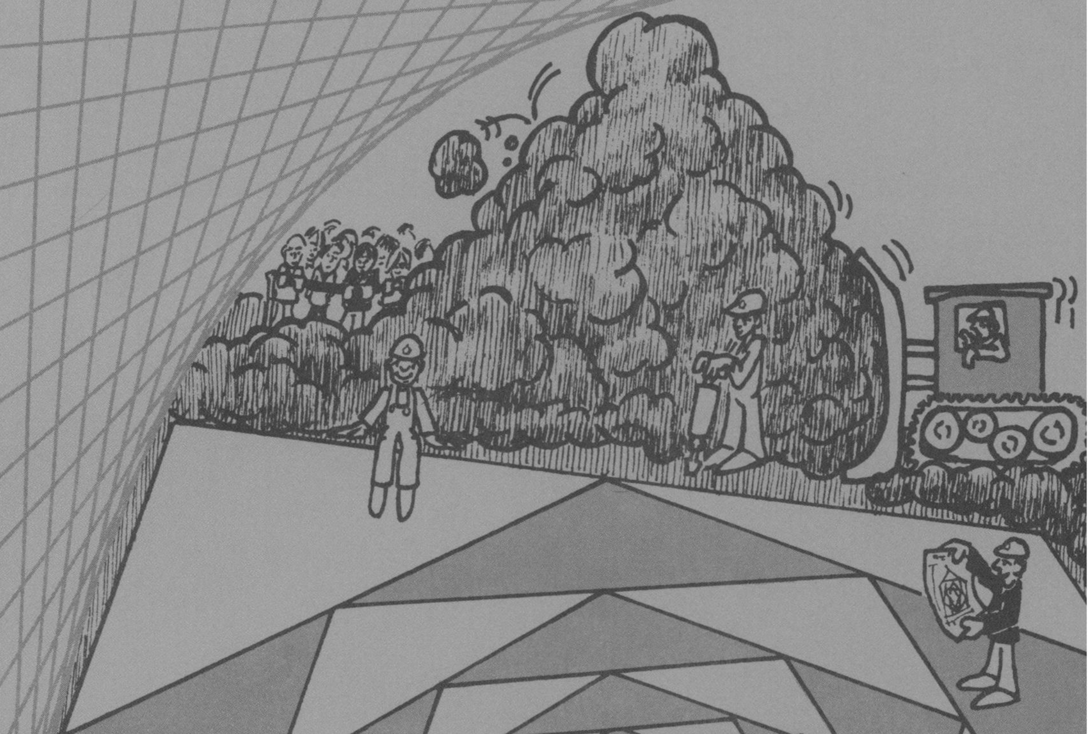
MATHEMATICS MAGAZINE



Vol. 52, No. 2
March, 1979

CODES • MULTIPLEX ADVANTAGES

ERDÖS PROBLEMS • WAYS TO VOTE

MATHEMATICS MAGAZINE

Volume 52, Number 2, March 1979

(ISSN—0025-570X)

COVER: A sequence of pentagons, each with vertices at the midpoints of the sides of the preceding pentagon, "converges" to a golden pentagon (see p. 102).

ARTICLES

NOTES

PROBLEMS

REVIEWS

NEWS AND LETTERS

EDITORIAL POLICY

Mathematics Magazine is a journal of collegiate mathematics designed to enrich undergraduate study of the mathematical sciences. The *Magazine* should be an inviting, informal journal emphasizing good mathematical exposition of interest to undergraduate students. Manuscripts accepted for publication in the *Magazine* should be written in a clear and lively expository style. The *Magazine* is not a research journal, so papers written in the terse "theorem-proof-corollary-remark" style will ordinarily be unsuitable for publication. Articles printed in the *Magazine* should be of a quality and level that makes it realistic for teachers to use them to supplement their regular courses. The editors especially invite manuscripts that provide insight into applications and history of mathematics. We welcome other informal contributions, for example, brief notes, mathematical games, graphics and humor.

Editorial correspondence should be sent to: Mathematics Magazine, Department of Mathematics, St. Olaf College, Northfield, Minnesota 55057. Manuscripts should be prepared in a style consistent with the format of *Mathematics Magazine.* They should be typewritten and double spaced on $8\frac{1}{2}$ by 11 paper. Authors should submit the original and one copy and keep one copy as protection against possible loss. Illustrations should be carefully prepared on separate sheets of paper in black ink, the original without lettering and two copies with lettering added; the printers will insert printed letters on the illustration in the appropriate locations.

Authors planning to submit manuscripts may find it helpful to obtain the more detailed statement of guidelines available from the editorial office.

ABOUT OUR AUTHORS

Paul Erdös ("Some Unconventional Problems in Number Theory"), Hungarian by birth and officially attached to the Hungarian Academy of Sciences, is in reality a mathematician to the world at large. His mathematical interests are catholic, including the education of child prodigies, number theory, geometry, and combinatorics. A constant traveller and around-the-clock mathematician, Erdös likes asking questions as much as answering them. In the paper published here he offers intriguing current questions in number theory.

Neil J. A. Sloane ("Multiplexing Methods in Spectroscopy") obtained his Ph.D. in electrical engineering from Cornell in 1967, and since 1969 has been a member of technical staff at Bell Labs, Murray Hill, New Jersey. He is the author of *A Handbook of Integer Sequences, A Short Course on Error-Correcting Codes,* and (with F. J. MacWilliams) *The Theory of Error-Correcting Codes.* About ten years ago the astronomer Martin Harwit asked Sloane the best way to encode a light spectrum: the result was a series of papers, to which the present article is an introduction. Sloane recently received the 1979 Chauvenet Prize for expository writing in mathematics (see p. 123).

Ian F. Blake ("Codes and Designs") has been in the Department of Electrical Engineering of the University of Waterloo since 1969. During this time his research interests have been in communication and in particular coding for both discrete and continuous channels. Such problems, which are essentially packing problems in finite and continuous spaces, respectively, inevitably lead to interesting algebraic and combinatorial structures. The present paper attempts to introduce the reader to the elements of this fascinating study for the finite case.

BUSINESS INFORMATION. Mathematics Magazine is published by the Mathematical Association of America at Washington, D.C., five times a year in January, March, May, September, and November. Ordinary subscriptions are $12 per year. Members of the Mathematical Association of America or of Mu Alpha Theta may subscribe at special reduced rates. Colleges and university mathematics departments may purchase bulk subscriptions (5 or more copies to a single address) for distribution to undergraduate students.

Subscription correspondence and notice of change of address should be sent to A. B. Willcox, Executive Director, Mathematical Association of America, 1529 Eighteenth Street, N.W., Washington, D.C. 20036. Back issues may be purchased, when in print, from P. and H. Bliss Co., Middletown, Connecticut 06457.

Advertising correspondence should be addressed to Raoul Hailpern, Mathematical Association of America, SUNY at Buffalo, Buffalo, New York 14214.

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

Some Unconventional Problems in Number Theory

A mélange of simply posed conjectures with frustratingly elusive solutions.

PAUL ERDŐS
Hungarian Academy of Sciences
University of Colorado
Boulder, CO 80309

I state some curious, unusual, and mostly unsolved problems in various branches of number theory.

Factorial Powers

1. Put $f(n)=\Sigma(1/p)$ for $p<n$ and $p\nmid\binom{2n}{n}$. In a previous paper **[6]** written with Graham, Ruzsa and Straus, we conjectured that there is an absolute constant C so that for all $n, f(n)<C$. I further conjectured for $n>4$, $\binom{2n}{n}$ is never squarefree. It is surprising that this simple conjecture presents so many difficulties.

Since $\binom{2n}{n}\equiv 0 \pmod 4$ except if $n=2^k$, we "only" have to prove that $\binom{2^{k+1}}{2^k}$ is divisible by the square of an odd prime for $k\geqslant 3$. But this does not seem easy. I conjecture that for $k>8$, 2^k is not the sum of distinct powers of 3. (However, $2^8=256=3^5+3^2+3+1$.) This conjecture would imply that for $k\geqslant 9$ $\binom{2^{k+1}}{2^k}\equiv 0 \pmod 3$, but as far as I see there is no method at our disposal to attack this conjecture. There is no doubt that for $n>n_0(k,\alpha)$, $\binom{2n}{n}\equiv 0 \pmod{p^\alpha}$ for some $p>k$. For $p>2, p^2\nmid\binom{342}{171}$ and there is a good chance that this is the greatest $\binom{2n}{n}$ with this property. In other words, for $n>171$, the number $\binom{2n}{n}$ is perhaps divisible by the square of an odd prime.

2. Let $M(n;k)=[n+1,\ldots,n+k]$ be the least common multiple of the integers $n+i$ for $1\leqslant i\leqslant k$. I conjecture that for $m\geqslant n+k, M(n;k)\neq M(m;k)$, or more generally for $l\geqslant k$ and $m\geqslant n+k, M(n;k)\neq M(m;l)$. Unfortunately, I do not see any method to attack these very attractive conjectures. Probably $M(n;k)=M(m;l)$ has very few solutions when $m\geqslant n+k$ and $l>1$. I only know $M(4;3)=M(13;2)$ and $M(3;4)=M(19;2)$. If, similarly, we put $A(n;k)=\Pi_{i=1}^{k}(n+i)$, then I conjecture also that $A(n;k)=A(m;l)$ probably has very few solutions for $m\geqslant n+k$ and $l>1$.

Suppose that $k\geqslant 3$ and $m\geqslant n+k$. Observe that then, for each $k, M(n;k)>M(m;k)$ has infinitely many solutions. Yet I cannot decide whether the same is true for $M(n;k)>M(m;k+1)$. (The referee found two solutions, namely $M(96;7)>M(104;8)$ and $M(132;7)>M(139;8)$.) Let n_k denote the smallest solution of $M(n;k)>M(m;k)$. Try to determine or estimate n_k. It is almost certainly the case that $n_k/k \to \infty$ and perhaps this will not be difficult to prove. It would be worthwhile to compute n_k for small values of k, for perhaps then one can formulate some reasonable conjectures. (Added in proof: $n_k/k \to \infty$ is indeed easy, but I have no good upper bound for n_k.)

Let u_k be the smallest integer for which $M(u_k,k) > M(u_k+1;k)$. It is easy to see that $u_k = (1+o(1))k$ and $u_k > k$. It seems to me that if $t < u_k$ and $T > t$ then $M(t;k) \leqslant M(T;k)$. Perhaps I overlooked a simple argument but I could not prove this.

Unusual Sieve Processes

3. Consider the integers of the form

$$n = ap^2 + b, \text{ where } a \geqslant 1, 0 \leqslant b < p, \text{ and } p \text{ is prime.} \tag{1}$$

It is easy to see by the sieve of Eratosthenes that almost all integers n are of the form (1), but I could not prove that every sufficiently large integer is of this form. In fact, this seems rather unlikely. In view of this I hoped that perhaps the equation

$$n = ak^2 + b, \text{ where } a \geqslant 1, 0 \leqslant b < k, k \text{ is an integer, } k \geqslant 2 \tag{2}$$

would be solvable for every sufficiently large n. Selfridge and Wagstaff made a preliminary computer search and in their opinion it is quite possible that (2) is not solvable for infinitely many n. It would be interesting to find some large (say of size 10^{13} or larger) values of n not of the form (2). Denote by $g(x)$ the number of integers $n < x$ not of the form (1), and by $G(x)$ those not of the form (2). Clearly $G(x) \leqslant g(x)$. It follows from the Brun-Selberg Sieve that $g(x) < c_1 x (\log x)^{-c_2}$. Probably $G(x) < x^c$ for $x > x_0(c)$ for some $c > 0$, and perhaps for all $c > 0$.

Let $u_1 < u_2, \ldots,$ be an infinite sequence of integers. It is probably not difficult to prove that the density of integers not of the form $n = au_i^2 + b$ for $a \geqslant 1$ and $0 \leqslant b < u_i$ exists and is positive if $\Sigma 1/u_i$ converges and is 0 otherwise. More generally (1) could be replaced by the equation

$$n = au_i^2 + b, \text{ where } a \geqslant 1 \text{ and } 0 \leqslant b < v_i \tag{3}$$

and one could try to find non-trivial conditions on $u_1 < u_2 < \cdots$ and $v_1 < v_2 < \cdots$ that (3) should be solvable for all sufficiently large n. I am not very hopeful of success. There is more hope if we only insist that almost all n should be of the form (3).

Barriers

4. Let $f(m) \geqslant 0$ for $1 \leqslant m < \infty$ be any positive function defined on the integers. Then n is called a **barrier** for $f(m)$ if for all $m < n$, $m + f(m) \leqslant n$. Clearly $\phi(m)$ and $\sigma(m)$ do not have barriers because they increase too fast. Let $V(m)$ be the number of distinct prime factors of m. Probably $V(m)$ has infinitely many barriers, but I am very far from being able to prove this. I cannot even prove that there is an $\varepsilon > 0$ for which $\varepsilon V(m)$ has infinitely many barriers. One could try to attack this problem by sieve methods, but it seems to me that these methods are not strong enough at present.

Let $\Omega(m)$ be the number of prime factors of m, multiple factors counted multiply. It seems certain that $\Omega(m)$ also has infinitely many barriers. But this, if true, is certainly unattackable by present day methods. Selfridge observed that $n = 99840$ is the largest barrier for $\Omega(m)$ less than 10^5. Selfridge and I then investigated $d(n)$, the number of divisors of n. Since $\max(d(n-1)+n-1, d(n-2)+n-2) \geqslant n+2$, the most we can hope here is that for infinitely many n,

$$\max_{m<n}(m + d(m)) = n + 2. \tag{4}$$

It is extremely doubtful whether (4) has infinitely many solutions. In fact it is quite possible that $\lim_{n\to\infty} \max_{m<n}(m + d(m) - n) = \infty$. Selfridge and I observed that 24 satisfies (4) and we convinced ourselves that if there is an $n > 24$ which satisfies (4), then this n must be enormously large, far beyond the range of our tables or computers.

It is not difficult to show that the product $F(m) = \Pi \alpha_i$ (where $m = \Pi p_i^{\alpha_i}$) of the number of prime factors of m has infinitely many barriers.

Translation Properties

5. Denote by A the sequence $1 \leqslant a_1 < a_2 < \cdots$. A is said to have the **translation property** if for every n there is a $t_n > 0$ so that u is in A if, and only if, $u + t_n$ is in A for every $1 \leqslant u \leqslant n$. It is not hard to show that the squarefree numbers have the translation property. (This must have been known, although perhaps it does not appear in this form in the literature.) More generally let $b_1 < b_2 < \cdots 1$ where $(b_i, b_j) = 1$ and $\Sigma 1/b_i < \infty$, and let A be the sequence of integers not divisible by any of the b's. It follows easily from the sieve of Eratosthenes that A has the translation property.

If the condition $\Sigma 1/b_i < \infty$ is dropped, the situation is much more complicated. If $\Sigma_{b_i < x}(1/b_i) = o(\log\log x)$, then it can be deduced by Brun's method that A has the translation property. If this condition is also dropped, I have no non-trivial result. I do not know if the integers which are sums of two squares have the translation property. I do not know what happens if we divide the primes into two disjoint classes $q_1 < q_2 < \cdots$; and $r_1 < r_2 < \cdots$, both having for every x more than $cx/\log x$ terms not exceeding x. Denote by $Q_1 < Q_2 < \cdots$ the integers composed of the primes $q_1 < q_2 < \cdots$. Can the sequence $Q_1 < Q_2 < \cdots$ have the translation property? Let $p_1 < p_2 < \cdots$ be the sequence of all primes. It is not hard to show that $p_k < p_{k+1} < \cdots$ can never have the translation property.

Let A have the translation property. The task of determining the smallest t_n which satisfies the definition of the translation property for A will not be easy. For example, if A is the sequence of squarefree numbers, I expect that $t_n > \exp n^c$. I am quite sure that t_n increases faster than polynomially; perhaps this will not be hard to prove.

6. It is extremely difficult to obtain results on the difference of consecutive primes. A well-known conjecture of Cramer states that

$$\limsup_{n\to\infty} \frac{p_{n+1} - p_n}{(\log n)^2} = 1.$$

This conjecture is completely unattackable by present day methods and I expect that it will stay in this class for a very long time.

Let $Q_1 < Q_2 < \cdots$ be the sequence of consecutive squarefree numbers. It is curious that $Q_{n+1} - Q_n$ is almost as difficult to study as $p_{n+1} - p_n$. The best upper bound is, as far as I know, still due to Richert and Rankin [**10, 11**] who proved that for every $\varepsilon > 0$ and $n > n_0, Q_{n+1} - Q_n < n^{2/9+\varepsilon}$. There is no doubt that this inequality holds with $2/9 + \varepsilon$, replaced by ε, but the proof is nowhere in sight. Perhaps $Q_{n+1} - Q_n < c\log n$ holds, but I am very doubtful. It is easy to see that

$$\limsup_{n\to\infty} (Q_{n+1} - Q_n)\log\log n(\log n)^{-1} \geqslant \pi^2/6$$

and as far as I know this has never been improved.

I proved (in [**5**]) that for $0 \leqslant \alpha \leqslant 2$,

$$\lim_{x\to\infty} \frac{1}{x} \sum_{Q_n < x} (Q_{n+1} - Q_n)^\alpha = c_\alpha \tag{5}$$

and Hooley recently proved [**8**] that (5) holds for $\alpha \leqslant 3$. No doubt (5) holds for every α.

Miscellaneous Problems

7. There are many unconventional problems connected with the divisors of n. R. R. Hall and I have a long paper on this subject. Here I just state a conjecture of mine which is more than forty years old: The density of integers n which have two divisors $d_1 < d_2 < (1+\varepsilon)d_1$ is 1 for every $\varepsilon > 0$. I can prove that the density exists, but cannot prove that it is 1, even for large values of ε. A much stronger and more recently formulated conjecture is as follows: Denote by $d^+(n)$ the number of integers k for which n has at least one divisor t where $2^k < t < 2^{k+1}$. Then for almost all integers $n, d^+(n)/d(n) \to 0$ as $n \to \infty$.

8. Let $h(n)$ be the smallest integer so that every μ, where $1 \leqslant \mu < n!$, is the sum of $h(n)$ or fewer distinct divisors of $n!$. I proved $h(n) \leqslant n$. The proof by induction is easy. No doubt very much more is true: $h(n) = o(n)$ and probably $h(n) = o(n^{\varepsilon})$ and hopefully $h(n) < (\log n)^c$ for some c. (I was lead to $h(n)$ by studying $\frac{a}{b} = \frac{1}{x_1} + \cdots + \frac{1}{x_k}$ where $x_1 < \cdots < x_k$ and k is minimal.)

9. An old conjecture of Strauss and myself states that for every $n \geqslant 3$

$$\frac{4}{n} = \frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3}$$

is solvable in integers x_1, x_2, x_3 where $1 \leqslant x_1 < x_2 < x_3$. This conjecture seems surprisingly difficult. A forthcoming paper of Strauss and Subbarao deals with some related questions.

10. Forty years ago I asked: does $x^x y^y = z^z$ have any nontrivial solutions in integers? Chao Ko found infinitely many solutions [**1**]; perhaps he found them all.

11. Put $p_{k+1} - p_k = d_k$ (where $\{p_k\}$ is the sequence of primes). Turan and I proved that both $d_{k+1} > d_k$ and $d_{k+1} < d_k$ have infinitely many solutions. But we could not prove that at least one of the inequalities $d_{k+2} > d_{k+1} > d_k$ and $d_{k+2} < d_{k+1} < d_k$ has infinitely many solutions. If our conjecture is false then, as we observed, there is a k_0 so that for $k > k_0, d_{k+1} - d_k$ alternates in sign. This is certainly not the case and perhaps the proof is not hard. I offer 100 dollars for a proof or disproof.

12. Let $1 = r_1 < \cdots < r_{\phi(n)} = n - 1$ be the $\phi(n)$ integers relatively prime to n. I conjectured nearly forty years ago that there is an absolute constant C so that

$$\sum_{i=1}^{\phi(n)-1} (r_{i+1} - r_i)^2 < C \frac{n^2}{\phi(n)}. \tag{6}$$

This does not look hard, but it has not yet been settled and I offer 250 dollars for a proof or disproof. No doubt (6) holds if 2 is replaced by any positive α (C must then be replaced by C_α). Hooley proved this for $\alpha < 2$.

This paper is based on remarks presented at the fifth annual Mathematics and Statistics Conference in October, 1977, at Miami University in Oxford, Ohio.

References

[1] Ko Chao, Note on the diophantine equation $x^x y^y = z^z$, J. Chinese Math. Soc., 2 (1940) 205–207 (Math. Rev. V. 2, p. 346).

[2] P. Erdös, On the density of some sequences of integers, Bull. Amer. Math. Soc., 54 (1948) 685–692.

[3] ———, On the difference of consecutive primes, Bull. Amer. Math. Soc., 54 (1948) 885–889.

[4] ———, On the equation $\frac{1}{x_1} + \cdots + \frac{1}{x_k} = \frac{a}{b}$ (in Hungarian), Mat. Lapok, 1 (1950) 192–210. (MR 13, p. 208.)

[5] ———, Some problems and results in elementary number theory, Publ. Math. Debrecen, 2 (1951) 103–109. (MR 13, p. 627.)

[6] ———, R. L. Graham, I. Z. Ruzsa and E. Straus, On the prime factors of $\binom{2n}{n}$, Math. Comp., 29 (1975) 83–92.

[7] ———, and P. Turan, On some new questions on the distribution of prime numbers, Bull. Amer. Math. Soc., 54 (1948) 271–278.

[8] C. Hooley, On the intervals between consecutive terms of sequences, Proc. Symp. Pure Math. XXIV. Analytic Number Theory, Amer. Math. Soc., 1973, 129–140.

[9] W. H. Mills, Number theory conference, Boulder, Colorado, 1959.

[10] R. A. Rankin, Van der Corput's method and the theory of exponent pairs, Quart. J. Math., 6 (Ser. 2) (1955) 147–153.

[11] H. E. Richert, On the difference between consecutive squarefree numbers, J. London Math. Soc., 29 (1954) 16–20.

Multiplexing Methods in Spectroscopy

Linear combinations of inaccurate measurements yield improved results from minimal laboratory data.

Neil J. Sloane
Bell Laboratories
Murray Hill, NY 07974

Over the past eight years a new technique known as Hadamard Transform Optics has been developed in spectroscopy and imaging. The basic idea is as follows. In order to determine the spectrum of a beam of light, for example, instead of measuring the intensity of each frequency component separately, the frequency components are combined in groups and the total intensity of each group is measured. Thus the different frequency components are **multiplexed**, and as a result the spectrum may be determined much more accurately. The best multiplexing schemes are based on Hadamard matrices, and in the most favorable case reduce the mean square error per frequency component by a factor proportional to n if there are n components.

We begin this paper with a description of multiplexing methods, then show how these are related to what statisticians call weighing designs, and finally describe the best multiplexing methods and the improvements they produce. Along the way several unsolved combinatorial problems will be described that arise from this work. A more detailed account of the subject may be found in [1].

Multiplexing Methods

Color photography is so much a part of everyday life that we tend to forget how unique a color picture really is. There are no direct counterparts of a color print in X-ray optics, nor is there any similar device in the infrared or radio domain. Even the simpler black-and-white photos have no direct analog in the far infrared or radio regions. To obtain an infrared picture we must scan an image of the scene with a single detector or receiver, whose output is then used to reconstruct the scene. In the near infrared, images can sometimes be obtained using arrays of detectors, or a vidicon or charge injection device. But these sensors operate only at certain wavelengths, and in any case are expensive.

These limitations can be overcome, and black-and-white pictures obtained, by using a technique known as **multiplexing**. In this technique the incident radiation is first separated into distinct bundles of rays, corresponding to different portions of the scene. Then certain combinations of these bundles are allowed to fall on the detector and the total intensity is recorded. After a series of say n suitably chosen combinations have been recorded, the individual intensities of n different bundles can be calculated, and a black-and-white picture obtained.

Multiplexing is also useful in spectroscopy. A conventional spectrometer sorts electromagnetic radiation into distinct bundles of rays, corresponding to different colors. Thus each bundle is labeled by the appropriate frequency, wavelength or wavenumber. The spectrum of the radiation is found by measuring the intensity of each bundle. Alternatively, the bundles can be multiplexed: instead of measuring the intensity of each bundle separately, we could measure the total intensity of various combinations of bundles. After measuring n suitably chosen combinations, the individual intensities of n different bundles can be calculated, and the spectrum obtained.

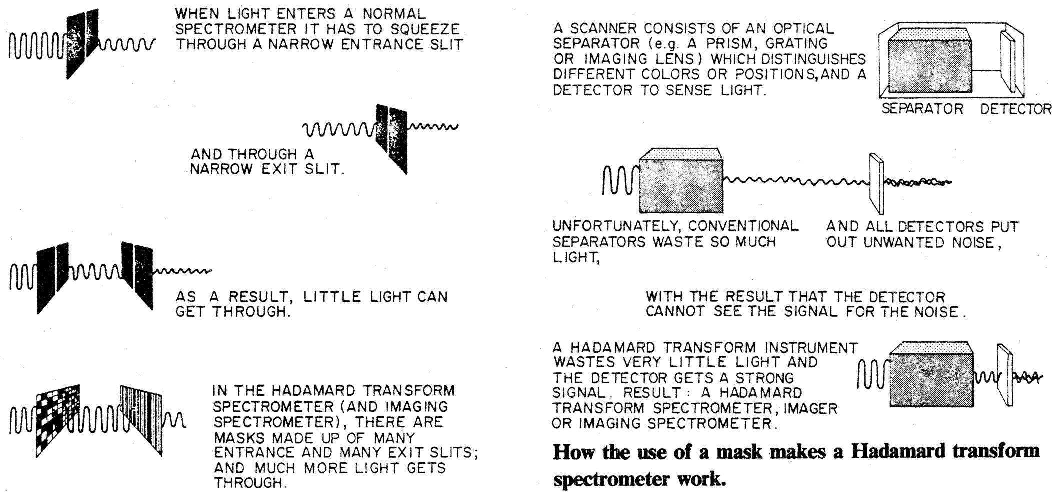
Finally, by combining these two forms of multiplexing—multiplexing radiation from different parts of a picture and from different frequencies—it is possible to reconstruct a color picture of the scene.

The reason experimenters resort to a technique as complicated as multiplexing is that it enables them to reduce the effect of noise in optical measurements. Any detector is a source of noise, no matter how carefully it is constructed. (Even when no radiation is present, a detector will produce spurious signals, which are often indistinguishable from the signals produced when light does fall on the detector.) An experimenter's task is to minimize the effects of this noise, and to reconstruct with as much fidelity as possible the intensity of radiation incident on the detector.

If the noise is independent of the strength of the incident signal, it may be advantageous to combine the radiation from a large number of bundles of rays, because then the total intensity of the light may provide a signal considerably greater than the detector noise. Thus the primary purpose of multiplexing is to maximize the radiant flux incident on the detector in order to improve the signal-to-noise ratio of the final intensity display. The final display may be a spectrum, a black-and-white picture, or a color picture. Although the optical apparatus needed to separate spectral components is different from that needed to separate spatial components, the principle is the same in the three cases.

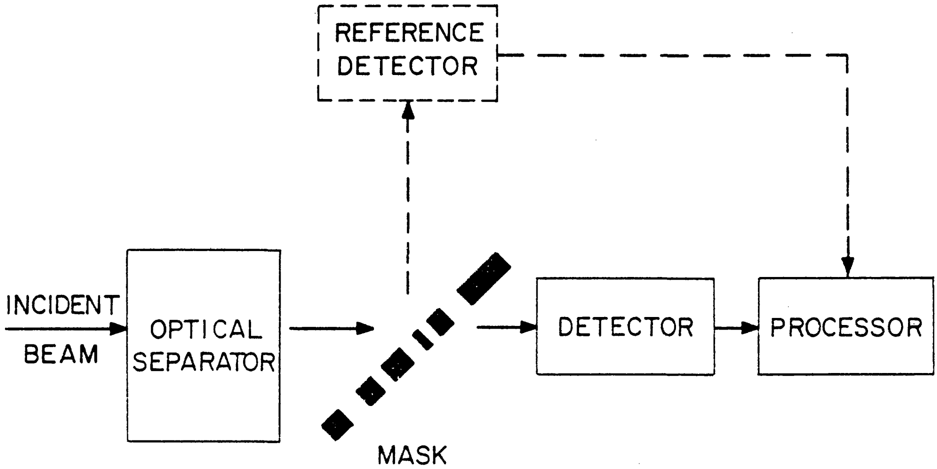
Two quite different multiplexing techniques are available. These can be loosely described by saying that the first uses interference techniques and Fourier transforms, while the second uses masks and discrete (often Hadamard) transforms. Examples of the first technique are (i) the Michelson interferometric spectrometer, in which an interferometer is used to modulate the intensities of different transmitted wavelengths (some of the many references dealing with this instrument are **[2]**-**[12]**), (ii) the method of aperture synthesis in radio astronomy, in which spatial maps of the sky are obtained by measuring the interference patterns between radio waves reaching two or more antennas from cosmic sources (see for example **[13]**), and (iii) the technique of holography, in which an interference pattern is stored and then used to reconstruct a spatial, even three-dimensional, image **[14]**.

The underlying principle in the second technique is the use of masks which either block or transmit light. As we shall see, the best masks for these instruments are constructed from matrices named after the French mathematician Jacques Hadamard (cf. **[15]**). We have therefore called this general class of instruments **Hadamard Transform Optics**. These instruments are capable of obtaining color pictures at any wavelength and over a wide range of spectral and spatial resolutions. FIGURE 1 gives a simple view of how one of those instruments works.


