial\phi \times \partial\boldsymbol{r}/\partial\lambda$ to the sphere has a nonzero constant absolute value at every point except at the poles. There, when one attempts to determine the normal by means of the expression for $\partial\boldsymbol{r}/\partial\phi \times \partial\boldsymbol{r}/\partial\lambda$ it vanishes. Nevertheless, one can picture a tangent plane at a pole, just as at any other point of the sphere, which necessarily defines a nonzero normal. In such a case one defines the normal $\boldsymbol{N}$ of a sphere of radius a as

$$N = \begin{cases} \dfrac{\partial \boldsymbol{r}}{\partial \phi} \times \dfrac{\partial \boldsymbol{r}}{\partial \lambda}, & |\phi| \neq \dfrac{\pi}{2} \\[2ex] a^2 \boldsymbol{k}, & \phi = \dfrac{\pi}{2} \\[2ex] -a^2 \boldsymbol{k}, & \phi = -\dfrac{\pi}{2} \cdot \end{cases}$$

With this definition $\boldsymbol{N}$ becomes a continuous function of its independent variables. This may even be more appreciated when one notices that although the normal itself vanishes at the poles, the unit normal $\boldsymbol{n} = \boldsymbol{N}/|\boldsymbol{N}|$ does not and its direction is that of $\boldsymbol{k}$.

Other types of singularities may not be removable. Those are inherent to the surface. An example of a common surface with such a point is a right-circular cone. The apex has an infinite number of normals depending on one of the parameters, and the unit normal is not unique either. The restriction $\lambda \neq \text{cst}$ is necessary as otherwise (3) could not be satisfied for $n_3 \neq 0$ and nonconstant ϕ. (When $\lambda = \text{cst}$ and $n_3 = 0$ (3) becomes $\cos \phi (n_1 \cos \lambda + n_2 \sin \lambda) = K$. This can only be satisfied for $K = n_1 \cos \lambda + n_2 \sin \lambda = 0$. But this by (1) implies (a) $\boldsymbol{r}' = 0$, (b) $\boldsymbol{n} = 0$, (c) $\boldsymbol{r}' \perp \boldsymbol{n}$. (a) and (b) are untenable because $\boldsymbol{r}'$ is on the sphere and the plane must have a nonzero normal. The remaining possibility $\boldsymbol{r}' \perp \boldsymbol{n}$ implies that the plane passes through the origin which is the center of the sphere. This is the condition for a great circle.

From (3) follows

$$\begin{aligned} n_1 \cos \lambda + n_2 \sin \lambda &= K \sec \phi - n_3 \tan \phi \\ &= \pm \sqrt{\{n_1^2 + n_2^2\}} \sin \left(\lambda + \tan^{-1} \frac{n_1}{n_2} \right) \qquad |\phi| < \frac{\pi}{2} \end{aligned}$$

whence

$$\lambda = \sin^{-1} \frac{K \sec \phi - n_3 \tan \phi}{\pm \sqrt{\{1 - n_3^2\}}} - \tan^{-1} \frac{n_1}{n_2} \qquad |\phi| < \frac{\pi}{2} \tag{4}$$

since $n_1^2 + n_2^2 + n_3^2 = 1$. λ is evidently a multivalued function of ϕ. Its range will be best determined through an examination of the circle in question as a function of ϕ. But until then let it be written formally

$$\begin{aligned} \sin \lambda &= \frac{K \sec \phi - n_3 \tan \phi}{\pm \sqrt{\{1 - n_3^2\}}} \; \frac{n_2}{\pm \sqrt{\{1 - n_3^2\}}} \\ &\quad - \frac{\pm \sqrt{\{(1 - n_3^2) - (K \sec \phi - n_3 \tan \phi)^2\}}}{\pm \sqrt{\{1 - n_3^2\}}} \; \frac{n_1}{\pm \sqrt{\{1 - n_3^2\}}} \\ &= \frac{n_2 (K \sec \phi - n_3 \tan \phi) \mp n_1 \sqrt{\{(1 - n_3^2) - (K \sec \phi - n_3 \tan \phi)^2\}}}{1 - n_3^2} \end{aligned}$$

and

$$\cos\lambda = \frac{n_1(K\sec\phi - n_3\tan\phi) \pm n_2\sqrt{\{(1-n_3^2) - (K\sec\phi - n_3\tan\phi)^2\}}}{1-n_3^2}.$$

Substituting these in (2) and casting the radicand in a different form the equation of the circle becomes

$$\begin{aligned} r = {} & \frac{n_1(K - n_3\sin\phi) \pm n_2\sqrt{\{(1-n_3^2)(1-K^2) - (Kn_3 - \sin\phi)^2\}}}{1-n_3^2}\,i \\ & + \frac{n_2(K - n_3\sin\phi) \mp n_1\sqrt{\{(1-n_3^2)(1-K^2) - (Kn_3 - \sin\phi)^2\}}}{1-n_3^2}\,j \\ & + \sin\phi k. \end{aligned} \tag{5}$$

It is clear that the radical cannot be allowed to become imaginary. Suppose that the radical vanishes at $\phi=\phi_c$. This yields

$$\sin\phi_c = Kn_3 \pm \sqrt{\{(1-n_3^2)(1-K^2)\}}.$$

Since the component along the k-axis is sin ϕ it becomes immediately evident that the "uppermost" point of the circle is

$$\sin\phi_{c_1} = Kn_3 + \sqrt{\{(1-n_3^2)(1-K^2)\}}$$

and the "lowermost" is

$$\sin\phi_{c_2} = Kn_3 - \sqrt{\{(1-n_3^2)(1-K^2)\}}$$

or vice versa. As ϕ is also a parameter of the surface it is clear that when ϕ_{c_i} is reached the sign of the radical should be changed and the values of ϕ retraced backwards from the direction which led to ϕ_{c_i}. When both ϕ_{c_i} are encountered the circle will have been completely described. Thus, if the upper sign of the radical corresponds to the "right side" of the circle then the lower sign pertains to the "left side," or vice versa.

The i and j components of r in (5) are $\cos\lambda\cos\phi$ and $\sin\lambda\cos\phi$, respectively. Since $\cos\phi$ is a single-valued function of ϕ and multiplies $\cos\lambda$ and $\sin\lambda$ it follows that the latter are multi-valued functions of ϕ. This can be traced back to the expression of λ in (4) having the radical with two signs. It thus becomes evident that the two signs may be necessary in order to describe the whole circle. This is an example of why one should be careful with the interpretation of multi-valued functions.

The representation in (5) fails when $n_1=n_2=0$. Geometrically this means that the plane containing the circle is parallel to the i-j plane. The unit normal then is $n=k$, $\phi=$constant and λ becomes indeterminate. This characterizes the parallels. In theory λ can be evaluated or defined by the limit of (4) as $n_1, n_2\to 0$. This would preserve the continuity of λ. The limit depends on the manner in which n_1 and n_2 approach 0. However, in such cases the circle may be represented as

$$r = (\cos\lambda i + \sin\lambda j)\cos\phi_0 + \sin\phi_0 k \tag{6}$$

with ϕ_0 a constant and λ the parameter.

It has been established before that $K=0$ yields a great circle. Equation (4) reduces then to

$$\lambda = \sin^{-1}\frac{-n_3\tan\phi}{\pm\sqrt{\{1-n_3^2\}}} - \tan^{-1}\frac{n_1}{n_2}. \tag{7}$$

Differentiating (4) one obtains

$$d\lambda = \pm\frac{d\phi}{\cos\phi\sqrt{\{(\cos\phi/n_3)^2-1\}}}$$

which is the differential equation of a geodesic on a sphere. (See, for example, R. Weinstock, *Calculus of Variation*, p. 27. A geodesic on a surface is a curve of zero geodesic or tangential curvature, or a curve whose normals are parallel to the surface normals.)

Elements of arc length along a parallel and a meridian are $\cos\phi d\lambda$ and $1\cdot d\phi$, respectively. Since the parallels and meridians are orthogonal the direction of the geodesic at any point ϕ can be computed from

$$\tan\psi = \frac{d\phi}{\cos\phi d\lambda} = \sqrt{\left\{\frac{\cos^2\phi}{n_3^2}-1\right\}},$$

where ψ is the angle the tangent to the geodesic forms with a parallel, the complement with respect to $\pi/2$ of the azimuth. Since $n_3^2\leqq 1$ and the domain of ϕ can be $[-\pi/2,\pi/2]$ it follows that for some $\phi=\phi_c$, $\cos\phi_c=n_3$ and $\psi=0$. In other words, at $\phi_c=\cos^{-1}n_3$ the geodesic becomes parallel to a parallel whose radius is n_3. It is easy to verify that if

$$\sin^2\phi_c = (Kn_3\pm\sqrt{\{(1-n_3^2)(1-K^2)\}})^2$$

then

$$\cos^2\phi_c = (n_3\sqrt{\{1-K^2\}}\mp K\sqrt{1-n_3^2})^2,$$

and when $K=0$, $\cos^2\phi_c=n_3^2$.

The general expression for the tangent to the circle of (5) is given by

$$\begin{aligned}\frac{dr}{d\phi} &= \frac{\cos\phi}{1-n_3^2}\left[-n_1n_3\pm\frac{n_2(Kn_3-\sin\phi)}{\sqrt{\{(1-n_3^2)(1-K^2)-(Kn_3-\sin\phi)^2\}}}\right]i\\ &+\frac{\cos\phi}{1-n_3^2}\left[-n_2n_3\mp\frac{n_1(Kn_3-\sin\phi)}{\sqrt{\{(1-n_3^2)(1-K^2)-(Kn_3-\sin\phi)^2\}}}\right]j+\cos\phi k.\end{aligned}\tag{8}$$

The *unit* tangent t is obtained by differentiating r with respect to arc length (of the circle) s:

$$t = \frac{dr}{ds} = \frac{dr}{d\phi}\ \frac{d\phi}{ds} \tag{9}$$

and $|t| = 1$. This yields

$$
\begin{aligned}
1 &= \left| \frac{\sqrt{\{1 - K^2\}} \cos\phi}{\sqrt{\{(1 - n_3^2)(1 - K^2) - (Kn_3 - \sin\phi)^2\}}} \right| \left| \frac{d\phi}{ds} \right|, \\
ds &= \pm \frac{\sqrt{(1 - K^2)}}{\sqrt{\{(1 - n_3^2)(1 - K^2) - (Kn_3 - \sin\phi)^2\}}} \cos\phi d\phi.
\end{aligned} \tag{10}
$$

