PROCEEDINGS

Computation Seminar

DECEMBER

1949

PROCEEDINGS

Computation Seminar

DECEMBER

1949

EDITED BY IBM APPLIED SCIENCE DEPARTMENT

CUTHBERT C. HURD, *Director*

INTERNATIONAL BUSINESS MACHINES CORPORATION

NEW YORK ✦ NEW YORK

Form 22-8342-0

PRINTED IN THE UNITED STATES OF AMERICA

FOREWORD

A COMPUTATION SEMINAR, sponsored by the International Business Machines Corporation, was held in the IBM Department of Education, Endicott, New York, from December 5 to 9, 1949. Attending the Seminar were one hundred and seven research engineers and scientists who are experienced both in applying mathematical methods to the solution of physical problems and in the associated punched card methods of computation. Consequently, these Proceedings represent a valuable contribution to the computing art. The International Business Machines Corporation wishes to express its appreciation to all those who participated in this Seminar.

CONTENTS

PARTICIPANTS

ALT, FRANZ L., *Associate Chief*
Computation Laboratory, National Bureau of Standards
Washington, D. C.

ARNOLD, KENNETH J., *Assistant Professor of Mathematics*
University of Wisconsin
Madison, Wisconsin

BAILEY, EDWARD W., *Statistician*
Y-12 Plant, Carbide and Carbon Chemicals Corporation
Oak Ridge, Tennessee

BARBER, E. A.
Engineering Laboratory, IBM Corporation
Endicott, New York

BAUER, S. H., *Professor of Chemistry*
Cornell University
Ithaca, New York

BELL, WILLIAM D., *Vice-President*
Telecomputing Corporation
Burbank, California

BELZER, JACK, *Mathematician*
Cryogenic Laboratory, Ohio State University
Columbus, Ohio

BENNETT, CARL A., *Statistician*
General Electric Company
Richland, Washington

BERMAN, JULIAN H., *Flutter Analyst*
Fairchild Aircraft Corporation
Hagerstown, Maryland

BINGHAM, RONALD H., *Research Specialist*
Ansco Division of General Aniline and Film Corporation
Binghamton, New York

BISCH, PAUL E., *Engineer*
In Charge of Special Structures, North American Aviation, Incorporated
Los Angeles, California

BRAGG, JOHN, *Professor of Chemistry*
Cornell University
Ithaca, New York

BRANDT, A. E., *Biometrician*
Atomic Energy Commission
New York, New York

BRILLOUIN, LEON, *Director*
Electronic Education, IBM Corporation
New York, New York

CLARK, H. KENNETH
Department of Pure Science, IBM Corporation
New York, New York

CONCORDIA, CHARLES, *Engineer*
General Electric Company
Schenectady, New York

COULTER, W. WAYNE, *Assistant Director of Research*
International Chiropractors Association
Davenport, Iowa

CROSSMAN, STUART L., *Group Supervisor*
Computing Laboratory, United Aircraft Corporation
East Hartford, Connecticut

CURL, GILBERT H., *Senior Physicist*
Navy Electronics Laboratory
San Diego, California

DISMUKE, NANCY M., *Mathematician*
Oak Ridge National Laboratory
Oak Ridge, Tennessee

DOCKERTY, STUART M., *Research Physicist*
Corning Glass Works
Corning, New York

DUKE, JAMES B., *Analytical Engineer*
Hamilton Standard Division, United Aircraft Corporation
East Hartford, Connecticut

DUNCOMBE, RAYNOR L., *Astronomer*
Nautical Almanac Division, U. S. Naval Observatory
Washington, D. C.

DYE, WILLIAM S. III, *Supervisor*
Tabulating Division, The Pennsylvania State College
State College, Pennsylvania

ECKERT, WALLACE J., *Director*
Department of Pure Science, IBM Corporation
New York, New York

FERBER, BENJAMIN, *Research Engineer*
Consolidated Vultee Aircraft Corporation
San Diego, California

FINLAYSON, L. D., *Process Control and Product Engineer*
Corning Glass Works
Corning, New York

GELLMAN, HARVEY, *Staff Mathematician*
Computation Centre, McLennan Laboratory, University of Toronto
Toronto, Ontario

GOODMAN, L. E., *Assistant Professor of Civil Engineering*
Graduate College, University of Illinois
Urbana, Illinois

GOTLIEB, CALVIN C., *Acting Director*
Computation Centre, McLennan Laboratory, University of Toronto
Toronto, Ontario

GREMS, MANDALAY D., *Analytical Engineer*
General Electric Company
Schenectady, New York

GROSCH, H. R. J., *Senior Staff Member*
Watson Scientific Computing Laboratory, IBM Corporation
New York, New York

HAMMER, PRESTON C., *Staff Member*
Los Alamos Scientific Laboratory, University of California
Los Alamos, New Mexico

HAMMING, RICHARD W., *Mathematician*
Bell Telephone Laboratories, Incorporated
Murray Hill, New Jersey

HANKAM, ERIC V.
Watson Scientific Computing Laboratory, IBM Corporation
New York, New York

HARDER, EDWIN L., *Consulting Transmission Engineer*
Westinghouse Electric Company
East Pittsburgh, Pennsylvania

HASTINGS, BRIAN T., *Project Engineer*
Computation Branch, Air Materiel Command
Wright Field, Dayton, Ohio

HASTINGS, CECIL JR., *Associate Mathematician*
The RAND Corporation
Santa Monica, California

HEISER, DONALD H., *Mathematician*
U. S. Naval Proving Ground
Dahlgren, Virginia

HENRY, HARRY C., *Chief*
Office of Air Research, Air Materiel Command, Wright Field
Dayton, Ohio

HERGET, PAUL, *Director*
Cincinnati Observatory, University of Cincinnati
Cincinnati, Ohio

HORNER, JOHN T., *Project Engineer*
Allison Division, General Motors Corporation
Indianapolis, Indiana

HUMMEL, HARRY H., *Associate Physicist*
Argonne National Laboratory
Lemont, Illinois

HUNTER, J. STUART, *Assistant Statistician*
University of North Carolina
Raleigh, North Carolina

HURD, CUTHBERT C., *Director*
Applied Science Department, IBM Corporation
New York, New York

JOHNSON, PHYLLIS C., *Statistician*
Y-12 Plant, Carbide and Carbon Chemicals Corporation
Oak Ridge, Tennessee

JOHNSON, WALTER H.
Applied Science Department, IBM Corporation
New York, New York

KAC, MARK, *Professor of Mathematics*
Cornell University
Ithaca, New York

KEAST, FRANCIS H., *Chief Aerodynamicist*
Gas Turbine Division, A. V. Roe, Canada, Limited
Malton, Ontario

KELLER, ALLEN
General Electric Company
Lynn, Massachusetts

KELLY, JOHN P., *Head*
Central Statistical Laboratory,
K-25 Plant, Carbide and Carbon Chemicals Corporation
Oak Ridge, Tennessee

KIMBALL, EVERETT, JR., *Research Associate*
Massachusetts Institute of Technology
Cambridge, Massachusetts

KING, GILBERT W., *Research Chemist*
Arthur D. Little, Incorporated, and Research Laboratory for Electronics
Massachusetts Institute of Technology, Cambridge, Massachusetts

KOCH, WARREN B., *Aeronautical Engineer*
Glenn L. Martin Company
Baltimore, Maryland

KRAFT, HANS, *Aerodynamicist*
Turbine Engineering Division, General Electric Company
Schenectady, New York

KRAWITZ, ELEANOR
Watson Scientific Computing Laboratory, IBM Corporation
New York, New York

KUNZ, KAISER S., *Associate Professor of Electrical Engineering*
Case Institute of Technology
Cleveland, Ohio

LEVIN, JOSEPH
Computation Laboratory, National Bureau of Standards
Washington, D. C.

LOTKIN, MARK, *Mathematician*
Ballistic Research Laboratories, Aberdeen Proving Ground
Aberdeen, Maryland

LOWE, JOHN, *Staff Assistant*
Engineering Tabulating, Douglas Aircraft Company, Incorporated
Santa Monica, California

LUCKEY, BONALYN A., *Engineering Assistant*
General Electric Company
Schenectady, New York

MADDEN, JOHN D., *Mathematician*
The RAND Corporation
Santa Monica, California

MALONEY, CLIFFORD J., *Chief*
Statistical Branch, Camp Detrick
Frederick, Maryland

MARSH, H. WYSOR, JR., *Chief Mathematics Consultant*
U. S. Navy Underwater Sound Laboratory
New London, Connecticut

MCPHERSON, JOHN C., *Vice-President*
IBM Corporation
New York, New York

MITCHELL, WILBUR L., *Mathematician*
Holloman Air Force Base
Alamogordo, New Mexico

MONROE, ROBERT J.
Institute of Statistics, University of North Carolina
Raleigh, North Carolina

MORRIS, PERCY T., *Technical Assistant to the Chief*
Statistical Division, U. S. Air Force, Wright Field
Dayton, Ohio

MORRISON, WINIFRED
The Texas Company
Beacon, New York

MORTON, J. E.
New York State School of Industrial and Labor Relations
Cornell University, Ithaca, New York

MOSHMAN, JACK, *Statistician*
U. S. Atomic Energy Commission
Oak Ridge, Tennessee

MYERS, FRANKLIN G., *Design Specialist*
Glenn L. Martin Company
Baltimore, Maryland

PENDERY, DONALD W.
Applied Science Department, IBM Corporation
New York, New York

POLACHEK, H., *Mathematician*
Naval Ordnance Laboratory, White Oak,
Silver Springs, Maryland

PORTER, RANDALL E.
Physical Research Unit, Boeing Airplane Company
Seattle, Washington

RICE, REX, JR., *Research Engineer*
Northrop Aircraft Company
Hawthorne, California

RICH, KENNETH C., *Mathematician*
Naval Ordnance Test Station
Inyokern, California

RIDER, WILLIAM B.
Applied Science Department, IBM Corporation
St. Louis, Missouri

RINALDI, LEONARD D., *Mathematician*
Cornell Aeronautical Laboratory, Incorporated
Buffalo, New York

ROCHESTER, NATHANIEL
Engineering Laboratory, IBM Corporation
Poughkeepsie, New York

SAMUEL, ARTHUR L.
Engineering Laboratory, IBM Corporation
Poughkeepsie, New York

SCHMIDT, CARL A., JR., *IBM Supervisor and Coordinator*
Fairchild Engine and Airplane Corporation
Hagerstown, Maryland

SCHUMACKER, LLOYD E., *Flight Research Engineer*
Flight Test Division, Headquarters Air Materiel Command
Wright Field, Dayton, Ohio

SCHUTZBERGER, HENRY, *Division Leader*
Test Data Division, Sandia Corporation
Albuquerque, New Mexico

SHELDON, JOHN
Applied Science Department, IBM Corporation
New York, New York

SHREVE, DARREL R.
Research Laboratory, The Carter Oil Company
Tulsa, Oklahoma

SMITH, ALBERT E., *Chemist*
Physics Department, Shell Development Corporation
Emeryville, California

SMITH, ROBERT W., *Mathematician*
U. S. Bureau of Mines
Pittsburgh, Pennsylvania

SONHEIM, DANIEL W., *Research Analyst*
Ordnance Aerophysics Laboratory, Consolidated Vultee
Aircraft Corporation, Daingerfield, Texas

SOROKA, WALTER W., *Associate Professor of Engineering Design*
College of Engineering, University of California
Berkeley, California

SOUTHWORTH, W. T., *Director*
Punched Card Applications, The State College of Washington
Pullman, Washington

SPENCER, ROBERT S., *Research Physicist*
Dow Chemical Company
Midland, Michigan

STEWART, ELIZABETH A.
Department of Pure Science, IBM Corporation
New York, New York

STULEN, FOSTER B., *Chief Structures Engineer*
Propeller Division, Curtiss Wright Corporation
Caldwell, New Jersey

THOMPSON, PHILIP M., *Physicist*
Hanford Works, General Electric Company
Richland, Washington

TOBEN, GREGORY J., *Supervisor*
IBM Group, Northrop Aircraft, Incorporated
Hawthorne, California

TUKEY, JOHN W., *Associate Professor of Mathematics*
Princeton University
Princeton, New Jersey

TURNER, L. RICHARD, *Coordinator of Computing Techniques*
Lewis Flight Propulsion Laboratory, NACA
Cleveland, Ohio

VERZUH, FRANK M., *Research Associate*
Electrical Engineering, Massachusetts Institute of Technology
Cambridge, Massachusetts

WAHL, ARTHUR M., *Advisory Engineer*
Westinghouse Electric Company
Pittsburgh, Pennsylvania

WETMORE, WARREN L., *Physicist*
Research Laboratory, Corning Glass Works
Corning, New York

WHEELER, BYRON W., JR.
Corning Glass Works
Corning, New York

WILSON, LEWIS R., JR.
Tabulating Department, Consolidated Vultee Aircraft Corporation
Fort Worth, Texas

WOLANSKI, HENRY S., *Aerodynamicist*
Consolidated Vultee Aircraft Corporation
Fort Worth, Texas

WOMBLE, AETNA K.
Department of Pure Science, IBM Corporation
New York, New York

YORKE, GREGORY B., *Statistician*
A. V. Roe, Canada, Limited
Malton, Ontario

YOWELL, EVERETT C., *Mathematician*
Institute for Numerical Analysis, National Bureau of Standards
Los Angeles, California

*The Future of High-Speed Computing**

JOHN VON NEUMANN

Institute for Advanced Study

A MAJOR CONCERN which is frequently voiced in connection with very fast computing machines, particularly in view of the extremely high speeds which may now be hoped for, is that they will do themselves out of business rapidly; that is, that they will out-run the planning and coding which they require and, therefore, run out of work.

I do not believe that this objection will prove to be valid in actual fact. It is quite true that for problems of those sizes which in the past—and even in the nearest past—have been the normal ones for computing machines, planning and coding required much more time than the actual solution of the problem would require on one of the hoped-for, extremely fast future machines. It must be considered, however, that in these cases the problem-size was dictated by the speed of the computing machines then available. In other words, the size essentially adjusted itself automatically so that the problem-solution time became longer, but not prohibitively longer, than the planning and coding time.

For faster machines, the same automatic mechanism will exert pressure toward problems of larger size, and the equilibrium between planning and coding time on one hand, and problem-solution time on the other, will again restore itself on a reasonable level once it will have been really understood how to use these faster machines. This will, of course, take some time. There will be a year or two, perhaps, during which extremely fast machines will have to be used relatively inefficiently while we are finding the right type and size problems for them. I do not believe, however, that this period will be a very long one, and it is likely to be a very interesting and fruitful one. In addition, the problem types which lead to these larger sizes can already now be discerned, even before the extreme machine types to which I refer are available.

Another point deserving mention is this. There will probably arise, together with the large-size problems which are in "equilibrium" with the speed of the machine, other and smaller, "subliminal" problems, which one may want to do on a fast machine, although the planning and programming time is longer than the solution time, simply because it is not worthwhile to build a slower machine for smaller problems, after the faster machine for larger problems is already available. It is, however, not these "subliminal" problems, but those of the "right" size which justify the existence and the characteristics of the fast machines.

Some problem classes which are likely to be of the "right" size for fast machines are of the following:

1. In hydrodynamics, problems involving two and three dimensions. In the important field of turbulence, in particular, three-dimensional problems will have to be primarily considered.

2. Problems involving the more difficult parts of compressible hydrodynamics, especially shock wave formation and interaction.

3. Problems involving the interaction of hydrodynamics with various forms of chemical or nuclear reaction kinetics.

4. Quantum mechanical wave function determinations —when two or more particles are involved and the problem is, therefore, one of a high dimensionality.

In connection with the two last-mentioned categories of problems, as well as with various other ones, certain new statistical methods, collectively described as "Monte Carlo procedures," have recently come to the fore. These require the calculation of large numbers of individual case histories, effected with the use of artificially produced "random numbers." The number of such case histories is necessarily large, because it is then desired to obtain the really relevant physical results by analyzing significantly large samples of those histories. This, again, is a complex of problems that is very hard to treat without fast, automatic means of computation, which justifies the use of machines of extremely high speed.

*This is a digest of an address presented at the IBM Seminar on Scientific Computation, November, 1949.

Some Methods of Solving Hyperbolic and Parabolic Partial Differential Equations

RICHARD W. HAMMING

Bell Telephone Laboratories

THE MAIN PURPOSE of this paper is to present a broad, non-mathematical introduction to the general field of computing the solutions of partial differential equations of the hyperbolic and parabolic types, as well as some related classes of equations. I hope to show that there exist methods for reducing such problems to a form suitable for formal computation, with a reasonable expectation of arriving at a usable answer.

I have selected four particular problems to discuss. These have been chosen and arranged to bring out certain points which I feel are important. The first problem is almost trivial as there exist well-known analytical methods for solving it, while the last is a rather complicated partial differential-integral equation for which there is practically no known mathematical theory.

To avoid details, I shall give only a brief introduction to the physical situation from which the equations came. Nor shall I dwell at all on the importance or meaning of the solutions obtained.

Lastly, I have chosen only equations having two independent variables, usually a space variable and a time variable. Similar methods apply to equations having three and more independent variables.

I have not attempted to define rigorously what is meant by hyperbolic or parabolic partial differential equations, nor shall I later. Instead, I intend to bring out certain common properties, and inferentially these properties define the classes of equations. In fact, from a computer's point of view it is the class of problems which is amenable to the same type of attack that provides the natural classification. It is on this basis that I have included a partial differenial-integral equation as the last example.

Each of the four problems is carried successively further toward its solution until, in the last example, I have given the detailed steps which were actually used.

If, in the rest of the paper, I do not mention any names, it should not be inferred that I did everything alone; on the contrary, I have at times played but a minor part in the entire effort.

The Wave Equation

The classic and best known example of a hyperbolic partial differential equation in two independent variables is the wave equation:

$$\frac{\partial^2 w}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 w}{\partial t^2} .$$

This is the equation which describes the propagation of signals, w, in one dimension, x. The signals progress in time, t, with a velocity, c. This equation is a linear equation and, as such, there is a large body of theory available for use in solving it. Thus, it is not likely that anyone would be called upon to solve it numerically except in very unusual circumstances. Nevertheless, I have chosen it as my first example, since I hope its simplicity and familiarity to you will aid in bringing out the main points I wish to make.

In solving partial differential equations it is customary to replace the given equations with corresponding difference equations, and then to solve these difference equations. Whether one looks at the approximation as being made once and for all and then solving the difference equations as exactly as possible, or whether one looks at the difference equations as being an approximation at every stage is a matter of viewpoint only. I personally tend to the latter view.

In the case at hand, the second differences are clearly used as approximations to the second derivatives. Such a choice immediately dictates an equally spaced rectangular net of points at which the problem is to be solved. Such a net is shown in Figure 1. The space coordinate, x, is vertical while the time coordinate, t, is horizontal. Thus, at a fixed time, t, we look along the corresponding vertical line to see what the solution is in space, x.

Suppose for the moment that a disturbance occurs at the upper point at time t. As time goes on the disturbance will spread out in space as shown in the figure. The space covered by the disturbance at any later time is indicated by the length of the corresponding vertical shading line at that time. The area of this disturbance in the figure is

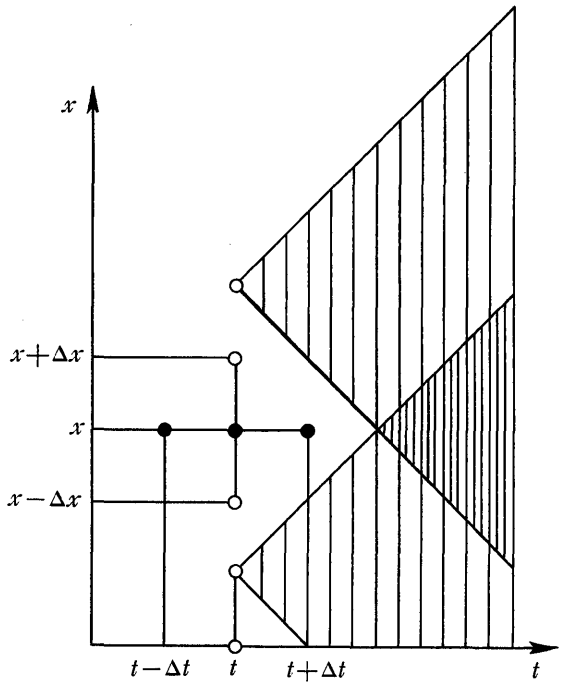


FIGURE 1. WAVE EQUATION $\frac{\partial^2 w}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 w}{\partial t^2}$

called the "cone of disturbance." The slopes of the bounding lines indicate the velocity of propagation c, and in this simple case they are straight lines. In the mathematical theory of partial differential equations the lines are called "characteristics."

The figure shows a second disturbance started at the same time, t, but at a lower point. This, too, spreads out as time goes on, and there finally occurs a time when the two cones overlap.

Consider, again, the given equation. The second difference in the x direction is calculated from the three points which are connected by the vertical line. This is to be equated to $1/c^2$ times the second difference in the time direction, which naturally uses the three solid black points. Suppose that the solution of the problem, up to the time t, is known, then we have an equation giving an estimate of the solution at one point at a later time, $t + \Delta t$.

Suppose, now, that the spacing in the x direction is kept as before, but the spacing in t is increased so as to predict as far as possible into the future. It should be obvious that the spacing in t cannot increase so far that the advanced point falls into the cones of disturbance of the first two points which are neglected. To do so is to neglect effects that could clearly alter the estimate of what is going to happen at such a point. Thus, it is found that, for a given spacing in the x direction, and a formula of a given span for estimating derivatives in the x direction, there is a maximum permissible spacing in the t direction, beyond which it is impossible to go and still expect any reasonable answer. In this simple case the condition may be written

$$\Delta t \leqq \frac{\Delta x}{c} .$$

Supposing that this condition has been satisfied, and also, the solution up to some time t is known, the above method may be used to advance the solution a distance Δt at all interior points. A point on the boundary cannot be so advanced. There must be independent information as to what is happening on the boundary. Such conditions are called "boundary conditions" and are usually given with the problem. The simplest kind of boundary condition gives the values of the dependent variable w by means of a definite formula. More complex situations may give only a combination of a function w and its derivative dw/dx. Such situations may require special handling when solving the problem in a computing machine, but in principle are straightforward.

A step forward at all points x is usually called a "cycle," and the solution of a problem consists of many, many repetitions of the basic cycle. The sole remaining question is that of starting the solution for the first cycle or two. Just as in the case of ordinary differential equations, this usually requires special handling and is based on some simple Taylor's series expansion of the solution. In practice, this step is often done by hand before the problem is put onto whatever machines are available.

As remarked before, this problem is not realistic, so having made a few points about spacing, boundary conditions, and initial, or starting, conditions, let us turn to a more complex problem.

THE TWO-BEAM TUBE

The two-beam tube is a tube with two electron beams going down the length of it together. Upon one beam of electrons a signal is imposed. The second beam, which has a slightly greater velocity, interacts with the first beam through their electric fields, and may be regarded as giving up some of its energy to the first beam. This in turn produces, one hopes, an amplification of the signal on the first beam.

The equations describing one particular idealization of such a tube are:

$$\left.\begin{aligned} \frac{\partial \rho_i}{\partial t} + \frac{\partial}{\partial x}(\rho_i v_i) &= 0 \\ \frac{\partial v_i}{\partial t} + v_i \frac{\partial}{\partial x}(v_i) &= \frac{\Psi}{2} \end{aligned}\right\} \quad i = 1, 2$$

$$\frac{\partial \Psi}{\partial x} = k^2 \Phi + (\rho_1 + \rho_2)$$

$$\frac{\partial \Phi}{\partial x} = \Psi$$

where the solution is to be periodic in time of period 1, and we are given information as to the state of affairs at the beginning of the tube, $x = 0$. The upper two equations for $i = 1$ describe one of the beams, while for $i = 2$ they describe the other beam. The lower two equations describe the interaction between the two beams of electrons.

I shall gloss over any questions of existence theorems for such a system and merely suppose that there is a solution. The information needed to start the problem at $x = 0$ comes from the " linear" theory which is not hard to find from a "linearized" form of the equations. We are here called upon to calculate the essentially nonlinear aspects of the tube.

The first reduction of the equations has already been performed before they were written as above, namely, that of transforming out of the equations all of the various constants and parameters of the problem that we could. In their present state the v_i of the equations give the velocities of the two beams measured in units of the mean velocity, the ρ_i the corresponding charge densities of the beams measured in mean charge density units, while the Φ and Ψ describe the electric field in suitable dimensionless units.

Since we are expecting a "wave-like" solution, it is convenient to transform to a moving coordinate system which moves with the expected mean velocity of the two beams. In such a coordinate system, the dependent variables ρ_i, v_i, Φ and Ψ may be expected to change slowly.

The equations obtained by such a transformation,

$$x = \sigma \qquad \sigma = x$$

or

$$t = \tau + \sigma \qquad \tau = t - x \ ,$$

are

$$\frac{\partial}{\partial\sigma}[\rho_i v_i] = \frac{\partial}{\partial\tau}[\rho_i(v_i - 1)]$$

$$\frac{\partial}{\partial\sigma}[v_i^2] = \frac{\partial}{\partial\tau}[(v_i - 1)^2] + \Psi$$

$$\frac{\partial\Psi}{\partial\sigma} = \frac{\partial\Psi}{\partial\tau} + (\rho_1 + \rho_2) + k^2\Phi$$

$$\frac{\partial\Phi}{\partial\sigma} = \frac{\partial\Phi}{\partial\tau} + \Psi \ ,$$

where the solution is still periodic in time τ with period 1.

In solving the usual hyperbolic type of equation, one advances step by step in time, but in this problem a periodic condition in time is given on the solution, and were the time to be advanced, it would be difficult to make the solution come out periodic. There would also be difficulty in finding suitable starting conditions. Instead of advancing in time, advancement is step by step down the length of the tube in the σ direction, using the periodic condition in τ to help estimate the derivatives in the τ direction at the ends of the interval. Thus, the periodic condition in effect supplies the boundary conditions.

One may calculate, if one wishes, the characteristic lines and determine the cones of disturbance, but in this case it must be looked at sidewise, as it were. Assuming that the solution is known for an interval of time, how far in space may the solution be predicted at a time corresponding to the mid-point of the time interval? If the cones of disturbance were to be calculated, it would be found, as is usual in nonlinear problems, that the velocity of propagation depends on the solution which is to be calculated. For example, a large shock wave of an explosion travels at a velocity that depends not only on the medium through which it passes, but also upon the amplitude of the shock wave itself.

Let us turn to the question of choosing a net of points at which we shall try to calculate an approximate solution. The use of a two-point formula in the τ direction for estimating the τ derivatives requires a great many points and produces a very fine spacing in the σ direction. If a four-point formula is chosen (a three-point one is hardly better than a two-point one for estimating first derivatives), the following is obtained,

$$f'(0) = \frac{f(-3/2) - 27f(-1/2) + 27f(1/2) - f(3/2)}{24\triangle\tau} + \epsilon,$$

with an error term of the order of

$$\epsilon \sim \frac{3}{640}(\triangle\tau)^4 f^{(V)}(\Theta) \ .$$

A formula like this is easiest to obtain by expanding each term about the mid-point of the interval in a Taylor's series with a remainder. Since in this moving coordinate system we expect a sinusoidal variation in time, the fifth derivative is estimated from the function

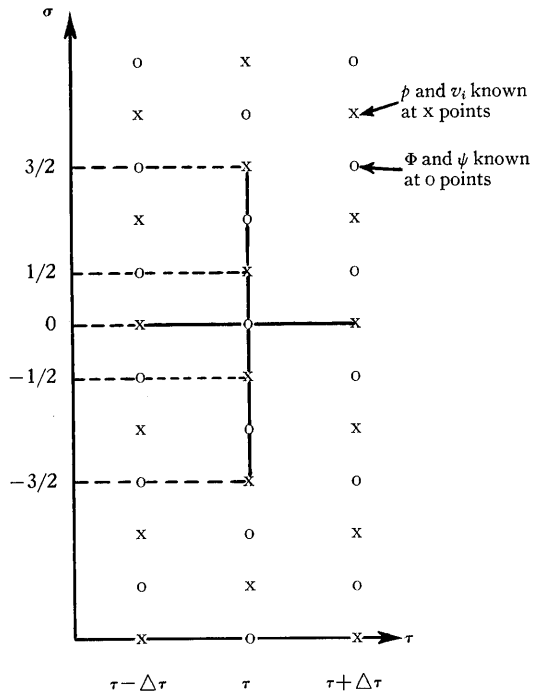
$$f = \sin 2\pi\tau \ .$$

In order to obtain an accuracy of about 1 part in 10^4 as the maximum error, it is necessary to choose 24 points in the τ direction—a most fortunate choice in view of the product $24\triangle\tau$ in the denominator of the estimate.

The statement that the maximum error at any one point is at most one part in 10^4 tells very little about the accumulated error due to many, many small errors, but as far as I know, there are no general methods which are both practical and close enough to help out in this type of situation. My solution to this dilemma is twofold:

1. To the person who proposed the problem, I pose questions such as, "If the solution is disturbed at a given point, will it tend to grow, remain the same, or die out?" At the same time, I try to answer the question independently, keeping an eye on the routine to be used in the solution of the problem.
2. Hope that a properly posed problem does have a solution, and that human ingenuity is adequate to the problem at hand.

Such a method may lead one astray on occasions, but with nothing else to fall back on, I feel that it is better than



$$f'(0) = \frac{f(-3/2) - 27f(-1/2) + 27f(1/2) - f(3/2)}{24(\triangle x)} + \epsilon$$

where $\epsilon \sim \frac{3}{640}(\triangle x)^4 f^{(V)}(\Theta)$

FIGURE 2

inactivity. One should, of course, make a diligent effort to resolve this dilemma, but pending that time, go ahead and hope for the best, ready at any time to suspect the solutions obtained.

Returning to our problem, some of the advantages of a four-point formula for estimating the derivatives in the τ direction have been listed. Let us look at the disadvantages in general. In the first place, except in periodic cases such as this one, where one can double back the solution from the top of the interval to add on to the bottom, the difficult problem of estimating derivatives near the boundaries arises. In the second place, one faces the task of assembling information from four different points to estimate the derivative at any one point. If, for example, the four values lie on four different IBM cards, it is not easy to get the information together. One method would be to calculate both 1 and 27 times the value of the function on each card and then on one pass through the accounting machine-summary punch equipment, using selectors and four counters to accumulate the running sums, punch out the proper totals along with suitable identification on each summary punched card.

To estimate the derivatives on the σ direction to advance one step, a simple two-point formula is used. Since both the four- and two-point formulas give estimates at the mid-points of the intervals, one is led to a "shifting net" of points as shown in Figure 2. Such a net leads to some slight troubles in the identification of points, but gives probably the least calculation where it is necessary to deal with many odd order derivatives. At least in this case, it certainly does. I have glossed over the accuracy of the estimate of the derivative in the σ direction, but in this case it was adequate, due to the fineness in the spacing necessary to satisfy the net spacing condition in $\triangle\sigma$ and $\triangle\tau$.

Let us drop this problem at this point and take up the next example.

A Parabolic Partial Differential Equation

The most common parabolic partial differential equation in two independent variables has the form

$$\frac{\partial B}{\partial t} = \frac{4\pi\gamma}{c^2}\frac{\partial^2 H}{\partial x^2} .$$

Such an equation describes the flow of heat, the diffusion of material, and the magnetization of iron.

In the particular case we shall discuss, a thin slab of permalloy is given, 2 mils thick and of infinite extent in the other two directions. This slab is subjected to an external field H which is changing in a sinusoidal fashion with frequency f. The question posed is that of determining the frequency of the external field such that B at the center of the slab rises to 90 per cent of its maximum possible value.

I would like to digress here for a moment to remark that it appears to me to be frequently the case that one is asked to determine a constant in the boundary conditions, or a parameter of the problem, such that the solution will have a given form. This is often a way of measuring a physical constant; indeed, when one finds a problem whose solution is sensitive to some parameter, then this may well provide a way of measuring that quantity with a high degree of precision.

Returning to the problem, it is immediately noted that in heat flow or diffusion there is no concept of velocity of propagation; hence the ideas of characteristics and cones of disturbance are of little help. Nevertheless, there is a condition on the spacing in x and t. To arrive at a necessary condition, suppose that at some point an error ϵ in H is committed, due, perhaps, to roundoff. This produces in turn an error of 2ϵ in the second difference. Following this through it is found that there is an error of

$$\frac{c^2\triangle t}{4\pi\gamma(\triangle x)^2} \cdot 2\epsilon$$

in the estimate of $B_n^{i+1/2}$, since a difference equation of the form

$$B_n^{i+1/2} = B_n^{i-1/2} + \frac{c^2\triangle t}{4\pi\gamma(\triangle x)^2}\triangle^2 H_n^i$$

is used. When the value of B is extrapolated to the point B_n^{i+1} for the next cycle the error becomes

$$\frac{c^2\triangle t}{4\pi\gamma(\triangle x)^2}\cdot 3\epsilon\ .$$

Using this to calculate the new H of the next cycle, it is found, on expanding in a Taylor's series and keeping only two terms,

$$H_n^{i+1}(\quad) = H_n^{i+1}(B_n^{i+1}) + \frac{c^2\triangle t}{4\pi\gamma(\triangle x)^2}\cdot 3\epsilon\frac{dH}{dB}\ .$$

If the original error ϵ is to produce an effect in the next cycle at the same point that is not greater than itself, then the following condition must be met,

$$\triangle t \leqq \frac{4\pi\gamma(\triangle x)^2}{3c^2}\frac{dB}{dH}\ .$$

This condition differs from that of hyperbolic equations in that it depends on the square of $\triangle x$. Thus, if the spacing in $\triangle x$ is halved, the $\triangle t$ must be divided by 4. This is typical of parabolic equations. The inequality takes care of a single roundoff, while if a roundoff at each point is assumed, an additional factor of 7/10 is needed on the right-hand side.

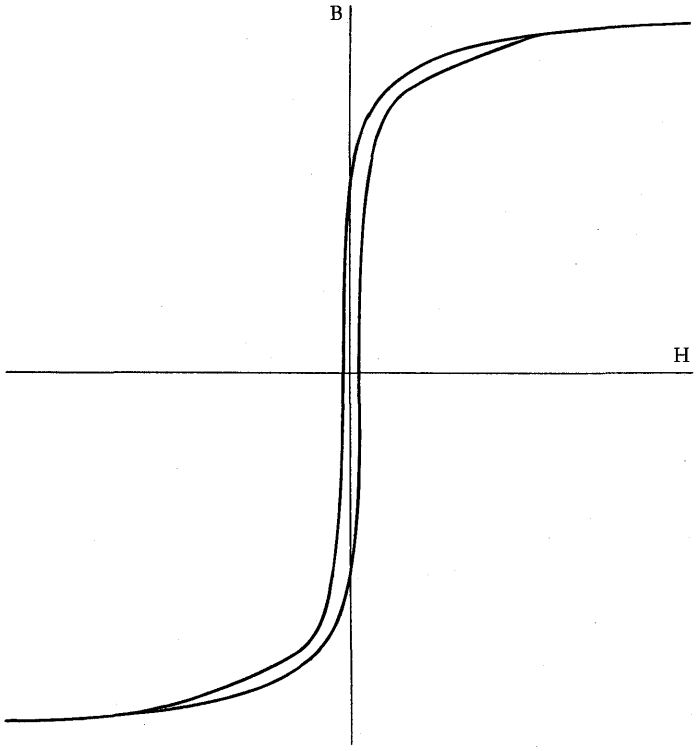


FIGURE 3. HYSTERESIS LOOP, MOLYBDENUM PERMALLOY

This condition is clearly necessary in order to expect anything like a reasonable solution; its sufficiency will be discussed later. Note the derivative dB/dH on the right-hand side.

The particular sample of permalloy discussed had a B-H curve, as shown in Figure 3. Recalling the importance of the derivative dB/dH in the net spacing condition, it is seen that as the problem progresses a very tight spacing must be accepted throughout the problem or else the spacing at various stages must be modified to take advantage of those places where the derivative is large. The latter was chosen.

In the early stages of the computation, while an attempt was made to obtain an idea of the order of magnitude of the frequency f, a crude subdivision of the slab into four zones marked by five points was made. By symmetry one half could be ignored so that, in fact, only three points needed to be followed. The outer point was driven by the boundary condition, while the two inner points followed according to the difference equations.

To test the method, first a B-H curve was used which was a straight line. The comparison with the analytical solution was excellent. To show the reality of the net spacing condition the problem was deliberately allowed to run past a point where the net should have been tightened. The results are shown in Figure 4. This oscillation is typical of what happens when a net spacing condition is violated, although sometimes it takes the form of a sudden exponential growth instead. Indeed, when such phenomena occur, one may look for a violation of some net spacing condition.

I have emphasized that the condition just derived is a necessary condition. There has been a lot of discussion lately as to whether this is a sufficient condition. Unfortunately, I do not have time here to go into this matter more fully. Instead, let me present some of the results obtained several years ago when we did this problem. Figures 5, 6, and 7 show a smooth transition on the solution as the frequency f was changed. Any errors are clearly systematic. The jumps in the inner points are due both to the shape of the B-H curve and the extremely coarse spacing of three points. When a finer spacing of five points was used (eight sections of the slab instead of four), much the same picture was found. The labor, of course, was eight times as much since there were twice as many points, and the $\triangle t$ was decreased by a factor of four. This crude spacing should indicate how much valuable information may be obtained from even the simplest calculations when coupled with a little imagination and insight into the computation.

There seems to me to be no great difficulty in setting up such a problem for machine computation; so I shall not go further except to note that in the original setup of the problem we provided for the fact that we would have to

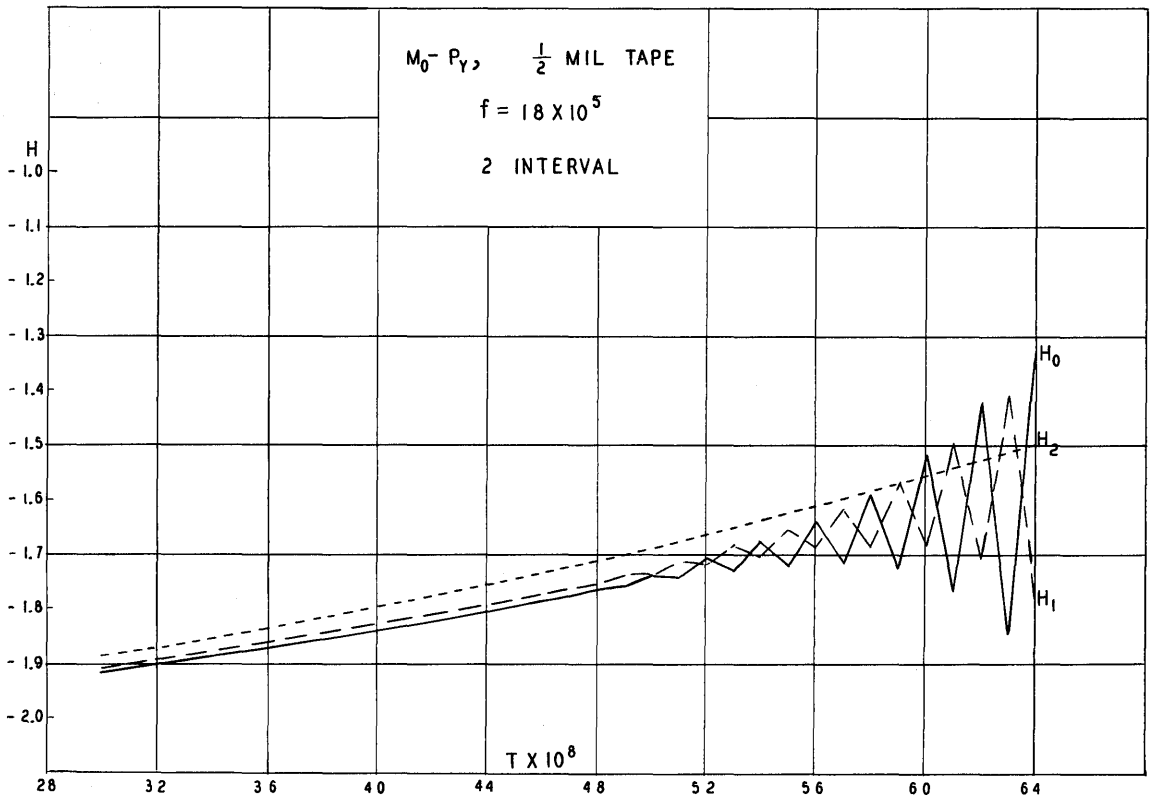


FIGURE 4

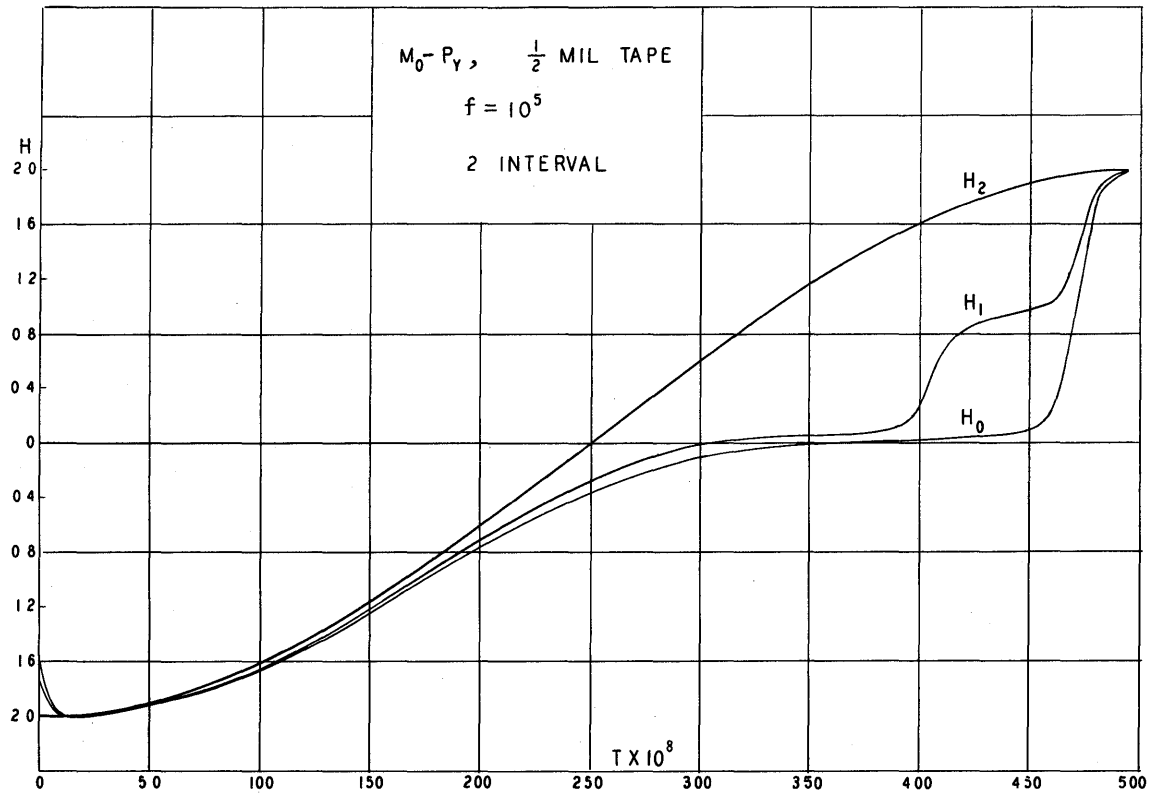


FIGURE 5

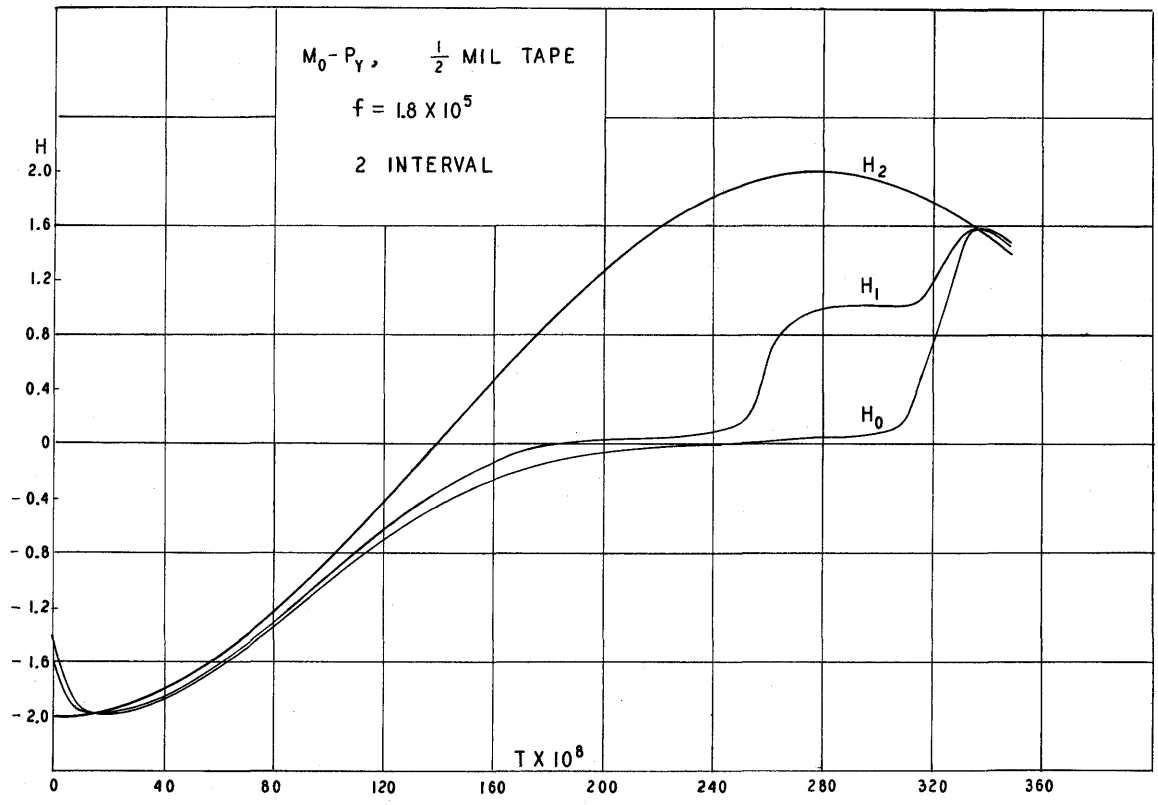


FIGURE 6

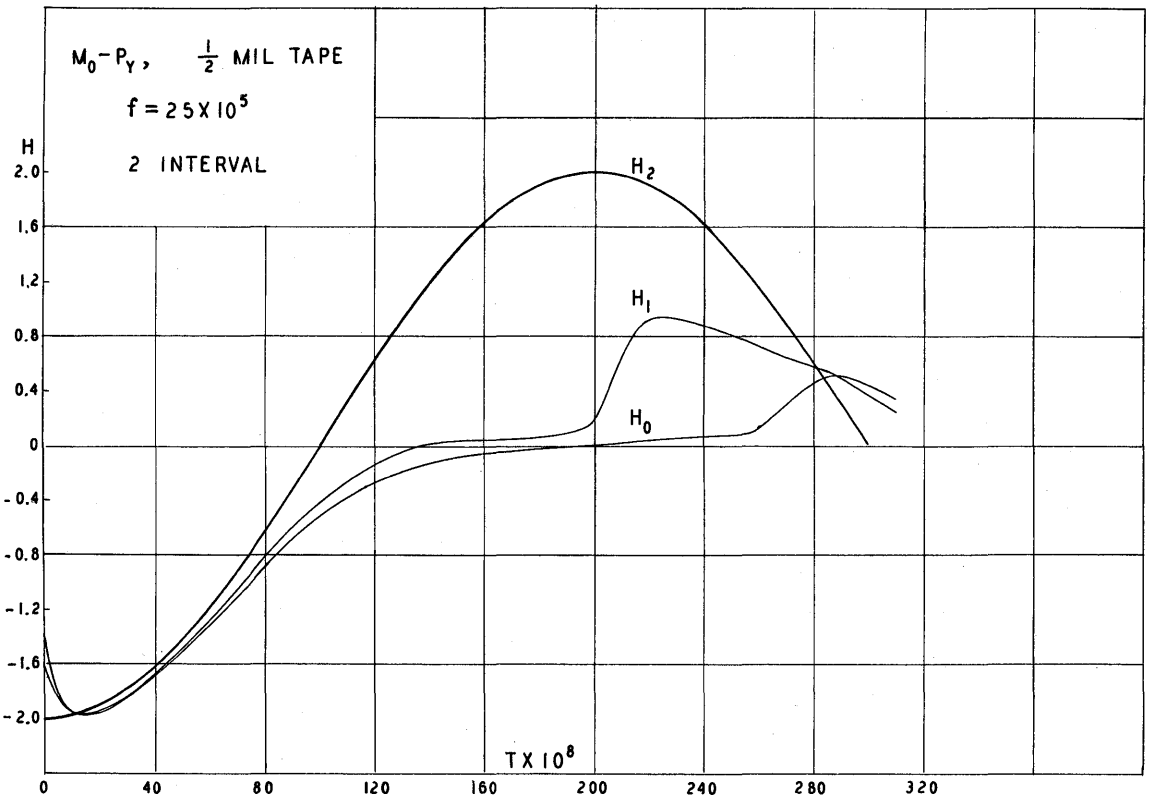


FIGURE 7

consult not one but a family of *B-H* curves, the one chosen for each point depending on its maximum saturation. This refinement was not included in the results shown before, and in any case it produces only a small effect.