WHEN LIGHT ENTERS A NORMAL SPECTROMETER IT HAS TO SQUEEZE THROUGH A NARROW ENTRANCE SLIT
AND THROUGH A NARROW EXIT SLIT.
AS A RESULT, LITTLE LIGHT CAN GET THROUGH.
IN THE HADAMARD TRANSFORM SPECTROMETER (AND IMAGING SPECTROMETER), THERE ARE MASKS MADE UP OF MANY ENTRANCE AND MANY EXIT SLITS; AND MUCH MORE LIGHT GETS THROUGH.
A SCANNER CONSISTS OF AN OPTICAL SEPARATOR (e.g. A PRISM, GRATING OR IMAGING LENS) WHICH DISTINGUISHES DIFFERENT COLORS OR POSITIONS, AND A DETECTOR TO SENSE LIGHT.
SEPARATOR
DETECTOR
UNFORTUNATELY, CONVENTIONAL SEPARATORS WASTE SO MUCH LIGHT,
AND ALL DETECTORS PUT OUT UNWANTED NOISE,
WITH THE RESULT THAT THE DETECTOR CANNOT SEE THE SIGNAL FOR THE NOISE.
A HADAMARD TRANSFORM INSTRUMENT WASTES VERY LITTLE LIGHT AND THE DETECTOR GETS A STRONG SIGNAL. RESULT : A HADAMARD TRANSFORM SPECTROMETER, IMAGER OR IMAGING SPECTROMETER.

How the use of a mask makes a Hadamard transform spectrometer work.

FIGURE 1.



The basic Hadamard transform multiplexing system.
FIGURE 2.

The basic Hadamard transform instrument consists of four essential components: an optical separator, an encoding mask, a detector and a processor (FIGURE 2). (More complex instruments may make use of additional components. In some devices, for example, two masks are used.) The separator may be nothing more than a lens which produces a focused image at the mask, and separates light arriving from different spatial elements of a scene. Or the separator may be a dispersing system (a prism or grating) which separates different frequency components of a beam and focuses them onto different locations on the mask.

In the instrument shown in FIGURE 2 the mask is made up of three types of elements. A particular location on the mask either transmits light to the main detector, absorbs the light, or reflects it towards a reference detector. In this way the corresponding element of the separated beam is modulated. If we record the difference between the reading of the main detector and the reference detector, the intensity of this element of the beam has been multiplied by $+1$, 0 or -1 respectively. Usually the reference detector in FIGURE 2 is omitted, however, and then the mask is made up of just two types of elements: open and closed slots. Each element of the beam is either transmitted or absorbed, and the single detector measures the sum of the transmitted elements.

If there are n unknown intensity values to be determined, at least n different measurements must be made, each with a different mask position. Three important questions in designing such an instrument are: (a) How should the mask be chosen? (b) How much does the instrument improve the quality of the measurements? and (c) How close to optimum is the chosen mask design? Such questions have been studied for many years in statistics under the name of weighing designs, as we shall see in the next section. (The application of weighing designs to optics seems to have been first pointed out in [**16**] and [**17**].)

Weighing Designs and Optical Multiplexing

A **weighing design** is a scheme for accurately weighing a number of objects by weighing them in groups rather than one at a time. (The idea appears to have originated with Yates [**18**].) This is a form of multiplexing, only now we are combining weights rather than intensities. In the optical analogue we "weigh" (i.e., measure the intensity of) different bundles of rays by weighing them in groups rather than individually. The benefits of this procedure in optics were pointed out by Fellgett ([**19**],[**20**]) and the resulting increase in accuracy is sometimes called the Fellgett or multiplex advantage.

First we give three very simple examples of weighing designs. Suppose four objects are to be weighed, using a balance that makes an error each time it is used. We assume that the balance has been adjusted as well as it can be, so that the average value of this error is zero. We indicate this by saying that $E\{e\}=0$ where e is the error, and E denotes **expected value**, (or, speaking loosely, the average value) over a large number of experiments. (For these and other terms from probability theory, see for example Cramér [**21**] or Papoulis [**22**].) We also assume that e is independent of the total weight on the balance, and that the errors e,e' in different measurements are **independent**, or in other words that $E\{ee'\}=0$. Of course the square of the error, e^2, is always non-negative, and we denote *its* average by σ^2: $E\{e^2\}=\sigma^2$. The expression σ^2 is called the **variance** of the error, and σ itself is the **standard deviation** of the error. In a good balance σ^2 is small.

Weighing Design I. Suppose the four objects are weighed separately. Let the true unknown weights of the objects be $\psi_1,\psi_2,\psi_3,\psi_4$, let the actual measurements obtained with the balance be $\eta_1,\eta_2,\eta_3,\eta_4$, and let the errors made by the balance be e_1,e_2,e_3,e_4. Then the four weighings give four equations:

$$\begin{aligned}\eta_1&=\psi_1+e_1,\\ \eta_2&=\psi_2+e_2,\\ \eta_3&=\psi_3+e_3,\\ \eta_4&=\psi_4+e_4.\end{aligned}\tag{1}$$

Because of the errors we cannot determine $\psi_1,\dots,\psi_4$ exactly, but must be content with estimates $\hat{\psi}_1,\dots,\hat{\psi}_4$ which we hope will be close to the true values. It is very plausible in this example that we should use $\eta_1,\dots,\eta_4$ themselves as the estimates. If we agree that the estimates should be linear functions of the measurements, and be **unbiased** (satisfy $E\{\hat{\psi}_1\}=\psi_1,\dots,E\{\hat{\psi}_4\}=\psi_4$) then this is indeed true. (See for example [**1**], [**21**], [**23**], [**24**]. However, it is possible to make a good argument for using other estimators even in this simple example—see [**25**], [**26**].)

Hence we use $\hat{\psi}_1=\eta_1=\psi_1+e_1,\dots,\ \hat{\psi}_4=\eta_4=\psi_4+e_4$ as estimates of $\psi_1,\dots,\psi_4$. The difference between the estimate $\hat{\psi}_i$ and the true value is $\hat{\psi}_i-\psi_i=e_i$. By hypothesis this has average value zero: $E\{\hat{\psi}_i-\psi_i\}=E\{e_i\}=0$, or $E\{\hat{\psi}_i\}=\psi_i$. Thus these estimates are unbiased. On the other hand the square of the error has average value $E\{(\hat{\psi}_i-\psi_i)^2\}=E\{e_i^2\}=\sigma^2$. In other words the **mean square error** in each weight is σ^2. The crucial observation made by Yates is that the mean square error can be reduced by weighing several objects at once.

Weighing Design II. For the second experiment we suppose that the balance is a chemical balance with two pans, and that the four weighings are made as follows:

$$\begin{aligned}\eta_1&=\psi_1+\psi_2+\psi_3+\psi_4+e_1,\\ \eta_2&=\psi_1-\psi_2+\psi_3-\psi_4+e_2,\\ \eta_3&=\psi_1+\psi_2-\psi_3-\psi_4+e_3,\\ \eta_4&=\psi_1-\psi_2-\psi_3+\psi_4+e_4.\end{aligned}\tag{2}$$

This means that in the first weighing all four objects are placed in the left-hand pan, in the second weighing objects 1 & 3 are in the left-hand pan and 2 & 4 in the right, and so on. Such a specification of which objects are to be weighed in each measurement is called a **weighing design**. (Thus equation (1) is also a weighing design, albeit a trivial one.)

In this case the best estimates for $\psi_1,\dots,\psi_4$ are found by solving equations (2) for $\psi_1,\dots,\psi_4$, and are given by

$$\begin{aligned}\hat{\psi}_1&=\frac{1}{4}(\eta_1+\eta_2+\eta_3+\eta_4) & \qquad \hat{\psi}_4&=\frac{1}{4}(\eta_1-\eta_2-\eta_3+\eta_4)\\ &\qquad\qquad\qquad\qquad\cdots\\ &=\psi_1+\frac{1}{4}(e_1+e_2+e_3+e_4) & &=\psi_4+\frac{1}{4}(e_1-e_2-e_3+e_4)\,.\end{aligned}$$

Again these are unbiased estimates, since $E\{\hat{\psi}_i-\psi_i\}=0$, but now the mean square error is, for example,

$$E\left\{(\hat{\psi}_1-\psi_1)^2\right\}=E\left\{\frac{1}{16}(e_1+e_2+e_3+e_4)^2\right\}=\frac{1}{4}\sigma^2;$$

similarly, $E\{(\hat{\psi}_i-\psi_i)^2\}=\frac{1}{4}\sigma^2$, for $i=1,\ldots,4$. Therefore by weighing the objects together the mean square error has been reduced by a factor of four!

Weighing Design III. Finally, for the third experiment we suppose that the balance is a spring balance with only one pan. Now only coefficients 0 and 1 can be used—either an object is weighed or it is not. A good method of weighing the four objects is:

$$\begin{array}{rl} \eta_1= & \quad\;\;\; \psi_2+\psi_3+\psi_4+e_1, \\ \eta_2= & \psi_1+\psi_2 \qquad\qquad\;\; +e_2, \\ \eta_3= & \psi_1 \qquad\; +\psi_3 \qquad\; +e_3, \\ \eta_4= & \psi_1 \qquad\qquad\quad +\psi_4+e_4. \end{array} \tag{3}$$

Thus first objects 2, 3 & 4 are weighed together, then 1 & 2, then 1 & 3, and finally 1 & 4. Solving (3) for $\psi_1,\ldots,\psi_4$ we find that the best estimates are given by

$$\hat{\psi}_1=\frac{1}{3}(-\eta_1+\eta_2+\eta_3+\eta_4)$$

$$\cdots\cdot$$

$$=\psi_1+\frac{1}{3}(-e_1+e_2+e_3+e_4).$$

Once again these are unbiased estimates, but now the mean square errors are

$$E\left\{(\hat{\psi}_1-\psi_1)^2\right\}=\frac{4\sigma^2}{9},$$

$$E\left\{(\hat{\psi}_2-\psi_2)^2\right\}=E\left\{(\hat{\psi}_3-\psi_3)^2\right\}=E\left\{(\hat{\psi}_4-\psi_4)^2\right\}=\frac{7\sigma^2}{9}.$$

The mean square errors have again been reduced, but by a smaller amount than in weighing design II.

The General Weighing Design

In the general case, suppose there are n unknown weights $\psi_1,\ldots,\psi_n$, and $m \geqslant n$ weighings are made, described by

$$\begin{array}{l} \eta_1=w_{11}\psi_1+\cdots+w_{1n}\psi_n+e_1 \\ \cdots\cdots\cdots \\ \eta_m=w_{m1}\psi_1+\cdots+w_{mn}\psi_n+e_m. \end{array} \tag{4}$$

For a chemical balance (or two-pan) design, the coefficients w_{ij} are $+1$, 0, or -1, specifying whether the object is to be placed in the left pan, in neither pan, or in the right pan. In a spring balance (or one-pan) design, the w_{ij} are 0 and 1, and specify whether the object is to be weighed or not.

These equations are simpler in matrix form. Let

$$\eta=\begin{bmatrix}\eta_1\\ \cdot\\ \vdots\\ \cdot\\ \eta_m\end{bmatrix},\quad \psi=\begin{bmatrix}\psi_1\\ \cdot\\ \vdots\\ \cdot\\ \psi_n\end{bmatrix},\quad e=\begin{bmatrix}e_1\\ \cdot\\ \vdots\\ \cdot\\ e_m\end{bmatrix}$$

be column vectors of measurements, unknowns, and errors, and let W be an $m\times n$ matrix with (i,j)th entry equal to w_{ij}. Equations (4) become

$$\eta = W\psi + e. \tag{5}$$

The matrix W gives a concise specification of the weighing design. For example, the matrices describing weighing designs (1), (2) and (3) are, respectively,

$$W=\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{bmatrix},\quad W=\begin{bmatrix}1&1&1&1\\1&-1&1&-1\\1&1&-1&-1\\1&-1&-1&1\end{bmatrix},\quad W=\begin{bmatrix}0&1&1&1\\1&1&0&0\\1&0&1&0\\1&0&0&1\end{bmatrix}.$$

The connection between weighing designs and multiplex optics is now straightforward. In the optical case the unknowns ψ_i represent intensities of individual spatial and/or spectral elements in a beam of radiation. In contrast to scanning instruments which measure the intensities one at a time, the multiplex optical system measures (i.e. weighs) several intensities (or ψ_i's) simultaneously. The η_i's now represent the readings of the detector (instead of the readings of the balance). Finally, the weighing design itself, W, is represented by the mask in FIGURE 2. More precisely, one row of W, which specifies which objects are present in a single weighing, corresponds to the row of transmitting, absorbing or reflecting elements shown in FIGURE 2. We usually refer to such a row as a **mask configuration.** The two types of weighing designs—chemical and spring balance designs—are realized by masks which contain either transmitting, absorbing and reflecting elements (for the chemical balance design) or simply open and closed slots (for the spring balance design). Note that the former case requires two detectors, as shown in FIGURE 2, whereas in the latter case the reference detector can be omitted.

Given the measurements $\eta_1,\ldots,\eta_m$, what should we use for the estimates $\hat{\psi}_1,\ldots,\hat{\psi}_n$ of the unknowns? Again restricting ourselves to linear unbiased estimates, the answer (see the above references) is that we use $\hat{\psi}=W^{-1}\eta$ if W has an inverse, or $\hat{\psi}=W^{+}\eta$ if it doesn't, where W^{+} is the **Moore-Penrose generalized inverse** ([27]-[33]). The most important example is when the columns of W are linearly independent. Then W^TW is an invertible $n\times n$ matrix (the T denotes transpose), and $W^{+}=(W^TW)^{-1}W^T$; hence $\hat{\psi}=(W^TW)^{-1}W^T\eta$.

Let $\varepsilon_j=E\{(\hat{\psi}_j-\psi_j)^2\}$ be the mean square error in the jth estimate and $\varepsilon=(\varepsilon_1+\cdots+\varepsilon_n)/n$ be the average mean square error. After some algebra we find from equation (5) that whether $\hat{\psi}$ is $W^{-1}\eta$ or $W^{+}\eta$, the average mean square error ε is related to W by the useful formula

$$\varepsilon=\frac{\sigma^2}{n}\operatorname{Trace}(W^TW)^{-1}, \tag{6}$$

where the trace of a matrix is the sum of the entries on the main diagonal. If W has an inverse, (6) states that ε is (σ^2/n) times the sum of the squares of the elements of W^{-1}.

We wish to make ϵ as small as possible. So the first question is: which W minimizes $\operatorname{Trace}(W^TW)^{-1}$? A partial answer is given below. (A weighing design W with the minimum ϵ is called **A-optimal.** For more about weighing designs see [34]-[38].)

Hadamard and S-Matrices

The best masks and weighing designs (the two are equivalent, as we have just seen) use Hadamard matrices for chemical balance designs and S-matrices for spring balance designs. We now proceed to define these matrices.

A **Hadamard matrix** H_n of order n is an $n\times n$ matrix of $+1$'s and -1's with the property that the scalar product of any two distinct rows is 0. Thus H_n must satisfy $H_nH_n^T=nI_n$, where I_n is the $n\times n$ identity matrix. Examples of Hadamard matrices of orders 1, 2, 4 are shown in FIGURE 3. If H_n is a Hadamard matrix, then so is any matrix obtained from H_n by multiplying any of the rows and columns by -1. In this way we can always suppose that H_n is arranged to have all elements of the first row and first column equal to +1. Such a Hadamard matrix is said to be **normalized**. For example, all the Hadamard matrices in FIGURE 3 are normalized.

$$H_1=[1],\quad H_2=\begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix},\quad H_4=\begin{bmatrix}1 & 1 & 1 & 1\\ 1 & -1 & 1 & -1\\ 1 & 1 & -1 & -1\\ 1 & -1 & -1 & 1\end{bmatrix},$$

$$H_8=\begin{bmatrix}1 & 1 & 1 & 1 & 1 & 1 & 1 & 1\\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1\\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1\\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1\\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1\\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1\\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1\\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1\end{bmatrix}$$

Examples of Hadamard matrices.

FIGURE 3.

To define an S-matrix, we begin with a normalized Hadamard matrix H_n of order n. Then an S-matrix of order $n-1$, S_{n-1}, is the $(n-1)\times(n-1)$ matrix of 0's and 1's obtained by omitting the first row and column of H_n and then changing +1's to 0's and −1's to 1's. The S-matrices of orders 1, 3, and 7 obtained from FIGURE 3 are shown in FIGURE 4. It is not difficult to show that an S-matrix of order n satisfies

$$S_nS_n^T=\frac{1}{4}(n+1)(I_n+J_n), \tag{7}$$

$$S_nJ_n=J_nS_n=\frac{1}{2}(n+1)J_n, \tag{8}$$

and

$$S_n^{-1}=\frac{2}{n+1}(2S_n^T-J_n), \tag{9}$$

where J_n is the $n\times n$ matrix of 1's. An $n\times n$ matrix of 0's and 1's is an S-matrix if and only if (7) and (8) are satisfied.

$$S_1=[1],\quad S_3=\begin{bmatrix}1 & 0 & 1\\ 0 & 1 & 1\\ 1 & 1 & 0\end{bmatrix},$$

$$S_7=\begin{bmatrix}1 & 0 & 1 & 0 & 1 & 0 & 1\\ 0 & 1 & 1 & 0 & 0 & 1 & 1\\ 1 & 1 & 0 & 0 & 1 & 1 & 0\\ 0 & 0 & 0 & 1 & 1 & 1 & 1\\ 1 & 0 & 1 & 1 & 0 & 1 & 0\\ 0 & 1 & 1 & 1 & 1 & 0 & 0\\ 1 & 1 & 0 & 1 & 0 & 0 & 1\end{bmatrix}$$

Examples of S-matrices.

FIGURE 4.

It is easy to prove that if a Hadamard matrix of order n exists, then n must be 1, 2, or a multiple of 4. It is generally believed that Hadamard matrices exist of every order which is a multiple of 4. A large number of different constructions are known (see **[39]**-**[41]**), and at the present time the smallest order which has not been constructed is 268.

$$S_7=\begin{bmatrix}1&1&1&0&1&0&0\\1&1&0&1&0&0&1\\1&0&1&0&0&1&1\\0&1&0&0&1&1&1\\1&0&0&1&1&1&0\\0&0&1&1&1&0&1\\0&1&1&1&0&1&0\end{bmatrix}$$

A cyclic S-matrix.
FIGURE 5.

For optical applications it is most convenient if the S-matrix is a **cyclic** matrix, i.e., has the property that each row is obtained by shifting the previous row one place to the left (with any overflow from the left coming in on the right). This considerably reduces the cost of the optical apparatus. The first two S-matrices in FIGURE 4 are cyclic, as is the matrix of order 7 shown in FIGURE 5. For the three known constructions of cyclic S-matrices see [1], [39], [42]-[45].

Hotelling's Theorem

In 1944 Hotelling ([35]; see also [1], [46]) established the following fundamental result: *If W is any $m \times n$ matrix with $m \geqslant n$ and with entries in the range $-1 \leqslant w_{ij} \leqslant 1$, then* $\operatorname{Trace}(W^TW)^{-1} \geqslant 1$. *Furthermore, equality holds if and only if $W^TW = mI_n$.* This implies that the average mean square error is bounded below by $\varepsilon \geqslant \sigma^2/m$. Let us consider the most important case, when W is square. Then $m=n$, so the theorem implies that $\varepsilon \geqslant \sigma^2/n$, and equality holds if and only if W is a Hadamard matrix of order n. Thus to attain this minimum value for ε, the mask entries w_{ij} must be ± 1 (corresponding to a chemical balance design), n must be 1, 2 or a multiple of 4, and a Hadamard matrix of order n must exist. (Note that mask entries between -1 and $+1$ are not used.) The average mean square error in each unknown is then reduced from σ^2 (if no multiplexing is used) to σ^2/n (multiplexing with $W = H_n$). For example, weighing design II illustrates the use of the Hadamard matrix H_4, and reduces ϵ to $\sigma^2/4$.

The S-Matrix Conjecture

If the mask entries are restricted to the range $0 \leqslant w_{ij} \leqslant 1$, which is the case of greatest interest in spectroscopy, much less is known. For simplicity we assume $m=n$. In place of Hotelling's theorem we make the following conjecture.

CONJECTURE. *If W is any $n \times n$ matrix with entries in the range $0 \leqslant w_{ij} \leqslant 1$, then*

$$\operatorname{Trace}(W^TW)^{-1} \geqslant \frac{4n^2}{(n+1)^2}. \tag{10}$$

Furthermore, equality holds if and only if W is an S-matrix.

If the conjecture is true, then in this case the average mean square error is bounded by $\epsilon \geqslant 4n\sigma^2/(n+1)^2$. For equality to hold, the mask entries must be 0's and 1's, $n+1$ must be 2 or a multiple of 4, Hadamard matrix of order $n+1$ must exist, and $W = S_n$. In this case the average mean square error is reduced from σ^2 (no multiplexing) to about $4\sigma^2/n$ (multiplexing with $W = S_n$).

For example, a weighing design which uses the S-matrix S_3 of FIGURE 4 would be: first weigh objects 1 & 3, then 2 & 3, and then 1 & 2. These weighings give three equations

$$\begin{aligned}\eta_1 &= \psi_1 \phantom{{}+\psi_2} + \psi_3 + e_1,\\ \eta_2 &= \phantom{\psi_1 + {}} \psi_2 + \psi_3 + e_2,\\ \eta_3 &= \psi_1 + \psi_2 \phantom{{}+\psi_3} + e_3,\end{aligned}$$

these estimates of the unknowns

$$\hat{\psi}_1 = \frac{1}{2}(\eta_1 - \eta_2 + \eta_3),$$

$$\hat{\psi}_2 = \frac{1}{2}(-\eta_1 + \eta_2 + \eta_3),$$

$$\hat{\psi}_3 = \frac{1}{2}(\eta_1 + \eta_2 - \eta_3),$$

and mean square errors of $\epsilon_j = 3\sigma^2/4$, for $j = 1,\ldots,3$. The average mean square error has been reduced by a factor of $3/4$ in agreement with our conjecture.

Although the conjecture has not been proved, it follows from (9) that choosing $W = S_n$ does attain (10), and it is known that no other matrix can do much better than S_n. More precisely, if W is any $n \times n$ matrix with $0 \leqslant w_{ij} \leqslant 1$, then ([1], [17])

$$\operatorname{Trace}(W^T W)^{-1} \geqslant \frac{4n}{(n+1)^{(n+1)/n}} \approx 4 \quad \text{for } n \text{ large.} \tag{11}$$

Since the values of n used in practice tend to be quite large (typically 255—see [1], [4]), for all practical purposes the conjecture is settled by this estimate.

Unsolved Problems

For constructing weighing designs and masks, we are interested in four classes of matrices W, and in each class the main problem is to find the minimum value of $p(W) = \operatorname{Trace}(W^T W)^{-1}$. These lead to the following unsolved problems.

Class (*i*). W is $n \times n$, invertible, with $-1 \leqslant w_{ij} \leqslant 1$. If n is a multiple of 4, show that a Hadamard matrix of order n always exists. For other values of n, what is the minimum value of $p(W)$? (For partial results, e.g., the use of conference matrices, see for example [17].)

Class (*ii*). W is $m \times n$, of rank n, with $-1 \leqslant w_{ij} \leqslant 1$. If m is a multiple of 4, then we can attain the minimum by using n columns of a Hadamard matrix of order m. What about other values of m? (Cf. [35].)

Class (*iii*). W is $n \times n$, invertible, with $0 \leqslant w_{ij} \leqslant 1$. Prove the S-conjecture. Or prove the weaker result that the minimum $\rho(W)$ can always be attained by a matrix of 0's and 1's. What if $n+1$ is not a multiple of 4?

Class (*iv*). W is $m \times n$ of rank n, with $0 \leqslant w_{ij} \leqslant 1$. If $m+1$ is a multiple of 4, is it best to take $W = S_{m,n}$, consisting of n columns of an S-matrix of order m, and attaining $p(W) = p(S_{m,n}) = 4n^2/(m+1)(n+1)$? For this class of matrices no lower bound on $p(W)$ (analogous to (11)) is known.

It would also be interesting to know the average or the maximum values of $p(W)$ in each class.

References

[1] M. Harwit and N. J. A. Sloane, Hadamard Transform Optics, Academic Press, New York (to appear).
[2] J. B. Bates, Fourier transform infrared spectroscopy, Science, 191 (1976) 31-37.
[3] E. D. Becker and T. C. Farrar, Fourier transform spectroscopy, Science, 178 (1972) 361-368.
[4] M. Born and E. Wolf, Principles of Optics, Pergamon Press, Oxford, 5th edition, 1975.
[5] M. Françon, Optical Interferometry, Academic Press, New York, 1966.
[6] P. R. Griffiths, Interferometry in the Seventies, Analyt. Chem., 46 (1974) 645A-654A.
[7] T. H. Maugh, Fourier transform: the revolution comes to infrared, Science, 191 (1976) 1250-1251.

[8] L. Mertz, Transformations in Optics, John Wiley, New York, 1965.
[9] A. A. Michelson, Light Waves and Their Uses, Univ. of Chicago Press, 1902, reprinted 1961.
[10] B. Rossi, Optics, Addison-Wesley, Reading, Mass., 1957.
[11] J. E. Stewart, Infrared Spectroscopy: Experimental Methods and Techniques, Marcel Dekker, New York, 1970.
[12] G. A. Vanasse and H. Sakai, Fourier spectroscopy, in Progress in Optics, edited by E. Wolf, North Holland, Amsterdam, VI (1967) 259-330.
[13] J. L. Steinberg and J. Lequeux, Radio Astronomy, McGraw-Hill, New York, 1963.
[14] J. B. DeVelis and G. O. Reynolds, Theory and Applications of Holography, Addison-Wesley, Reading, Mass., 1967.
[15] J. Hadamard, Résolution d'une question relative aux déterminants, Bull. Sci. Math., (2) 17 (1893) 240-248.
[16] A. G. Marshall and M. B. Comisarow, Fourier and Hadamard transform methods in spectroscopy, Analyt. Chem., 47 (1975) 491A-504A.
[17] N. J. A. Sloane and M. Harwit, Masks for Hadamard transform optics, and weighing designs, Appl. Opt., 15 (1976) 107-114.
[18] F. Yates, Complex experiments, J. Roy. Statist. Soc. Supp., 2 (1935) 181-247.
[19] P. Fellgett, The Theory of Infrared Sensitivities and Its Application to Investigations of Stellar Radiation in the Near Infrared, Ph.D. Thesis, Cambridge Univ., 1951.
[20] P. Fellgett, Conclusions on multiplex methods, J. de Physique, Colloque C2, 28 (1967) 165-171.
[21] H. Cramér, Mathematical Methods of Statistics, Princeton Univ. Press, 1946.
[22] A. Papoulis, Probability, Random Variables, and Stochastic Processes, McGraw-Hill, New York, 1965.
[23] R. Deutsch, Estimation Theory, Prentice Hall, Englewood Cliffs, N. J., 1965.
[24] N. J. A. Sloane, T. Fine, P. G. Phillips, and M. Harwit, Codes for multislit spectrometry, Appl. Opt., 8 (1969) 2103-2106.
[25] B. Efron, Biased versus unbiased estimation, Advances in Math., 16 (1975) 259-277.
[26] B. Efron and C. Morris, Stein's paradox in statistics, Scientific American, 236 (No. 5, 1977) 119-127.
[27] A. Ben-Israel and A. Charnes, Contributions to the theory of generalized inverse, J. Soc. Indust. Appl. Math., 11 (1963) 667-699.
[28] A. Ben-Israel and T. N. E. Greville, Generalized Inverses: Theory and Applications, John Wiley, New York, 1974.
[29] M. Z. Nashed, editor, Generalized Inverses and Applications, Academic Press, New York, 1976.
[30] R. Penrose, A generalized inverse for matrices, Proc. Cambridge Philos. Soc., 51 (1955) 406-413.
[31] R. Penrose, On best approximate solution of linear matrix equations, Proc. Cambridge Philos. Soc., 52 (1956) 17-19.
[32] C. M. Price, The matrix pseudoinverse and minimal variance estimates, SIAM Review, 6 (1964) 115-120.
[33] C. R. Rao and S. K. Mitra, Generalized Inverse of Matrices and Its Applications, John Wiley, New York, 1971.
[34] K. S. Banerjee, Weighing Designs for Chemistry, Medicine, Economics, Operations Research, Statistics, Marcel Dekker, N. Y. 1975.
[35] A. V. Geramita and J. S. Wallis, Orthogonal designs III: Weighing matrices, Utilitas Math., 6 (1974) 209-236.
[36] H. Hotelling, Some improvements in weighing and other experimental techniques, Ann. Math. Statist., 15 (1944) 297-306.
[37] J. Kiefer, Optimum experimental designs, J. Roy. Statist. Soc., Ser. B, 21 (1959) 272-319.
[38] D. Raghavarao, Constructions and Combinatorial Problems in Design of Experiments, John Wiley, N. Y. 1971.
[39] M. Hall, Jr., Combinatorial Theory, Blaisdell, Waltham, Mass., 1967.
[40] R. J. Turyn, Hadamard matrices, Baumert-Hall units, four-symbol sequences, pulse compression and surface wave encodings, J. Combinatorial Theory, 16A (1974) 313-333.
[41] W. D. Wallis, A. P. Street and J. S. Wallis, Combinatorics: Room Squares, Sum-Free Sets, Hadamard Matrices, Lecture Notes in Mathematics 292, Springer-Verlag, New York, 1972.
[42] L. D. Baumert, Cyclic Hadamard matrices, JPL Space Programs Summary, Vol. 37-43-IV (1967) 311-314.
[43] L. D. Baumert, Cyclic Difference Sets, Lecture Notes in Math. 182, Springer-Verlag, New York, 1971.
[44] S. W. Golomb, editor, Digital Communications with Space Applications, Prentice-Hall, Englewood Cliffs, N. J., 1964.
[45] R. Thoene and S. W. Golomb, Search for cyclic Hadamard matrices, JPL Space Programs Summary, Vol. 37-40-IV (1966) 207-208.
[46] E. D. Nelson and M. L. Fredman, Hadamard spectroscopy, J. Opt. Soc. Amer., 60 (1970) 1664-1669.
[47] R. D. Swift, R. B. Wattson, J. A. Decker, Jr., R. Paganetti, and M. Harwit, Hadamard transform imager and imaging spectrometer, Appl. Opt., 15 (1976) 1595-1609.