In order to determine the proper sign one can consider the behavior of the tangent or, preferably, the unit tangent, from a geometric point of view, or to note that arc length should be a monotone increasing function. Geometrically the unit tangent should turn around the circle continuously. In terms of the k-component this means that if between ϕ_{c_1} and ϕ_{c_2} curve of constant width it is positive, then between ϕ_{c_2} and ϕ_{c_1} curve of constant width it must be negative. This shows the necessity of changing the sign when passing the points ϕ_{c_1} and ϕ_{c_2} as in the other radicals of (5). Hence, when substituting (10) in (9) one obtains

$$
\begin{aligned}
t = {} & \frac{1}{(1 - n_3^2)(1 - K^2)^{1/2}} [\mp n_1 n_3 \sqrt{\{(1 - n_3^2)(1 - K^2) - (Kn_3 - \sin\phi)^2\}} \\
& \qquad + n_2(Kn_3 - \sin\phi)] i \\
& + \frac{1}{(1 - n_3^2)(1 - K^2)^{1/2}} [\pm n_2 n_3 \sqrt{\{(1 - n_3^2)(1 - K^2) - (Kn_3 - \sin\phi)^2\}} \\
& \qquad - n_1(Kn_3 - \sin\phi)] j \\
& \pm \frac{1}{(1 - K^2)^{1/2}} [\sqrt{\{(1 - n_3^2)(1 - K^2) - (Kn_3 - \sin\phi)^2\}}] k.
\end{aligned} \tag{11}
$$

Analogously to the case of the normal and unit normal to the sphere, the unit tangent (11) exists everywhere along the circle while the tangent (8) itself does not; (11) was obtained by choosing the positive sign of (10). In reality it may be necessary to choose the opposite one. This will depend on the direction cosines of the unit normal to the plane n. They will also determine in which of the intervals $[\phi_{c_1}, \phi_{c_2}]$ or $[\phi_{c_2}, \phi_{c_1}]$ $ds/d\phi$ is positive or negative and thus indicate the proper sign to be chosen. In view of the monotonicity of s one should interpret

$$
s - s_0 = \int_{s_0}^{s} d\sigma = \int_{\phi_6}^{\phi} \sqrt{\left\{ \frac{(1 - K^2)}{(1 - n_2)^3(1 - K^2) - (Kn_3 - \sin\theta)^2} \right\}} \cos\theta d\theta \tag{12}
$$

as

$$
\begin{aligned}
& \int_{\phi_0}^{\phi_c} \sqrt{\left\{ \frac{(1 - K^2)}{(1 - n_3^2)(1 - K^2) - (Kn_3 - \sin\theta)^2} \right\}} \cos\theta d\theta \\
& + \int_{\phi}^{\phi} \sqrt{\left\{ \frac{(1 - K^2)}{(1 - n_3{}^2)(1 - K^2) - (Kn_3 - \sin\theta)^2} \right\}} \cos\theta d\theta,
\end{aligned}
$$

whenever the interval of integration contains ϕ_c; (12) can be expressed in terms of an elementary function. Thus

$$s = \sqrt{\{1 - K^2\}}\left[\sin^{-1}\frac{\sin\phi - Kn_3}{\sqrt{\{(1 - n_3^2)(1 - K^2)\}}} - \sin^{-1}\frac{\sin\phi_0 - Kn_3}{\sqrt{\{(1 - n_3^2)(1 - K^2)\}}}\right] + s_0$$

Of course s should always be positive.

There are several ways of finding the radius of the circle. One is to use the fact that the curvature is a constant and its reciprocal is the radius. This can be done by finding an expression for the curvature vector and computing the reciprocal of its absolute value. The curvature vector is

$$\begin{aligned}\frac{dt}{ds} &= \frac{dt}{d\phi}\frac{d\phi}{ds}\\ &= \frac{1}{(1 - n_3^2)(1 - K^2)}\,[-n_1n_3(Kn_3 - \sin\phi)\\ &\qquad \pm n_2\sqrt{\{(1 - n_3^2)(1 - K^2) - (Kn_3 - \sin\phi)^2\}}]i\\ &\quad + \frac{1}{(1 - n_3^2)(1 - K^2)}\,[-n_2n_3(Kn_3 - \sin\phi)\\ &\qquad \pm n_1\sqrt{\{(1 - n_3^2)(1 - K^2) - (Kn_3 - \sin\phi)^2\}}]j\\ &\quad + \frac{1}{1 - K^2}\,[Kn_3 - \sin\phi]k.\end{aligned} \tag{13}$$

This leads to

$$\left|\frac{dt}{ds}\right| = \frac{1}{\sqrt{\{1 - K^2\}}}$$

so that the radius is $R = \sqrt{\{1 - K^2\}}$ which is certainly plausible for a great circle. Another way is to resort to geometric intuition and reason that the points on the circle corresponding to ϕ_{c_1} and ϕ_{c_2} are diametrically opposite. Therefore the radius should be given by

$$\tfrac{1}{2}\left|\, r(\phi_{c_2}) - r(\phi_{c_1})\,\right| .$$

At this point the center of the circle r_c is readily determined:

$$\begin{aligned} r_c &= r(\phi_{c_1}) + \tfrac{1}{2}[r(\phi_{c_2}) - r(\phi_{c_1})] = \tfrac{1}{2}[r(\phi_{c_2}) + r(\phi_{c_1})]\\ &= Kn_1i + Kn_2j + Kn_3k = Kn.\end{aligned} \tag{14}$$

This too is plausible for a great circle. Having determined, among other things, the domain of ϕ in terms of n and K one may return to equation (4) and establish the range of λ. It should be realized that (4) by itself defines the circle but in an ambiguous way since λ is a multivalued function of ϕ. The equation must be subordinated to the circle it is required to describe. To remove the ambiguity let it be assumed that $n_3 \geqq 0$. This is no loss of generality at all as the sense of a normal to a surface can be assumed arbitrarily. There is also no serious loss of generality by assuming $K>0$ as the equations and arguments are easily modified for $K<0$. (The case $K=0$ is easily inferred from the subsequent discussion.) Consequently there exist the following possible relationships between K and n_3:

$$\text{(a) } K < n_3, \qquad \text{(b) } K > n_3, \qquad \text{(c) } K = n_3.$$

Each of these has a geometrical interpretation and a corresponding effect on the variable part of (4). They will now be examined.

The above trichotomy implies that equation (1) is satisfied by a vector $r=0i+0j+z_0k$ such that

$$\text{(a) } z_0 < 1; \qquad \text{(b) } z_0 > 1; \qquad \text{(c) } z_0 = 1.$$

This means that, dividing the sphere into "left" and "right" hemispheres,

(a) the plane cuts the sphere in both hemispheres and λ ranges exactly once over an interval of 2π; (λ is univalent throughout its domain)
(b) for some such division, the plane cuts the sphere in only one hemisphere and λ ranges twice over an interval of length less than π; (λ is multivalent throughout its domain except at two points)
(c) the plane cuts the sphere through a pole and λ ranges once over an open interval (one point removed) of length π; (λ is univalent throughout its domain, with the pole removed).

(A function $f(x)$ is univalent in $[a, b]$ if for every $x_1, x_2 \in [a, b]$, $f(x_1)=f(x_2)$ implies $x_1=x_2$.)

These conclusions can also be deduced by noting that K and n_3 are the cosines of the angles between the directions of r' and n and n and k.

For a fixed sign of the radical the expression

$$\text{(15)} \qquad \frac{K\sec\phi - n_3\tan\phi}{\pm\sqrt{\{1-n_3^2\}}} = \frac{K - n_3\sin\phi}{\pm\sqrt{\{(1-n_3^2)\}}\cos\phi}$$

vanishes when $\sin\phi=K/n_3$ and attains an extreme when $\sin\phi=n_3/K$. Under this condition it is also true that

$$\text{(16)} \qquad \frac{K - n_3\sin\phi_{c_1}}{\pm\sqrt{\{1-n_3^2\}}\cos\phi_{c_1}} = \frac{K - n_3\sin\phi_{c_2}}{\pm\sqrt{\{1-n_3^2\}}\cos\phi_{c_2}} = \pm 1.$$

This suggests that in the open interval (ϕ_{c_2}, ϕ_{c_1}) (14)

(a) Cannot attain a nonzero extreme.
(b) Cannot vanish.
(c) Can neither vanish nor attain an extreme.

and in view of (15) that at some point, (14)

(a) Vanishes.
(b) Attains a nonzero extreme.
(c) Does not vanish and does not attain an extreme.

With this information on hand λ is readily cleared of ambiguities inasmuch as $\tan^{-1}[n_1/n_2]$ is well defined by the signs of n_1 and n_2.

(a) Starting at

$$\phi_{c_1} = \sin^{-1}[Kn_3 + \sqrt{\{(1-K^2)(1-n_3^2)\}}] \quad \text{(principal value)}$$

with the positive sign of the radical in (4) λ varies from

$$\sin^{-1}\frac{K - n_3(Kn_3 + \sqrt{\{(1-K^2)(1-n_3^2)\}})}{(n_3\sqrt{\{1-K^2\}} - K\sqrt{\{1-n_3^2\}})\sqrt{\{1-n_3^2\}}} - \tan^{-1}\frac{n_1}{n_2} = \frac{\pi}{2} - \tan^{-1}\frac{n_1}{n_2}$$

until the variable part vanishes for $\sin\phi = Kn_3$ and λ becomes $\tan^{-1}[n_1/n_2]$. At this point the sign of the radical is changed and when ϕ_{c_2} is encountered, λ becomes $-\pi/2 - \tan^{-1}[n_1/n_2]$. This describes the second quarter of the circle, the one "below" the first. Going on (not back) to ϕ_{c_1} (retracing ϕ) another branch of $\sin^{-1} u$ is chosen, one that is characterized by $\sin^{-1} 0 = -\pi$. For $u = -1$ this branch yields $\sin^{-1}[-1] = -\pi/2$. Since at ϕ_{c_2} the variable part of (4) is equal to -1 this branch takes over and when $\sin\phi = K/n_3$ is reached again (in the process of describing the third quarter of the circle) λ becomes $-\pi - \tan^{-1}[n_1/n_2]$. The sign of the radical is now changed back to its original state. As ϕ_{c_1} is approached the argument (4) varies from 0 to 1 and this corresponds to $\sin^{-1} u$ going from $-\pi$ to $-3\pi/2$. λ thus ranges over an interval of length 2π.

(b) Starting at ϕ_{c_1} with the positive radical, λ is $\pi/2 - \tan^{-1}[n_1/n_2]$. It reaches the extreme

$$\sin^{-1}[\sqrt{\{K^2 - n_3^2\}}/\sqrt{\{1-n_3^2\}}] - \tan^{-1}[n_1/n_3]$$

for $\sin\phi = n_3/K$. Continuing with ϕ to ϕ_{c_2}, λ retraces its values until it reaches $\pi/2 - \tan^{-1}[n_1/n_2]$ again at ϕ_{c_2}. Choosing now a branch of $\sin^{-1} u$ (without changing the sign of the radical) represented by $\sin^{-1} 1 = \pi/2$ and $\sin^{-1} 0 = \pi$, λ reaches now the extreme

$$\pi - \sin^{-1}[\sqrt{\{K^2 - n_3^2\}}/\sqrt{\{1-n_3^2\}}] - \tan^{-1}[n_1/n_2]$$

changes direction and returns back to $\pi/2 - \tan^{-1}[n_1/n_2]$. It thus ranges twice over an interval of length

$$\pi - 2\sin^{-1}[\sqrt{\{K^2 - n_3^2\}}/\sqrt{\{1-n_3^2\}}].$$

(c) The evaluation of λ can be made exactly as in the former case. Since $K = n_3$ it follows from the last expression in (b) that the interval covered by λ is of length π.