The Traveling Wave Tube

The last example I wish to consider is that of the traveling wave tube. A traveling wave tube consists of a helix of wire wound around an evacuated cylinder. The pitch of the helix reduces the effective velocity of the field due to an impressed electric current in the helix to around 1/10 that of light. Inside the helix is a beam of electrons going at a velocity slightly greater than the electromagnetic wave from the helix. As in the two-beam tube, the stream of electrons interacts with the field and gives up some of its energy to magnify the signal impressed on the helix.

The equations describing one particular idealization of such a tube are

$$\frac{dA(y)}{dy} = -\frac{1}{2\pi\epsilon}\int_0^{2\pi} \sin\phi(\theta, y)\,d\theta$$

$$\eta(y) = -\frac{1}{2\pi\epsilon A(y)}\int_0^{2\pi} \cos\phi(\theta, y)\,d\theta$$

$$\frac{\partial}{\partial y} q(\theta, y) = A(y)\sin\phi(\theta, y)$$

$$\frac{\partial}{\partial y}\phi(\theta, y) = k + \eta(y) + 2\epsilon q(\theta, y)\ .$$

These equations have already been transformed over to a coordinate system moving with the wave. The y coordinate measures, in a sense, the length down the tube, while the θ measures the time.

If the equations are examined more closely, it is seen that, for each θ, the lower two equations must be solved in order to move the solution of q and ϕ one step down the tube in the y direction. The sine and cosine of ϕ are then summed to obtain numbers depending on the fundamental frequency. The higher harmonics are clearly being neglected. These upper equations in turn supply the coefficients for the lower equations. This neglect of the higher harmonics was justified on the physical grounds that the helix damped them out very strongly. As a check, the amount of the second, third, fourth, and fifth harmonics in the beam was calculated later, and it was found that they could indeed be neglected.

The first step is to make a transformation so that the parameters ϵ and k drop out of the equations.

Proceeding much as in the two-beam tube, it was decided that 16 points would provide an adequate picture in the θ direction. Thus, there are 16 pairs of equations like the lower ones to solve. In addition, it was desired to solve eight such problems for different parameter values (which appear in the initial conditions and enter in the "linear" solution used to start the nonlinear problems). This gives 128 pairs of equations to be solved at each cycle, a situation very suitable for IBM equipment. On the other hand, the upper two equations only occur once per problem, or eight times in all, which makes them unsuitable for machine solution. Thus, the solution of the lower equations and calculation of the sums corresponding to the integrals of the upper equations were accomplished on IBM machines, while the rest of the upper equations were solved by hand calculations. Included in the hand calculations was a running check that may be found by integrating the equations over a period and finding the ordinary equations that govern the mean values.

With a spacing chosen in one variable, how is the spacing to be chosen in the other? In this case, there is no theory of characteristics, in fact, very little mathematical theory available at all. The obvious was done. A number of spacings were tried, with a crude net space in $\triangle\theta$, and the maximum permissible $\triangle y$, at that stage of the problem, was determined experimentally. Then a spacing $\triangle y$ was chosen, comfortably below this limit, although not so far as to make too much work, and the calculation started with the hope that this would either be adequate for the entire problem or that the effect would show up in a noticeable manner in the computations. No obvious anomalies appeared; so presumably the spacing of 1/10 unit in y was adequate. The net chosen was rectangular with every other y value being used to calculate the ϕ and q, while at the other values the A and η were evaluated. This produces central difference formulas which are the most accurate. The old cycle in ϕ^- and q^- was labeled by a $-$, the current values of A° and η° by a $^\circ$, and the predicted values of ϕ^+ and q^+ by a $+$. The new values of A^{++} and η^{++} were labeled $++$.

To set up a system of difference equations corresponding to the lower equations: first, an estimate of q^+ was obtained, then this was used to find a reliable value of ϕ^+; finally, using this ϕ^+, a better estimate of q^+ was obtained. The difference equations describing this situation are

$$\phi^+ = \phi^- + \frac{1}{10}\left(\eta^\circ + q^- + \frac{A^\circ}{10}\sin\phi^-\right) \qquad L = \Sigma \sin\phi^+$$

$$q^+ = q^- + \frac{A^\circ}{10}\left(\sin\phi^- + \sin\phi^+\right) \qquad M = \Sigma\, q^+\sin\phi^+$$

$$N = \Sigma\cos\phi^+\ .$$

The difference equations corresponding to the upper equations have not been shown, but it has been indicated that the solution of both equations was made to depend on three sums labeled L, M, and N.

To simplify matters in finding the sine and cosine of ϕ, the units of measurement of angle were changed from radians to 1/1000 part of a circle. The trigonometric functions for such angles can be found by consulting tables in decigrades

at every fourth decigrade. The advantage of such a unit is that the integral part of the angle automatically gives the number of rotations, and the fractional part gives the value at which to enter the table.

Consider the basic cycle of computation. It is obvious that the accounting machine-summary punch will be best adapted to the summing of the quantities leading to L, M, and N. This is the natural point to start a cycle, since the cards from the summary punch will have the minimum amount of information, leaving the rest of the space on the cards for future calculations. These cards will be called detail cards.

Each detail card needs to be identified uniquely. To do this the problem number, the card number which is its θ value, and the cycle number which is its y value, are given. The information that the detail cards must carry at this stage to describe the problem is the current values of ϕ^- and q^-. In addition, it is convenient to have the value of the sine of the old angle, $\sin \phi^-$.

The master, or rate, cards—the information for which comes from the hand calculations—must have identification consisting of the problem number and the cycle number, and the values of the two dependent variables A° and η°. The procedure is:

1. Key punch the eight master cards and sort them into their appropriate places, a matter of one sort on one column.

2. Multiply with crossfooting to obtain the quantity,

$$q^{+(\text{estimate})} \sim \frac{A^\circ}{10} \sin \phi^- + q^- \ ,$$

which is an estimate of the q at the new cycle.

3. Another multiplier-crossfoot operation produces

$$\phi^+ = \phi^- + \frac{\eta^\circ}{10} + \frac{A^\circ}{100} \sin \phi^- + \frac{q^-}{10}$$

which is the new value of ϕ^+. Now the sine and cosine of ϕ must be found.

4. Sort on ϕ for three digits,

5. Collate in the table values of the trigonometric functions,

6. and 7. Using the multiplier, linearly interpolate the values of sine and cosine of ϕ. Each may be obtained with a single pass through the multiplier, provided there are only five figures in the table values and three in their first differences. The algebraic signs may be picked up from the master cards and held up for the detail cards which follow, so that with a suitable complement punching circuit the value and its algebraic sign may be punched.

8. Collate again to remove the table cards and at the same time put the table back in proper order. (Incidentally, the same control panel is used for both operations on the collator.)

9. Resort the detail cards so that they are again in order, both as to the card number and the problem number, a matter of a three-column sort.

10. Multiply-crossfoot to obtain the new value of q from the formula

$$q^+ = q^- + \frac{A^\circ}{10} (\sin \phi^- + \sin \phi^+) \ .$$

11. Multiply to obtain $q^+ \sin \phi^+$.

12. List the calculated values, the three sums L, M, and N, and summary punch the cards for the next cycle.

If these operations are gathered together, it is found that there are 6 passes through a 601 type multiplier, three sorts for a total of 7 columns, two passes through the collator, a key punching of 8 cards, and one pass through an accounting machine-summary punch for each cycle.

We used our own accounting department machines with a 601 multiplier modified to have sign control, and three selectors. We operated only at times when they were not doing their main task of getting out the pay checks! Needless to say, no arguments ever arose as to priority on the use of the machines.

Conclusion

Let me summarize the points I hope I have made. First and foremost, there is a large class of problems where the relative size of the net spacing chosen is of fundamental importance. Where there is no known mathematical theory, or where one is ignorant of it, one may still proceed on an experimental basis and watch for either violent oscillations or sudden exponential growths to indicate where the going is not safe.

Second, it is not hard to set up a method of computation for a given problem, and one can estimate the accuracy at any step by some such device as using a Taylor's expansion with a remainder. The harder problem of propagation and compounding of errors I have not answered at all definitely, but have suggested that prudence, physical intuition, and faith will provide one with a suitable guide.

Lastly, it is not hard to work out the details of a basic cycle if one keeps in mind the amount of information that must be available at certain stages, watches the flow of information, and has the courage to try to work out the details of a plan. When it comes to comparing alternate methods I presume that one can count operations, judge reliabilities, etc., of the various alternates. There may be better ways than you have thought of, but don't let that stop you! If the method is sound, economically worth while, then you are justified in going ahead. You don't need super computing machines, although they are nice to have; you can go ahead with what you have at the moment and obtain useful and valuable results.

DISCUSSION

Dr. Hammer: The choosing of networks in comparison to the interval of the variables sometimes can be avoided by using a different system of integration; that is, an implicit calculation in which perhaps some of the variables are first found by explicit integration, and then they are recalculated.

Dr. Hamming: You are thinking of the Riemann method, no doubt, or the von Neumann method of getting a difference equation which involves the present values and simply adjacent values one cycle forward.

Dr. Hammer: Yes. One essentially calculates all the values at the same time, and then the condition you mentioned can be violated to some extent.

Dr. Herget: The graphical way in which you portray the effect of $(\triangle t)^2$ to $\triangle x$ is very good, and I think it has been stated in some of these meetings before that to be safe for the convergence involved in this process, $\triangle t$ should be about half of $\triangle x$. Isn't that right?

Dr. Hamming: You can't say any $\triangle x$ and $\triangle t$. It depends on the scale of the variables used. If the variable is multiplied by 10, the spacing would be changed numerically. The condition is stated in terms of the velocity of propagation of signal in the hyperbolic case, and in the parabolic case one considers the derivative dB/dH.

Dr. Grosch: I would like to ask a question about the nature of the oscillations encountered when the condition is violated. Have you made any effort to see, if you will pardon a nontechnical term, what the mean curve through those oscillations does? Does it follow the solution?

Dr. Hamming: Yes, it does.

Dr. Grosch: That is an interesting point.

Dr. Hamming: If you examine Figure 4, you can see this is true.

Dr. Grosch: We had a situation like that arise back in 1946 when we were using 601's, and in our case the condition on the very short $\triangle x$ and $\triangle t$ was not a simple constant but a sort of variable of the column. We had this oscillation happen just a few x intervals from the end; we tried a fudging method of this sort, and it seemed to work out all right.

Dr. Alt: I think the situation concerning propagation of the local errors is not as hopeless as you indicate. One can use Green's function in order to study the propagation of errors whenever Green's function is available. If it is not in there, at least one can try to linearize it. We have tried that in a nonlinear problem, too, and it worked out. It can be done. I felt that the problem was simple enough that one didn't even consider publishing it, because it was just a straightforward application of the Green's function.

Dr. Hamming: I agree with what you say completely if your problem is either linear or your solution is reasonably close to linear with a perturbation, but when you encounter essentially nonlinear effects, Green's function will tie you up hopelessly when it is essentially the nonlinear part you want. That was the problem in all the examples that I showed; not to get the linear problem with the perturbation, but to get the essential nonlinear effects and see where they entered, how much they entered, and where they cut off.

Numerical Solution of Partial Differential Equations

EVERETT C. YOWELL

National Bureau of Standards

THE USUAL METHODS for determining numerical solutions of partial differential equations with specified boundary conditions are based on the approximation of the differential equation by a difference equation. In the case of the two-dimensional Laplace's equation, which is the only one I will speak of today, the differential equation is $\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = 0$. The standard approximation is $\frac{\Delta_x^2 f}{(\Delta x)^2} + \frac{\Delta_y^2 f}{(\Delta y)^2} = 0$ where $\Delta_x^2 f$ is the second difference in the x direction and $\Delta_y^2 f$ is the second difference in the y direction.

The exact relation between the differential operator and the difference operator is an infinite series in the difference operator,

$$\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = \frac{1}{(\Delta x)^2}\left(\Delta_x^2 f - \frac{1}{12}\Delta_x^4 f + \frac{1}{90}\Delta_x^6 f - \dots\right) + \frac{1}{(\Delta y)^2}\left(\Delta_y^2 f - \frac{1}{12}\Delta_y^4 f + \frac{1}{90}\Delta_y^6 f - \dots\right)$$

and the standard approximation amounts to cutting off this series after the first term in x and in y. If only second differences are to be used, it is obviously advisable that the interval Δx should be chosen sufficiently small so that the term $\frac{1}{12}\frac{\Delta_x^4 f}{(\Delta x)^2}$ becomes negligible. If this leads to too small an interval, we try to recover the accuracy by including more terms of the series in the approximation. The validity of this procedure needs rigorous justification, but it presents a practical computational approach.

Now differences are linear relations between function values at adjacent points. Hence, any method which works basically on the difference equation will be a method dealing with values of the function at specified points within the boundaries. These points are generally chosen systematically to cover the interior of the region with a regular grid. We shall consider only a square grid, as it is more easily adapted to machine computations than are the triangular or hexagonal grids.

The direct approach to the problem is to write out the difference equation for each point of the grid. Since we are dealing with a square grid, $\Delta x = \Delta y$, we shall call this mesh distance h. Then the difference equation at the point $x = x_i$, $y = y_j$ will become:

$$\frac{1}{h^2}[f(x_{i-1},y_j) - 2f(x_i,y_j) + f(x_{i+1},y_j) + f(x_i,y_{j-1}) - 2f(x_i,y_j) + f(x_i,y_{j+1})] = 0$$

or symbolically

$$\frac{1}{h^2}\begin{bmatrix} & +1 & \\ +1 & -4 & +1 \\ & +1 & \end{bmatrix} = 0 .$$

There is one such equation for each point of the grid. Each equation may involve only interior points, or interior points and boundary points. If the boundary points are considered as known and transferred to the right sides of the equations in which they occur, then there is a system of n equations, one for each grid point, involving n unknowns, the values of the function at each grid point, which completely defines the function at the grid points within the boundary. In the present case, it can easily be shown that a unique solution of these equations always exists.

There is one great drawback to this approach to the numerical solution of a differential equation, and that is the number of equations in the system. Consider a relatively simple heat conduction problem. We have a cube, 10 cm. on each edge, at a uniform temperature of 0°C. We place this cube in contact with a heat source along one face. The temperature of the source is some function of both coordinates. We insulate one of the adjacent faces and then inquire as to the distribution of temperature over the free faces ten seconds after contact is made. This problem will reduce to a four-dimensional case, three space dimensions and one time dimension. If a ten-point grid is introduced in each dimension, a system of 10,000 equations in 10,000 unknowns results. And, while a large number of coefficients will be zero, this is not a problem to be approached with equanimity. Although this example

was designed to show how rapidly the number of equations can increase, and is not the type of problem that would be solved in practice, problems of the same order of magnitude are available in the physically interesting problems whose solutions are being sought today.

A second method of attack is the relaxation method of Sir Richard Southwell. Here one guesses at the value of the function at each grid point and then systematically improves the guess. The values of the function are substituted into Laplace's difference equation, and the result will in general differ from zero. This difference, or residual, is computed for each grid point. The largest residual in the entire field is now located, and the value of the function at that point altered in such a way that the residual becomes zero. This is equivalent to adding one quarter of the residual to the residual at each of the four adjacent points, leaving the rest of the field unaffected. The field is again scanned for the largest residual, and it is reduced to zero by changing the value of the function at that point. The process is continued until all the residuals become small, one or two units in the last place.

As a hand computing method, relaxation has many advantages. It deals with only a few points at a time, it involves very simple operations, and it converges to the solution of the difference equation rather rapidly. And there are variations—over-relaxing and under-relaxing, group relaxing and block relaxing—which increase the speed of convergence. As a machine method, many of these advantages are lost. The speed of convergence depends on relaxing the largest residual at each step. Hence, the entire residual field must be scanned before each operation to locate this largest residual. This scanning for size is still a very inefficient operation, particularly when it is interposed between every set of five additions. Then, too, the block and group relaxations, which so speed up the convergence in hand computing, are very difficult to apply using automatic computing machinery.

Another method related to the relaxation method is Liebmann's smoothing method. Once again, we start with the basic difference formula

$$\frac{1}{h^2}[f(x_{i-1},y_j) + f(x_{i+1},y_j) + f(x_i,y_{j-1}) + f(x_i,y_{j+1}) - 4f(x_i,y_j)] = 0.$$

If now we multiply the equation by h^2 and then transfer $4f(x_i,y_j)$ to the right-hand side, we have an equation defining $f(x_i,y_j)$ in terms of the four adjacent values of the function. The method consists in guessing the value of the function at each grid point, and then applying the smoothing formula to each point of the grid. The entire field is smoothed again and again until no changes are introduced in the function values to the degree of accuracy required.

This method has some advantages and some disadvantages. Its main disadvantage is its slow speed of convergence. Its advantages are that it deals with only a few points at a time, that it involves only simple operations, and it is adaptable to machine computations. M. Karmes, of New York City, has done this, reporting in 1943 on an adaptation of this method to 601 multipliers. His machine method is straightforward, one quarter of the value of the function at the four neighboring points being summed to give the value at the central point. To assemble correctly the cards to be summed, Karmes prepares four decks of work cards. Each of these contain $\frac{1}{4}f(x_i,y_j)$, but they differ in that a second argument is introduced in each deck. One contains $(i-1,j)$; one, $(i+1,j)$; one, $(i,j-1)$; and one, $(i,j+1)$. The four decks are now sorted together on the second argument and summed, summary punching the new value of the function at each grid point. The deck with the new function values is then reproduced four times, the second arguments are put in, and the process is repeated. This cycle is continued until the function values converge within the required accuracy.

A method similar to Liebmann's method, but better adapted to machine computation, has been devised by Dr. Milne and tested on the 604 electronic calculators at the Institute for Numerical Analysis. Dr. Milne was seeking to avoid the sorting problem that led Karmes to the use of four decks of cards. He added two difference operators, each satisfying Laplace's difference equation, together. The first is the usual

$$\frac{1}{h^2}\begin{bmatrix} & 1 & \\ 1 & -4 & 1 \\ & 1 & \end{bmatrix} = 0,$$

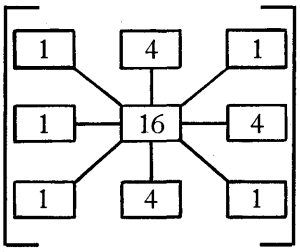
while the second is basically the same operators rotated 45°,

$$\frac{1}{h^2}\begin{bmatrix} \frac{1}{4} & & \frac{1}{4} \\ & -1 & \\ \frac{1}{4} & & \frac{1}{4} \end{bmatrix} = 0$$

Multiplying each by four and summing, we obtain

$$\frac{1}{h^2}\begin{bmatrix} 1 & 4 & 1 \\ 4 & -20 & 4 \\ 1 & 4 & 1 \end{bmatrix} = 0.$$

The trick now is to multiply through by h^2 and then add $36f(x_i,y_j)$ to both sides of the equation. This gives



$$= 36f(x_i,y_j).$$

This equation is now factorable, and we can define two operators U and V such that

$$U f(x_i,y_j) = \frac{1}{6}[f(x_{i-1},y_j) + 4f(x_i,y_j) + f(x_{i+1},y_j)]$$

$$V f(x_i,y_j) = \frac{1}{6}[f(x_i,y_{j-1}) + 4f(x_i,y_j) + f(x_i,y_{j+1})].$$

If these are applied successively to the n^{th} approximation to the function values at the grid points, they will yield the $n + 1^{st}$ approximation. Or,

$$f^{n+1}(x_i,y_j) = UV f^n(x_i,y_j) = VU f^n(x_i,y_j).$$

The last equation indicates that the operators commute and that rows or columns may be smoothed first.

This method works nicely on the 604 electronic calculating punch. For each iteration, the cards must be fed through the machine twice, once in row sort and once in column sort. At the end of the second run, a new set of function values will have been computed for each grid point.

The example tested a 9 x 10 rectangular grid with values of arctan x/y given along the boundaries. We have available 10 place values of arctan x/y, so that a check was possible on the speed of convergence. The smoothing was applied first by rows and then by columns, although this choice was completely arbitrary.

The wiring of the 604 control panel was simple and straightforward. The value of the function was read into factor storage 3 and 4, and a 4 and a 6 were emitted into the MQ and factor storage 2, respectively, on the read cycle. The analysis chart reads as shown below:

		Factor Storage				*MQ*	*Counter*	*General Storage*			
Step	*Operation*	1	2	3	4			1	2	3	4
Read			6	$f(x_i,y_j)$		4					
1.	add $f(x_i,y_j)$			RO			RI				
2.	add $f(x_i,y_{j-2})$						RI			RO	
3.	Mult $4f(x_i,y_{j-1})$							RO			
4.	Divide sum by 6 (Expanded division if more than five digits are used)	RO									
5.	Transfer $f(x_i,y_{j-1})$							RO		RI	
6.	Transfer $f(x_i,y_j)$			RO				RI			

The result of this operation is to punch $V f(x_i,y_{j-1})$ on the (i,j) card. The two last transfers set up the operation for the next point. This arrangement of storage units will handle any size numbers up to eight digits, and that should include all problems of practical interest today. There is no question of the function values growing too large, as the maximum and minimum values must occur on the boundary.

The wiring of the 521 control panel is a little more complicated, as it was desirable to make the control panel automatically change itself for the differences between the first and second runs. There are two problems that must be handled on the 521 control panel. The first is that the input field for the second run is the same as the output field of the first run. And the second is the shift in argument.

The card layout is as follows. In column 1, punch the row identification i. In column 2, punch the column identification j. In columns 3-8, punch the original value of the function. After smoothing along a row, punch the answer in columns 9-14; and after the next smoothing along a column, punch the answer in columns 15-20.

The first problem, then, is to read from columns 3-8 on the first run and punch into columns 9-14. On the second run, read from columns 9-14 and punch into columns 15-20. This is done through punch selectors, the normal side being wired to the first run read and punch fields, and the transferred side being wired to the second run read and punch fields. The selectors are transferred by a y in column 80, a punch which is introduced on the first run by wiring from the emitter through the normal side of a punch selector to the punch magnet for column 80. The selectors which switch the read fields should be controlled through their immediate pickup, while the selectors that control the punch field should be controlled through their delayed pickup.

The shift in argument is easily handled. On the first run, the j identification is gang punched backward into the following card, while on the second run, the i identification is similarly gang punched back one card. Column 2 is wired from the second reading brushes through a normal

point of a punch selector to the punch magnet of column 22. Column 1 is wired from the second reading brushes through a transferred point of the same punch selector to the punch magnet of column 21. This selector is then controlled through its delayed pickup by the y in column 80.

Two points might be examined in a little greater detail. The first of these has to do with the smoothing of the boundary values during the first run. As the cards are going through in row sort, the first and last rows will be entirely boundary cards, and the values punched into these cards will be the smoothed boundary values rather than the true boundary values. This is necessary for the correct application of the formula, as a consideration of the function at the point (1,1) will show. Suppose first that we do not smooth the first row. Then we will have available at the end of the first run these values on their corresponding cards:

i	j	function
0	1	$f(0,1)$
1	1	$\frac{1}{6}[f(1,0) + 4f(1,1) + f(1,2)]$
2	1	$\frac{1}{6}[f(2,0) + 4f(2,1) + f(2,2)]$.

At the end of the second run, the answer punched in the card for the point (1,1) will be

$$\frac{1}{36}[6f(0,1) + 4f(1,0) + 16f(1,1) + 4f(1,2) + f(2,0) + 4f(2,1) + f(2,2)]$$

which is equivalent to the true expression only if $f(i,j)$ is linear along the boundary $i = 0$.

If, on the other hand, we smooth the first row, we will have available at the end of the first run these values on their corresponding cards:

i	j	function
0	1	$\frac{1}{6}[f(0,0) + 4f(0,1) + f(0,2)]$
1	1	$\frac{1}{6}[f(1,0) + 4f(1,1) + f(1,2)]$
2	1	$\frac{1}{6}[f(2,0) + 4f(2,1) + f(2,2)]$.

At the end of the second run, the answer punched on the card for the point (1,1) will be the correct expression

$$\frac{1}{36}[f(0,0) + 4f(0,1) + f(0,2) + 4f(1,0) + 16f(1,1) + 4f(1,2) + f(2,0) + 4f(2,1) + f(2,2)].$$

Thus the use of a smoothed boundary value in the second run actually is necessary for the successful evaluation of the smoothing formula at all points of the grid.

The second point is the use of a single delay in transferring the selector, which governs the gang punching of the i identification on the second run. Standard practice for gang punching through a selector is to use a double delay so that the card containing the pickup punch will be passing the second reading station when the selector transfers. In this case, all cards have the pickup punch. Thus, use of a double delay would transfer the selector from the time the first card is passing the second reading station until the last card is passing the reading station. Use of a single delay will transfer the selector from the time the first card is passing the punching station until the time the last card is passing the punching station. Either type of delay will give the correct gang punching result in this case; so a single delay was used as a simpler method.

The complete sequence of operations now can be summarized. A deck of cards containing the boundary values is reproduced a large number of times. A deck of cards containing the initial values of the function at the interior points is key punched. This deck and one of the boundary decks are then sorted on columns 2 and 1. This puts the cards in order of column number within rows. The cards are then run through the 604 and again sorted on column 22. This orders the cards on rows within columns. The first and last columns are removed, as they contain spurious smoothed values. The remaining cards are again run through the 604 and then sorted on column 21. The first and last rows are removed, as they again contain spurious values. The remaining cards are reproduced, reproducing columns 21 and 22 into columns 1 and 2, and columns 15-20 into columns 3-8. These new cards form the deck for the interior points in the next approximation. They are combined with a new boundary deck, and the process is repeated.

For the example we tested, one cycle of operations on ninety cards took about five minutes. More time must be allowed for reproducing new boundary decks, but certainly ten steps an hour can be accomplished if a 604, reproducer, and sorter are set aside for the problem. And then an occasional check must be made of the convergence of the solution. We listed the ninety cards after every tenth iteration and examined two successive lists for changes. In about sixty iterations, we had reached an accuracy of about two units in the fourth place.

This same example was tested in our hand computing section, using a mixture of smoothing and block relaxing. The field was smoothed three times, then block relaxed. This cycle was repeated three times, and then three additional smoothings were made. At the end of these twelve smoothings and three block relaxings, answers were reached that were closer to the true answers than had been reached in the 60-odd iterations by punched card machine.

The great difference in the speed of convergence is due to the use of block relaxing. An intuitive idea of the reason for this is gained by considering the basic action of the smoothing operator. Now Dr. Milne's smoothing operator will work just as well on the residuals as on the functional values. The residuals are defined with respect to

this operator in a similar fashion to the way residuals were defined with respect to Liebmann's operator. Now, consider the original residual field and the effect of the smoothing operator on it. If the residual were plotted vertically against the x and y coordinates of the points and a surface passed through the ends of the residuals, a three-dimensional model similar to a mountain would be obtained. As the original guesses were not good, plus and minus errors would be found, large and small errors, and the mountain would be rough-covered with peaks and valleys. A few applications of the smoothing operator will level off the peaks and fill in the valleys, producing a smooth instead of a rough mountain. The outstanding deviation from smoothness will come at the boundary, where the elevation of the mountain goes to zero. And beyond the boundary, a flat, level plane stretches to infinity in all directions. The task of the smoothing operator is to erase this lack of smoothness at the boundary by forcing the entire mountain out through the boundary. And as the altitude of the mountain decreases, the slope at the boundary approaches closer and closer to zero, and less and less of the residual is removed with each iteration. Hence, the convergence is rather poor, because the operator is most efficient at smoothing and inefficient at forcing residuals through the boundary.

This situation is completely upset when block relaxing is added as a further tool. Now one smoothes for a while until a smooth mountain is formed. Then one traces a few approximate contour lines along the mountain. The area between any two contour lines is then dropped in altitude by the mean altitude of the two adjacent contour lines. This then removed the bulk of the mountain, leaving small peaks and ditches. These are rapidly smoothed over by use of the smoothing operator, and again the bulk of the mountain is carted off by the block relaxation.

While the use of block relaxing together with smoothing provides a rapidly convergent way of solving Laplace's equation, it is at present not set up for machine computation. The need for drawing contour lines and the interdependence of neighboring points makes it very difficult to set up for automatic calculation on present day calculators. There are undoubtedly ways of accomplishing the same thing without using block relaxation in its standard manner, but these must be found by further investigations and offer problems which I hope Dr. Milne will investigate during his next stay at the Institute for Numerical Analysis.

DISCUSSION

[This paper and the following one by Dr. Harry H. Hummel were discussed as a unit.]

An Eigenvalue Problem of the Laplace Operator

HARRY H. HUMMEL

Argonne National Laboratory

IN A PAPER presented at the November, 1949, meeting at Endicott, Flanders and Shortley[1] discussed the solution of the equation

$$\nabla^2 \psi = \alpha \psi . \qquad (1)$$

Here α is an eigenvalue, and the problem is to find its highest value and the corresponding fundamental eigenfunction ψ for homogeneous boundary conditions. This paper will describe the solution of this problem on IBM machines for a two-dimensional region consisting of a square with a square hole cut out of it (Figure 1). The function is set equal to zero at the outer and inner boundaries of the region.

The solution of (1) is accomplished by transforming to a difference equation over a two-dimensional network of points in the usual way (Figure 1). Set $\psi_{x+1,y} + \psi_{x-1,y} + \psi_{x,y+1} + \psi_{x,y-1} \equiv \Sigma\, \psi_{x,y}$, where $\psi_{x,y}$ is the value of the function at (x,y). Then the difference form of (1) is

$$\frac{(\Sigma - 4)\, \psi_{x,y}}{h^2} = \bar{\alpha}\, \psi_{x,y} . \qquad (2)$$

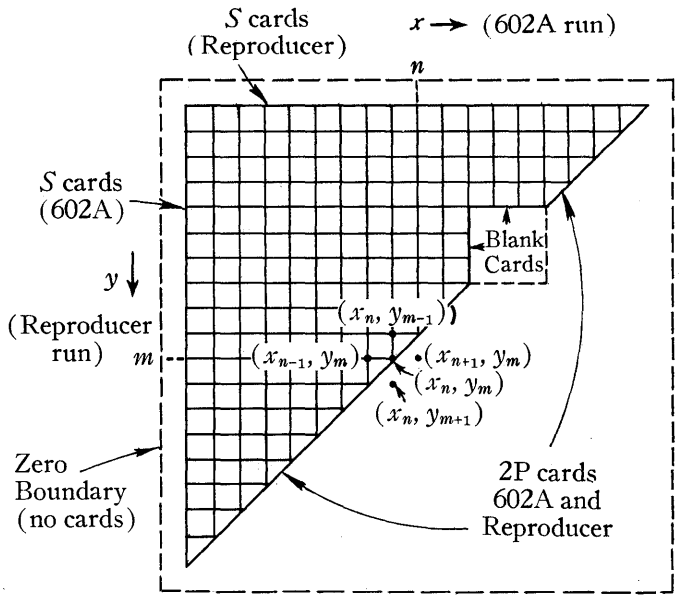


FIGURE 1. LAYOUT OF CARDS

Here h is $(x_{n+1} - x_n) = (y_{m+1} - y_m)$, the net spacing in the difference problem. $\bar{\alpha}$ is the eigenvalue of the difference problem, assumed equal to α.

By defining

$$\omega \psi \equiv \frac{1}{4} \Sigma\, \psi_{x,y}$$

and

$$\lambda = 1 + \frac{\bar{\alpha} h^2}{4}$$

equation (2) becomes

$$\omega \psi_{x,y} = \lambda \psi_{x,y} . \qquad (3)$$

Since the highest value of $\bar{\alpha}$ is desired, the highest value of λ is also desired. The number of homogeneous equations (3) is equal to the number N of points (x,y) in the network, which is, therefore, the number of eigenvectors of the equations (3). For this algebraic problem it is known[1] that $-1 < \lambda < 1$, and also that the set of eigenvectors is complete. A solution ψ_n will, of course, consist of N numbers $\psi_{x,y}$, one for each point of the network, and will correspond to an eigenvalue λ_n.

Then a first approximate solution function ψ can be analyzed in terms of the eigenvectors ψ_n

$$\psi = \sum_{n=1}^{N} C_n \psi_n . \qquad (4)$$

Operating with $(\omega - a)/(1 - a)$, where a is a real number such that $-1 \leqq a < 1$,

$$\frac{\omega - a}{1 - a} \psi = \sum_{n=1}^{N} C_n \left(\frac{\lambda_n - a}{1 - a} \right) \psi_n . \qquad (5)$$

Thus, it is seen that, by performing this operation a number of times with various values of a, the amplitudes of higher eigenfunctions may be reduced as much as desired relative to that of the fundamental mode ψ_1, the eigenvalue λ_1, which is usually nearly 1. The value of $(\lambda - a)/(1 - a)$ as a function of λ and a is shown in Figure 2. Flanders and Shortley[1] discuss the selection of values of a; concluding that the greatest efficiency is achieved by choosing the a values as the roots of the Tschebyscheff Polynomial of order equal to the number of iterations it is desired to carry out.

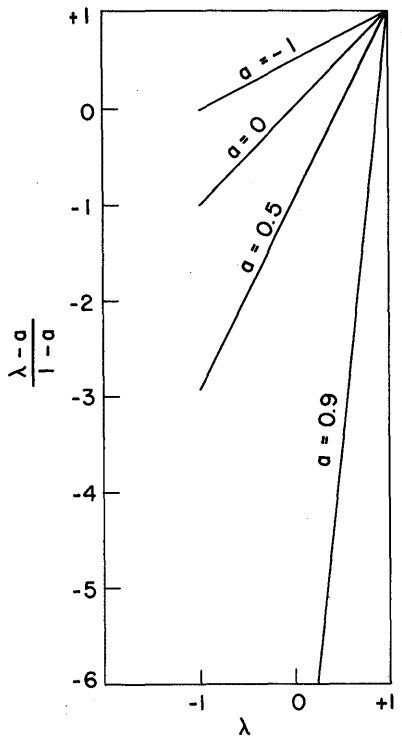


FIGURE 2

It has been found convenient to apply this polynomial, assumed to be of even order with roots symmetrical about 0, as follows:

Operate twice with the fundamental operator ω to obtain $\omega^2\psi$, and then form the linear combination $[(\omega^2 - a^2)/(1 - a^2)]\psi$, thus using the roots $\pm a$. Using this result, one again iterates twice with ω and repeats the process for another value of a, etc., until the roots are all used. It is desirable to use the roots in an order that will prevent high frequency oscillations from building up excessively, as the number of digits carried in the iterations may be exceeded at some points.

The remainder of this paper will be devoted to a discussion of the application of the fundamental operator ω to a function ψ. The problem is to compute the average of the four nearest neighbors for each point of the network. This is done simultaneously for all points of the network, and the resulting set of values is the new function $\omega\psi$. The network of points is shown in Figure 1. It is necessary to cover only half the square because of symmetry.

In performing this operation on the machines a card is provided for each point of the network. The two directions have been labeled x and y as shown in Figure 1; the x and y identification numbers are punched on each card. To run the cards through a machine consecutively in the y direction, they are sorted first on y, then x, and vice versa.

A new deck is used for each iteration. The sequence of operations in an iteration is as follows:

1. By a gang punching and reproducing operation on the reproducer with the cards running in the y direction, the new function and its y neighbors are punched in the new deck from the old deck on which the new function has just been calculated. Both decks have been sorted on y, then x.
2. New deck is sorted on y.
3. Deck is run through the 602-A consecutively on x, allowing the x neighbors to be read from cards ahead and behind. Thus, the average of four neighbors can be calculated.
4. Deck is sorted on x.
5. Cards are listed to check for errors.
6. New function is reproduced and gang punched into still another deck, starting another iteration.

The following special cards are used and necessitate control wiring on the machines (Figure 1).

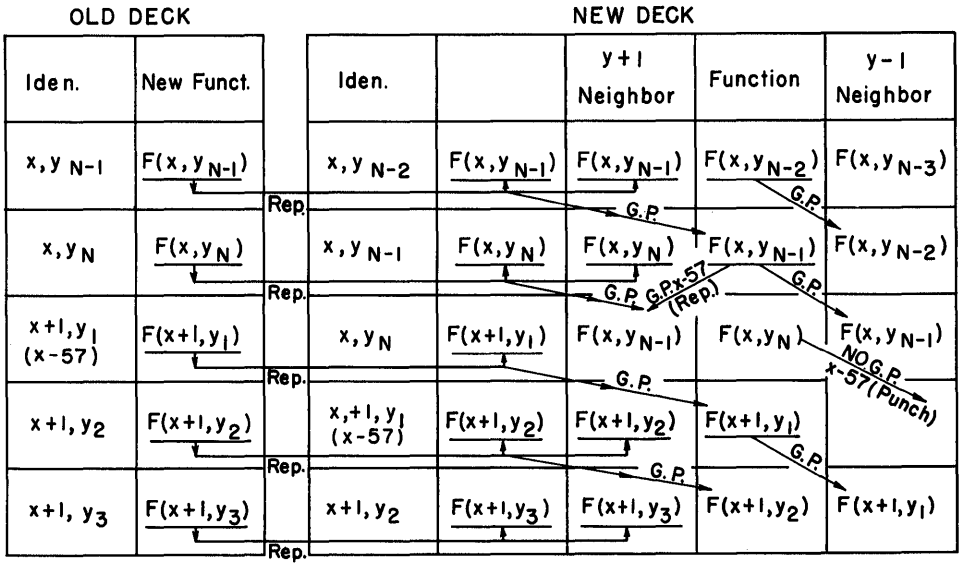
S or successor cards: For these the predecessor neighbor is eliminated. These occur next to the outer zero boundary, for which no card is provided.

2P cards: For these cards, which occur on the diagonal line of symmetry, the predecessor is substituted for the successor, which is not in the deck. That is, for the point x_n, y_m on the diagonal, on the y run one substitutes (x_n, y_{m-1}) $(=x_{n+1}, y_m)$ for x_n, y_{m+1}, and on the x run substitutes (x_{n-1}, y_m) $(=x_n, y_{m+1})$ for x_{n+1}, y_m, thus obtaining the proper neighbors.

Blank cards: These are used for the zero boundaries of the hole. Punching is suppressed for them on the 602-A so that they provide zero neighbors for adjacent points of the network.

The operations in the reproducer run are shown in Figure 3. Note that the cards in the old deck must run one ahead of those in the new one. It is desirable to have auxiliary identification on one deck or the other so that identifications may be compared. The cards shown are at the end of one row and the beginning of another, illustrating the operation of the *2P* and *S* controls in the y direction.

The programming of the 602-A is shown in Figure 4, and control panel wiring is shown in Figure 5, page 32. Here is formed 0.25 times the sum of the neighbors (denoted as Σ).