Codes and Designs

Combinatorial designs formed by patterns of subsets provide insight into the existence of useful codes.

IAN F. BLAKE
University of Waterloo
Waterloo, Ontario, Canada N2L 3G1

Algebraic coding theory is now over twenty-five years old. It originally arose in response to the celebrated noisy coding theorem of Shannon [1] which bounded the error performance obtainable on a discrete channel, but gave little indication of how that performance might be achieved. Since its inception the interactions of coding theory with other branches of mathematics, particularly group theory and combinatorics, have continually deepened. The result is a fascinating subject drawing results and techniques from a variety of areas to investigate a problem of some practical importance.

In this article I plan to consider the connection between coding and certain combinatorial designs. I first establish some fundamental notions of coding, constructing many of the codes that will be needed later. Much of the discussion will be restricted to binary codes since that is the case of most interest to us. The weight enumeration of codes is important in finding designs among the codes, as well as being an interesting subject in its own right, so some relevant results of this theory are given. We then discuss combinatorial designs of special interest from a coding point of view, and prove several of their properties. The results of these sections are then brought together to prove a collection of theorems due to E. F. Assmus and H. F. Mattson which indicate how designs can be obtained from "good" codes. In the final section of the paper I have tried to indicate generalizations and recent advances on these topics, and have included a discussion of the literature.

The scope of this article is somewhat ambitious, so I have been selective as to what is proven and what is merely stated. My aim is to give as much of the flavor of the subject and to expose the reader to as many of the techniques of coding as possible, without becoming overly obsessed with any one topic. There have recently appeared two other papers which survey coding and combinatorics ([2], [3]); they are both deeper and more extensive treatises on the subject and readers who find this paper of interest should certainly read these other two. The paper by Assmus and Mattson [2] was particularly useful in preparing this article, which is largely a survey of their work.

Codes

To motivate the study of coding, we consider the simple model of a communication system shown in FIGURE 1. In this model a message m_i is chosen for transmission from among the M possible messages $m_1, m_2, \ldots, m_M$. The coder converts this message into a binary sequence of length N, $x_i = (x_{i1}, x_{i2}, \ldots, x_{iN})$, $x_{ij} \in \{0,1\}$, the codeword for message m_i. At each unit of time the channel accepts one of these binary symbols and transmits it to the receiver either correctly, with probability $1-p$, or in error with probability p. Such a channel is called a binary symmetric channel (BSC) and is shown in FIGURE 2. The decoder, which has a list of the possible codewords x_i, $i = 1, 2, \ldots, M$, on receiving the N digits y decides, according to the minimum probability of error criterion, which message it believes was sent. If it decides that message m_j was sent, when in fact m_i was sent, and $i \neq j$, then a decoding error was made, and this happens

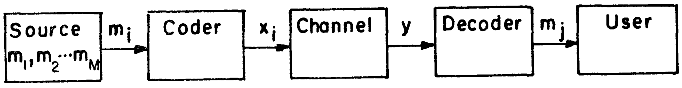


FIGURE 1

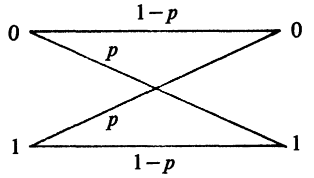


FIGURE 2

with probability P_e. The capacity of the BSC is $C = 1 + p\log_2 p + (1-p)\log_2(1-p)$ and the rate of the code R is defined as $R = (\log_2 M)/N$. The fundamental theorem of Shannon then states that for any $\delta > 0$, if $R < C$, then there will exist a code with rate $R' > R$ for N sufficiently large such that $P_e < \delta$. The implication is that we are able to communicate with as small an error probability as desired, keeping the rate fixed, by choosing a code length large enough. The cost of this performance is increased complexity and delay in the coding and decoding operations.

The proof of this remarkable theorem employed random coding arguments. Briefly stated, it chose the M codewords from among the 2^N possible at random, according to some probability distribution. The resulting probability of error was bounded over the ensemble of all possible codes of size M, and there must exist at least one code which satisfies this average bound. Of course, such a random coding procedure is not very useful for practical purposes and methods to construct codes whose performance met the derived bound were sought. The result is algebraic coding theory.

We begin by defining the basic structures and approaches to coding. Other approaches have been attempted with varying degrees of success and interest but have not flourished to the point where they are widely studied. The basic structure in which essentially all coding work takes place is that of a finite dimensional vector space over a finite field. Denote by F_q the finite field with q elements, q a power of a prime, which is unique up to isomorphism. At times we will need properties of this field which may not be well known but, rather than give an exposition of them now, we will summon them as needed. Let F_q^n denote the vector space of dimension n over F_q, realized as the set of n-tuples over F_q, which is also unique up to isomorphism.

The (Hamming) **weight** $w(x)$ of an element $x \in F_q^n$ is the number of non-zero coordinate positions of x and the (Hamming) **distance** $d(x,y)$ between elements x and y of F_q^n is the number of coordinate places in which they differ, and so $d(x,y) = w(x-y)$. An **(n,M,d,q) code** C is a subset of F_q^n for which $M = |C|$, the cardinality of C, and $d = \min\{d(x,y) | x,y \in C,\ x \neq y\}$. If C is a subspace of F_q^n of dimension k, it is referred to as a **linear code** or a **linear (n,k,d,q) code**. Usually q is fixed and d is unknown or not needed in a particular argument, and in such cases C is referred to simply as a linear (n,k) code. For a linear code $d = \min\{w(x) | x \in C,\ x \neq 0\}$, as is easily seen.

The problem of coding is, for a given n and M (and q), to pack the codewords in such a way as to maximize d. Conversely given n and d, the problem is to maximize M. If $e = [(d-1)/2]$ (where $[x]$ indicates the integer part of x), define the sphere of radius e about the point $x \in F_q^n$ by $S_e(x) = \{y \in F_q^n | d(x,y) \leqslant e\}$. Then clearly $|S_e(x)| = \sum_{j=0}^{e} \binom{n}{j}(q-1)^j$, since there are $(q-1)^j$ elements of F_q^n, differing from x in j given positions and $\binom{n}{j}$ ways of choosing these j positions, $0 \leqslant j \leqslant e$. In the packing problems just referred to, the set of all such spheres centered at the codewords are nonintersecting and it follows immediately that for an (n,M,d,q) code, $M|S_e(x)| \leqslant q^n$, where $e = [(d-1)/2]$. This result is referred to as the **sphere packing** or **Hamming-Rao bound**, and any code which meets this bound with equality is called **perfect**. Notice that d must be odd for a perfect code.

Since for any code with minimum distance d, the spheres of radius $e = [(d-1)/2]$ centered at the codewords are non-intersecting, we can conceptually appreciate the minimum distance decoding algorithm: if the received n-tuple $y \in F_q^n$ lies in the sphere of radius e about $c \in C$ then y is decoded to c. If fewer than e errors are made in transmission, then the decoding will be correct. If more than e errors are made in transmission, then either y will be decoded to the incorrect codeword (it lies in a sphere of radius e about some codeword not equal to the transmitted codeword) or else it lies in no sphere of radius e about a codeword in which case some other strategy, such as requesting a retransmission, is invoked. From these considerations the sphere packing nature of the coding problem is seen. In practice, of course, other techniques are used to effect the decoding and these are not considered here.

To proceed with the discussion of linear codes, we first observe that if C is an (n,k) linear code, then the code can be thought of as the row space of a $k \times n$ matrix over F_q, where the rows of G are chosen as any set of k linearly independent codewords of C. The matrix G, the **generator matrix** of the code, can be row reduced to a systematic form $G' = [I_k : A]$ where A is a $k \times (n-k)$ matrix. If the coding operation is described as the matrix product $c = iG'$, where i is a k-tuple information sequence over F_q and c, the codeword corresponding to it, then the first k positions of c form i, the remaining $(n-k)$ positions being parity checks on them. A code where the information bits appear explicitly is called **systematic**.

If C is a linear code, we can define the useful notion of a **dual code** (or space) in the usual way as

$$C' = \left\{ y \in F_q^n \,|\, (x,y) = \sum_{i=1}^{n} x_i y_i = 0, \forall x \in C \right\}$$

where the inner product is calculated in F_q. The dual code C' is itself a linear code. (The concept of the dual of a nonlinear code is noticeably absent.) For a real vector space, with the dual of a subspace defined in a similar manner, it is true that $C \cap C' = \{0\}$ and $\dim C + \dim C' = n$. For F_q^n however, it is only true that $\dim C + \dim C' = n$. For example if $C \subset F_2^4$ is

$$\begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix},$$

then $C = C'$, a peculiar property for those used to the geometrical notions of real spaces.

A generator matrix H for the dual C' of the linear (n,k) code C is an $(n-k) \times n$ matrix whose row space is C' and is called the **parity check** matrix for C. By definition it satisfies the equation $GH^T = 0$. If G is in the systematic form $[I_k : A]$ then $H = [-A^T : I_{n-k}]$. A useful property of the parity check matrix H to keep in mind is the following. If the codeword $x \in C$ is not all-zero, then $xH^T = 0$, but since xH^T is a linear combination of columns of H this implies a linear dependence relation among the columns of H, corresponding to the nonzero positions of x. As a consequence of this, if no set of $(d-1)$ columns of H are linearly dependent, the code C has distance at least d. We can also obtain a bound on the code from these arguments. Since the rank of H is at most $n-k$, it can at best have every set of $n-k$ columns independent and so $d-1 \leqslant n-k$ or $d \leqslant n-k+1$ which is the **Singleton bound** of coding. Any code which satisfies this bound with equality is usually referred to as an **optimal** code. Note the artful use of the dual code to deduce information about the code itself.

An example of these ideas might be helpful. Consider the binary linear (7,4) code C with parity check matrix

$$H = \begin{bmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix} \tag{1}$$

and notice that every nonzero 3-tuple appears as a column of H. Since any two columns are linearly independent, the minimum distance of C is 3. Another way of viewing this code (which is the original construction of Hamming [4]) is to think of each codeword in the form $(p_1, p_2, i_1, p_3, i_2, i_3, i_4)$ where i_1, i_2, i_3 and i_4 are information digits and p_1, p_2 and p_3 parity check digits. These digits satisfy the F_2 equations

$$p_1+i_1+i_2+i_4=0$$
$$p_2+i_1+i_3+i_4=0$$
$$p_3+i_2+i_3+i_4=0$$

For future reference the codewords of C are

000 0000	111 0000	110 0110	011 0011
110 1001	100 0011	101 1010	101 0101
010 1010	010 0101	011 1100	001 0110
100 1100	001 1001	000 1111	111 1111

Since

$$M|S_1(x)|=2^4\left(\sum_{i=0}^{1}\binom{7}{i}\right)=2^7=|F_2^7|$$

the code is perfect and called the binary (7,4) Hamming code.

It is not hard to generalize this example and describe the class of all Hamming codes which we might as well do now. Construct a $k\times n$ parity check matrix H over F_q by choosing as its columns all $(q^m-1)/(q-1)=n$ m-tuples over F_q such that no two columns are scalar multiples of one another. The code C with parity check matrix H has length $n=(q^m-1)/(q-1)$, dimension $n-m$ and distance 3 and is again perfect since

$$M|S_1(x)|=q^{n-m}\left(\sum_{i=0}^{1}\binom{n}{i}(q-1)^i\right)=q^n=|F_q^n|.$$

It has recently been shown ([5], [6]) that there are only two other linear perfect codes, which we will meet towards the end of this section, and that all other perfect codes will be nonlinear with exactly the same parameters (length, distance, and size) as the Hamming codes.

The coordinate positions of a linear code C of length n are often labelled with the elements $\{0,1,\ldots,n-1\}$. An operation which turns out to be useful is to extend C, to what we shall call C_e, by adding an overall parity check to make the sum of all coordinate positions zero. The extra position is often labelled with the symbol ∞ as a result of some computations on the code with linear fractional groups. If C is an (n,k) code, then C_e is an $(n+1,k)$ code and if H is the parity check matrix for C, then the parity check matrix for C_e is

$$\begin{array}{cc} & \begin{array}{ccccccccc} 0 & 1 & . & . & . & . & . & n-1 & \infty \end{array} \\ H'= & \left[\begin{array}{cccccccc|c} & & & H & & & & & 0 \\ & & & & & & & & 0 \\ & & & & & & & & \cdot \\ & & & & & & & & \cdot \\ & & & & & & & & \cdot \\ & & & & & & & & 0 \\ \hline 1 & 1 & . & . & . & . & . & 1 & 1 \end{array}\right]. \end{array}$$

Over nonbinary fields the appended position is sometimes filled with a constant times the sum of the other positions in order to satisfy certain considerations about the dual of the extended code. The extension of the (7,4) Hamming code C leads to a code C_e of length 8. The 7 vectors of weight 3 in C have weight 4 in C_e and those of weight 4 in C have zeros appended. Thus C_e has 14 vectors of weight 4, 1 of weight 0 and 1 of weight 8 and is an (8,4) code with distance 4.

The next important notion to introduce is that of a cyclic code. To do so, notice that by rearrangement of its columns, the parity check matrix H of (1) could be written as

$$H'=\begin{bmatrix} 1 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{bmatrix}$$

and in this form every cyclic shift of a codeword is also a codeword in both C and its dual C'.

(By a cyclic shift of $(c_0,c_1,\dots,c_{n-1})$ is meant $(c_{n-1},c_0,\dots,c_{n-3},c_{n-2})$). We define a cyclic code as a linear code with the property that every cyclic shift of a codeword is also a codeword, i.e., it is invariant under cyclic permutations of the coordinate positions. These codes have been by far the most widely studied and most of the interesting codes are either cyclic or extended cyclic codes.

In studying cyclic codes it is convenient to use a polynomial notation: to a codeword $c=(c_0,c_1,\dots,c_{n-1})$, we associate the polynomial $c(x)=c_0+c_1x+\cdots+c_{n-1}x^{n-1}$. A cyclic shift of this codeword corresponds to the polynomial $xc(x)$ modulo x^n-1. A cyclic code corresponds to an ideal in the polynomial ring $F_q[x]/(x^n-1)$. In such a ring every ideal is a principal ideal with a generator $g(x)$ which is necessarily a divisor of x^n-1. If the cyclic subspace is of dimension k, then the degree of $g(x)$, its generator polynomial, is $n-k$. The code C is thus described by the polynomial ideal

$$C=\{a(x)g(x)|\deg(a(x))\leqslant k-1, a(x)\in F_q[x]\}.$$

If x^n-1 has s irreducible factors over F_q, then there are precisely 2^s cyclic codes of length n over F_q, since any divisor of x^n-1 generates a cyclic code.

At this point we shall briefly review some facts about polynomials over finite fields. The multiplicative group F_q^* of F_q is cyclic and a generator of it is called a **primitive element**. Every finite field always has at least one primitive element and has, in fact, $\phi(q-1)$ primitive elements where ϕ is Euler's totient function. An **irreducible polynomial** over F_q is one which cannot be expressed as a product of two polynomials of lower degree. An irreducible polynomial of degree k over F_q always divides $x^{q^k}-x$ and this polynomial factors into the product of all irreducible polynomials whose degrees divide k. An irreducible polynomial $f(x)$ over F_q will be called **primitive** if $f(x)|x^{q^k-1}-1$ but $f(x)\nmid x^s-1$ for $s<q^k-1$. If β is an element of an extension field F_{q^k} of F_q, then the minimum polynomial $m_\beta(x)$ of β over F_q is that monic (coefficient of highest power of x is unity) polynomial over F_q of lowest degree which has β as a root. For any polynomial $f(x)\in F_q[x]$, $f(x^q)=f^q(x)$ and so if β is a root of $f(x)$, so is $\beta^q,\beta^{q^2},\dots$ and the set of all such roots are called **conjugates** of β. If $K=\{\beta,\beta^q,\beta^{q^2},\dots\beta^{q^{s-1}}\}$, where $\beta^{q^s}=\beta$ and $\beta^{q^i}\neq\beta$ for $0<i<s$, then $m_\beta(x)=\Pi_{i=0}^{s-1}(x-\beta^{q^i})$ and, since $m_\beta(x)|x^{q^k}-x$, s divides k. A monic primitive polynomial of degree k over F_q is the minimum polynomial of some primitive element in F_{q^k}.

We observed above that the $(7,4)$ binary Hamming code could always be made cyclic. In general if α is a primitive element of F_2^m, then the Hamming code of length 2^m-1 and dimension 2^m-m-1 has $m_\alpha(x)$, a primitive polynomial, as a generator polynomial. For Hamming codes over F_q, if α is a primitive element of F_{q^m}, then the generator polynomial of the $((q^m-1)/(q-1),\ (q^m-1)/(q-1)-m)$ Hamming code is the minimum polynomial of the element α^{q-1}, provided $(q-1)$ is relatively prime to m ([7]).

If $g(x)$ is the generator polynomial of a cyclic (n,k) code C and $g(x)h(x)=x^n-1$, then C' is also cyclic and, as is easily verified, has $x^{n-k}h(1/x)$ as a generator polynomial.

It will be useful to describe another class of codes for later reference, the quadratic residue codes. Let n be an odd prime and α an nth root of unity in an extension field of F_q. Let R be the set of quadratic residues in F_n, i.e., $R=\{x\in F_n|\exists a: a^2=x, x\neq 0\}$, and let $\bar{R}$ be the set of nonzero elements of F_n not in R. Define the two polynomials

$$g_1(x)=\prod_{r\in R}(x-\alpha^r)\qquad g_2(x)=\prod_{r\in\bar{R}}(x-\alpha^r)$$

and assume that $q^{(n-2)/2}\equiv 1$ (modulo n), which implies that q is a quadratic residue in F_n. In the binary case it is sufficient that $n=\pm 1$ modulo 8. Since a residue times a residue is a residue, $qR=R$ and similarly $q\bar{R}=\bar{R}$. Thus $g_1(x)$ contains as roots the conjugate of every root, and is a polynomial over F_q, as is $g_2(x)$. The codes generated by $g_1(x)$ and $g_2(x)$ are called quadratic residue codes of length n with dimension $(n+1)/2$.

We will be interested in two quadratic residue codes in particular. Of course $x^n-1=(x-1)g_1(x)g_2(x)$ and $g_i(x)|x^n-1$, $i=1,2$. The first code of interest is binary, has length 23 and generator polynomial either

$$g_1(x)=1+x+x^5+x^6+x^7+x^9+x^{11} \quad \text{or} \quad g_2(x)=1+x^2+x^4+x^5+x^6+x^{10}+x^{11},$$

and these polynomials are irreducible over F_2. It can be shown that the cyclic (23,12) code generated by either $g_1(x)$ or $g_2(x)$ has minimum distance 7 and, since $|S_3(x)|=\sum_{i=0}^{3}\binom{23}{i}=2^{11}$ and $M|S_3(x)|=2^{12}\cdot 2^{11}=2^{23}=|F_2{}^{23}|$, the code is perfect.

The other code of interest has length $n=11$ over $F_3=\{0,1,-1\}$. The polynomials $g_1(x)$ and $g_2(x)$ are

$$g_1(x)=x^5-x^3+x^2-x-1 \quad \text{and} \quad g_2(x)=x^5+x^4-x^3+x^2-1$$

and either of these generates an (11,6) code with distance $d=5$. Since $|S_2(x)|=\sum_{i=0}^{2}\binom{11}{i}2^i=3^5$ and $M|S_2(x)|=3^6\cdot 3^5=3^{11}=|F_3^{11}|$ the code is, again, perfect.

These two quadratic residue codes are the only perfect codes (either linear or nonlinear) with $n>d>3$ which can exist. As an interesting historical sidelight, they were both found in 1949 by Golay [8] who first established the possibility of their existence by searching the Pascal triangle of binomial coefficients (or, in the case of the ternary code, a modified version of it), and gave parity check matrices for them without any explanation of how they were arrived at. As mentioned earlier the only other perfect codes which can exist are nonlinear codes with $d=3$ with the same length and size as the Hamming codes. In subsequent sections we will be interested in the extended Golay codes which have remarkable combinatorial structure. The whole class of quadratic residue codes has been extensively studied, yielding bounds on their minimum distance and characterizations of their automorphism groups. The above information will, however, be sufficient for our purposes.

The Weight Enumeration of Codes

Intuitively, the distance structure of the code determines the quality and hence the effectiveness of the code. From this information parameters such as the probability of error when using the code on a discrete channel can be computed. In general the task of completely describing the distance structures is too complicated, so we settle for the simpler one of determining the number of codewords at distance i from codeword $x\in C$, denoted by $A_i(x)$. Clearly $A_i(0)$ is just the number of codewords of weight i in the code. If we translate the code by adding a codeword y to every codeword, the code is left invariant and we conclude that $A_i(y)=A_i(0)=A_i$. In other words, if we imagine ourselves standing on a codeword and looking around at other codewords, then the view will be the same from every codeword. We define the **weight enumerator** of a linear (n,k) code by the bivariate polynomial

$$A(x,y)=\sum_{i=0}^{n}A_ix^iy^{n-i}, \qquad A_0=1, \qquad \sum_{i=1}^{n}A_i=q^k-1$$

where, again, A_i is the number of codewords of weight i or the number of codewords at distance i from a given codeword. The weight enumerator does not uniquely define the code in that two different codes can have the same weight enumerator. It is, nonetheless a convenient and useful descriptor for a code.

As examples of weight enumerators, we consider the codes already introduced. The (7,4) Hamming code with distance 3 has the weight enumerator

$$h(x,y)=x^7+7x^4y^3+7x^3y^4+y^7$$

and the (8,4) extension of this code has the weight enumerator

$$H(x,y)=x^8+14x^4y^4+y^8.$$

The (23,12) cyclic binary Golay code with minimum distance 7 has the weight enumerator

$$g(x,y)=x^{23}+253x^{16}y^7+506x^{15}y^8+1288x^{12}y^{11}+1288x^{11}y^{12}+506x^8y^{15}+253x^7y^{16}+y^{23}$$

and the (24,12) extension of this code has the weight enumerator

$$G(x,y)=x^{24}+759x^{16}y^{8}+2576x^{12}y^{12}+759x^{8}y^{16}+y^{24}.$$

Notice that the weight enumerators of the (8,4) and (24,12) extended codes imply that they are self-dual.

One of the cornerstones of coding theory, due to F. J. MacWilliams, relates the weight enumerator of a linear (n,k) code C to that of its dual C'. Specifically if $A(x,y)$ is the weight enumerator of C and $A'(x,y)=\sum_{i=0}^{n} A_i' x^i y^{n-i}$ that of C', then the MacWilliams identities state that

$$A(x,y)=\left(\frac{1}{q^{n-k}}\right)A'(y-x,y+(q-1)x). \tag{2}$$

An equivalent form of this polynomial relation, which can be obtained by expanding the polynomials and comparing coefficients, is

$$\sum_{i=0}^{n-j}\binom{n-i}{j}A_i=q^{k-j}\sum_{i=0}^{j}\binom{n-i}{n-j}A_i' \qquad j=0,1,\ldots,n.$$

The matrix $\lambda=(\lambda_{ij})$, where $\lambda_{ij}=\binom{n-i}{j}$, $0\leqslant i,j\leqslant n$, can be reduced to a van der Monde matrix by an obvious sequence of elementary row operations and hence is nonsingular. The fact that the weight enumerator of a code uniquely determines that of its dual is often useful.

An immediate and important consequence of the MacWilliams identities for our requirements is the following observation. If d' is the minimum distance of C', then

$$\sum_{i=0}^{n-j}\binom{n-i}{j}A_i=q^{k-j}\binom{n}{n-j} \qquad j=0,1,\ldots,d'-1 \tag{3}$$

since $A_0'=1$ and $A_i'=0$, $1\leqslant i\leqslant d'-1$. If the code C has only s nonzero weights and $s\leqslant d'$, then this set of d' equations in s unknowns has a unique solution.

As an example of this situation, let C be an optimal (n,k) code over F_q; i.e., its minimum distance d is $(n-k+1)$, and let G and H be a generator matrix and parity check matrix for it, respectively. Now every set of k columns of G must be linearly independent, since otherwise there would be a nonzero codeword with zeros in these k positions, giving a codeword of weight at most $(n-k)$, a contradiction of the fact that its weight is $(n-k+1)$. This fact implies that the minimum distance of C', an $(n,n-k)$ code, is at least $k+1$, and this says that C' is also optimal. With $d'=k+1$, equations (3) form a set of k equations in the unknowns $A_d,A_{d+1},\ldots,A_n$ and since $d=n-k+1$, we have equations in k unknowns which can be solved uniquely. It is not a difficult matter to calculate (see **[15]**) that if C is an (n,k) optimal code over F_q, then

$$A_{n-i}=\sum_{r=i}^{k-1}(-1)^{r-i}\binom{r}{i}\binom{n}{r}(q^{k-r}-1) \qquad i=0,1,\ldots k-1. \tag{4}$$

A simple example (**[9]**) of an optimal code over F_3 is the (4,2) Hamming code, of distance 3, with generator matrix

$$G=\begin{bmatrix}1 & 1 & 1 & 0\\ 0 & 1 & 2 & 1\end{bmatrix}$$

and weight enumerator $A(x,y)=8x^3y+y^4$, as can be verified by generating the row space of G.

A case of particular interest from a combinatorial point of view occurs when the code is self-dual, i.e., $C=C'$. A binary self-dual code must have the weight of every codeword divisible by two. It can happen, however, that the weight of every codeword is divisible by 4 as the extended (8,4) Hamming and (24,12) Golay codes indicate. An interesting result due to Gleason and Pierce (and quoted in **[10]**) states that if C is a self-dual code over F_q for which every codeword is divisible by c, then only four cases are possible, namely, $(q,c)=(2,2)$, $(2,4)$, $(3,3)$ and $(4,2)$. This is quite a deep result, requiring substantial machinery to prove.

The condition that a code is self-dual places a severe restriction on the form that its weight enumerator may have. We consider only the binary case for which we notice from (2) that if $A(x,y)$ is the weight enumerator of an $(n,n/2)$ self-dual code (n is even) then

$$A(x,y)=\frac{1}{2^{n/2}}A(y-x,y+x)=A\left(\frac{y-x}{\sqrt{2}},\frac{y+x}{\sqrt{2}}\right).$$

In addition, since only even weights can appear in the code, $A(x,y)=A(x,-y)$. It follows that $A(x,y)$ is invariant under the transformations

$$\frac{1}{\sqrt{2}}\begin{bmatrix}-1 & 1\\ 1 & 1\end{bmatrix} \quad\text{and}\quad \begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix}$$

and hence is invariant under the group of matrices generated by these two, which happens to be the dihedral group of order 16 [9]. Using the theory of invariants for such polynomials, it is possible to show the following: if C is a self-dual binary code of length n, then its weight enumerator is a linear sum of products of the polynomials $f(x,y)=x^2+y^2$ and $H(x,y)=x^8+14x^4y^4+y^8$ (the weight enumerator of the extended Hamming code). In other words,

$$A(x,y)=\sum_{2r+8s=n} a_{rs}f(x,y)^r H(x,y)^s.$$

If we impose the additional restriction that every codeword is divisible by 4, rather than just having even weight, then

$$A(x,y)=\sum_{8r+24s=n} a_{rs}H(x,y)^r G(x,y)^s$$

where $G(x,y)$ is the weight enumerator of the extended Golay code. In this case the length of the code is always divisible by 8. Similar results are available for the other two cases, i.e., $q=3$ with all weights divisible by 3, and $q=4$ with all weights divisible by 2.

Combinatorial Designs

The subject of combinatorial designs is rather comprehensive. In fact I intend to discuss only two such designs, namely orthogonal arrays and t-designs. This does not mean that other combinatorial objects, such as projective planes, projective and Euclidean geometries, Latin squares, Hadamard matrices etc., are not important and relevant to coding work. It only means that in this article I prefer to restrict attention to these two.

An (M,n,q,r,μ) **orthogonal array** is an $M\times n$ array over an alphabet with q symbols with the property that any r of its columns contain each of the q^r r-tuples of the alphabet exactly μ times. The parameter r is called the strength of the array and μ the index of the array; of course, $M=\mu q^r$.

An example of an orthogonal array is the ternary (4,2) code encountered earlier (it is actually a Hamming code over F_3):

$$\begin{array}{cccc}
0 & 0 & 0 & 0\\
1 & 1 & 1 & 0\\
0 & 1 & 2 & 1\\
2 & 2 & 2 & 0\\
0 & 2 & 1 & 2\\
1 & 2 & 0 & 1\\
1 & 0 & 2 & 2\\
2 & 0 & 1 & 1\\
2 & 1 & 0 & 2
\end{array}$$

This is an example of a (9,4,3,2,1) orthogonal array, since running down any two of its columns, every ordered pair of elements from F_3 is encountered exactly once. These arrays are widely studied, but since we are only interested in their relationship to codes, we defer further discussion of them to the next section.

The other combinatorial object of interest to us is the t-design. A t-(v,k,λ) ***t*-design** is a collection of k-sets, called blocks, from a set V of v distinct elements such that any t-set of V appears in precisely λ blocks. A 2-design is more commonly referred to as a **balanced incomplete block design**. When $\lambda=1$ a t-design is referred to as a **Steiner system** and when $\lambda=1$ and $k=3$, it is called a **Steiner triple system**.