It should be understood that the above procedures to define λ are by no means the only ones possible. The starting points as well as the choices of the signs of the radicals are arbitrary. As long as λ is made a continuous function of ϕ through its functional elements the goal is attained. Notice should also be taken of the fact that in (b) and (c) it was not necessary to change the sign of the radical. The reason is that the interval of λ in these cases did not exceed π.

Part 2. For the second representation let (3) be solved for ϕ. Then

$$(17)\qquad \pm\sqrt{\{n_3^2 + (n_1\cos\lambda + n_2\sin\lambda)^2\}}\,\sin\left(\phi + \tan^{-1}\frac{n_1\cos\lambda + n_2\sin\lambda}{n_3}\right) = K$$

and

$$(18)\qquad \phi = \sin^{-1}\frac{K}{\pm\sqrt{\{n_3^2 + (n_1\cos\lambda + n_2\sin\lambda)^2\}}} - \tan^{-1}\frac{n_1\cos\lambda + n_2\sin\lambda}{n_3}$$

Before determining the proper sign and branches for a given circle it may be instructive to consider the special case $n_3 = 0$, $K \neq 0$. This is a circle symmetrical about the equatorial plane. The solution of (3) for ϕ becomes then

$$(19)\qquad \phi = \cos^{-1}\frac{K}{n_1\cos\lambda + n_2\sin\lambda}\,.$$

It is clear that λ must be so restricted that $|K| \leqq |n_1 \cos\lambda + n_2 \sin\lambda|$. Since $n_3 = 0$ the amplitude of the denominator in (18) is unity and its period is 2π. Therefore there exist four values of λ, say λ_{c_i}, $i = 1, 2, 3, 4$, for which the absolute value of the denominator becomes $|K|$. At these points λ must be retraced when ϕ or the circle is described. Evidently only two of them are pertinent to the given circle and, at these points, ϕ must be assigned the value 0 since any other value satisfying (18) would lie outside the permissible range $(-\pi/2, \pi/2)$ of ϕ.

Now if for $i = 1, 2$, $n_1 \cos\lambda_{c_i} + n_2 \sin\lambda_{c_i} = K$, then for $i = 3, 2$, $n_1 \cos\lambda_{c_i} + n_2 \sin\lambda_{c_i} = -K$ and conversely. Therefore, a choice of λ_{c_i}, $i = 3, 2$, as denoted above, would mean that

$$\phi_{c_i} = \cos^{-1}\frac{K}{n_1\cos\lambda_{c_i} + n_2\sin\lambda_{c_i}} = \cos^{-1}[-1] = \pm(2k+1)\pi, \quad k = 0, 1, 2, \cdots,$$

which is outside $(-\pi/2, \pi/2)$.

The codomain of ϕ can now be determined. Starting with $\lambda_{c_i}(\phi_{c_1} = 0)$, λ should proceed towards λ_{c_2} via the shortest path, i.e., without encountering λ_{c_3} and λ_{c_4}. Along this path ϕ reaches an extreme ϕ_{e_i} and returns to 0 at λ_{c_2}. At this point another branch of $\cos^{-1} u$ must be chosen, one that is characterized by $\cos^{-1} 1 = 0$, $\cos^{-1} 0 = -\pi/2$ if for the first branch $\cos^{-1} 1 = 0$ and $\cos^{-1} 0 = +\pi/2$ was true, and conversely.

Suppose now that $n_3 \neq 0$ but $|n_3|$ is very small, and that n_1 and n_2 are essentially as before. The circle is then not too different from the one in the previous case. Since $n_3 \neq 0$ there is no difficulty in using (17). But before this is attempted let (16) be examined.

In view of the choice of n_3 the amplitude of $n_1 \cos\lambda + n_2 \sin\lambda = A(\lambda)$ is essentially unity for two values of λ, say λ_{a_i}, $i = 1, 2$. As $\phi \in (-\pi/2, \pi/2)$ the choice of branch for $\tan^{-1}[A(\lambda)/n_3]$ in the neighborhood of λ_{a_i} must be such that $\tan^{-1}[A(\lambda_{a_i})/n_3] \approx \pm\pi/2$. Consequently, as ϕ varies nearly symmetrically about 0, $\phi + \tan^{-1}[A(\lambda)/n_3]$ essentially varies symmetrically about $+\pi/2$ for the choice of the upper sign and about $-\pi/2$ for the lower. Hence, $\sin(\phi+\tan^{-1}[A(\lambda)/n_3])$ follows suit about ± 1 correspondingly. Since $\sin x$ does not change sign in $[0, \pi]$ or in $[-\pi, 0]$ it follows that $\sin(\phi+\tan^{-1}[A(\lambda)/n_3])$ does not change sign in the course of describing the circle. As the sign of K is fixed the sign of the radical in (16), and therefore in (17), must be fixed for a given circle. Thus it appears so far that the choice of the proper sign in (17) depends on the sign of K and the choice of the branch for $\tan^{-1}[A(\lambda)/n_3]$, and once it is chosen it is not to be altered in the process of describing the circle. Note that the case $n_3=0$ did not give rise to an ambiguous sign.

This argument may now be applied to a circle in general by observing that it may be obtained from one for which $n_3=0$ by a continuous transformation of the constants. (A meridian may require special handling.) Assuming, with no loss of generality, that $K>0$, and choosing $[0, \pi/2]$ for the codomain of $\tan^{-1}[A(\lambda)/n_3]$ and the positive sign of the radical,

$$\begin{aligned}\sin\phi &= \frac{Kn_3 \mp (n_1\cos\lambda + n_2\sin\lambda)\sqrt{\{n_3^2 + (n_1\cos\lambda + n_2\sin\lambda)^2 - K^2\}}}{n_3^2 + (n_1\cos\lambda + n_2\sin\lambda)^2}\\ \cos\phi &= \frac{K(n_1\cos\lambda + n_2\sin\lambda) \pm n_3\sqrt{\{n_3^2 + (n_1\cos\lambda + n_2\sin\lambda)^2 - K^2\}}}{n_3^2 + (n_1\cos\lambda + n_2\sin\lambda)^2}\end{aligned} \tag{20}$$

and the Cartesian form of the circle becomes from (2) and (19)

$$\begin{aligned} r = {} & (\cos\lambda i + \sin\lambda j)\\ & \cdot\frac{K(n_1\cos\lambda + n_2\sin\lambda) \pm n_3\sqrt{\{n_3^2 + (n_1\cos\lambda + n_2\sin\lambda)^2 - K^2\}}}{n_3^2 + (n_1\cos\lambda + n_2\sin\lambda)^2}\\ & + \frac{Kn_3 \mp (n_1\cos\lambda + n_2\sin\lambda)\sqrt{\{n_3^2 + (n_1\cos\lambda + n_2\sin\lambda)^2 - K^2\}}}{n_3^2 + (n_1\cos\lambda + n_2\sin\lambda)^2}k.\end{aligned} \tag{21}$$

This equation reduces to (6) for $n_1=n_2=0$ ($n_3=1$) by noting that $\sqrt{\{1-K^2\}}$ corresponds then to $\cos\phi_0$. Equations (19) or (20) afford a way to determine the branch points by applying to it the reasoning applied to (5)—the radicals cannot be allowed to become imaginary. Therefore the solution of $n_3^2+(n_1\cos\lambda+n_2\sin\lambda)^2-K^2=0$ for λ, λ_{c_1} and λ_{c_2} gives the branch points if they exist. $\lambda_{c_1}=\text{cst}$ and $\lambda_{c_2}=\text{cst}$ are the meridians which are tangent to the circle. The handling of the dual sign when branch points are encountered follows the same pattern as discussed previously.

Unlike the previous representation, $\frac{1}{2}|r(\lambda_{c_2})-r(\lambda_{c_1})|$ is generally not the radius. (It is the radius when $n_3=0$.) It can be found from the absolute value of the difference between (20) evaluated at λ_0 with the upper signs and the lower signs, λ_0 being the λ-coordinate of the center of the circle. It corresponds to the point at which ϕ in (16) attains its extremes ϕ_{e_i}, and is the solution of $d\phi/d\lambda=0$ ((21) below), which is $\lambda=\tan^{-1}[n_2/n_1]$. Substituting this in (20) gives for the upper sign

$$\mathbf{r}_2(\lambda_0) = (n_1 i + n_2 j)\left(K + n_3 \frac{\sqrt{\{1 - K^2\}}}{\sqrt{\{1 - n_3\}}}\right) + (Kn_3 - \sqrt{\{1 - K^2\}}\sqrt{\{1 - n_3^2\}})k$$

and for the lower sign

$$\mathbf{r}_1(\lambda_0) = (n_1 i + n_2 j)\left(K - n_3 \frac{\sqrt{\{1 - K^2\}}}{\sqrt{\{1 - n_3\}}}\right) + (Kn_3 + \sqrt{\{1 - K^2\}}\sqrt{\{1 - n_3^2\}})k$$

and

$$\tfrac{1}{2}\left| \mathbf{r}_2(\lambda_0) - \mathbf{r}_1(\lambda_0) \right| = \sqrt{\{1 - K^2\}}.$$

The expressions for the tangent and unit tangent, normal and unit normal of (2) in terms of λ are cumbersome and will not be given here. Should it be necessary to develop them it is best not to differentiate (19) directly with respect to λ but rather to derive $d\mathbf{r}/d\lambda$ from (2), substitute therein

$$(22)\quad \frac{d\phi}{d\lambda} = \left[\frac{(n_1 \cos\lambda + n_2 \sin\lambda)K}{\pm\sqrt{\{n_3^2 + (n_1 \cos\lambda + n_2 \sin\lambda)^2 - K^2\}}} - n_3\right] \frac{n_2 \cos\lambda - n_1 \sin\lambda}{n_3^2 + (n_1 \cos\lambda + n_2 \sin\lambda)^2}$$

and (19).

Acknowledgement of the final proof reading of the text is made to Mr. Robert Diamond.

THE EXPONENTIAL REPRESENTATION OF UNITARY MATRICES

R. F. RINEHART, Case Institute of Technology

A classical result (e.g. see [1]) is:

THEOREM. *If S is a skew hermitian matrix, then e^S is unitary. Conversely, any unitary matrix can be represented as e^S, for some skew hermitian matrix S.*

Proofs of this theorem usually invoke the diagonalizability of a unitary matrix as well as the fact that the characteristic roots lie on the unit circle. The purpose of the present note is to show that the theorem is essentially obvious from the standpoint of matrix function theory, without knowledge of these more advanced properties.