X-57 OLD DECK 2P (y Direction) Reproducer Control
X-57 NEW DECK S (y Direction) Punch Control

REPRODUCER RUN (y)

FIGURE 3

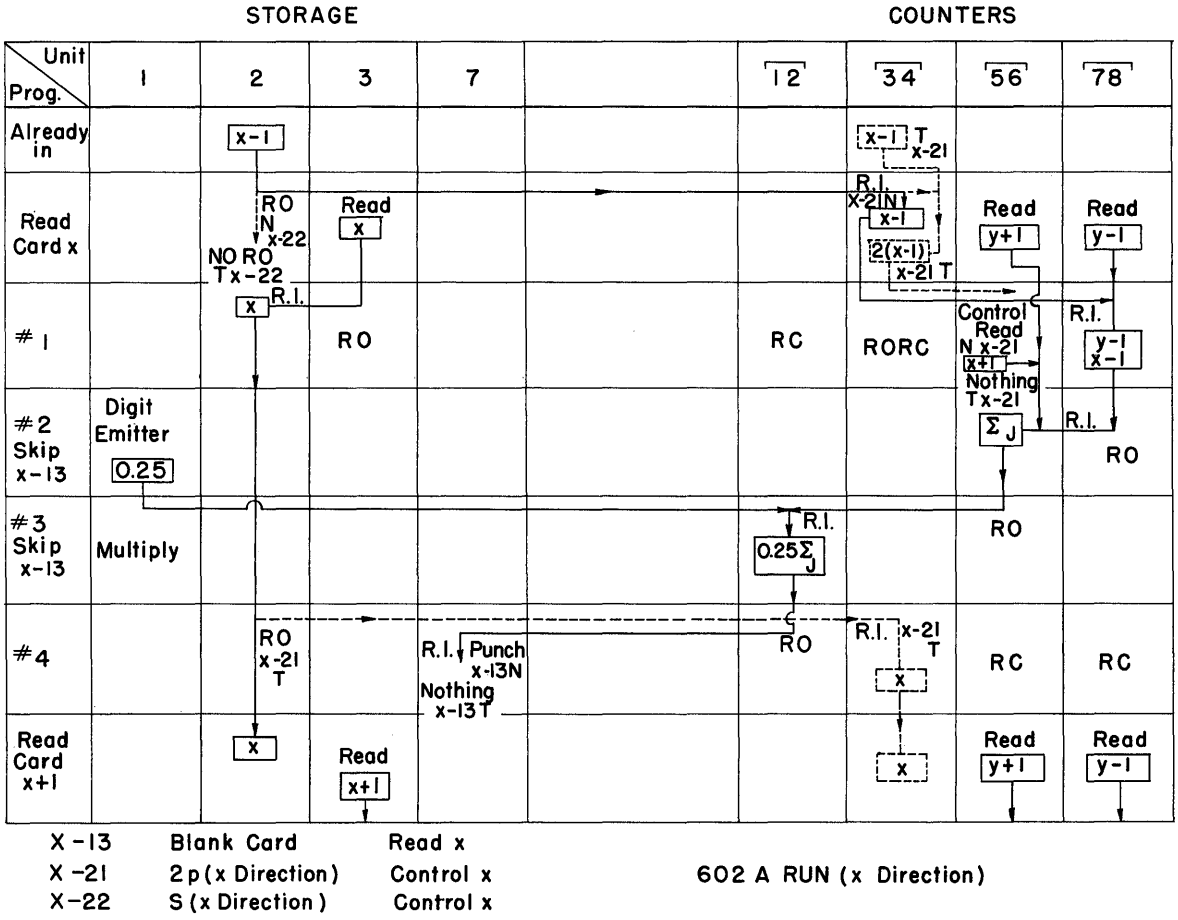


FIGURE 4

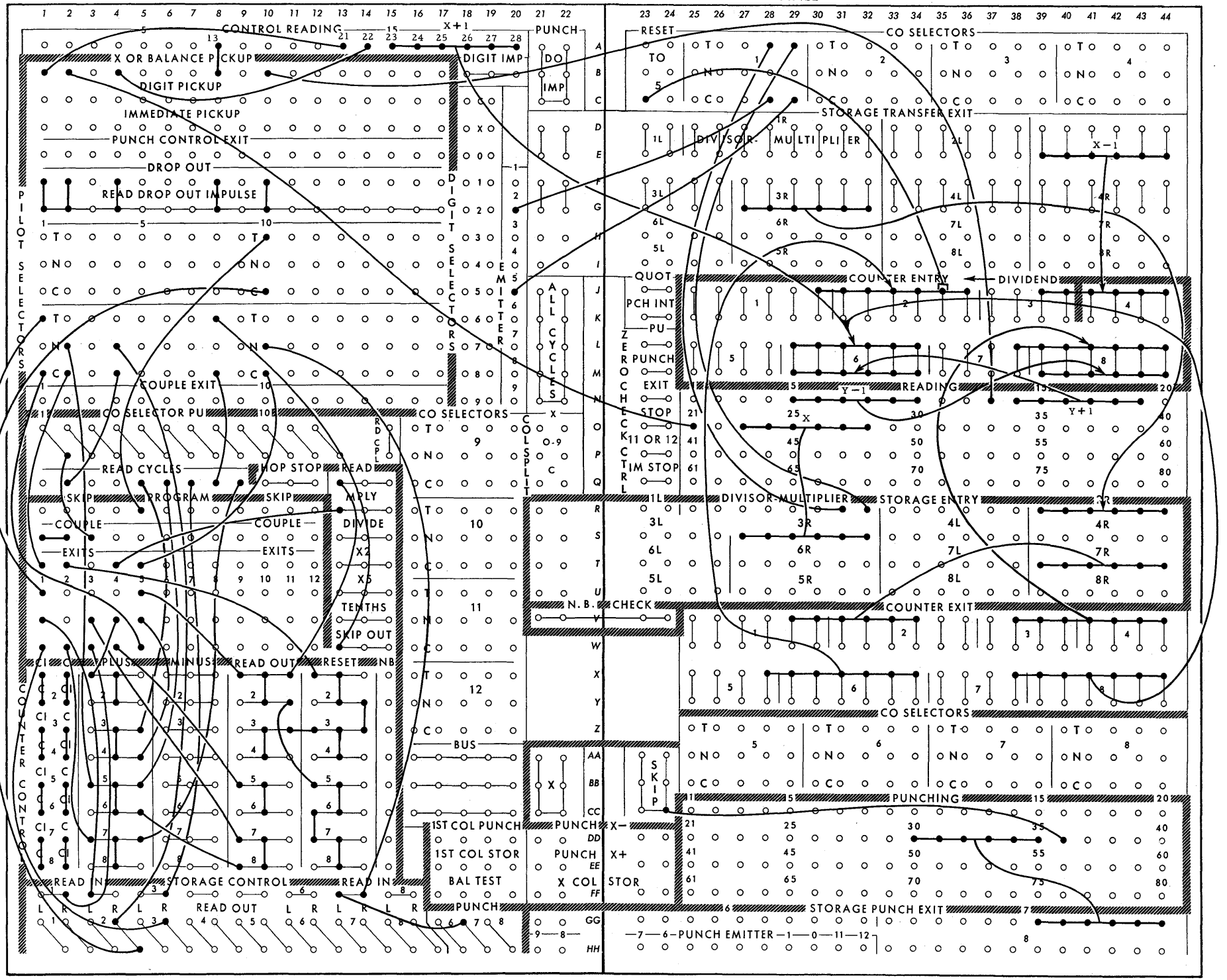

CONTROL READING
X+1
PUNCH
RESET
CO SELECTORS
X OR BALANCE PICKUP
DIGIT IMP
DIGIT PICKUP
IMMEDIATE PICKUP
PUNCH CONTROL EXIT
STORAGE TRANSFER EXIT
DROP OUT
READ DROP OUT IMPULSE
DIVISOR-MULTIPLIER
PILOT SELECTORS
DIGIT SELECTORS
EMITTER
ALL CYCLES
QUOT
COUNTER ENTRY
DIVIDEND
PCH INT
PU
ZERO CHECK CTRL
PUNCH EXIT
READING
COUPLE EXIT
CO SELECTOR PU
CO SELECTORS
RDR FEED
COL SPLIT
11 OR 12
TIM STOP
READ CYCLES
HOP STOP
READ IMPLY
SKIP PROGRAM SKIP
DIVIDE
COUPLE EXITS
COUPLE EXITS
X2
X5
TENTHS
SKIP OUT
N.B. CHECK
COUNTER EXIT
PLUS
MINUS
READ OUT
RESET
NB
BUS
1ST COL PUNCH
PUNCH
1ST COL STOR
PUNCH
BAL TEST
X COL STOR
PUNCHING
READ IN
STORAGE CONTROL
READ IN
READ OUT
STORAGE PUNCH EXIT
PUNCH EMITTER
COUNTER CONTROL

FIGURE 5

The simplest means of calculating the eigenvalue λ is simply to take the new total of the function for all points over the old total. This gives $\lambda = \int\omega\psi/\int\psi$. A more accurate value of λ for the fundamental when higher modes are present may be obtained by forming $\int\psi(\omega\psi)/\int\psi^2$.

Special IBM techniques described in this paper were developed by Mr. James Alexander of the Argonne National Laboratory.

REFERENCE

1. D. A. Flanders and George Shortley, "Eigenvalue Problems Related to the Laplace Operator," *Seminar on Scientific Computation, November, 1949.*

DISCUSSION

Professor Kunz: I would like to point out that there is some difference between this and Laplace's equation, in this sense: In Laplace's equation we have two items of error: (1) How close do we approximate the differential equation? And (2), how close are we to the solution of the difference equation? We have those two, plus the fact that the definition of λ from the difference equation is not the definition of λ that is in the differential equation. This may be seen by considering the simple case of the vibration, let's say, of a drumhead that is one by one. The eigenvalue in this case is $2\pi^2$. The appropriate finite difference equations can be solved exactly. In fact, for any number of points taken, even though the number of interior points is only four or nine, the distribution is the same; the actual distribution. It is a sine sine distribution of the drumhead. So no error is made in approximating the differential equation by a difference equation, as far as the characteristic function is concerned. But the definition of λ is now $\triangle^2\psi/h^2\psi$, which is not the proper definition in terms of the differential equation.

I might mention in that connection that you can obtain a much better result by not iterating further or taking more points, but using simply a higher approximation to the Laplacian.

Dr. Hummel: Yes; you do obtain a more accurate answer.

Mr. Turner: The difficulties in convergence in Dr. Yowell's method and Dr. Hummel's method are associated. As was pointed out, the characteristic solutions of the equation $\nabla^2\psi + \lambda\psi$ form a complete system. Suppose the equation were $\nabla^2\psi = 0$ and in carrying out the numerical operations at some point ij, instead of getting zero, we had ϵ_{ij}. Then, I think, it is quite obvious that these ϵ_{ij}'s or the errors can be composed of linear combinations of the eigenvectors associated with this problem. There are just as many ϵ's as there are points, and there are just as many eigenvectors as there are points in your numerical difference equation.

Therefore, when a set of errors, a set of residuals, occurs, which forms a repetitive pattern, it turns out that they are actually composed of a combination either dominated by one particular eigenvector or made up of a linear combination of the eigenvectors corresponding to one particular eigenvalue.

If we were to go through the operations (I have had it happen in actual numerical calculations that I get a set of residuals) and after operating on them, all we did was to change the magnitude of the residuals but didn't succeed in changing their distribution, then in that case it corresponds rather clearly. If we treat the set of simultaneous equations—that is, the matrix coefficients—as purely a matrix operator, then what we have done is carry out the operation $A\,\epsilon = \lambda\,\epsilon$. That is practically the same thing as happened in Dr. Yowell's paper.

If we will carry out this very simple operation we may, in one step, eliminate the dominant phase of these errors. First, at each point we form the sum of the absolute values of the errors. We also take each of the errors, each of the residuals, and operate on it with our matrix of coefficients.

In other words, with the errors treated as the initial function ψ, let's suppose that the result of operation equals $\lambda\psi$. Let's call it some quantity ν. Now, if we will form a second sum; let's call this one A-B, which is equal to $\Sigma(\nu|\epsilon|/\epsilon)$. This sum is to permit us to determine whether a particular eigenvector, which has perhaps both positive and negative signs, is dominant. Then an approximate value of the λ is B/A, because in the operation on ϵ to obtain $\lambda\epsilon$, that is A on ϵ to get ν, we have multiplied it by the latent root of the matrix of coefficients, which happens to correspond to the dominant eigenvector or combination of eigenvectors having nearly the same latent root. Once we have found λ we can correct our original ψ. If we now have errors of any substantial amount, corresponding to the higher eigenvalues, this will produce a roughness corresponding to a small variation from point to point which the subsequent smoothing process will eliminate quite rapidly.

Mr. Kelly: Has anyone had any experience along these lines of staying within the forced considerations of your mesh? Has anyone observed any forced oscillations of the type Dr. Hamming has found? We observed it staying within the mesh by a factor of 10 to 1, and still observing forced oscillations.

Dr. Hummel: This business of oscillation depends on the range of the eigenvalues, doesn't it? If you know what the range of the eigenvalues is, you can certainly choose the mesh in such a way that you won't get the oscillation. You can always, of course, change your variables in such a way that you don't get the oscillations. This has been found true at least for the solution of the diffusion equation.

Dr. Alt: I have a question for Dr. Hummel in connection with the process of speeding up your convergence by this trick. The division by $(1-a)$ is not essential; that is just to bring the eigenvalues back into scale. But subtracting the constant a reminds me of something that I have seen in the literature that I am not sure is the same thing. It is in a paper by Aitken about 1937. It is the last paragraph of a very long paper, and is easily overlooked. What he mentions is this: Suppose A is a matrix and x is a vector, and that you are trying to solve the equation $Ax - \lambda x = 0$ for the largest value of λ. If you replace the matrix A by $A - a$ this matrix has the eigenvalue $(\lambda - a)$. Some of the methods for finding the λ's converge with a speed which depends on the ratio of the largest to the second largest λ. We are trying to choose a so as to maximize that ratio. But, as you mentioned, you have to make sure that some of the smaller a's don't become large in that process.

I did not hear what you said about getting around that; but there is an answer given by Aitken. Suppose your eigenvalues are $\lambda_1, \lambda_2, \ldots, \lambda_n$, and suppose they are all real and arranged in size here. What you want to subtract is the mean of the second largest and the smallest eigenvalue. It is important to choose it this way. After this, the second root and the smallest root become equal in size and opposite in sign, and the ratio of λ_1 to either of those is maximized. When the eigenvalues are complex, it is a little more complicated. But you can see geometrically what point you have to choose for a.

Dr. Hummel: That is essentially what I have in mind.

Dr. Alt: There is a very brief mention of this simple method by Aitken in *Proceedings Royal Society of Edinburgh, 1936-37.*

Professor Kunz: I would like to point out an even earlier work. There is an article by R. G. D. Richardson in 1911, which is one of the earliest works on stresses in a masonry dam. He considers the choice of a in quite some detail. It has been disapproved of by those who did not know of it.

Dr. Hamming: This point was discussed quite extensively at the last meeting. Flanders went a little bit further in his discussions than has been indicated here.

In the first place, you are not restricted to a linear expression. You can resort to polynomial expressions. What they apparently had done—Flanders, Shortley, and others—was to use the Legendre polynomials, which, as you know, have many roots spread out fairly low and rise sharply at the ends near 1, so that it multiplies this factor and keeps all the rest of the bounds down. In private conversation afterward, we pointed out that he should not have used Legendre polynomials but should have used Tchebyscheff's equal ripple polynomials. In that fashion you are not restricted to this. You simply form a polynomial combination of about the order you want, which would be a Tchebyscheff polynomial—to keep the function down over the whole range, to keep all these down while working only on your maximum—and the degree has to be restricted so that your eigenvalue is not caught over the first zero of the polynomial which you are using.

A Numerical Solution for Systems of Linear Differential Equations Occurring in Problems of Structures

PAUL E. BISCH

North American Aviation, Incorporated

THE PROBLEMS of engineering in which such systems are found and which are successfully solved are:

Determination of natural modes of free oscillations of structures.[1]

Determination of stresses in indeterminate structures.[2]

In general these problems cover all the variations of an actual structure; therefore, the classical solutions are impractical for the equations at hand.

There is only one variation in this solution between one class of problems and the other; so the method will be briefly sketched for the first class (oscillations) for which it is more extensive. Let the differential equation be

$$A_n(x)\frac{d^n y}{dx^n} + \ldots + A_1(x)\frac{dy}{dx} + A_0(x)\,y = 0$$

where the A's are functions of x and may contain a characteristic number λ, or ω^2.

The method, which is presented in detail in reference 1 and reference 2, is briefly described here. The problem includes n boundary conditions and when one boundary condition is used in the differential equation, another equation, called a secondary boundary condition, is obtained. There are n such equations. Altogether there are $2n$ boundary conditions.

The unknown y is then written

$$y = \sum_{i=1}^{s} C_i\,Y_i(x)$$

where C_i are factors to be determined and $Y_i(x)$ are polynomials in x which satisfy the n boundary conditions and the n secondary boundary conditions.

These polynomials $Y_i(x)$ have a form

$$B_1(i)\,x^{a_i+p} + \ldots + B_{2n+1}(i)\,x^{a_i+p(n+1)}$$

where a_i and p are selected in a simple manner, and the $B(i)$'s are obtained from recurrence formulae obtained from the $2n$ boundary conditions. In general a_i and p are the same for the same type of problem, and the $B(i)$'s change very little with the coefficients of the differential equation. The polynomials for the many cases of bending and torsion oscillations are all to be found in reference 1.

If this approximate y and its successive derivatives are used in the differential equation, a function $\epsilon(x)$ is obtained which represents the error, as a correct y would make the left-hand side vanish.

The two boundaries are called x_1 and x_2, and one equation for the solution of the C_i's is obtained from

$$\int_{x_1}^{x_2} \epsilon(x)\,Y_i(x)\,dx = 0\,.$$

There are s such equations, and they are homogeneous in C_i. They can be solved for $s-1$ of them as functions of the s'th, provided that the determinant of the coefficients of the C_i's vanishes. This condition provides an equation of the s power of λ, the s roots of which are positive.

For any root λ_i there results a set of coefficients C_i and therefore an approximate solution y_i of the problem, or mode of oscillation.

This method has many advantages which cannot be pointed out here. It can easily be set up for tabular or IBM calculations. When the $A_n(x)$ are random curves, the integrations can be rapidly made by increments on the IBM machines, thus making the method very general.

Its accuracy is very satisfactory. For instance, it is only necessary to make $s = 3$ in order to obtain the first two modes y_1 and y_2 with an accuracy consistent with the engineering problem.

It can also be said that the preceding integral equations happen also to satisfy the equation of least work for this first class of problems.

On account of its simple algebraic form it is possible, as shown in Sections 1-T and 1-B of reference 1, to solve in advance the problem for a large family of cases.

Other problems of structures which have the same type of differential equations but where the characteristic number λ has another meaning can be similarly solved.

THE SECOND CLASS covers two-dimensional problems of indeterminate structures. An important one is the determination of stresses in box structures as in reference 2. The

key of this problem is the solution of a set of n linear differential equations such as

$$A_m(x)\frac{d^2y_m}{dx^2} + B_m(x)\frac{dy_m}{dx} + C_m^{m-1}(x)\,y_{m-1} + C_m^m(x)\,y_m + C_m^{m+1}(x)\,y_{m+1} = D_m(x)$$

with $m = 1, 2, \ldots, n$. The A, B, and C are known functions of x, and the D's are known functions of x and of other independent variables.

This problem has $2n$ boundary conditions, 2 per equation. Moreover, when these conditions are used in the differential equations written at the boundaries x_1 and x_2, $2n$ secondary boundary conditions are obtained. This makes a total of $4n$ boundary conditions.

The unknown y_m is approximated by

$$y_m = Y_o^m(x) + \sum_{i=1}^{s} C_i^m\,Y_i^m(x)$$

as in the first class of problems.

The polynomials $Y_o^m(x)$ are made to satisfy completely the boundary conditions. Their form is

$$Y_o^m(x) = A_o^m + B_o^m x + C_o^m x^2 + D_o^m x^3.$$

The coefficients A, B, C, and D turn out to be easily computed as linear functions of the independent variables present in the previous $D_m(x)$, and of the boundary values.

The polynomials $Y_i^m(x)$ are set to satisfy the $4n$ boundary equations obtained when the right-hand terms made of independent variables and boundary values are made equal to zero. Their form is

$$Y_i^m = A_i^m x^{a_i+p} + B_i^m x^{a_i+2p} + C_i^m x^{a_i+3p} + D_i^m x^{a_i+4p}$$

where a_i and p are easy to determine and the A, B, C, and D are found from recurrence formulae obtained from the $4n$ mentioned equations.

For the box structure problem of reference (2) the functions Y^m $(i = 0, 1, 2, \ldots, s)$ can be found completely determined in this reference.

The coefficients C_i^m in y_m are determined as in the first class of problems. However, the ns integral equations obtained for the determination of the same number of C_i^m coefficients are not homogeneous, and their direct solution gives these coefficients as linear forms of the independent variables of the problem.

Considerable simplification is obtained by the use of an auxiliary variable z which varies from 0 to 1 between boundaries.

For the problem of reference 2 it was found with sets of six linear differential equations that the solution checked closely the solution based on consideration of the minimum square errors obtained by using the integrals

$$\int_{x_1}^{x_2} \epsilon_m(x)\frac{d\epsilon_m}{dC_i^m}\,dx = 0$$

although $d\epsilon_m/dC_i^m$ is different from $Y_i^m(x)$.

Moreover, these solutions check well the solutions obtained by classical methods for sets of 6 linear equations with constant coefficients, although s was taken as 1 which represents the simplest solution.

In general it will be found that these solutions, in addition to their satisfactory accuracy, are several times shorter than the classical solutions, even for the simplest cases of such systems.

The general solution of a given problem can often be carried algebraically up to the integrated form of the C equations, thus giving a compact solution which can be carried out simultaneously on IBM machines for several cases of the same problem in the time required for one case.

Finally, the writer believes that the same method could be successfully extended to linear partial differential equations, although he knows of no such application to have been made to date.

REFERENCES

1. *North American Aviation Report NA-5811,* "Method of determination of the frequencies, deflection curves and stresses of the first three principal oscillations in torsion and bending of aircraft tapered structures."
2. *North American Aviation Report NA-48-310,* "Determination of stress distribution and rigidity of multi-cell box structures."
3. Aero. Res. Committee Reports and Memoranda:
   R & M 1799–Approximation to Functions and to the Solutions of Differential Equations.
   R & M 1798–Galerkin's Method in Mechanics and Differential Equations.
   R & M 1848–The Principles of the Galerkin's Method.

Matrix Methods

KAISER S. KUNZ

Case Institute of Technology

WE SHALL START with a set of simultaneous linear equations as being something familiar, and I shall restrict my discussion to three equations in three unknowns x_1, x_2, and x_3. The generalization to larger sets should be clear. We can write the coefficients of these unknowns by using a single letter a with two subscripts; thus

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= b_3 \ . \end{aligned} \tag{1}$$

Because of the presence of the constants b_1, b_2, and b_3 (at least one of these is assumed different from zero), this set of equations is said to be nonhomogeneous.

If the determinant of (1)

$$\Delta \equiv \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} \neq 0 \ ; \tag{2}$$

then a solution exists and by Cramer's rule is

$$\left.\begin{aligned} x_1 &= \frac{1}{\Delta}(b_1A_{11} + b_2A_{21} + b_3A_{31}) \\ x_2 &= \frac{1}{\Delta}(b_1A_{12} + b_2A_{22} + b_3A_{32}) \\ x_3 &= \frac{1}{\Delta}(b_1A_{13} + b_2A_{23} + b_3A_{33}) \end{aligned}\right\} . \tag{3}$$

Here A_{ij} is the cofactor of a_{ij}, i.e.,

$$A_{ij} = (-1)^{i+j}\mathcal{A}_{ij} \tag{4}$$

where $\mathcal{A}_{ij}$, the minor of a_{ij}, is the determinant obtained from Δ, by crossing out the row and the column in which a_{ij} occurs; thus

$$A_{12} = -\begin{vmatrix} \cancel{a_{11}} & \cancel{a_{12}} & \cancel{a_{13}} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = -\begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} .$$

Matrices and Matrix Products

Introducing the concept of a matrix and of a matrix product, one can write (1) in the form

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} \tag{5}$$

or simply as

$$ax = b \ , \tag{6}$$

where a, x, and b are termed matrices and stand, respectively, for the sets of numbers included in the parentheses of (5). A matrix is conceived of as a complex of all the numbers in the parentheses; thus, if any of these numbers is changed, the matrix is changed.

Two matrices are equal only if they have the same number of rows, the same number of columns, and corresponding elements (numbers) are equal. Matrices of the type of x and b are called column matrices or vectors. In order that (5) be equivalent to (1), the product of the matrix a and the vector x must be the vector

$$\begin{pmatrix} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 \end{pmatrix} .$$

This may be summarized by saying that the elements of the product are

$$(ax)_i = \sum_{k=1}^{3} a_{ik}x_k \ . \tag{7}$$

In the same manner, the solutions x_1, x_2, x_3 given by (3) can be expressed by the matrix equation

$$x = a^{-1}b \ , \tag{8}$$

where the square matrix a^{-1} has elements

$$a^{-1}_{ij} = \frac{1}{\Delta}A_{ji} \ . \tag{9}$$

The matrix a^{-1} is called the inverse of the matrix a. The numerical determination of the elements of the inverse of a given matrix is one of the important problems of numerical analysis.

Another basic problem is the finding of the product of two matrices when the number of rows and/or columns is large. The product of a matrix A by a matrix B can be taken only if the number of columns of A, say p, is equal to the number of rows of B. If this condition is met, the product is a matrix C having the same number of rows as A, say m, and the same number of columns as B, say n. The elements of C are

$$C_{ij} = \sum_{k=1}^{p} A_{ik}B_{kj} \ , \tag{10}$$

where $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$. Clearly (7) is a special case, $n = 1$, of this rule.

A matrix having n rows and n columns is termed a square matrix of degree n. It is easily verified that the product of a square matrix and a vector involves n^2 multiplications, and the product of two square matrices of nth degree requires n^3 multiplications.

Numerical Solution of Simultaneous Linear Equations and Matrix Inversion

The numerical solution of a set of simultaneous linear equations by means of (3) is usually thoroughly impractical when the number of equations n is of the order of ten or more. This is due to the excessively large number of multiplications required, namely, $(n+1)!\,\gamma(n)$, where $1 \leqq \gamma(n) < e-1$, $(e = 2.718\ldots)$. For $n = 10$, this means about 70,000,000 multiplications. Compare this number with about $1/3\ n^3$ multiplications, or actually about 410 when $n = 10$, needed for the elimination method.

It is convenient in dealing with simultaneous linear equations to work with a matrix composed of all the given constants,

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & b_1 \\ a_{21} & a_{22} & a_{23} & b_2 \\ a_{31} & a_{32} & a_{33} & b_3 \end{pmatrix}, \tag{11}$$

the so-called augmented matrix. From a computational point of view this set of numbers represents the original equations written with detached coefficients, much as is done in evaluating a polynomial by synthetic division. Equality marks can be imagined before each of the elements of the last column.

It is clear from this point of view that the augmented matrix, obtained from a given matrix by subtracting from the elements of any row some multiple of the corresponding elements of a second row, is equivalent to the given matrix. By equivalent is meant that the solutions of the set of equations represented by these two augmented matrices are the same.

Elimination Method

In the elimination method $1/a_{11}$ is computed, and all the elements of the first row are multiplied by this number.[a] The first element of this row is now unity; by subtracting from each element of each of the other rows the proper multiple of the corresponding elements of the first row, one obtains the matrix.

$$\begin{pmatrix} 1 & a'_{12} & a'_{13} & b'_1 \\ 0 & a'_{22} & a'_{23} & b'_2 \\ 0 & a'_{32} & a'_{33} & b'_3 \end{pmatrix}$$

If the equation represented by the first row is dropped momentarily from consideration, two equations in two unknowns remain. Therefore, we can deal with these elements in the same way, thereby obtaining a 1 in place of the a'_{22} and a zero in place of the a'_{32}. It is possible, then, to drop consideration of the second row. It is clear that by continuing this process the matrix can be reduced to the form

$$\begin{pmatrix} 1 & \bar{a}_{12} & \bar{a}_{13} & \bar{b}_1 \\ 0 & 1 & \bar{a}_{23} & \bar{b}_2 \\ 0 & 0 & 1 & \bar{b}_3 \end{pmatrix}, \tag{12}$$

in which the $\bar{a}_{ij}$ are equal to 1 along the diagonal and zero below. We will refer to (12) as the triangular form.

Starting with n equations in n unknowns, this reduction will require n divisions and $(1/6)n\,(n+1)\,(2n+1)$ multiplications. This assumes, of course, that there are no ones or zeros in the equations. For large n, there are about $(1/3)n^3$ multiplications. The n divisions are negligible.

Having obtained the triangular form (12), the last row may be multiplied by $\bar{a}_{23}$ and subtracted from the second row, and by $\bar{a}_{13}$ and subtracted from the first row. This will cause zeros to appear where $\bar{a}_{13}$ and $\bar{a}_{23}$ are in (12). Similarly by multiplying the new second row by $\bar{a}'_{12}$, the new $\bar{a}_{12}$ obtained from the above process, a zero is established at that location. The equations are now in the diagonal form

$$\begin{pmatrix} 1 & 0 & 0 & b_1^* \\ 0 & 1 & 0 & b_2^* \\ 0 & 0 & 1 & b_3^* \end{pmatrix}, \tag{13}$$

where the stars simply indicate new constants obtained by the process. The solution of these equations and therefore of the given equations is $x_1 = b_1^*$, $x_2 = b_2^*$, and $x_3 = b_3^*$.

The number of multiplications needed to go from (12) to (13) is only $(1/2)n\,(n-1)$ or, for large n, approximately $(1/2)n^2$. Thus, nearly all the computation is needed to obtain the equations in triangular form. In fact, for large n, the number of multiplications needed for the whole elimination method is still approximately $(1/3)n^3$.

While considering this method, it is convenient to see how it may be carried over to the inversion of matrices. The technique is very much the same, but instead of working with the augmented matrix (11), in which the b's are known numbers that can be multiplied, subtracted, etc., the b's are treated in the same way as the x's and each one is provided with a separate column. The augmented matrix now appears as follows:

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & 1 & 0 & 0 \\ a_{21} & a_{22} & a_{23} & 0 & 1 & 0 \\ a_{31} & a_{32} & a_{33} & 0 & 0 & 1 \end{pmatrix}. \tag{14}$$

The numbers in the first three columns again represent the coefficients of x_1, x_2, and x_3, respectively, while the numbers in the last three columns represent the coefficients of b_1, b_2, and b_3. Equality marks can be imagined between the third and fourth columns. Thus, one can look upon (14) as the equations (1) in skeleton form.

[a] Actually, one needs first to arrange the equations in such a way that a_{11} is one of the larger diagonal terms so that multiplying by $1/a_{11}$ does not reduce the number of significant figures.

Without proving it, if the nine a's are operated upon in exactly the same way as before, by means of operations on individual rows of (14), letting the numbers accumulate as they will in the b columns, the following matrix is obtained

$$\begin{pmatrix} 1 & \bar{a}_{12} & \bar{a}_{13} & \bar{c}_{11} & 0 & 0 \\ 0 & 1 & \bar{a}_{23} & \bar{c}_{21} & \bar{c}_{22} & 0 \\ 0 & 0 & 1 & \bar{c}_{31} & \bar{c}_{32} & \bar{c}_{33} \end{pmatrix}. \tag{15}$$

At this stage the matrices of the coefficients of the x's and of the b's are each in triangular form. The number of multiplications needed for this step is about $(1/2)n^3$.

Now multiply the second row by $\bar{a}_{12}$ and subtract it from the first row, thereby introducing a zero at the location of $\bar{a}_{12}$. Similarly, by working with the third row, zeros can be obtained at the $\bar{a}_{13}$ and $\bar{a}_{23}$ positions. By continuing this process, the matrix can be written in the form

$$\begin{pmatrix} 1 & 0 & 0 & c_{11}^* & c_{12}^* & c_{13}^* \\ 0 & 1 & 0 & c_{21}^* & c_{22}^* & c_{23}^* \\ 0 & 0 & 1 & c_{31}^* & c_{32}^* & c_{33}^* \end{pmatrix}. \tag{16}$$

The number of multiplications needed for this step is again about $(1/2)n^3$.

The three equations represented by (16) express x_1, x_2, and x_3, respectively, in terms of the b's. Since the c_{ij}^* are the coefficients of the b's in these expressions, they are the elements of the inverse matrix [see equation (8)]. The amount of computation needed to invert a matrix in this manner is indicated roughly by the n^3 multiplications, needed for the two steps above. The n divisions and many additions required are customarily neglected.

Having the inverse of the matrix a, the vector b can be multiplied by it to obtain the numerical values of the x's. As seen earlier, this requires n^2 multiplications. To solve k sets of equations with the same coefficients a_{ij}, but with different b_i requires about $n^3 + kn$ multiplications. On the other hand, to solve the sets of equations individually would require about $(k/3)n^3$ multiplications.

Therefore, if you have at least four sets of equations to solve, with the same a_{ij} but different b_i, and the number of equations in the sets is twelve or more, it is better to first invert the matrix a. Also, it is necessary to invert the matrix if the b's are not specified numerically or are treated as variables.

Now, I would like to pass to other methods of finding the solution of a set of linear equations. As we have seen, the change in procedure from this, to the finding of an inverse of a matrix, is not very difficult. One simply makes provision for keeping the b's separated, instead of allowing them to add together as one proceeds.

There are a great many methods which are closely related to the elimination method. Although some of these are very good for certain purposes, such as the Crout method for a desk-type calculator, I shall lump them with the elimination method. The methods I shall treat are chosen mainly for their interest.

Square Root Method

The square root method is not completely general. It assumes that the matrix a of the coefficients of the x's is symmetric. By this is meant that any coefficient a_{ji} is always equal to a_{ij}. This limitation on the form of a is not as serious as one might suppose, since many of the matrices arising from physical problems are symmetric.

In this method, rather than obtain the intermediate triangular form (12) by some operations on the rows of the augmented matrix, we assume it and determine the conditions that such an assumption imposes. In particular, we assume that the matrix a can be expressed as the product of a triangular matrix S and its transpose S', a matrix obtained from S by interchanging rows and columns. Or, more precisely (17)

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \dots a_{1n} \\ a_{12} & a_{22} & a_{23} \dots a_{2n} \\ a_{13} & a_{23} & a_{33} \dots a_{3n} \\ \cdot & \cdot & \cdot \quad \cdot \\ a_{1n} & a_{2n} & a_{3n} \dots a_{nn} \end{pmatrix} = \begin{pmatrix} s_{11} & 0 & 0 & \dots 0 \\ s_{12} & s_{22} & 0 & \dots 0 \\ s_{13} & s_{23} & s_{33} \dots 0 \\ \cdot & \cdot & \cdot \quad \cdot \\ s_{1n} & s_{2n} & s_{3n} \dots s_{nn} \end{pmatrix} \begin{pmatrix} s_{11} & s_{12} & s_{13} \dots s_{1n} \\ 0 & s_{22} & s_{23} \dots s_{2n} \\ 0 & 0 & s_{33} \dots s_{3n} \\ \cdot & \cdot & \cdot \quad \cdot \\ 0 & 0 & 0 \dots s_{nn} \end{pmatrix},$$

which is just the equation $a = S'S$.

Therefore, in place of (5) we write

$$\begin{pmatrix} s_{11} & 0 & 0 & \dots 0 \\ s_{12} & s_{22} & 0 & \dots 0 \\ s_{13} & s_{23} & s_{33} \dots 0 \\ \cdot & \cdot & \cdot \quad \cdot \\ s_{1n} & s_{2n} & s_{3n} \dots s_{nn} \end{pmatrix} \begin{pmatrix} k_1 \\ k_2 \\ k_3 \\ \cdot \\ k_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ \cdot \\ b_n \end{pmatrix}, \tag{18}$$

where the k matrix is the matrix product,

$$\begin{pmatrix} s_{11} & s_{12} & s_{13} \dots s_{1n} \\ 0 & s_{22} & s_{23} \dots s_{2n} \\ 0 & 0 & s_{33} \dots s_{3n} \\ \cdot & \cdot & \cdot \quad \cdot \\ 0 & 0 & 0 \dots s_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \cdot \\ x_n \end{pmatrix} = \begin{pmatrix} k_1 \\ k_2 \\ k_3 \\ \cdot \\ k_n \end{pmatrix}. \tag{19}$$

The elements s_{ij} are obtained from the known elements of a by the use of the equations

$$\begin{aligned} s_{11} &= \sqrt{a_{11}} \\ s_{1j} &= \frac{a_{1j}}{s_{11}} \\ s_{ii} &= \sqrt{a_{ii} - \sum_{r=1}^{i-1} s_{ri}^2}\,, \; i \geqq 2 \\ s_{ij} &= \frac{1}{s_{ii}}\left(a_{ij} - \sum_{r=1}^{i-1} s_{ri}\, s_{rj}\right), \; j > i \geqq 2\,. \end{aligned} \tag{20}$$

These equations result from multiplying the matrices S' and S in (17) and equating corresponding elements on the two sides of that equation.

Having the s_{ij}, (18) can be solved, very easily, for the k_i and (19) for the x_i, since these equations are in diagonal form. These steps require only about $2n^2$ multiplications. The total number of multiplications for the whole process is approximately $(1/6)n^3$.

This method requires, therefore, only about half as many operations as the elimination method; however, it is applicable only to symmetric matrices. While the taking of a square root is not usually as simple as a multiplication or a division, the process requires only n square roots and n divisions, which are negligible for sizeable n. The coding of this for automatic calculators has been tried, I believe. (Some difficulty in this respect arises from the need, at times, to take a square root of a negative number, thus introducing $i = \sqrt{-1}$ into our computations. The numbers, however, are either real or pure imaginary.)

Iterative Methods

The next method I would like to consider is an iterative method, a method of successive approximation, or sometimes called the Gauss-Seidel method. In discussing this method I should like to treat a particular example. Again, I shall restrict myself to three equations in three unknowns. The particular equations I shall consider are:

$$\left.\begin{aligned} 25x_1 + 2x_2 + x_3 &= 69 \\ 2x_1 + 10x_2 + x_3 &= 63 \\ x_1 + x_2 + 4x_3 &= 43 \end{aligned}\right\} . \tag{21}$$

I have used a mathematician's prerogative to deal with nice round numbers. Seldom will you be called upon to work with such convenient numbers.

Often new methods are developed by starting with very bold approximations. Here we shall initially assume that all the coefficients, that are less than 4, are small enough to be neglected. This makes it possible to write down at once as a zeroth approximation,

$$\left.\begin{aligned} 25x_1^{(0)} &= 69 \\ 10x_2^{(0)} &= 63 \\ 4x_3^{(0)} &= 43 \end{aligned}\right\} . \tag{22}$$

The equations (21) are approximated, thus, by a set of equations in diagonal form.

This brings up a problem in terminology. Does one start with a zeroth or a first approximation? Well, I have adopted the following answer. You normally employ the designation first approximation, except when making a wild guess that cannot properly be justified; then it is a zeroth approximation. Surely in this case it is a zeroth approximation, which shall be designated by the matrix

$$x^{(0)} = \begin{pmatrix} x_1^{(0)} \\ x_2^{(0)} \\ x_3^{(0)} \end{pmatrix} = \begin{pmatrix} 2.76 \\ 6.3 \\ 10.75 \end{pmatrix} . \tag{23}$$

For purposes of comparison, the correct answer is

$$x = \begin{pmatrix} 2 \\ 5 \\ 9 \end{pmatrix} . \tag{24}$$

Surprisingly enough, the zeroth approximation gives the right order of magnitude for the x's.

Now, having some idea of the size of the individual x's, we can go back to (21) and correct for the off-diagonal terms. Thus, the first approximation is written:

$$\left.\begin{aligned} 25x_1^{(1)} &= 69 - 2x_2^{(0)} - x_3^{(0)} \\ 10x_2^{(1)} &= 63 - 2x_1^{(0)} - x_3^{(0)} \\ 4x_3^{(1)} &= 43 - x_1^{(0)} - x_2^{(0)} \end{aligned}\right\} . \tag{25}$$

The right-hand sides are known; so $x^{(1)}$ can be found, which turns out to be

$$x^{(1)} = \begin{pmatrix} 1.826 \\ 4.673 \\ 8.485 \end{pmatrix} .$$

This still differs considerably from the answer given in (24), but progress is being made.

Having a better answer for x, we are in position to make a still better estimate of the correction terms in (25); therefore, we can obtain a better approximation $x^{(2)}$. This process can be repeated as often as desired. Thus, in general

$$\left.\begin{aligned} 25x_1^{(i+1)} &= 69 - 2x_2^{(i)} - x_3^{(i)} \\ 10x_2^{(i+1)} &= 63 - 2x_1^{(i)} - x_3^{(i)} \\ 4x_3^{(i+1)} &= 43 - x_1^{(i)} - x_2^{(i)} \end{aligned}\right\} , \tag{26}$$

and these equations can be written

$$x^{(i+1)} = \begin{pmatrix} x_1^{(i+1)} \\ x_2^{(i+1)} \\ x_3^{(i+1)} \end{pmatrix} = \begin{pmatrix} 2.76 \\ 6.3 \\ 10.75 \end{pmatrix} + \begin{pmatrix} 0 & -0.08 & -0.04 \\ -0.2 & 0 & -0.1 \\ -0.25 & -0.25 & 0 \end{pmatrix} \begin{pmatrix} x_1^{(i)} \\ x_2^{(i)} \\ x_3^{(i)} \end{pmatrix} . \tag{27}$$

At the sixth approximation, we have

$$x^{(6)} = \begin{pmatrix} 2.002 \\ 5.0004 \\ 9.0006 \end{pmatrix} .$$

These values are close enough so that we do not need to apologize for them, and, clearly, the iteration can be continued with consequent improvement of the results, as long as desired. This is a Gauss-Seidel process. Whether the process converges depends on how large the diagonal terms are, compared to the off-diagonal terms.

A slightly different technique may also be employed. Instead of pruning off all the non-diagonal terms, just those terms above (or below) the diagonal may be removed. This reduces the equations for the zeroth approximation to triangular form. To solve these, only n^2 multiplications are needed instead of the $(1/3)n^3$ needed for the original equations.

This method is due to Morris. He showed that if the matrix of the coefficients of a is positive semidefinite, which means that the associated quadratic form is either positive or zero, then the process converges. The iteration equations here are

$$\left.\begin{aligned} 25x_1^{(i+1)} &= 69 - 2x_2^{(i)} - x_3^{(i)} \\ 10x_2^{(i+1)} &= 63 - 2x_1^{(i+1)} - x_3^{(i)} \\ 4x_3^{(i+1)} &= 43 - x_1^{(i+1)} - x_2^{(i+1)} \end{aligned}\right\} . \tag{28}$$

Having the ith approximation of all three x's, the $(i+1)$ approximation of x_1 is obtained from the first equation. Then, knowing $x_1^{(i+1)}$, the second equation can be solved for $x_2^{(i+1)}$ and thereafter the third equation for $x_3^{(i+1)}$.

Going back to the question of the convergence of the iteration equations (26), this equation can be written in matrix form as (29)

$$\begin{pmatrix} 25 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 4 \end{pmatrix} \begin{pmatrix} x_1^{(i+1)} \\ x_2^{(i+1)} \\ x_3^{(i+1)} \end{pmatrix} = \begin{pmatrix} 69 \\ 63 \\ 43 \end{pmatrix} - \begin{pmatrix} 0 & 2 & 1 \\ 2 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} x_1^{(i)} \\ x_2^{(i)} \\ x_3^{(i)} \end{pmatrix}.$$

This in turn may be written

$$E\,x^{(i+1)} = b - H\,x^{(i)}\,, \tag{30}$$

where E and H are the square matrices shown. Note that this latter equation could be obtained directly from equations (21), which can be written in the matrix form $A\,x = b$, by observing that $A = E + H$.

If both sides of (30) are multiplied by the inverse of E,

$$E^{-1} = \begin{pmatrix} 0.04 \\ 0.10 \\ 0.25 \end{pmatrix},$$

we obtain the equation

$$x^{(i+1)} = E^{-1}\,b - E^{-1}\,H\,x^{(i)}. \tag{31}$$

Now, since $E^{-1}\,b = x^{(0)}$, the solution of the diagonal equations, $E\,x^{(0)} = b$, given in (22), and letting $F = -E^{-1}\,H$, equation (31) can be written

$$x^{(i+1)} = x^{(0)} + F\,x^{(i)}\,. \tag{32}$$

Written out, this is just equation (27).

In particular (33)

$$\left.\begin{aligned} x^{(1)} &= x^{(0)} + F\,x^{(0)} &&= (1+F)\,x^{(0)} \\ x^{(2)} &= x^{(0)} + F\,x^{(1)} &&= (1+F+F^2)\,x^{(0)} \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ x^{(n)} &= x^{(0)} + F\,x^{(n-1)} &&= (1+F+F^2+\ldots+F^n)\,x^{(0)} \end{aligned}\right\};$$

thus, the convergence of the process reduces to the question of the convergence of the series

$$\sum_{i=0}^{\infty} F^i\,x^{(0)}\,. \tag{34}$$

The latter can be shown to require that the characteristic values of F, in absolute value, be all less than unity. These characteristic values are the possible values of the constant λ in the equation

$$F\,\psi = \lambda\,\psi\,, \tag{35}$$

where

$$\psi = \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix}$$

is any vector chosen so as to satisfy (35). Such a vector is called a characteristic vector.

Generally there will be three ψ's and three corresponding λ's, that satisfy this equation.

Thus, in place of (35) we may write

$$F\,\psi_i = \lambda_i\,\psi_i\,,\ i = 1,2,3. \tag{36}$$

For n equations the number, of course, will be n instead of 3.

I shall not prove the above requirement for convergence, except in the following plausible way. Let us assume that the characteristic vectors ψ_1, ψ_2, ψ_3 are three linearly independent vectors, so that $x^{(0)}$ can be expanded in terms of them,

$$x^{(0)} = k_1\,\psi_1 + k_2\,\psi_2 + k_3\,\psi_3\,, \tag{37}$$

where the k's are suitable constants. Then

$$F\,x^{(0)} = k_1\,F\,\psi_1 + k_2\,F\,\psi_2 + k_3\,F\,\psi_3\,,$$

or by (36)

$$F\,x^{(0)} = k_1\,\lambda_1\,\psi_1 + k_2\,\lambda_2\,\psi_2 + k_3\,\lambda_3\,\psi_3\,.$$

Repeated application of F to $x^{(0)}$, therefore, gives

$$F^i\,x^{(0)} = k_1\,\lambda_1^i\,\psi_1 + k_2\,\lambda_2^i\,\psi_2 + k_3\,\lambda_3^i\,\psi_3\,,$$

and hence, if

$$|\,\lambda_i\,| < 1\,,\ \text{for } i = 1, 2, \text{and } 3,$$
$$\lim_{i\to\infty} F^i\,x^{(0)} = 0.$$

Moreover, if $|\,\lambda_1\,| > |\,\lambda_2\,|$ and $|\,\lambda_1\,| > |\,\lambda_3\,|$ and n is sufficiently large $F^{n+1}\,x^{(0)} \doteq k_1\,\lambda_1^{n+1}\,\psi_1 \doteq \lambda_1\,F^n\,x^{(0)}$. Thus, the series (34), as far as convergence is concerned, acts like a geometric series with a ratio given by λ_1. Since $|\,\lambda_1\,| < 1$, this series should converge.

Finding Characteristic Values of Matrices

The above discussion of the iteration methods points up the need to find the characteristic values λ_i for a matrix. This fundamental problem is very interesting. Let me make several observations concerning it.

Let F again be the matrix under consideration, but let it be used to represent any matrix being studied. We require the characteristic values λ_i of equations (35) and (36). For this purpose, let us introduce the unit matrix,

$$I = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \tag{38}$$

then (35) can be written

$$(F - \lambda I)\,\psi = 0\,. \tag{39}$$

In expanded form, (39) is written as follows

$$\begin{pmatrix} F_{11}-\lambda & F_{12} & F_{13} \\ F_{21} & F_{22}-\lambda & F_{23} \\ F_{31} & F_{32} & F_{33}-\lambda \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = 0\,. \tag{40}$$

Equation (40) represents three homogeneous linear equations for the components c_1, c_2, and c_3 of ψ.

A solution of these equations, other than trivial solution $c_1 = c_2 = c_3 = 0$, is possible only if the determinant of the coefficients is zero, that is

$$D(\lambda) \equiv \begin{vmatrix} F_{11}-\lambda & F_{12} & F_{13} \\ F_{21} & F_{22}-\lambda & F_{23} \\ F_{31} & F_{32} & F_{33}-\lambda \end{vmatrix} = 0\,. \tag{41}$$

Equation (41) leads to a polynomial equation of the nth degree in λ (here $n = 3$); therefore, there are at most n characteristic values.

One of the methods for finding the largest of the characteristic values, say λ_1, is indicated by our previous discus-

sion. If ψ is any initial vector, which can be expanded in terms of the characteristic vectors, as $x^{(0)}$ was in (37), then repeated application of F leads to a vector that is nearly ψ_1 times a constant. When one reaches this point, any further multiplication by F multiplies this vector by λ_1, essentially. There are several methods based on this fact.

Another method, one with which you may not be familiar, is to solve for the roots of $D(\lambda)$ directly by the use of the method of false position; that is, we substitute some value of λ in the determinant and evaluate the determinant numerically. We repeat this for several neighboring λ's. This gives us a few points on the plot of $D(\lambda)$ versus λ. If $D(\lambda)$ has opposite signs for two values λ_a and λ_b, then, since $D(\lambda)$ is a polynomial and hence continuous in λ, it must be zero somewhere between these values. The method of false position estimates this value by assuming a linear variation of $D(\lambda)$ between λ_a and λ_b, and is conveniently coded for automatic computation.