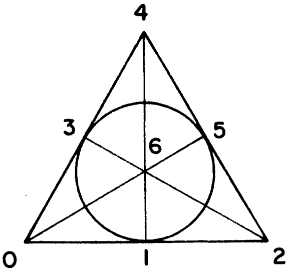


FIGURE 3

The Fano geometry, shown in FIGURE 3, is a simple example of a 2-design. In this design the set V is $\{0,1,2,3,4,5,6\}$ and each block contains the numbers on a line, the circle counting as a line. The blocks are $\{1,3,5\}$, $\{0,3,4\}$, $\{0,1,2\}$, $\{0,5,6\}$, $\{1,4,6\}$, $\{2,3,6\}$ and $\{2,4,5\}$, and any 2-set is contained in precisely one block. For later reference we note that we can identify each block of this system with a binary 7-tuple with the coordinates labelled with elements of V and containing a one in the positions of the block and zeros elsewhere. The Fano geometry gives the following correspondence:

$\{1,3,5\}$	0 1 0 1 0 1 0	$\{1,4,6\}$	0 1 0 0 1 0 1
$\{0,3,4\}$	1 0 0 1 1 0 0	$\{2,3,6\}$	0 0 1 1 0 0 1
$\{0,1,2\}$	1 1 1 0 0 0 0	$\{2,4,5\}$	0 0 1 0 1 1 0
$\{0,5,6\}$	1 0 0 0 0 1 1		

Notice that these 7-tuples are precisely the weight 3 codewords of the (7,4) Hamming code described earlier.

Much of the next section revolves around certain properties of t-designs, so because it is convenient, we will develop them here. By definition, the number of blocks in a t-(v,k,λ) design containing a given t-set is λ. An easy calculation shows that if λ_i is the number of blocks containing a given i-set, then

$$\lambda_i=\frac{\lambda\binom{v-i}{t-i}}{\binom{k-i}{t-i}},\qquad i=0,1,\dots t,$$

where $\lambda_t=\lambda$ and λ_0 is the number of blocks in the design.

The following counting argument, which appears in **[11]** and is actually a generalization of one given in **[12]**, gives useful structural information on a t-design. Let S be an arbitrary s-set of V and let $y_i(S)$ be the number of blocks of the t-design which intersects S in exactly i elements. We perform a double counting argument on the number of times an r-set appears. The s-set contains $\binom{s}{r}$ r-sets and each of these appears in λ_r blocks. On the other hand, if I is an i-set occurring as the intersection between S and a block, then it contributes $\binom{i}{r}$ r-sets as the intersection between S and a block, then it contributes $\binom{i}{r}$ r-sets to the count and $\binom{i}{r}y_i(S)$ such r-sets arise in this manner. Hence,

$$\sum_{i=r}^{s}\binom{i}{r}y_i(S)=\binom{s}{r}\lambda_r\quad\text{for}\quad r=0,1,\dots,\min(s,t).\tag{5}$$

This is a set of $1+\min(s,t)$ equations in $(s+1)$ unknowns and when $s\leqslant t$, they have the unique solutions, independent of S,

$$y_l(S)=y_l=(-1)^l\sum_{r=0}^{s}(-1)^r\binom{r}{l}\binom{s}{r}\lambda_r,\qquad l=0,1,\dots,s.\tag{6}$$

If $s>t$, the equations (5) will not in general have a unique solution, but in the particular case that $s=t+1$, we can multiply the "r" equation by $(-1)^r$ and, by summing over r, obtain

$$y_0(S)+(-1)^t y_{t+1}(S)=\sum_{r=0}^{t}(-1)^r\binom{t+1}{r}\lambda_r. \tag{7}$$

Since the right hand side of this equation is independent of S, we conclude that the left hand side is independent of S also.

Our first use of these relationships is to obtain a new design from a given one. Denote by $\mathfrak{B}$ the set of blocks in a t-(v,k,λ) design and by $\overline{\mathfrak{B}}$ the complements of these blocks, i.e., $\overline{\mathfrak{B}}=\{(V\setminus B):B\in\mathfrak{B}\}$. We claim that this is a t-design. To prove this, it is only necessary to show that the number of blocks containing a t-set is a constant, independent of the particular t-set chosen. But this number is just the number of blocks of $\mathfrak{B}$ containing no elements of the t-set; by equation (6) with $l=0$ and $s=t$, this number is y_0 and is indeed independent of the particular t-set chosen. Consequently the blocks in $\overline{\mathfrak{B}}$ form a t-$(v,v-k,y_0)$ t-design.

As an example of the complement of a design, the complement of the design obtained from the Fano geometry has the blocks $\{0,2,4,6\}$, $\{1,2,5,6\}$, $\{3,4,5,6\}$, $\{1,2,3,4\}$, $\{0,2,3,5\}$, $\{0,1,4,5\}$ and $\{0,1,3,6\}$. Since the (7,4) Hamming code contained the vector with all 1's, the complement of the weight 3 codewords are the weight 4 codewords and these correspond to the above blocks. It is a 2-(7,4,2) 2-design.

In the previous section we described how a code could be extended in a natural way by adding an overall parity check. Under certain circumstances we can also extend a t-design, and the two operations are analogous in this case. Suppose that $(V,\mathfrak{B})$ is a t-$(2k+1,k,\lambda)$ design with t even. We now adjoin the element ∞ to V to form $V'=V\cup\{\infty\}$. If we denote by $\overline{\mathfrak{B}}$ the set of complements of the blocks of $\mathfrak{B}$ in V and by B' the set $\{B\cup\{\infty\}:B\in\mathfrak{B}\}$, then we claim that the set of blocks $\overline{\mathfrak{B}}\cup\mathfrak{B}'$ is a $(t+1)$-$(2k+2,k+1,\lambda)$ design. The proof is straightforward. Consider a $(t+1)$-set T. If T contains ∞, then it is contained in no blocks of $\overline{\mathfrak{B}}$ and is contained in exactly λ blocks of $\mathfrak{B}'$, so T is contained in exactly λ blocks of $\overline{\mathfrak{B}}\cup\mathfrak{B}'$. If T does not contain ∞, then it is contained in exactly $y_0(T)$ blocks of $\overline{\mathfrak{B}}$ and exactly $y_{t+1}(T)$ blocks of $\mathfrak{B}'$ and, by equation (7), noting that t is even, we have

$$y_0+y_{t+1}=\sum_{r=0}^{t}(-1)^r\binom{t+1}{r}\frac{\lambda\binom{2k+1-r}{t-r}}{\binom{k-r}{t-r}}=\frac{\lambda}{\binom{2k+1-t}{k-t}}\sum_{r=0}^{t}(-1)^r\binom{t+1}{r}\binom{2k+1-r}{k-r}=\lambda.$$

Consequently $\overline{\mathfrak{B}}\cup\mathfrak{B}'$ is a $(t+1)$-$(2k+2,k+1,\lambda)$ $(t+1)$-design on V'.

The extended design obtained from the Fano geometry is a 3-(8,4,1) design with the blocks

$$\begin{array}{llll}
\{\infty,1,3,5\} & \{\infty,2,3,6\} & \{0,2,4,6\} & \{0,2,3,5\}\\
\{\infty,0,3,4\} & \{\infty,2,4,5\} & \{1,2,5,6\} & \{0,1,4,5\}\\
\{\infty,0,1,2\} & \{\infty,0,5,6\} & \{3,4,5,6\} & \{0,1,3,6\}\\
\{\infty,1,4,6\} & & \{1,2,3,4\} &
\end{array}$$

If we form the corresponding binary 8-tuples, they would be the weight 4 codewords in the extended (8,4) Hamming code.

We have established a few of the relevant properties of t-designs and, not incidentally, have used codes to illustrate them. We now consider the situation in greater generality and, in the next section, give the theorems of Assmus and Mattson which form the basis of what is known on the relationships between codes and designs.

The Assmus-Mattson Theorems

Intuitively, one would guess that a "good" code, one that is "densely packed", should in most cases be highly structured and, for the lower block lengths, this certainly seems to be true. Perhaps the extreme examples of this are the perfect codes where the spheres of radius e cover F_q^n, yet do not intersect. We will show in a moment that such codes do in fact yield t-designs and that the examples of the previous sections are not just isolated cases. There is a rather simple relationship between orthogonal arrays and codes and, before tackling the more interesting case of t-designs, we dispose of it. All the theorems and proofs of this section, except for the last paragraph, are due to Assmus and Mattson ([13], [14]).

Let C be a linear (n,k) code and C' the dual $(n,n-k)$ code with distance d'. If G is a generator matrix for C, and hence a parity check matrix for C', then since the minimum distance of C' is d', any set of $(d'-1)$ columns of G are linearly independent. If D is a particular collection of $(d'-1)$ columns of G, then we can form a new generator matrix G^* for C, by elementary row operations on G, such that G^* contains the identity matrix in the upper $(d'-1)$ rows of G^* in the positions of D. It follows immediately that in the row space of G^* over F_q, and hence in the row space of G, every $(d'-1)$-tuple over F_q appears the same number of times in D, namely $q^{k+1-d'}$. Thus any linear (n,k) code over F_q with dual distance d' forms a $(q^k,n,q,d'-1,q^{k+1-d'})$ orthogonal array. This observation, which appears unassuming, has actually been used very effectively in coding. (We will have some more comments on it in the section on Further Reading below.)

Let us turn now to the relationship between t-designs and codes. If we label the coordinate positions of a code of length n with n distinct elements, then a codeword of weight w can be identified with the set of coordinate positions where it is nonzero, this set being referred to as the **support** of the codeword. The phrase "the set of vectors of weight w in a code supports a t-design" will then mean that if we form the blocks which are the w-sets of the supports of vectors of weight w, the set of distinct blocks obtained will form a t-design. Over F_q every scalar multiple of a codeword is a codeword and so each support appears at least $(q-1)$ times, but to form the design we take only distinct representatives. The first theorem obtaining t-designs from codes is trivial but sets the stage.

THEOREM 1 [14]. *A linear (n,k) code over F_q is optimal if and only if the maximum weight vectors support a trivial design.*

Proof. Suppose first that C is a linear (n,k) code with distance d and that for every d-set of coordinate positions, there is a codeword with this support. We wish to show that the code is optimal, i.e., that $d=n-k+1$. Let C_1 be the subcode of C spanned by all the minimum weight codewords and C_1' its dual. Since C_1 has distance d, every set of $d-1$ columns of any generator matrix of C_1' is independent, and the dimension is $(d-1)$. The dimension of C_1 is therefore $n-(d-1)$ which is less than or equal to $k: n-d+1 \leqslant k$. But we saw earlier that $k \leqslant n-d+1$ and so $k=n-d+1$.

Now suppose we have an optimal linear (n,k) code C with minimum distance $d=n-k+1$. We have to show that every d-set of coordinate positions is the support of a minimum weight codeword. But from equation (4) the number of minimum weight codewords is $A_{n-k+1}=A_d=\binom{n}{d}(q-1)$ and since there are $\binom{n}{d}$ possible supports, there must be a codeword (and its $(q-1)$ scalar multiples) for every such possible support since, if two codewords of weight d had the same support and were not scalar multiples, among their linear combinations would be codewords of weight less than d—a contradiction.

The next two theorems delve a little deeper and are concerned with perfect codes and their extensions.

THEOREM 2 [13]. *A linear code C with minimum distance $d=2e+1$ is perfect if and only if its minimum weight codewords support an $(e+1)$-$(n,d,(q-1)^e)$ design.*

Proof. Suppose first that the code C is perfect, implying that for any $x \in F_q^n$ there is a unique codeword $c \in C$ such that x is in $S_e(c)$, the sphere of radius e about c. Consider any $(e+1)$-set E of coordinate places and note that there are $(q-1)^{e+1}$ elements of F_q^n with support E. Each of these elements is within distance e of some codeword which implies that the codeword is of weight at most $d=2e+1$ and hence equal to $2e+1$. Thus the codeword agrees with the $(e+1)$-tuple on E. As above, any two codewords of weight d must be scalar multiples of one another. Thus there are, up to scalar multiples, $(q-1)^e$ codewords with support which covers E. Since this number is independent of E, these supports form the design.

Now suppose that the supports of the minimum weight codewords form an $(e+1)$-$(n,d,(q-1)^e)$ design. We want to show that every $x \in F_q^n$ is within distance e of some codeword. Suppose $x \in F_q^n$ is the element of smallest weight not within distance e of some codeword (and so $w(x) \geqslant e+1$ as the all zero n-tuple is an element of C). Let E be an $(e+1)$-set of coordinates from the support of x. Since the codewords of weight d form an $(e+1)$-design, there are $(q-1)^e$ blocks of the design containing E and, putting back scalar multiples, $(q-1)^{e+1}$ codewords of weight $d=2e+1$ with support containing E. Exactly one of these must be identical to x on E and so the weight of $x-c$ is at most $w(x)-1$. However, if x is at a distance at least $(e+1)$ from a codeword, then $x-c$ is at a distance at least $(e+1)$ from a codeword and so $x-c$ is a vector of weight at most $w(x)-1$ and distance at least $(e+1)$ from a codeword. This contradicts the choice of x, which completes the proof.

Over the field F_2 this theorem yields Steiner systems and, in this case, we can obtain a restricted version of it for the extended code C_e.

THEOREM 3. [13]. *Let C_e be an extended perfect linear code over F_2 of length $n+1$ and distance $d+1=2e+2$. Then its set of minimum weight vectors support an $(e+2)$-$((n+1),(d+1),1)$ design.*

Proof. From Theorem 2 the vectors of weight $d=2e+1$ in C form a Steiner system. Denote the added coordinate position containing the overall parity check by ∞ and consider an $(e+2)$-set E of coordinate positions. If E contains ∞ then, since d is odd, exactly one codeword of weight d in C covers $E\backslash\{\infty\}$ and hence exactly one word in C_e covers E. If E does not contain ∞, then it can at most be contained in the support of a vector of weight d in C (or equivalently, the support of a vector of weight $(d+1)$ in C_e containing ∞). If it is not contained in such a support, we claim it must be contained in the support of a vector of weight $(d+1)$ in C. The binary vector of weight $(e+2)$ equivalent to E is within a sphere of radius e about some codeword, and, by assumption, this codeword cannot be of weight d. It follows immediately that it must be of weight $2e+2$ and is unique and its support covers E.

This last theorem actually applies to nonlinear codes as well. We briefly examine the t-designs obtained from the perfect codes mentioned earlier. The Hamming codes over F_q have parameters $n=(q^m-1)/(q-1)$, $k=n-m$, $d=3$ and, by Theorem 2, yield 2-$(n, 3, (q-1))$ 2-designs. In the case that $q=2$ these are Steiner triple systems; the extended code has minimum distance 4 and the codewords of weight 4 yield a 3-$(n+1, 4, 1)$ design, called a Steiner quadruple system. The (11, 6) Golay code over F_3 has minimum distance 5 and the codewords of weight 5 support a 3-(11, 5, 4) 3-design. In fact, this design is actually a 4-(11, 5, 1) Steiner system, the theorem not being quite strong enough to give this result. (The following theorem will remedy this.) If we extend this code by adding a coordinate position containing minus the sum of the other coordinate positions, a (12, 6) code is obtained with minimum distance 6. The codewords of weight 6 form a 5-(12, 6, 1) Steiner system. The binary (23, 12) Golay code has minimum distance 7 and the codewords of weight 7 yield a 4-(23, 7, 1) Steiner system. The extension of this code yields a (24, 12) code with minimum distance 8 and a 5-(24, 8, 1) Steiner system.

The two Steiner systems which are 5-designs, obtained from the two Golay codes, are the only ones known. Many other 5-designs with λ greater than one are known, but there are no known t-designs (Steiner systems or not) with $t \geqslant 6$. Informed opinion seems equally divided as to whether or not they exist.

The following theorem is perhaps the focal point of much of the work in this area. The arguments are simple yet powerful and involve a clever use of the MacWilliams identities. Before giving the theorem we clear up one point. In the previous theorems we used the fact that two codewords x, y of weight d with the same support in a linear code must be scalar multiples of one another since otherwise we could find $\alpha \in F_q$ such that $0<w(x-\alpha y)<d$, contradicting the fact that the code has minimum distance d. For the next theorem we need an extension of this property. First observe that if $w(x)=v$, then x contains at least $[v/(q-1)]+1$ like nonzero

elements of F_q. If x and y are two codewords of weight v with the same support, in a linear code with minimum distance d, then it is possible to find an $\alpha \in F_q$ such that $w(x-\alpha y) \leqslant v - ([v/(q-1)]+1)$. If this quantity is less than d, then x and y would have to be scalar multiples of one another. For the final theorem we let C be a linear (n,k) code with minimum distance d and C' the dual $(n,n-k)$ code with minimum distance e. Denote by v_0 and w_0 the largest integers satisfying the inequalities

$$v_0-\left(\left[\frac{v_0}{q-1}\right]+1\right)<d \quad \text{and} \quad w_0-\left(\left[\frac{w_0}{q-1}\right]\right)<e$$

respectively. For binary codes, $v_0 = w_0 = n$. The important point to keep in mind is that two codewords in C of weight less than or equal to v_0 with the same support are scalar multiples of one another and similarly for codewords of weight less than or equal to w_0 in C'.

THEOREM 4 [14]. *Suppose that the number of nonzero weights of C' which are less than or equal to $n-t$ is itself less than or equal to $d-t>0$. Then, for each weight v, $d \leqslant v \leqslant v_0$, the vectors of weight v in C yield a t-design and for each weight w, $e \leqslant w \leqslant$ min $(n-t, w_0)$, the vectors of weight w in C' yield a t-design.*

Proof. The statement about C' is easier to prove than that about C and we begin with it. The proof will show that the complement of the supports of vectors of weight w in C' form a t-design and so the supports themselves form a t-design.

Designs from C'. Let T be a t-set of coordinates and let C^T be the code of length $(n-t)$ obtained by deleting positions of T. Let $C'^{0@T}$ be the code obtained by considering only those vectors in C' with zeros in positions of T. Now C^T and $C'^{0@T}$ are certainly orthogonal and $C'^{0@T}$ is an $(n-t,\ n-t-k)$ code. By assumption $d>t$ and if $x^T, y^T \in C^T$ were identical, then the vectors $x, y \in C$ corresponding to x^T, y^T would have distance at most $t<d$, a contradiction. It follows that $|C^T| = q^k$ and the dual of C^T is $C'^{0@T}$. Now let $W = \{w_1, w_2, \ldots, w_{d-t}\}$ be the possible nonzero weights of $C'^{0@T}$ and note that the minimum weight of C^T is at least $d-t$. Applying the MacWilliams identities to C and C' yields

$$\sum_{j\in W}\binom{n-t-j}{\mu}A_j'^{0@T}=q^{n-t-k-\mu}\binom{n-t}{\mu}-\binom{n-t}{\mu}, \qquad \mu=0,1,\ldots,d-t-1,$$

a set of $d-t$ equations in at most $d-t$ unknowns, where $A_j'^{0@T}$ is the number of codewords of weight j in $C'^{0@T}$. These equations can be solved uniquely for the weight distribution of $C'^{0@T}$ and from this the weight distribution of C^T can be obtained and both of these are independent of the T set chosen.

Now let E_v be the supports of all codewords of weight v in C', $v \leqslant \min(w_0, n-t)$, and let $\bar{E}_v$ be the complements of these supports. The number of sets in $\bar{E}_v$ containing T is just $1/(q-1)$ times the number of codewords in $C'^{0@T}$ of weight v, a number which we have seen is independent of T. Thus the $(n-v)$-sets of $\bar{E}_v$ form a t-design and hence by a previously established property, the v-sets of E_v form a t-design.

Designs from C. Let D_d be the set of supports of codewords of weight d in C. The number of d-sets in D_d containing a given t-set of T is $1/(q-1)$ times the number of vectors of weight $(d-t)$ in C^T, which is independent of T. The d-sets of D_d form a t-design. To establish the result for the weights $d<v\leqslant v_0$, we use induction and suppose that supports of codewords of all weights $d \leqslant v < v' \leqslant v_0$ yield t-designs. Let $D_{v'}$ be the set of supports of codewords of weight v'. The number of subsets of $D_{v'}$ containing T is $1/(q-1)$ times the number of codewords of C^T of weight $v'-t$ which come from codewords of weight v' in C. The total number of weight $v'-t$ in C^T is independent of T. From the assumption that all weights less than v' yield t-designs, it follows that the number of vectors of weight $v'-t$ in C^T coming from codewords of weight less than v' is independent of T. Thus $D_{v'}$ yields a t-design.

The method of proof of this theorem is particularly interesting. The use of the MacWilliams identities and the observation that codewords of weight less than v_0 with the same support must be scalar multiples, form the basis of the proof.

The use of the theorem is best illustrated with examples. The most interesting application of it has been to self-dual codes with gaps in their weight enumerators. Let us first resolve the problem with the (12, 6) Golay code over F_3. This code only has codewords of nonzero weights 6, 9 and 12 and, choosing $t=5$, the number of nonzero weights less than or equal to $n-t=12-5$ is 1 which is itself less than or equal to $d-t=6-5=1$. In this case the codewords of weight 6 support a 5-(12, 6, 1) Steiner system and the codewords of weight 9 yield a trivial design containing all 9 subsets of 12 elements. Restricting this extended code by considering those codewords with a one in a particular position, and knowing that this code is an extension of an (11, 6) code, shows that the codewords of weight 5 in that code do in fact form a 4-design, a result not immediately available from Theorem 2.

It is also interesting to reconsider the binary (24, 12) Golay code. The only nonzero weights in this code are 8, 12, 16 and 24 and, letting $t=5$, the number of nonzero weights less than or equal to $24-5$, is itself less than or equal to $8-5$, which is 3. Thus, the supports of each weight form 5-designs and are respectively 5-(24, 8, 1), 5-(24, 12, 48) and 5-(24, 16, 78) designs **[14]**. It is easy to show that the code must contain the codeword which is all ones and so the complement of a weight 8 codeword is a weight 16 codeword and the corresponding designs are complementary. The design obtained from the weight 12 codewords is self-complementary.

The (47, 24) quadratic residue code over F_2 has minimum distance 11 and its extension has only weights 12, 16, 20, 24, 28, 32, 36 and 48. It follows from the MacWilliams identity that any binary self-dual linear code with distance 12, length 48, and the weight of any codeword divisible by 4 has a unique weight enumerator, which we don't give here. For $t=5$ the number (7) of nonzero weights less than or equal to $n-t=48-5$ is equal to $d-t=12-5$ and so the codewords of every weight support 5-designs. Their parameters are recorded in **[14]** and we simply note that since the code contains the all-ones codeword, the designs from words of weights 12, 16 and 20 are complementary to those of weights 36, 32 and 28 respectively, while the design with $k=24$ is self-complementary.

To conclude the section we outline an argument, originally due to Delsarte [20], which shows the existence of t-designs among the supports of codewords of C, using only the fact that C is an orthogonal array. Let C be a linear (n,k,d,q) code with s nonzero weights and let C' have minimum distance d' and s' nonzero weights. Let $u\in F_q^n$ have weight $t<d$ and denote by $\lambda_\tau(u)$ the number of codewords of C of weight τ which agree with u on the t nonzero positions of u. Since C is an orthogonal array of strength $d'-1$, we can enumerate in two ways on the number of appearances in codewords of C of vectors of weight $t+j$ in F_q^n, $0\leqslant j\leqslant d'-1-t$, which agree with u on its non-zero positions:

$$\sum_{\tau}\binom{\tau-t}{j}\lambda_\tau(u)=\binom{n-t}{j}(q-1)^j q^{k-t-j},\qquad j=0,1,\ldots,d'-1-t.$$

The summation of the left hand side is over all s nonzero weights of C. This is a set of $d'-t$ equations in the s unknowns $\lambda_\tau(u)$. If $d'>s$, we can choose $t=d'-s$ to obtain s equations in s unknowns. Since the transformation matrix is nonsingular (it can be obtained from a van der Monde matrix by a simple sequence of elementary row operations), a unique solution for $\lambda_\tau(u)$ exists for each τ, implying that the codewords of each nonzero weight of C support a $(d'-s)$-design since $\lambda_\tau(u)$ is then independent of the particular vector u of weight $(d'-s)$ chosen. Similarly, the codewords of each nonzero weight in C' support $(d-s')$-designs. MacWilliams and Sloane [21] show that for binary codes, the nonzero weights of *both* C and C' support t-designs where $t=\max\,(d-\bar{s}', d'-\bar{s})$, where $\bar{s}$ is $(s-1)$ if C contains the all ones codeword and s otherwise and $\bar{s}'$ is defined similarly. The argument is simple, elegant and powerful and well illustrates the interesting relationships between coding and combinatorics.

Further Reading

What we have presented here is the state of affairs that existed in about 1972. Since that time developments have taken place in two directions, and both have proven interesting. In the first direction workers felt that, since perfect codes had proven so useful in producing t-designs, perhaps the restriction of being perfect could be relaxed in some controlled manner without losing t-designs. This is the thought behind the definition of uniformly packed codes ([**16**], [**17**] and [**18**], although they have different definitions in two of these references) and nearly perfect codes [**19**]. This approach has in fact met with some success.

The other direction is the result of the revolutionary work of Delsarte [**20**], which is certainly one of the most important pieces of work in coding theory over the past few years. Before this work appeared there were several highly structured nonlinear codes known which supported t-designs, which did not fit into any of the known theories. Attempts were made to define the dual of a nonlinear code, but these did not seem to help. Delsarte worked with the distance distribution of a code (rather than the weight distribution) and defined a transformation on it. With this distribution and its transform he defined four parameters. If the code is linear, the distance distribution reduces to its weight distribution and the transform of it reduces to a scalar multiple of the weight distribution of the dual code. The four parameters in this case are the distance d of D, the number s of nonzero weights of C, the distance d' of C', and the number s' of nonzero weights in C'. The surprising thing is that many of the results obtained for linear codes have strong equivalents for nonlinear codes, using the four parameters of the nonlinear codes.

For the readers interested in pursuing the subject, the original paper of Delsarte [**20**] makes excellent reading. The book by MacWilliams and Sloane [**21**] also contains a detailed survey of much of this work. The two survey papers by Assmus and Mattson [**2**] and van Lint [**3**] are also of interest. Beyond these references the papers tend to be rather narrow and detailed, and further reading should be guided by the references in [**2**], [**3**], or [**21**].

References

[1] C. E. Shannon, A Mathematical Theory of Communication, Bell System Tech. J., 27 (1948) 379–423, 623–656.

[2] E. F. Assmus Jr. and H. F. Mattson Jr., Coding and Combinatorics, SIAM Rev., 16 (1974) 349–388.

[3] J. H. van Lint, Combinatorial Designs Constructed from or with Coding Theory, in Information Theory: New Trends and Open Problems, edited by G. Longo, CISM Courses and Lectures 219, Springer-Verlag, Wien (1975).

[4] R. W. Hamming, Error Detecting and Error Correcting Codes, Bell System Tech. J., 29 (1950) 147–150.

[5] A. Tietäväinen, On the Nonexistence of Perfect Codes over Finite Fields, SIAM J. Appl. Math., 24 (1973) 88–96.

[6] A. Tietäväinen and A. Perko, There are no Unknown Perfect Binary Codes, Ann. Univ. Turku., Ser. A, 148 (1971) 3–10.

[7] W. W. Peterson and E. J. Weldon Jr., Error-Correcting Codes, MIT Press, Cambridge (1972).

[8] M. J. E. Golay, Notes on Digital Coding, Proc. IRE., 37 (1949) 657.

[9] N. J. A. Sloane, Weight Enumerators of Codes, Mathematical Centre Tracts, 55 (1974) 111–138.

[10] F. J. MacWilliams, C. L. Mallows and N. J. A. Sloane, Generalizations of Gleason's Theorem on Weight Enumerators of Self-Dual Codes, IEEE Trans. Information Theory, 18 (1972) 794–805.

[11] W. O. Alltop, Extending t-Designs, J. Combinatorial Theory (A), 18 (1975) 177–186.

[12] N. S. Mendelsohn, Intersection Numbers of t-Designs, Studies in Pure Mathematics, Academic Press, New York (1971), 145–150.

[13] E. F. Assmus Jr. and H. F. Mattson Jr., On Tactical Configurations and Error-Correcting Codes, J. Combinatorial Theory, 2 (1967) 243–257.

[14] E. F. Assmus Jr. and H. F. Mattson Jr., New 5-Designs, J. Combinatorial Theory, 6 (1969) 122–151.

[15] J. -M. Goethals, A Polynomial Approach to Linear Codes, Philips Res. Repts., 24 (1969) 145–159.

[16] L. A. Bassalygo, G. V. Zaitsev and N. V. Zinovev, Uniformly Packed Codes, Problems of Information Transmission, 10 (1974) 6–10 (English Translation).

[17] J. -M Goethals and H. C. A. van Tilborg, Uniformly Packed Codes, Philips Res. Repts., 30 (1975) 9–36.

[18] N. V. Semakov, V. A. Zinovev and G. V. Zaitsev, Uniformly Packed Codes, Problems of Information Transmission, 7 (1971) 30–39. (English Translation).

[19] J. -M. Goethals and S. L. Snover, Nearly Perfect Binary Codes, Disc. Math., 3 (1972) 65–68.

[20] P. Delsarte, Four Fundamental Parameters of a Code and their Combinatorial Significance, Information and Control, 23 (1973) 407–438.

[21] F. J. MacWilliams and N. J. A. Sloane, The Theory of Error Correcting Codes, North-Holland Publishing Co., Amsterdam (1977).