We recall certain pertinent properties of matric primary functions (i.e., the extension of functions of a complex variable to complex matrices). Let $f(z)$ be a function of a complex variable, and $f(Z)$ its extension to $n\times n$ complex matrices. Then

I. If $f(z)$ is defined at the characteristic roots of A and analytic at the repeated characteristic roots, then $f(A)$ is well defined, [3].

II. Any functional identity satisfied by scalar functions is satisfied by their matric extensions [2].

$$\mathbf{r}_2(\lambda_0) = (n_1\mathbf{i} + n_2\mathbf{j})\left(K + n_3\frac{\sqrt{\{1 - K^2\}}}{\sqrt{\{1 - n_3\}}}\right) + (Kn_3 - \sqrt{\{1 - K^2\}}\sqrt{\{1 - n_3^2\}})\mathbf{k}$$

and for the lower sign

$$\mathbf{r}_1(\lambda_0) = (n_1\mathbf{i} + n_2\mathbf{j})\left(K - n_3\frac{\sqrt{\{1 - K^2\}}}{\sqrt{\{1 - n_3\}}}\right) + (Kn_3 + \sqrt{\{1 - K^2\}}\sqrt{\{1 - n_3^2\}})\mathbf{k}$$

and

$$\tfrac{1}{2}\left| \mathbf{r}_2(\lambda_0) - \mathbf{r}_1(\lambda_0)\right| = \sqrt{\{1 - K^2\}}.$$

The expressions for the tangent and unit tangent, normal and unit normal of (2) in terms of λ are cumbersome and will not be given here. Should it be necessary to develop them it is best not to differentiate (19) directly with respect to λ but rather to derive $d\mathbf{r}/d\lambda$ from (2), substitute therein

$$(22)\quad \frac{d\phi}{d\lambda} = \left[\frac{(n_1\cos\lambda + n_2\sin\lambda)K}{\pm\sqrt{\{n_3^2 + (n_1\cos\lambda + n_2\sin\lambda)^2 - K^2\}}} - n_3\right]\frac{n_2\cos\lambda - n_1\sin\lambda}{n_3^2 + (n_1\cos\lambda + n_2\sin\lambda)^2}$$

and (19).

Acknowledgement of the final proof reading of the text is made to Mr. Robert Diamond.

THE EXPONENTIAL REPRESENTATION OF UNITARY MATRICES

R. F. RINEHART, Case Institute of Technology

A classical result (e.g. see [1]) is:

THEOREM. *If S is a skew hermitian matrix, then e^S is unitary. Conversely, any unitary matrix can be represented as e^S, for some skew hermitian matrix S.*

Proofs of this theorem usually invoke the diagonalizability of a unitary matrix as well as the fact that the characteristic roots lie on the unit circle. The purpose of the present note is to show that the theorem is essentially obvious from the standpoint of matrix function theory, without knowledge of these more advanced properties.

We recall certain pertinent properties of matric primary functions (i.e., the extension of functions of a complex variable to complex matrices). Let $f(z)$ be a function of a complex variable, and $f(Z)$ its extension to $n\times n$ complex matrices. Then

I. If $f(z)$ is defined at the characteristic roots of A and analytic at the repeated characteristic roots, then $f(A)$ is well defined, [3].

II. Any functional identity satisfied by scalar functions is satisfied by their matric extensions [2].

III. $f(A^T) = [f(A)]^T$, where the superscript T denotes "transpose" [3].

IV. If $f(z)$ satisfies $f(\bar{z}) = f(z)$, then $f(\overline{A}) = f(A)$ [2].

Now let S be skew hermitian, i.e., $S^* = -S$. Since e^z satisfies the condition of IV, $e^S(e^S)^* = e^S e^{S^*}$ by III and IV, and $e^S e^{S^*} = e^S e^{-S} = e^0 = I$ by II. Hence e^S is unitary.

Conversely, let U be unitary, i.e., (1) $U^* = U^{-1}$. Let $\log z$ denote the customary principal value of the logarithm function, and suppose initially that U has no negative real eigenvalues. Then (1) implies $\log U^* = \log U^{-1} = -\log U$ by II. Now since $\log z$ satisfies the condition of IV in the complex plane with the negative real axis deleted, III and IV imply $\log U^* = (\log U)^* = -\log U$. Hence $\log U$ is skew hermitian and by II, $S = e^{\log U}$.

It remains to dispose of the case where U has one or more negative real eigenvalues, since in this case $\log U^* = (\log U)^*$ does not follow from III and IV. This is easily accomplished by a classical trick. Let $V = e^{i\theta} U$. Then V is again unitary, and its eigenvalues are the eigenvalues of U multiplied by $e^{i\theta}$, i.e., rotated through the angle θ. Choose θ so that no eigenvalue of V is negative real. Then by the preceding paragraph $V = e^S$ where S is skew hermitian, and $U = e^{-i\theta} e^S = e^{S - i\theta I}$, by II, and $S - i\theta I$ is skew hermitian, and U has the required representation.

The same proof can be used for the real version of the theorem, with "skew-hermitian" replaced by "real, skew" and "unitary" replaced by "orthogonal," for orthogonal matrices which (in the second half of the theorem) have no negative real eigenvalues. The classical result (e.g., see [4]) that such an exponential representation is possible for any proper orthogonal matrix (i.e., possessing an even number of eigenvalues equal to -1), requires a deeper analysis than the above approach provides.

It may also be noted that the classical result that $U = (I+S)^{-1}(I-S)$ is unitary for S skew hermitian, similarly becomes obvious from the point of view of properties I–IV, of matric primary functions.

Research supported by ONR grant.

References

1. H. L. Hamburger and M. E. Grimshaw, Linear transformations in n-dimensional vector space, Cambridge University Press, 1951.

2. H. Richter, Über Matrixfunktionen, Math. Ann., 122 (1950) 6–34.

3. R. F. Rinehart, The equivalence of definitions of a matric function, Amer. Math. Monthly, 62 (1955) 395–413.

4. H. Schwerdtfeger, Introduction to linear algebra and the theory of matrices, P. Noordhoff, Groningen, the Netherlands, 1950.

OPEN PROBLEMS OF INTEREST IN APPLIED MATHEMATICS

HENRY WINTHROP, University of South Florida

I. Introduction. College undergraduates who study applied mathematics find that the areas of application, although numerous, can generally be subsumed under two major categories. These are applications in the fields of the

III. $f(A^T) = [f(A)]^T$, where the superscript T denotes "transpose" [3].

IV. If $f(z)$ satisfies $f(\bar{z}) = f(z)$, then $f(\overline{A}) = f(A)$ [2].

Now let S be skew hermitian, i.e., $S^* = -S$. Since e^z satisfies the condition of IV, $e^S(e^S)^* = e^S e^{S^*}$ by III and IV, and $e^S e^{S^*} = e^S e^{-S} = e^0 = I$ by II. Hence e^S is unitary.

Conversely, let U be unitary, i.e., (1) $U^* = U^{-1}$. Let log z denote the customary principal value of the logarithm function, and suppose initially that U has no negative real eigenvalues. Then (1) implies $\log U^* = \log U^{-1} = -\log U$ by II. Now since log z satisfies the condition of IV in the complex plane with the negative real axis deleted, III and IV imply $\log U^* = (\log U)^* = -\log U$. Hence $\log U$ is skew hermitian and by II, $S = e^{\log U}$.

It remains to dispose of the case where U has one or more negative real eigenvalues, since in this case $\log U^* = (\log U)^*$ does not follow from III and IV. This is easily accomplished by a classical trick. Let $V = e^{i\theta} U$. Then V is again unitary, and its eigenvalues are the eigenvalues of U multiplied by $e^{i\theta}$, i.e., rotated through the angle θ. Choose θ so that no eigenvalue of V is negative real. Then by the preceding paragraph $V = e^S$ where S is skew hermitian, and $U = e^{-i\theta} e^S = e^{S - i\theta I}$, by II, and $S - i\theta I$ is skew hermitian, and U has the required representation.

The same proof can be used for the real version of the theorem, with "skew-hermitian" replaced by "real, skew" and "unitary" replaced by "orthogonal," for orthogonal matrices which (in the second half of the theorem) have no negative real eigenvalues. The classical result (e.g., see [4]) that such an exponential representation is possible for any proper orthogonal matrix (i.e., possessing an even number of eigenvalues equal to -1), requires a deeper analysis than the above approach provides.

It may also be noted that the classical result that $U = (I+S)^{-1}(I-S)$ is unitary for S skew hermitian, similarly becomes obvious from the point of view of properties I–IV, of matric primary functions.

Research supported by ONR grant.

References

1. H. L. Hamburger and M. E. Grimshaw, Linear transformations in n-dimensional vector space, Cambridge University Press, 1951.

2. H. Richter, Über Matrixfunktionen, Math. Ann., 122 (1950) 6–34.

3. R. F. Rinehart, The equivalence of definitions of a matric function, Amer. Math. Monthly, 62 (1955) 395–413.

4. H. Schwerdtfeger, Introduction to linear algebra and the theory of matrices, P. Noordhoff, Groningen, the Netherlands, 1950.

OPEN PROBLEMS OF INTEREST IN APPLIED MATHEMATICS

HENRY WINTHROP, University of South Florida

I. Introduction. College undergraduates who study applied mathematics find that the areas of application, although numerous, can generally be subsumed under two major categories. These are applications in the fields of the natural sciences and applications in the fields of the social sciences. There are, however, contexts of genuine interest in contemporary research, which, although of practical significance, have received *relatively little attention* in developments in applied mathematics. Among these are the problems associated with the mathematical theory of war and problems associated with the spread of novel social behavior, that is, the mathematical theory of diffusion. This latter area is concerned, for instance, with such matters as the spread of rumor, the spread in the adoption of a novel, social pastime like Canasta, the spread in the adoption of newly coined words, slang or otherwise, the propagation of new attitudes in the general population, etc. Some work on the mathematical theory of war has been done by Rashevsky [3], Richardson [4], and others. There are, however, so many different ways of approaching this subject that I propose in this paper to give one example of this which, I feel sure, will be of interest to the college undergraduate. Work on the mathematical theory of diffusion has been undertaken by Dodd [1], Rapoport [2] and other members of the Chicago School of Mathematical Biophysics, Winthrop [5, 6], and many others. I shall furnish one example of a diffusion theory model in this paper which has thus far not appeared in print. My purpose in presenting one example each of models in these two areas which are not being extensively cultivated, is to interest the college mathematics major in the value of applying his knowledge to these relatively novel contexts. Such work sooner or later will not only advance considerably certain interesting areas in the field of applied mathematics but will also serve at the same time to contribute towards the solution of problems which are of decided importance to specialties which are only beginning their development today.