Since each evaluation of $D(\lambda)$ requires evaluating a determinant of the nth order, and this requires $(1/3)n^3$ multiplications, the process is open to serious objections. If it is at all feasible, it is desirable to evaluate the coefficients of powers of λ in $D(\lambda)$, since once this is done, the task of obtaining $D(\lambda)$ for some value of λ is reduced to n multiplications. The great advantage of methods of this sort is that all of the characteristic values can be evaluated at least in principle, and not just the largest.

I shall close by pointing out one of the simple ways in which one may obtain an upper bound for the absolute value of the largest characteristic value λ_1. This can be accomplished by considering equation (40), which must be satisfied. For the moment, assume that $| c_1 |$ is the largest of the three numbers $| c_1 |$, $| c_2 |$, and $| c_3 |$, the absolute values of the components of the characteristic vector ψ corresponding to λ_1. Then from the first equation arising from (40), we have

$$\lambda_1 c_1 = F_{11} c_1 + F_{12} c_2 + F_{13} c_3$$

or

$$| \lambda_1 | \leqq \frac{|F_{11}||c_1| + |F_{12}||c_2| + |F_{13}||c_3|}{|c_1|} \leqq |F_{11}| + |F_{12}| + |F_{13}| .$$

This is just the sum of the absolute values of elements of the first row of F.

If $|c_2|$ is the largest of the constants, it can be shown, from the second equation, that $|\lambda_1|$ is less than the sum of the absolute values of the elements in the second row. Likewise for $|c_3|$ largest, the absolute values of the elements in the third row are summed. Without making any assumptions as to the relative sizes of the $|c_i|$, nevertheless, the following rule can be stated: If the absolute values of the elements of each individual row are summed and the largest sum chosen, this sum must exceed the largest characteristic value. This upper limit is very helpful.

There are, of course, still other schemes for determining an upper bound on the size of the characteristic values.

Inversion of an Alternant Matrix

BONALYN A. LUCKEY

General Electric Company

AN ALTERNANT MATRIX is of the form as shown in Figure 1. It is a square matrix of order n in which the elements of each row are increasing powers from 0 to $N-1$ of A_N. Thus the elements of the first column are one.

According to Aitken,[1] the reciprocal of such a matrix can be written. The elements of columns 1, 2 and n are shown in Figures 2, 3, 4, respectively.

$$\begin{bmatrix} 1 & A_1 & A_1^2 & \cdot & \cdot & A_1^{N-1} \\ 1 & A_2 & A_2^2 & \cdot & \cdot & A_2^{N-1} \\ 1 & A_3 & A_3^2 & \cdot & \cdot & A_3^{N-1} \\ 1 & A_4 & A_4^2 & \cdot & \cdot & A_4^{N-1} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & A_N & A_N^2 & \cdot & \cdot & A_N^{N-1} \end{bmatrix}$$

FIGURE 1

1ST COLUMN

$$\pm \frac{A_2 \cdot A_3 \cdot A_4 \cdot A_5 \ldots A_N}{(A_1-A_2)\,(A_1-A_3)\,(A_1-A_4)\ldots(A_1-A_N)}$$

$$\cdot$$
$$\cdot$$
$$\cdot$$

$$+ \frac{A_2A_3 + A_2A_4 + \ldots + A_2A_N + A_3A_4 + \ldots + A_3A_N + \ldots + A_{N-1}A_N}{(A_1-A_2)\,(A_1-A_3)\,(A_1-A_4)\ldots(A_1-A_N)}$$

$$- \frac{A_2 + A_3 + A_4 + A_5 + \ldots + A_N}{(A_1-A_2)\,(A_1-A_3)\,(A_1-A_4)\ldots(A_1-A_N)}$$

$$+ \frac{1}{(A_1-A_2)\,(A_1-A_3)\,(A_1-A_4)\ldots(A_1-A_N)}$$

FIGURE 2

2ND COLUMN

$$\pm \frac{A_1 \cdot A_3 \cdot A_4 \cdot A_5 \ldots A_N}{(A_2-A_1)\,(A_2-A_3)\,(A_2-A_4)\ldots(A_2-A_N)}$$

$$\cdot$$
$$\cdot$$
$$\cdot$$

$$+ \frac{A_1A_3 + A_1A_4 + \ldots + A_1A_N + A_3A_4 + \ldots + A_3A_N + \ldots + A_{N-1}A_N}{(A_2-A_1)\,(A_2-A_3)\,(A_2-A_4)\ldots(A_2-A_N)}$$

$$- \frac{A_1 + A_3 + A_4 + A_5 + \ldots + A_N}{(A_2-A_1)\,(A_2-A_3)\,(A_2-A_4)\ldots(A_2-A_N)}$$

$$+ \frac{1}{(A_2-A_1)\,(A_2-A_3)\,(A_2-A_4)\ldots(A_2-A_N)}$$

FIGURE 3

NTH COLUMN

$$\pm \frac{A_1 \cdot A_2 \cdot A_3 \cdot A_4 \ldots A_{N-1}}{(A_N-A_1)\,(A_N-A_2)\,(A_N-A_3)\ldots(A_N-A_{N-1})}$$

$$\cdot$$
$$\cdot$$
$$\cdot$$

$$- \frac{A_1A_2A_3 + A_1A_2A_4 + \ldots + A_1A_2A_{N-1} + A_1A_3A_4 + \ldots + A_1A_3A_{N-1} + \ldots + A_{N-3}A_{N-2}A_{N-1}}{(A_N-A_1)\,(A_N-A_2)\,(A_N-A_3)\ldots(A_N-A_{N-1})}$$

$$+ \frac{A_1A_2 + A_1A_3 + \ldots + A_1A_{N-1} + A_2A_3 + \ldots + A_2A_{N-1} + \ldots + A_{N-2}A_{N-1}}{(A_N-A_1)\,(A_N-A_2)\,(A_N-A_3)\ldots(A_N-A_{N-1})}$$

$$- \frac{A_1 + A_2 + A_3 + A_4 + \ldots + A_{N-1}}{(A_N-A_1)\,(A_N-A_2)\,(A_N-A_3)\ldots(A_N-A_{N-1})}$$

$$+ \frac{1}{(A_N-A_1)\,(A_N-A_2)\,(A_N-A_3)\ldots(A_N-A_{N-1})}$$

FIGURE 4

Note that all of the denominators in any one column are identical. The numerators are formed by taking various combinations of the values of $A_N, A_{N-1}, A_{N-2}, \ldots, A_0$. The forms for denominators and numerators are shown:

Denominator $\prod_{\substack{r=1 \\ r\neq i}}^{r=N} (A_i - A_r)$

i = column index

π = product of indicated quantities

Numerator $\sum \prod_{\substack{r=1 \\ r\neq i}}^{r=N} A_r$

Terms taken $N-1, N-2, \ldots, 3, 2, 1, 0$ at a time for rows $1, 2, 3, \ldots, N$, respectively.

It might be of interest to consider the number of terms in any numerator. This can be done by finding the number of combinations of $(N-1)$ things taken $(N-j)$ at a time. For example, in a matrix of order 13, the numerator of the sixth row would contain 792 terms:

$$ {}_{N-1}C_{N-j} = \frac{(N-1)!}{(N-j)!\,[(N-1)-(N-j)]!} $$

N = order of matrix

j = row index

The algebraic signs of the terms in the even-numbered rows, counting from the bottom up, are negative.

If $\triangle A$ is constant and positive, where $\triangle A$ is the difference between successive terms of column 2 of the alternant matrix, the denominators can be simplified:

$$ \triangle A = (A_j - A_{j-1}) = \delta $$

$$ D_1 = (-1)^{N+i}\,(N-1)!\;0!\;\delta^{N-1} $$

$$ D_2 = (-1)^{N+i}\,(N-2)!\;1!\;\delta^{N-1} $$

$$ D_3 = (-1)^{N+i}\,(N-3)!\;2!\;\delta^{N-1} $$

$$ \cdot $$

$$ \cdot $$

$$ \cdot $$

$$ D_{N-1} = (-1)^{N+i}\;1!\;(N-2)!\;\delta^{N-1} $$

$$ D_N = (-1)^{N+i}\;0!\;(N-1)!\;\delta^{N-1} $$

$D_1, D_2, D_3, \ldots, D_N$ are denominators of columns $1, 2, 3, \ldots, N$, respectively.

To facilitate the use of IBM punched card machines, the solution was written in a different form. This eliminates finding as many products and combinations of products as in previous forms. The first, second, and Nth columns for the inverse matrix are shown in Figures 5, 6, and 7.

$$ P = A_1 \cdot A_2 \cdot A_3 \cdot A_4 \ldots A_N $$

1ST COLUMN

Row 1: $\dfrac{P}{A_1 D_1}$

Row 2: $\dfrac{P}{A_1 D_1}\left[\dfrac{1}{A_N}+\dfrac{1}{A_{N-1}}+\dfrac{1}{A_{N-2}}+\cdots+\dfrac{1}{A_3}+\dfrac{1}{A_2}\right]$

Row 3: $\dfrac{P}{A_1 D_1}\left[\dfrac{1}{A_{N-1}}\cdot\dfrac{1}{A_N}+\dfrac{1}{A_{N-2}}\left(\dfrac{1}{A_N}+\dfrac{1}{A_{N-1}}\right)+\dfrac{1}{A_{N-3}}\left(\dfrac{1}{A_N}+\dfrac{1}{A_{N-1}}+\dfrac{1}{A_{N-2}}\right)\right.$
$\left.+\cdots+\dfrac{1}{A_2}\left(\dfrac{1}{A_N}+\dfrac{1}{A_{N-1}}+\dfrac{1}{A_{N-2}}+\cdots+\dfrac{1}{A_3}\right)\right]$

Row 4: $\dfrac{P}{A_1 D_1}\left\{\dfrac{1}{A_{N-2}}\cdot\dfrac{1}{A_{N-1}}\cdot\dfrac{1}{A_N}+\dfrac{1}{A_{N-3}}\left[\dfrac{1}{A_{N-1}}\cdot\dfrac{1}{A_N}+\dfrac{1}{A_{N-2}}\left(\dfrac{1}{A_N}+\dfrac{1}{A_{N-1}}\right)\right]\right.$
$+\dfrac{1}{A_{N-4}}\left[\dfrac{1}{A_{N-1}}\cdot\dfrac{1}{A_N}+\dfrac{1}{A_{N-2}}\left(\dfrac{1}{A_N}+\dfrac{1}{A_{N-1}}\right)+\dfrac{1}{A_{N-3}}\left(\dfrac{1}{A_N}+\dfrac{1}{A_{N-1}}+\dfrac{1}{A_{N-2}}\right)\right]$
$+\cdots+\dfrac{1}{A_2}\left[\dfrac{1}{A_{N-1}}\cdot\dfrac{1}{A_N}+\dfrac{1}{A_{N-2}}\left(\dfrac{1}{A_N}+\dfrac{1}{A_{N-1}}\right)+\dfrac{1}{A_{N-3}}\left(\dfrac{1}{A_N}+\dfrac{1}{A_{N-1}}+\dfrac{1}{A_{N-2}}\right)\right.$
$\left.\left.+\cdots+\dfrac{1}{A_3}\left(\dfrac{1}{A_N}+\dfrac{1}{A_{N-1}}+\dfrac{1}{A_{N-2}}+\cdots+\dfrac{1}{A_4}\right)\right]\right\}$

Row N: $\dfrac{P}{A_1 D_1}\left[\dfrac{1}{A_2}\cdot\dfrac{1}{A_3}\cdot\dfrac{1}{A_4}\cdots\dfrac{1}{A_{N-2}}\cdot\dfrac{1}{A_{N-1}}\cdot\dfrac{1}{A_N}\right]=\dfrac{1}{D_1}$

FIGURE 5

2ND COLUMN

$$\text{Row } 1: \frac{P}{A_2 D_2}$$

$$\text{Row } 2: \frac{P}{A_2 D_2}\left[\frac{1}{A_1}+\frac{1}{A_N}+\frac{1}{A_{N-1}}+\cdots+\frac{1}{A_3}\right]$$

$$\text{Row } 3: \frac{P}{A_2 D_2}\left[\frac{1}{A_N}\cdot\frac{1}{A_1}+\frac{1}{A_{N-1}}\left(\frac{1}{A_1}+\frac{1}{A_N}\right)+\frac{1}{A_{N-2}}\left(\frac{1}{A_1}+\frac{1}{A_N}+\frac{1}{A_{N-1}}\right)+\cdots+\frac{1}{A_3}\left(\frac{1}{A_1}+\frac{1}{A_N}+\cdots+\frac{1}{A_4}\right)\right]$$

$$\text{Row } 4: \frac{P}{A_2 D_2}\left\{\frac{1}{A_{N-1}}\cdot\frac{1}{A_N}\cdot\frac{1}{A_1}+\frac{1}{A_{N-2}}\left[\frac{1}{A_N}\cdot\frac{1}{A_1}+\frac{1}{A_{N-1}}\left(\frac{1}{A_1}+\frac{1}{A_N}\right)\right]+\frac{1}{A_{N-3}}\left[\frac{1}{A_N}\cdot\frac{1}{A_1}+\frac{1}{A_{N-1}}\left(\frac{1}{A_1}+\frac{1}{A_N}\right)+\frac{1}{A_{N-2}}\left(\frac{1}{A_1}+\frac{1}{A_N}+\frac{1}{A_{N-1}}\right)\right]+\cdots+\frac{1}{A_3}\left[\frac{1}{A_N}\cdot\frac{1}{A_1}+\frac{1}{A_{N-1}}\left(\frac{1}{A_1}+\frac{1}{A_N}\right)+\frac{1}{A_{N-2}}\left(\frac{1}{A_1}+\frac{1}{A_N}+\frac{1}{A_{N-1}}\right)+\cdots+\frac{1}{A_4}\left(\frac{1}{A_1}+\frac{1}{A_N}+\frac{1}{A_{N-1}}+\cdots+\frac{1}{A_5}\right)\right]\right\}$$

$$\text{Row } N: \frac{P}{A_2 D_2}\left[\frac{1}{A_3}\cdot\frac{1}{A_4}\cdots\frac{1}{A_{N-1}}\cdot\frac{1}{A_N}\cdot\frac{1}{A_1}\right]=\frac{1}{D_2}$$

FIGURE 6

NTH COLUMN

$$\text{Row } 1: \frac{P}{A_N D_N}$$

$$\text{Row } 2: \frac{P}{A_N D_N}\left[\frac{1}{A_{N-1}}+\frac{1}{A_{N-2}}+\frac{1}{A_{N-3}}+\cdots+\frac{1}{A_2}+\frac{1}{A_1}\right]$$

$$\text{Row } 3: \frac{P}{A_N D_N}\left[\frac{1}{A_{N-2}}\cdot\frac{1}{A_{N-1}}+\frac{1}{A_{N-3}}\left(\frac{1}{A_{N-1}}+\frac{1}{A_{N-2}}\right)+\frac{1}{A_{N-4}}\left(\frac{1}{A_{N-1}}+\frac{1}{A_{N-2}}+\frac{1}{A_{N-3}}\right)+\cdots+\frac{1}{A_1}\left(\frac{1}{A_{N-1}}+\frac{1}{A_{N-2}}+\frac{1}{A_{N-3}}+\cdots+\frac{1}{A_3}+\frac{1}{A_2}\right)\right]$$

$$\text{Row } 4: \frac{P}{A_N D_N}\left\{\frac{1}{A_{N-3}}\cdot\frac{1}{A_{N-2}}\cdot\frac{1}{A_{N-1}}+\frac{1}{A_{N-4}}\left[\frac{1}{A_{N-2}}\cdot\frac{1}{A_{N-1}}+\frac{1}{A_{N-3}}\left(\frac{1}{A_{N-1}}+\frac{1}{A_{N-2}}\right)\right]+\frac{1}{A_{N-5}}\left[\frac{1}{A_{N-2}}\cdot\frac{1}{A_{N-1}}+\frac{1}{A_{N-3}}\left(\frac{1}{A_{N-1}}+\frac{1}{A_{N-2}}\right)+\frac{1}{A_{N-4}}\left(\frac{1}{A_{N-1}}+\frac{1}{A_{N-2}}+\frac{1}{A_{N-3}}\right)\right]+\cdots+\frac{1}{A_1}\left[\frac{1}{A_{N-2}}\cdot\frac{1}{A_{N-1}}+\frac{1}{A_{N-3}}\left(\frac{1}{A_{N-1}}+\frac{1}{A_{N-2}}\right)+\frac{1}{A_{N-4}}\left(\frac{1}{A_{N-1}}+\frac{1}{A_{N-2}}+\frac{1}{A_{N-3}}\right)+\cdots+\frac{1}{A_2}\left(\frac{1}{A_{N-1}}+\frac{1}{A_{N-2}}+\frac{1}{A_{N-3}}+\cdots+\frac{1}{A_3}\right)\right]\right\}$$

$$\text{Row } N: \frac{P}{A_N D_N}\left[\frac{1}{A_1}\cdot\frac{1}{A_2}\cdot\frac{1}{A_3}\cdots\frac{1}{A_{N-3}}\cdot\frac{1}{A_{N-2}}\cdot\frac{1}{A_{N-1}}\right]=\frac{1}{D_N}$$

FIGURE 7

$P = A_1A_2A_3 \ldots A_N$ $D_1, D_2, \ldots, D_N$ are the denominators of columns 1 to N, respectively.

For actual calculation procedure the values of P,

$D_1, D_2, D_3, \ldots, D_N$

$1/A_1, 1/A_2, 1/A_3, \ldots, 1/A_N$

$P/A_1D_1, P/A_2D_2, \ldots, P/A_ND_N$ are calculated first. Note that the last values are the elements of the first row of the reciprocal matrix.

N decks of cards are made up, each containing N cards. The decks are made up from a circular arrangement of the values of $1/A_N$, $1/A_{N-1}$, ..., $1/A_1$. The last card in each deck is $1/A_i$ replaced by P/A_iD_i. These cards also contain the values offset gang punched on the following cards, as shown below.

1ST COLUMN

$\frac{1}{A_N}$

$1/A_{N-1}$ $1/A_N$ $1/A_{N-1}\,(1/A_N)$

$1/A_{N-2}$ $1/A_{N-1}$ $1/A_{N-2}\,(1/A_N + 1/A_{N-1})$

$1/A_{N-3}$ $1/A_{N-2}$ $1/A_{N-3}\,(1/A_N + 1/A_{N-1} + 1/A_{N-2})$

. .

. .

. .

$1/A_2$ $1/A_3$

P/A_1D_1 $1/A_2$ $P/A_1D_1\,(1/A_N + 1/A_{N-1} + \ldots + 1/A_2)$

2ND COLUMN	3RD COLUMN	NTH COLUMN
$1/A_1$	$1/A_2$	$1/A_{N-1}$
$1/A_N$	$1/A_1$	$1/A_{N-2}$
$1/A_{N-1}$	$1/A_N$	$1/A_{N-3}$
.	$1/A_{N-1}$	.
.	.	.
.	.	.
$1/A_3$	$1/A_4$	$1/A_1$
P/A_2D_2	P/A_3D_3	P/A_ND_N

The first group of values is used as multiplier while the second group is used as multiplicand after being progressively accumulated. The product for first column is shown below. The last card in the deck contains the element of the second row.

For each successive row, the first group of values is reproduced, and the products are offset gang punched on following card. This process is done $N - 1$ times. Note that each time the process is completed, the number of cards in each deck with products other than 0 decreases one, until the decks for the last row, the only card containing a product, will be the last card which is the P/A_ND_N card.

1ST COLUMN

$1/A_N$	———
$1/A_{N-1}$	———
$1/A_{N-2}$	$1/A_{N-1}\left(\frac{1}{A_N}\right)$
$1/A_{N-3}$	$1/A_{N-2}\left(\frac{1}{A_N}+\frac{1}{A_{N-1}}\right)$
$1/A_{N-4}$	$1/A_{N-3}\left(\frac{1}{A_N}+\frac{1}{A_{N-1}}+\frac{1}{A_{N-2}}\right)$
.	.
.	.
.	.
.	.
$1/A_2$	$1/A_3\left(\frac{1}{A_N}+\frac{1}{A_{N-1}}+\ldots+\frac{1}{A_4}\right)$
P/A_1D_1	$1/A_2\left(\frac{1}{A_N}+\frac{1}{A_{N-1}}+\ldots+\frac{1}{A_3}\right)$

REFERENCE

1. A. C. AITKEN, *Determinants and Matrices* (London: Oliver, 1945).

Matrix Multiplication on the IBM Card-Programmed Electronic Calculator

JOHN P. KELLY

Carbide and Carbon Chemicals Corporation

THE ONLY TYPE of calculation to be considered here is matrix multiplication. Time has not permitted any concrete work to be done on matrix inversion, but I shall have a few comments to offer on this later. In matrix work there are only simple arithmetic operations. This multiplication demands only two steps:

Operation 1: This is a shifting from electronic storage A (FS 1-2) into the electronic counter. For this problem, channel A is permanently connected to FS 1-2 (assigned) and channel B to FS 3-4 (assigned).

Operation 2: This is an 8-digit by 8-digit multiplication with the results in the electronic counter.

The problem to be considered is the following matrix multiplication:

$$AB = \begin{pmatrix} a_{11} & a_{12} \cdots \\ a_{21} & a_{22} \cdots \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} \cdots \\ b_{21} & b_{22} \cdots \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \end{pmatrix} = \begin{pmatrix} c_{11} & c_{12} \cdots \\ c_{21} & c_{22} \cdots \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \end{pmatrix} = C$$

The general term of the product matrix takes on the following form:[1]

$$c_{ij} = \sum_k a_{ik}\, b_{kj}$$

that is, each term is the scalar product of a row and column vector. There are two practical methods of calculation. One is to perform all the multiplications involving one row of the left-hand matrix; this generates an entire row of the product matrix. The other is to perform all the multiplications involving one column of the right-hand matrix, thus obtaining an entire column of the product matrix. The latter has been chosen for reasons which will become obvious.

Calculations

The machine is loaded with the elements of one column of the right-hand matrix, the instructions for which will be found in Table I. The zero instructions for channels A and B indicate card reading. In Operation 1, as mentioned above, the quantity in FS 1-2 is shifted to the counter. The instruction for channel C is the code number for the storage register or accounting machine counter in which the indicated item will be held. The problem considered here is simple enough so that no shift is required. The elements of rows 1-5 are stored in the accounting machine counters, using an adding code of 7.

The entire bank 1 and all but one register in bank 2 of the mechanical storage (941) has been used, in addition to 5 accounting machine counters. Of the two remaining counters, number 1 is too small and number 2 will be used to accumulate totals. Thus, the method to be used here can be used for multiplication when the right matrix contains 21 or fewer rows. For larger matrices, elements containing fewer digits might be used with the registers split.

Table II shows the general layout for the deck of cards to be used, and a detailed description of the clearing cards.

Table III contains the description of one row of the left-hand matrix A. Each element of this matrix is card read over channel A. The corresponding element from the right-hand matrix B is called from storage on channel B. The operation 2 is multiplication, with the 72 in channel C adding the product in counter group 2.

It should be pointed out that, with the exception of the clearing cards, only one row and one column have to be programmed. All other rows or columns, as the case may be, take on the same form.

The clearing cards can be eliminated by changing the channel B coding of the last row of the A matrix to counter read out and reset (8), instead of just read out (7). The deck to be run through consists of column 1 of matrix B and all of matrix A; column 1 of the product matrix C will be obtained. The row number of matrix A is used as a control field to clear counter 2. The products that make up C_{11} will be listed; however, all other elements will be tabulated.

THE STANDARD methods of matrix inversion—any of the elimination methods to a triangle or to a diagonal directly—involve approximately eight to sixteen thousand intermediate results, depending on whether you carry the unit matrix in your calculation.

Let's take for granted that we do not carry it along. We will just carry along the 400 elements of the 20 × 20 matrix in the inversion. Even this involves 8,000 summary punchings, which take in the neighborhood of 1.5 seconds each as compared to a 0.4 second card cycle. The obvious way is to avoid taking intermediate results out as much as possible.

One way of accomplishing this is through the use of the enlargement method. To review this: one starts with the inverse of the upper left-hand element, which can be enlarged to the inverse of the upper 2 × 2 matrix through simple multiplications, additions, and subtractions. It is a function of the inverse of the single element, the additional column, row, and diagonal element as shown below:

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}^{-1} = f(a_{11}^{-1}, a_{12}, a_{21}, a_{22}) .$$

This enlargement then proceeds, increasing the matrix by 1 row and 1 column in each step. Another alternative is the use of second order enlargement, which increases the order by 2 in each step. This general method appears to have several advantages; summary punching and machine card passes are reduced considerably, and the problem of significant digits can be avoided by checking the intermediate inverses and iterating them for greater accuracy if necessary.

TABLE I

RIGHT-HAND MATRIX COLUMN CODE

Row	*Column*	*A*	*Oper.*	*B*	*C*	*A-Entry*
1	*j*	00	1	00	73	$b_{1,j}$
2					74	$b_{2,j}$
3					75	$b_{3,j}$
4					76	$b_{4,j}$
5					77	$b_{5,j}$
6					11	$b_{6,j}$
7					12	$b_{7,j}$
8					13	$b_{8,j}$
9					14	$b_{9,j}$
10					15	$b_{10,j}$
11					16	$b_{11,j}$
12					17	$b_{12,j}$
13					18	$b_{13,j}$
14					21	$b_{14,j}$
15					22	$b_{15,j}$
16					23	$b_{16,j}$
17					24	$b_{17,j}$
18					25	$b_{18,j}$
19					26	$b_{19,j}$
20					27	$b_{20,j}$

All blank spaces indicate the same entry as that used in first line.

TABLE II

CLEARING CARDS

Card No.	*A*	*Oper.*	*B*	*C**
1	81			00
2	82			
3	83			
4	84			
5	85			
6	86			
7	87			

*This first card channel C clears the 604 electronic counter.

GENERAL DECK

(Obtains *j*th column of product matrix)

1. Seven clearing cards
2. *j*th column of B (20 Cards)
3. Matrix A in row order (400 Cards)
4.* 2 blank cards

*These are necessary to print the list result when an intermediate control break is used.

TABLE III

LEFT-HAND MATRIX ROW CODE

Row	*Column*	*A*	*Oper.*	*B*	*C*	*A-Entry*
i	1	00	2	73	72	$a_{i,1}$
	2			74		$a_{i,2}$
	3			75		$a_{i,3}$
	4			76		$a_{i,4}$
	5			77		$a_{i,5}$
	6			11		$a_{i,6}$
	7			12		$a_{i,7}$
	8			13		$a_{i,8}$
	9			14		$a_{i,9}$
	10			15		$a_{i,10}$
	11			16		$a_{i,11}$
	12			17		$a_{i,12}$
	13			18		$a_{i,13}$
	14			21		$a_{i,14}$
	15			22		$a_{i,15}$
	16			23		$a_{i,16}$
	17			24		$a_{i,17}$
	18			25		$a_{i,18}$
	19			26		$a_{i,19}$
	20			27		$a_{i,20}$

All blank spaces indicate the same entry as that used in first line.

REFERENCE

1. KAISER S. KUNZ, "Matrix Methods," pages 37-42.

*Machine Methods for Finding Characteristic Roots of a Matrix**

FRANZ L. ALT

Computation Laboratory, National Bureau of Standards

THE PURPOSE of this paper is to describe a few expedients which can be applied to computation of characteristic roots of matrices by means of punched card machines. In the course of two problems of this kind, recently handled by the Computation Laboratory of the National Bureau of Standards, some of these methods or variants of methods were actually tried out on cards, and some others were considered and laid out without actually being carried through. In both cases the general type of method used was suggested by the originator of the problem.

THE FIRST of these examples was of the conventional type: given a matrix A of order n (in the example, $n = 14$), with elements a_{ik} representing the approximate distribution of elastic forces in an idealized airplane wing, to find the three characteristic roots with greatest moduli. For finding the dominant root (i.e., the one with greatest modulus) there is the common method of starting with a trial vector $y^{(0)} = (y_1^{(0)}, y_2^{(0)}, \ldots, y_n^{(0)})$, and multiplying it repeatedly by the given matrix. Thus, $y^{(k)} = Ay^{(k-1)}$. This method is described, e.g., by Frazer, Duncan and Collar.[1] It is an excellent method for punched card machines, since the multiplication of a vector by a matrix can be carried out very simply. One minor trouble that arises is that after repeated multiplication the numbers fall outside the range of decimal digits which had been allotted to them on the machine. This is prevented by "norming" the vector after each multiplication by the matrix. We accomplished the norming by making the last component of the vector equal to 1 after each step. Thus (with a slight change in notation) we set

$$\bar{y}^{(k)} = Ay^{(k-1)}\,;\, y^{(k)} = \frac{1}{\bar{y}_n^{(k)}} \cdot \bar{y}^{(k)}\,.$$

This method would fail in case the matrix were such that the last component of the characteristic vector happens to be very small compared to the other components. Obviously, other norming methods could be used which avoid this failure. However, it seems preferable to have a simple method, which fails once in a hundred cases but saves work in the remaining 99 cases. With this norming convention, the factors $\bar{y}_n^{(k)}$ converge to the dominant characteristic root, and the vectors $y^{(k)}$ to a corresponding characteristic vector.

The computations were performed on the 602-A calculator. The 602 or 604 would have been equally suitable, since there is no great amount of number storage. A machine with two card feeds, such as the Aberdeen IBM Relay Calculators, would have been superior, because in this case it would have been possible to feed the matrix cards into one card feed and the vector cards into another. Since we had no such machine available in our laboratory, we proceeded as follows:

The matrix elements are punched into a deck of cards, one element to a card. This deck is reproduced as many times as we expect to have iterations. Before starting any one iteration, one of these decks is prefaced by a small deck containing the latest approximation to the characteristic vector (in the case of the first deck, the chosen trial vector $y^{(0)}$), the combined deck is sorted by columns, the vector elements are gang punched into the matrix cards, then the deck is followed by a set of summary cards, sorted by rows, and put through the 602-A for performing the matrix multiplication. This operation produces the unnormed, new approximation to the characteristic vector, punched into the summary cards. These are then sorted out and put through the 602-A again for the norming process.

To obtain the second characteristic root, the method of "sweeping-out" the first characteristic vector is used. That is to say, proceed exactly as for the first root, but after each iteration subtract from the iterative vector $y^{(k)}$ a certain multiple of the first characteristic vector. The same process can be carried out for subsequent characteristic roots. In these cases it is desirable to punch each component of each iterative vector in several successive summary cards, one for each of the previous characteristic vectors to be swept out.

In the actual example carried out in our case, there were additional computing steps brought about as a result of the

*The preparation of this report was sponsored by the Office of Air Research, USAF.

fact that the equations of motion of the airplane wing were referred to a moving coordinate system. This requires an adjustment after each iteration; the computation is similar to the sweeping-out of earlier characteristic vectors.

The vectors converge reasonably well, except in cases where there are two characteristic roots with equal or almost equal moduli. We did not run into any such cases. Nevertheless, we felt it useful to speed up the convergence of the process. A method which we used for this purpose is the one described by A. C. Aitken[2] and called by him "the delta-square process." It consists in taking three successive approximations to the desired characteristic root, say, v_{t-1}, v_t, and v_{t+1}, and extrapolating from them to the desired root by using the expression

$$\frac{v_{t+1}\, v_{t\ 1} - v_t^2}{v_{t+1} - 2v_t + v_{t-1}}\ .$$

The same method can be applied to find directly a close approximation to the characteristic vector.

Another method, which we discussed but have not yet used, consists in subtracting from all terms of the principal diagonal of the matrix a suitable constant, chosen in such a way as to increase the ratio between the dominant and subdominant characteristic root. (The subdominant root is the one with second-largest modulus.) Suppose, for example, that it is known that all roots are real (as, for instance, in the case of symmetric matrices with real coefficients), and the largest root is estimated to be 10, the second largest 9, and the smallest 1. By subtracting 5 from all elements of the principal diagonal of the matrix, a matrix is obtained whose characteristic roots are smaller by 5 than those of the original matrix; that is, the largest root has become about 5, the second largest 4 and the smallest -4. The ratio of largest to second-largest, in absolute value, is now 5:4, whereas previously it was 10:9. Since the speed of convergence of the iteration process tends to increase with the size of this ratio, the process is likely to converge faster for the modified matrix. In general, the constant to be subtracted, in case all roots are real, is the arithmetic mean between the root nearest the dominant root and the one farthest away from it. It is necessary, of course, to have estimates of these roots in order to apply this method.

ONE VERY OFTEN encounters matrices which might be called "almost triangular." The name "triangular" shall be applied to matrices in which all elements above the principal diagonal are zero. By "almost triangular" is meant a matrix which has only a few nonzero elements above the principal diagonal, and those are all bunched close to the diagonal. To be exact, an nth order matrix A with elements a_{ik} will be called "almost triangular of degree t" if $a_{ik} = 0$ for $k - i > t$, where t is some integer between 0 and $n - 1$. There is no restriction on the elements below the principal diagonal. However, some of the statements which will be made toward the end of this paper apply only to "almost diagonal" matrices, which are defined analogously by $a_{ik} = 0$ for $|k - i| > t$; that is to say, both above and below the principal diagonal all elements except those close to the diagonal are zero.

For a completely triangular matrix, that is, $t = 0$, no computation is required. The characteristic roots are equal to the elements of the principal diagonal.

Now take the case $t = 1$. Consider the matrix $A - \lambda I$ as the matrix of a system of simultaneous homogeneous equations.

$$\begin{aligned}
(a_{11} - \lambda)\, x_1 + \quad & a_{12}\, x_2 & = 0 \\
a_{21}\, x_1 + (a_{22} - \lambda)\, x_2 + \quad & a_{23}\, x_3 & = 0 \\
\cdot \quad \cdot \quad \cdot \quad & \cdot \quad \cdot \quad \cdot \quad \cdot & \\
a_{n1}\, x_1 + \ldots \quad & + (a_{nn} - \lambda)\, x_n & = 0
\end{aligned}$$

Our problem is to find those values of λ for which the system has a solution. Because of the "almost triangular" character of the matrix, the first equation contains only the first two unknowns, the second equation only the first three unknowns, generally the kth equation, only the first $k + 1$ unknowns. For simplicity of presentation, let us assume first that $a_{ik} \neq 0$ for $k - i = 1$. Let us choose a particular value of λ and ask ourselves whether the system of linear homogeneous equations has a non-trivial solution for this particular λ, that is, a solution in which not all of the unknowns are equal to 0. It can easily be seen that because of our assumption that $a_{i,\,i+1} \neq 0$, the value of the first unknown in any non-trivial solution is not zero. And since the system is homogeneous, an arbitrary nonzero value can be assigned to x_1, for example $x_1 = 1$. Now substitute $x_1 = 1$ in the first equation and obtain x_2, then substitute x_1 and x_2 in the second equation and obtain x_3, etc., down to the $(n-1)$st equation from which the value of x_n is obtained. If, now, all these values $x_1, x_2, \ldots, x_n$ are substituted into the nth equation, this equation may or may not be satisfied. The result of the substitution in the left-hand side of the equation is, of course, a function of the particular value of λ chosen initially, and it may be designated by $E(\lambda)$. If, and only if, $E(\lambda) = 0$, λ is one of the characteristic roots of the matrix. Now, the value of $E(\lambda)$ for a number of different λ's may be computed, until enough values of the function $E(\lambda)$ are obtained to determine its zeros, either graphically or by inverse interpolation or some other method.

This method of obtaining characteristic roots was described by Myklestad[3,4] and Prohl.[5] They described, independently of each other, the application to two different engineering problems, but apparently without noticing its general applicability and import. To Dr. A. Gleyzal goes the credit for having noticed this and for having generalized the method to cases of $t > 1$.

If these substitutions are carried out for a number of different values of λ, let us see how the values of the unknowns x_1, x_2, etc., depend on λ. Of course, x_1 is chosen arbitrarily, and it does not matter how it is chosen, as long as the same x_1 is substituted with each value of λ. For the

sake of definiteness, let $x_1 = 1$. It is seen at once that x_2 is a linear function of λ; specifically, $x_2 = -(a_{11} - \lambda)/a_{12}$. Similarly, it is seen, by substitution, that x_3 is a quadratic function of λ, and generally that x_k is a polynomial in λ of degree $k - 1$. This property is of very great assistance in checking the computations. It is also evident that $E(\lambda)$ is a polynomial of degree n and, as such, has n zeros. (These, of course, coincide with the n characteristic roots of our matrix.) Theoretically, therefore, it is sufficient to compute $E(\lambda)$ for $n + 1$ values of λ. Thereafter the polynomial can be extrapolated indefinitely by differencing. In fact, however, one might lose too many significant digits in this process. It is, therefore, preferable to compute $E(\lambda)$ for more than n values of λ, say, something like $2n$ values, and to distribute these as far as possible equidistantly over an interval which spans all expected values of λ. I might also mention that, in the example with which we were concerned, the question was not to find all characteristic values, nor to find the largest or smallest in absolute value, but to find those characteristic values which are located in a given interval. For this type of problem this method is particularly well suited.

Sometimes one finds a matrix which is not almost-triangular as it stands, but which can be brought into almost-triangular form by suitable rearrangement of rows and columns. (It is necessary to apply the same permutation to both rows and columns in order to preserve the characteristic roots.) In the case of any matrix with many zeros, it is worth looking for this possibility.

Let us, now, drop the requirement that all $a_{i,\,i+1}$ be different from zero and assume that one of these elements, say $a_{j,\,j+1}$ is equal to 0. In this case we can obtain the characteristic roots of the matrix A by partitioning it into a matrix A_1 of order j, consisting of the elements in the upper left-hand corner of A, and a matrix A_2 of order $n - j$, consisting of the elements in the lower right-hand corner of A. Each root of A is a root of either A_1 or A_2, and vice versa. The roots of A_1 and A_2 can be found by the method described above. If there are several zeros among the elements just above the principal diagonal, the matrix A is partitioned into a correspondingly larger number of submatrices.

We now turn to the case $t > 1$. The essential features of the method can be fully explained for $t = 2$. Generalization for greater values of t will be self-evident. The basic idea is again to attempt to solve the homogeneous linear system of equations for a number of particular values of λ.

The system now has the form

$$\begin{aligned}
(a_{11} - \lambda)\,x_1 + \quad & a_{12}\,x_2 + a_{13}\,x_3 & = 0 \\
a_{21}\,x_1 + (a_{22} - \lambda)\,x_2 + \quad & a_{23}\,x_3 + a_{24}\,x_4 & = 0 \\
\cdot \quad \cdot \quad \cdot \quad & \cdot \quad \cdot \quad \cdot \quad \cdot & \cdot \quad (1) \\
a_{n-2,\,1}\,x_1 + \ldots \quad & + a_{n-2,\,n}\,x_n & = 0 \\
a_{n-1,\,1}\,x_1 + \ldots \quad & + a_{n-1,\,n}\,x_n & = 0 \\
a_{n1}\,x_1 + \ldots \quad & + (a_{nn} - \lambda)\,x_n & = 0
\end{aligned}$$

As before, consider first the case in which all $a_{i,\,i+2}$ are different from zero. For any given λ, in the case $t = 1$ it was necessary to start with an assumed value of x_1. For $t = 2$, it is necessary to start with assumed values for both x_1 and x_2. For example, start with $x_1 = 1$, $x_2 = 0$ and obtain successively from the linear equations the values of $x_3, x_4, \ldots, x_n$. By this time the first $n - 2$ equations have been used. Now all the x's can be substituted into the last two equations. This gives a value $e_1(\lambda)$ for the left-hand side of the $(n-1)$st equation and a value $f_1(\lambda)$ for the left-hand side of the nth equation.

We now repeat the process with two different initial values for x_1 and x_2, say, $x_1 = 0$, $x_2 = 1$. For the left-hand side of the two last equations two new values $e_2(\lambda)$ and $f_2(\lambda)$ are obtained. In general, all four values e_1, e_2, f_1, and f_2 will be different from zero. To find the characteristic values of the matrix, notice the following: any non-trivial solution of a system of equations will start with certain two values $x_1 = a_1$, $x_2 = a_2$. (If all $a_{i,\,i+2} \neq 0$, then x_1 and x_2 do not vanish simultaneously.) The values of the remaining unknowns in such a solution are expressible as linear combinations of the unknowns, obtained in the two basic solutions before, with weights a_1 and a_2. Substitution of these unknowns in the last two equations gives two values $e(\lambda) = a_1\,e_1(\lambda) + a_2\,e_2(\lambda)$, $f(\lambda) = a_1\,f_1(\lambda) + a_2\,f_2(\lambda)$. For a non-trivial solution, $e(\lambda)$ and $f(\lambda)$ must vanish simultaneously. For this to be possible it is necessary (and sufficient) that the determinant

$$D(\lambda) = \begin{vmatrix} e_1(\lambda) & e_2(\lambda) \\ f_1(\lambda) & f_2(\lambda) \end{vmatrix}$$

be equal to 0. This determinant is a function of λ whose zeros coincide with the characteristic roots of the matrix. Just as before, we select a number of values of λ and evaluate the determinant $D(\lambda)$ for each of them and then find the roots of $D(\lambda)$, either graphically or by interpolation.

To summarize, the steps taken are as follows:

Assume	$x_1^{(1)} = 1$,	$x_1^{(2)} = 0$,	$x_1 = a_1$
Assume	$x_2^{(1)} = 0$,	$x_2^{(2)} = 1$,	$x_2 = a_2$
Solve eq. 1 of (1)	$x_3^{(1)}$	$x_3^{(2)}$	$x_3 = a_1 x_3^{(1)} + a_2 x_3^{(2)}$
Solve eq. 2 of (1)	$x_4^{(1)}$	$x_4^{(2)}$	$x_4 = a_1 x_4^{(1)} + a_2 x_4^{(2)}$
. . .	. . .	. . .	. . .
Solve eq. $n-2$ of (1)	$x_n^{(1)}$	$x_n^{(2)}$	$x_n = a_1 x_n^{(1)} + a_2 x_n^{(2)}$
Substitute in eq. $n-1$ of (1)	e_1	e_2	$e = a_1 e_1 + a_2 e_2$
Substitute in eq. n of (1)	f_1	f_2	$f = a_1 f_1 + a_2 f_2$

The terms in the last column above need not be computed; they are listed here only to illustrate the explanation. The values of $x_k^{(1)}$ and $x_k^{(2)}$, considered as functions of λ, are polynomials, so that their computation can be checked

by differencing. Likewise the determinant $D(\lambda)$ is a polynomial. It is possible to prove directly that $D(\lambda)$ is of degree n in λ. This proof is not given here, since it is not needed for the argument.

Cases in which one or more of the coefficients $a_{i,\,i+2}$ vanish need to be treated specially. It seems most economical not to complicate the machine method of computation by allowing for these degenerate cases, but rather to treat these cases separately when they arise.

In the general case of any value of t, the determinant $D(\lambda)$ will be of order t. Considered as a function of λ, it is always a polynomial of degree n, regardless of its order.

Since the evaluation of the determinants of higher order is laborious, the method given here is recommended primarily for the cases of $t = 1, 2$, or 3. These are just the cases which are most likely to occur in practice.

The performance of computations on punched card machines is straightforward. The coefficients of the matrix are punched into cards, one coefficient to a card. From this deck, a series of decks is prepared, so that t decks are available for each value of λ to be used. All of these are identical with the original deck, except that the value of λ has been subtracted from the numbers in the principal diagonal. The cards containing the coefficients $a_{i,\,i+t}$ are characterized, by a special punch, as summary cards. It is expedient, but not necessary, to divide each row by $a_{i,\,i+t}$, so as to make the latter coefficient equal to unity. No cards are required for coefficients which are equal to zero. The computations proceed in a number of identical steps, one step for each unknown x_j. We are going to describe one of these steps. Suppose that the value of x_j has just been computed, by using the cards of row $j - t$. There is one such x_j for each deck of matrix cards, i.e., for each value of λ and each combination of assumed values $x_1, \ldots, x_t$. Each x_j is punched into the corresponding summary card.

Now sort all matrix cards on the column number, selecting each jth column. Automatically, for each deck, the summary card is in front and is followed by the remaining cards of the jth column. Feed the cards into the multiplier (either the 602, 602-A or the 604 could be used), use the value of x_j, as read from the summary card, as a group multiplier, and punch the products $a_{ij}x_j$ into the card corresponding to a_{ij}. (Alternatively, it would have been possible to use the reproducer instead of the multiplier and to gang punch the values x_j themselves instead of the product.)

Next, select the row $j - t + 1$. In this row each card except the summary card has a product, $a_{ik}x_k$ (or in the alternative procedure a value x_k), previously punched into it. This is so because the cards of the jth column have been punched in the preceding step, the cards of earlier columns have been punched in earlier steps, the cards in column $j + 1$ are the summary cards in this row, and cards for columns following $j + 1$ do not exist in this row, since all corresponding coefficients are 0.

Now feed the cards into the machine and add all the products (in the alternative procedure, the products are formed in this step and added at the same time). When the machine reaches a summary card, it punches the sum of all products. This is the value of x_{j+1}. Then select all these summary cards, place them in front of the matrix decks, discard all other cards of row $j - t + 1$, and from here on this sequence of operations is repeated.

The polynomials $e(\lambda)$ and $f(\lambda)$ are evaluated in the same way. The fact that x_j is a polynomial in λ can be used conveniently for checking the computations by taking differences of sufficiently high order. Finally, if $t > 1$, the determinants of order t have to be evaluated, either manually or by machine, depending on how many there are. This in turn depends on the order of the matrix and the size of the interval being searched for characteristic roots.