NOTES

Elementary Counterexamples in Infinite Dimensional Inner Product Spaces

NORTHRUP FOWLER III
Hamilton College
Clinton, NY 13323

It is surprising to find that while most authors of elementary linear algebra texts are careful to distinguish between those results that hold for general vector spaces and those that are true only in the finite dimensional case, they rarely do the same for inner product spaces. For example, if S and W are subspaces of a finite dimensional inner product space V, it is easy to prove that

$$\begin{aligned}&\text{(i) } (S^{\perp})^{\perp}=S,\\ &\text{(ii) } S\oplus S^{\perp}=V,\\ &\text{(iii) } S^{\perp}+W^{\perp}=(S\cap W)^{\perp},\end{aligned}$$

yet none of these three statements are true in general for infinite dimensional inner product spaces. The standard infinite dimensional counterexample to (i) and (ii) is the subspace S of $C[0,1]$ consisting of polynomials with the usual integral inner product. Since establishing that $S^{\perp}$ is the trivial (zero) subspace in this example requires the Weierstrass approximation theorem, it is not capable of proof in elementary courses. The purpose of this note is to provide a more elementary infinite dimensional inner product space that contains simple counterexamples to the three basic properties listed above.

Let R be the set of real numbers and N be the set of non-negative integers. Then define V to be the set of all sequences of elements of R that have only a finite number of non-zero terms. In other words, V is the set of all functions f from N to R for which there is some $n\in N$ such that $f(m)=0$ for all $m>n$. Give V the usual (coordinatewise) addition and scalar multiplication. Choose as the canonical basis for V the set $\eta=\{e_n|n\in N\}$ where e_n is the sequence with a "1" in the $(n+1)$st coordinate and a "0" elsewhere. Define the function $\langle,\rangle$ from $V\times V$ into R by $\langle u,v\rangle=\sum_{k=0}^{n}u_kv_k$ where n is sufficiently large such that $u_m=v_m=0$ for all $m>n$. It is easy to show that $(V,+,\cdot,\langle,\rangle)$ is an infinite dimensional real inner product space and that η is an orthonormal basis for V. Indeed, V is a natural generalization of R^n; it is just the direct sum of denumerably many copies of R.

Let $\{a_i\}_{i=0}^{\infty}$ be the sequence $1,1,2,6,42,1806,\ldots$, where, for $i\geq 2$, $a_i=a_{i-1}(a_{i-1}+1)$. Then it is easy to show by induction that $a_{k+1}=1+\Sigma_{i=1}^{k}a_i^2$. Now define a sequence w_i of vectors in V recursively by

$$\begin{aligned}w_0&=e_0+e_1,\\ w_1&=e_0-e_1+e_2,\\ w_2&=e_0-e_1-2e_2+e_3,\\ w_3&=e_0-e_1-2e_2-6e_3+e_4,\\ &\vdots\\ w_n&=e_0-\left(\sum_{i=1}^{n}a_ie_i\right)+e_{n+1}.\end{aligned}$$

Then $\{w_i\}_{i=0}^{\infty} \cup \{e_0\}$ is a basis for V. Moreover, the vectors $\{w_i\}_{i=0}^{\infty}$ are orthogonal. To see this, assume $1 \leqslant k < n$ (the case $k=0$ is immediate) and write

$$w_k = e_0 - \left(\sum_{i=1}^{k} a_i e_i \right) + e_{k+1},$$

and

$$\begin{aligned} w_n &= e_0 - \left(\sum_{i=1}^{n} a_i e_i \right) + e_{n+1} \\ &= e_0 - \left(\sum_{i=1}^{k} a_i e_i \right) - a_{k+1} e_{k+1} - \left(\sum_{i=k+2}^{n} a_i e_i \right) + e_{n+1}. \end{aligned}$$

Then

$$\langle w_k, w_n \rangle = 1 + \sum_{i=1}^{k} a_i^2 - a_{k+1} = a_{k+1} - a_{k+1} = 0.$$

If C is a collection of vectors in V, we will denote the linear span of C—the set of all (finite) linear combinations of elements of C—by $L(C)$. Let $S = L(\{w_i\}_{i=0}^{\infty})$, the linear span of $\{w_i\}_{i=0}^{\infty}$ and suppose that $x \in S^{\perp}$. Then $x \in V$ implies that there is a q such that x is a linear combination of $e_0, e_1, \ldots, e_q$. Let p be the least such q; then $w_0, w_1, \ldots, w_{p-1}, w_p - e_{p+1}$ are $p+1$ mutually orthogonal elements in $L(\{e_0, \ldots, e_p\})$, and hence form a basis for $L(\{e_0, \ldots, e_p\})$. Since $\langle x, w_i \rangle = 0$ for all $i \geqslant 0$, and since x (when expressed as a linear combination of elements in η) has a zero coordinate with respect to e_{p+1}, it follows that $\langle x, w_p - e_{p+1} \rangle = 0$. Thus x is orthogonal to each element in the basis $w_0, w_1, \ldots, w_{p-1}, w_p - e_{p+1}$; hence x is the zero vector. Thus $S^{\perp} = \{0\}$, and consequently $(S^{\perp})^{\perp} \neq S$, and $S \oplus S^{\perp} \neq V$.

To find proper subspaces S_1 and S_2 of V such that $(S_1 \cap S_2)^{\perp} \neq S_1^{\perp} + S_2^{\perp}$, define the vectors u_i and v_i recursively as follows:

$$\begin{array}{ll} u_0 = e_0 + e_1 & v_0 = e_0 + e_1 \\ u_1 = e_0 - e_1 + e_2 & v_1 = e_0 - e_1 + e_3 \\ u_2 = e_0 - e_1 - 2e_2 + e_4 & v_2 = e_0 - e_1 - 2e_3 + e_5 \\ \vdots \qquad \vdots & \vdots \qquad \vdots \\ u_n = e_0 - e_1 - \left(\displaystyle\sum_{i=1}^{n-1} a_{i+1} e_{2i} \right) + e_{2n}, & v_n = e_0 - e_1 - \left(\displaystyle\sum_{i=1}^{n-1} a_{i+1} e_{2i+1} \right) + e_{2n+1}. \end{array}$$

Let $S_1 = L(\{u_i\})$ and $S_2 = L(\{v_i\})$. Then it is easy to verify that

(1) $\{u_i\}$ and $\{v_i\}$ are infinite linearly independent sets;
(2) $L(\{e_0, e_3, e_5, \ldots\})$ and $L(\{e_0, e_2, e_4, \ldots\})$ are complementary spaces for S_1 and S_2 respectively;
(3) $S_1 \cap S_2 = L(\{e_0 + e_1\})$;
(4) $S_1^{\perp} = L(\{e_3, e_5, e_7, \ldots\})$, $S_2^{\perp} = L(\{e_2, e_4, e_6, \ldots\})$;
(5) $(S_1 \cap S_2)^{\perp} = L(\{e_0 - e_1, e_2, e_3, e_4, \ldots\})$;
(6) $S_1^{\perp} + S_2^{\perp} = L(\{e_2, e_3, e_4, e_5, \ldots\})$.

Properties (5) and (6) together yield the desired counterexample to statement (iii) above.

We should point out that the examples here really are elementary since the vectors w_i, for example, are constructed by letting $w_0 = e_0 + e_1$ and defining the rest in order by making the simplest choices of coefficients to preserve orthogonality. This elementary nature is the result of our choice of V as the direct sum of denumerably many copies of R and the fact that at any stage of our construction all the necessary information is obtained directly from finite linear combinations of previously defined objects. A similar purely algebraic construction works for any formally real field.

Approval Voting: A 'Best Buy' Method for Multi-candidate Elections?

Samuel Merrill
Wilkes College
Wilkes Barre, PA 18703

Often a voter is confronted with an election in which more than two candidates are running for the same office. Yet he is normally required to cast a ballot for only one of them, and is thus deprived of registering any preferences among the others. Such an election may award a winning plurality to a candidate who is the first choice of a minority, while other candidates may enjoy approval by a larger proportion of the electorate and could, if elected, serve with a wider mandate.

Let us consider an alternative form of voting, called **approval voting** by Weber [**18**] and Brams [**3**], under which each voter votes for as many of the candidates as he chooses but casts no more than one vote for any candidate. The candidate receiving the most votes wins. This system would be simple to implement, yet avoids many of the shortcomings of the simple plurality method when there are more than two candidates, only one of whom is to be elected.

One of the easiest cases in which to see the effect approval voting might have had is the 1970 New York Senatorial race. There were three candidates: Ottinger (Democrat), Goodell (Republican–Liberal), and Buckley (Conservative). As it turned out, Buckley won with 39% of the vote, followed by Ottinger with 37%, and Goodell with 24%. It is reasonable to speculate that many if not most supporters of the two candidates perceived to be liberal (Ottinger and Goodell) would have voted for both under approval voting. This would almost certainly have led to a victory for Ottinger and possibly a third-place finish for Buckley.

This leads us to the question: How should the rational voter select the subset of candidates to vote for under approval voting? Assuming n voters and k candidates $c_1,\ldots,c_k$, let us make the following assumptions:

(A) Each voter v defines a real-valued function f on the set of candidates, such that the quantity $f(c_i)-f(c_j)$ is intended to represent the utility or value to voter v of having candidate c_i elected instead of candidate c_j.

(B) The total approval vote for each candidate is a random variable with the same probability distribution as that for any other candidate.

(C) n is large enough that the probability of an m-way tie ($m>2$) is negligible, relative to that for a 2-way tie.

(In order to define the function f postulated in assumption (A), one may assume that the voter has a preference ordering on the set of *pairs* (c_i,c_j), $i,j=1,\ldots,k$, satisfying the six axioms given in Luce and Raiffa [**12**, pp. 25–28]. It follows that there exists a linear utility function u over the set of risky alternatives arising from this set of pairs (c_i,c_j) [**12**, p. 30]. We write u_{ij} for $u(c_i,c_j)$. It is intended that u_{ij} represent the value to voter v of having candidate c_i elected, instead of candidate c_j. If one further assumes that $u_{ij}=u_{im}+u_{mj}$ for all $i,j,m=1,\ldots,k$, then the one-parameter family of solutions to the $k-1$ equations $f_i-f_{i+1}=u_{i,i+1}$, $i=1,\ldots,k-1$, satisfies $u_{ij}=f_i-f_j$ for all $i,j=1,\ldots,k$. It then follows easily that $u_{ii}=0$ and $u_{ji}=-u_{ij}$ for all $i,j=1,\ldots,k$. Setting $f(c_i)=f_i$ (for a fixed choice of the parameter) yields the desired function f. Hence, once a preference ordering on the set of pairs satisfying Luce and Raiffa's axioms is admitted, the crucial assumption in (A) is that embodied in the equation $u_{ij}=u_{im}+u_{mj}$.)

Letting S denote the set of candidates voted for by v, define the **total utility** for v by

$$V(S)=\sum[f(c_i)-f(c_j)] \tag{1}$$

where the summation is over all $i\in S$ and $j\notin S$. I am motivated here by the assumption that a voter can exercise power only if his votes are decisive in the sense that for some pair of candidates c_i and c_j, he can break a tie for first place if n is odd (or produce such a tie if n is even) which might occur among all other votes before his are counted. In this case the value to him of his vote is proportional to $f(c_i)-f(c_j)$ provided that he votes for c_i but not c_j. Total utility is the sum of these utilities over all pairs of candidates.

Suppose voter v has decided to vote for the candidates in set S and wishes to know if he could improve his total utility by also voting for another candidate c. He observes that $V(S)$ and $V(S\cup\{c\})$ have the same summands in (1) except for those involving c. Thus

$$\begin{aligned}V(S\cup\{c\})-V(S)&=\sum_{c_j\notin S}[f(c)-f(c_j)]-\sum_{c_j\in S}[f(c_j)-f(c)]\\&=kf(c)-\sum_{j=1}^{k}f(c_j).\end{aligned}$$

Hence he will improve total utility by voting for c precisely if

$$f(c)>\frac{1}{k}\sum_{j=1}^{k}f(c_j). \tag{2}$$

It follows that voter v achieves maximal total utility by voting for the set of those candidates c for which (2) holds. This result was also obtained independently by Weber [**18**].

In words this says that a rational voter under assumptions (A), (B), and (C) for approval voting should vote for all candidates whom he rates above the average of those running. For example, if a citizen, voting in a four-way contest, rates candidates 10, 8, 7, and 0, he should vote for the top three since all three rate above the average (6.25).

Now let us relax assumption (B) by permitting different probability distributions for the number of votes received by the various candidates. In this case, multiply each summand in (1) by p_{ij}, the probability that voter v be decisive (with respect to first place) for the pair of candidates c_i and c_j (let $p_{ii}=0$). It follows that the voter will improve his total utility by voting for c_i precisely if

$$f(c_i)>\sum_{j=1}^{k}q_{ij}f(c_j) \tag{2'}$$

where $q_{ij}=p_{ij}\Big/\sum_{m=1}^{k}p_{im}$.

Don't bother to memorize this formula for use in the voting booth. Remember that, generally speaking, the larger values of q_{ij} will correspond to the stronger candidates, at least if more than one has a good chance to win. Hence the voter's rule of thumb in this setting would be to vote for candidates whom he rates above the average of his ratings where that average is weighted according to the strength of the candidates.

It follows from (2′) that under approval voting, it is never to the voter's advantage to withhold a vote for his first choice while voting for a less preferred candidate. This need not be true for a simple plurality ballot. For example, in the 1970 New York Senate race, a voter who preferred Goodell slightly to Ottinger, but strongly opposed Buckley, might find it in his interest to vote instead for Ottinger because the latter was thought to have a better chance to win. However, if there are four or more candidates, for any pair of candidates rated strictly between the voter's first and last choices, there always exist circumstances in which a rational approval voter should vote for the less preferred candidate while withholding a vote from the more preferred. For this result and a proof that approval voting is still "more sincere" than any other single-ballot non-ranked voting system, see Brams and Fishburn [**6**, Th. 3]. The development there and in [**3**] and [**5**] is based on ordinal utility (in contrast to the cardinal utility model presented here).

At first sight, it might appear that the advantages of approval voting could be augmented by permitting a voter a fixed number of votes which he can apportion among the candidates as he pleases. But in a single winner race, this is an illusion. By adapting the model to this form of balloting, I will show that a voter maximizes his total utility by casting all of his votes for one of the candidates.

Let $b(c_i)$ be the number of votes for candidates c_i cast by voter v, $i=1,\dots,k$. Modifying the definition used before, let us define total utility by

$$V(b)=\sum_{i<j}\left[b(c_i)-b(c_j)\right]\left[f(c_i)-f(c_j)\right]p_{ij}. \tag{3}$$

Note that if i and j are unrestricted, the sum in (3) doubles. What happens if voter v decides to shift a single vote from candidate c_j to candidate c_i? Denote by b' this altered apportionment function. Since $V(b)$ and $V(b')$ have the same summands except for those involving c_i and c_j,

$$2[V(b')-V(b)]=\sum_{q=1}^{k}\left[f(c_i)-f(c_q)\right]p_{iq}-\sum_{q=1}^{k}\left[f(c_j)-f(c_q)\right]p_{jq}.$$

Hence voter v will increase his total utility precisely if the first sum on the right exceeds the second sum. It follows by induction that voter v attains maximal utility by putting all of his eggs in one basket, i.e., giving all of his votes to the candidate c_i for which

$$\sum_{q=1}^{k}\left[f(c_i)-f(c_q)\right]p_{iq} \tag{4}$$

is largest. Thus for rational voters, this system is equivalent to its special case, the simple plurality ballot! In general, the candidate determined by criterion (4) need not be the voter's first choice, although it is if (B) holds. The reader is invited to apply this criterion to an example, such as the New York Senate race.

This problem can also be interpreted as a linear program in the variables $b(c_i)$. The rational voter desires to maximize total utility under the constraint that $\Sigma\, b(c_i)$ not exceed the fixed number of votes permitted each voter. The feasible region is a k-dimensional tetrahedron and the maximum must occur at an extreme point, all of whose coordinates but one are zero.

Approval voting is, of course, not the only alternative which takes account of voter preferences among the candidates. The **Borda count** requires each voter to rank the candidates 1st, 2nd, 3rd, etc. The candidate with the smallest rank sum wins. Ranked preferences can also be used to seek a **Condorcet winner** (if one exists), i.e., a candidate who would win a majority in a two-way contest against any of the other candidates. These and other methods are discussed in [**1**, **7**, **13**, **16**, and **17**]. For a particularly readable account, see chapter 10 of [**13**].

The **Copeland method** [**7**; **17**, pp. 26–27] awards victory to the candidate who can win the most pairwise contests. This insures the election of the Condorcet winner if one exists. However, in the case of either 3 or 4 candidates, if no Condorcet winner exists (and no two candidates receive the same number of votes) the Copeland method fails to determine a winner (this takes a little thought). Thus, there must be at least 5 candidates for Copeland's contribution to be of value.

Yet another attempt to utilize preferences is the run-off, and its extension, the elimination contest. For example, the election of the House Majority Leader in December, 1976, (see [**15**]) was conducted by a succession of ballots on each of which the candidate with the fewest votes was eliminated until only one candidate—the winner—remained. The first (and second) ballots were easily won by Philip Burton, an outspoken liberal, who would thus have been elected under a simple plurality system. On the second ballot, Burton received 107 votes, Jim Wright of Texas

95, and Richard Bolling of Missouri 93. Bolling was eliminated. To the surprise of nearly everyone, probably including Wright, on the third and final ballot, Wright beat Burton by a single vote, 148 to 147.

Had this election been conducted under approval voting, it seems reasonable to speculate that perhaps half of the Burton and Wright supporters might have rated middle-of-the-roader Bolling above average, and voted for him along with their first preferences. On the other hand, Wright and Burton could have added precious few approval votes from each other's camps. It is perhaps likely that Bolling would have won, receiving the approval of a healthy majority of the Democratic Caucus.

Good and Tideman **[8]** suggest a method which uses the voters' preference relations to position the candidates in a multidimensional attribute space, permitting estimates of the aggregate cardinal utilities of the electorate. However, even for three candidates, determination of the winner requires a computer.

Each of these alternatives, other than approval voting, suffers from complexity, in that the voter must determine a complete or nearly complete preference profile. Except for the last, these methods take no account of cardinal utility, whereas approval voting does, to the extent that in the present model the criterion of whom to vote for depends on cardinal utility. For example the rational approval voter votes differently if he rates three candidates 10, 9, and 0 rather than 10, 1, and 0.

Let us look briefly at approval voting when there is more than one vacancy to be filled in a given office (such as a school board). Assume, as before, that a voter can vote for as many candidates as he wishes. If there are p vacancies, the important contest is that between pth and $(p+1)$st place in the results of the balloting. If (B) holds, the definition of total utility given in (1) is unchanged, so that, just as before, the rational voter should vote precisely for those candidates whom he rates above average for those running. Thus, in this case, the voter's decision is unaffected by the number of vacancies to be filled. This form of approval voting is already used for school boards and some other races except that the voter cannot vote for more candidates than there are vacancies. How the rational voter should vote under this constraint or if (B) does not hold is left as an exercise for the reader.

Certain variants of this last method can be used to serve a purpose in a sense opposite to that of approval voting. In multi-member districts it is often desirable to use a voting system which assures minority parties some representation. For example, **cumulative voting** **[2]** (used to elect members of the Illinois House of Representatives) achieves this purpose by permitting the voter to apportion among the candidates a number of votes equal to the number of vacancies (usually 3 in each district). Elementary algebra shows that a disciplined minority whose proportion of the electorate exceeds $1/(1+M)$ where M is the number of vacancies in the district can assure itself of the election of one member in that district. (Cumulative voting serves no purpose in a single-vacancy election, as we have seen.) **Limited voting** **[2]** (used for Pennsylvania County Commissioners) achieves a similar purpose by permitting each voter fewer votes than the number of vacancies.

Approval voting might be most useful in primary elections, particularly Presidential primaries, where there are often many candidates from which it is desired to choose a single nominee who has broad support within his party. Such a system could be easily introduced in Presidential primaries in one or two states on an experimental basis.

For further discussion of arguments in favor of approval voting, see Brams **[5]**, Brams and Fishburn **[6]**, Kellett and Mott **[11]**, and Weber **[18]**. Kellett and Mott also give the results of a poll of Presidential candidates using approval voting. Joslyn **[10]** and Merrill **[14]** use polling data collected by the Survey Research Center in the fall of 1972 to analyze the effect different voting systems might have had on the outcome of the 1972 Democratic Presidential primaries. These analyses suggest that almost any of the methods described above, with the notable exception of simple plurality voting, would probably have led to greater success in the primaries for Humphrey than for McGovern.

The analysis of optimal voting strategies using linear programming is pursued in [14] for a variety of voting methods. Weber [18] develops a measure of effectiveness of voting systems assuming a random society. According to this measure, approval voting is found slightly more effective than the Borda system for the case of three candidates and both methods are found substantially more effective than the simple plurality ballot for any number of candidates greater than two.

The basic outline of the model developed here was presented by the author at the second session of the NSF Chautauqua-Type Short Course on Mathematical Modeling and Voting directed by Professor William Lucas at Syracuse University, March, 1977. The results embodied in equations (2′) and (4) were obtained independently by Hoffman [9] and the author, as a follow-up to this presentation.

References

[1] D. Black, The Theory of Committees and Elections, Cambridge University Press, Cambridge, 1958.

[2] G. S. Blair, Cumulative Voting: An Effective Electoral Device for Fair and Minority Representation, Ann. N. Y. Acad. Sci., 219 (1973) 20–26.

[3] S. Brams, One Man, N Votes, MAA Modules in Appl. Math., Cornell University, Ithaca, N. Y., 1976.

[4] ———, Paradoxes of Politics, the Free Press, New York, 1976.

[5] ———, Comparison of Voting Systems, Innovative Instructional Unit, Amer. Pol. Sci. Assoc., Washington, D. C., 1978.

[6] S. Brams and P. Fishburn, Approval Voting (1977), Amer. Pol. Sci. Rev., forthcoming.

[7] P. C. Fishburn, A Comparative Analysis of Group Decision Methods, Behavioral Science, 16 (1971) 538–544.

[8] I. J. Good and T. N. Tideman, From Individual to Collective Ordering through Multidimensional Attribute Space, Proc. Roy. Soc. London Ser. A, Vol. 347 (1976) 371–385.

[9] D. Hoffman, A Model for Sophisticated Voting (1977), mimeographed.

[10] R. A. Joslyn, The Impact of Decision Rules in Multi-Candidate Campaigns: The Case of the 1972 Democratic Presidential Nomination, Public Choice, 25 (1976) 1–17.

[11] J. Kellett and K. Mott, Presidential Primaries: Measuring Popular Choice, Polity, Vol. IX, No. 4 (1977) 528–537.

[12] R. D. Luce and H. Raiffa, Games and Decisions, Wiley, New York, 1957.

[13] J. Malkevitch and W. Meyer, Graphs, Models, and Finite Mathematics, Prentice-Hall, Englewood Cliffs, New Jersey, 1974.

[14] S. Merrill, A Unified Framework for Multi-candidate Voting Systems via Linear Programming, in preparation.

[15] The New York Times, Dec. 7, 1976.

[16] D. Rae, The Political Consequences of Electoral Laws, Yale University Press, New Haven, 1971.

[17] W. H. Riker and R. G. Niemi, The Choice of Voting Systems, Scientific American, 234 (June 1976) 21–27.

[18] R. J. Weber, Comparison of Voting Systems (1977), Econometrica, forthcoming.

Sequences of Polygons

R. J. Clarke
University of Adelaide
Adelaide, South Australia 5001

Let P be a convex polygon and let TP be the polygon whose vertices are the midpoints of the sides of P. We are interested in the sequence $P, TP, T^2P, \ldots$. If P is a triangle, TP is similar to P. If P is a quadrilateral, TP is a parallelogram; thereafter T^mP is similar to TP if m is odd, to T^2P if m is even. For a pentagon, the situation is less simple. However, experimentation soon convinces us that T^mP approaches a limiting shape, and that alternate members of this sequence "point" in opposite directions. We shall prove these facts for a general convex polygon P.

It is convenient to consider each vertex v_i of P as a complex number and to take

$$P=\begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$$

as a vector in $\mathbf{C}^n$. We may then think of the operator T as a matrix; in fact

$$T=\frac{1}{2}\begin{bmatrix} 1 & 1 & 0 & \ldots & & 0 \\ 0 & 1 & 1 & 0.. & & 0 \\ 0 & 0 & 1 & & & \\ & & & \ddots & & \\ & & & & \cdot & \\ 1 & 0 & \cdot & \cdot & \cdot\,0 & 1 \end{bmatrix}.$$

Define matrices C and A by

$$C=\begin{bmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ & & & \ddots & 1 \\ 1 & 0 & & \cdots & 0 \end{bmatrix}$$

and $A=C^{-1}T^2$, so that

$$A=\frac{1}{4}\begin{bmatrix} 2 & 1 & 0 & \cdot\cdot & 0 & 1 \\ 1 & 2 & 1 & \cdot\cdot & 0 & 0 \\ 0 & 1 & 2 & \cdot\cdot & 0 & 0 \\ \vdots & & & & & \\ 1 & 0 & 0 & \cdot\cdot & 1 & 2 \end{bmatrix}.$$

On considering the cases $n=4$ and $n=5$ we see that the sequence $\{A^mP\}$ is a judicious one to look at: the T^2 in A enables us to look at every other polygon in the sequence $\{T^mP\}$, and the C^{-1} labels the vertices of T^2P conveniently (see FIGURE 1).

Let $\zeta=\exp(2\pi i/n)$. Then A has eigenvalues

$$\lambda_r=\tfrac{1}{4}(2+\zeta^r+\zeta^{-r}),\ \text{for } r=0,1,2,\ldots n-1$$

and eigenvectors

$$P_r=\begin{bmatrix} 1 \\ \zeta^r \\ \zeta^{2r} \\ \vdots \\ \zeta^{(n-1)r} \end{bmatrix},$$

since the ith entry of AP_r is

$$\begin{aligned} \tfrac{1}{4}(\zeta^{(i-2)r}+2\zeta^{(i-1)r}+\zeta^{ir}) &= \lambda_r\zeta^{(i-1)r} \\ &= \lambda_r\times i\text{th entry of } P_r. \end{aligned}$$

Since the eigenvectors of A are linearly independent, they form a basis for C^n, so we may write $P=\sum_{r=0}^{n-1}\alpha_rP_r$ where $\alpha_r\in C$. Then $A^mP=\sum_{r=0}^{n-1}\lambda_r^m\alpha_rP_r$. If $0<r<n$, then $|\lambda_r|<\lambda_0=1$. So $A^mP\to\alpha_0P_0$. An easy calculation shows that α_0P_0 is the centroid of P. Thus the sequence $\{A^mP\}$ and hence the sequence $\{T^mP\}$ tends to the centroid of P. It seems difficult to prove this point geometrically.

However, we are really interested in the limit of the "shape" of A^mP. To this end we define the **similarity class** $[P]$ of any polygon P to be the set of all polygons similar to P. If $P_1,P_2,\ldots$ is a sequence of polygons, the statement "$[P_n]\to[P]$" will mean that for some positive numbers

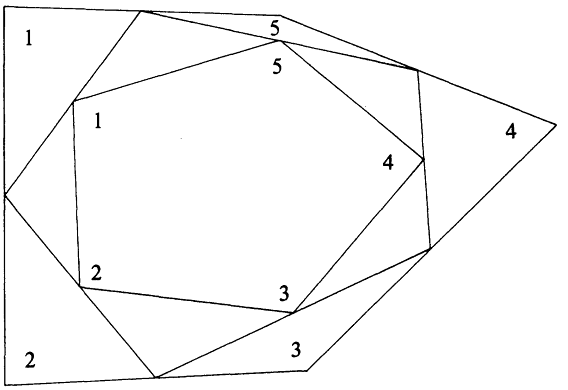


FIGURE 1.

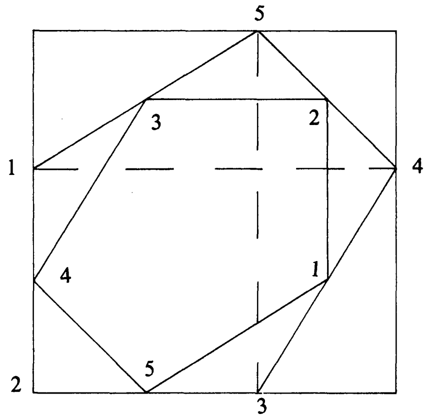


FIGURE 2.

$k_1, k_2, \ldots,$ we have $\lim\limits_{n\to\infty} k_n P_n = P$. We may assume that $\alpha_0 = 0$. Now among the λ_i we have the relations $\lambda_1 = \lambda_{n-1}$, and $|\lambda_r| < |\lambda_1|$ whenever $1 < r < n-1$. So if α_1 or α_{n-1} is non-zero, $A^m P/\lambda_1^m \to \alpha_1 P_1 + \alpha_{n-1} P_{n-1}$. Thus $[A^m P] \to [\alpha_1 P_1 + \alpha_{n-1} P_{n-1}]$.

We now show that for any convex polygon P we have α_1 or α_{n-1} non-zero. Let r be the smallest integer such that either α_r or α_{n-r} is non-zero. Then $A^m P/\lambda_r^m \to \alpha_r P_r + \alpha_{n-r} P_{n-r}$. Hence the polygon on the righthand side is convex. Let e_k be the kth edge of $\alpha_r P_r + \alpha_{n-r} P_{n-r}$. Then $e_k = \alpha_r \Delta_k + \alpha_{n-r} \Delta_k'$, where $\Delta_k = \zeta^{r(k+1)} - \zeta^{rk}$ is the kth edge of P_r and $\Delta_k' = \zeta^{-r(k+1)} - \zeta^{-rk}$ is the kth edge of P_{n-r}. Then $\Delta_{k+1} = (\zeta^r + \zeta^{-r})\Delta_k - \Delta_{k-1}$ for $1 \leqslant k \leqslant n-1$. Define a sequence of polynomials $F_m(x)$ by $F_0(x) = 0$, $F_1(x) = 1$, $F_2(x) = x$, and $F_{m+1}(x) = xF_m(x) - F_{m-1}(x)$ for $m \geqslant 1$. Then by induction we may show that

$$\Delta_k = F_k(\zeta^r + \zeta^{-r})\Delta_1 - F_{k-1}(\zeta^r + \zeta^{-r})\Delta_0, \qquad 1 \leqslant k \leqslant n,$$

and the same relationship holds among the Δ_k' and among the e_k.