II. An Analysis Of Conflict. In all conflict situations we deal with an opposition between two forces, F_1 and F_2. These may be persons, groups, institutions, nations, or coalitions. The nature of the forces in opposition does not affect the structure of a conflict situation and therefore does not change substantially the nature of the analysis required. What does affect that analysis, however, is first, the nature of the physical or psychological inputs and outputs which determine the status of the conflict at any given point in time, and second, the rate of inflow of these inputs and outflow of the outputs. Since the typical conflict situation of large-scale significance is war, let us provide an analysis of the structure of one type of combat situation, which reveals the manner in which inputs and outputs and their flows serve to determine the outcome of the conflict.

Let us assume a battle situation in which two forces, F_1 and F_2, are locked in combat. The one significant set of parameters which will determine the outcome, are the following:

Let ρ_1^m represent the replacement rate of men for force F_1.
Let ρ_2^m represent the replacement rate of men for force F_2.
Let ρ_1^d represent the replacement rate of materiel for F_1.
Let ρ_2^d represent the replacement rate of materiel for F_2.
Let f_1 represent the firepower per man per unit time in F_1.
Let f_2 represent the firepower per man per unit time in F_2.

Let f_c represent the average maximum value of f_i.
Let μ_1^m represent the man-loss coefficient per unit of materiel spent by F_1.
Let μ_2^m represent the man-loss coefficient per unit of materiel spent by F_2.
Let μ_1^d represent the materiel-loss coefficient per unit of materiel spent by F_1.
Let μ_2^d represent the materiel-loss coefficient per unit of materiel spent by F_2.

Assume all the parameters above are constant. Letting M_1 and M_2 represent the amount of men which forces F_1 and F_2 have *at any time*, t, and D_1 and D_2 the corresponding amounts of materiel *at any time*, t, we have

$$\frac{dM_1}{dt} = \overset{m}{\rho_1} - \overset{m}{\mu_1} M_2 f_2. \tag{1}$$

$$\frac{dD_1}{dt} = \overset{d}{\rho_1} - \overset{d}{\mu_1} M_2 f_2 - M_1 f_1. \tag{2}$$

$$\frac{dM_2}{dt} = \overset{m}{\rho_2} - \overset{m}{\mu_2} M_1 f_1 \tag{3}$$

$$\frac{dD_2}{dt} = \overset{d}{\rho_2} - \overset{d}{\mu_2} M_1 f_1 - M_2 f_2. \tag{4}$$

The quantities M_1f_1 and M_2f_2 represent the total firepower available for forces, F_1 and F_2, respectively. The expressions M_1 and M_2 are the men available at time, t. The total firepower is equal to the expenditure of some variable number of units of materiel.

The maximum average value of this expenditure is $D_i/M_i = f_c$. D_i and M_i may be of sufficient magnitude to be numerically greater than f_c, but f_c is a maximum constraint imposed upon D_i/M_i in virtue of the physical limitations governing the amount of materiel a man can expend. For various reasons, however, as in the case of insufficient available materiel, D_i/M_i may be less than f_c.

Equations (2) and (4) represent the fact that materiel is lost in two ways: materiel which is destroyed by the enemy and materiel which is lost through utilization in combat. It should be borne in mind that the differential, dM_i, represents a change in the value of M_i due to replacements *plus* change in the number of men who are rendered *hors de combat* (killed, wounded, or taken prisoner) at time dt. Equations (1) and (3), (2) and (4) are structurally symmetrical and reflect the interaction of F_1 and F_2 upon each other. The solution of (1) will therefore be isomorphic with the solution of (3). Suppose we wish to solve for (1) to obtain M_1 as a function of t. We can express (1) as

$$\dot{M}_1 = \overset{m}{\rho_1} - \overset{m}{\mu_1} f_2 M_2. \tag{5}$$

Then

$$\ddot{M}_1 = -\overset{m}{\mu_1} f_2 \dot{M}_2 = -\overset{m}{\mu_1} f_2 [\overset{m}{\rho_2} - \overset{m}{\mu_2} f_1 M_1] = \ddot{M}_1 - A_1 M_1 = -A_2, \tag{6}$$

where $A_1 = \mu_1^m \mu_2^m f_1 f_2$ and $A_2 = \mu_1^m f_2 \rho_2^m$.

Let $\dot{M}_1 = p$. Then

$$(7)\qquad \text{(a)}\ \ddot{M}_1 = \dot{p} = \frac{dp}{dt} = \frac{dp}{dM_1}\cdot\frac{dM_1}{dt} = p\,\frac{dp}{dM_1} = A_1M_1 - A_2,$$

and

$$\text{(b)}\ p(dp) = (A_1M_1 - A_2)dM_1.$$

Integrating, we have

$$(8)\qquad \text{(a)}\ \int p\,dp = \int A_1M_1dM_1 - \int A_2dM_1,$$

or

$$\text{(b)}\qquad p^2/2 = \frac{1}{2}A_1M_1^2 - A_2M_1 + \beta_1 = \frac{1}{2}\left(\frac{dM_1}{dt}\right)^2.$$

To evaluate β_1 we note that when $t=0$ the value of M_1 is $M_1(0)$. Therefore from equation (1) we have

$$(9)\qquad M_1(0) = \overset{m}{\rho}_1 - \overset{m}{\mu}_1 f_2M_2(0).$$

Hence,

$$(10)\qquad \begin{aligned}\beta_1 = {} & \frac{1}{2}\left\{(\overset{m}{\rho}_1)^2 - 2\overset{m}{\mu}_1f_2\overset{m}{\rho}_1M_2(0) + (\overset{m}{\mu}_1)^2f_2^2[M_2(0)]^2\right\}\\ & - \frac{A_1}{2}[M_1(0)]^2 + A_2M_1(0).\end{aligned}$$

From (8)(b) we have

$$(11)\qquad \frac{dM_1}{dt} = (A_1M_1^2 - 2A_2M_1 + \beta_1)^{1/2} = A^{1/2}.$$

But (11) is of the form $dX/(C_1X^2 - C_2X + C_3)^{\frac{1}{2}}$ since β_1, by definition, is a sum of fixed parameters.

Integrating (11) we have

$$\int \frac{dM_1}{(A_1M_1^2 - 2A_2M_1 + \beta_1)^{1/2}} = \int dt$$

which equals

$$(12)\qquad \frac{1}{(A_1)^{1/2}}\ln\left[A^{1/2} + M_1A_1^{1/2} - \frac{A_2}{A_1^{1/2}}\right] = t + \beta_2.$$

We can evaluate β_2 by noting that when $t=0$, $M_1 = M_1(0)$.

$$(13)\qquad \therefore\ \beta_2 = \frac{1}{(A_1)^{1/2}}\ln\left[A(0)^{1/2} + A_1^{1/2}M_1(0) - \frac{A_2}{A_1^{1/2}}\right].$$

We then have

$$\begin{aligned} t &= (A_1)^{-1/2} \ln \left[A^{1/2} + M_1 A_1^{1/2} - A_2 A_1^{-1/2}\right] \\ &\quad - (A_1)^{-1/2} \ln \left[A(0)^{1/2} + A_1^{1/2} M_1(0) - A_2 A_1^{-1/2}\right]. \end{aligned} \tag{14}$$

In (14) all quantities are known except M_1 and t. If we therefore wish to determine the magnitude of M_1 at any time, t, we need only substitute the value of t to obtain the desired result.

By a general type of analysis of this sort, extended to equations (2) through (4), we can solve for the properties of the system, obtaining oscillatory phenomena with respect to attrition. This entire type of analysis could now be repeated assuming the ρ's variable and the μ's variable under specified conditions. Further than this we could specify a series of functions, ρ_i and μ_i, for the *various* kinds of men and materiel which a modern army needs. We could likewise allow for a set, f_i, that would depend upon different combinations of men and materiel. But even this would be insufficiently elastic for it provides no *modus operandi* for retreat. But were a mechanism provided for attack and retreat the essential logistic quality of a military campaign would still be lacking. In planning to win a war a force decides that only if a certain minumum area is occupied will *final* victory be certain. Every fighting army on one side has as its objective not only to win a series of localities, but if we call these localities, $L_1, L_2, \cdots, L_r$, a necessary condition is that when L_1 is gained the winning force remaining plus its various replacement rates for men and materiel be sufficient to meet and vanquish the enemy force at L_2, in the light of the replacement rates for various kinds of men and materiel at the disposal of the enemy force at L_2, and likewise for $L_3, \cdots, L_r$. Thus a sort of Gestalt logistic which involves a series of campaigns both in space and time, which are serially dependent upon each other, is required. Even this is still far from sufficient to approximate military realities unless one side provides for (a) intangibles—such as various kinds of morale function on both sides due to losses, enemy propaganda, etc.; (b) surprise strategies; (c) new weapons; (d) differential aspects of military intelligence; (c) mistakes from one's own forces, etc. However, all such factors can in principle be introduced. It remains to effect the necessary treatment and we can look forward to this possibility for innovations in applied mathematics as time goes on.

III. The spread of behavior by steady and changing sources. All novel behavior (b_n) which is spreading in a population, may be spreading strictly on an atomistic basis, that is, from person to person only, or it may be behavior which is propagated strictly from a central source, that is, one person who moves around and transmits b_n to those whom he contacts, such as, for instance, a healer employing the "laying on of hands" in a revivalist atmosphere or a radio commentator changing his listeners' attitudes overnight by the information he alone can provide. The first type of propagation can be called *atomistic diffusion*, while the second type can be called *gross diffusion*. Where b_n spreads in both ways we can legitimately speak of *mixed diffusion*. An interesting problem which arises in this connection, and has importance in many contexts is that of calculating the total number of persons who will show the novel behavior (b_n) at time, t, when it is spreading by mixed diffusion.

In order to approach a problem of this type we shall introduce some simplifying but somewhat realistic assumptions. These are:

(1) Novel social behavior, b_n, is spreading from a gross or central source, S_G.

(2) Due to the charismatic influence of the central source only those converts to b_n from this source will transmit the new behavior. What this means concretely in our second example is that the novel behavior is adopted and spread only if it has been picked up from the charismatic, central source. If it is picked up from a convert, rather than S_G, *it is retained but is not salient enough to prompt the receiver to spread it, himself.*

(3) The time function of contact with the population by the central source, which we shall call $G(t)$, is a constant, that is to say, $G(t)=K$.

(4) The percentage of $G(t)$ which adopts b_n from S_G may be constant or variable and is given by $r(t)\leq 1$. Thus $r(t)G(t)$ will always give the effective number of converts at time, t.

(5) We either assume (1) that for stated intervals of time, S_G does not make any repetitious contacts with his potentially convertible population or (2) that $r(t)G(t)$ is an empirically establishable function which already represents the new and nonrepeated converts at time, t, so that the effect of repetitious contacts by S_G is already weeded out.

(6) We also either assume (1) that transmitters of b_n who have adopted it from S_G do not make repetitious contacts with their prospective converts or (2) that the sociality function which gives the variation in their contacts over time already reflects the time rate of change of only those of their social contacts who have not been met with previously.