A considerable gain in efficiency over this method can be accomplished in the important special case of almost-diagonal matrices, for moderate size of t. In this case each row of the matrix contains at most $2t + 1$ nonzero elements. All sorting of cards is eliminated, the entire computation is performed in a single run through the 602-A multiplier. Where formerly the computation for a particular value of j was carried out in succession for all λ's before going on to the next j, in this case all cards pertaining to one coefficient deck (i.e., to the same λ and to the same choice of $x_1, \ldots, x_t$), are kept together, arranged by rows. At each step of the substitution, not more than $2t + 1$ different unknowns x_i are needed. These are all stored in the machine, the cards of a row are fed in, the coefficients a_{ik} read off the cards, and multiplied by the corresponding x_i, and the products accumulated and punched into the summary card of the row. Thereafter, the first of the stored x's is discarded, and each subsequent x is moved to the storage location of the preceding one. The last storage location is filled with the x which has just been computed. Now the machine is ready to receive the cards of the next row, and the whole process is carried out without ever stopping the machine. In our work so far, this method has been planned but not yet tried out on cards.

REFERENCES

1. R. A. Frazer, W. J. Duncan, and A. R. Collar, *Elementary Matrices* (Cambridge Univ. Press, 1938), see espec. pp. 134 and 140-141.
2. A. C. Aitken, "Studies in Practical Mathematics," *Proc. Roy. Soc. Edin. 57* (1936-37).
3. N. O. Myklestad, "New Method of Calculating Natural Modes of Uncoupled Bending Vibration of Airplane Wings and Other Types of Beams," *Jour. Aeronaut. Soc. 11,* no. 2, (Apr., 1944), pp. 153 to 162.
4. N. O. Myklestad, "New Method of Calculating Natural Modes of Coupled Bending-Torsion Vibration of Beams," *Trans. Am. Soc. Mech. Engrs. 67,* no. 1 (Jan., 1945), pp. 61 to 67.
5. M. A. Prohl, "A General Method for Calculating Critical Speeds of Flexible Rotors," *Jour. Appl. Mech.* (Sept., 1945), pp. 142 to 148.

DISCUSSION

Mr. Kimball: Dr. Alt mentioned that they had only one experience in the iterative multiplication. In 1945, using a complex node matrix of size 14 by 14, we obtained convergence to four digits in 33 steps of iteration, about half an hour for each step, using the 601 multiplier.

Mr. Bell: Concerning practical problems that arise in evaluating the formulas that were derived: if you have a system of simultaneous equations and reduce the first to a triangular matrix and then by back substitution to a diagonal matrix, you have essentially two procedures, and this complicates the machine work.

We have found that there is a critical point beyond which you would consider the back solution and that the order of that matrix is quite high. I would say something like the 15th order, at least. The advent of the 604 has made the straightforward approach much simpler. In other words, instead of working on the first column and then eliminating it, leave it in the matrix.

Another point is that if you divide and make your elements 1 immediately, you are dividing by numbers whose size may be quite small, and that may make the size of the numbers go outside the limits of your field.

We have found that a method which protects us in that respect is to leave the numbers as a number. Your equation is then of a form where you subtract from each element a ratio multiplied by a number, and then the correction tends to be small, if the dividing term is large, which will keep your numbers within size.

We have done some work with the iterative methods without a great deal of success. We have found that the conditions of convergence are more difficult to determine than just going straight into the problem and trying to get a solution.

In what we have done practically, in trying an iterative process, we have set it up and started it running. If we don't get solutions, if it begins to diverge, we stop, assuming that it is divergent.

It seems to me that essentially those processes are designed where you do not have a machine that is capable of a grinding operation, such as the IBM machines. So that we almost always set the problems up for a direct solution.

One other thing is that in problems of the form where you have a matrix that is symmetrical on both sides, and other special matrix forms where there are mathematical techniques that will give you much fewer operations, it means that you must have different procedures and different methods for your operators, and that always slows you down. We have aimed to do as much of our matrix work by this one simple process as possible; and, although the number of mathematical operations can be unnecessarily large, the elapsed time is very much reduced, rather than trying to be elegant at every step.

Chairman Hurd: A very good point.

Solution of Simultaneous Linear Algebraic Equations Using the IBM Type 604 Electronic Calculating Punch

JOHN LOWE

Douglas Aircraft Company, Incorporated

MANY METHODS exist for solving simultaneous equations with punched card accounting machines. The one presented here takes advantage of the speed and flexibility of the 604 electronic calculator. A 10th order matrix can be inverted in one hour by use of this method, which compares with approximately eight hours through use of relay multipliers. Furthermore, the method is extremely simple.[a]

The basic reduction cycle consists of: sort (650 cards per minute), reproduce (100 cards per minute), sort (650 cards per minute), and calculate (100 cards per minute). This cycle must be repeated a number of times equal to the number of equations.

Theory

Several variations of the basic elimination method can be used with the machine procedure outlined. The one described requires no back solution and is well suited to machine methods. It is well known and will be described very briefly.

The equations may be expressed in matrix notation as $AX = C$. C and X may have, of course, any number of columns. If A^{-1} is desired, C becomes I and X becomes A^{-1} (see reference 1).

The object of the calculation is to operate on the matrices A and C, considered as equations, so as to reduce A to a unit matrix, thus reducing C to X.

Let M be the augmented matrix composed of A and C. Choose any row, k, of M and form M' such that

$$m'_{kj} = \frac{m_{kj}}{m_{kk}}$$

$$m'_{ij} = m_{ij} - m_{ik}\frac{m_{kj}}{m_{kk}}, i \neq k \ .$$

The kth column of M' is zero, excepting the kth row which is unity. Therefore, no cards are made for the kth column of M'.

Form M'' from M' using the above equations, but a different row for k. If this process is repeated until each row has been used and all the columns of A eliminated, the columns of C will have been reduced to X.

[a]The value of the determinant of the matrix of coefficients can be obtained as a by-product of the process. See reference 1.

For best accuracy and to insure that all numbers stay within bounds, the elements of M should be close to unity, and, if possible, the principal diagonal elements of A should be larger than the other elements.

A column of check sums (negative sums of each row) appended to M provides an easy and complete check on the work. These check sums can be calculated by machine, but if they are manually calculated and written as a part of M they provide an excellent check on the key punching. Also, experience has shown that the agreement of the final check sums with X is an index to the accuracy of X.

Machine Procedure

Layout

The following card fields are necessary:

A. Row (of M)
B. Column (of M)
C. Order (initially n and reduced by one each cycle until it has become zero)
D. Common 12 punch
E. Pivotal column 11 punch
F. Pivotal row 12 punch
G. Next pivotal row 12 punch
H. Product or quotient
I. Cross-foot or dividend
J. Multiplicand or divisor

Procedure, Using Rows in Order

1. Start with M punched in fields (A), (B), (C) and amounts in (H).
2. Sort to column. Emit 11 in (E) of column 1.
3. Place column 1 in front and sort to row. Emit 12 in (G) of row 1.
4. Reproduce cards. Emit 12 in (D) of all cards. Reproduce (A) to (A), (B) to (B), and (G) to (F). Reproduce (H) to (I) except that (H) of the pivotal column cards is gang punched in (J). Emit 11 in (E) of the first card of each gang punched group (column 2 in this case). It is advisable to pass blanks on the punch side for the 11 in (E) masters. See note (3).

5. Sort the new cards to column [row 1 with 12 in (F) should automatically be the first card of each column].
6. Calculate on the 604. On the 12 in (F) masters, calculate $(I/J) = Q$, punch in (H) and store. On the following detail cards, calculate $I-QJ$ and punch in (H). Gang punch 12 in (G) of row 2 from (D) by means of digit selectors. Gang punch $(n-1)$ in (C). See note (3).
7. Sort to row. If check sums are carried, cards may be tabulated controlling on row. All rows should sum to zero except the pivotal row which should sum to -1. Round-off errors will appear, of course.
8. At this point, these facts exist:
   a. The cards are in order by row with column 2 first in each row [column 1 was not reproduced in step (4)].
   b. Column 2 has an 11 in (E) which was emitted in (4).
   c. Row 2 has a 12 in (G) which was gang punched in (6). Therefore, the cards may be reproduced again as in (4), sorted as in (5), this time placing row 2 in front, multiplied as in (6) gang punching $(n-2)$ in (C) and 12 in row 3, and sorted and checked as in (7).

The process then consists of repeating this basic cycle: sort, reproduce, sort, calculate, until the order has been reduced to zero. Then all the columns of A will have disappeared, all the rows will sum to -1, and C will have become X. For a final check, multiply AX and compare with C.

NOTES:[b]

1. The use of digit selectors in (6) can be obviated by placing the next pivotal row behind the pivotal row before (5) and gang punching the 12 from (F) to (G) in (6). It is felt that use of the digit selectors offers less chance for error.
2. In (6), using a standard 40-program 604, the following limits on size of numbers seem to exist:

Divisor:	10 digits
Quotient:	10 digits
Dividend:	12 digits
Multiplicand:	10 digits
Cross-foot:	11 digits

The question then arises as to the best way to apportion these digits between decimals and whole numbers. Eight decimals is a good choice for many problems, and seven would provide for all but the most disadvantageous cases. Since the 521 control panel is the same for any number of decimals, it may be advisable to have two or more calculator control panels.

3. In order to divide by ten digits, the ten-digit divisor is split into its first eight digits, x, and its last two digits, y. Then
   $$a(x+y)^{-1} = ax^{-1} - ax^{-2}y + ax^{-3}y^2 - \ldots .$$
   Only the first two terms of this series are calculated. If eight decimals are carried, $y<10^{-6}$, and eight-decimal accuracy is obtained if $x \geqq .1$.

The following procedure provides eight-decimal accuracy when $x<1$.

   a. As terms are calculated on the 604, they are checked, and if $\geqq 1.$, an 11 is punched (not shown on schedule of fields).
   b. This 11 punch is gang punched to field (J) in the reproducing operation.
   c. In the next calculation, if this 11 punch is absent in the divisor field, the divisor and dividend fields are shifted two places to the left in reading. Thus, y becomes zero, and eight-decimal accuracy is obtained at all times.

After the first reproduction, or if pivotal rows are chosen manually (see note 5.), it is necessary to emit this 11 punch in the divisor field if the divisor is $\geqq 1$.

4. If several sets of equations are being handled simultaneously, time can be saved by not sorting case in step (7) but making case the major sort in step (5).
5. The nature of the equations may be such that rows and columns cannot be pivoted in order as outlined, but must be chosen so as to give the smallest quotients. In this event, the 11 in (E) and 12 in (G) must be emitted prior to the reproduction and their automatic insertion discarded at a sacrifice in speed.
6. It is usually not economical to check every cycle on the accounting machine. Errors should be rare and will carry forward if they occur. One compromise is to let a given cycle check while the next one is being processed.

[b]The writer will be glad to supply copies of the planning charts for the 604 and reproducer control panels used in this procedure. Address John Lowe, Douglas Aircraft Company, Inc., Engineering Department, Santa Monica, California.

REFERENCES

1. WILLIAM EDMUND MILNE, *Numerical Calculus* (Princeton University Press, 1949), p. 26.
2. FRANK M. VERZUH, "The Solution of Simultaneous Linear Equations with the Aid of the 602 Calculating Punch," *Mathematical Tables and Other Aids to Computation*, III, No. 27 (July 1949), pp. 453-462.

DISCUSSION

Mr. Turner: What do you do if B_{22} happens to be very small?

Mr. Lowe: In the manner I described, you would actually pick the starting rows and column in sequence—that is, the first column in the first row and the second column in the second row, and so forth. It isn't necessary to do that. You can pick any one you want. In picking, pick the one that would give you the most advantageous numbers. In particular, we usually try to pick the one that gives the smallest quotients in doing this division.

Mr. Wolanski: We have a method that is similar to this, but we always use the element that is greatest; we cannot say the first row or the first column. In the first column we use an element which is the largest; when we do eliminate and get B_{21}, and B_{31} equals zero, we start in on our second column, and we pick the element that is the largest.

Mr. Lowe: Our method for finding out if the numbers get too big is simply to punch out a few more numbers than we can use the next time and then sight-check the cards.

Mr. Bell: In our handling of this problem we try to remove judgment from the operation which the operator performs. We don't want him to have to look at it and evaluate and decide which term to use. So, in handling matrices—usually, in groups—we simply start up from the main diagonal. Perhaps ten per cent of the problems will go bad. We take that ten per cent and start down the main diagonal, and maybe ten per cent of those go bad. Well, then we have 1/100th left over, of the total working volume, and those we actually evaluate and select proper big terms in order to make it behave. But by doing that the mass of the work is handled in a routine way.

Rational Approximation in High-Speed Computing

CECIL HASTINGS, JR.

The RAND Corporation

THIS is a brief report on a study that is being made at RAND on the use of rational approximation in high-speed computing. The work we report upon was largely stimulated, in the first place, through appearance of the IBM Type 604 Calculating Punch, and our work was given further impetus by the reported development of the IBM Card-Programmed Electronic Calculator.

The opportunity of doing away with the use of card tables thus presented itself to us, and we began to prepare for the day when compact approximate expressions would take their place in the art of digital computing. The subject of rational approximation then became a matter of increasing importance. We note in passing that proper use of the 604 can eliminate the use of tables to a considerable extent. Thus, to give an example, one can evaluate a fifth (or even higher) degree polynomial in single-card computation on a 604. The machine will then read an arbitrary value of x from a detail card and punch out $P(x)$ upon the same card. This capability may be used to compute, for example, 5-decimal sines over the quadrant from such a polynomial expression. Simple rational expressions may also be computed in single card computation. Logarithms, square roots, and many special functions have been computed directly on the 604 with considerable success. With the card-programmed electronic calculator, our needs for compact approximate rational expressions to univariate functions will be much increased, for in a sequence of calculations we shall often be required to "look up" a random value of a function in order to continue, and a bulky table is useless here. Some attention is also being given to the problem of multivariate approximation. Our study is largely of an empirical nature. We are compiling data on many instances of rational approximation. A few comments on this data follow. In the remaining sections, a number of random topics pertinent to the study are discussed briefly.

By the term "primitive approximation" we essentially mean the most accurate approximation that can be achieved in a given set of circumstances. We shall call the graph of

$$\epsilon(x) = f^*(x) - f(x) , \qquad (1)$$

the "error curve" of approximation. The reader will notice that the several error curves displayed in this paper have been truly leveled. That is, our approximations are "primitive." Our parameter values are then, of necessity, determined to an excessive number of figures for practical purposes. These may be cut back to obtain "working" approximations. Each primitive approximation will be described by an accurate error curve, the primitive parameter values, location of roots r_i, location of extremal values e_i, and the common extremal values d.

Thus, we approximate the common logarithmic function $\log x$ over $(1/\sqrt{10}, \sqrt{10})$ by the form

$$\xi = \frac{x-1}{x+1} \qquad (2)$$

$$\log^* x = C_1 \xi + C_3 \xi^3 + C_5 \xi^5 + C_7 \xi^7 ,$$

and record the following data:

$C_1 = .8685{,}5434$	$r_1 = 1.000$	$e_1 = 1.204$	$d = .0000{,}0206$
$C_3 = .2911{,}5068$	$r_2 = 1.446$	$e_2 = 1.722$	
$C_5 = .1536{,}1371$	$r_3 = 2.028$	$e_3 = 2.348$	
$C_7 = .2113{,}9497$	$r_4 = 2.656$	$e_4 = 2.920$	
	$r_5 = 3.098$	$e_5 = 3.162 = \sqrt{10}$	

The example is an interesting and perhaps useful one. We notice that

$$\eta = \frac{x-\sqrt{10}}{x+\sqrt{10}} \qquad (3)$$

$$\log^* x = .5 + C_1 \eta + C_3 \eta^3 + C_5 \eta^5 + C_7 \eta^7$$

is an equally good approximation to $\log x$ over the full interval $(1, 10)$.

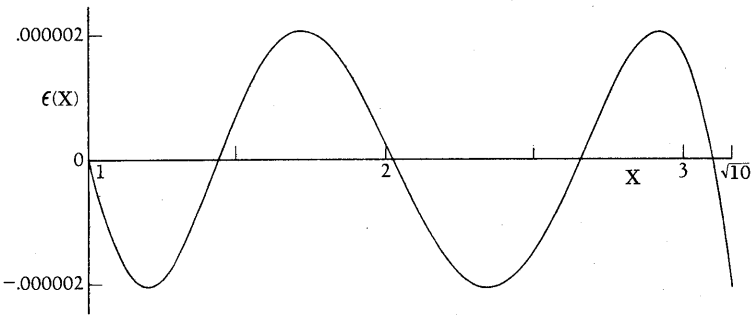


FIGURE 1

Each function approximated is studied with respect to some reasonable sequence of forms, and not as an isolated case, lest one gain this impression from the example above. We will thus obtain empirical data concerning rates of convergence, location of roots, location of extremals, and convergence of parameter values.

The Fitting of Rational Forms

There are several important ways in which the problem of fitting a rational form may be linearized. One method is that of specification of roots. Another is that of specification of extremal deviations.

We hope to make some useful comments upon the location of roots and extremals. Perhaps our final observations will merely be of an empirical nature. It has become apparent to us that, in comparable situations, one error curve will be much like another of the same order. Thus, consider the examples shown in Figures 2 and 3. (We shall actually rule out approximations of the type below as being non-computable. The next section will throw some light on this aside.)

$$W(x) = \frac{e^{-x}}{(1+e^{-x})^2} \text{ by } \frac{1}{4.03809 + .85142x^2 + .13826x^4}$$

$$E'(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \text{ by } \frac{1}{2.56581 + .71158x^2 + .81788x^4}$$

While the functions approximated have quite different behavior, the error curves are remarkably alike except for scale.

Our first trial in making a rational fit is usually by specification of roots. That is, we impose upon the form those values of x at which the error curve is to cross the x-axis.

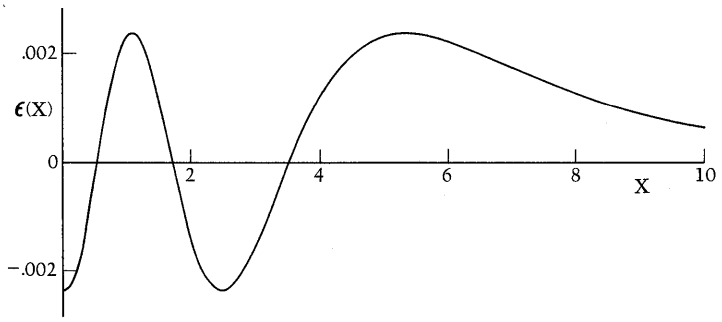


FIGURE 2

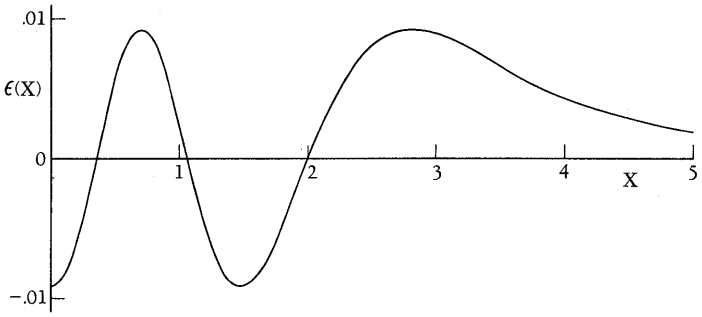


FIGURE 3

Then we compute the error curve on an evaluation sheet, and if we have a true feeling for the situation, all peaks will be of comparable magnitude.

An evaluation sheet is simply a work sheet containing a list of key values of the function $f(x)$ being approximated and a suitable computing setup for use in evaluation of corresponding values of a given approximation $f^*(x)$. A final column allows for the evaluation of $\epsilon(x)$.

To obtain the quality of approximation that we desire to exhibit, an adequate number of accurate function values must be incorporated in the evaluation sheet. Generally, we select about fifty or sixty values of the independent variable in the construction of a given evaluation sheet, and these are given to about three or more figures than the most accurate approximation we expect to obtain. With the possible exception of a few extraneous values, our tabulated key values of a function $f(x)$ will be for equally spaced intervals either in x or in some simply related transformed variable $\xi = \xi(x)$.

Once a crude approximation has been obtained and the corresponding error curve determined, we may employ the more refined method of specification of extremals to level the peaks as we have done in the several examples given above. Several computing cycles may be required to achieve the quality of approximation desired.

In speaking of extremal deviations, we naturally refer to the error curve of approximation. Generally an extremal of an error curve will be either a true minimum or a true maximum. Possible exceptions occur only at end points of the interval of interest, in well-behaved cases. In general there will be one more extremal to an error curve than there are free parameters in the form. Thus, the two error curves given above each have four extremals, and there are three parameters (b_0, b_2, b_4) in the form. Or looking at the matter in another light, there are seven extremals, and six parameters, as the approximation holds over $(-\infty, +\infty)$.[a] Thus, a final error curve results from the correct determination of $n + 1$ extremal locations e_i or, more helpfully,

[a] And odd power terms are missing.

from the correct determination of n extremal locations e_i and one common deviation d, that is the absolute magnitude of the error curve at each of the $n + 1$ extremals.

While our problem is then accurately described as one of solving a system of transcendental equations in $n + 1$ unknowns e_i and d, actually our problem amounts to nothing more than that of solving a single transcendental equation in d. And this is quite a simple matter. In practice, the values of e_i^* that we "read" from a given intermediate error curve are used unchanged in the cycle that follows. The drift of location is usually very small, and the results are somewhat insensitive to the e_i^*. (An abbreviated numerical process is used in lieu of curve plotting, and hence "read" is in quotes.)

The Problem of Sensitivity

In almost all cases of rational approximation, we run into a problem of sensitivity. Thus, a function $f(x)$ may behave decently, have an adequate rational approximation of the form

$$f^*(x) = \frac{N(x)}{D(x)} \tag{4}$$

and yet the components $N(x)$ and $D(x)$ may behave very badly, changing in joint fashion by many orders of magnitude as x ranges over the interval of approximation. Fortunately, however, there always seems to be a happy solution to this kind of difficulty.

Thus, in making a 4-decimal ($d = .000139$) approximation to

$$\phi(x) = \frac{1 - e^{-x}}{x} \tag{5}$$

over $(0, \infty)$, we cast the result into the desensitized form

$$\xi = \frac{1}{1 + .3606032x} \tag{6}$$

$$\phi^*(x) = \frac{.3671626\,\xi - .2272232\,\xi^2 + .8601996\,\xi^3}{1 - 1.3562710\,\xi + 1.6148087\,\xi^2 - .2585377\,\xi^3}$$

in which a $D(1) = 1$ condition has been imposed on the second form by proper choice of parameter in the first form.

It is easy to see that (6) is an extremely tractable approximation—what we shall call a computable approximation. The variable ξ is limited to the interval $(0, 1)$ of variation, all coefficients and individual terms in the expression are of reasonable magnitude, and the denominator of the second form behaves very well. This may be seen in Figure 4.

In very many instances, the least sensitive representation that may be obtained is obtained through imposing a $D(a) = D(b) = 1$ condition on the denominator when an approximation is made over (a,b).

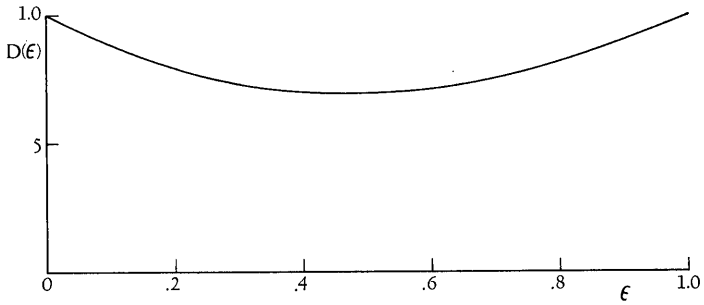


FIGURE 4

The Problem of Best Fit

Our criterion of best fit will be that of minimum deviation. Can we then say that our sinusoidal error curves give evidence, in each instance, that no better fit can be obtained? We now give an interesting result for polynomial forms. The result is, undoubtedly, a familiar one to workers in the field of polynomial approximation.

Consider the polynomial form

$$y = a_1x^{p_1} + a_2x^{p_2} + \ldots + a_nx^{p_n}, \tag{7}$$

with n free parameters a_i, and specified powers p_i that are distinct integers $0 \leqq p_1 < p_2 < \ldots < p_n$, and $n + 1$ points (x_i, y_i), $i = 0(1)n$, for which $0 < x_0 < x_1 < \ldots < x_n$. We introduce the residual notation

$$T_i = a_1x_i^{p_1} + \ldots + a_nx_i^{p_n} - y_i, \tag{8}$$

and let A_i denote the value of the determinant obtained from the $n \times n + 1$ matrix

$$\begin{bmatrix} x_0^{p_1}, x_1^{p_1}, \ldots, x_n^{p_1} \\ x_0^{p_2}, x_1^{p_2}, \ldots, x_n^{p_2} \\ \cdot \quad \cdot \quad \cdot \\ \cdot \quad \cdot \quad \cdot \\ \cdot \quad \cdot \quad \cdot \\ x_0^{p_n}, x_1^{p_n}, \ldots, x_n^{p_n} \end{bmatrix} \tag{9}$$

by deletion of the x_i column. A brief argument will show that all A_i are positive.

A proof is by induction on n. Expand any A_i by elements of the last column, and refer to Descartes' rule of signs. There are at most $n - 1$ roots to the polynomial in the last column variable, and these are each of the remaining column variables. Our polynomial may then be factored in such a fashion that all factors are easily seen to be positive, and the result follows.

Now, we notice that minimization of

$$\sum_{i=0}^{n} A_i T_i^2 , \tag{10}$$

with respect to the a_i, yields the equations of condition

$$T_i = (-1)^i T_0 . \tag{11}$$

Let us denote the residuals in this particular case by the notation T_i^*, and write

$$T^* = | T_i^* | . \tag{12}$$

Let T_i^{**} stand for the residuals that result from any other choice of parameters a_i, and write

$$M = \operatorname*{Max}_i \left| T_i^{**} \right| . \tag{13}$$

It then follows that

$$T^{*2} \sum_{i=0}^{n} A_i = \sum_{i=0}^{n} A_i T_i^{*2} < \sum_{i=0}^{n} A_i T_i^{**2} \leqq M^2 \sum_{i=0}^{n} A_i \tag{14}$$

which says that $T^* < M$. The result may be stated as a theorem.

THEOREM: *The minimum deviation solution is such that all deviations are equated in absolute value, and the signs of the deviations alternate, unless the form passes exactly through the $n + 1$ points, in which case all deviations are zero.*

This result tells us, in the case of polynomial approximation, that if we obtain an error curve with $n + 1$ extremals that obey the conditions of the above theorem, then the approximation cannot be improved. For no other instance of the same form can approximate those particular $n + 1$ points of the curve to be approximated as closely as the instance in question.

Iterated Rational Forms

Certain limiting cases of rational approximation may also be of practical interest. Thus, we may find a form of the type

$$f(ax) = R[f(x)] , \tag{15}$$

in which $R(\lambda)$ is a rational function of λ, quite adaptable in special instances. Here a stands, in general, for a positive constant sizeably greater than unity. Quite obviously, the problem of fitting such a form is just that of making a usual rational fit.

Thus, consider the function that satisfies

$$H\left(\frac{3}{2}x\right) = 3H^2(x) - 2H^3(x) , \tag{16}$$

and has the power series expansion

$$H(x) = \frac{1}{2} + \left(\frac{x}{\sqrt{2\pi}}\right) - \frac{16}{15}\left(\frac{x}{\sqrt{2\pi}}\right)^3 + \frac{1024}{975}\left(\frac{x}{\sqrt{2\pi}}\right)^5 + \ldots \tag{17}$$

about the origin. Here $H(0) = 1/2$ is a fixed point, and $H'(0)$ is an imposed condition. $H(x)$ rather closely approximates the Gaussian error function

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} dt , \tag{18}$$

over $(-\infty, \infty)$ as may be seen in the difference curve plotted below. Note that $F(x)$ and $H(x)$ both have the same type of odd symmetrical behavior, so that our difference curve need but be given over $(0, \infty)$, Figure 5.

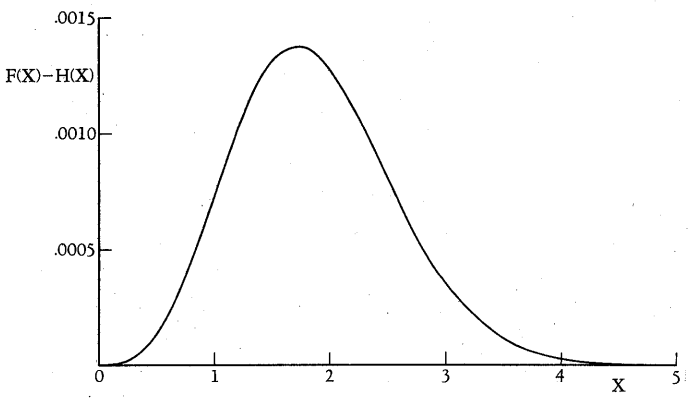


FIGURE 5

Multivariate Approximations

We are interested in the subject of a generalized empirical analysis in which multivariate forms of a very general type are studied. These are non-parametric forms in which the unknown quantities are univariate functions of the separate independent variables.

By considering basic functional forms, such as

$$h^*(x,y) = \frac{f_1(x) + g_1(y)}{1 + f_2(x) g_2(y)} , \tag{19}$$

we decompose the problem of multivariate approximation into its fundamentals. Here $f_i(x)$ and $g_i(y)$ are unknown univariate functions to be determined so that $h^*(x,y)$ shall approximate a specified function $h(x,y)$ over a region of interest as well as possible.

Thus, if the form (19) is inadequate, in an instance, no parametric form of similar structure can be adequate in the same instance. Considerable work has been done on the problem of determining required key values of the unknown functions in such forms. Once tables of values are available for each unknown function, the art of univariate approximation may be employed to complete the fitting process.

A further example of a general functional form is the slide rule form

$$h(z^*) = f(x) + g(y) \tag{20}$$

in which $z^* = z^*(x,y)$ is to approximate $z(x,y)$ over a

specified region. Thus, in a large machine computation that we undertook for a division of RAND, we approximated

$$z(x,\gamma) = \sqrt{\left(\frac{2}{\gamma-1}\right)\left(1-x^{(\gamma-1)/\gamma}\right)} \qquad (21)$$

to 3 decimals over $1.2 \leqq \gamma \leqq 1.4$, $.05 \leqq x \leqq 1$ by such a form, and prepared turning point tables of $f(x)$, $g(\gamma)$, and $h(z)$. These were used with efficient success.

Of great practical interest, in the approximation of well behaved bivariate tables, is the simple form

$$z^*(x,y) = f_1(x)\,g_1(y) + f_2(x)\,g_2(y) + \ldots + f_n(x)\,g_n(y)\,. \qquad (22)$$

To give an impressive example, for the reader to learn the power of such a form in the fitting of useful tables, we note that

$$V(h,q) = \frac{1}{2\pi}\int_o^h \int_o^{\frac{qx}{h}} e^{-\frac{1}{2}(x^2+y^2)}\,dy\,dx\,, \qquad (23)$$

may be approximated by

$$V^*(h,q) = f_1(h)\,g_1\left(\frac{q}{h}\right) + f_2(h)\,g_2\left(\frac{q}{h}\right) \qquad (24)$$

over $0 \leqq h \leqq \infty$, $0 \leqq q/h \leqq 1$ to 4 decimals. This suffices, with the aid of a table of the error integral, to yield all values of $V(h,q)$.

REFERENCES

1. C. Lanczos, "Trigonometric Interpolation of Empirical and Analytical Functions," *Journal of Mathematics and Physics*, Vol. XVII, No. 3, September, 1938.
2. H. H. Germond, "Miscellaneous Probability Tables," AMP Note No. 14 (unclassified).

DISCUSSION

Professor Kunz: I am interested in how this scheme, which you use for getting a polynomial approximation, compares with using Tchebyscheff polynomials. Do you use that technique?

Mr. Hastings: We often use that method in fitting polynomials.

Professor Kunz: The Tchebyscheff gives you the best fit in the same sense that you have used it here?

Mr. Hastings: Not exactly. It gives it to first order.

Dr. King: Dr. Tukey and I have a theorem that when a man gets associated with computing machinery he seems to spend a lot of time discovering previous fields of mathematics. The first part of your talk is really the theory of Pades' tables, for which there is quite an extensive literature, although not one that one comes across very commonly. I don't mean to detract at all from the value of what you have done, but to point out, rather, that there are existing theorems. In fact, one of the difficulties with much of the mathematical literature is that it is not too useful when you want to put it on machines.

In particular, I would like to point out that the polynomial approximation has one especially simple form; when you try to approximate a function by a continued fraction, you get exactly that form, and the theory of the Pade tables was to take that particular form and adjust the coefficients in the polynomials to get a better approximation. In other words, a continued fraction is a relationship between the coefficients of the polynomial; while the Pade theorem approach was to try and investigate what are the best polynomials to use. I might say that, as far as I know, they never came to any good conclusion, either as to what you mean by "best" or how you get there!

Mr. Hastings: Thank you.

Dr. Tukey: I would like to ask the speaker if he would agree with the following position: That the real purpose of computing a table is so that you can get a good approximation and throw the table you just computed away.

Mr. Hastings: I probably would keep the table, even if I did not use it.

The Construction of Tables

PAUL HERGET

Cincinnati Observatory

THE FIRST IDEA I wish to mention is a very simple one that comes from calculus. This idea occurred to me years ago when we had only a 601 and were unable to perform division.

Suppose our purpose is to tabulate an analytic function, $f(x) = X$, at equal intervals of the argument, x, where X represents the numerical value placed in the table for each entry. This method is to be applied in the cases where there exists an inverse function, $F(X) = x$, and will be practicable, if it is simpler to compute $F(X)$ than it is to compute $f(x)$ by some direct means.

The problem may be stated as follows: If the series for $\sin \phi$ can be expanded on a program cycle or a series of programs, then if the inverse function is needed, the series is expanded until it equals the sine, thus giving the angle. If we write $F(X_0 + e) - x = 0$ and expand the expression by Taylor's theorem and use Newton's method of approximation, $e = \dfrac{x - F(X_0)}{F'(X_0)}$ is obtained, where X_0 is some approximation to the correct value. Also the following equations are true:

$$dX = f'(x)dx$$
$$dx = F'(X)dx$$
$$\frac{dX}{dx} = \frac{1}{F'(X)} = f'(x) = \frac{\Delta^1}{w}$$

Now write

$$X_1 = X_0 + e = X_0 + [x - F(X_0)]\,\frac{\Delta^1}{w}. \qquad (1)$$

This is a first-order approximation, but it is applied only to the residual which is supposed to be small. If a set of approximate X_0's can be obtained for the entries to be put into the table, this iterative equation is computed instead of attempting to compute $f(x)$. This is on the assumption that $F(X_0)$ can be computed more readily than $f(x)$. In the case of an arc sine table, for example, this means that one must have, necessarily, a table of sines to use, or a sine series, or something similar.

If the table is to be at equal intervals of the argument, presumably the entries are in units of a certain decimal place, so that $1/w$ simply requires proper placing of the decimal point, and Δ^1 is obtained by differencing the values to be placed in the table at that stage. If necessary, a first approximation can be obtained as illustrated in Figure 1.

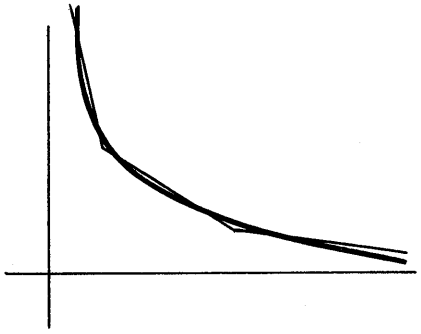


FIGURE 1

The true curve is approximated by a series of chords. Each value of X_0 is obtained by summary punching with progressive totals, tabulating blank cards, and accumulating a constant first difference from the digit emitter. Some hand computing is necessary to find the points where the constant emitted first difference should be changed, and the number of changes is balanced against the size of the allowable error of X_0.

The division by $F'(X_0)$ has been replaced by a multiplication by Δ^1, and equation (1) is iterated repeatedly until the final values are reached. Although this procedure may not appear attractive, it is for the purpose of eliminating the division. It may be applied if a table, say, to four places is available, and a table to eight places is needed. Perhaps some of Mr. Hastings' inverse functions could be generated more easily than the direct functions. It would have an application there.

Next, I would like to indicate the results in two simple cases. If a reciprocal table is to be obtained, then $f(x) = X = 1/x$, and $f'(x)$ is eliminated, as derived from this expression, instead of substituting Δ^1/w in equation (1). Then $X = X_0 + (1 - xX_0)X_0$. It is not necessary to summary punch any differences. An iterative formula is obtained which would have been obtained more easily some

other way! In the second case, if $f(x) = x^{-1/2}$, the same procedure is applied, obtaining

$$X = X_0 + (0.5 - 0.5xX_0^2)X_0 .$$

This is a quantity which arises when transforming from rectangular to polar coordinates. It is necessary to divide by the square root of r^2.

Now I would like to say a few things about the construction of a sine table. My remarks apply beautifully to the sine and cosine function, and contain some ideas which may be extended to other functions. Let Figure 2 represent the quantities in a sine table at equal intervals of the argument.

$\phi - w$	$\sin(\phi - w)$	$A \sin(\phi - w)$	. . .	
		Δ^1	Δ^3	. . .
ϕ	$\sin\phi$	$A \sin\phi$	$A^2 \sin\phi$	
		Δ^1	Δ^3	. . .
$\phi + w$	$\sin(\phi + w)$	$A \sin(\phi + w)$	. . .	

FIGURE 2

The second difference opposite $\sin\phi$ is

$$\sin(\phi + w) - 2\sin\phi + \sin(\phi - w) = -2(1 - \cos w)\sin\phi = A \sin\phi .$$

Similarly, the fourth difference becomes $A^2 \sin\phi$, which is rigorous, and not an approximation. This is a property of the sine function. The above suggests that if interpolation is desired, the best procedure is to use Everett's interpolation formula. This will reduce to:

$$\sin(\phi + nw) = \sin\phi\left[m - \frac{m(1^2 - m^2)}{3!}A + \frac{m(1^2 - m^2)(2^2 - m^2)}{5!}A^2 + \ldots\right] + \sin(\phi + w)\left[n - \frac{n(1^2 - n^2)}{3!}A + \frac{n(1^2 - n^2)(2^2 - n^2)}{5!}A^2 + \ldots\right]. \quad (2)$$

This process of interpolating for a sine between two given values means that each of the values is multiplied by its corresponding square brackets, which is, in general, different from the ordinary concept of interpolation.

It is seen from trigonometry that the square brackets in equation (2) which have been derived by means of working with Everett's interpolation formula, have closed expressions, namely

$$\frac{\sin(1 - n)w}{\sin w} \text{ and } \frac{\sin nw}{\sin w} .$$

Here I would like to point out something which is just a curiosity. Suppose $\phi = 0$, then the first line of equation (2) will drop out. Consider the second line. One of the first things the teacher tries to emphasize when this subject is reached in trigonometry is that sin nw is not equal to n sin w, but you see that, with the exception of the higher order terms in the series, it is true. So that is a good way to confuse everybody!

I shall describe briefly how we constructed an eight-place sine table, in what we considered a most efficient manner of arriving at a set of punched cards for the table. Ninety cards were key punched, proof read, and differenced, each containing the sine of an integral degree. If the argument is in degrees and if an eight-place sine table is desired, then the interval must be 0.°01 in order to have linear interpolation. This yields a second difference of three units in the eighth place. Since the second difference coefficient is never greater than one eighth, the error is less than a half unit in the last place.

If the expression in the square brackets of equation (2) was computed one hundred times for $n = 0.01, 0.02$, etc., and the results used to form the one hundred interpolates, the basic idea of the process can be seen. Now this complete set of square brackets is much like a sine and cosine table itself, because after going from $n = 0.0$ to 0.5 for the first bracket, the second half from 0.5 to 1.00 may be obtained by reading up the column, as one reads down one column to get the sine from 0° to 45° and then back up the cosine column to obtain values from 45° to 90°.

In the present case it means that when one multiplication has been made of a square bracket times sin ϕ, it is used once for a given value of m in one degree and again for the same value of n in the next degree. Although every single sine entry is obtained as the sum of two products, there are only as many products as there are entries in the table, because each one is used twice.

In practice the entries are not formed directly by equation (2) but the square brackets are differenced, the products formed, and then the interpolates are built up by progressive totals over each range of one degree. This process enables a multiplication with eight-figure factors (the normal capacity of a 601) and still protects the end figures against accumulated roundings. The differences of the square brackets are of the form:

$$0.01 + A\Delta^1 E_2 + A^2\Delta^1 E_4 + \ldots$$

If this expression is evaluated to 12 decimal places, there will be only 8 significant figures beside the leading 0.01. $\Delta^1 E_2$ means the first difference of the Everett second difference coefficients. The multiplication of sin ϕ by 0.01 is accomplished by crossfooting, and the rest is multiplied in the usual way. This allows two decimal places for protection of the end figures, owing to the progressive totals 100 times, and two extra places in the computation to insure the correct rounding to the closest half unit.

Nine thousand multiplications were performed, using 100 group multipliers, and the work was arranged with three fields to the card. Then the cards were sorted and tabu-

lated; 3,000 summary cards yielded the final values for the table. There was an automatic check as each sine of an integral degree was reproduced at the end of that interval. The final table was then reproduced, and it was necessary to punch the first differences of the rounded values in order to interpolate these. The final check was the differencing of the first differences, which were inspected. The entire operation took about twenty hours.

You will perceive readily, from Figure 2, that a sine table may be constructed simply by multiplying $A \sin \phi = \Delta^2 (\sin \phi)$ and building up step by step. The process has the disadvantage of accumulated rounding off errors, just as in numerical integration; thus, more places must be carried as a protection. In fact, this is the process of numerical integration in the case of this simple function. That leads me to make one other comment: by means of numerical integration it is possible to tabulate these functions, or to generate the functions so that cards are punched.

There is no need to discuss the subject further, since it is all covered in the subject of numerical integration. The ease with which equations can be integrated depends upon how much protection can be obtained against the accumulation errors, and how well the integrands, needed for the integration, can be computed.

THE ABOVE concludes the remarks with respect to tables, which have equal intervals of the argument. The remaining remarks will apply to optimum interval tables, and I am not sure whether or not it is necessary to describe the essential property of an optimum interval table. It amounts to this: First, consider the simplest way of interpolating a table linearly: $f_n = f_0 + n\Delta^1$. Of course, in all the tables that everyone uses from high school on, the interval of the argument, w, is in some unit in the decimal system and one never thinks of the whole role that is played by the fraction n. It is necessary to consider what happens if the interval has some odd value, for example 0.2. Actually $n = (x - x_0)/w$ and we shall write our interpolation formula in the form

$$f_n = \left(f_0 - \frac{x_0}{w}\Delta^1 \right) + x\left(\frac{\Delta^1}{w}\right) = F_0 + x D_1 \ . \quad (3)$$

What does this mean in numbers? This is illustrated in the following example:

x_0	f_0	Δ^1	F_0	D_1
12.0	0.17318		0.17318	1110
		222		
12.2	0.17540		0.17316	1120
		224		
12.4	0.17764		0.17312	1130
		226		
12.6	0.17990			

If $x = 12.325, f(x) = 0.17316 + 0.325(1120) = 0.17680$. It is obvious, at a glance, that since 0.325 lies between 0.2 and 0.4, one is not supposed to multiply directly by the digit 3. There must be something wrong here. If we had applied this operation to the f_0 and Δ^1 columns, it would be wrong. But the F_0 and D_1 columns have been adjusted so as to compensate in advance for the error we would otherwise make. This means that all the card columns corresponding to .325 are wired directly to the multiplier unit and the machine takes no cognizance of whether the interval is ordinary or unusual.

That is all there is to an optimum interval table. You see, you fool the machine by giving it a table which doesn't make sense. The control panel is wired in a way that doesn't make sense. Generally, this theorem doesn't hold, but in this case the two nonsenses make sense.

What I would like to show you is a very unsophisticated way of looking at the construction of such a table, if a table with second order interpolation is desired. The generalizations are fairly obvious. If one wishes to have an interpolation formula of the form

$$f_n = F_0 + N D_1 + N^2 D_2 \quad (4)$$

the way in which we shall approach the problem is as follows: Suppose the intervals of the argument, which we are to use, have been established, the interval may be changed whenever the function's variation warrants, but within the restriction that the combinations of intervals must always end in a cycle of ten units of the left-hand position of the multiplier N. Thus, you may have combinations like 0.2, 0.3, 0.2, 0.3 or 0.3, 0.3, 0.4 or 0.5, 0.5. We shall call the interval length $2a$ and the value of the argument at the center of the interval x_0. Now expand the function in a Taylor's series about the mid-point. In the present case, instead of trying to write down completely general equations, I shall write down the results as if we are interested in constructing a table of $f(x) = x^{-3/2}$. This is a table that is needed repeatedly in dynamical astronomy because we have $1/r^2$, which is the law of gravitation, and another factor x/r, which is the projection onto the x-axis (and similarly for y and z). Thus, r^3 always will be in the denominator and it is easy to obtain $r^2 = x^2 + y^2 + z^2$. We then have:

$$f(x) = x^{-3/2},$$

$$f^{\mathrm{i}} = -\frac{3}{2}x^{-5/2} = -\frac{3}{2}\frac{f(x)}{x},$$

$$f^{\mathrm{ii}}(x) = +\frac{15}{4}x^{-7/2} = +\frac{15}{4}\frac{f(x)}{x^2},$$

$$f^{\mathrm{iii}}(x) = -\frac{105}{8}x^{-9/2} = -\frac{105}{8}\frac{f(x)}{x^3},$$

$$f^{\mathrm{iv}}(x) = +\frac{945}{16}x^{-11/2} = +\frac{945}{16}\frac{f(x)}{x^4},$$

$$f(x) = f(x_0 + h) = \frac{1}{x_0^{3/2}}\left[1 - \frac{3}{2}\frac{h}{x_0}\left\{1 + \frac{35}{24}\frac{h^2}{x_0^2}\right\} + \frac{15}{8}\frac{h^2}{x_0^2}\left\{1 + \frac{63}{48}\frac{h^2}{x_0^2}\right\} + \cdots\right]$$

The terms h^2/x_0^2 in the braces are actually the third and fourth derivative terms which cannot be included because the interpolation formula is to be only quadratic. However, since these terms are always positive, they shall be used as

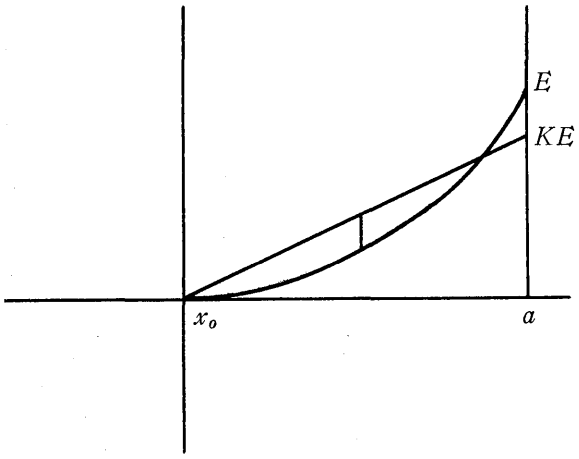


FIGURE 3

illustrated in Figure 3. Let E be the total error committed at $x = x_0 + a$ if we neglect the third order term completely. Let KE be the fractional part of this error which is taken into account if the cubic term is replaced by a linear term, as shown. Then the remaining neglected error is $h^3E - hKE$. This error has a maximum (shown by the short vertical line) at $h = \sqrt{K/3}$. If the error is set at this point equal in magnitude and with opposite sign to the error at the end of the interval, we have

$$\sqrt{\frac{K}{3}}\left(\frac{K}{3} - K\right)E = (K-1)E \text{ and } K = 3/4\,.$$

If we analyze the quartic term in the same way, we obtain $K = 2(\sqrt{2} - 1)$. Our interpolation formula becomes

$$f(x) = \frac{1}{x_0^{3/2}}\left[1 - \frac{3}{2}\frac{1}{x_0}\left\{1 + \frac{3}{4}\frac{35}{24}\frac{a^2}{x_0^2}\right\}h + \frac{15}{8}\frac{1}{x_0^2}\left\{1 + \frac{(\sqrt{2}-1)}{1}\frac{63}{24}\frac{a^2}{x_0^2}\right\}h^2\right].$$

Since the two braces are so nearly alike, we may use, without sensible error, $1 + 1.09375\ a^2/x_0^2$ for each of them.