Consider the linear mapping of the plane that takes e_0 to Δ_0 and e_1 to Δ_1. By the linear relationships above, the mapping takes e_k to Δ_k, i.e., it takes the polygon $\alpha_r P_r + \alpha_{n-r} P_{n-r}$ to P_r. Now the property of being convex is preserved under a linear transformation of the plane. But P_r is not convex unless $r = 1$, so the result follows: either α_1 or α_{n-1} is non-zero.

We may call a polygon P a **limit polygon** if AP is similar to P. We now know that the limit polygons are those of form

$$P = \alpha_1 P_1 + \alpha_{n-1} P_{n-1}. \tag{1}$$

These polygons can be characterized geometrically: they are just those polygons P, with vertices $v_1, \ldots, v_n$, for which there is a constant c with

$$v_{i+2} - v_{i+1} = c(v_{i+1} - v_i), \qquad 1 \leqslant i \leqslant n. \tag{2}$$

We can easily verify that a polygon of form (1) has property (2) with $c = 4\lambda_1 - 1 = 1 + 2\cos(2\pi/n)$, as $\lambda_1 = \frac{1}{4}(\exp(2\pi i/n) + \exp(-2\pi i/n) + 2$. On the other hand one easily sees that a polygon with property (2) is a limit polygon.

The case $n = 5$ is interesting. Here $c = 1 + 2\cos(2\pi/5) = \frac{1}{2}(1 + \sqrt{5}\,)$, the ubiquitous "golden ratio"! FIGURE 2 shows a "golden pentagon" and its associated golden rectangles inscribed in a square.

We now show that if n is odd, then the sequence of similarity classes $[T^m P]$ has the same limit as the sequence $[A^m P]$. Define an operation S by $SP = C^{(n-1)/2} TP$. Here we are just labelling the vertices of TP so that the first vertex lies between the $\frac{1}{2}(n+1)$th and $\frac{1}{2}(n+3)$rd vertices of P. Clearly $S^2 = A$. Now we have

$$SP_r = \tfrac{1}{2}\left(\zeta^{\frac{1}{2}(n-1)r} + \zeta^{\frac{1}{2}(n+1)r}\right)P_r, \qquad \text{for } r = 1, 2, 3, \ldots, n-1.$$

So if $P=\sum_{r=1}^{n-1}\alpha_r P_r$,

$$SP=\frac{1}{2}\sum_{r=1}^{n-1}\alpha_r\left(\zeta^{\frac{1}{2}(n-1)r}+\zeta^{\frac{1}{2}(n+1)r}\right)P_r.$$

So, with $\lambda_1=\frac{1}{2}(2+\zeta+\zeta^{-1})$,

$$\left(\frac{1}{\lambda_1}A\right)^m SP\rightarrow\frac{1}{2}\left(\zeta^{\frac{1}{2}(n-1)}+\zeta^{\frac{1}{2}(n+1)}\right)(\alpha_1P_1+\alpha_{n-1}P_{n-1})$$
$$=-\left(\cos\frac{\pi}{n}\right)(\alpha_1P_1+\alpha_{n-1}P_{n-1}).$$

Hence the sequence $[S^mP]$ has the same limit as $[A^mP]$.

The problem of midpoint polygons allows a natural generalization as follows. Let $f=a_0+a_1x+\cdots+a_dx^d$ be a polynomial with non-negative real coefficients whose sum is 1. Let n be an integer greater than d and let $T=a_0I+a_1C+\cdots+a_dC^d$, where C is the matrix defined above. Then we ask if there are integers k and l such that for all n, and for all convex n-gons P, the sequence $[A(k,l)^mP]$ converges to a similarity class of convex polygons, where $A(k,l)=C^{-k}T^l$. If the answer is yes, we call $f(x)$ **convergent**. We have found above, in the midpoint problem, that the polynomial $\frac{1}{2}+\frac{1}{2}x$ is convergent.

Define a function $g(x)$ by $g(x)=x^{-k}f(x)^l$. Then A has eigenvalues $g(\zeta^r)$, for $r=0,\ldots,n-1$, where $\zeta=\exp(2\pi i/n)$, and eigenvectors P_r as above. If we write $P=\sum_{r=0}^{n-1}\alpha_rP_r$ as usual we have $AP=\sum_{r=0}^{n-1}\alpha_rg(\zeta^r)P_r$.

By our assumption on the polynomial $f(x)$ we have $g(1)=1$ and $|g(\zeta^r)|<1$ if $r=1,2,3,\ldots,n-1$. So $A^mP\rightarrow\alpha_0P_0$ as before. Assume $\alpha_0=0$. Then necessary and sufficient conditions for the sequence $[A^mP]$ to have the required limit are that

$$g(\zeta)=g(\zeta^{-1}) \tag{3}$$

and

$$|g(\zeta^r)|<|g(\zeta)| \text{ if } r=2,3,\ldots,n-2. \tag{4}$$

We can rewrite these as

$$\zeta^{-k}f^l(\zeta)=\zeta^kf^l(\zeta^{-1}) \tag{3$'$}$$

and

$$|f(\zeta^r)|<|f(\zeta)| \text{ if } r=2,3,\ldots,n-2. \tag{4$'$}$$

Now (3)′ implies that $x^{-k}f^l(x)-x^kf^l(x^{-1})$ is divisible by $x-\zeta$ for every primitive nth root of unity ζ, for all $n>d$. Hence $x^{-k}f^l(x)=x^kf^l(x^{-1})$. As $f^l(x)$ has degree dl, we have that dl is even and $k=dl/2$. So $x^{-d/2}f(x)=x^{d/2}f(x^{-1})$. Thus $a_r=a_{d-r}$ for $r=0,1,2,\ldots,d$. Such a polynomial is called a **reciprocal polynomial**. If the degree d is even, we may satisfy (4), by taking $l=1$ and $k=d/2$, while if d is odd, we must take $l=2$ and $k=d$.

THEOREM. *A polynomial $f(x)$ of degree d is a convergent polynomial if and only if $f(x)$ is reciprocal and $|f(\zeta^r)|<|f(\zeta)|$ for $r=2,3,\ldots,n-2$, where $n>d$ and $\zeta=\exp(2\pi i/n)$.*

It follows, for instance, that the sequence of polygons obtained by trisecting the sides of a given polygon does not usually converge in the above sense.

We conclude by mentioning a related but apparently more difficult problem. Let P be a convex pentagon and let P' be the pentagon formed by the intersections of the diagonals of P. What can we say about the sequence $P,P',P'',\ldots$? This problem is non-linear and therefore harder than the previous problems, although it may well have a similar solution. We have been able to make little progress on it.

How Analytic Functions Preserve Closeness

RICHARD K. WILLIAMS
Southern Methodist University
Dallas, TX 75275

Topological dynamics is an area of topology in which one studies the geometric properties of certain groups of transformations which had their origins in classical dynamics and systems of differential equations. Henri Poincaré is generally credited with originating the subject. He was probably the first to solve dynamics problems as problems in topology. Poincaré's work was later abstracted, systematized, and carried on by G. D. Birkhoff.

An area of current research in Topological Dynamics is the study of the somewhat elusive "expanisve property". Roughly speaking, the expansive property deals with the ability of functions to "spread out" points by means of repeated application of the functions, i.e., iterations. This concept will be precisely defined later.

In this paper, various types of expansiveness will be defined, and it will then be shown that in most cases, analytic functions are not expansive. The results are both new and easy to prove, and the paper gives the reader with a limited background in topology an introduction to an area of current research. In addition the proofs are good exercises in complex analysis.

If f is a homeomorphism of a metric space (X,d) onto itself, then f is said to be **expansive** if there exists $\delta>0$ such that $x,y \in (X,d)$, $x\neq y$ implies $d(f^n(x),f^n(y))>\delta$ for some integer n. If f is only assumed to map (X,d) continuously into itself, the expansive property can be adjusted by calling f **positively expansive** if the integer n is non-negative. The concept of expansiveness can also be weakened by requiring that for each $x \in (X,d)$, there is a $\delta(x)>0$ such that if $y \in (X,d)$, $y\neq x$, then $d(f^n(x), f^n(y))>\delta(x)$ for some integer n. In this case, f is called **pointwise expansive**. A similar modification allows one to weaken positively expansive to **pointwise positively expansive**. Clearly, "expansive" and "positively expansive" are stronger than "pointwise expansive" and "pointwise positively expansive", respectively, and for homeomorphisms onto, i.e. bijections, "positively expansive" is stronger than "expansive".

We now illustrate these concepts with several examples. Let us take (X,d) to be the real line with the usual distance function in the following two examples. First, if $f(x)=2x$, f is a homeomorphism of (X,d) onto itself, and since $f^n(x)=2^nx$, and $d(f^n(x),f^n(y))=2^n|x-y|$, we see that f is both expansive and positively expansive. (Choose δ to be *any* positive real number.) If $f(x)=2x$ for $x\geqslant 0$ and $x/2$ for $x<0$, f is easily seen to be expansive, but not positively expansive.

If we take (X,d) to be $\{1/n\colon n=1,2,3,\dots\}$ with the usual distance function, and if $f(x)=x$, then f is a homeomorphism of (X,d) onto itself which is pointwise expansive and pointwise positively expansive, but not expansive.The reader may refer to **[1]** through **[10]** for results and further examples on these concepts.

We now give our final generalizations. Let F be a family of functions, each mapping (X,d) into itself. We call F an **expansive family** if there exists $\delta>0$ such that $x,y \in (X,d)$, $x\neq y$ implies $d(f(x),f(y))>\delta$ for some $f \in F$. A similar generalization allows one to define **pointwise expansive families**. Clearly "expansive family" implies "pointwise expansive family", and these two concepts are implied by their respective predecessors, since if f is either expansive or pointwise expansive, we can take $F=\{f^n\colon n=0,\pm 1,\dots\}$, and if f is either positively expansive or pointwise positively expansive, we can take $F=\{f^n\colon n=0,1,\dots\}$.

These definitions are easily understood by anyone who knows what a metric space is, but some very simple questions about these properties remain unanswered. For instance, it has been known for over 18 years (see **[7]** and **[8]**) that there does not exist an expansive homeomorphism on the closed unit disk, but the existence or non-existence of an expansive homeomorphism on the closed unit ball in three dimensions is still an open question.

In this paper we will specialize all functions to be analytic functions of a complex variable, and by proving several non-existence theorems in this special case, perhaps we will gain some insight relative to the more general cases.

The following theorem is the key result of this paper, all of the other results being either direct or indirect consequences of it.

THEOREM. *If F is an equicontinuous family of functions, then F is not pointwise expansive at any point.*

Proof. Fix $x_0 \in (X,d)$, and let $\delta > 0$. Choose $\eta > 0$ such that $d(x,x_0) < \eta$ and $x \in (X,d)$ imply $d(f(x),f(x_0)) < \delta$ for each $f \in F$. This is clearly contrary to expansiveness at x_0.

COROLLARY 1. *There does not exist a pointwise expansive family of analytic functions on a bounded open set.*

Proof. Let F be a family of analytic functions, each mapping a bounded open set G to itself. Since the members of F are uniformly bounded, F is a normal family. (See **[1]**, p. 216.) Thus, by the Arzela-Ascoli Theorem (**[1]**, p. 214), F is equicontinuous on each compact subset of G. The result follows from the Theorem.

If we use the word **region** to denote an open set, together with some or all of its boundary points, we have the following additional corollary to the Theorem.

COROLLARY 2. *There does not exist a pointwise expansive family of analytic functions on a bounded region.*

Proof. By the Open Mapping Theorem, if a non-constant analytic function maps a region R into itself, it maps the interior of R into itself. Thus, if there were a pointwise expansive family of analytic functions on R, we could delete any constant functions from the family, and the resulting family would be expansive on the interior of R, contradicting Corollary 1.

The reader may find it an interesting exercise to prove Corollary 1 directly by using the Cauchy Integral Formula.

Since "pointwise expansive family" is the weakest of all the expansive properties mentioned above, it is clear that if R is a bounded region, there exists neither an expansive family of analytic functions on R, nor a positively expansive analytic function on R, nor a pointwise positively expansive analytic function on R, nor an expansive analytic homeomorphism on R. However, there *do* exist expansive homeomorphisms on the open unit disk. In fact, the homeomorphism may be taken to be positively expansive (see **[9]**, p. 660). While it may be possible to simplify the slightly complicated example in **[9]**, we see that the homeomorphism cannot be analytic.

It is natural to investigate the possibility of generalizing Corollaries 1 and 2; e.g., can the word "bounded" be omitted? The following example shows that this generalization is impossible. If $f(z) = 2z$, where $\operatorname{Im} z > 0$, and if $F = \{f^n\colon n = 1,2,\dots\}$, it is easy to see that f is a positively expansive homeomorphism on the upper half plane, and that F is thus an expansive family of analytic functions on the upper half plane. This example also shows that the "bounded open set" in Corollary 1 cannot even be replaced by "conformally equivalent to a bounded open set", since the upper half plane is conformally equivalent to the open unit disk.

Neither can we generalize Corollary 2 by replacing "bounded region" by "bounded set", for if $f(z) = z$ on $\{0,1\}$, and if $F = \{f^n\colon n = 1,2,\dots\}$, then F *is* a pointwise expansive family of analytic functions on a bounded (even compact) set. In fact, f is a positively expansive homeomorphism.

In dynamics, one studies motions of particles which are governed by a system of differential equations, subject to certain conditions. For example, if a particle is at point p at time $t = 0$, and if $x(p,t)$ denotes the position of the particle at time t, then $x(p,t)$ will satisfy certain ordinary differential equations (the differentiation being with respect to t), and $x(p,0) = p$. Generally, the function $x(p,t)$ is called a **motion** and $\{x(p,t)\colon t \geqslant 0\}$ is called a **trajectory**. There are usually other conditions imposed on $x(p,t)$, but this need not concern us here. If two particles which are close at time $t=0$ remain close forever, then the system enjoys a certain kind of **stability.**

Suppose the particles all belong to a bounded region in the plane, and let $x(p,t)$ be analytic in p for each fixed t. If $F=\{x(p,t)\colon\ t\geqslant 0\}$, then F is a family of analytic functions, so by Corollary 2, F is not pointwise expansive. Thus, for each $\delta>0$, there exists $\eta>0$ such that $d(p_1,p_2)=d(x(p_1,0),x(p_2,0))<\eta$ implies $d(x(p_1,t),x(p_2,t))<\delta$ for all t. Therefore, in a certain sense, analytic motions are stable.

References

[1] L.V. Ahlfors, Complex Analysis, McGraw-Hill, New York, 1966.

[2] B.F. Bryant, Expansive self-homeomorphisms of a compact metric space, Amer. Math. Monthly, 69(1962) 386–391.

[3] ______, On expansive homeomorphisms, Pacific J. Math., 10(1960) 1163–1167.

[4] W.L. Reddy, The existence of expansive homeomorphisms on manifolds, Duke Math. J., 32(1965) 627–632.

[5] ______, Pointwise expansion homeomorphisms, J. London Math. Soc., 2(1970) 232–236.

[6] ______, On positively expansive maps, Math. Systems Theory, 6(1972) 76–81.

[7] J.F. Jacobsen and W.R. Utz, The non-existence of an expansive homeomorphism on a closed 2-cell, Pacific J. Math., 10(1960) 1319–1321.

[8] R.K. Williams, Some theorems on expansive homeomorphisms, Amer. Math. Monthly, 73(1966) 854–856.

[9] ______, Some results on expansive mappings, Proc. Amer. Math. Soc., 26(1970) 655–663.

[10] ______, Linearization of expansive homeomorphisms, General Topology and its Applications, 6(1976) 315–318.

Separation of Points in the Plane

Steven H. Lameier
Thomas More College
Fort Mitchell, KY 41017

Edward P. Merkes
University of Cincinnati
Cincinnati, OH 45221

Let $K=\{k_1,k_2,\ldots,k_n\}$ be n distinct points in the Euclidean plane. A partition $[A,B]$ of K is two nonempty subsets A and B of K, such that $A\cap B=\varnothing$ and $A\cup B=K$. There are $2^{n-1}-1$ partitions of K.

A partition $[A,B]$ of K is said to **separate** K if there exist two perpendicular lines such that A and B lie in the interior of opposite quarter planes determined by these lines. We denote a separation of K by (A,B). When $n=3$, there are at least two partitions of K that are separations. All partitions of three points $\{k_1,k_2,k_3\}$ are separations if the triangle $k_1k_2k_3$ has only acute angles. When $n=4$, there are no separations if the points are located at the vertices of a square. For each of the integers $m=1$, 2, 3, or 4, there is an arrangement of four points in the plane such that there are exactly m separations. More generally, we devote this note to proving the following result.

Theorem. *There are at most n separations of n distinct points in the Euclidean plane. When* $n>2$ *there is a set of n distinct points in the plane such that there are exactly n separations.*

The second part of this theorem is easy: there are exactly n separations of the points

$(0,1),(0,-1),(2,0),(3,0),\ldots,(n-1,0)$ for each integer $n>2$. We prove the first part of the theorem by induction after certain preliminary results are established.

LEMMA 1. *If (A,B) is a separation of K and if A_1 and B_1, respectively, are nonempty proper subsets of A and B, then the partition $[A_1\cup B_1,K\backslash(A_1\cup B_1)]$ is not a separation of K.*

Proof. By a suitable choice of axes, we can assume that the set A lies in the first quadrant and the set B is in the third quadrant. Any line that separates $A_1\cup B_1$ and $K\backslash(A_1\cup B_1)$ into opposite half-planes must contain points in the first quadrant and points in the third quadrant. There cannot, therefore, be two such lines that are perpendicular. Hence, $[A_1\cup B_1,K\backslash(A_1\cup B_1)]$ is not a separation of K.

Let $\eta(P)$ denote the number of separations of a set P consisting of a finite number of points in the plane.

LEMMA 2. *If K is any set of four points in the plane, then $\eta(K)\leqslant 4$.*

Proof. Let $K=\{k_1,k_2,k_3,k_4\}$. The partitions of K are (1) $[k_1,K\backslash k_1]$, (2) $[k_2,K\backslash k_2]$, (3) $[k_3,K\backslash k_3]$, (4) $[k_4,K\backslash k_4]$, (5) $[k_1\cup k_2,k_3\cup k_4]$, (6) $[k_1\cup k_3,k_3\cup k_4]$, and (7) $[k_1\cup k_4,k_2\cup k_3]$. We show first that the first four of these partitions cannot all be separations of K. Indeed, if (1), (2), and (3) are separations of K, then $k_1k_2k_3$ must be a triangle with only acute angles. If k_4 is on or inside this triangle, then (4) is not a separation since an angle between the lines joining k_4 and the vertices of this triangle exceeds $\pi/2$. If k_4 is outside this triangle, $k_1k_2k_3k_4$ are vertices of a quadrilateral with a corner angle exceeding $\pi/2$. Thus, some k_j $(1\leqslant j\leqslant 4)$ cannot be separated from the other three points. Hence, at most three of the partitions (1), (2), (3), and (4) can be separations of K. Next, if any of the three partitions (5), (6), (7) is a separation of K, then the other two cannot be separations by Lemma 1. Thus, $\eta(K)\leqslant 4$.

There is a natural correspondence between partitions $[A,B]$ of a set K of n points in the plane and pairs of partitions of K', where $K'=k\cup K$ and the point $k\notin K$, given by $[A\cup k,B]$, $[A,B\cup k]$. If $[A,B]$ is not a separation of K, then neither of the corresponding partitions of K' can be separations of K'. There is exactly one partition of K' that does not correspond to some partition of K and it is $[k,K]$.

LEMMA 3. *Let the point $k\notin K=\{k_1,k_2,\ldots,k_n\}$, $n>2$. There is at most one separation (A,B) of K such that both $(A\cup k,B)$ and $(A,B\cup k)$ are separations of $K'=k\cup K$.*

Proof. Assume (A,B) is a separation of K such that both corresponding partitions of K' are separations. Any partition of K, other than $[A,B]$, has one of the following forms: (a) $[A_1,K\backslash A_1]$, (*b*) $[B_1,K\backslash B_1]$, (*c*) $[A_1\cup B_1,K\backslash(A_1\cup B_1)$, where A_1 and B_1 respectively are nonempty proper subsets of A and B. By Lemma 1, the partition (c) is not a separation of K. Hence, neither of its corresponding partitions of K' are separations. Of the two partitions of K' corresponding to (a), the partition $[A_1\cup k,K\backslash A_1]$ is not a separation of K' by Lemma 1 and the hypothesis that $(A,K\backslash(k\cup A))$ is a separation of K'. Similarly $[B_1\cup k,K\backslash B_1]$ is not a separation of K'. Hence, there is no second separation of K such that both corresponding partitions of K' are separations.

LEMMA 4. *Let K be a set of n points in the plane, where $n>2$, and let $k\notin K$. If $\eta(K)=n$ and C is a nonempty subset of K such that (k,K), $(k\cup C,K\backslash C)$, $(k\cup K\backslash C,C)$ are separations of $K'=k\cup K$, then there is a separation $(D,K\backslash D)$ of K such that neither $[k\cup D,K\backslash D]$ nor $[k\cup K\backslash D,D]$ are separations of K'.*

Proof. There exist two separations $(A,K\backslash A)$, $(B,K\backslash B)$ such that either A and B are proper subsets of C or of $K\backslash C$ and for which there are points $a\in A$, $b\in B$ where $a\notin B$, $b\notin A$. Indeed, if n_1 is the number of points in C and if two such sets A,B do not exist, then every pair of separations $(A_1,K\backslash A_1)$, $(A_2,K\backslash A_2)$, where $A_1\subset C$, $A_2\subset C$, must satisfy either $A_1\subset A_2$ or $A_2\subset A_1$. Since there are n_1 points in C, there can be at most n_1-1 separations of K of this form and distinct from C. Similarly, there are at most $n-n_1-1$ separations of K of the form $(A_1,K\backslash A_1)$

where A_1 is a proper subset of $K\setminus C$. Now all separations of K are of the form $(A_1,K\setminus A_1)$ where either $A_1\subset C$ or $A_1\subset K\setminus C$ by Lemma 1. Hence, there are $(n_1-1)+(n-n_1-1)+1=n-1$ separations of K, contrary to the assumption $\eta(K)=n$.

We can, therefore, assume there are two separations $(A,K\setminus A)$, $(B,K\setminus B)$, where A,B are subsets of $K\setminus C$ and there are points $a\in A$, $b\in B$ such that $a\notin B$, $b\notin A$. By the hypothesis, $[k,a\cup b\cup c]$, $[k\cup c,a\cup b]$, and $[k\cup a\cup b,c]$ are separations of $K_4=\{a,b,c,k\}$ for each $c\in C$. By Lemma 2, there is at most one more separation of K_4. Neither $[k\cup a,b\cup c]$ nor $[k\cup b,a\cup c]$ is a separation of K_4 by Lemma 1. Hence, at most one of the partitions $[k\cup b\cup c,a]$, $[k\cup a\cup c,b]$ is a separation of K_4. These facts imply at most one of the partitions $[k\cup K\setminus A,A]$, $[k\cup K\setminus B,B]$ is a separation of K' and neither of the partitions $[k\cup A,K\setminus A]$, $[k\cup B,K\setminus B]$ is a separation of K'. Therefore, one of the separations, $(B,K\setminus B)$ say, has the property that neither $[k\cup B,K\setminus B]$ nor $[k\cup K\setminus B,B]$ is a separation of K'.

We can now finish the proof of the main theorem. Assume for any set K of distinct points in the plane, $n>2$, that $\eta(K)\leqslant n$. Let $k\notin K$, $K'=k\cup K$. If there is no nonempty set $C\subset K$ such that $(k\cup C,K\setminus C)$ and $(k\cup K\setminus C,C)$ are separations of K', then each separation of K corresponds to at most one separation of K'. Since the partition $[k,K]$ can also be a separation, we conclude $\eta(K')\leqslant\eta(K)+1\leqslant n+1$ in this case. On the other hand, if there is such a subset C of K, then there is no second such subset of K distinct from C or $K\setminus C$ by Lemma 3. Furthermore, by Lemma 4, either $\eta(K)<n$ or there is a separation $(B,K\setminus B)$ of K such that neither of the corresponding partitions of K' are separations when (k,K) is a separation of K'. Hence $\eta(K')\leqslant n+1$ in any case and the proof is complete.

Think-A-Dot: A Useful Generalization

Michael Gemignani
Indiana University-Purdue University
Indianapolis, IN 46205

In September of 1967 there appeared in this Magazine an article [1] by Benjamin Schwartz entitled "Mathematical Theory of Think-A-Dot." In that paper Schwartz analyzed a then-popular game in which the player drops a marble into one of three holes in the top of a box, thereby changing the pattern on the face of the box. The object of the game is to use a succession of marble drops to change a starting pattern into some pattern which had been selected as a goal. Schwartz deduced rules to determine which patterns could be obtained from any given pattern.

Think-A-Dot is a fairly primitive game with a three row pattern

$$\begin{array}{ccccc}\times&&\times&&\times\\&\times&&\times&\\\times&&\times&&\times\end{array}$$

in which each × can be one of two possible colors. In the actual game, only some of the possible colored patterns can actually be achieved. The purpose of this paper is to generalize the game of Think-A-Dot, to solve a particular instance of that generalization, and to suggest that this may not be so unrelated to applications as it may seem at first glance.

A generalized Think-A-Dot box has a vertical face on which several rows of "entries" appear. While the entries may be color-coded dots, we consider them to be represented by integers $0,1,2,\ldots,n-1$. To operate Think-A-Dot, the player drops a marble into a port at the top of the box. This act changes the pattern of colors (numbers) on the face of the box according to whatever particular rule of transformation is associated with that particular action. The transformation rule is a function of the particular pattern displayed at the time the marble is dropped, as well as the port into which the marble is dropped. The transformation rules are "built into" the game itself, and, in this general situation, can be specified at will by the designer of the game. Usually, one would wish at most one entry to be changed in each row, but even this is not an absolute requirement.

Two patterns are "equivalent" if one can be obtained from the other by a sequence of marble drops. The problem we shall consider is to determine, for any generalized game of Think-A-Dot, if two patterns are equivalent, and to count the number of equivalence classes.

We solve this problem initially for a pattern consisting of a 3×3 matrix with entries in Z_3, the ring of integers mod 3, with the following transformation rule: when a marble hits an entry e, it will add 1 mod 3 to that entry, then drop to the next row, shifting e columns mod 3. (The columns are numbered $0,1,2$ from left to right.) In other words, if the marble is in column j and sees entry e, it will leave behind a new entry $(e+1)\bmod 3$ as it drops to the next row into column $(j+e)\bmod 3$. (After dropping through all 3 rows, it falls out of the box.) For example, starting with

$$\begin{matrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 1, \end{matrix}$$

dropping the marble into the middle column will produce the array

$$\begin{matrix} 1 & 1 & 0 \\ 0 & 2 & 0 \\ 1 & 1 & 2. \end{matrix}$$

We shall consider the following questions for this particular game.

(1) Can the transformations thus described be characterized algebraically?

(2) How many different patterns can be obtained from a given pattern by a sequence of marble drops?

(3) How can we tell if two patterns are equivalent?

Let D_j represent the transformation associated with dropping the marble into the jth column. Then the juxtaposition D_kD_j denotes the composite transformation formed by first dropping a marble into column j and then dropping a marble into column k. Our first important result is that $D_jD_k=D_kD_j$: *the order in which marbles are dropped makes no difference in the outcome.* Both D_j and D_k add 1(mod 3) to whatever entries they affect and leave every other entry unaffected. If D_j and D_k act on different entries of a particular row, then, clearly, for that row, they commute. If they act on the same element of a row, they each add 1(mod) to it; so, again, they commute for that row. Therefore they commute for every row and hence for the entire pattern.

Now consider D_0. No matter what entry is initially in column 0 in the first row, three successive drops into column 0, i.e., D_0^3, has the effect of adding 1(mod 3) to each entry of the second and third rows and returning the column 0 entry of row 1 to its original value. Also, D_0^6 has the effect of adding 2(mod 3) to each entry of the second and third rows and leaving the column 0 entry of row 1 unchanged. Likewise, D_0^9 returns the matrix to its original configuration. Analogous properties hold for D_1 and D_2. We therefore conclude that each D_j is of order 9, but that $D_0^3=D_1^3=D_2^3$.

Since we have already proved that the composition of transformations is commutative (in the special case under consideration), we will use additive rather than multiplicative notation for what follows since it will somewhat simplify our discussion. Thus, now D_2+2D_0 will denote dropping the marble twice into the 0th column followed by a drop into column 2. The previous discussion then leads us to conclude that the set of transformations is isomorphic to $T=(Z_9\oplus Z_9\oplus Z_9)/H$, where H is the subgroup generated by (3,0,0), (0,3,0), and (0,0,3). This group T contains exactly 81 elements. Therefore, given a matrix P, exactly 81 matrices can be obtained from P by a sequence of marble drops; moreover, each of these matrices can be obtained by some sequence of marble drops from any other matrix in this class. Since there are 3^9 possible 3×3 matrices with entries in Z_3 and each equivalence class contains 3^4 members, there are $3^5=243$ distinct equivalence classes.