(7) We designate the sociality function of effective converts mentioned in (6) above, as $\phi(t)$. This is a phase function which is identical for all *atomistic transmitters*, regardless of the actual calendar value of t at which they may begin their own conversion activity.

In the light of the preceding seven assumptions, we can set forth a tableau which will reflect the increments of b_n at time, t. Let Δ_i represent the increment at $t=i$ and $N(k)$, the total number of converts at $t=k$, created by the process of mixed diffusion described above. We then have

$$\begin{aligned}
&\Delta_0 = 1\\
&\Delta_1 = r(1)K\\
&\Delta_2 = r(2)K + r(1)K\phi(1)\\
(1)\qquad &\Delta_3 = r(3)K + r(1)K\phi(2) + r(2)K\phi(1)\\
&\Delta_4 = r(4)K + r(1)K\phi(3) + r(2)K\phi(2) + r(3)K\phi(1)\\
&\quad\vdots\\
&\Delta_i = r(i)K + r(1)K\phi(i-1) + r(2)K\phi(i-2) + \cdots + r(i-1)K\phi(1)
\end{aligned}$$

from which it may be easily seen that

$$\begin{aligned}(2)\qquad N(k) = K\Bigg[\sum_{1}^{k} r(t) + r(1)\sum_{1}^{k-1}\phi(t) + r(2)\sum_{1}^{k-2}\phi(t) + \cdots \\ + r(n)\sum_{1}^{k-n}\phi(t) + \cdots + r(k-1)\phi(1)\Bigg] + 1\end{aligned}$$

which in turn is equivalent to

$$(3)\qquad N(k) = K\left\{ \sum_{1}^{k} r(t) + \sum_{j=1}^{k-1}\left[r(j)\sum_{1}^{k-j}\phi(t)\right]\right\} + 1.$$

This same process of mixed diffusion can clearly be studied under other assumptions. Among these would be the following: (1) Both types of converts can transmit b_n; (2) $G(t)$ and $r(t)$ are both constant, both variable or one constant and one variable; (3) $\phi(t)$ is a constant or reflects concurrent phase effects, that is to say, all transmitters for any value of t are at the same phase of $\phi(t)$ for the same value of calendar time; (4) S_G and other transmitters cease to transmit at the same or different times; (5) converts have identical or variable circles of acquaintance and in the case of the latter assumption cease to be effective transmitting agents after variable periods of time of effective conversion; and so on and so forth.

Diffusion models of this sort, if explored by mathematical methods which make provision for a variety of realistic psychological and sociological parameters, can increase our knowledge concerning the spread of new behavior considerably. The phenomena of social diffusion comprise a context of inquiry which, in comparison with many others of contemporary interest, have been only scantily explored. Such phenomena, apart from the possible social utility they may come to provide sooner or later, enlarge the concerns of applied mathematics. The undergraduate mathematics major venturing into this area will find the problems intriguing and the scope for ingenuity and creativity rewarding, indeed.

References

1. S. C. Dodd, Mimeographed reports from the Washington Public Opinion Laboratory, Project Revere, 1952.

2. A. Rapoport, Spread of information through a population with sociostructural bias: III. Suggested Experimental Procedures, Bull. Math. Biophys., 16 (1954) 75–81.

3. N. Rashevsky, Mathematical theory of human relation: an approach to a mathematical biology of social phenomena, Bloomington, Indiana: Principia Press, 1947, p. 202.

4. L. F. Richardson, Generalized foreign politics, Brit. J. Psychol., Monog. Suppl., 1939, No. 23.

5. H. Winthrop, A kinetic theory of socio-psychological diffusion, Journal of Social Psychology, 22 (1945) 31–60.

6. ———, A theory of behavioral diffusion. A contribution to the mathematical biology of social phenomena. Unpublished thesis submitted to the faculty of the New School for Social Research in partial fulfillment of the requirements for the degree of Doctor of Philosophy, 1953, p. 357.

PROBLEMS AND SOLUTIONS

EDITED BY ROBERT E. HORTON, Los Angeles City College

Readers of this department are invited to submit for solution problems believed to be new that may arise in study, in research, or in extra-academic situations. Proposals should be accompanied by solutions, when available, and by any information that will assist the editor. Ordinarily, problems in well-known textbooks should not be submitted. Solutions should be submitted on separate, signed sheets. Figures should be drawn in India ink and exactly the size desired for reproduction. Send all communications for this department to Robert E. Horton, Los Angeles City College, 855 North Vermont Avenue, Los Angeles 29, California.

PROPOSALS

544. *Proposed by Huseyin Demir, Middle East Technical University, Ankara, Turkey.*

Solve the cryptarithm (alphametic)

$$ONE+TWO+SIX=NINE$$

in the base 10, with the following conditions:

a) $ONE<TWO<SIX$

b) $2|TWO$, $6|SIX$, $9|NINE$ where $a|b$ means "a divides b."

545. *Proposed by C. Stanley Ogilvy, Hamilton College, New York.*

A curve is given by the parametric equations $x=2t/(1+t^2)$, $y=(1-t^2)/(1+t^2)$. What is the geometric meaning of the parameter t?

546. *Proposed by D. Rameswar Rao, Secunderabad, India.*

Solve in integers the equation $x^2-y^2=X^6-Y^6$.

547. *Proposed by Daniel I. A. Cohen, Midwood High School, Brooklyn, New York.*

Prove that the arithmetic mean of the nth powers of a set of numbers is never less than the nth power of the arithmetic mean of the numbers.

548. *Proposed by Leo Moser, University of Alberta.*

Show that if a and c are positive reals and b and d positive integers with $b\geq d$, then $(a-1)b\geq(c-1)d$ implies $a^b\geq c^d$.

549. *Proposed by Murray S. Klamkin, SUNY at Buffalo, New York.*

The solution of the Clairaut equation $y=xy'+F(y')$ is obtained by setting $y'=c$ which gives $y=cx+F(c)$. Determine the most general first order differential equation in which the solution can be obtained in this manner.

550. *Proposed by Brother U. Alfred, St. Mary's College, California.*

The following is a set of equations in which n consecutive integers and $n+1$ consecutive integers have equal sums of squares.

$$3^2+4^2=5^2$$

$$10^2+11^2+12^2=13^2+14^2$$

$$21^2+22^2+23^2+24^2=25^2+26^2+27^2$$

a) What would be the first term on the left for the case of n and $n-1$ consecutive integers?
b) Prove that your result holds in the general case.

SOLUTIONS

Late Solutions

516, 518. *Dee Fuller, University of Georgia.*

518, 519. *Alan Sutcliffe, Knottingley, Yorkshire, England*

N-Sided Box

523. [September 1963] *Proposed by Gilbert Labelle, Université de Montréal, Canada.*

It is well known that if, from a square of side A, we cut at each of its vertices a square of side $A/6$ and fold the resulting figure to form an open box, the box will have maximum volume. What must we do if the box of maximum volume is to be cut from a regular n-gon instead of a square?

I. *Solution by Robin S. McDowell, Los Alamos, New Mexico.*

An n-gon of side A can be folded into a box if on each side two cuts are made perpendicular to that side and at a distance b from each vertex, each cut to be extended until it meets the cut from the other side of the same vertex. The base of the box is then an n-gon similar to the original n-gon but with sides $A-2b$, and since the interior angles of the n-gon are $\pi(1-2/n)$, the height of the box is $b \cot(\pi/n)$. Using the usual formula for the area of a regular n-gon, we obtain for the volume of the box $V=\frac{1}{4}nb(A-2b)^2 \cot^2(\pi/n)$, $dV/db =\frac{1}{4}n(A-2b)(A-6b) \cot^2(\pi/n)$, and $b=\frac{1}{2}A$ or $b=A/6$; obviously only the second result is correct. Thus the cuts are started at a distance $A/6$ from each vertex, as in the case of the square, regardless of the number of sides.

The volume of the box thus formed is $V=(1/54)nA^3 \cot^2(\pi/n)$, and as $n \rightarrow \infty$, V approaches a limiting value of $(4\pi/27)r^3$, where r is the radius of the circumscribed circle.

II. *Solution by Harry W. Hickey, Alexandria, Virginia.*

Given a regular n-gon of side A, start cuts at distance x from each end of every side, to form a box. The height of the box is proportional to x, the base is a regular n-gon of side $(A-2x)$, and the volume is proportional to $x(A-2x)^2$, which is independent of n. So if setting $x=A/6$ will maximize the volume for $n=4$, it will do so for any other n.

Also solved by Marc Aronson, University of Florida; Dermott A. Breault, Sylvania Electronics Systems, Waltham, Massachusetts; B. A. Hausmann, West Baden College, Indiana; Felix Lo, St. Mary's College, California; Michael J. Pascual, Watervliet Arsenal, New York; C. W. Trigg, San Diego, California; Lowell Van Tassel, San Diego, California; and the proposer.

Van Tassel pointed out that the problem has appeared in *The Mathematics Student Journal*, May, 1961, Volume 8, Number 4.

A Triangular Inequality

524. [September 1963] *Proposed by L. Carlitz, Duke University.*

Let I denote the incenter, R the radius of the circumcircle, and r the radius of the inscribed circle of the triangle ABC. Show that $6r \leqq AI+BI+CI \leqq 3R$. Moreover, equality holds in either place if and only if ABC is equilateral.

Solution by W. J. Blundon, Memorial University of Newfoundland.

To prove the first inequality, we apply the Erdös-Mordell inequality to the incenter I of the triangle $A_1A_2A_3$ to obtain

$$IA_1 + IA_2 + IA_3 \geq 2(r + r + r) = 6r.$$

Since the orthocenter H of the triangle $A_1A_2A_3$ is the incenter of the pedal triangle $H_1H_2H_3$ and since the circumradius of this triangle is $\frac{1}{2}R$, the second inequality is equivalent to $HH_1+HH_2+HH_3 \leq \frac{3}{2}R$. To prove this we use the Euler inequality $R \geq 2r$ and also the relation $HA_1+HA_2+HA_3=2r+2R$. Applying the Erdös-Mordell inequality to the point H of the triangle $H_1H_2H_3$, we have

$$HH_1 + HH_2 + HH_3 \leq \tfrac{1}{2}(HA_1 + HA_2 + HA_3) = r + R \leq \tfrac{3}{2}R.$$

Also solved by Leon Bankoff, Los Angeles, California, and the proposer.

Unequal Partitions

525. [September 1963] *Proposed by Francis L. Miksa, Aurora, Illinois.*

It is known that the magic total, T, in a magic square of order n is $T=n(n^2+1)/2$. In how many ways can T be partitioned into n unequal parts, those parts to be chosen only from the consecutive integers from 1 to n^2?

Solution by Alan Sutcliffe, Knottingley, Yorkshire, England.