Now we may expect that the interval which may be used is somewhat more favorable than that which would be determined on the basis of neglecting the third and fourth order terms completely. I shall let Dr. Grosch explain the way in which the intervals are obtained. What is done as a rule of thumb is to write $192e = w^3 f^{\text{iii}}(x)$, where e is the admissible error, usually one-half unit in the last place. Then the size of the third derivative will control the value of the interval which may be used.

At this stage we have

$$f(x) = f_0 + h f_1 + h^2 f_2\,. \qquad (5)$$

We are still faced with one other problem before we are finished: h is counted from the middle of the interval; so we shall write $n - a = h$. Then n is counted in the same units as h, but from the beginning of the interval. But if the value of the argument at the beginning of the interval is not zero, but A, then let $N - A = n$ where N is the number which is actually used as the multiplier in equation (4). Making these substitutions in (5) in order to reduce it to the form of (4), we find that all of the following relations exist. I shall write down only the end results.

$$D_2 = f_2\,, \qquad D_1 = f_1 - 2(a + A)D_2$$
$$F_0 = f_0 - (a + A)\,D_1 - (a + A)^2 D_2\,.$$

This is about the simplest way, of which I could think, to present the development from the function and a Taylor's series to the final results entered in the table.

DISCUSSION

Dr. King: I would like to make some much more general remarks. It is a good thing for a lot of people to work on these problems so as to make sure that the best method finally comes out.

On the other hand, there is a point where it is inefficient to have too many people, and I would like to ask the speakers whether they think the last word has been more or less said on optimum interval tables, and, if so, I am sure there are some particular little details that could be improved. So I would like to hear from them whether they think the time is ripe for people to get together and have one system of optimum interval tables.

Dr. Grosch: I think, honestly, we can say that the polynomial case for the single variable is just about under control now. By the time you go to two variables it becomes so complicated that it may not be worth investigating until we have some big bi- or trivariates that we just have to make. It is really a beastly job, even in the linear case. I have made some explorations in that direction and don't feel at all satisfied. In the univariate case, I don't think there is much we can do beyond this inverse matrix business, and the reason I am so sure of it is this: that if you pick any interval (and you may pick a wrong interval, because several terms in the Taylor series are contributing; higher order terms, as Dr. Herget shows, are being added with lower order terms and so forth); but if you pick an interval under any assumption whatsoever, Mr. Hastings' comment of yesterday is the key to the whole situation that the error curve for a certain degree of approximation is going to look just about the same. It will change a little bit. He said Tchebyscheff was the zeroth order approximation to that curve. It will change a little, but the position of the extrema is very stable. Therefore, you are going to make an error of ϵ at the position where you think those extrema are going to occur; and, even if the function doesn't behave quite the way its next higher term indicates it should, the extrema aren't going to shift very much. Therefore, your value of the actual error curve obtained when you use the table will not be more than a tenth of a per cent or a hundredth of a per cent greater than theoretical ϵ,

unless you come to a curve so very bad that the error curve doesn't look anything like a Tchebyscheff polynomial; and, of course, we can always invent such curves. But I think they are going to be quite hard to invent.

I also expect that the rational function is going to have a very stable error curve, what Professor Tukey referred to as the Tchebyscheff polynomial for rational functions. But I don't have that as yet and I don't know whether Mr. Hastings has.

Professor Kunz: I think one of the important things in this talk is that Dr. Grosch has reminded us that there are other ways of interpolating. Just as a very trivial suggestion: if you take a hyperbola and pass it through three points, this gives you second order interpolation in a more general sense. Some of you who haven't tried similar things might like to try it. One of the interesting properties is that you try inverse interpolation with such a function, and it is just as easy as direct interpolation. You can obtain second order inverse interpolation very nicely. I have used this in quite a few cases, and it sometimes yields a very nice fit to curves, particularly if they have a singularity somewhere in the region.

It is just a suggestion to become sort of initiated to these reciprocal differences which are a little forbidding and are awfully hard to integrate.

Professor Tukey: I cannot agree with Professor Kunz in the use of the word "interpolation." The essential point about this is that we have given up interpolating, just as we have given up expanding in series. We are trying to get something that is a good fit, and that is a different problem.

Mr. Bell: While we are on the subject of tables, I would like to point out another way of getting a fit to curves of various sorts. It is a widespread opinion among engineers that a problem which involves curves of some sort cannot be done on punched cards. I am talking, of course, about engineers who have hearsay knowledge of IBM equipment.

Now, this is not true. All you have to do is read a lot of points. With the points you can obtain first differences, set up a linear interpolation, which can be done quite quickly. Of course, this is completely non-elegant, but very practical. Many times you have whole families of curves. We, in our organization, are fortunate in having rapid ways of reading such data. We can read, say, a thousand points from families of curves in maybe an hour and be ready to go on a problem without mathematics.

A Description of Several Optimum Interval Tables

STUART L. CROSSMAN

United Aircraft Corporation

THE SEVERAL optimum interval tables included in this paper were constructed for linear interpolation on the IBM Type 602 Calculating Punch in an endeavor to accelerate the process of table look-up. In each case a critical table of thousands of lines was reduced to a table of fewer than two hundred lines. The number of lines per table might have been still further reduced by using a higher order of interpolation, but this was not desirable since the interpolating time on the type 602 calculating punch is approximately proportional to the order of interpolation.

The following tables were constructed:

Table	*Function*	*Accuracy*	*Range and Interval*	*Size*
A	e^t	$1 \cdot 10^{-5}$	$t = -1.7000(.0114$	
	e^{-t}	$1 \cdot 10^{-4}$	$-.0102)-0.4000$	104 cards
B	$1-r^{2/7}$	$1 \cdot 10^{-5}$	$r = .30000(.00350$	
	$\frac{(1-r^{2/7})^{\frac{1}{2}}}{r^{2/7}}$	$1 \cdot 10^{-5}$	$-.00013).99900$	192 cards
C	Arc cosh x	$1 \cdot 10^{-4}$	$x = 1.0002(.0001$ $-1.0500)27.3600$	132 cards

Tables A and B each consist of two functions with a common argument. This arrangement was convenient in that both functions of a given table were needed, at the same time, in the particular problem for which the table was constructed. Only one sorting operation is necessary to file the table cards with the detail cards in preparation for the interpolation of both functions. Including two functions in a table with a common argument may result in a slightly larger number of lines than would be obtained in either of the functions if each were optimized independently. However, the additional lines are of little consequence, since an entire sorting operation is eliminated.

The tables were constructed using a method developed by Dr. Herbert R. J. Grosch of the Watson Scientific Computing Laboratory.

The method consists of dividing the interval (a, b) upon which the required function is to be approximated, into a number of sub-intervals, upon which the function is replaced by straight lines of the form (Figure 1)

$$f(x) = b_i + m_i\,(x-x_i)\,. \qquad (1)$$

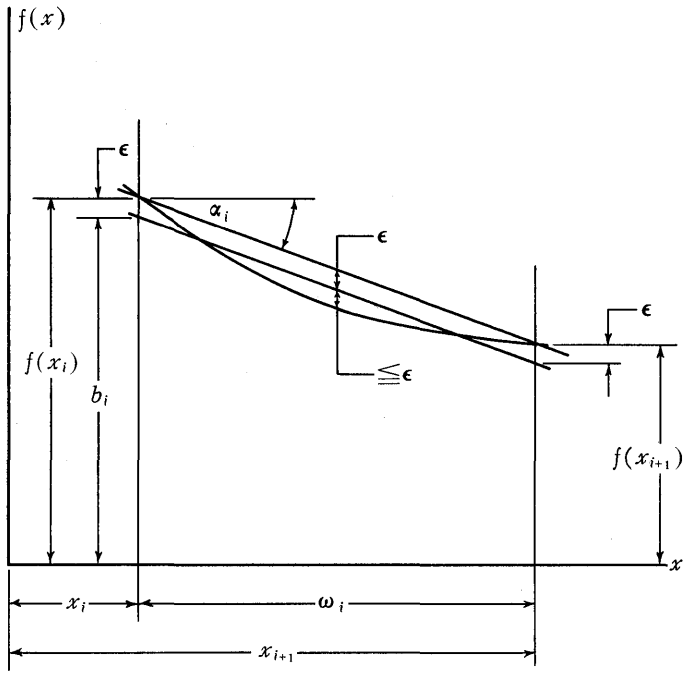


FIGURE 1

For the functions included in this paper it is convenient that b_i be referenced to the vertical axis at x_i for each interval, thereby limiting its magnitude. The optimum sub-intervals are determined by the expression

$$\omega = \frac{4\sqrt{\epsilon}}{\left|\frac{d^2f}{dx^2}\right|^{\frac{1}{2}}}\,, \qquad (2)$$

where

ω = the tabular interval

$\frac{d^2f}{dx^2}$ = the 2nd derivative of the function, $f(x)$

ϵ = the maximum theoretical error between the approximate value and the true value of the function.

The number of lines, N, can be found approximately from the expression

$$N(a, b) = \int_a^b \frac{dx}{\omega}\,. \qquad (3)$$

Since each of the tables was constructed in the same manner, a description of Table B will illustrate the details of construction.

This table was constructed such that the functions are everywhere represented on the interval to an accuracy of 1×10^{-5} (total error).

The total error consists of the sum of the theoretical error (ϵ) between the straight line and the function, the rounding error when interpolating (1), and the departure of b_i from theory due to rounding. Hence, for each of the two functions

$$f(r) = 1 - r^{2/7} \tag{4}$$

$$F(r) = \frac{(1 - r^{2/7})^{\frac{1}{2}}}{r^{2/7}} \tag{5}$$

the value of ϵ is as follows:

$$1 \times 10^{-5} = \epsilon + .5 \times 10^{-5} + .05 \times 10^{-5},$$
$$\epsilon = .45 \times 10^{-5}. \tag{6}$$

It will be noted that the inclusion of "the rounding error when interpolating (1)" insures that the rounded interpolated value of the function agrees within 1×10^{-5} of the true unrounded value of the function.

An examination of the second derivatives of the functions discloses the direction of increasing ω, i.e., the successive values of ω increase as the absolute values of the derivative decrease. For the function $f(r)$, the intervals increase from left to right throughout the range of the function. However, for $F(r)$ the intervals increase from left to right to the inflection point, at $r = .73115$, for the left-hand portion of the curve, and from right to left for the right-hand portion of the curve. Hence, for the function $F(r)$ the two sections of the curve are treated independently and the intervals calculated from $r = .30000$ and $r = .99900$ toward the inflection point.

The second derivative of $f(r)$ and the value of ϵ when substituted in (2) give

$$\omega = 18.782971 \times 10^{-3} r^{6/7}. \tag{7}$$

Beginning at $r = .30000$ and substituting in (7), the first interval was computed. The interval is added to the value of r to establish the new r for calculating the next interval. The process is repeated through $r = .99900$. Each value of ω is calculated to seven places, but the last two places are dropped and the unrounded five-place value used to establish the starting point for the next interval.

The second derivative of $F(r)$ and the value of ϵ when substituted in (2) give

$$\omega = \frac{59.396970\, r^{8/7}\, (1 - r^{2/7})^{\frac{3}{2}} \times 10^{-3}}{(8r^{4/7} - 27r^{2/7} - 18)^{\frac{1}{2}}}. \tag{8}$$

The intervals were then computed for each portion of the curve. As the value of r approached the inflection point, the intervals increase and become infinite at the inflection point. The intervals which cross the inflection point are shortened to end at that point.

A comparison of the intervals for the two functions disclosed that the intervals for $F(r)$ were smaller than for $f(r)$ over the complete range, and the controlling factor in establishing the tabular values of r for both functions. Then the values of the functions for each argument were calculated with seven-place logarithm tables.

The final step in the preparation of the table involved the calculation of the interpolation coefficients for each interval. This was accomplished on the type 602 calculating punch.

The straight lines which approximate the function, $f(r)$, have the particular form

$$f(r) = b_i - m_i (r - r_i), \qquad r_i \leqq r < r_{i+1}$$

where

$$m_i = \tan \alpha_i = \frac{f(r_i) - f(r_{i+1})}{r_{i+1} - r_i},$$

and

$$b_i = f_i - \epsilon .$$

Analogous formulas hold for the function $F(r)$, except that $B_i = F_i + \epsilon$ for the right-hand portion of the curve, since this portion of the curve is concave in the opposite direction.

The values of r, $f(r)$, and $F(r)$ were punched in a deck of cards and the values m_i and b_i for $f(r)$, and M_i and B_i for $F(r)$ calculated in two passes of the cards in the calculating punch. The values of b_i, B_i are rounded to six places and the values of m_i, M_i rounded to seven places. Only the five-place value of r, six-place values of b_i, B_i, and seven-place values of m_i, M_i are used in the final table deck for interpolating.

In practice the table cards are sorted in front of the detail cards on a common field r. A single interpolating control panel calculates the values $f(r)$ on one pass and $F(r)$ on a second pass. The second pass is preceded by an x-punched lead card which controls the reading and punching in the proper fields.

Three man-weeks were required to construct the table, but a five-place critical table of 69,900 cards was reduced to an interpolation table of 192 cards. Less time was required for preparation of this table than to key punch and verify a critical table. Use of the interpolation table greatly accelerates the process of table look-up.

DISCUSSION

Dr. Herget: Suppose that we are going to have a six-significant-figure table. There is no reason why still another significant figure cannot be added in the first term, for the purpose of avoiding the round-off error from this particular term; and, in addition to that, include an extra five, which is the half adjust for rounding.

Also, when these intervals are estimated, another method can be used. Take such a function as was shown; it was obviously easier to compute that than to compute any of its derivatives. Next, compute the function at fairly large equal intervals over the whole range and obtain the second difference. These numbers, divided by the square root of the reduced interval, give an estimate of the second derivative times the square of the interval.

Table Interpolation Employing the IBM Type 604 Electronic Calculating Punch

EVERETT KIMBALL, JR.

Massachusetts Institute of Technology

MUCH OF THE WORK done in the tabulating section of the U. S. Navy's Operations Evaluation Group involves small numbers of cards. Most of the jobs also require table look-up of such common functions as exponentials, logarithms and the trigonometric ratios. Often, however, it is necessary to process from 10 to 90 times as many master cards as detail cards, in seeking these functions, when employing the conventional table made up of arguments at equal intervals. The programming feature of the IBM Type 604 Calculating Punch has introduced the opportunity of interpolating intermediate values lying between entries of a table, and, of course, each person has his own favorite technique. Many systems of interpolation have been developed, and each appears to have advantages over others.

The objective of punched card tables, it seems to us, is one in which there are very few entries in order to reduce card handling time, and in which sufficient accuracy can be developed. In an attempt to reach this objective, we have developed a method of table construction and table interpolation which lends itself ideally to the small volume tabulating installation fortunate enough to possess a 604. Specifically, the OEG has developed abbreviated tables which permit accurate interpolation up to 8 digits.

Since this paper is on the application of punched card techniques to a mathematical problem, rather than an elegant development of a mathematical formula, the mathematics leading up to the formulation of the expression, used as a tool in this system, will be glossed over quickly.

The table chosen for illustration is the natural log table with the range of argument from 1.000 to 9.999, inclusive. Conventional equal interval argument tables for this function employ 9,000 cards, yet this method develops a file of but 327 cards. The conventional table can be of any accuracy required; the one developed is correct to 8×10^{-9}.

Mr. E. B. Gardner, in an unclassified memorandum to the director of research, OEG, has developed a three-term formula for the interpolation of unequal interval arguments, which is nicely adaptable to the 604. Essentially, this process is one employing a constant (pre-selected) difference between successive values of the function, rather than between the arguments. A formula based on the average of the Gauss "forward" and "backward" interpolation formulas, using divided differences, was developed.

The formula developed was:

$$Ux = A + Bx + Cx^2 ,$$

where Ux is the function of U of the argument x, and where $m < l \leqq x \leqq a < b$ and

$$A = \frac{U_l + U_a}{2} - \frac{l + a}{2}\,\Delta U_l + alC \qquad \Delta U_l = \frac{U_a - U_l}{a - l}$$

$$B = \Delta U_l - (a + l)C \qquad \text{and}$$

$$C = \frac{\Delta^2 U_m + \Delta^2 U_l}{2} \qquad \Delta^2 U_l = \frac{\Delta U_a - \Delta U_l}{b - l}$$

Once the initial interval between the values of the function has been determined, and the data recorded in some form for key punching, the cards were key punched, showing only the entry and the function of that entry. For ease in identification, these two fields will be called 1 and 2, respectively.

The detail processing of the master cards follows stepwise:

1. "Forward" gang punch fields 1 and 2 into fields 3 and 4. This process of "forward" gang punching, possible on the 521 punch, is a process which formerly required a reverse sequence sort before running through the conventional gang punch. This process transfers the information from a second card, into the first.
2. Compute, on the 604, ΔU, as shown in the formula above, punching the result into field 5.
3. "Forward" gang punch field 3 to field 6, and field 5 to field 7 (see card form).
4. Compute $\Delta^2 U$, punching the result in field 8.
5. "Backward" (conventional) gang punch $\Delta^2 U$ into field 9.

The card layout now contains the following information:

Field	1	2	3	4	5	6	7	8	9
	l	U_l	a	U_a	ΔU_l	b	ΔU_a	$\Delta^2 U_l$	$\Delta^2 U_m$

which is all the data needed to compute the three terms of the interpolation formula.

In three passes on the 604, C, B, and A are computed, in that order, punching the results as C 12.34×10^{-8}, B 12345.67×10^{-8}, A $123456789. \times 10^{-8}$. It is not necessary to show the exponent of 10, which has been introduced only to keep the decimal point on the card during programming, and is not actually used in the computation.

The usual steps of sorting and merging are employed to place the detail cards behind the proper master card (equal or lower), on the basis of the argument. The detail card argument, a five-place number, is merged against the entry, on the basis of the first four digits only. The value of the function to the fifth place of the argument will be interpolated.

The problem posed in the interpolation process was, primarily, a storage problem: 20 digits from a master card must be read into the 604, used, and held until the next master card was read.

The storage grouping finally employed was:

*FS*1-2, the first (left-hand) 8 digits of the factor A.

*FS*3, blank, for use during the computation of the interpolated value.

*FS*4, the 4 digits of C, and, in the right-hand position, the units position of A.

*GS*1-2, the seven digits of B.

*GS*3-4, for use during the computation of the interpolated value.

MQ, for the argument, X.

Program step:

1. *FS*4. *RO* to *Ctr* plus.
2. Counter *R&R* from 2nd to *GS*3-4. This pair of steps eliminates the A_9 digit.
3. Multiply *GS*3-4 by X, producing CX.
4. *RO Ctr* from 3rd to *GS*3-4, retaining seven digits.
5. *RO GS*3-4 to the *Ctr* 3rd, subtracting, to leave a residue of the 8th and 9th digits of CX in the counter.
6. *R&R Ctr* to *FS*3.
7. Multiply *FS*3 by X, producing the partial 8th through 14th digits of CX^2.
8. *R&R* from 5th position of *Ctr* to *FS*3, thus dropping decimals which have no effect on the final CX^2 calculation.
9. Multiply *GS*3-4 by X, producing the 1st through 12th positions of CX^2 (partial from 8 through 12).
10. *RO FS*3 into 3rd *Ctr*, completing the value of CX^2, digits 1-10.
11. *RO* from 5th position of *Ctr* to *GS*3-4 (digits 1-8 of CX^2).
12. *R&R* from 2nd position of *Ctr* to *FS*3 (digits 9-11).
13. Multiply *GS*1-2 by X, producing BX (digits 1-12).
14. *RO GS*3-4 into 5th of *Ctr*.
15. *RO FS*3 to *Ctr* 3rd.
16. *RO FS*1-2 (A, digits 1-8).
17. *RO FS*4 through shift read into 3rd, *RI* into *FS*3. Note: This shifts out the unwanted values of C which were to the left of the 9th digits of A.
18. *RO FS*3 into counter.
19. 1/2 adjust.

Punch from counter.

Sign control is necessary because A and/or B and/or C may be negative. The sign of A is carried in the factor storage sign hub 2, and controls normally the first 8 digits of A. The sign of B is carried normally in general storage 2, and operates normally. The sign of C is carried in the factor storage 4, and operates normally, although it introduces a difficulty in processing the sign for the 9th digit of A which is also carried in the same storage unit. If a negative sign is punched with an x punch, and if the master cards are also so punched, the solution is possible. By wiring the sign control columns for both A and C, in addition to the wiring indicated above, from first reading brushes, through column splits to the transferred points of a punch selector under the control of a master card indication punch, thence to the I pickup of two pilot selectors, the basic control is established for the sign control for the first detail card following the master card. By matching signs, as in multiplication, a third pilot selector may be picked up when, and only when, the 9th term of A is to add, as shown in Figure 1.

When this third pilot selector is normal, A_9 is to subtract. In order to effect this control on detail cards other than the first of a group, the 2nd reading brushes covering these same control columns are read, because the signs from the master card have been gang punched as the detail cards pass the punch station.

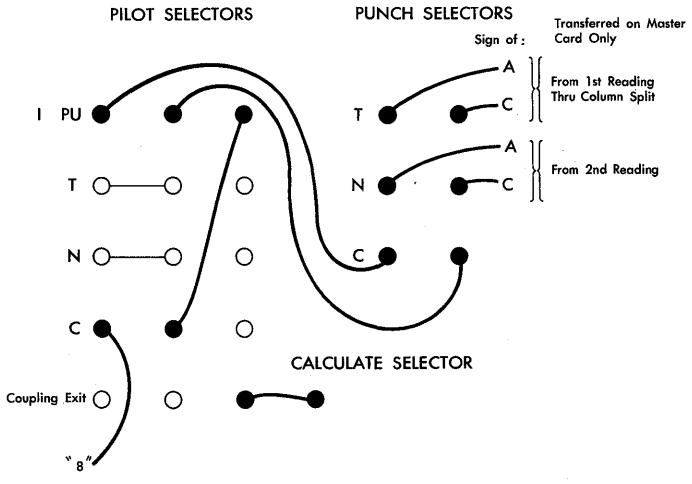


FIGURE 1

*An Algorithm for Fitting a Polynomial through n Given Points**

F. N. FRENKIEL H. POLACHEK

Naval Ordnance Laboratory, White Oak, Maryland

THE PROBLEM of fitting a polynomial through n given points has numerous applications in mathematical physics and engineering. In particular, this problem has become of considerable importance in the field of high-speed computation techniques where polynomial expressions are commonly used to replace any prescribed functions for interpolation purposes. The well-known Lagrangian interpolation formula, of course, can be used for this purpose. However, the expression obtained in this manner is not in a form usually suitable for computation, since it consists of a sum of products, each of which involves the variables at the known points. In this paper we seek a simple procedure or algorithm for calculating (once and for all) the coefficients $a_1, a_2, a_3, \ldots, a_n$ of the various terms of the polynomial which is given in the form,

$$y = a_1 + a_2 x + a_3 x^2 + \ldots + a_n x^{n-1} \qquad (1)$$

and which passes through n arbitrary points.

To make the results more accessible to the reader, we shall first state (without proof) the principal expressions used in our algorithm. Then we shall illustrate the use of this algorithm for a typical problem, and finally we shall give a mathematical derivation of the basic equations. We shall use abbreviated notation for certain familiar functions involving the n letters $x_1, x_2, x_3, \ldots, x_n$. We state these here.

Definition 1

$$ {}_kp^n = \frac{1}{(x_k-x_1)(x_k-x_2)\ldots(x_k-x_{k-1})(x_k-x_{k+1})\ldots(x_k-x_n)} $$

where

$$ {}_1p^1 = 1 $$

Definition 2

Elementary symmetric functions, p_i^n, on the n letters,

$x_1, x_2, \ldots, x_n$ taken i at a time.

Definition 3

Elementary symmetric functions, ${}_kp_i^n$, on the $(n-1)$ letters,

$x_1, x_2, \ldots, x_{k-1}, x_{k+1}, \ldots, x_n$ taken i at a time.

Definition 4

Sum, q_i^n, of all possible distinct products, $x_1^{\alpha_1} x_2^{\alpha_2} \ldots x_n^{\alpha_n}$ such that $\alpha_1 + \alpha_2 + \ldots + \alpha_n = i$ (α_j = positive integer or zero), i.e.

$$\begin{aligned} q_2^2 &= x_1^2 + x_1x_2 + x_2^2 \\ q_2^3 &= x_1^2 + x_2^2 + x_3^2 + x_1x_2 + x_1x_3 + x_2x_3 \\ q_3^2 &= x_1^3 + x_2^3 + x_1^2x_2 + x_2^2x_1 \\ q_5^1 &= x_1^5 \\ q_1^5 &= p_1^5 = x_1 + x_2 + x_3 + x_4 + x_5 \ . \end{aligned}$$

Basic Relations

Given n points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, it is required to obtain expressions for the coefficients $a_1, a_2, \ldots, a_n$ of the polynomial (1), which passes through these points. In terms of the functions defined above, these expressions may be given as follows

$$\begin{aligned} a_1 &= {}_1p^1y_1 - q_1^1a_2 - q_2^1a_3 - q_3^1a_4 - \ldots - q_{n-1}^1a_n \\ a_2 &= {}_1p^2y_1 + {}_2p^2y_2 - q_1^2a_3 - q_2^2a_4 - \ldots - q_{n-2}^2a_n \\ a_3 &= {}_1p^3y_1 + {}_2p^3y_2 + {}_3p^3y_3 - q_1^3a_4 - \ldots - q_{n-3}^3a_n \\ &\vdots \\ a_i &= {}_1p^iy_1 + {}_2p^iy_2 + \ldots + {}_ip^iy_i - q_1^ia_{i+1} \ldots - q_{n-i}^ia_n \\ &\vdots \\ a_n &= {}_1p^ny_1 + {}_2p^ny_2 + \ldots + {}_np^ny_n \ . \end{aligned} \qquad (2)$$

It will be noted that in these expressions only a_n is given in terms of known quantities, while a_{n-1} may be computed in turn from known quantities and a_n; a_{n-2} from a_{n-1} and a_n; etc. However, this can be carried out at the very end with very little additional work, after all the coefficients have been determined. The main job is to compute the p's and q's. We shall illustrate how this can be done systematically and with relatively little computational work for the case of a polynomial passing through seven points. From this example the general procedure for n points will be apparent.

*This project was sponsored by the Office of Naval Research.

Illustrative Example

We will apply our algorithm to the points $x_1 = 1$, $x_2 = 2^2, \ldots, x_7 = 7^2$; we will not fix the values of $y_1, y_2, \ldots, y_7$. Substituting in equations (2), we obtain

$$a_1 = {}_1p^1 y_1 - q_1^1 a_2 - q_2^1 a_3 - q_3^1 a_4 - q_4^1 a_5 - q_5^1 a_6 - q_6^1 a_7$$
$$a_2 = {}_1p^2 y_1 + {}_2p^2 y_2 - q_1^2 a_3 - q_2^2 a_4 - q_3^2 a_5 - q_4^2 a_6 - q_5^2 a_7$$
$$a_3 = {}_1p^3 y_1 + {}_2p^3 y_2 + {}_3p^3 y_3 - q_1^3 a_4 - q_2^3 a_5 - q_3^3 a_6 - q_4^3 a_7$$
$$a_4 = {}_1p^4 y_1 + {}_2p^4 y_2 + {}_3p^4 y_3 + {}_4p^4 y_4 - q_1^4 a_5 - q_2^4 a_6 - q_3^4 a_7$$
$$a_5 = {}_1p^5 y_1 + {}_2p^5 y_2 + {}_3p^5 y_3 + {}_4p^5 y_4 + {}_5p^5 y_5 - q_1^5 a_6 - q_2^5 a_7$$
$$a_6 = {}_1p^6 y_1 + {}_2p^6 y_2 + {}_3p^6 y_3 + {}_4p^6 y_4 + {}_5p^6 y_5 + {}_6p^6 y_6 - q_1^6 a_7$$
$$a_7 = {}_1p^7 y_1 + {}_2p^7 y_2 + {}_3p^7 y_3 + {}_4p^7 y_4 + {}_5p^7 y_5 + {}_6p^7 y_6 + {}_7p^7 y_7$$

The computation of the p's is straightforward. We can compute them column-wise. Thus, for instance,

$${}_1p^1 = 1, \quad {}_1p^2 = \frac{1}{x_1 - x_2}, \quad {}_1p^3 = \frac{1}{(x_1 - x_2)(x_1 - x_3)}, \text{ etc.}$$

For the computation of the q's the following procedure based on the relation (which will be proved later)

$$q_s^t = x_t q_{s-1}^t + q_s^{t-1} \tag{3}$$

should be used. First compute the q's in the first row, i.e.

$q_1^1 = x_1 = 1, q_2^1 = x_1^2 = 1, \ldots, q_6^1 = x_1^6 = 1.$

Secondly, compute the q's on the first diagonal (containing q's), i.e.

$q_1^1 = x_1 = 1, q_1^2 = x_1 + x_2 = 5, \ldots,$

$$q_1^6 = x_1 + x_2 + x_3 + x_4 + x_5 + x_6 = 91.$$

Finally, compute all other values of q by use of equation (3). Thus the remaining q's in the second row may be computed, in sequence, as follows:

$$q_2^2 = q_2^1 + x_2 q_1^2 = 21$$
$$q_3^2 = q_3^1 + x_2 q_2^2 = 85$$
$$\vdots$$
$$q_5^2 = q_5^1 + x_2 q_4^2 = 1365.$$

The q's in the third row are given by

$$q_2^3 = q_2^2 + x_3 q_1^3 = 147$$
$$q_3^3 = q_3^2 + x_3 q_2^3 = 1408$$
$$q_4^3 = q_4^2 + x_3 q_3^3 = 13013.$$

Similarly, the q's in all the succeeding rows may be obtained by this method. The values of the various coefficients for the above problem are given in the table below.

Proof of Basic Relations (2) *and* (3)

We shall now proceed to derive our basic relations (2) and (3) upon which our algorithm is based. We begin with the well-known Lagrange's interpolation formula which may be written in the form

$$y_1 = {}_1p^n [(x-x_2)(x-x_3)\ldots(x-x_n)]\, y_1 + {}_2p^n [(x-x_1)(x-x_3)\ldots(x-x_n)] y_2 + \ldots + {}_np^n [(x-x_1)(x-x_2)\ldots(x-x_{n-1})]\, y_n. \tag{4}$$

An obvious method of transforming Lagrange's equation (4) to the form of equation (1) is to multiply each of the above products and sum the coefficients of the like powers of x. This results in the following expressions for the coefficients $a_1, a_2, \ldots, a_n$:

$$(-1)^{n-1} a_1 = {}_1p^n\, {}_1p^n_{n-1}\, y_1 + {}_2p^n\, {}_2p^n_{n-1}\, y_2 + \ldots + {}_np^n\, {}_np^n_{n-1}\, y_n$$
$$(-1)^{n-2} a_2 = {}_1p^n\, {}_1p^n_{n-2}\, y_1 + {}_2p^n\, {}_2p^n_{n-2}\, y_2 + \ldots + {}_np^n\, {}_np^n_{n-2}\, y_n$$
$$\vdots$$
$$(-1)^{n-i} a_i = {}_1p^n\, {}_1p^n_{n-i}\, y_1 + {}_2p^n\, {}_2p^n_{n-i}\, y_2 + \ldots + {}_np^n\, {}_np^n_{n-i}\, y_n \tag{5}$$
$$\vdots$$
$$+ a_{n-2} = {}_1p^n\, {}_1p^n_2\, y_1 + {}_2p^n\, {}_2p^n_2\, y_2 + \ldots + {}_np^n\, {}_np^n_2\, y_n$$
$$(-1) a_{n-1} = {}_1p^n\, {}_1p^n_1\, y_1 + {}_2p^n\, {}_2p^n_1\, y_2 + \ldots + {}_np^n\, {}_np^n_1\, y_n$$
$$+ a_n = {}_1p^n\, y_1 + {}_2p^n\, y_2 + \ldots + {}_np^n y_n$$

These relations differ from equations (2) inasmuch as the coefficients are given directly in terms of $x_1, x_2, \ldots, x_n$; $y_1, y_2, \ldots, y_n$. From equations (5) it follows that relations (2) will hold if we can demonstrate that

$$(-1)^{n-i} a_i = {}_1p^n\, {}_1p^n_{n-i}\, y_1 + {}_2p^n\, {}_2p^n_{n-i}\, y_2 + \ldots + {}_np^n\, {}_np^n_{n-i}\, y_n = {}_1p^i y_1 + {}_2p^i y_2 + \ldots + {}_ip^i y_i$$
$$\vdots$$
$$+ (-1)^{n-i} q_1^i ({}_1p^n\, {}_1p^n_{n-i-1}\, y_1 + \ldots + {}_np^n\, {}_np^n_{n-i-1}\, y_n) \tag{6}$$
$$\vdots$$
$$+ q_{n-i-1}^i ({}_1p^n\, {}_1p^n_1\, y_1 + {}_2p^n\, {}_2p^n_1\, y_2 + \ldots + {}_np^n\, {}_np^n_1\, y_n) - q_{n-i}^i ({}_1p^n\, y_1 + {}_2p^n\, y_2 + \ldots + {}_np^n\, y_n)$$

$x_1 = 1$	$a_1 =$	$y_1 -$	$a_2 -$	$a_3 -$	$a_4 -$	$a_5 -$	$a_6 -$	a_7
$x_2 = 4$	$a_2 =$	$-\frac{1}{3} y_1 +$	$\frac{1}{3} y_2 -$	$5a_3 -$	$21a_4 -$	$85a_5 -$	$341a_6 -$	$1365a_7$
$x_3 = 9$	$a_3 =$	$\frac{1}{24} y_1 -$	$\frac{1}{15} y_2 +$	$\frac{1}{40} y_3 -$	$14a_4 -$	$147a_5 -$	$1408a_6 -$	$13013a_7$
$x_4 = 16$	$a_4 =$	$-\frac{1}{360} y_1 +$	$\frac{1}{180} y_2 -$	$\frac{1}{280} y_3 +$	$\frac{1}{1260} y_4 -$	$30a_5 -$	$627a_6 -$	$11440a_7$
$x_5 = 25$	$a_5 =$	$\frac{1}{8640} y_1 -$	$\frac{1}{3780} y_2 +$	$\frac{1}{4480} y_3 -$	$\frac{1}{11340} y_4 +$	$\frac{1}{72576} y_5 -$	$55a_6 -$	$2002a_7$
$x_6 = 36$	$a_6 =$	$-\frac{1}{302400} y_1 +$	$\frac{1}{120960} y_2 -$	$\frac{1}{120960} y_3 +$	$\frac{1}{226800} y_4 -$	$\frac{1}{798336} y_5 +$	$\frac{1}{6652800} y_6 -$	$91a_7$
$x_7 = 49$	$a_7 =$	$\frac{1}{14515200} y_1 -$	$\frac{1}{5443200} y_2 +$	$\frac{1}{4838400} y_3 -$	$\frac{1}{7484400} y_4 +$	$\frac{1}{19160064} y_5 -$	$\frac{1}{86486400} y_6 +$	$\frac{1}{889574400} y_7$

On the other hand, expression (6) will hold if we can prove the following two sets of identities

$$q^i_{n-i} - {}_kp^n_1\, q^i_{n-(i+1)} + {}_kp^n_2\, q^i_{n-(i+2)} - \cdots + (-1)^{n-(i+1)}\, {}_kp^n_{n-(i+1)}\, q^i_1 + (-1)^{n-i}\, {}_kp^n_{n-i} = (x_k - x_{i+1})(x_k - x_{i+2}) \cdots (x_k - x_n)$$

with $k = 1, 2, \ldots, i$ (7)

$$q^i_{n-i} - {}_kp^n_1\, q^i_{n-(i+1)} + {}_kp^n_2\, q^i_{n-(i+2)} - \cdots + (-1)^{n-(i+1)}\, {}_kp^n_{n-(i+1)}\, q^i_1 + (-1)^{n-i}\, {}_kp^n_{n-i} = 0$$

$k = 1, 2, \ldots, n$. (7′)

This is accomplished by carrying out the expansion of the ratio

$$\frac{(x-x_1)(x-x_2)\cdots(x-x_{k-1})(x-x_{k+1})\cdots(x-x_n)}{(x-x_1)(x-x_2)\cdots(x-x_k)\cdots(x-x_i)} = \frac{(x-x_{i+1})\cdots(x-x_n)}{(x-x_k)} \quad (8)$$

in two different ways and by comparing the coefficient of $(1/x)$.[a] By virtue of the remainder theorem the right-hand side of equation (8) may be written

$$\frac{(x-x_{i+1})\cdots(x-x_n)}{(x-x_k)} = \text{polynomial in } x + \frac{(x_k-x_{i+1})\cdots(x_k-x_n)}{x} + \frac{(x_k-x_{i-1})\cdots(x_k-x_n)}{x^2}x_k + \cdots . \quad (9)$$

On the other hand,

$$(x-x_1)(x-x_2)\cdots(x-x_{k-1})(x-x_{k+1})\cdots(x-x_n) = x^{n-1} - {}_kp^n_1\, x^{n-2} + {}_kp^n_2\, x^{n-3} - \cdots ; \quad (10)$$

and

$$\frac{1}{(x-x_1)(x-x_2)\cdots(x-x_i)} = \frac{1}{x^i}\frac{1}{\left(1-\frac{x_1}{x}\right)}\frac{1}{\left(1-\frac{x_2}{x}\right)}\cdots\frac{1}{\left(1-\frac{x_i}{x}\right)}$$

$$= \frac{1}{x^i}\left[1 + \frac{x_1}{x} + \left(\frac{x_1}{x}\right)^2 + \cdots\right]\left[1 + \frac{x_2}{x} + \left(\frac{x_2}{x}\right)^2 + \cdots\right]\cdots\left[1 + \frac{x_i}{x} + \left(\frac{x_i}{x}\right)^2 + \cdots\right]$$

$$= \frac{1}{x^i} + \frac{1}{x^{i+1}}q^i_1 + \frac{1}{x^{i+2}}q^i_2 + \cdots . \quad (11)$$

The coefficients of $1/x$ obtained by multiplying (10) by (11) are, precisely the left-hand side of identities (7) and (7′). The right-hand sides of equations (7) and (7′) are given by the coefficients of $1/x$ in (9). If $k = 1, 2, \ldots, i$ we obtain the expression given in (7). On the other hand, if $k = i+1, i+2, \ldots, n$ then the coefficients of $1/x$ in (9) are zeros.

Finally, we must justify the validity of the recursion relationship for the q functions given in equation (3). This can be accomplished by a simple argument based on the definition of these functions. q^{t-1}_s is the sum of all distinct products on the letters $x_1, x_2, \ldots, x_{t-1}$ raised to exponents $\alpha_{i_1}, \alpha_{i_2}, \ldots, \alpha_{i_{t-1}}$ such that $\Sigma\alpha_i = s$. q^t_s is a similar sum involving in addition the letter x_t.

In order to obtain q^t_s it is obvious that we must include in our summation all products contained in q^{t-1}_s. These do not involve the letter x_t. In addition, we must insert all products such that $\Sigma\alpha_i = s$ which do contain x_t. It is apparent that all terms $x_t\, q^t_{s-1}$ possess this property. Also, it is seen easily that all terms involving x_t, and such that $\Sigma\alpha_i = s$, are obtained in this manner. For, assume $x^{\beta_1}_1 x^{\beta_2}_2 \ldots x^{\beta_t}_t$ is a term such that $\Sigma\beta_i = s$, $\beta_t \geqq 1$, is not contained in $x_t\, q^t_{s-1}$. Then $x^{\beta_1}_1 x^{\beta_2}_2 \ldots x^{\beta_{t-1}}_t$ is not contained in q_{s-1}. This is contrary to the definition of q^t_{s-1} which must contain all terms such that Σ exponents $= s - 1$.

[a] For this elegant proof of identities (7) and (7′) the authors are indebted to D. Shanks of the Naval Ordnance Laboratory.

*The Monte Carlo Method and Its Applications**

M. D. DONSKER MARK KAC

Cornell University

CERTAIN PROBLEMS leading to complicated partial or integro-differential equations have recently been approached and some actually solved by utilizing various probability techniques and sampling methods. Collectively, these methods have become known as the Monte Carlo Method.

The problems to which Monte Carlo techniques have been applied seem to be divided into two types. Typical of the first type is the problem of neutrons diffusing in material media in which the particles are subjected not only to certain deterministic influences but to random influences as well. In such a problem, the Monte Carlo approach consists in permitting a "particle" to play a game of chance, the rules of the game being such that the actual deterministic and random features of the physical process are, step by step, exactly imitated by the game. By considering very large numbers of particles, one can answer such questions as the distribution of the particles at the end of a certain period of time, the number of particles to escape through a shield of specified thickness, etc. One important characteristic of the preceding approach is that the functional equation describing the diffusion process is by-passed completely, the probability model used being derived from the process itself.

A more sophisticated application of Monte Carlo Methods is to the problem of finding a probability model or game whose solution is related to the solution of a partial differential equation, or, as in the present paper, to determine the least eigenvalue of a differential operator by means of a sampling process. As an example of how the latter problem might be attacked, we quote from a paper of Metropolis and Ulam:[1]

"For example, as suggested by Fermi, the time independent Schrödinger equation

$$\triangle\phi(x, y, z) = (\lambda - V)\,\phi(x, y, z)$$

could be studied as follows. Reintroduce time dependence by considering

$$u(x, y, z, t) = \phi(x, y, z)\, e^{-\lambda t} \ ;$$

then, u will obey the equation

$$\frac{\partial u}{\partial t} = \triangle u - Vu \ .$$

This last equation can be interpreted, however, as describing the behavior of a system of particles each of which performs a random walk, i.e., diffuses isotropically and at the same time is subject to multiplication, which is determined by the value of the point function V. If the solution of the latter equation corresponds to a spatial mode multiplying exponentially in time, the examination of the spatial part will give the desired $\phi(x, y, z)$ — corresponding to the lowest 'eigenvalue' λ."[a]

The main purpose of the present paper is to present an alternative method for finding the lowest eigenvalue and corresponding eigenfunction of Schrödinger's equation. The chief difference between the two approaches is that ours involves only a random walk eliminating entirely the multiplicative process. This alteration in the model seems to simplify the numerical aspects of the problem, especially if punched card equipment is to be used. Apart from the possible numerical simplification, the method is based on a mathematical theory which in itself is of some interest.

The Mathematical Theory

Let $x_1, x_2, x_3, \ldots$ be independent identically distributed random variables each having mean 0 and standard deviation 1 and let $s_k = x_1 + x_2 + \ldots + x_k$. Under certain general assumptions on $V(x)$, the most severe of which is that $V(x)$ be non-negative, it can be shown[2] that the limiting distribution function $\sigma(\alpha,t)$ of the random variable

$$\frac{1}{n}\sum_{k\leqq nt} V\left(\frac{s_k}{\sqrt{n}}\right) \qquad (1)$$

is such that

$$\int_0^\infty \int_0^\infty e^{-\alpha-st}\, d_\alpha\sigma(\alpha,t)dt = \int_{-\infty}^\infty \psi(x)dx \ , \qquad (2)$$

*This paper (except for the two appendices) was written while the authors were associated with the National Bureau of Standards at the Institute for Numerical Analysis. It appears in the *Journal of Research of the National Bureau of Standards* under the title "A Sampling Method for Determining the Lowest Eigenvalue and the Principal Eigenfunction of Schrödinger's Equation." The preparation of the paper was sponsored (in part) by the Office of Naval Research.