To determine if two matrices are equivalent, we seek a canonical form to facilitate direct comparison. It is clear that from any matrix P we can obtain an equivalent matrix Q which has 0 as each entry in the first row. We call such a matrix with 0 as each entry in the first row **quasi-canonical.**

Suppose that Q_1 and Q_2 are each quasi-canonical matrices. If they are equivalent, then the sequence of marble drops that transforms Q_1 into Q_2 must have the form $uD_0+vD_1+wD_2$, where u, v, and w are each divisible by 3 since it is only in this way that each entry in the first row of Q_2 could be 0. But this, in turn, implies that we have added either 0, 1, or 2 to each entry of the second and third rows. Thus two quasi-canonical matrices Q_1 and Q_2 are equivalent if and only if Q_2 can be obtained from Q_1 by adding some one integer $j \pmod 3$ to each entry of the second and third rows of Q_1.

Adding 1 to each entry of the second and third rows is the same as applying $3D_0$; adding 2 to each entry of the second and third rows is the same as applying $6D_0$; while adding 0, of course, is the same as applying I, the identity transformation. By adding 0, 1, or 2, whichever is appropriate, to each entry of the second and third rows, we can see that each quasi-canonical matrix is equivalent to a quasi-canonical matrix with 0 as the column 0 entry of the second row.

We say that a matrix is in **canonical** form if it is quasi-canonical and has 0 as the column 0 entry of the second row. It follows from the preceding discussion that the canonical matrices are in 1-1 correspondence with the equivalence classes. Thus, two matrices are equivalent if and only if they are both equivalent to the same canonical matrix.

We next generalize our results from matrices with entries in Z_3 to matrices with entries in Z_n. Let P be an $n\times n$ matrix with entries in Z_n and columns numbered from 0 to $n-1$. When a marble hits an entry e, it will add $1 \bmod n$ to that entry, then drop to the next row, shifting e columns $\bmod n$. As before, a matrix is quasi-canonical if its first row is zero, and canonical if in addition the column 0 entry of its second row is also zero. Then by direct generalization of the Z_3 analysis we can establish the following results:

(a) The set T of transformations is isomorphic to the direct sum of n copies of Z_{n^2} modulo its subgroup generated by $(n,0,\ldots,0),(0,n,0,\ldots,0),\ldots,(0,0,\ldots,0,n)$.

(b) There are n^{n+1} matrices equivalent to any given matrix, and n^{n^2-n-1} equivalence classes.

(c) Two quasi-canonical matrices Q_1 and Q_2 are equivalent if and only if Q_2 can be obtained from Q_1 by adding some one integer $j \pmod n$ to each entry of the second through nth rows of Q_1.

(d) The canonical matrices are in 1-1 correspondence with the equivalence classes, so any two matrices are equivalent if they are equivalent to the same canonical matrix.

By varying the configuration of the rows, the admissible entries for the rows, and the transformation rules, an infinite number of distinct problems of the type actually solved in this paper can be generated. Such problems may arise in a practical context in certain situations; for example, where there are banks of switches, each switch having a finite number of settings, where resetting one switch in the first bank automatically effects changes in switch settings in other banks; or where there is a collection of groups of storage bins, where the addition of an item to a bin in one group determines which bin in the next group will subsequently receive an item and where the bins are emptied when the reach some finite capacity.

Reference

[1] B. L. Schwartz, Mathematical Theory of Think-A-Dot, this MAGAZINE, Vol. 40, No. 4 (September 1967) 187–193.

PROBLEMS

Dan Eustice, Editor
Leroy F. Meyers, Associate Editor
The Ohio State University

Proposals

To be considered for publication, solutions should be mailed before November 1, 1979.

1066. Consider the following children's game ("clock"): k copies of well-shuffled cards numbered $1,2,3,\ldots,L$ are distributed in boxes labeled $1,2,3,\ldots,L$, with exactly k cards per box. (For the game as normally played, a standard deck of playing cards is used, with $L=13$ and $k=4$.) At the start of the game the top card in box 1 is drawn. If the value of this card is j ($j=1,2,3,\ldots,L$) we proceed to box j, draw the top card and go to the box so numbered, draw the top card, and so on. The objective of the game (a 'win') is to draw all cards from every box before being directed to an empty box. Characterize all winning distributions of cards, and find the probability of a win. [*Eric Mendelsohn & Stephen Tanny, University of Toronto.*]

1067. Problem P. M. 11 on the first William Lowell Putnam Competition, April 16, 1938, was to find the length of the shortest chord that is normal to the parabola $y^2=2ax, a>0$, at one end. A calculus solution is quite straightforward. Give a completely "non-calculus" solution. [*M. S. Klamkin, University of Alberta.*]

1068. Given a simple closed curve S, let the "navel" of S denote the envelope of the family of lines that bisect the area within S.

(a) If S is a triangle, find sharp upper and lower bounds for the ratio of the area within the navel of S to the area within S.

(b)* If S bounds a convex set, find a sharp upper bound for this ratio.

(c)* If S is arbitrary, find a sharp upper bound for this ratio. [*James Propp, Great Neck, New York.*]

1069. Suppose $f:\mathbf{R}\to\mathbf{R}$ is continuous and for every rational q there exists an n with $f^n(q)=0$ (the nth iterate of f). Prove or disprove: For every real number t there is an n such that $f^n(t)=0$. [*F. David Hammer, University of California at Davis.*]

1070.* Let $p_1+p_2+\cdots+p_k=1$ be a sum of $k\geqslant 2$ probabilities and let M_n, for $n=1,2,\ldots$, be the multinomial distribution based on these probabilities and n trials. Event A_n occurs if, during the n trials, no possible outcome of the experiment occurs in two consecutive trials. Find the sum $\Sigma_{n=1}^{\infty}P(A_n)$. What are the convergence criteria for this sum to exist? [*Thomas E. Elsner & Joseph C. Hudson, General Motors Institute.*]

1071. Player A rolls $n+1$ dice and keeps the highest n. Player B rolls n dice. The higher total wins, with ties awarded to Player B.

(a) For $n=2$, show that Player A wins and find his probability of winning.

(b*) Find the smallest value of n for which Player B wins. [*Joseph Browne, Onondaga Community College.*]

Quickies

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

Q658. If $a,b>0$, prove that $a^b+b^a>1$. [*M.S. Klamkin, University of Alberta.*]

Q659. Show that for each complex number b the polynomial $P(z)=z^4+32z+b$ has a zero in $\{z\colon \operatorname{Re}(z)\geqslant 1\}$. [*Peter Ørno, The Ohio State University.*]

Solutions

Strictly Increasing **November 1977**

1027. Let $f(k)$ be a real-valued function on the nonnegative integers. Suppose that $f(0)=0$ and that $f(k)$ is a convex function. That is, for $k\geqq 1$, $f(k)$ is less than the average of $f(k-1)$ and $f(k+1)$. For integers $k, 1\leqq k\leqq n$, define

$$F_n(k)=f(k)q+f(r) \quad \text{for } n=kq+r, \qquad 0\leqq r<k.$$

Prove that, for fixed n, $F_n(k)$ is strictly increasing for $1\geqq k\geqq n$. [*Daniel B. Shapiro, The Ohio State University.*]

Solution: Since $f(k)$ is a convex function, $f(k)-f(k-1)<f(k+1)-f(k)$. Write $f(k)-f(k-1)=a_k$; then $a_k<a_{k+1}$ for every k. We have

$$F_n(k)=f(k)q+f(r) \text{ with } 0\leqslant r<k,$$

$$F_n(k+1)=f(k+1)p+f(s) \text{ with } 0\leqslant s<k+1,$$

and $n=kq+r=(k+1)p+s$ or $k(q-p)+r-s=p$. Therefore

$$\begin{aligned} F_n(k+1)-F_n(k) &= p(f(k+1)-f(k))+f(s)-(q-p)f(k)-f(r) \\ &= pa_{k+1}+\sum_{j=1}^{s} a_j-(q-p)\sum_{j=1}^{k} a_j-\sum_{j=1}^{r} a_j \\ &= pa_{k+1}-X. \end{aligned}$$

Now there are $(q-p)k+r-s$ or p terms added in X, none of them exceeding a_k. Therefore $pa_{k+1}-X\geqslant pa_{k+1}-pa_k>0$. Thus $F_n(k)$ is strictly increasing for $1\leqslant k\leqslant n$.

Gillian W. Valk
Tucker, Georgia

Also solved by G. A. Heuer, J. M. Stark, and the proposer.

1030. a. Solve the following functional-differential equation for the complex-valued differentiable function f:

$$f(s+t)=f(s)+f(t)-f'(s)f'(t) \text{ for all real } s \text{ and } t,$$

and $f(0)=0$.

b. If the real part of $f(t)$ is non-positive for all real t, but f is not identically zero, show that $f(t)=0$ only if $t=0$.

(This problem arose in connection with "infinitely divisible distributions" in probability.) [*G. Edgar, The Ohio State University.*]

Solution : (a) Let f satisfy

$$(*) \qquad f(s+t)=f(s)+f(t)-f'(s)f'(t) \text{ for all real } s \text{ and } t.$$

with

$$f(0)=0. \tag{1}$$

Setting $s=t=0$ in (*) yields

$$f'(0)=0. \tag{2}$$

If $f'(t)=0$ for all t, then $f\equiv 0$. So assume $f'(t_0)\neq 0$ for some t_0. With $s=t_0$, (*) may be rewritten as

$$f'(t)=f'(t_0)^{-1}\{f(t_0)+f(t)-f(t_0+t)\}.$$

This equation shows that f' has as many derivatives as f. Hence, f is infinitely differentiable. This justifies subsequent differentiations of (*).

Now differentiate (*) with respect to s to get

$$(**) \qquad f'(s+t)=f'(s)-f''(s)f'(t),$$

and set $s=0$ to get $f'(t)=-f''(0)f'(t)$. Since $f'(t)\not\equiv 0$, we conclude that

$$f''(0)=-1. \tag{3}$$

Differentiating (**) with respect to both t and s, and setting $s=0$ yields

$$(***) \qquad f'''(t)=-f'''(0)f''(t).$$

If $f'''(0)=0$, then $f'''(t)\equiv 0$. Conditions (1), (2) and (3) then imply $f(t)=-\frac{1}{2}t^2$, and this does satisfy (*).

So assume $f'''(0)=z\neq 0$. Then the solution of (***) with the initial conditions (1), (2), (3) is

$$f(t)=z^{-2}\{1-zt-e^{-zt}\}. \tag{4}$$

Substitution shows this is a solution of (*) for any $z\neq 0$. We therefore obtain the three solutions $f(t)\equiv 0, f(t)=-\frac{1}{2}t^2$, and (4).

(b) This is clearly true for $f(t)=-\frac{1}{2}t^2$. So let $f(t)$ be given by (4) with $z\equiv x+iy\neq 0$, and set $r=x^2+y^2$. Then

$$\begin{aligned} r\,\mathrm{Re}f(t) &= -xt+\frac{x^2-y^2}{r}-e^{-xt}\left\{\frac{x^2-y^2}{r}\cos yt-\frac{2xy}{r}\sin yt\right\} \\ &= -xt+r^{-1}(x^2-y^2)-e^{-xt}\cos(yt+\theta), \end{aligned}$$

where $r\cos\theta=x^2-y^2$, and $r\sin\theta=2xy$.

If $x\neq 0$ and $y\neq 0$, then the last term above will dominate the behavior (for large $-xt$), and $\mathrm{Re}f(t)$ will change sign. If $y=0$, then

$$f(t)=x^{-2}\{1-xt-e^{-xt}\}.$$

Since. $e^u>1+u$ for $u\neq 0$ (u real), the result follows in this case.

If $x=0$, then

$$f(t)=-y^{-2}\{1-\cos yt+i(\sin yt-yt)\}.$$

Thus $\text{Re}f(t) \leqslant 0$. In this case, $f(t)=0$ implies $yt = \sin yt$ so that $t=0$.

Hence, in all cases, $\text{Re}f(t) \leqslant 0$ implies the only zero of $f(t)$ is $t=0$.

Eli L. Isaacson
New York University

Also solved by J. M. Stark and the proposer. Partial solutions by B. D. Aggarwala & C. Nasim (Canada), H. Kappus (Switzerland), and James T. Smith.

A Standing Problem **January 1978**

1031. There are n people, numbered consecutively, standing in a circle. First #2 sits down, then #4, #6, etc., continuing around the circle with every other standing person sitting down until just one person is left standing. What is his number? (For example, with $n=6$, the seating order is 2, 4, 6, 3; 1 and 5 is left standing.) [*Richard A. Gibbs, Fort Lewis College.*]

Solution I: Write $n=2^i+j$ where $0 \leqslant j \leqslant 2^i - 1$. Then seat the people numbered $2,4,6,\ldots,2j$. This leaves 2^i people standing, beginning with the person numbered $2j+1$; call him Stan. Now continue to seat people until you get back to Stan. It is easy to see that 2^{i-1} people will be left standing, starting with Stan again. On every subsequent pass of the circle half of those standing will be left standing with Stan always the first among them. Stan's the man.

Mike Chamberlain
U.S. Naval Academy

Solution II: Let $F(n)$ be the number of the person left standing. Clearly

$$F(1)=F(2)=1. \qquad (1)$$

We claim that for $n \geqslant 2$

$$F(n+1)=\begin{cases} F(n)+2 & \text{if } F(n)<n, \\ 1 & \text{if } F(n)=n, \end{cases} \qquad (2)$$

so that if $2^k \leqslant n < 2^{k+1}$, then

$$F(n)=2(n-2^k)+1. \qquad (3)$$

Consider $n+1$ people standing, numbered clockwise 1 through $n+1$. After number 2 sits down, let each person standing be given a new number, beginning with person number 3 getting number 1 and continuing consecutively clockwise. Then the person left standing has a new number of $F(n)$. Hence (2) follows.

Finally, (3) follows because (1) and (2) determine F uniquely and the F given in (3) satisfies (1) and (2).

Russell Lyons, Undergraduate
Case Western Reserve University

Also solved by B. D. Aggarwala, C. Nasim, & J. Schaer (Canada), Mangho Ahuja, S. F. Barger, J. C. Binz (Switzerland), Carol A. Blomstrom, Michael Collins, Michael W. Ecker, Milton P. Eisner, Thomas E. Elsner, Robert S. Fisk, Herta T. Freitag, W. W. Funkenbusch, Michael Goldberg, Kenneth M. Gustin, F. David Hammer, Eli L. Isaacson, Steve Komie, Lew Kowarski, Robert T. Kurosaka, Mary S. Krimmel, Dennis Lichtman, Peter W. Lindstrom, Lance Littlejohn & Blair Spearman, J. M. Metzger, Zane C. Motteler, William Myers, George A. Novacky, Jr., Boon-Yian Ng, (Malaysia), Jim Peterson, Bob Prielipp, Howard W. Pullman, Reinhard Razen (Austria), James J. Reynolds, Steve Ricci, David Singmaster (England), Scott Smith, Rolf Sonntag, J. M. Stark, Ronald S. Tiberio, L. van Hamme (Belgium), Michael Vowe (Switzerland), Ann E. Watkins, Fred E. Watkins II, James D. Watson, Mead C. Whorton, Jr., Harry Zantopulos, and the proposer. Singmaster and Vowe provided the reference to the Josephus Problem in Ball and Coxeter, Mathematical Recreations and Essays, 12th ed., University of Toronto Press, pp. 32–36.

A Müntz-like Condition **January 1978**

1032. Let $l_1(x)=\log x$, $l_2(x)=\log\log x$, and $l_k(x)=\log l_{k-1}(x)$. Let $N(k)$ be the first integer n such that $l_k(n)>1$. When k is fixed, the integral test shows that the series

$$(\#)\qquad \sum_{n=N(k)}^{\infty}\frac{1}{nl_1(n)l_2(n)\cdots(l_k(n))^p}$$

diverges for $p=1$ and converges for $p>1$. R. P. Agnew [Amer. Math. Monthly, 54 (1947), 273–274; Selected papers on calculus, MAA 1969, pp. 348–349] called attention to the result that (#) is very slowly divergent if $p=1$ and k (the number of logarithmic factors in (#)) is no longer fixed but depends on n, being taken as large as possible so that all the logarithms exceed 1, i.e., so that $l_k(n)>1$ but $l_{k+1}(n)<1$. With this choice of $k=k(n)$, how large can $p=p(k)$ be before the series becomes convergent? (Will $p=2$ or $p=k$ suffice?) [*R. P. Boas, Northwestern University.*]

Solution: We will show that the series (#) is convergent if and only if $\Sigma_{k=1}^{\infty}1/p(k)$ is convergent, and hence neither $p=2$ or $p=k$ will suffice.

In proving the result we will use the following lemma. Let $0<x_0<x_1<\cdots<x_n<\cdots$ where (to avoid picky details) we assume $x_i-x_{i-1}>5$.

LEMMA 1: *Let $f(x)>0$ be defined and bounded on $[x_0,\infty)$, and be continuous and decreasing on $(x_i,x_{i+1}]$, $i=0,1,2,\ldots$, with right-hand limits $\lim_{x\to x_i^+}f(x)\equiv f(x_i+)$. Assume that $\Sigma_{i=1}^{\infty}f(x_i+)<\infty$. Then $\Sigma_{n=[x_1]}^{\infty}f(n)<\infty$ if and only if $\int_{x_0}^{\infty}f(x)dx<\infty$.*

Proof. If $\int_{x_0}^{\infty}f(x)dx=\infty$, then

$$\sum_{n=[x_1]}^{\infty}f(n)\geqslant\sum_{i=1}^{\infty}\sum_{j=[x_i]+1}^{[x_{i+1}]-1}f(j)\geqslant\sum_{i=1}^{\infty}\int_{[x_i]+1}^{[x_{i+1}]}f(x)dx$$

$$\geqslant\int_{[x_1]}^{\infty}f(x)dx-\sum_{i=1}^{\infty}\max(f(x_{i-1}+),f(x_i+))$$

$$\geqslant\int_{[x_1]}^{\infty}f(x)dx-2\sum_{i=0}^{\infty}f(x_i+)=\infty.$$

If $\int_{x_0}^{\infty}f(x)dx<\infty$, then

$$\sum_{n=[x_1]}^{\infty}f(n)=\sum_{i=1}^{\infty}(f([x_i])+f([x_i]+1))$$

$$+\sum_{i=1}^{\infty}\sum_{n=[x_i]+2}^{[x_{i+1}]-1}f(n)\leqslant\sum_{i=1}^{\infty}(f(x_{i-1}+)+f(x_i+))+\sum_{i=1}^{\infty}\int_{[x_i]+1}^{[x_{i+1}]-1}f(x)dx$$

$$\leqslant 2\sum_{i=1}^{\infty}f(x_i+)+\int_{x_0}^{\infty}f(x)dx<\infty.$$

This completes the proof of the lemma.

We now take $x_0=e, x_1=e^{x_0}, x_2=e^{x_1},\ldots,x_n=e^{x_{n-1}},\ldots$ and let $f(x)=[xl_1(x)l_2(x)\cdots(l_k(x))^p]^{-1}$ where k is the largest integer such that $l_k(x)>1$ and $l_{k+1}(x)\leqslant 1$. It is easily shown that $k=i$ when $x_{i-1}<x\leqslant x_i$. Making use of the fact that for $i\geqslant j$, $l_j(x_i)=x_{i-j}$, we have

$$f(x_i+)=\frac{1}{x_i\left(\prod_{j=1}^{i}l_j(x_i)\right)(l_{i+1}(x_i))^p}=\frac{1}{\left(\prod_{j=0}^{i}x_j\right)1^p}=\frac{1}{\prod_{j=0}^{i}x_j}.$$

From the ratio test it follows that $\Sigma_{i=0}^{\infty} f(x_i+) < \infty$. Since the hypotheses of Lemma 1 are satisfied the series $\Sigma_{n=[x_1]}^{\infty} f(n)$ (and hence $\Sigma_{n=1}^{\infty} f(n)$) is convergent if and only if $\int_{x_0}^{\infty} f(x)dx < \infty$. Using the fact that

$$\frac{d}{dx} l_k(x) = \frac{1}{x}\left(\prod_{i=1}^{k-1} l_i(x)\right)^{-1}$$

we have for $p \neq 1$,

$$\int_{x_{i-1}}^{x_i} f(x)dx = \frac{(l_i(x))^{-p+1}}{1-p}\bigg|_{x_{i-1}}^{x_i} = \frac{1-e^{-p+1}}{p-1}.$$

Thus

$$\int_{x_0}^{\infty} f(x)dx = \sum_{k=1}^{\infty} \frac{1-e^{-p(k)+1}}{p(k)-1}.$$

For this series to converge it is necessary that $p(k) \to \infty$. Now for large $p(k) > 0$ we have $1/(2p) \leqslant (1-e^{-p+1})/(p-1) \leqslant 2/p$. Thus (with $p(k) > 0$), $\Sigma_{k=1}^{\infty} 1/p(k)$ is convergent, and hence $\Sigma_{n=1}^{\infty} f(n)$ is convergent if and only if $\Sigma_{k=1}^{\infty} 1/p(k)$ is convergent.

Peter W. Lindstrom
St. Anselm's College

Also solved by J. M. Stark and the proposer.

Answers

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

Q658. Since the inequality is obviously valid if either a or $b \geqslant 1$, it suffices to consider $a = 1-x$ and $b = 1-y$, where $0 < x,y < 1$. Our inequality now becomes

$$\frac{1-x}{(1-x)^y} + \frac{1-y}{(1-y)^x} > 1.$$

By the mean value theorem,

$$(1-x)^y = 1 - \frac{xy}{(1-\theta x)^{1-y}} \leqslant 1 - xy.$$

Hence,

$$\frac{1-x}{(1-x)^y} + \frac{1-y}{(1-y)^x} \geqslant \frac{1-x+1-y}{1-xy} = \frac{2-x-y}{1-xy} - 1 + 1 = \frac{(1-x)(1-y)}{1-xy} + 1 > 1.$$

The stated inequality appears in D. S. Mitrinović, *Analytic Inequalities*, Springer-Verlag, Heidelberg, 1970, p. 281, where a more involved proof is given.

Q659. $P'(z) = 4z^3 + 32$ has zeros at -2, $2e^{\pi i/3}$, and $2e^{-\pi i/3}$. By the Gauss-Lucas Theorem these zeros are in the closed convex hull of the zeros of $P(z)$. Thus, at least one zero of $P(z)$ has real part greater than or equal to $2\cos\frac{\pi}{3} = 1$.

REVIEWS

Paul J. Campbell, Editor
Beloit College

Pierre Malraison, Editor
Control Data Corp.

Assistant Editor: Eric S. Rosenthal, Princeton University. Articles and books are selected for this section to call attention to interesting mathematical exposition that occurs outside the mainstream of the mathematics literature. Some reviews of books are adapted from the Telegraphic Reviews in the American Mathematical Monthly.

PRIME-80: Proceedings of a Conference on Prospects in Mathematics Education in the 1980's, MAA; 84 pp, $3.50 (P).

A multifaceted assessment of the current state of collegiate mathematics and its future. The recommendations may not be surprising, but they provide an important guide for the MAA and college educators. Throughout there is an emphasis on the importance of students' pre-college preparation: "There is much evidence available that it is extremely difficult, if not impossible, for colleges to make up in any reasonable amount of time for academic deficiencies of entering college students."

Overbye, Dennis, *The wizard of space and time*, Omni 1 (February 1979) 44-47.

Popular-level "gee whiz" interview with Stephen Hawking covering space-time, black holes and the big bang, in addition to interesting biographical detail.

Tobias, Sheila, Overcoming Math Anxiety, Norton, 1978; 278 pp, $10.95.

Tobias, not herself a mathematician but once a math avoider, explores in depth the causes of math anxiety, particularly among women. "Men have math anxiety too, but it disables women more." Although no alleged inherent differences in mathematical talent between women and men stand up to close scrutiny, differences in sex-role socialization do provide an explanation of observed divergences in interest and ability. In an individual case, however, apart from intelligence, motivation, and interest, it is temperament that may make the difference between achievement and failure: "How people cope with uncertainty, whether they can tolerate a certain amount of floundering, whether they are willing to take risks, what happens to their concentration when an approach fails, and how they feel about failure." All these are affected by socialization; they suggest a place to start, not only to remove math anxiety and learn math, but to learn about oneself.

Kogelman, Stanley and Warren, Joseph, Mind Over Math, Dial Pr, 1978; xii + 239 pp, $8.95.

The cover urges, "Put yourself on the road to success by freeing yourself from math anxiety." Writing for those who have at least tentatively decided to overcome fear and loathing of math, the authors emphasize awareness of emotional factors leading to anxiety and re-examination of negative feelings about mathematics. The book is based on their experience in helping groups of adults to help overcome math anxiety.

Roark, Anne C., *Solution to a 124-year-old problem creates a controversy among scholars*, Chron. of Higher Educ. (February 20, 1979) 3-4.

Account of the controversy attending philosopher Thomas Tymoczko's argument in the *Journal of Philosophy* that Haken and Appel's computer-based proof of the four color theorem represents a revolutionary shift in the definition of mathematical proof, from the *a priori* to the empirical, from the rule of demonstration to the rule of authority.

Steen, Lynn Arthur (Ed.), Mathematics Today: Twelve Informal Essays, Springer-Verlag, 1978; viii + 367 pp, $12.

The book's objective is to tell "the intelligent nonmathematician something of the nature, development and use of mathematical concepts." The essays, on very diverse topics, are at or somewhat above the *Scientific American* level. Only an intelligent nonmathematician can tell whether the book attained its objective; but most mathematicians should find at least some of the essays interesting and informative.

Perl, Teri, Math Equals: Biographies of Women Mathematicians + Related Activities, A-W, 1978; vi + 250 pp, (P).

The primary purpose of this book is to provide role models in mathematics. Interesting illustrated biographical sketches (Hypatia, du Châtelet, Agnesi, Germain, Somerville, Lovelace, Kovalerskaya, Young, Noether) are each followed by exposition of elementary mathematics related to the woman's area of research, together with accompanying problems and workbook-like activities.

Pearce, Peter and Pearce, Susan, Polyhedra Primer, Van N-Rein, 1978; viii + 134 pp, $5.95 (P).

Aimed at readers requiring a knowledge of spatial geometry. An illustrated glossary of over 200 terms, arranged in order of increasing complexity, covering polygons, tessellations, polyhedra, space filling and open packings. Great for browsing. Good index. Skimpy bibliography. Unsophisticated and non-mathematical but a steal at the price.

Honsberger, Ross, Mathematical Morsels, Dolciani Math. Expos., No. 3, MAA, 1978; xii + 249 pp.

"A showcase for some of mathematics' minor miracles"--a compendium of some of the bright ideas of elementary mathematics in action, as displayed in scores of elementary problems culled primarily from the *American Mathematical Monthly*. Honsberger has rewritten and embellished the presentations of the solutions.

Gries, David and Misra, J., *A linear sieve algorithm for finding large prime numbers*, Comm. Assoc. Comp. Mach. 21 (December 1978) 999-1003.

An algorithm for finding primes between 1 and n, which executes in time proportional to n. Linearity arises from using a sieve which never attempts to remove a non-prime which was removed earlier. The proof that the algorithm works and is in fact linear uses some nice number theory.

Bezuska, S., Kenney, M. and Silvey, L., Tessellations, The Geometry of Patterns, Creative Pub, 1977; vi + 169 pp, $6.50 (P).

Using tessellations as their central theme and concept, the authors lead the reader to attempt more and more complicated tessellations and learn about geometry from the success or failure of those attempts.

Carlbom, I. and Paciorek, J., *Planar geometric projections and viewing transformations*, ACM Comp. Surveys 10 (December 1978) 465-503.

How to implement some of the standard geometric projections of three-dimensional objects on a flat surface within the constraints of the Proposed Graphics Standard. (The entire issue is devoted to that proposal.) Homogeneous coordinates and linear transformations are used to describe mathematically the projections, and derive from the mathematical description an implementation.

Boyse, J.W., *Interference detection among solids and surfaces*, Comm. Assoc. Comp. Mach. 22 (January 1979) 3-9.

Solutions to two problems: internal representation of solids in a digital computer, and computation of whether or not those solids intersect. As interactive computer graphics becomes a more widespread tool, automated solution of the above problems becomes more desirable. The author looks at static intersections, and tests for collision between a moving and a stationary object.

Bremermann, H.J., *Complexity and transcomputability*, in Duncan, Ronald and Weston-Smith, Miranda (Eds.), The Encyclopedia of Ignorance, Pergamon Pr, 1977; $30, $15 (P), pp. 168-174.

Here is an article on the other meaning of complex, dealing with computational complexity and physical limits to speed of computers. An absolute upper bound is 1.35 x 10^{47} bits per second per gram of computer, assuming all the mass of the computer is spent on signal energy.

Brams, Steven J., The Presidential Election Game, Yale U Pr, 1978; xix + 242 pp, $15, $3.95 (P).

Primaries, conventions, the election: Brams applies scientific modeling in a popular style to give an analytic treatment of strategy in the race for the U.S. Presidency, his main tools being modern game theory and decision theory. He concludes with recommendation of abolition of the Electoral College and institution of a new form of voting called approval voting. Other topics considered are the bandwagon curve and a game-theoretic analysis of the White House tapes case.

Berry, Donald A. and Regal, Ronald R., *Probabilities of winning a certain carnival game*, American Statistician 32 (November 1978) 126-129.

Probabilities in the notorious ripoff carnival game of "razzle" (sometimes called "Play Football," "Auto Races," "Double-Up," or "Razzle Dazzle") are calculated to show the game is essentially unbeatable. The calculations exhibit interesting small differences between naive approximations and the exact probabilities.

Morrison, Philip, *On broken symmetries*, in Wechsler, Judith (Ed.), On Aesthetics in Science, MIT Pr, 1978; pp. 54-70.

Symmetry is often hailed as a manifestation of mathematical order and perfection in the world. But every macroscopic symmetry is broken: the world is finite and "immersed in disturbance." Morrison calls us to an appreciation of broken symmetry as "making visible both sides [order and imperfection] of the act of becoming."