1. The number of ways, $p_j(m, n)$, that m can be partitioned into n unequal parts, none greater than j, is equal to the coefficient of t^m in

$$t^{\frac{1}{2}n(n+1)} \cdot \frac{(1-t^j)(1-t^{j-1})(1-t^{j-2})\cdots(1-t^{j-n+1})}{(1-t^n)(1-t^{n-1})(1-t^{n-2})\cdots(1-t)}.$$

This is given as Problem 7 in Chapter 6 of "An Introduction to Combinatorial Analysis" by John Riordan (New York, 1958), and the following proof is based on that outlined there.

Proof. $p_j(m, n)$ is the coefficient of a^nt^m in the expansion of

$$G_j(t, a) = (1 + at)(1 + at^2)(1 + at^3) \cdots (1 + at^j),$$

since each partition will contribute just one to the coefficient. They will be partitions of m since we have the coefficient of t^m; they will be partitions into n parts since we have the coefficient of a^n; they will be unequal partitions since each bracket can contribute only once to each partition, and no part will be greater than j since the largest term is $1+at^j$.

Now

$$\begin{aligned}(1 + at)G_j(t, at) &= (1 + at)(1 + at^2)(1 + at^3) \cdots (1 + at^j)(1 + at^{j+1}),\\ &= (1 + at^{j+1})G_j(t, a).\end{aligned}$$

Let

$$G_j(t,\,a) = \sum_{n=0}^{n=j} U_j(t,\,n)a^n,$$

so that $U_j(t,\,n)$ is the coefficient of a^n in the expansion of $G_j(t,\,a)$, and hence $p_j(m,\,n)$ is the coefficient of t^m in $U_j(t,\,n)$. Then we have

$$(1+at)\sum_{n=0}^{n=j} U_j(t,\,n)a^nt^n = (1+at^{j+1})\sum_{n=0}^{n=j} U_j(t,\,n)a^n.$$

Equating coefficients of a^n in this gives

$$U_j(t,n)t^n + tU_j(t,n-1)t^{n-1} = U_j(t,n) + t^{j+1}U_j(t,n-1).$$

$$\therefore\ U_j(t,n) = t^n\cdot\frac{(1-t^{j-n+1})}{(1-t)}\cdot U_j(t,n-1),$$

$$= t^nt^{n-1}\cdot\frac{(1-t^{j-n+1})(1-t^{j-n+2})}{(1-t)(1-t^2)}\cdot U_j(t,n-2).$$

This process may be continued until we have

$$U_j(t,n) = t^{n+(n-1)+\cdots+1}\cdot\frac{(1-t^{j-n+1})(1-t^{j-n+2})\cdots(1-t^j)}{(1-t)(1-t^2)\cdots(1-t^n)}\cdot U_j(t,\,0).$$

Since $U_j(t,\,0)=1$, this proves the result.

2. In the problem as stated, we have $j=n^2$ and $m=\frac{1}{2}n(n^2+1)$. Hence the number of different lines that can appear in nth order magic squares is the coefficient of $t^{\frac{1}{2}n^2(n-1)}$ in the expansion of

$$\frac{(1-t^{n^2})(1-t^{n^2-1})(1-t^{n^2-2})\cdots(1-t^{n^2-n+1})}{(1-t^n)(1-t^{n-1})(1-t^{n-2})\cdots(1-t)}.$$

The values of this coefficient for $n=1$, 2, 3, 4 and 5 are 1, 2, 8, 86 and 1394.

A Prime Generator

527. [September 1963] *Proposed by Sidney Kravitz, Dover, New Jersey.*

It is known that $f(n)=n^2-n+41$ yields prime numbers for $n=1, 2, \cdots, 40$. Find a sequence of 40 consecutive values of n for which $f(n)$ is composite.

Solution by Lawrence A. Ringenberg, Eastern Illinois University.

Let $a_1=41, a_2=43, \cdots, a_{40}=1601$, i.e. let $a_k=k^2-k+41$ for $k=1, 2, \cdots, 40$. Let $A=a_1\cdot a_2\cdot a_3\cdots a_{40}$. Then for $k=1, 2, 3, \cdots, 40$, we have

$$f(A+k) = (A+k)^2 - (A+k) + 41 = A^2 + (2k-1)A + a_k.$$

Since a_k divides A, it follows that A is composite.

It is easy to see how this result may be extended to find a sequence (of consecutive values of n) of arbitrary length for which $f(n)$ is composite.

Also solved by Murray Berg, Standard Oil Company of California; W. J. Blundon, Memorial University of Newfoundland; L. Carlitz, Duke University; Gloria Gottesfeld, AVCO, Wilmington, Massachusetts; B. A. Hausmann, West Baden College, Indiana; Erwin Just and Norman Schaumberger, Bronx Community College, New York; Samuel S. Kutler, St. John's College, Maryland; Gilbert Labelle, Université de Montréal; Barry Litvack, University of Michigan; M. Raghavachari, University of California, Berkeley, California; Alan Sutcliffe, Knottingley, Yorkshire, England; Dale Woods, Northeast Missouri State Teachers College; and the proposer.

Hausmann noted that a smaller value of A, namely $A=f(1)\cdot f(2)\cdots\cdot f(20)$ can be used if we take for k the 40 consecutive numbers $-19, -18, \cdots, +20$ as $f(-k)=f(k+1)$. Also if we use the original A and let $k=-39, -38, \cdots, +40$, then $f(A+k)$ will not be prime for at least 80 consecutive numbers.

Litvack commented as follows. One of the most interesting theorems in elementary number theory is the theorem: *There are arbitrarily large gaps in the sequence of primes.* This theorem first appeared in the book, "Théorie des Nombres," by Edouard Lucas in 1891. The theorem can be restated as follows: *If $f(n)=n$ then one can find a sequence of k consecutive values of n for which $f(n)$ is composite.* In view of this restatement, one could consider the following generalization: If $f(n)=an^2+bn+c$, with obvious restrictions on the coefficients, then one can find a sequence of k consecutive values of n for which $f(n)$ is composite. To prove this let $x=f(1)\cdot f(2)\cdots f(k)$. Then $f(x+i)=ax^2+2aix+bx+f(i)$, and $f(x+i)$ is divisible by $f(i)$ for $i=1, 2, \cdots, k$. The sequence of k consecutive values is $x+1, x+2, \cdots, x+k$. This completes the proof.

$f(n)$ in the theorem may be replaced by an arbitrary polynomial, with the obvious restrictions on the coefficients. A similar proof holds. This theorem gives a solution to Problem 527 as a special case. One question comes to mind: Are there any other classes of functions or sequences for which this property holds?

A Pythagorean-like Relation

528. [September 1963] *Proposed by Dewey C. Duncan, East Los Angeles College.*

In a right trihedral tetrahedron the square of the area of the face opposite the right trihedral angle is equal to the sum of the square of the areas of the other three faces. (A right trihedral angle has three right angles for its face angles.)

Solution by V.F.I., University of California, 1927.

This is a particular case of Carnot's theorem (if I am not mistaken); namely, if a, b, c, d, are the areas of the faces of *any* tetrahedron, then

$$a^2 = b^2 + c^2 + d^2 - 2bc\cos(b, c) - 2cd\cos(c, d) - 2bd\cos(b, d),$$

in which (b, c) represents the interior dihedral angle between the faces with areas b and c, respectively, etc. I "discovered" this theorem several years ago and was greatly disappointed that Carnot had beaten me to it.

Proof. By projecting the faces, orthogonally, b, c, d, upon the face a, one obtains

$$b\cos(b, a) + c\cos(c, a) + d\cos(d, a) = a$$

and, similarly,

$$a\cos(a, b) + c\cos(c, b) + d\cos(d, b) = b$$
$$a\cos(a, c) + b\cos(b, c) + d\cos(d, c) = c$$
$$a\cos(a, d) + b\cos(b, d) + c\cos(c, d) = d.$$

Substituting the values of $\cos(a, b)$, $\cos(a, c)$, $\cos(a, d)$, obtained from the last three equations, into the first equation, we get the desired result. The solution of the above proposed problem is obtained by setting the three indicated cosines equal to zero. Incidentally, the above four equations are homogeneous in a, b, c, d; therefore,

$$\begin{vmatrix} -1 & \cos(b, a) & \cos(c, a) & \cos(d, a) \\ \cos(a, b) & -1 & \cos(c, b) & \cos(d, b) \\ \cos(a, c) & \cos(b, c) & -1 & \cos(d, c) \\ \cos(a, d) & \cos(b, d) & \cos(c, d) & -1 \end{vmatrix} = 0,$$

and yields the relationship among the interior dihedral angles of any tetrahedron, as is mentioned in Salmon's "Solid Analytical Geometry."

Also solved by Marc Aronson, University of Florida; Merrill Barneby, University of North Dakota; Dermott A. Breault, Sylvania Electronics Systems, Waltham, Massachusetts; Brother T. Brendan, St. Mary's College, California; Harry W. Hickey, Alexandria, Virginia; Gilbert Labelle, Université de Montréal; Charles Lewis, Swarthmore College, Pennsylvania; C. C. Oursler, Belleville, Illinois; C. W. Trigg, San Diego, California; and the proposer.

Center of Curvature

529. [September 1963] *Proposed by Huseyin Demir, Middle East Technical University, Ankara, Turkey.*

A cycloid (cardioid) rolls on a straight line without sliding. Prove that the locus of the center of curvature of the curve at the point of tangency is a circle (ellipse).

Solution by P. R. Nolan, Department of Education, Dublin, Ireland.

Cycloid. Taking the regular case and putting $\omega t = \alpha$ we have

$$x = a(\alpha - \sin \alpha), \qquad y = a(1 - \cos \alpha).$$

By the usual methods, the arc length from the origin is given by

(i) $$S_\alpha = 4a(1 - \cos \alpha/2)$$

and the radius of curvature by

(ii) $$P_\alpha = 4a \sin \alpha/2 \cdot$$

Now if one arch of the cycloid rolls once along the y axis, the coordinates of the center of curvature at the point of tangency will be (P_α, S_α). Therefore from (i) and (ii), its locus is

$$x^2 + (y - 4a)^2 = (4a)^2$$

which is a *semicircle*, negative values of $x(P)$ not being admissible, unless we consider the next arch to roll back along the y axis to complete the locus-circle.

Cardioid. In polar coordinates

$$r = a(1 - \cos \theta).$$

As before, this gives

$$\text{(i)} \qquad S_\theta = 4a(1 - \cos\theta/2)$$

and

$$\text{(ii)} \qquad P_\theta = (4a/3)\sin\theta/2.$$

Now if the cardioid rolls once along the upper edge of the x axis, the coordinates of the center of curvature at the point of tangency will be (S_θ, P_θ). Therefore from (i) and (ii), its locus is

$$(x - 4a)^2 + 9y^2 = (4a)^2$$

which is the upper *half of an ellipse*, negative values of $y(P)$ not being admissible, unless the same cardioid is also rolled along the lower edge of the axis.

Also solved by the proposer.