[a]To the best of our knowledge, this method has not been tried out numerically.

where $\psi(x)$ is the fundamental solution of the differential equation

$$\frac{1}{2}\frac{d^2\psi}{dx^2} - [s + V(x)]\,\psi = 0\ , \tag{3}$$

subject to the conditions

$$\psi(x) \to 0 \qquad x \to \pm\infty$$
$$|\psi'(x)| < M \qquad x \neq 0$$
$$\psi'(+0) - \psi'(-0) = 2\ .$$

The fundamental solution $\psi(x)$ of (3) is expressible in terms of the normalized eigenfunction $\{\psi_j(x)\}$ and eigenvalues λ_j of the one-dimensional Schrödinger eigenvalue problem[b]

$$\frac{1}{2}\frac{d^2\psi}{dx^2} - V(x)\,\psi(x) = -\lambda\psi \qquad \lambda > 0 \tag{4}$$

as

$$\psi(x) = \sum_j \frac{\psi_j(0)\,\psi_j(x)}{s + \lambda_j}\ . \tag{5}$$

Thus, from (5) and (2)

$$\int_{-\infty}^{\infty}\psi(x)dx = \int_0^{\infty}\int_0^{\infty} e^{-\alpha - st}\,d_\alpha\sigma(\alpha,t)\,dt = \sum_j \frac{\psi_j(0)\int_{-\infty}^{\infty}\psi_j(x)dx}{s+\lambda_j}\ . \tag{6}$$

Inverting (6) with respect to s there results

$$\int_0^{\infty} e^{-\alpha}\,d_\alpha\sigma(\alpha,t) = \sum_j e^{-\lambda_j t}\,\psi_j(0)\int_{-\infty}^{\infty}\psi_j(x)dx\ , \tag{7}$$

and therefore the following expression for λ_1 is obtained,

$$\lambda_1 = \lim_{t\to\infty} -\frac{1}{t}\log\int_0^{\infty} e^{-\alpha}\,d_\alpha\sigma(\alpha,t)\ . \tag{8}$$

If in (7) all terms in the expansion but the first are neglected,

$$\log\int_0^{\infty} e^{-\alpha}\,d_\alpha\sigma(\alpha,t) \sim \log\psi_1(0)\int_{-\infty}^{\infty}\psi_1(x)dx - \lambda_1 t\ ,$$

or

$$\lambda_1 \sim \frac{\log\left\{\psi_1(0)\int_{-\infty}^{\infty}\psi_1(x)dx\right\}}{t} - \frac{1}{t}\log\int_0^{\infty} e^{-\alpha}\,d_\alpha\sigma(\alpha,t)\ . \tag{9}$$

Thus, if, by choosing a finite t, an attempt is made to calculate λ_1 from (8), there are two sources of error. The first, usually a small source of error, is from the exponentials neglected in the expansion (7). The second, and more important, is from neglecting the term $\frac{1}{t}\log\left\{\psi_1(0)\int_{-\infty}^{\infty}\psi_1(x)dx\right\}$. This latter source of error is especially significant since, as will be apparent shortly, it is impractical from other points of view to take t very large. All of this difficulty may be obviated by considering (7) for two distinct values of t, say t_1 and t_2; then, if the exponentials after the first are neglected as before on dividing there results,

$$\lambda_1 \sim \frac{1}{t_2 - t_1}\log\frac{\int_0^{\infty} e^{-\alpha}\,d_\alpha\sigma(\alpha,t_1)}{\int_0^{\infty} e^{-\alpha}\,d_\alpha\sigma(\alpha,t_2)}\ . \tag{10}$$

The Monte Carlo process consists in the calculation $\sigma(\alpha,t_1)$ and $\sigma(\alpha,t_2)$ by a sampling process.

If, instead of $\sigma(\alpha,t)$, the limiting distribution $\sigma_\xi(\alpha,t)$ of the random variable is considered,

$$\frac{1}{n}\sum_{k\leq nt} V\left(\xi + \frac{s_k}{\sqrt{n}}\right), \tag{11}$$

then $\sigma_\xi(\alpha,t)$ also satisfies (2) and (3), but now the condition $\psi'(+0) - \psi'(-0) = 2$, is replaced by $\psi'(\xi+) - \psi'(\xi-) = 2$. Therefore, repeating steps (4)-(7)

$$\int_0^{\infty} e^{-\alpha}\,d_\alpha\sigma_\xi(\alpha,t) = \sum_j e^{-\lambda_j t}\,\psi_j(\xi)\int_{-\infty}^{\infty}\psi_j(x)dx\ . \tag{12}$$

Thus,

$$\frac{\psi_1(\xi)}{\psi_1(0)} = \lim_{t\to\infty}\frac{\int_0^{\infty} e^{-\alpha}\,d_\alpha\sigma_\xi(\alpha,t)}{\int_0^{\infty} e^{-\alpha}\,d_\alpha\sigma(\alpha,t)}\ , \tag{13}$$

so that the principal eigenfunction also can be calculated.

The extension of the preceding method to multidimensional Schrödinger equations is immediate. It is in these cases that the method will probably prove to be most useful since, unlike the standard variational procedures, the extension to several dimensions seems to cause comparatively little difficulty. For illustrative purposes, consider Schrödinger's equation in three dimensions. Here three independent sequences

$$X_1, X_2, X_3, \ldots$$
$$Y_1, Y_2, Y_3, \ldots$$
$$Z_1, Z_2, Z_3, \ldots$$

of independent identically distributed random variables, each having mean 0 and standard deviation 1, must be considered. Let $s_{xk} = X_1 + X_2 + \ldots + X_k$, and s_{yk}, s_{zk} have the obvious meanings. Consider the limiting distributions $\sigma(\alpha,t)$ and $\sigma(\xi,\eta,\zeta;t)$ of the random variables

$$\frac{1}{n}\sum_{k\leq nt} V\left(\frac{s_{xk}}{\sqrt{n}}, \frac{s_{yk}}{\sqrt{n}}, \frac{s_{zk}}{\sqrt{n}}\right)$$

and

$$\frac{1}{n}\sum_{k\leq nt} V\left(\xi + \frac{s_{xk}}{\sqrt{n}}, \eta + \frac{s_{yk}}{\sqrt{n}}, \zeta + \frac{s_{zk}}{\sqrt{n}}\right),$$

respectively.

[b] From here on all the steps are formal. In all cases of physical interest they can be justified rigorously.

In exactly the same way as in the one-dimensional case

$$\lambda_1 \sim \frac{1}{t_2 - t_1} \log \frac{\int_0^\infty e^{-\alpha}\, d_\alpha \sigma(\alpha, t_1)}{\int_0^\infty e^{-\alpha}\, d_\alpha \sigma(\alpha, t_2)}$$

and

$$\frac{\psi_1(\xi, \eta, \zeta)}{\psi_1(0,0,0)} = \lim_{t\to\infty} \frac{\int_0^\infty e^{-\alpha}\, d_\alpha \sigma(\xi, \eta, \zeta; t)}{\int_0^\infty e^{-\alpha}\, d_\alpha \sigma(\alpha, t)} .$$

So far, the theory was carried out under the assumption that the potential function V was non-negative. In most cases of physical interest this is not so. For the hydrogen atom, for instance

$$V(x, y, z) = -\frac{\text{const.}}{\sqrt{x^2 + y^2 + z^2}} .$$

However, the modification is easy, although, as yet, all the points of mathematical rigor have not been clarified.

The formula for the lowest eigenvalue now becomes

$$\lambda_1 \sim \frac{1}{t_2 - t_1} \log \frac{\int_{-\infty}^0 e^{-\alpha}\, d_\alpha \sigma(\alpha, t_1)}{\int_{-\infty}^0 e^{-\alpha}\, d_\alpha \sigma(\alpha, t_2)}$$

and a corresponding modification needs to be made in the formula for the principal eigenfunction. We have not yet tested numerically any case with a negative potential function, but we hope to be able to report in the near future.

Numerical Examples and Discussion

The Monte Carlo procedure used here consists in the calculation of the distribution function $\sigma(\alpha,t)$ by a sampling process; the principal eigenvalue is then calculated from (10). For the purposes of numerical illustration consider two examples, $V(x) = x^2$ and $V(x) = |x|$. In both of these cases the eigenvalues are known; hence there is a check on the accuracy of our procedure. In order to calculate $\sigma(\alpha,t)$, say when $V(x) = x^2$, it is seen from (1) that the limiting distribution must be considered as $n \to \infty$ of

$$\frac{1}{n^2} \sum_{k \leqq nt} s_k^2 . \tag{14}$$

This means from our point of view that we must (a) choose a distribution for the X's; (b) choose a sufficiently large n; (c) select an appropriate t; (d) calculate for nt X's the normalized sum (14); (e) repeat (d) many times so that the empirical distribution may be obtained from these many samples.

Although, under the conditions mentioned previously, the distribution function $\sigma(\alpha,t)$ is independent of the distribution of the X's, the actual numerical calculation of $\sigma(\alpha,t)$ is expedited by choosing the distribution of the X's to be the Bernoulli distribution, i.e.,

$$P(X = 1) = P(X = -1) = \tfrac{1}{2} .$$

The sequence of random variables $X_1, X_2, X_3, \ldots$ is then a sequence of + 1's and − 1's such as might be obtained in coin tossing. This is conveniently and rapidly achieved on a calculating machine by considering sequences of random digits, counting even digits + 1 and odd digits − 1.

The value of n to be used must be large enough so that the empirical distribution function calculated is close to the theoretical limiting distribution function $\sigma(\alpha,t)$. From (10) it is seen that the two values of t, t_1 and t_2, to be used in the calculation of λ, must be large enough so that the exponential terms neglected are sufficiently small. However, since the sample size is nt, the desire to make both n and t large must be tempered by practical considerations. The number of samples to be used must be large enough so that the empirical distribution adequately represents $\sigma(\alpha,t)$. Before discussion of these points in more detail, let us consider an actual numerical computation. The following data for $V(x) = x^2$ and $V(x) = |x|$ were calculated from a certain set of random digits[c] on the IBM Type 604 Electronic Calculating Punch. For both x^2 and $|x|$ n was selected to be 400, $t_1 = 5$, $t_2 = 3.75$, and 100 samples were used. In column A_1 of Table I is tabulated

$$\frac{1}{8000}\sum_{k=1}^{1500} |s_k| , \quad A_2 = \frac{1}{8000}\sum_{k=1}^{2000} |s_k| ,$$

$$B_1 = \frac{1}{160000}\sum_{k=1}^{1500} s_k^2 , \quad B_2 = \frac{1}{160000}\sum_{k=1}^{2000} s_k^2 .$$

Since each RAND random number card contains 50 random digits and 2000 digits are needed to form one sample, a set of 4000 RAND cards was sufficient for this experiment. It takes 20 minutes on the 604 to calculate A_1 and A_2 for one sample (similarly for B_1 and B_2) so that it takes approximately 35 hours to secure the following data.

In order to calculate λ_1 from (10) the values of

$$\int_0^\infty e^{-\alpha}\, d_\alpha \sigma(\alpha, t_1) \quad \text{and} \quad \int_\infty^0 e^{-\alpha}\, d_\alpha \sigma(\alpha, t_2)$$

are needed. Both of these integrals were calculated numerically from the data in Table I by adding the exponentials of the entries in appropriate columns. For the case $V(x) = |x|$, the values of the integrals were obtained from

$$\frac{1}{100}\sum_{j=1}^{100} e^{-A_{1j}} \text{ and } \frac{1}{100}\sum_{j=1}^{100} e^{-A_{2j}} \text{ and for } V(x) = x^2 \text{ from}$$

$$\frac{1}{100}\sum_{j=1}^{100} e^{-B_{1j}} \text{ and } \frac{1}{100}\sum_{j=1}^{100} e^{-B_{2j}} .$$

In the case $V(x) = |x|$

[c] This set of random digits was prepared by the RAND Corporation, Santa Monica, California.

the true lowest eigenvalue to two places is .81 and in the case $V(x) = x^2$ it is $(\sqrt{2}/2) = .71$.

	first 50 samples	*second 50 samples*	*all 100 samples*
$\|x\|$	.83	.79	.81
x^2	.80	.69	.75

TABLE I

	A_1	A_2	B_1	B_2
1.	1.332	2.101	.738	1.290
2.	3.808	5.707	5.295	8.502
3.	2.795	5.545	3.460	9.610
4.	8.723	13.917	30.722	52.616
5.	4.169	4.766	6.123	6.602
6.	1.195	2.136	.598	1.453
7.	6.674	12.112	15.258	39.058
8.	4.103	5.242	5.826	7.004
9.	5.751	8.840	10.440	18.143
10.	4.250	4.981	6.069	6.679
11.	2.909	5.643	3.844	10.044
12.	2.834	3.416	3.602	4.097
13.	1.888	2.194	1.595	1.700
14.	2.022	2.337	1.510	1.638
15.	1.680	3.908	1.184	5.289
16.	7.700	12.712	24.769	44.936
17.	3.228	4.973	3.563	6.145
18.	1.844	2.654	1.523	2.183
19.	2.376	5.275	2.017	8.854
20.	4.533	5.640	6.380	7.616
21.	4.209	7.543	7.735	16.703
22.	3.847	5.590	6.666	9.204
23.	5.206	6.314	9.148	10.600
24.	1.962	2.698	1.859	2.923
25.	3.523	4.807	5.084	6.482
26.	3.605	5.397	5.014	7.680
27.	9.168	13.692	27.973	44.461
28.	5.625	7.898	11.528	16.010
29.	3.215	4.152	3.362	4.170
30.	3.293	3.972	4.136	4.677
31.	2.710	3.683	3.006	4.025
32.	3.638	7.386	6.114	17.413
33.	7.928	11.567	25.084	36.001
34.	3.475	3.846	4.396	4.550
35.	1.958	2.594	1.457	1.948
36.	5.356	7.216	8.888	11.781
37.	3.297	6.395	3.800	11.555
38.	1.883	2.682	1.387	1.995
39.	4.849	10.020	8.396	30.233
40.	2.282	3.067	2.258	2.882
41.	4.253	6.957	6.744	13.502
42.	7.559	8.834	17.143	18.702
43.	1.122	2.587	.500	2.634
44.	1.433	3.602	.869	4.855
45.	2.984	3.477	2.851	3.161
46.	1.677	3.158	1.025	3.341
47.	5.145	9.697	11.941	28.826
48.	4.982	9.140	9.201	23.092
49.	1.675	2.152	1.129	1.472
50.	1.580	3.375	1.147	3.807
51.	2.606	3.621	2.254	3.247
52.	4.303	5.883	7.987	10.104
53.	7.553	12.092	19.632	36.427
54.	3.138	5.703	3.968	9.325
55.	3.343	4.327	4.264	5.380
56.	1.423	2.315	.770	1.533
57.	7.996	11.458	21.405	31.585
58.	1.916	2.744	1.253	1.959
59.	1.490	2.161	.789	1.404
60.	3.536	5.382	4.125	7.461

TABLE I (*Continued*)

	A_1	A_2	B_1	B_2
61.	6.278	11.060	16.187	34.779
62.	1.174	1.709	.573	.919
63.	5.272	9.919	10.463	27.825
64.	3.576	5.277	6.021	8.707
65.	2.686	5.081	3.577	8.307
66.	6.166	8.785	14.056	19.627
67.	4.412	5.548	6.095	7.182
68.	1.732	4.664	1.170	8.720
69.	2.730	4.970	2.440	6.639
70.	1.331	2.772	.765	2.642
71.	3.668	5.318	5.730	8.001
72.	2.948	3.801	3.248	4.104
73.	5.279	8.931	12.455	23.513
74.	11.284	15.784	41.684	58.525
75.	2.183	2.787	2.086	2.455
76.	4.298	7.427	9.139	17.157
77.	1.412	2.893	.894	3.490
78.	2.002	2.567	1.602	2.024
79.	2.619	3.780	2.981	4.205
80.	2.695	4.616	2.974	6.077
81.	4.706	6.672	8.330	11.487
82.	5.517	7.981	9.846	14.805
83.	1.755	2.099	1.066	1.207
84.	2.084	2.876	1.735	2.504
85.	1.574	2.743	.993	2.256
86.	5.818	10.444	13.322	30.522
87.	4.387	7.844	9.037	18.695
88.	3.666	5.244	4.416	6.541
89.	7.111	13.311	19.834	50.822
90.	6.280	8.548	13.403	17.587
91.	4.872	5.836	9.230	10.023
92.	3.035	3.323	2.865	2.973
93.	3.608	4.697	4.256	5.357
94.	3.977	4.964	5.215	6.242
95.	3.614	7.867	5.757	20.434
96.	9.193	11.183	27.789	31.075
97.	4.108	5.282	5.918	7.374
98.	4.355	5.744	7.774	10.223
99.	1.898	2.661	1.507	2.013
100.	9.683	17.154	35.803	80.543

An interesting feature of the data is that column B_2, for example, may be thought of as arising from $n = 2000$, $t = 1$; $n = 1000$, $t = 2$, etc., as well as from $n = 400$, $t = 5$. The larger n is taken the closer will be the empirical distribution to $\sigma(\alpha,t)$. The value of $\int_0^\infty e^{-\alpha}\, d_\alpha\sigma(\alpha,t)$ in the case $t = 1$ and $V(x) = x^2$ can be calculated exactly and, to three places, is .678. Calculating this integral from column B_2 with $n = 2000$, $t = 1$ there results .685.

Instead of using Bernoulli distributed variables, one might use other distributions. One definite advantage of Bernoulli distributed variables is that the computation utilizes only the crudest properties of the random digits, i.e., whether they are even or odd. One possible advantage for certain other distributions is that n might not have to be taken so large. In particular this should be true if Gaussian distributed variables are used. RAND Gaussian deviates were used in constructing Table II. Here $t_1 = 3.75$, $t_2 = 5$ were chosen as before, but now $n = 100$. This means sam-

ples of size 500 instead of 2000 and, therefore, a total machine computation time of 9 hours for 100 samples.

Envisaging the possibility of calculating the second eigenvalue, we considered the quantities

$$10^{-4}\sum_{k=1}^{375}(s_k+5)^2 \text{ and } 10^{-4}\sum_{k=1}^{500}(s_k+5)\ , \qquad (15)$$

which correspond to taking $\xi = 0.5$ in (11). This should not make any difference in the calculation of the lowest eigenvalue; consequently, Table II can be utilized in the same way as Table I. It should be borne in mind, however, that columns C_1 and C_2 of Table II represent experimental values of the quantities (15) with the s_k's being sums of Gaussian deviates. Asterisks on the entries of Table II indicate that s_{375} or s_{500} was negative. Although this information is unnecessary for the purpose of calculations of the lowest eigenvalue, it is used in the calculation of the next eigenvalue. How this can be done is explained briefly in section 3.

TABLE II

	C_1	C_2		C_1	C_2
1.	3.963	13.997	44.	7.595	13.096
2.	2.348*	2.721	45.	2.336	2.862
3.	1.889*	2.231	46.	1.931*	2.579*
4.	3.817*	4.372*	47.	4.507*	8.294*
5.	2.395	2.802*	48.	8.581*	11.301*
6.	12.467	27.683	49.	4.698*	8.220*
7.	1.068*	1.115*	50.	10.154	16.884
8.	.578*	.967*	51.	8.828	15.387
9.	1.504*	1.760	52.	.997*	1.604*
10.	4.249	9.787	53.	8.582	12.838
11.	1.751*	3.316*	54.	2.730*	3.249*
12.	9.922	13.975	55.	10.471	13.140
13.	5.680	6.649	56.	1.456*	1.845*
14.	3.348	4.684	57.	8.230*	11.924*
15.	13.431	24.769	58.	5.515*	8.146*
16.	1.473*	2.615*	59.	2.202*	3.182*
17.	45.262	74.304	60.	2.129*	9.650*
18.	1.157*	3.109	61.	49.929	84.307
19.	.906*	3.644*	62.	3.095	6.220
20.	4.601	12.641	63.	2.585*	3.867*
21.	8.356*	13.084*	64.	1.318*	1.906*
22.	.308*	1.874*	65.	2.241*	9.102*
23.	2.111*	4.428*	66.	9.737	12.185
24.	2.035*	2.726	67.	9.738*	11.347*
25.	1.625*	3.898	68.	4.604	5.589
26.	1.879*	5.713	69.	20.364	29.707
27.	1.333	8.961	70.	1.097*	1.895*
28.	1.769	2.319*	71.	5.891	13.391
29.	1.906*	2.417*	72.	9.019*	12.815*
30.	7.522	7.979*	73.	1.843	6.280
31.	14.934	15.493*	74.	7.961	11.431
32.	4.282*	13.747*	75.	1.109	1.926
33.	1.356*	2.718	76.	2.417*	3.037
34.	4.448	8.329	77.	2.402*	2.797*
35.	1.306*	2.112*	78.	2.043*	4.214*
36.	5.465*	12.396*	79.	.582*	.752
37.	2.918	9.006	80.	2.429	4.381
38.	35.821	56.626	81.	8.584	13.034
39.	9.814*	11.707*	82.	2.793*	4.659*
40.	1.811	4.628	83.	1.557	5.023
41.	3.374	6.460	84.	1.006*	1.677*
42.	1.633	6.578	85.	5.327*	15.531*
43.	6.976	8.705	86.	2.058	7.307
87.	12.481	12.803*	144.	1.421*	3.652*
88.	.925*	2.403	145.	2.669*	3.889*
89.	15.519	19.879	146.	1.125*	2.071
90.	1.456*	4.009*	147.	3.439*	5.824*
91.	5.244	10.047	148.	1.269*	5.651
92.	2.497	2.931	149.	7.949	14.252
93.	16.700	17.382	150.	2.808*	4.434*
94.	4.471*	8.156*	151.	6.413	12.573
95.	2.139*	2.580*	152.	5.936*	10.432*
96.	1.196*	2.273*	153.	9.788	18.093
97.	23.014	28.420	154.	1.251	2.707
98.	6.959*	7.022*	155.	34.799	63.494
99.	3.374*	8.651*	156.	.841*	1.742*
100.	1.600*	2.250*	157.	1.461*	1.615*
101.	4.626*	7.794*	158.	7.220	21.125
102.	5.195*	10.215*	159.	1.275*	2.359*
103.	6.837*	12.160*	160.	9.355	19.565
104.	1.913	4.412*	161.	3.546	4.040*
105.	2.276*	2.770	162.	1.238*	2.304*
106.	1.935*	3.983*	163.	.975*	1.283
107.	3.560	5.808	164.	1.980	3.659*
108.	2.857*	7.384*	165.	14.620*	17.565*
109.	3.399*	4.484*	166.	2.378*	2.865
110.	6.917*	27.107*	167.	2.279	6.279
111.	38.661	56.479	168.	10.846	13.754
112.	1.891	2.000*	169.	1.341*	3.673*
113.	8.506	9.646*	170.	1.861*	2.729*
114.	4.209	10.461	171.	28.588*	53.358*
115.	33.194	44.247	172.	2.943	3.087*
116.	.869*	1.289	173.	1.694	2.568
117.	3.946	15.204	174.	1.729	2.516*
118.	1.680*	2.995*	175.	.844*	1.449*
119.	8.178*	9.056	176.	3.441*	5.380*
120.	27.621	51.650	177.	.712	.838*
121.	8.444*	15.836*	178.	1.222*	3.469*
122.	13.254	16.467	179.	20.852*	42.263*
123.	1.755*	5.246*	180.	7.575*	16.552*
124.	36.902	70.134	181.	2.042*	3.058
125.	7.567*	18.553*	182.	6.797	7.343*
126.	14.725	38.551	183.	1.536	4.499
127.	.983*	2.572	184.	1.235	1.472*
128.	24.529	31.660	185.	3.541*	4.272
129.	2.042*	2.385*	186.	5.394*	6.149*
130.	3.587*	9.858*	187.	18.589	28.787
131.	1.335*	1.755	188.	8.372	12.628
132.	2.142*	3.801*	189.	13.470*	23.775*
133.	16.361	27.759	190.	1.196*	2.851*
134.	8.193	9.301	191.	3.507	7.712
135.	2.719*	5.141*	192.	21.489	43.372
136.	4.767*	12.737*	193.	1.038*	1.109*
137.	42.230	108.038	194.	3.591*	3.864
138.	41.236	78.423	195.	2.295*	2.529*
139.	8.243	9.911*	196.	1.044*	2.320
140.	31.522	36.393	197.	4.865*	6.759*
141.	2.600	2.881*	198.	1.184	11.852
142.	1.932	2.055*	199.	1.228*	1.969*
143.	8.562	10.813	200.	9.346*	22.204*

Using the data in Table II and again (10), we computed the following approximations to λ_1 (actual value = .71).

	first 50 samples	*second 50 samples*	*third 50 samples*	*fourth 50 samples*	*all 200 samples*
λ_1	.82	.72	.83	.64	.74

For both $V(x) = |x|$ and $V(x) = x^2$ all the eigenvalues are known, so that in the two illustrative examples above it was easy to choose appropriate values for t_1 and t_2. The proper value for n and the appropriate number of samples were determined experimentally. In general, when all eigenvalues are unknown, the following rule-of-thumb procedure seems to be feasible. Having first made a guess at an appropriate n and t, and a certain number of samples, compute λ_1. Repeat the computation now for the same n, the same number of samples and successively larger t's until the calculated values of λ_1 become stable to the desired number of places. If they do not become stable, the number of samples must be increased. Keeping two values of t, for which the calculated λ's had the stable value, increase n and see if the calculated value of λ changes. If not, n is sufficiently large. If it does change, increase n until a new stability appears. This stable value is then the appropriate approximation to λ_1. The mere fact that stability is observed means the number of samples is sufficiently large.

The Second Eigenvalue

If the principal eigenfunction is even, then it is possible to extend the theory of section 1 in such a way that the calculation of the second eigenvalue becomes feasible. Without going into any details, we just state the pertinent result: Let

$$E_t^+(\xi) = \lim_{n\to\infty} E\left\{ e^{-\frac{1}{n}\sum_{k\leqq nt} V\left(\xi+\frac{s_k}{\sqrt{n}}\right)} \,\middle|\, s_{nt} > 0 \right\} \tag{16}$$

and

$$E_t^-(\xi) = \lim_{n\to\infty} E\left\{ e^{-\frac{1}{n}\sum_{k\leqq nt} V\left(\xi+\frac{s_k}{\sqrt{n}}\right)} \,\middle|\, s_{nt} < 0 \right\}, \tag{17}$$

the mathematical expectations on the right-hand sides being conditional expectations under the conditions $s_{nt} > 0$ and $s_{nt} < 0$, respectively.

Thus

$$\frac{1}{2}\{E_t^+(\xi) - E_t^-(\xi)\} = \sum_{j=2}^{\infty} e^{-\lambda_j t}\,\psi_j(\xi) \int_{-\infty}^{\infty} \rho(x)\,\psi_j(x)\,dx\,, \tag{18}$$

where

$$\rho(x) = \begin{cases} +1, & x > 0 \\ -1, & x < 0\,. \end{cases}$$

If t_1 and t_2 are sufficiently large,

$$\lambda_2 \sim -\frac{1}{t_2 - t_1} \log \frac{E_{t_2}^+(\xi) - E_{t_2}^-(\xi)}{E_{t_1}^+(\xi) - E_{t_1}^-(\xi)}\,. \tag{19}$$

From the discussion of section 2 it should be clear how one applies (19) when the data of Table II are available. One must see to it that ξ is so chosen that $\psi_2(\xi) \neq 0$; otherwise a higher eigenvalue may have been calculated. From the data of Table II the following is obtained

$$\lambda_2 \sim 1.1\,,$$

whereas the exact value is $\sqrt{2} = 1.41$. The poor agreement could have been expected in view of low accuracy in the calculation of λ_1.

In conclusion, we wish to thank Dr. E. C. Yowell of the National Bureau of Standards for wiring the control panels and for excellent supervision of all the punched card work.

Appendix I

We give here an intuitive approach to the mathematical theory of section 1. This approach was suggested to us by Dr. G. A. Hunt.

Consider the simple random walk in steps

$$\pm\,\Delta\ (\Delta = 1/\sqrt{n})$$

each step being of duration τ. At each stage the particle has the probability $\tau\,V(s_k\,\Delta) = (1/n)\,V(s_k/\sqrt{n})$ of being destroyed, where $s_k\,\Delta$ is the displacement of the particle after time $k\,\tau$ (k steps).

In the limit as $n\to\infty$ ($\Delta\to 0$, $\tau\to 0$, $\Delta^2/\tau = 1$) we are led to a continuous diffusion process with destruction of matter governed by the function $V(x) \geqq 0$.

The probability $Q(x,t)dx$ that the particle will be found between x and $x + dx$ at time t can be found by calculating the Green's function (fundamental solution) of the differential equation

$$\frac{\partial Q}{\partial t} = \frac{1}{2}\frac{\partial^2 Q}{\partial x^2} - V(x)Q\,,$$

i.e., that solution of the equation which for $t\to 0$ satisfies

$$Q(x,t)\to\delta(x)\,.$$

The integral

$$\int_{-\infty}^{\infty} Q(x,t)\,dx$$

represents the probability that the particle will survive during the time interval $(0,t)$.

In terms of eigenvalues and normalized eigenfunctions of the Schrödinger's equation (4) we can express $Q(x,t)$ as follows:

$$Q(x,t) = \sum_j e^{-\lambda_j t}\,x_j(0)\,x_j(x)\,.$$

It is, of course, understood that $V(x)$ is assumed to yield a discrete spectrum. In the case when continuous spectrum is also present the formula has to be modified but the calculations of the lowest eigenvalue are, in general, not affected. Finally,

$$\int_{-\infty}^{\infty} Q(x,t)\,dx = \sum_j e^{-\lambda_j t}\,\psi_j(0) \int_{-\infty}^{\infty} \psi_j(x)\,dx\,,$$

and it remains to verify that

$$\int_{-\infty}^{\infty} Q(x,t)\,dx = \int_0^{\infty} e^{-\alpha}\,d_\alpha\sigma(\alpha;t)\,.$$

First note that the expectation (average) of

$$e^{-\frac{1}{n}\sum_{k\leqq nt} V\left(\frac{s_k}{\sqrt{n}}\right)} = e^{-\tau\sum_{k\leqq nt} V(s_k\,\Delta)}$$

approaches

$$\int_0^\infty e^{-\alpha}\, d_\alpha \sigma(\alpha;t)$$

as $n\to\infty$ (because the distribution function of $\frac{1}{n}\sum_{k\leqq nt} V\left(\frac{s_k}{\sqrt{n}}\right)$ approaches $\sigma(\alpha;t)$ and $V(x)\geqq 0$).

On the other hand, using the approximation

$$e^{-\tau V(s_k\Delta)} \sim 1-\tau V(s_k\Delta)$$

note that

$$e^{-\tau \sum_{k\leqq nt} V(s_k\Delta)}$$

is approximately the probability of survival of the particle if its consecutive displacements are $s_1\Delta, s_2\Delta, \ldots, s_{nt}\Delta$. In taking the expectation of

$$e^{-\tau \sum_{k\leqq nt} V(s_k\Delta)}$$

we average the probability of survival over all possible choices of successive displacements (all possible paths) and thus obtain the unconditional probability of survival. This unconditional probability of survival approaches, as $n\to\infty$, the integral

$$\int_{-\infty}^{\infty} Q(x,t)dx \ .$$

On the other hand, it also approaches

$$\int_0^\infty e^{-\alpha}\, d_\alpha \sigma(\alpha;t)$$

and consequently

$$\int_0^\infty e^{-\alpha}\, d_\alpha \sigma(\alpha;t) = \int_{-\infty}^{\infty} Q(x,t)dx$$

$$= \sum_j e^{-\lambda_j t}\psi_j(0)\int_{-\infty}^{\infty}\psi_j(x)dx \ .$$

Although it has been assumed $V(x)\geqq 0$, all considerations are applicable to potentials which are bounded from below. Although atomic and molecular potentials become negatively infinite, they can be cut off sufficiently low without changing appreciably the eigenvalues and the eigenfunctions.

Appendix II

In this appendix we sketch a Monte Carlo Method for finding the lowest frequency of a vibrating membrane. In mathematical terms, find the lowest eigenvalue of the equation

$$\frac{1}{2}\Delta u + \lambda u = 0 \ ,$$

valid in a region Ω and subject to the boundary condition

$$u = 0 \ ,$$

on the boundary Γ of Ω. (This corresponds to the case of the clamped membrane.) The lowest frequency is the square root of twice the lowest eigenvalue.

Cover the region Ω with a square net, the side of the square being Δ. Starting from a point $(x_0, y_0) = (m\Delta, n\Delta)$ inside Ω, consider a two-dimensional random walk in which a point is allowed to move to each of its four nearest neighbors, the choices being equiprobable (probability 1/4). The duration of each step is τ related to Δ by the equation

$$\frac{\Delta^2}{\tau} = 1 \ .$$

Consider the boundary Γ of Ω as an absorbing barrier. This means that whenever the particle, performing the random walk, leaves the region Ω it is destroyed.

In the limit as $\Delta\to 0$, the probability $Q(x_0, y_0; t)$ that the particle will survive during the time interval $(0,t)$ can be obtained as follows:

$$Q(x_0, y_0; t) = \iint_\Omega P(x_0, y_0 \mid x, y; t)\, dx\, dy \ ,$$

where $P(x_0, y_0 \mid x, y; t)$ is the fundamental solution of the differential equation $\frac{\partial P}{\partial t} = \frac{1}{2}\Delta P$,

subject to the boundary condition

$$P = 0 \text{ on } \Gamma \ ,$$

and the initial condition

$$P(x_0, y_0 \mid x, y; t) \to \delta(x-x_0)\,\delta(y-y_0) \ ,$$

as $t\to 0$.

This fundamental solution can be expressed in terms of the eigenvalues and normalized eigenfunctions of the membrane problem as follows:

$$P(x_0, y_0 \mid x, y; t) = \sum_j e^{-\lambda_j t}\psi_j(x_0, y_0)\,\psi_j(x, y) \ .$$

Thus

$$Q(x_0, y_0; t) = \sum_j e^{-\lambda_j t}\psi_j(x_0, y_0)\iint_\Omega \psi_j(x, y)dx\, dy \ ,$$

and

$$\lambda_1 = -\lim_{t\to\infty}\frac{\log Q(x_0, y_0; t)}{t} \ .$$

The probability $Q(x_0, y_0; t)$ can be calculated by a sampling method in the following way. Start N_0 independent particles from (x_0, y_0) and let each of them perform the random walk described above. Watch each particle and count the number N_t of those particles which have not left the region during the time interval $(0,t)$.

Then

$$Q(x_0, y_0; t) \sim \frac{N_t}{N_0}$$

and

$$\lambda_1 \sim -\frac{1}{t}\log\frac{N_t}{N_0} \ .$$

The practicality of this method has not been tested.

References

1. N. Metropolis and S. Ulam, "The Monte Carlo Method," *Journal of the American Stat. Asso.* (September, 1949), pp. 335-341.
2. Mark Kac, "On Distributions of Certain Wiener Functionals," *Trans. Am. Math. Soc. 65* (1949), pp. 1-13.

DISCUSSION

Mr. Bisch: I was very much interested in the clear discussion and the problem you chose, which is part of a problem we have to solve. My first question concerns that function, $V(x)$ or $V(x, y, z)$. Could you have that function determined experimentally—in other words, not expressed algebraically?

Professor Kac: Yes.

Mr. Bisch: The second question is about the boundary: Could you in this method leave the boundary as a simple unknown temporarily?

Professor Kac: To which problem are you referring? Do you mean here there is no boundary?

Mr. Bisch: There is no boundary in a problem of the membrane. You made the boundary zero. In other words, your deflection was zero. Could you leave that deflection temporarily unknown as a quantity like U_0?

Professor Kac: I think so. Actually, all I can say with certainty is the following: If you have the other boundary condition, $(du/dn) = 0$, the other classical condition, then all you have to do is to have the boundary not absorbing but reflecting. Now, the mixed boundary $au + b\ (du/dn) = 0$ can again be done in principle by making the boundary partially reflecting, partially absorbing. When you come to the boundary you play an auxiliary game, which would decide whether you are going to throw the particle out or keep it in the game. You see, this is only an eigenvalue problem. Consequently, you cannot just say that the solution must be $f(x)$ on the boundary, because that would not give a characteristic value problem, and this is designed primarily to find eigenvalues. On the other hand, if it comes to Laplace's equation with a prescribed condition, then Dr. Yowell will speak about a random walk method which will do that. In fact, they have some experiments in the case of Laplace's equation in a square.

Dr. King: One should emphasize the point of view of accuracy. I don't believe there is any hope of getting, say, six significant figures out of a Monte Carlo Method.

Professor Kac: Agreed.

Dr. King: But I disagree a bit with your point of view that it is worth doing, even in the very simplest cases, if you are not interested in accuracy. I think for the same amount of work you could work the harmonic oscillator with other methods and get one significant figure or two.

Professor Kac: If I understood you correctly, you are saying that apart from the practical applications it is interesting because of the principle involved. Is that correct?

Dr. King: Yes.

Professor Kac: There I definitely agree.

Dr. King: I think it is still practical, though, apart from being an amusing experimentation; it is a practical method if you are interested in an engineering problem; so you only need a couple of figures.

Professor Kac: With that I agree. As a matter of fact, it has another advantage, actually, which bore some fruit, not particularly exciting, but this way of looking at it produces even results of some theoretical interest.

For instance, I am able—although I won't do it here because it will be a bit too technical and probably too tedious—to justify the so-called W K B method, at least one aspect of it, by diffusion analogies; and there are other new viewpoints. If you can look at something from different points of view, it is certainly helpful, and often practical.

On the other hand, for someone who has the tough attitude ("You give me the lithium molecule to ten per cent, or I won't take the method"), of course, one would still have to see what one can do, and, actually, I agree with you. What I am trying to do is to bend backwards in being cautious about it.

Professor Tukey: It seems to me that there is a point to be made that came out in the discussion at the last conference. That is, that one point of view for the use of Monte Carlo in the problem is to quit using Monte Carlo after a while. That, I think, was the conclusion that people came to then. That was the natural evolution and perhaps the desirable thing. After you play Monte Carlo a while, you find out what really goes on in the problem, and then you don't play Monte Carlo on that problem any more.

I think the thing to suggest here is that, by the time people have played Monte Carlo on the lithium atom, perhaps, or the lithium molecule, or something more complicated, people will get to the place where they won't be playing this simple Monte Carlo any more; they will be playing Monte Carlo in some peculiar space where you have obtained approximations to the wave functions as your coordinates, and not x, y, and z; and then you will start to get more for a given amount of machine time.

This is going to get arbitrarily complicated. When you start to cross-breed Monte Carlo with a standard method—and that isn't too far away in the handling of hard problems—you are going to have to do something like that.

Professor Kac: There is confirmation of that because of rather extensive calculations performed with relatively simple equipment built for that purpose at Cornell, by Dr. Wilson and collaborators, in connection with his study of cosmic ray showers. They found the Monte Carlo Method most valuable because it showed them what goes on. I mean the accuracy was relatively unimportant. The five per cent or the seven per cent accuracy they obtained could be considered low; but all of a sudden they got a certain analytic picture from which various guesses could be formulated, some of them of a purely analytical nature, which later on turned out to verify very well. As a matter of fact, that is certainly one aspect of Monte Carlo that should be kept in mind. I agree that one of the purposes of Monte Carlo is to get some idea of what is going on, and then use bigger and better things.

*A Punched Card Application of the Monte Carlo Method**

P. C. JOHNSON F. C. UFFELMAN

Carbide and Carbon Chemicals Corporation

IN THIS PAPER a description will be given of a punched card technique for applying the Monte Carlo Method to certain neutron problems. It should be kept in mind that this procedure is designed for a machine installation consisting of a standard type 602 or 604 calculating punch, a collator, a reproducer, a sorter, and an accounting machine. Other combinations of machines would require a different approach, and an installation with a card-programmed electronic calculator would use an entirely different technique from the one about to be described.

In any event, the problem may be stated as follows:

Assuming a monochromatic point source of neutrons of given energy within an infinite medium of known constituents, hypothetical case histories for a number of these neutrons will be built up as they undergo a series of random collisions with the constituents of the selected medium. These collisions result in either an absorption or an elastic scattering of the neutrons, and the main work of this problem is to follow the selected neutrons through successive collisions until they either are absorbed or fall below a certain energy level.

For each collision of each neutron the following will be recorded: the type of collision undergone; the energy loss per collision, $\triangle u$; the total energy loss at the end of the current collision, u; the z-component of the distance traveled, ρ, between collisions, $\triangle z$; the z-component of the total distance traveled, z; the angle of deflection after the collision, ω; and the direction cosine with the z-axis, μ.

When these data have been recorded, they may be used for various statistical studies to determine such things as, how far neutrons of a given energy may be expected to penetrate a certain substance. An example of how these data may be used will be given subsequently.

To begin any such series of calculations, an arbitrary number of neutrons—say 1,000 numbered from 000 to 999—are selected and permitted, on punched cards, to go off in a random direction, travel a random distance, hit a type of particle determined in a random fashion, and lose an amount of energy determined also in a random way. For each neutron, a collision card is made bearing the neutron number, the collision number, and all the data (distance, direction, etc.) pertaining to that particular collision. Then the neutrons are again permitted to go off in a random direction, travel a random distance, hit some type of particle, and lose energy. These data are recorded on a second set of collision cards, one for each neutron; in addition, summary data for the neutron's history to date are computed, such as total distance traveled along the z-axis during both collisions, total energy lost during both collisions, and direction cosine with the z-axis. In this manner the neutrons are carried from collision to collision, until all of them either are absorbed in undergoing a collision, or drop through successive collisions, below a stipulated energy level, at which point they are no longer useful for the purposes of the problem.

In the sequence of calculations described here, the energy loss, u, and the direction cosine, μ (both of which are independent of distance), will be calculated first for each collision of all neutrons. The distance, z, which is directly dependent on μ and indirectly upon u, is then calculated for successive collisions of each neutron; z could be calculated at the same time that u and μ are calculated, if the capacity of the machines used so permitted, or u, μ, and z might be calculated in separate, successive operations in that order if machine capacity is limited. It is understood that in actual practice many of the operations which will be described as separate steps should be combined; the method of combination will depend on the machines available, the training of the operator handling the problem and, of course, the particular problem involved (magnitude of numbers, etc.).

For this particular example, one thousand neutrons will be traced in a medium of two constituents. An elastic scattering as the result of a collision with an atom of one kind will be designated a type 1 collision, with the other type of atom a type 2 collision. Any collision resulting in absorption will simply be called absorption, without regard to type of atom hit.

Three decks of cards are assembled for the preparation and actual computation of the problem described. They are:

1. Probability cards (card layout 1)
2. Master cards (card layout 2)
3. Collision cards (card layout 3)

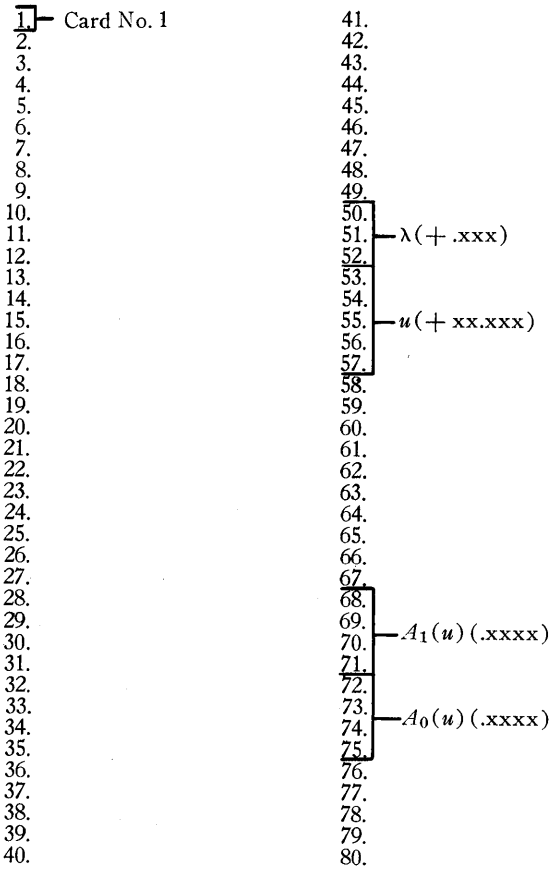
*This paper was presented by Edward W. Bailey.

CARD LAYOUT 1

Title: PROBABILITY CARD

Card No. 1 Card Color: Solid Blue

Source: Probability Tables Computed for the Particular Problem



$A_1(u)$ is the probability of a collision of type 1 or of absorption. $A_0(u)$ is the probability of absorption. The probability of a collision of type 2 is 1.0000—$A_1(u)$. All values on the card are key punched from tables.

Probability Cards

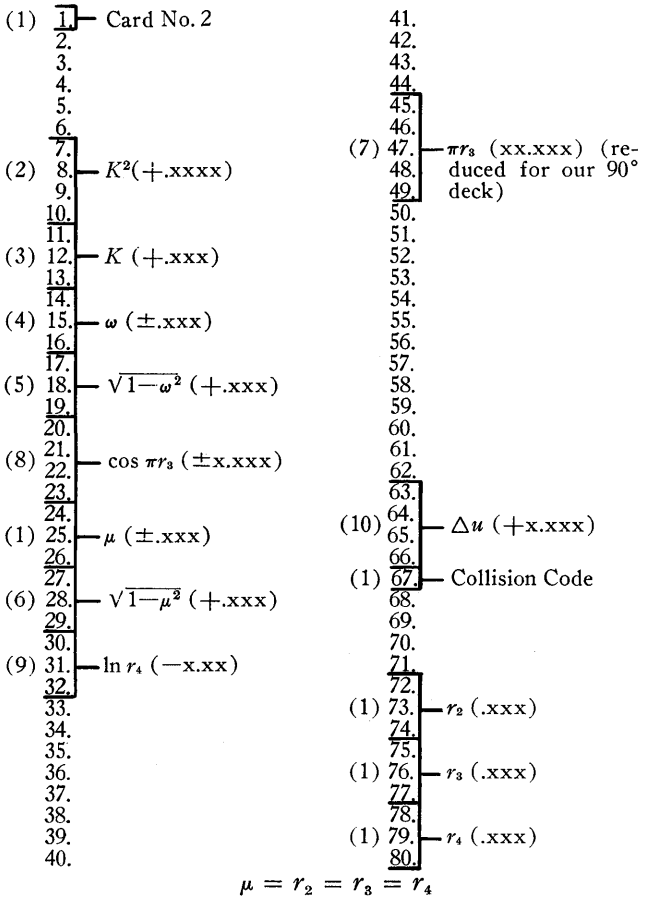
The probability cards (card layout 1), of which there are some twenty or thirty, are key punched from probability tables calculated for the particular problem at hand. Each card contains, in columns 53-57, a value u which is the lower value of the energy range which the card represents, and in columns 68-71 and 72-75 the probabilities associated with the particular energy range, of absorption or scattering with the different types of particle involved. Thus, $A_1(u)$ is the probability of a type 1 scattering or of absorption, and $A_0(u)$ is the probability of absorption. The probability of a type 2 scattering is then 1.000 − $A_1(u)$. Probability card 1 also contains the value λ which is the mean free path, or probability of scattering within a given distance, associated with the energy range.