Bell, Alex G., The Machine Plays Chess, Pergamon Pr, 1978; x + 114 pp, $10.

Historical survey of chess-playing machines, with behind-the-scenes details of recent world computer chess championships.

Parlett, Beresford, *Progress in numerical analysis*, SIAM Review 20 (June 1978) 443-456.

The author distinguishes between "inner-directed" and "outer-directed" research in numerical analysis, the subject having been entirely the former before the computer but now growing quietly in the latter manner. He analyzes "the tower of scientific computation" into component levels ranging from assembly language to in-house production codes, noting the growth and growth-potential of each with specific examples.

Dieudonné, Jean, Panorama des mathématiques pures. Le choix bourbachique, Gauthier-Villars (US Distr: SMPF, 14 E. 60th St., NY 10022), 1977; xv + 302 pp, 150 F.

An introductory (though high-level) survey of contemporary mathematics as seen through the eyes of the Seminar Bourbaki. This volume offers capsule descriptions of many major achievements of contemporary mathematics, with notes on the principal discoverers and bibliographic information, with particular emphasis on recent expository conferences. A good deal of effort has gone into indicating the relations among the achievements presented and between these achievements and science.

Beeler, Jeff, *Persistent teens pull off coup*, Computer World 12 (November 28, 1978) 1.

Two high school students fight parents and school administrators to find new largest prime number at Cal State University at Hayward. The new number: $2^{21,701} - 1$. The old record: $2^{19,937} - 1$.

Ramsey, F.P., Foundations; Essays in Philosophy, Logic, Mathematics and Economics, Humanities Pr, 1978; viii + 287 pp, $20.50.

Much of this book was first published in 1931 as The Foundations of Mathematics and Other Logical Essays; this edition omits a few essays in favor of new items such as his two papers in mathematical economics: *A mathematical theory of saving*, and *A contribution to the theory of taxation*. In addition to essays defending logicism, his paper on the combinatorial result now known as Ramsey's Theorem is included. The remaining half of the book is devoted to the contributions he made to philosophy before his death at age 26 in 1930.

Cole, Sam, *Modelling the international order*, Applied Mathematical Modelling 2:2 (June 1978) 66-76.

Examination of the techniques and usefulness of six world models of different types, including the "limits to growth" model.

Raeside, D.E., *et al.*, *Medical application of Fourier analysis*, SIAM Review 20 (October 1978) 850-854.

The article describes a successful use of Fourier analysis in classification of medical echocardiograms (made via ultrasound).

Foster, James E., Mathematics in Diversion, Fulton County Pr, 1978; xiii + 220 pp.

An engagingly-written ramble through the main kinds of mathematical puzzles and brainteasers--by no means a bare-bones puzzle book! It is marred, however, by a small error of attribution ("Jourdain curve theorem"), a crackpot "proof" of the Four Color Theorem attributed to Ball ("for some reason mathematicians choose to ignore it"), and other injudicious remarks.

Infeld, Leopold, Whom the Gods Love: The Story of Evariste Galois, NCTM, 1978; xvii + 323 pp, $9.78.

Reprint of the 1948 edition of the romantic novel.

NEWS & LETTERS

SCHATTSCHNEIDER ELECTED EDITOR

Doris J. Schattschneider of Moravian College was elected Editor of *Mathematics Magazine* for the five-year term 1981-85 at the January 1979 meeting of the Mathematical Association of America. Professor Schattschneider is currently an Associate Editor of the *Magazine*, and is the author (with Wallace Walker) of *M.C. Escher Kaleidocycles* (Ballentine Books, 1977).

SUMMER WORKSHOP

The Northeast Section of the M.A.A. will sponsor a summer workshop at the University of Maine, Orono 04469, June 18-22, 1979, concerning Applications of Mathematics in Medicine and Biology. Topics to be covered include discrete population dynamics, harvesting problems in discrete population models, deterministic models for communicable diseases, Monte Carlo models for common source epidemics. The principal lecturer will be Maynard Thompson, Professor of Mathematics at Indiana University, and co-author (with Daniel P. Maki) of *Mathematical Models and Applications*. Two lectures will be given each day. Other lectures on special topics will be arranged. The cost of the workshop, including room and board, is $110 for M.A.A. members, $120 for non-members.

MAKE JUNE A SINGULAR MONTH

A conference on catastrophe theory and its applications will be held June 4-8, 1979 at Salisbury State College, Salisbury, MD. The principal lecturer will be biologist Alexander Woodcock of Williams College, co-author of *Catastrophe Theory* (Dutton, 1978). This conference is sponsored by the Maryland-D.C.-Virginia Section of the MAA in order to make available to teachers in two- and four-year colleges important advances in applicable mathematics. For further information write to B.A. Fusaro, Department of Mathematical Sciences, Salisbury State College, Salisbury, MD 21801.

THE LAST LATE MAGAZINE?

Subscribers to *Mathematics Magazine* must by now assume that we are perpetually behind schedule. The November 1978 issue came out in March 1979, as did the January issue. This March issue will be out in April. However, the May issue will, we hope, be out at the end of May, and thereafter all issues should be on schedule. These delays have been caused primarily by the near-collapse of the compositors last fall, but since they have recovered we hope that the *Magazine* schedule will do likewise. We apologize to our readers for the confusion and uncertainty that these delays have caused.

A NON-VACUOUS APOLOGY

The word "Hoover" means many things to many people. Unfortunately when we wrote the cover caption for the November issue the name Herbert filled the vacuum in our minds. Obviously we should have written about the former director of the Federal Bureau of Investigation, not about the former president (or a household appliance, or a ball bearing). We hope the only consternation caused by this gaffe is ours alone.

A QUICKER QUICKIE

Quickie 466, November 1969, stated: If AB and BA are both identity matrices, then A and B are both square matrices. A more elegant solution than my original one or the subsequent one (this *Magazine*, November 1970) by J.L. Brenner is the following proof by Dave Lovelock, University of Arizona:

Let A and B be m x n and n x m matrices, respectively. Then, $AB = I_m$

and $BA = I_n$. It follows immediately that the traces are given by $Tr(AB) = m$ and $Tr(BA) = n$. Also, if $A = (a_{rs})$ and $B = (b_{rs})$, then

$$Tr(AB) = \sum_{r=1}^{m} \sum_{s=1}^{n} a_{rs} b_{sr}$$

and

$$Tr(BA) = \sum_{i=1}^{n} \sum_{j=1}^{m} b_{ij} a_{ji} .$$

Sincé these two expressions are the same (change the dummy indices $i \to s$, $j \to r$) we must have $m = n$.

Murray S. Klamkin
University of Alberta
Edmonton, Alberta
Canada T6G 2G1

1979 CHAUVENET PRIZE

Dr. Neil J.A. Sloane of Bell Laboratories, Murray Hill, NJ was awarded the 1978 Chauvenet Prize for noteworthy exposition for his paper "Error-Correcting Codes and Invariant Theory: New Applications of a Nineteenth Century Technique." This prize, represented by a certificate and a check for $500, is the twenty-seventh award of the Chauvenet Prize since its inception by the Mathematical Association of America in 1925.

Sloane was born in Beaumaris, Wales, on October 10, 1939. He received the B.E.E. and B.A. degrees (with Honors) from the University of Melbourne in 1959 and 1960, and the M.S. and Ph.D. degrees from Cornell University, Ithaca, NY in 1964 and 1967. From 1956 to 1961 Dr. Sloane worked for the Postmaster General's Department of the Commonwealth of Australia, and from 1963 to 1965 he was a research assistant with the Cognitive Systems Research Program at Cornell University. In 1967 he became an assistant professor of Electrical Engineering at Cornell, and became a member of the Technical Staff of Bell Labs.

Dr. Sloane is engaged in research in coding theory, communication theory, and combinatorial mathematics. He is the author of four books: *A Handbook of Integer Sequences* (Academic Press, 1973); *A Short Course on Error-Correcting Codes* (Springer-Verlag, 1975); *The Theory of Error-Correcting Codes*, with F.J. MacWilliams, (Elsevier/North-Holland, 1977), and *Hadamard Encoding of Optical Instruments: Spectrometers and Imaging Devices*, with Martin Harwit (Academic Press, in preparation). Dr. Sloane is an editor of the *SIAM Journal on Applied Mathematics*, and the editor-in-chief of the *IEEE Transactions on Information Theory*. He is a fellow of the IEEE, and a member of the American Mathematical Society and the Mathematical Association of America.

UNDERGRADUATE RESEARCH PROGRAMS

The National Science Foundation announced recently 127 grants for Undergraduate Research Participation (URP) for the summer of 1979; five of these programs are in mathematics. Since most of these programs accept applications from students at other institutions, we list below the project director and address for each of these five programs. Numbers in the left margin indicate the number of available stipends: students may earn up to $1000 for summer research in URP programs. Students interested in applying for these programs should communicate directly with the project directors.

7 Dr. K.L. De Bouvere
Department of Mathematics
University of Santa Clara
Santa Clara
California 95053

6 Dr. E.F. Stueben
Department of Mathematics
Illinois Institute of Technology
Chicago
Illinois 60616

3 Dr. Joseph A. Gallian
Department of Mathematical Sciences
University of Minnesota, Duluth
Duluth
Minnesota 55812

10 Dr. Robert Z. Norman
Department of Mathematics
Dartmouth College
Hanover
New Hampshire 03755

7 Dr. Milos A. Dostal
Department of Mathematics
Stevens Institute of Technology
Hoboken
New Jersey 07030

SOLUTIONS TO THE 1978 PUTNAM EXAM

In January we printed in this column questions from the 1978 Putnam Examination. To assist those who have been puzzling over these problems, we provide here hints and answers. The official report on the results of the competition, including names of winners and complete sample solutions, will be published later this year in the American Mathematical Monthly.

A-1. Let A be any set of 20 distinct integers chosen from the arithmetic progression 1, 4, 7, ... , 100. Prove that there must be two distinct integers in A whose sum is 104.

Sol. Each of the twenty integers of A must be in one of the eighteen disjoint sets $\{1\}$, $\{52\}$, $\{4, 100\}$, $\{7, 97\}$, ..., $\{49, 55\}$. The result follows from the pigeon-hole principle.

A-2. Let $a, b, p_1, p_2, \ldots, p_n$ be real numbers with $a \neq b$. Define $f(x) = (p_1-x)(p_2-x)(p_3-x)\ldots(p_n-x)$. Show that the determinant of

$$\begin{bmatrix} p_1 & a & a & a & \ldots & a & a \\ b & p_2 & a & a & \ldots & a & a \\ b & b & p_3 & a & \ldots & a & a \\ b & b & b & p_4 & \ldots & a & a \\ \cdot & \cdot & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & & \cdot & \cdot \\ b & b & b & b & \ldots & p_{n-1} & a \\ b & b & b & b & \ldots & b & p_n \end{bmatrix}$$

equals $[bf(a)-af(b)]/(b-a)$.

Sol. A solution can be given using induction. For the inductive step, subtract the second column entries from those in the first column and expand by cofactors of the new first column.

A-3. Let $p(x) = 2 + 4x + 3x^2 + 5x^3 + 3x^4 + 4x^5 + 2x^6$. For k with $0 < k < 5$, define

$$I_k = \int_0^\infty \frac{x^k}{p(x)}\, dx.$$

For which k is I_k smallest?

Sol. By letting $x = 1/t$, one finds that $I_k = I_{4-k}$. By the arithmetic-geometric means inequality,

$$\frac{x^k+x^{4-k}}{2} \geq \sqrt{x^k x^{4-k}} = x^2 .$$

It follows that $I_k = (I_k+I_{4-k})/2 \geq I_2$; that is, I_k is smallest when $k = 2$.

A-4. A "by-pass" operation on a set S is a mapping from $S \times S$ to S with the property $B(B(w,x), B(y,z)) = B(w,z)$ for all w, x, y, z in S.

(a) Prove that $B(a,b) = c$ implies $B(c,c) = c$ when B is a bypass.

(b) Prove that $B(a,b) = c$ implies $B(a,x) = B(c,x)$ for all x in S when B is a bypass.

(c) Construct a table for a bypass operation B on a finite set S with the following three properties:

(i) $B(x,x) = x$ for all x in S.

(ii) There exist d and e in S with $B(d,e) = d \neq e$.

(iii) There exist f and g in S with $B(f,g) \neq f$.

Sol. (a) If $B(a,b) = c$ then $B(c,c) = B(B(a,b),B(a,b)) = B(a,b) = c$.

(b) If $B(a,b) = c$ then $B(a,x) = B(B(a,b),B(x,x)) = B(c,B(x,x)) = B(B(c,c),B(x,x)) = B(c,x)$.

(c) Let I and J be sets, each with more than one element, and let S denote the cartesian product $I \times J$. Define the operation B by $B((i,j),(h,k)) = (i,k)$.

A-5. Let $0 < x_i < \pi$ for $i = 1, 2, \ldots, n$, and set $x = (x_1+x_2+\ldots+x_n)/n$. Prove that

$$\prod_{i=1}^{n} \frac{\sin x_i}{x_i} \leq \left(\frac{\sin x}{x}\right)^n.$$

Sol. Let $g(x) = \ln[(\sin x)/x]$. Since $g''(x) < 0$ for $0 < x < \pi$, the graph of $g(x)$ is concave downward. Therefore

$$\frac{1}{n}\sum_{i=1}^{n} g(x_i) \leq g\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = g(x).$$

This implies that

$$\prod_{i=1}^{n} \frac{\sin x_i}{x_i} = e^{\Sigma g(x_i)} \leq e^{ng(x)} = \left(\frac{\sin x}{x}\right)^n .$$

A-6. Let n distinct points in the plane be given. Prove that fewer than $2n^{3/2}$ pairs of them are unit distance apart.

Sol. Let $f(n)$ be the maximum number of pairs with unit distance apart from a set of n points in a plane. Then $f(1) = 0$, $f(2) = 1$, $f(3) = 3$, and $f(4) = 5$.

Suppose we have an array of $n \geq 2$ points which realizes $f(n)$ pairs unit distance apart. Suppose each point has at least k points (in the set) unit distance away and that one of them, namely x_0, has exactly k points unit distance away. Let these points be $x_1,\ldots,x_k$.

For $i = 0,1,\ldots,k$, let C_i be the circle with radius 1 and center x_i. For $i > 0$, C_i goes through x_0 and at most two other x_j. Therefore, there are at least k-3 points of the array other than $x_0,\ldots,x_k$ on C_i. Any one of these k-3 points can appear on at most one other C_j since there are only two unit circles through x_0 and another point. Hence

$$n \geq 1 + k + \frac{k(k-3)}{2} = 1 + \frac{k(k-1)}{2} .$$

Thus k is such that the triangular number $k(k-1)/2$ is less than n; we let the largest such integer k be k_n.

The array with x_0 removed has at most $f(n-1)$ pairs unit distance apart; therefore $f(n) \leq f(n-1) + k_n$. By repeating this argument we find that $f(n) \leq k_2 + k_3 + \ldots + k_n$.

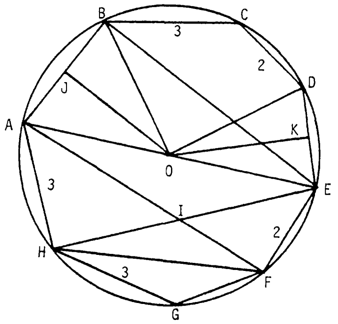
The definition of k_n implies that $k_n = t$ for $\binom{t}{2} < n \leq \binom{t+1}{2}$. From this it follows that $k_n \leq 1 + \sqrt{2(n-1)}$. Hence

$$\begin{aligned} f(n) &\leq k_2 + k_3 + \ldots + k_n \\ &\leq n-1 + \sqrt{2}(\sqrt{1}+\sqrt{2} + \ldots + \sqrt{n-1}) \\ &\leq n-1 + \sqrt{2}\int_1^n \sqrt{x}\, dx \\ &= n-1 + 2\sqrt{2}(n^{3/2}-1)/3 \\ &\leq n + .95n^{3/2} < 2n^{3/2} . \end{aligned}$$

B-1. Find the area of a convex octagon that is inscribed in a circle and has four consecutive sides of length 3 units and the remaining four sides of length 2 units. Give the answer in the form $r + s\sqrt{t}$ with r, s, and t positive integers.

Sol. The area is the same as for an octagon inscribed in a circle with sides alternately 3 units and 2 units in length. There are many ways to proceed. One approach is to observe that all angles measure $3\pi/4$ and one can augment the octagon into a square with sides of length $3 + 2\sqrt{2}$ by extending the sides of length 3 (isosceles right triangles are added on each of the four sides of length 2). Hence the desired area is

$$(3+2\sqrt{2})^2 - 4\cdot\tfrac{1}{2}\cdot\sqrt{2}\cdot\sqrt{2} = 13 + 12\sqrt{2} .$$



Alternatively, $\Delta HGF \cong \Delta HIF$ by ASA. Angles HAF and HEF are 45°, and angles AHE and AFE are right angles. Thus $HI = 3$, $IF = 2$, $AI = 3\sqrt{2}$ and $IE = 2\sqrt{2}$, so $BE = HE = 3 + 2\sqrt{2}$ and $AD = AF = 2 + 3\sqrt{2}$. Since $\Delta AJO \sim \Delta ABE$ and $AO = \frac{1}{2}AE$, it follows that $OJ = (3+2\sqrt{2})/2$. Similarly, $OK = (2+3\sqrt{2})/2$. Thus, the area of the octogon is $4(\frac{1}{2}\cdot 3\cdot(3+2\sqrt{2})/2) + 4(\frac{1}{2}\cdot 2\cdot(2+3\sqrt{2})/2) = 13 + 12\sqrt{2}$.

B-2. Express

$$\sum_{n=1}^{\infty} \sum_{m=1}^{\infty} \frac{1}{m^2n + mn^2 + 2mn}$$

as a rational number.

Sol. Let S be the desired sum. Then

$$S = \sum_{n=1}^{\infty} \frac{1}{n(n+2)}[(1-\frac{1}{n+3}) + (\frac{1}{2} - \frac{1}{n+4}) + \ldots]$$

$$= \sum_{n=1}^{\infty} \frac{1}{n(n+2)}[1 + \frac{1}{2} + \ldots + \frac{1}{n+2}]$$

$$= \frac{1}{2} \sum_{n=1}^{\infty} (\frac{1}{n} - \frac{1}{n+2})(1 + \frac{1}{2} + \ldots + \frac{1}{n+2})$$

$$= \frac{1}{2}[(1-\frac{1}{3})(1+\frac{1}{2}+\frac{1}{3}) + (\frac{1}{2} - \frac{1}{4})(1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}) + \ldots]$$

$$= \frac{1}{2}[(1+\frac{1}{2}+\frac{1}{3}) + \frac{1}{2}(1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}) + (\frac{1}{3\cdot 4}+ \frac{1}{4\cdot 5}+\ldots)+$$

$$(\frac{1}{3\cdot 5}+\frac{1}{4\cdot 6}+ \ldots)]$$

$$= \frac{7}{4} .$$

B-3. The sequence $\{Q_n(x)\}$ of polynomials is defined by

$$Q_1(x) = 1 + x, \; Q_2(x) = 1 + 2x,$$

and, for $m \geq 1$, by

$$Q_{2m+1}(x) = Q_{2m}(x) + (m+1)xQ_{2m-1}(x) ,$$

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

Let x_n be the largest real solution of $Q_n(x) = 0$. Prove that $\{x_n\}$ is an increasing sequence and that $\lim_{n\to\infty} x_n = 0$.

Sol. Clearly, $x_1 = -1$, $x_2 = -1/2$. An easy induction shows that each Q_n is positive for $x \geq 0$. Hence $x_n < 0$, if Q_n has zeros. Assume inductively that $x_1 < x_2 < \ldots < x_{2m-1} < x_{2m}$. Then $Q_{2m-1}(x) > 0$ for $x > x_{2m-1}$. In particular, $Q_{2m-1}(x_{2m}) > 0$. Hence

$$Q_{2m+1}(x_{2m}) = Q_{2m}(x_{2m}) + (m+1)x_{2m}Q_{2m-1}(x_{2m})$$
$$= (m+1)x_{2m}Q_{2m-1}(x_{2m}) < 0 .$$

This implies that $Q_{2m+1}(x) = 0$ for some $x > x_{2m}$, i.e., $x_{2m+1} > x_{2m}$. Similarly, one shows that $x_{2m+2} > x_{2m+1}$.

Let $a = -1/(m+1)$. Using the given recursive definition of the $Q_n(x)$, one finds that

$$Q_{2m+2}(a) = Q_{2m+1}(a) - Q_{2m}(a) = -Q_{2m-1}(a) .$$

Hence at least one of $Q_{2m+2}(a)$ and $Q_{2m-1}(a)$ is nonpositive. Thus either $x_{2m+2} \geq a$ or $x_{2m-1} \geq a$. But each of these implies that both $x_{2m+2} \geq -1/(m+1)$ and $x_{2m+3} \geq -1/(m+1)$. It follows that $-2/n \leq x_n < 0$ for all n and then that $\lim_{n\to\infty} x_n = 0$.

B-4. Prove that for every real number N, the equation

$$x_1^2 + x_2^2 + x_3^2 + x_4^2 =$$
$$x_1x_2x_3 + x_1x_2x_4 + x_1x_3x_4 + x_2x_3x_4$$

has a solution for which x_1, x_2, x_3, x_4 are all integers larger than N.

Sol. Clearly $(1,1,1,1)$ is a solution. Thinking of x_1, x_2, x_3 as fixed, the equation is quadratic in x_4 and one sees that the x_4 of a solution can be replaced by $x_4' = x_1x_2 + x_1x_3 + x_2x_3 - x_4$ to obtain a new solution when $x_4' \neq x_4$.

Also, the x_i may be permuted arbitrarily since the equation is symmetric in the x_i. Thus we may assume that $x_4 \leq m = \min(x_1, x_2, x_3)$. Also assume that each $x_i \geq 1$. Then $x_4' \geq 3m^2 - m > m$. This implies that one can start with the solution $(1,1,1,1)$ and through repeated use of the procedures stated above obtain a solution with each x_i an integer greater than N.

B-5. Find the largest A for which there exists a polynomial $P(x) = Ax^4 + Bx^3 + Cx^2 + Dx + E$, with real coefficients, which satisfies $0 \leq P(x) \leq 1$ for $-1 \leq x \leq 1$.

Sol. With $Q(x) = [P(x) + P(-x)]/2$, the condition becomes $0 \leq Q(x) = Ax^4 + Cx^2 + E \leq 1$ over $-1 \leq x \leq 1$. Letting $x^2 = y$, this becomes $0 \leq R(y) = Ay^2 + Cy + E \leq 1$ over $0 \leq y \leq 1$. The maximum value of A corresponds to the parabola which goes through the points $(0,1)$, $(1/2,0)$, and $(1,1)$: $4y^2 - 4y + 1$. That is, the maximum A is 4.

B-6. Let p and n be positive integers. Suppose that the numbers $c_{h,k}$ ($h=1,2, \ldots , n$; $k=1,2, \ldots , ph$) satisfy $0 \leq c_{h,k} \leq 1$. Prove that $\Sigma(c_{h,k}/h)^2 \leq 2p\Sigma c_{h,k}$, where each summation is over all admissible ordered pairs (h,k).

Sol. Let

$$a_h = \left(\sum_{k=1}^{ph} c_{h,k}\right)/h .$$

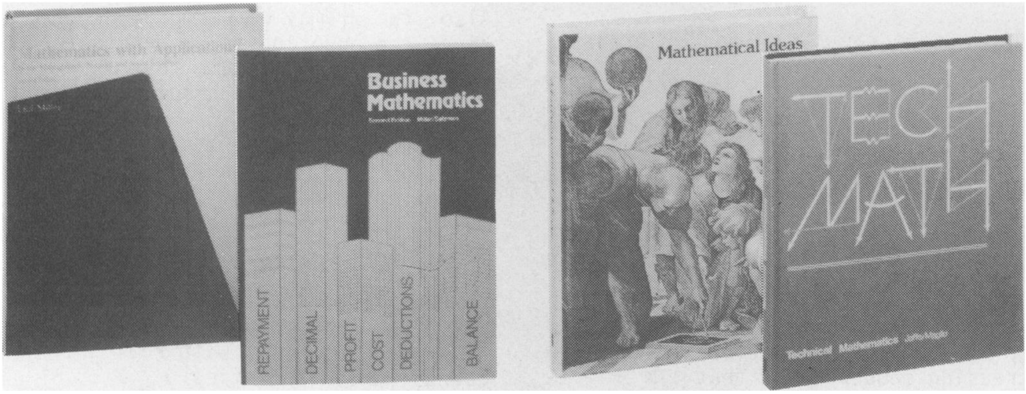
Clearly, $0 \leq a_h \leq p$. It follows by induction on n that

$$\sum_{n=1}^{n} a_h{}^2 \leq 2p \sum_{h=1}^{n} (ha_n) ,$$

and this is equivalent to the assertion of the problem.

More Problems...

From basics to applications—



Mathematics with Applications

In the Management, Natural, and Social Sciences
Second Edition
Lial / Miller
672 pages, hardbound $15.95
Almost 40% more problems than the nearest competitor—more word problems and almost 25% more drill problems. 15 cases, examples and applications throughout, glossary, and complete testing program. Instructor's Guide with Tests, and Selected Solutions, and Study Guide with 27 Computer Programs.

Business Mathematics

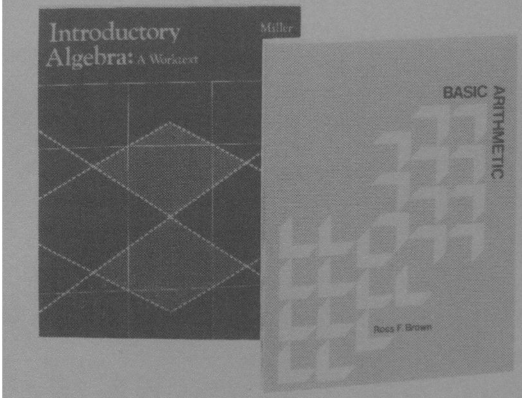
Second Edition
Miller / Salzman
432 pages, paperback $11.95
Lots of exercises, including word problems, plus quizzes and a complete testing program. Learning objectives, glossary, arithmetic review, and new chapters. Test-packed Instructor's Guide, and Tutorial Audio Tape cassettes.

Mathematical Ideas

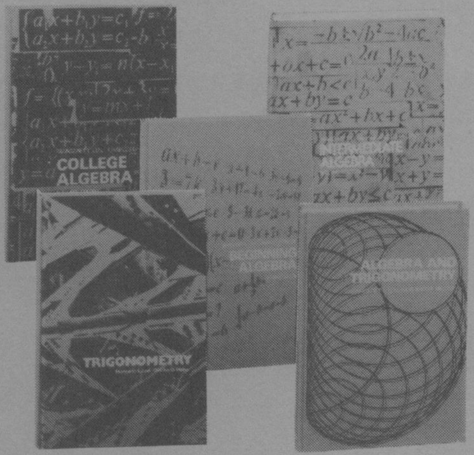
Third Edition
Miller / Heeren
512 pages, hardbound $14.95
For liberal arts students: 2600 exercises, over 500 worked-out examples, lots of everyday consumer math, emphasis on problem-solving skills, new chapters and chapter sequence, chapter summaries and tests, historical insights, and new illustration program. Expanded Instructor's Guide.

Technical Mathematics

Jaffe / Maglio
576 pages, hardbound $14.95
Word problems in three categories—general, technical, and shop, for practice with real job problems. Large number of topics, worked-out examples, arithmetic review. Complete Instructor's Guide with tests.



Introductory Algebra: A Worktext
Miller
BASIC ARITHMETIC
Ross F. Brown



COLLEGE ALGEBRA
INTERMEDIATE ALGEBRA
BEGINNING ALGEBRA
TRIGONOMETRY
ALGEBRA AND TRIGONOMETRY

THE CARUS MATHEMATICAL MONOGRAPHS

The Monographs are a series of expository books intended to make topics in pure and applied mathematics accessible to teachers and students of mathematics and also to non-specialists and scientific workers in other fields.

These numbers are currently available:

1. *Calculus of Variations,* by G. A. Bliss.
2. *Analytic Functions of a Complex Variable,* by D. R. Curtiss.
3. *Mathematical Statistics,* by H. L. Rietz.
4. *Projective Geometry,* by J. W. Young.
6. *Fourier Series and Orthogonal Polynomials,* by Dunham Jackson.
8. *Rings and Ideals,* by N. H. McCoy.
9. *The Theory of Algebraic Numbers* (Second edition), by Harry Pollard and Harold G. Diamond.
10. *The Arithmetic Theory of Quadratic Forms,* by B. W. Jones.
11. *Irrational Numbers,* by Ivan Niven.
12. *Statistical Independence in Probability, Analysis and Number Theory,* by Mark Kac.
13. *A Primer of Real Functions* (Second edition), by Ralph P. Boas, Jr.
14. *Combinatorial Mathematics,* by H. J. Ryser.
15. *Noncommutative Rings,* by I. N. Herstein.
16. *Dedekind Sums,* by Hans Rademacher and Emil Grosswald.
17. *The Schwarz Function and its Applications,* by Philip J. Davis.
18. *Celestial Mechanics,* by Harry Pollard.

One copy of each Carus Monograph may be purchased by individual members of the Association for $9.00 each; additional copies and copies for nonmembers are priced at $12.50 each. (Orders for under $10.00 must be accompanied by payment. Prepaid orders will be delivered postage and handling free.)

Orders should be sent to:

MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth Street, N.W.
Washington, D.C. 20036

THE MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth Street, N.W.
Washington, DC 20036

MATHEMATICS MAGAZINE VOL. 52, NO. 2, MARCH 1979