Comment on T55

T55. [November 1962] *Comment by Josef Andersson, Vaxholm, Sweden.*

In the first place the third term of y must be $5x^9/6$ and the following two terms decrease in importance as integers. Then if we let

$$y_1 = (6x^{11} + 3\cdot 11x^{10} + 11\cdot 5x^9 - 3\cdot 11x^3 + 5x)/66 = N/66,$$

it is sufficient to verify that $N \equiv 0 \pmod{66}$ for integral values of x. But, then $N \equiv 0 \pmod 2$, $N \equiv x^9 - x = (x^3)^3 - x \equiv x^3 - x \equiv 0 \pmod 3$ and $N \equiv 6(x^{11} - x) \equiv 0 \pmod{11}$. Besides suppose $\phi_n(x)$, $n = 1, 2, \cdots$ are the Bernoulli polynomials defined by Maclaurin's expansion

$$\frac{e^{xt} - 1}{e^t - 1} = x + \phi_1(x)\frac{t}{1!} + \phi_2(x)\frac{t^2}{2!} + \cdots$$

we verify that

$$\text{(1)} \qquad y(x) = \phi_{10}(x) + x^{10}$$

if $x = 2, 3, \cdots$. It is clear that $\phi_n(x) = 1^n + 2^n + \cdots + (x-1)^n$ where in particular $y(x) = \sum_1^x r^{10}$ is integral because of its characteristics.

Since $\phi_n(1) = 0$, the sum $\sum_1^1 r^{10}$ can be identified with $y(1)$. But do $y_1(0) = 0$ and $\sum_1^0 r^{10}$ designate the same number? And what about $\sum_1^{-1} r^{10}$, $\sum_1^{-2} r^{10}$, $\cdots$?

Therefore it appears that the solution based on (1) is applicable only to positive values of x. This can be shown by applying the general relation $\phi_n(1-x) = (-1)^{n-1}\phi_n(x)$, which we prove easily, that $y(x) = \phi_n(x) + x^n$ remains integral for x integral and non-positive.

QUICKIES

Q331. Solve in integers the Diophantine equation $2x(x+1)=y(y+2)$. [*Submitted by Monte Dernham.*]

Q332. Factor x^5-5x^2+2.

[*Submitted by M. S. Klamkin*]

Q333. For what integers can we find two others so that the three form a Pythagorean triplet $(a^2+b^2=c^2)$?

[*Submitted by Sylvan Eiman*]

Q334. Determine the ratio

$$\sum_{r=0}^{n} r\binom{n}{r}^p \div \sum_{r=0}^{n} \binom{n}{r}^p.$$

[*Submitted by M. S. Klamkin*]

Q335. Let ϕ denote Euler's totient and $\phi^m(x)=\phi[\phi^{m-1}(x)]$ where x is a positive integer. Let n be the least positive integer for which $\phi^n(x)=1$. Find all x such that $[\phi^w(x), \phi^v(x)]=1$, $w\neq v$ and $w=0, 1, 2, \cdots, n$, $v=0, 1, 2, \cdots, n$.

[*Submitted by Gilbert Labelle*]

(Answers on page 83)

TRICKIES

A trickie is a problem whose solution depends upon the perception of the key word, phrase, or idea rather than upon a mathematical routine. Send us your favorite trickies.

T58. Prove that if

$$\sum_{k=1}^{n} k! = m^2,$$

then

$$\sum_{k=1}^{m} k! = n^2.$$

[*Submitted by David L. Silverman*]

T59. Determine a function $\phi(x, y)$ such that the set of points (x, y) satisfying $\phi(x, y)=0$ has area 1.

[*Submitted by M. S. Klamkin*]

T60. A person was directed to the downtown side of an unfamiliar subway station. He desired to get on the first car. Which end of the platform should he walk to, assuming that there are no signs, signal lights, or trains in the station to cue him?

[*Submitted by M. S. Klamkin*]

(Solutions on page 82)

CLASSROOM NOTE ON $e^{i\alpha}$

J. PRITCHETT, Rutherford College of Technology

An introductory lesson on the relationship between $\cos \alpha + i \sin \alpha$ and $e^{i\alpha}$ to a class who have just been shown De Moivre's Theorem:

"We have already shown that

$$(\cos \alpha + i \sin \alpha)(\cos \beta + i \sin \beta) = \cos (\alpha + \beta) + i \sin (\alpha + \beta)$$

and

$$(\cos \alpha + i \sin \alpha)^n = \cos n\alpha + i \sin n\alpha.$$

You will notice that if we write $\cos \alpha + i \sin \alpha = f(\alpha)$ that $\cos \alpha + i \sin \alpha$ has the property

$$(f(\alpha))^n = f(n\alpha) \quad \text{and} \quad f(\alpha) \times f(\beta) = f(\alpha + \beta).$$

Can anyone think of any other function which obeys this law?"

With a little bit of luck, someone mentions indices.

"So that we have, if

$$F(\alpha) = a^\alpha$$

then $F(\alpha)^n = F(n\alpha)$ and $F(\alpha) \times F(\beta) = F(\alpha + \beta)$.

Is it a coincidence that $\cos \alpha + i \sin \alpha$ and a^α obey the same laws, or are they the same function?

Let us examine $\cos \alpha + i \sin \alpha$ a little further.

If we differentiate it we get $i(\cos \alpha + i \sin \alpha)$, i.e., apart from the factor i, it appears to be unchanged by differentiation.

Can anyone remember any other function with this property?"

Again someone will remember e^x or even e^{ax}.

"If we differentiate $e^{i\alpha}$ we get $ie^{i\alpha}$.

It appears very likely in fact that $\cos \alpha + i \sin \alpha$ and $e^{i\alpha}$ are the same function."

SOLUTIONS

S58. $m = n = 1$ and $m = n = 3$ are obvious solutions of both equations. For any value of n other than 1 or 3 the terminal digit of $\sum$ is clearly 3, whence $\sum$ cannot be a square.

S59. $\phi(x, y) = |x-1| + |x+1| + |y-1| + |y+1| - 4 = 0$. This set consists of all points in and on the square with vertices at $(\pm 1, \pm 1)$.

S60. In the United States, he should walk in a direction such that the uptown tracks are kept on his left. Presumably, in London, it would be in the opposite direction. That is, if the trains run the same way as the automobile.

(Trickies on page 126)

ANSWERS

A331. If $y=D-1$ we may write $x^2+(x+1)^2=D^2$, a familiar Pythagorean relation where D is restricted to the odd denominators of the convergents to $\sqrt{2}$. Then $x=(\pm\sqrt{\{2D^2-1\}}-1)/2$.

Then the first five solutions in positive integers are x, $y=3$, 4; 20, 28; 119, 168; 696, 984; 4059, 5740; $\cdots$.

A332.

$$\begin{array}{rrrrrr}
x^5 & - x^4 & - x^3 & & & \\
& + x^4 & - x^3 & - x^2 & & \\
& & + 2x^3 & - 2x^2 & - 2x & \\
& & & - 2x^2 & + 2x & + 2 \\
\hline
x^5 & & & - 5x^2 & & + 2
\end{array}$$

so $x^5-5x^2+2=(x^2-x-1)(x^3+x^2+2x-2)$.

A333. All Pythagorean triplets can be formed from $a=m^2-n^2$, $b=2mn$, $c=m^2+n^2$ for $m>n\geq 1$. Since the differences of consecutive squares exhaust all odd numbers ≥ 3, and since, for $n=1$, $2m$ exhaust all even numbers ≥ 4, any integer ≥ 3 can be part of a triplet. The values 1 and 2 obviously fail. In particular given a odd, $b=(a^2-1)/2$, $c=(a^2+1)/2$; given a even, $b=a^2/4-1$, $c=a^2/4+1$.

A334. The ratio $n/2$ follows immediately from

$$\sum_{r=0}^{n} r\binom{n}{r}^p = \sum_{r=0}^{n} (n-r)\binom{n}{r}^p \quad \text{or from}$$

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

A335. Since $\phi^n(x)=0$, this implies that $\phi^{n-1}(x)=1$, 2. (Then $n=1$ implies $x=1$ or 2.) But $\phi^{n-1}(x)=1$, 2 implies that $\phi^{n-2}(x)=3$, 4, or 6, but only 3 is acceptable. (Then $n=2$ implies $x=3$.) Finally $\phi^{n-2}(x)=3$ has no solution for $\phi^{n-3}(x)$; then $x=1$, 2, or 3.

(Questions on page 126)

AN APPLICATION OF CONTINUANTS

S. L. BASIN, Sylvania Electronic Systems, Mountain View, California

1. Introduction. The following note on continuants (defined below) is intended to demonstrate their usefulness in studying recurrent sequences. In particular, the correspondence between a class of *simple continuants* and the well known Fibonacci sequence is demonstrated. Several continuant identities, discovered between the years 1853 and 1880, listed by T. Muir [1] are given together with their corresponding Fibonacci identities.

COMBINATORIAL MATHEMATICS

BY HERBERT J. RYSER, SYRACUSE UNIVERSITY

(CARUS MONOGRAPH #14)

. . . investigates the mathematical theory of the discrete. This subject began long ago and many of the early problems were studied primarily for their recreational or aesthetic appeal. But today combinatorial mathematics is recognized as a lively branch of modern algebra with important applications in pure and applied science. The main topics discussed by the author include an analysis of elementary concepts, the principles of inclusion and exclusion, recurrence relations, Ramsey's theorem, systems of distinct representatives, matrices of zeros and ones, orthogonal Latin squares, finite projective planes, Hadamard matrices, combinatorial designs, and difference sets. The Monograph makes readily available a wealth of material that is scattered in a variety of sources. General structure theorems are emphasized throughout. The proofs of a number of the theorems are new and every effort is made to interrelate and unify the varied topics under consideration. The Monograph also describes some of the major unsolved problems and contains a sizeable bibliography of the current literature.

Each member of the Association may purchase one copy of this Carus Monograph at the special price of $2.00. Orders should be addressed to:

Mathematical Association of America
SUNY at Buffalo (University of Buffalo)
Buffalo, New York 14214

Additional copies of Monograph 14 for members and copies for nonmembers may be purchased at $4.00 from:

JOHN WILEY AND SONS
605 Third Avenue
New York, New York 10016

MAA STUDIES IN MATHEMATICS

Volume I: Studies in Modern Analysis

R. C. Buck, Editor

Volume II: Studies in Modern Algebra

Edited by A. A. Albert

This series is intended to bring to the mathematical community expository articles at the collegiate and graduate level on recent developments in mathematics. Volume I was published in 1962 and Volume II in 1963.

Each member of the Association may purchase one copy of each volume of the Studies at $2 per volume. Orders with remittance should be addressed to: Mathematical Association of America, SUNY at Buffalo (University of Buffalo), Buffalo, New York 14214.

Additional copies and copies for non-members may be purchased at $4 per volume from Prentice-Hall, Inc., Englewood Cliffs, New Jersey 07631.

MATHEMATICS MAGAZINE Vol. 37, No. 2, Mar.–Apr., 1964