CARD LAYOUT 2

Title: MASTER CARD

Card No. 2 Card Color: Solid Red

Source: Various Tables and Computations



The number of the operation in which a field is derived is given in parentheses to the left of the card columns. See "Master Cards" Operations for step involved.

Master Cards

The deck of master cards, of which there are 2,000 in this example (card layout 2), contains all the tabled values, such as functions of angles and natural logarithms, which must be searched from time to time during the problem, as well as some calculated values which appear over and over again, and hence can be intersperse gang punched more readily than calculated each time. The arguments on the master cards are μ, r_2, r_3, and r_4, and on any one card the four arguments are the same: they are put in the four different fields as a convenience to the operator, the fields on the master card being then lined up with those on the collision card. Since all of the arguments are three-digit numbers, our master deck will consist of 1,000 cards representing the 1,000 different choices of a three-digit number (000-999, inclusive) and having a collision type code of 1,

and another 1,000 cards representing all possible random digits, having a collision code of 2 (refer to master card table, step 1). In addition to the 4 three-digit arguments, the master card will also contain the following which are numbered in accordance with the step number on the table of operations referring to the master card.

Step No.	*Function*
2	$K^2 = 1 - 4r_2M(M+1)^{-2}$, where the value of M is a constant depending on the collision type, is calculated and punched in the master cards;
3	K is then intersperse gang punched into the master cards from a square root deck;
4	$\omega = [(M+1)K - (M-1)/K]/2$ is calculated on the 604 and punched;
5	$\sqrt{1-\omega^2}$ is intersperse gang punched from a sin cos deck;
6	$\sqrt{1-\mu^2}$ is intersperse gang punched from a sin cos deck;
7	r_3 is multiplied by π, reduced to an angle less than 90°, the absolute value of whose cosine equals that of πr_3, and the angle is punched in the card while the sign is punched in column 23;
8	$\cos \pi r_3$ is intersperse gang punched into columns 20-23 from the sin cos deck;
9	$\ln r_4$ is intersperse gang punched from an ln deck;
10	$\Delta u = -\ln k^2$ is intersperse gang punched from an ln deck.

This completes the data on the master cards. It should be noted that a number of the steps could be combined in actual computation, such as steps 1, 2, 4, and 7. Also, in the absence of suitable function decks, some of the functions could be calculated instead of intersperse gang punched.

CARD LAYOUT 3

Title: COLLISION CARD—
BASIC "RANDOM NUMBER" CARD

Card No. 3 Card Color: Plain Manila

Source: Computations and Random Number Tables

(1)	1.	— Card No. 3	
	2.		
	3.		
	4.		
	5.		
	6.		
	7.		
	8.		
	9.		
	10.		
	11.		
	12.		
	13.		
	14.		
	15.		
	16.		
	17.		
	18.		
	19.		
	20.		
(2)	21.	— $\cos \pi r_3$ (±x.xxx)	
	22.		
	23.		
	24.		
	25.		
	26.		
	27.		
	28.		
	29.		
	30.		
(3)	31.	— $\ln r_4$ (−x.xx)	
	32.		
	33.		
	34.		
	35.		
	36.		
	37.		
	38.		
	39.		
	40.		
	41.		
	42.		
	43.		
	44.		
	45.		
	46.		
(1)	47.	— Serial Number	
	48.		
	49.		
	50.		
	51.		
	52.		
	53.		
	54.		
	55.		
	56.		
	57.		
	58.		
	59.		
	60.		
	61.		
	62.		
	63.		
	64.		
	65.		
	66.		
	67.		
	68.		
(1)	69.	— r_1 (.xxxx) (to determine collision type)	
	70.		
	71.		
	72.		
(1)	73.	— r_2 (.xxx) (to determine energy loss)	
	74.		
	75.		
(1)	76.	— r_3 (.xxx) (to determine azimuth)	
	77.		
	78.		
(1)	79.	— r_4 (.xxx) (to determine distance traveled between collisions)	
	80.		

The random numbers r_1, r_2, r_3, and r_4 are independent of each other. The number of the operation in which a field is derived is given in parentheses to the left of the card columns. See "Collision Cards" Operations for step involved.

Collision Cards

The collision cards (card layout 3) are begun as "random number" cards. An estimate is made, on the basis of how many neutrons are involved and how many collisions are expected, of the number of collision cards which will be needed.

The following steps are then taken to obtain the information required concerning the case history of each neutron. These steps are numbered in accordance with the steps listed on the table of operations for the collision cards.

Step No.	*Function*
1	The estimated number of collision cards required is made up with card number (3) in column 1 and random digits, which will be the basis for the random choices to be made for each neutron, in columns 68-80. The first four of the random digits are called r_1, the next three r_2, the next three r_3, and the last three r_4. The source of the random numbers, which may be either taken from tables or calculated, should be recorded to avoid choosing them in the same way in a future problem of the same nature. The cards should be numbered serially to preserve the order.
2	Since $\cos \pi r_3$ and $\ln r_4$ are dependent only on random numbers, they may be put in all of the basic collision, or "random number," cards before calculation is started. The cards are sorted to order of r_3, match-merged with the master deck, and $\cos \pi r_3$ intersperse gang punched into the random number cards.
3	Similarly, $\ln r_4$ is intersperse gang punched from the master deck into the random number cards. The random number, or basic collision, cards then look like card layout 3 and are ready to be developed into com-

plete collision cards. They are next sorted back to serial number order to restore the randomness of the random numbers.

4 The first 1,000 of the random number cards are numbered consecutively from 000 to 999 in columns 2-4; 01 emitted into columns 5-6; 0's emitted into field 53-57; and $\mu = 1 - 2r_0$, where r_0 is chosen from the other random numbers on the card, is calculated and punched in columns 24-26. Each of these cards will then represent one of the thousand neutrons as it undergoes its first collision, having an initial energy loss u of zero and having a random direction cosine μ.

5 The thousand cards are sorted to order of μ, match-merged with the master deck, and $\sqrt{1-\mu^2}$ intersperse gang punched into the collision cards.

6 The cards are then sorted into ascending order of the value of energy loss, u, in columns 53-57 (which will, of course, be zero for all cards for the first collision, but not thereafter) and match-merged with the probability cards on columns 53-57, so that for each energy range there is a probability card followed by all the collision cards in the particular range. Collision cards having an energy level less than the specified limit are given a collision code of 3 and set aside, since they will undergo no more collisions. λ is intersperse gang punched from the probability cards into the remaining collision cards.

7 The merged deck of probability cards and collision cards is run through the collator, reading from each probability card the probability $A_0(u)$ of absorption and selecting from the following collision cards all cards whose $r_1 < A_0(u)$. These cards are coded 0 in column 67 and set aside, since they will undergo no more collisions. The remainder of the collision cards, which are still merged with the probability cards, are recollated, this time comparing r_1 with $A_1(u)$. Collision cards whose $r_1 \leqq A_1(u)$ are coded 1 in column 67 to indicate a type 1 collision, while cards whose $r_1 > A_1(u)$ are coded 2 to indicate a type 2 collision.

8 Collision cards having a code of 1 or 2 are sorted to the order of r_2 by collision code (67, 72-74) and match-merged on those columns with the master deck. K^2, K, ω, $\sqrt{1-\omega^2}$ and $\triangle u$ (all of which depend directly on the collision type and r_2) are intersperse gang punched from the master cards into the collision cards. The master deck is then put together again while the collision cards are being sorted to the order of neutron number (for convenience in case a particular neutron has to be checked in the future).

9 A random number card (card layout 3) is merged behind each collision card to become the second collision card for that particular neutron.

10 The merged deck is put in the 604 and μ_i, $\sqrt{1-\mu_i^2}$, ω_i, $\sqrt{1-\omega_i^2}$, $\cos \pi r_{3i}$ are read, where i is the collision number; $\mu_{i+1} = \mu_i\,\omega_i - \sqrt{1-\mu_i^2}\,\sqrt{1-\omega_i^2}\cos \pi\, r_{3i}$ is calculated and punched in the random number card following the collision card. The collision number is also read from the collision card, increased by one, and punched in the following card. The neutron number is carried forward directly from the collision card to the following random number card, whereupon the random number card becomes the new collision card for the neutron. u_i and $\triangle u$ are read from the ith collision card, added together and punched as u_{i+1} in the ith card and as u_i in the $i+1$th card; that is, the total energy loss after the ith collision is the same as the total energy loss before the $i+1$th collision.

11 The cards are then sorted on the collision number, the first collision deck set aside temporarily, and the deck of second collision cards taken through steps 5 to 11, thus generating a deck of third collision cards, and so on until all the neutrons have disappeared.

There is, then, a collision card for each collision each neutron has undergone. This card contains the total energy loss of the neutron after the collision and its direction cosine. To complete the data only the total distance traveled after each collision is needed.

12 To get this value all the collision cards are merged or sorted into order of collision number within neutron number (columns 2-4, 5-6). The first collision card for neutron 000 is read, and $\rho = \lambda \ln r_4$ and $\triangle z = \rho\mu$ are calculated. ρ and $\triangle z$ are punched in the first collision card; $\triangle z$ is punched in the z field in the first collision card and stored as well.

13 The second collision card for the same neutron is read, ρ and $\triangle z$ calculated and punched, $\triangle z$ added to the z stored from the previous collision and the new z value punched in the second collision card; that is, for any given neutron $z_{i+1} = \triangle z_{i+1} + z_i$, where i is the collision number. When the z values are computed, calculation of the data pertaining to individual collisions of each neutron is completed (card layout 4). These data can then be grouped and combined for whatever statistical studies are desired.

Complete case histories for each neutron are readily available and may be used in the solution of problems of various boundary conditions involving the constituents of the specific problem. Final compilation of the data takes the form of frequency distributions and factorial moments of these distributions. Two examples of such distributions are:

1. Distribution of the total distance traveled from the source before absorption or reaching a given energy level.
2. Distribution of the energy loss at each collision.

The factorial moments for these distributions are computed for curve fitting and calculation of various parameters of these distributions. Calculation of these moments

involves the use of standard accounting machines and presents no problems.

CARD LAYOUT 4

Title: COLLISION CARD

Card No. 3 Card Color: Plain Manila

Source: Card 3

(Random Number Card, see Card Layout 3)

(1)	1.	Card No. 3	41.		
	2.		42.	(12)	z (±xx.xx)
(4) (10)	3.	Particle No.	43.		
	4.		44.		
(4) (10)	5.	Collision No.	45.		
	6.		46.		
	7.		47.	(1)	Serial No.
(8)	8.	K^2 (+.xxxx)	48.		
	9.		49.		
	10.		50.		
	11.		51.	(6)	λ (+.xxx)
(8)	12.	K (+.xxx)	52.		
	13.		53.		
	14.		54.		
(8)	15.	ω (±.xxx)	55.	(4) (10)	u_i (+xx.xxx)
	16.		56.		
	17.		57.		
(8)	18.	$\sqrt{1-\omega^2}$ (+.xxx)	58.		
	19.		59.		
	20.		60.	(10)	u_{i+1} (+xx.xxx)
(2)	21.	cos πr_3 (±x.xxx)	61.		
	22.		62.		
	23.		63.		
	24.		64.	(8)	Δu (+x.xxx)
(4) (10)	25.	μ (±.xxx)	65.		
	26.		66.		
	27.		67.	(7)	Collision Code
(5)	28.	$\sqrt{1-\mu^2}$ (±.xxx)	68.		
	29.		69.		
	30.		70.	(1)	r_1 (.xxxx)
(3)	31.	ln r_4 (−x.xx)	71.		
	32.		72.		
	33.		73.	(1)	r_2 (.xxx)
	34.		74.		
	35.		75.		
(12)	36.	ρ (x.xx)	76.	(1)	r_3 (.xxx)
	37.		77.		
	38.		78.		
(12)	39.	Δz (±x.xx)	79.	(1)	r_4 (.xxx)
	40.		80.		

Type of Cards: MASTER CARDS

Operation	*Formulation*	*Machine Operations Involved*
1. Make Basic Master Cards		On 604 generate the consecutive numbers 000-999, punching the number for each card in columns 24-26, 72-74, 75-77 and 78-80. Emit a "2" into column 1 and a collision code of "1" into column 67. Make a similar deck with a "2" in column 67.
2. Calculate and Punch K^2	$K^2 = 1 - 4r_2M(M+1)^{-2}$	Calculate on 604, controlling on collision code to emit proper value of M.
3. Punch K in Cards		Sort the cards to the order of K^2, match-merge with a square root deck, and intersperse gang punch K into the master cards.
4. Calculate and Punch ω	$\omega = [(M+1)K-(M-1)/K]/2$	Calculate on 604, controlling on collision code to emit proper value of M.
5. Punch $\sqrt{1-\omega^2}$ in Cards		Sort the cards to the order of ω, match-merge with a sin-cos deck (matching ω with the sine) and intersperse gang punch $\sqrt{1-\omega^2}$ (cos) into the master cards.

The number of the operation in which a field is derived is given in parentheses to the left of the card columns. See "Collision Cards" Operations for steps involved.

DISCUSSION

Mr. Turner: I am not familiar enough with the details of this calculation to know whether it is possible to go back and reconstruct, shall we say, the x coordinates. Can that be done?

Mr. Bailey: Not in this particular problem. I am not very familiar with the work which is being done now. It may be that one of those actually doing the work could answer that question.

Miss Johnson: We can go back on this problem and reconstruct the x coordinates but it wouldn't be practical. We are working with the problem now where all the different coordinates, and time, are being calculated as we go, and we are using six constituents or six different types of atoms instead of only two.

Chairman Hurd: In this problem there was no interest in the x coordinates, Mr. Turner.

Mr. Turner: I was thinking of its use to get the spectrum, that is, the scattering, for the lower energy particles. If you could get the x coordinates, you could rotate your space and from the same data get the distribution for lower energy particles.

Mrs. Dismuke: That is the idea, of course, in the problem we are doing now. We started at higher energies. In the particular problem, which Mr. Bailey described, we started at such a low energy that our population probably wouldn't be big enough.

Operation	*Formulation*	*Machine Operations Involved*
6. Punch $\sqrt{1-\mu^2}$ in Cards		Same operation as above, except match μ instead of ω with sine.
7. Calculate and punch πr_3 so that we can pull cos πr_3 from our 90° deck.		Calculate πr_3 directly if $r_3 \leqq .500$; calculate $180° - \pi r_3$ if $r_3 > .500$ and punch X in column 23 to indicate a negative value.
8. Punch cos πr_3 in Cards.		Sort to the order of πr_3, match-merge with a 90° sin-cos deck and intersperse gang punch cos πr_3 into the master cards.
9. Punch $ln\ r_4$ in Cards.		Sort to the order of r_4, match-merge with ln deck and intersperse gang punch $ln\ r_4$ into master cards.
10. Punch $\triangle u$ in Cards.	$\triangle u = -ln\ K^2$	Sort to the order of K^2, match-merge with ln deck and intersperse gang punch $ln\ K^2$ into master cards, omitting sign since ln is negative and we want $-ln\ K^2$.
Type of Cards: COLLISION CARDS		
1. Make Basic Collision, or "Random Number," Cards.		Determine approximately how many collision cards will be needed and reproduce, or calculate, random numbers into columns 68-80 of that many blank cards. At same time, emit a "3" into column 1 of these cards and number the cards serially in cols. 45-49. Call cols. 68-71 r_1, 72-74 r_2, 75-77 r_3, and 78-80 r_4.
2. Punch in each "Random Number" Card the Cosine of a Random Angle of Deflection.	cos πr_3	Sort "random number" cards to the order of r_3 in cols. 75-77, match-merge with our master deck on cols. 75-77, and intersperse gang punch cos πr_3 into the random number cards. (Use only half of the master deck: the collision code 1 or the collision code 2 cards, since cos πr_3 is independent of collision type.)
3. Punch on each "Random Number" Card a ln picked at random.	$ln\ r_4$	Sort "random number" cards to the order of r_4 in cols. 78-80, match-merge with half of the master deck (say, the collision code 1 half) on cols. 78-80 and intersperse gang punch $ln\ r_4$ into the random number cards. Sort the random number cards to the order of cols. 45-49 to restore randomness.
4. Pick 1000 neutrons to undergo their first collision with no previous energy loss, and start them in a random direction.	Collision no. = 01 u = 00.000 $\mu = 1 - 2r_0$ for first collision only	On the 604 generate the consecutive numbers from 000-999, punching the number for each card in columns 2-4. Emit "01" into cols. 5-6 Emit "00.000" into cols. 53-57 Choose r_0 from r_1, r_2, r_3, and r_4: say the third digit of each number, and calculate $\mu = 1 - 2r_0$. (μ will be chosen in this fashion only on the first collision cards.)
5. Punch $\sqrt{1-\mu^2}$ in the Collision Cards.		Sort the collision cards to the order μ, match-merge them with the master deck, and intersperse gang punch $\sqrt{1-\mu^2}$ into the collision cards.
6. Punch λ, or the mean free path, in the cards.	λ is the probability of scattering in a given range and is dependent on the energy range into which the neutron falls.	Sort the collision cards to the order of u (which will be 00.000 on the collision cards for the first collision but not thereafter), merge them with the probability cards on columns 53-57 so that, in the merged deck, there will be a probability card for a certain energy range, and behind it will be all the collision cards with energy u_i within that range, then the next probability card and the collision cards within its range, etc. Check the sequence of the merged deck and then intersperse gang punch λ from the probability cards to the collision cards. Do not sort the probability cards and collision cards apart.

Operation	*Formulation*	*Machine Operations Involved*
7. Determine whether each neutron (a) fails to undergo another collision because of low energy; (b) is absorbed; (c) undergoes a collision of type 1 or (d) undergoes a collision of type 2. If not (a) then (b), (c), or (d) is a random choice.	(a) If energy loss, u, of the neutron exceeds a certain value the neutron undergoes no more collisions; (b) if $r_1 < A_0$ the neutron is absorbed; (c) if $A_1 \geqq r_1 \geqq A_0$ the neutron undergoes a type 1 collision; (d) if $r_1 > A_1$ the neutron undergoes a type 2 collision.	Cards of group (a) are removed from the deck by hand and "3" punched in them as the collision code. Operations (b), (c), and (d) are all done on the collator. Cards falling in group (b) are given a collision type code of "0"; cards falling in group (c) are coded "1"; and cards falling in group (d) are coded "2." Collision cards having a code of "3" or "0" are removed from the deck since the neutrons these cards represent will undergo no more collisions, while cards having a code of "1" or "2" are carried on through the collision.
8. Pick a random energy loss for each neutron and associated with this, a value of ω and $\sqrt{1-\omega^2}$	Use r_2 and collision code for random choice of $K^2 = 1-4r_2M(M+1)^{-2}$, $\omega = [(M+1)K-(M-1)/K]/2$, $\triangle u = -\ln K_2$	Sort collision cards to order of r_2 by collision code (cols. 67, 72-74), match-merge with the master deck on columns 67, 72-74, and intersperse gang punch K^2, K, ω, $\sqrt{1-\omega^2}$ and $\triangle u$ from the master cards to the collision cards. Sort the collision cards to order of neutron number (for future convenience) while putting the master deck back together.
9. Pick a set of random numbers to be used in calculating the next collision cards for each neutron.		Merge a "random number" card (see card layout 3) behind each collision card, the random numbers on the card to be used in calculating data for the next collision.
10. a. Calculate μ_{i+1} for the next collision.	$\mu_{i+1} = \mu_i\omega_i + \sqrt{1-\mu_i^2}\sqrt{1-\omega_i^2}\cos\pi r_3$	Read μ, ω, $\sqrt{1-\omega^2}$, $\sqrt{1-\mu^2}$, $\cos\pi r_3$ from the old collision cards, calculate μ_{i+1} and punch it in the new collision cards.
b. Punch the neutron number and new collision number in the new card.	Coll. No. $i+1$ = Coll. No. $i+1$	Intersperse gang punch the neutron number from the old collision card to the new collision card. Read the collision number from the old card, increase it by one, and punch it in the new collision card.
c. Calculate u_{i+1} and punch it in both the old and the new collision cards for each neutron.	$u_{i+1} = u_i + \triangle u$	Read u_i and $\triangle u$ from the old collision card, calculate u_{i+1} and punch it in columns 58-62 of the old collision card and in columns 53-57 of the new collision card.
11. Sort collision cards on the collision number and repeat steps 5-11 until all neutrons have disappeared.		
12. Pick a random distance ρ for each neutron to travel and calculate the distance $\triangle z$ traveled along the z-axis.	$\rho = \lambda \ln r_4$, $\triangle z = \rho\mu$, $z_{i+1} = \triangle z + z_i$	Sort or merge cards to order of collision number by neutron number (cols. 2-4, 5-6). Read first collision card for first neutron; calculate and punch ρ, $\triangle z$, and z, storing z. Read second collision card for first neutron; calculate and punch ρ, $\triangle z$, and z, etc.
13. Repeat operation 12 for all collisions of all neutrons starting each neutron at $z = 0$.		

A Monte Carlo Method of Solving Laplace's Equation

EVERETT C. YOWELL
National Bureau of Standards

DURING the first meeting of this seminar we discussed the solution of Laplace's equation in two dimensions by smoothing techniques. As was pointed out at that time, the smoothing process is only one approach to the problem. It was an approach that was being tested at the Institute for Numerical Analysis because it used a simple, iterative routine that a calculating machine could easily be instructed to follow.

A second experimental calculation that we performed, in seeking a simple method of solving Laplace's equation, was a test of a method suggested by Dr. Feller. The basis for this method is completely different from the other methods we mentioned. Dr. Feller was seeking a probability approach to the problem. That is, some sort of a random process is set up such that the probability distribution of the answer given by the process obeys the same differential equation as that of the physical problem. It is to be hoped that the random process offers a simpler computing scheme than any of the direct approaches to the solution of the differential equation, thus making the computational task simpler and less time-consuming.

In the case of Laplace's equation in two dimensions, this random process is merely a two-dimensional random walk. In such a walk, uniform steps are made along one or the other of the coordinate directions, the choice being purely random, and in either a positive or a negative direction, the choice once again being random. If such a walk is started at a point inside the boundary of a region, it will eventually end up at the boundary. The number of steps will vary for two different walks starting at the same point, but the average number of steps can be computed. Suppose, now, a large number of walks from a single interior point is made. Each time a boundary is reached, the process is stopped, the value of the function on the boundary is recorded, and the process is repeated from the same starting point. If, after a large number of walks, an average of all the boundary values is noted, that average will approximate the value of the function at the starting point and will converge to that value as the number of walks increases to infinity.

This process was tested on the IBM Type 604 Electronic Calculating Punch. A region bounded by the four lines $x = y = 0$ and $x = y = 10$ was selected and a unit step was made in the random walk. The boundary values were $f(10,y) = f(x,10) = 0$, $f(0,y) = 100 - 10y$, $f(x,0) = (10 - x)^2$.

The random variables were introduced into the problem according to the evenness or oddness of the digits in a table of random digits prepared and furnished us by RAND Corporation.

The wiring of the 521 control panel was very simple. The machine was to make two choices, depending on the evenness or oddness of two random digits. Hence, two columns of the random digit cards were wired through two digit selectors, and the outputs of all the even digits of each selector were wired together. These two even outputs were then used to transfer two calculate selectors.

The initial value of the coordinates of the starting point was introduced by a special leader card. This carried a control punch in column 79, and the original coordinates in columns 1-4. These values were read directly to factor storage units 1 and 3 and also to general storage units 1 and 3, and these units were controlled to read only on an x punch in column 79.

The final value of the answer was punched on a trailer card, which carried a control punch in column 80. The sum of the boundary values, and a tally of the number of times the boundary was reached, were read out of general storage 2 and 4 and punched in the trailer under control of the y in column 80.

The wiring of the 604 control panel was more complicated. The analysis chart is given below.

Factor Storage				MQ	Counter	General Storage			
1	2	3	4			1	2	3	4
x		y							

Read Step	*Suppress*	*Operation*
1.		Counter RI +, GS 1 read out if calc selector 1 energized, GS 3 read out if not.
2.		Emit "1," RI 2nd, counter add if calc selector 2 energized, subtract if not.
3.		Counter RO. GS 1 RI if calc selector 1 energized, GS 3 RI if not.
4.		Emit "1," RI 3rd, counter subtract.
5.		Balance Test.
6.	N.	Counter read out and reset.
7.	N.	GS 2 read out.
8.	N.	Emit "1," counter add.
9.	N.	Counter read out and reset, GS 2 read in.
10.	N.	FS 1 read out, GS 1 read in.
11.	N.	FS 3 read out, GS 3 read in.
12.	P.	Emit "1," RI 3rd, counter add.
13.	P.	Emit "1," counter subtract, balance test.

Read Step	*Suppress*	*Operation*
14.	P.	Counter read out and reset.
15.	P.	Emit "1," counter add, RI 3rd.
16.	P.	Counter subtract, RO GS 3 if calc selector 1 energized, RO GS 1 if not.
17.	P.	Read out and reset counter, MQ RI.
18.	P.	Emit "1," RI 3rd, counter add only if calc selector 1 energized.
19.	P.	GS 3 RI, counter RO and RS if calc selector 1 energized, GS 3 RI, MQ RO if not.
20.	P.	Multiply, GS 3 RO.
21.	P.	½ adjust, RI 2nd.
22.	P.	RO GS 4, RI 3rd, counter add.
23.	P.	RO and RS counter, GS 4 RI, RO 3rd.
24.	P.	GS 2 RO, counter add.
25.	P.	Emit "1," counter add.
26.	P.	Counter read out and reset, GS 2 RI.
27.	P.	FS 1 read out, GS 1 read in.
28.	P.	FS 3 read out, GS 3 read in.
29.		Counter read out and reset.

The first three steps compute the coordinate of the next point in the random walk. Either the x or the y coordinate is adjusted, according to one of the two random digits. And the adjustment is either positive or negative, according to the other random digit. Steps four and five test to see if the upper bound in x or y has been reached. If a negative balance test occurs, then the point is still under the upper bound, and steps 6 through 11, which correct the tally and reset the coordinate units to their original values, must be suppressed. Steps 12 and 13, which test for the lower boundary are permitted to take place, and the new balance test supersedes the old one. Now a negative test is a sign that the boundary has been reached; so steps 14 through 28 —which compute the value of the function at the lower bound, correct the tally, and reset the x and y storage units to their original values—are suppressed on a plus balance test.

Three possible conditions can arise from the balance tests. If the first test is positive, the upper boundary has been reached. Then steps 6 through 11 occur, and the remaining steps are suppressed. If the first test is negative, the upper boundary has not been reached, and a second test must be performed to see if the lower boundary has been reached. So steps 6 through 11 are suppressed, and steps 12 and 13 occur. If the balance test on step 13 is positive, the lower boundary has not been reached, and steps 14 through 28 are suppressed. If it tests negatively, the lower boundary has been reached, and steps 14 through 28 occur. In all cases, step 29 is taken to reset the counter at the end of each computation.

The operating procedure was as follows: From the computation of the mean length of the random walk, an estimate was made of the number of steps needed to give a specified number of walks. This many random digit cards were selected, a leader card was put in with the coordinates of the starting point, and a trailer card in which the answer was to be punched. This deck was then run through the 604, and the sum of the boundary values and the tally of the number of times the boundary was reached was punched in the final card. The quotient of these two numbers gave the value of the function at the starting point.

The tests indicate that the method will give the correct answers, but the speed of convergence is very slow. The smoothing method converges as $(1/n)$, where n is the number of times the field is smoothed. In this case, we mean that the difference between the approximate solution and the true solution is proportional to $(1/n)$. But in the Monte Carlo Method, the convergence is proportional to $(1/\sqrt{n})$. And the convergence here is a statistical convergence; that is, the probable error is proportional to $(1/\sqrt{n})$, where n is the number of random walks. With statistical convergence, one can never guarantee that the solution is correct, but one can state that the probability is high that the solution is correct.

It is obvious that the control panels described will not be applicable if the boundary values of the function are at all complicated functions of the coordinates. The method can still be used if the following changes are made: A trailer card is inserted behind each random digit card, and the coordinates of the point are punched in the trailer card. Or, similarly, a few columns of the random digit deck are reproduced into work cards, and these are run through the 604, punching the coordinates of the point in each card. The boundary values can be internally tested as long as the boundary is a rectangle, for two sides can always be made zero by a suitable translation of axes, and the other two constants can be stored in the 604. Hence, the x and y storage units can be reset to their initial value whenever the boundary is reached. If the boundary is not rectangular, then the mean length of the random walk and the dispersion of the mean length can be computed and enough random digit cards assigned to each random walk so that any preassigned percentage of fixed length walks will cross the boundary. A sufficient number of walks are made so that a reasonable number of boundary crossings are available.

The cards from the 604 in either of these methods will contain the coordinates of the points of the walk. In the case of rectangular boundaries, the boundary point can be found by a sorting operation. In the case of the non-rectangular boundary, the sorting operation must be followed by a selection of the first point of each walk that crosses the boundary. In any case, the functional value at each boundary point can then be gang punched in the proper boundary cards and an average made.

The great drawback of this statistical method is its slow speed of convergence. This should not cause the method to be discarded, for the ideas of Monte Carlo procedures are still new and ways may still be found to speed up the convergence of the solution. It is also true that the speed of

convergence of the Monte Carlo Method is not affected by the dimensionality of the problem, and this may prove to be a very great advantage in problems involving three or more dimensions. Finally, Monte Carlo procedures may be very important in problems where a single number is required as an answer (such as the solution at a single point, or a definite integral involving the solution of the differential equation). In these cases the entire solution would have to be found by smoothing techniques. With these limitations and advantages recognized, the simplicity of the procedure certainly makes the Monte Carlo Method worth considering for the solution of a partial differential equation.

DISCUSSION

Mr. Turner: When punching the position of each successive point in a random walk from a given starting point (in order to find the value of the function at the given point) can one consider the particular boundary value as one term of the partial sums for each of the points that was crossed in the random walk? That is, would it be a safe process with a large number of walks, to consider each one of the points, which was crossed as a starting point, for one of the other points in the problem?

Dr. Yowell: This question is now being investigated by Dr. Wasow and Dr. Acton at the Institute for Numerical Analysis. Their preliminary findings, about which I hesitate to say very much, indicate that this can be done, but it raises the standard deviation by a great deal, although it gives the same mean.

Mr. Turner: Another suggestion is that instead of starting from the interior, start from the boundary; then the problem of when the boundary is reached does not arise. In other words, use the symmetry of the functions. Start from the boundary and walk into the interior.

Professor Kac: That defeats the purpose. The result is the harmonic function at the starting point. So if one starts at the boundary, then one gets what he already knows, because the boundary value is given.

Mr. Turner: What I meant was: start at the boundary and then generate a sequence of points that are associated randomly with that particular boundary point (a sequence of interior points). In the final operation sort the cards on the point coordinates, run them through the accounting machine, and sum the values that are on the boundary cards; in other words, gang punch each boundary value into each one of these cards that was generated by the random walk.

Dr. Yowell: The question that occurs to me is, if one starts at any two boundary points, is the relative frequency of reaching a particular interior point the same as the relative probability of reaching the two boundary points by random walks originating at the particular interior point?

Professor Tukey: It seems to me that Mr. Turner has a very interesting suggestion, and that is to start at the boundary and walk n steps inside, until the boundary is reached. One then has two n pairs of numbers, each pair consisting of a boundary point and an interior point, and I understand the suggestion is to take a lot of walks of this type and cross-tabulate the number of times a pair (consisting of a certain boundary point and a certain interior point) occurs. It is an interesting speculation.

Professor Kunz: I would like to remark that we have taken a lot of liberties here. These probabilities around the edges (the Green's functions, which we are calculating here) must be evaluated for every interior point, n^2 in number; they must be multiplied in each case by every boundary value, $4n$, and that is too much work already.

Mr. Bell: Consider doing the problem on a type 604. The standard 604, of course, won't program repeat, but it has been pointed out that internal wiring can be altered to cause the machine to repeat the program cycles indefinitely.

The 604 has a pulse rate of 50,000 pulses a second. It takes about 25 pulses to do a particular operation, except multiplication and division, and they take a few more. Suppose we assume 100 cycles; which means 2,500 pulses, so that 20 of these testing operations a second can be performed. We can thus make about 1,200 such tests in a minute.

Thus, my question is: Would it be reasonable to take a 604, and, with quite minor modifications, load it up, let it run these random walks where all the computing is being done down to reaching a boundary condition, at which point a new card feeds and the computing continues?

Here is an opportunity to get the speed that is really available with electronic computing. I am certain that the circuits that would have to be changed are minor, a matter of a few hours' work, and a standard 604 could do this.

Mr. Sheldon: If this is to be done, it will be necessary to make the 604 compute its own random numbers since the random numbers must get into the machine in some way.

Mr. Bell: I wonder if that is a limitation. It is being done in the 602-A. It would depend, of course, on the functions; but a great deal can be done with the 604, and people who have had little or no experience with it are often amazed at how much can be done in spite of the limited storage available.

Dr. Yowell: We have actually tried generating random digits on the 604. We have 100,000 of them generated on a 5 by 8 representative multiplication, much of it the type that has been used on other machines, of squaring a ten-digit number and extracting the central ten digits. These have been subjected to the usual tests, and all we can say so far is that the digits are extremely lumpy and the question of how homogenized the digits have to be, in order to give good results, is something that should certainly be considered before generating random digits with a small multiplier or multiplicand.

Chairman Hurd: I think the idea sounds promising.

Further Remarks on Stochastic Methods in Quantum Mechanics

GILBERT W. KING

Arthur D. Little, Incorporated

I HAVE REMARKED that people working with computing machines frequently rediscover old mathematics. My philosophy is that with a computing machine one can forget classical mathematics. It is true that if one knows a little analysis, it can be worked in, but it is not necessary. However, Professor Kac[1] indicates that it looks as though I am going to have to learn some new mathematics after all, about Wiener functionals, to really understand what I am trying to do with the Monte-Carlo Method.

Mr. Bell has introduced the idea of a "DUZ" board, which is a control panel by which all sorts of calculations can be done without having to worry about wiring for each problem. The Monte-Carlo Method is really a "DUZ" method; and can do many kinds of problems without one having to think up new schemes of attacking each one.

In particular, I like to feel that one contribution the Monte-Carlo Method has made (and the Monte-Carlo Method is probably only the *first* new method we are going to have for computing machines), is the by-passing of specialized mathematical formulation, such as differential equations. From the viewpoint of a physicist or a chemist there really is no differential equation connected with the problem. That is just an abstraction that allows the problem to be solved by means of the last few hundred years of mathematics. With a computing machine one cannot use analysis directly. Here is an opportunity (for the physicist and the chemist) of putting the real problem right on the machine without having to go through differential equations.

Although Professor Kac[1] and Dr. Yowell[2] have illustrated this much more thoroughly, I would like to discuss again the very simplest kind of problem, namely, diffusion. We can discuss diffusion in one dimension, although I will indicate there is no fundamental difference in approach for many dimensions (in contrast with some analytical methods).

Consider a source of material such as a dye in a capillary in which the dye can diffuse either way. We wish to find out the distribution of these molecules after time, t. We are accustomed to say that the concentration varies with time according to its gradient in concentration along the x axis, with a proportionality constant D, the diffusion coefficient.

$$\frac{\partial u}{\partial t} = D\frac{\partial^2 y}{\partial x^2}$$

This differential equation is a simple one, and can be integrated. However, the physical problem need not be solved this way. The physical process may be considered directly. A molecule at the origin is subject to Brownian motion. (This is what makes diffusion work, and is implied in the differential equation.) That is, in time, $\triangle t$, it is going to move $\triangle x$ either in one direction or the other. In another interval of time, $\triangle t$, it is again going to move either forward or backward. The molecule describes a random path of $\triangle x$ in each interval of time, $\triangle t$, and after n steps it will arrive at some point x.

Another particle also follows a random path and arrives at a different place. If this process is carried out for a thousand particles, the distribution of particles at each value of x is a step function which, in the limit, will be exactly the solution of the differential equation. In the one-dimensional case the particles distribute themselves according to a normal distribution, which spreads out and becomes flatter with time.

These calculations can be carried out very easily on a computing machine, because when a particle moves $\triangle x$ and $\triangle t$, it is only necessary to add $\triangle x$ to its x coordinate, with a sign determined by a random number which can be fed into the machine. So, fundamentally, this is a very straightforward method. It might be crude in complicated problems, but one does not have to think very hard to set up the problem.

More difficult problems, which also have a physical basis, can be put directly on the machines, by-passing the differential equations. For instance, in quantum mechanics we are interested in the behavior of chemical systems. By a system is meant here an assembly of particles whose x, y, z coordinates completely describe the system. Thus, if the stationary distribution of these coordinates is known the stationary states of the system are known, and if that variation with time is known, chemical reactions can be

described. If the x, y, z coordinates of each particle are used to map out a multi-dimensional space the system can be described uniquely by a point in this space. A change of the system with time corresponds to a motion of this point in the dimensional space. From the point of view of the Monte-Carlo Method, the point is a particle. It is an abstract particle and does not refer to the electrons or nuclei in the system. In quantum mechanics these abstract particles do not have to diffuse in the classical sense; they can jump around in the configuration space. These jumps are called transitions. In this way a very characteristic feature of quantum mechanics can be introduced into the formulation of the problem. However, to simplify presentation and tie the Monte-Carlo Method to what has already been discussed at this seminar, we shall take the point of view that these transitions are over a small range and are governed by a partial differential equation. This is the so-called time-dependent Hamiltonian equation:

$$i\hbar \frac{\partial \psi(x,t)}{\partial t} = H\psi(x,t)$$

with solutions

$$\psi(x)\, e^{-i\lambda t/\hbar} .$$

This can be looked upon as a diffusion equation in a space with complex coordinates. It has periodic solutions. However, by making a transformation to a new time $(it/\hbar)$ it reduces to an ordinary differential equation with real variables of the type that Professor Kac[1] described:

$$\frac{\partial u}{\partial t} = \nabla^2 u - V(x)u$$

with solutions

$$u(x,t) = \psi(x)\, e^{-\lambda t} .$$

The first two terms correspond to the ordinary diffusion equation which I have discussed. The third term is additional. The diffusion process corresponds to the same kind of random walk, only now when a particle reaches a point, x, it is multiplied by a factor $e^{-V(x)\,dt}$. Thus, if the potential is positive, the particle has less weight when it arrives at the region. When $V(x)$ is negative it has more weight. Thus, the particles, beside diffusing, actually increase or decrease in weight. The diffusion phenomenon governed by this type of differential equation, that is by Schrödinger's equation, corresponds to the particles diffusing from the origin and distributing themselves, at first, in a curve similar to the normal one obtained for ordinary diffusion. The effect of the potential energy is such that as the particles diffuse far out they decrease in weight very quickly, whereas the particles diffusing near the origin do not decrease in weight very fast; so that the quantum mechanical distribution function gets cut off at the extremes and dies out uniformly throughout the whole region. The distribution is, in fact, the actual wave function in terms of the coordinates, dying out with a logarithmic decay equal to the eigenvalue. The exponential decay of the wave function causes some difficulty in accuracy on computing machines. To avoid this difficulty a modified potential function can be used,

$$V'(x) = V(x) - \lambda^0$$

where λ^0 is an approximation to the lowest eigenvalue. With this modified potential function in the exponential the total number of particles does not decrease. Thus, the particles diffuse out to form an approximation to the wave function whose area remains constant and which rapidly settles down to the proper value.

Employing the Monte-Carlo Method in this way, we have been able to set up a simple problem, namely that of the harmonic oscillator on the 602-A calculating punch, which is not an elaborate machine. We have actually been able to insert only one card and have computing carry on for a whole day. When enough random walks have been made, a button is pressed, and the sum of the weight factors divided by the number of walks is punched. This is the eigenvalue. It is rather interesting to see that, using random numbers during the day's calculations, a very definite number has been obtained, namely, the eigenvalue of the harmonic oscillator, which is correct at least to a certain number of significant figures. If a real source of random numbers was available, the process could be repeated on another day in which an entirely different set of numbers would have passed through the machine, and the same eigenvalue, within statistical fluctuations, would be obtained. Thus, the computations are carried out entirely with random numbers. Even in one calculation, the numbers at the beginning and the end are entirely independent, statistically. There happens to be a sort of Markoff process so that the numbers over a period of two or three minutes are related to each other, although the numbers over a period of several hours are quite independent.

I want, now, to make a connection of the Monte-Carlo Method to the other methods of solving differential equations discussed at this seminar. To do this we can discuss merely the ordinary diffusion equations and leave out the potential terms which is a characteristic of quantum mechanics. The diffusion equation must be written as a difference equation:

$$u(n,m) = \frac{1}{2}\Big\{u(n-1,m-1) + u(n+1,m+1)\Big\}$$
$$x = n\triangle x$$
$$t = m\triangle t$$
$$2\triangle t = \triangle x^2$$

which states that the number of particles at x, after one step, is equal to one half the probability of the particles being either to the left or to the right. The set of difference

equations, one for each value of x, form a matrix equation, where the distribution along x is represented as a vector:

$$\vec{U}(m) = T\,\vec{u}(m-1)$$

$$T = \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} & . & . & . & . \\ . & \frac{1}{2} & 0 & \frac{1}{2} & . & . & . \\ . & . & \frac{1}{2} & 0 & \frac{1}{2} & . & . \end{pmatrix}.$$

The matrix T operates on the original distribution $u(0)$, which will be assumed to correspond to particles only at the origin and, therefore, has an element $= 1$ at $x = 0$, and 0 elsewhere. Multiplication of the vector by the matrix T corresponds to the diffusion in a short time, $\triangle t$, the operation of the square of this matrix on the original vector to taking walks of two steps. The nth power of the matrix corresponds to taking n steps. Each term in each element of the resulting vector corresponds to a single walk, and the grand total of all the terms in all the elements of the vector corresponds to all possible walks.

The method can be generalized by allowing the particles not only to move dx and dt, but to move ndx and ndt where the chances of moving $2dx$, $3dx$ and $4dx$, etc., are governed by some distribution law. It is clear that this distribution law is exactly that required to express the second derivative in terms of the function to the left and to the right:

$$\frac{\partial^2 u}{\partial x^2} = \Sigma\, a_k\, u(x+k\triangle x) \qquad k = -K \text{ to } +K\ .$$

Physically, this corresponds to particles having, instead of a mean free path, a distribution of free paths, and the idea is introduced that a particle can make a transition from one state to another and not merely diffuse over a single interval, dx. In this case, the matrix T has $2K$ diagonals, but although the expressions for the elements of T u are more complicated, it is still true that all possible random walks correspond to the nth power of this matrix, so that the situation is exactly the same as described in the elementary diffusion problem.

The characteristic vectors of this matrix can be found in the usual way of iterating the matrix. Thus, it is seen that the Monte-Carlo Method of solving a differential equation, when carried to the limit of all possible random walks, becomes the recommended method of finding the characteristic vectors of a matrix. It is interesting to see the Monte-Carlo Method as a "DUZ" method, in the sense that it works with a 2 by 2 or a 10,000 by 10,000 matrix. The advantage of the Monte-Carlo Method is that, instead of computing all terms in all elements of the nth power of a matrix, only a sample of, say, 1,000 need be taken to obtain results to two or three significant figures. It is, therefore, quite clear that if all possible random walks were taken, a distribution would be obtained which would be exactly defined by the iterative method of solving difference equations.

DISCUSSION

Dr. Brillouin: There is a problem in connection with all these applications of the Monte-Carlo Method that has been in my mind for some time, and I would like to ask a question. One of the difficulties in using the machine is that you have to repeat the computation a number of times, at least twice, to be sure that the machine doesn't make mistakes. How can you repeat twice, the same random walk? How do you make the checks on the Monte-Carlo Method on the machine?

Dr. King: As I pointed out, that is fundamentally impossible if you have a real source of random numbers. However, the Monte-Carlo Method could be carried out very conveniently if there were a hub on every IBM control panel that said "random number" on it, and supplying a random number. To check the results one would merely repeat the whole problem, using entirely different random numbers. The eigenvalues obtained should, of course, be the same as the first time through, within a statistical fluctuation depending on the number of random walks taken. However, this method does not allow for faulty wiring or machine failures. To make sure that no mistakes of this type have been made, we have adopted the procedure of recording the random numbers used with every step. We then repeat the whole procedure, using the random numbers in reverse order. In other words, we allow the particles to walk in the opposite direction from the first case. This usually means that the function of the types of random numbers be changed so that a fairly reliable check of machine methods has been made.

REFERENCES

1. MARK KAC, "The Monte Carlo Method and Its Applications," pp. 74-81.
2. EVERETT YOWELL, "A Monte Carlo Method of Solving Laplace's Equation," pp. 89-91.

Standard Methods of Analyzing Data

JOHN W. TUKEY

Princeton University

I SHOULD LIKE to be able to make you statisticians in one brief exposure, but it seems unwise to try. We are going to go over some methods that form sort of a central core of the statistical techniques that are used today, trying to do it in such a way that when someone comes to you—wanting this or that computed—you may have a better understanding of why these particular things were chosen.

By and large, we are not going to discuss the formulas that would actually be used in the computation (although we shall occasionally refer to those used in hand computation). I will leave that to Dr. Monroe[1] and Dr. Brandt.[2] I am going to discuss these methods in terms of how it is easiest to think about them.

My purpose, then, is to supply background: statistical, algebraic and perhaps intuitive.

Interpreting Data

We shall have more to do with models than you might expect—quantitive models for what might have happened. This will seem strange at first glance, for most of us usually keep this aspect of interpreting numbers in our subconscious. But the whole of modern statistics, philosophy and methods alike, is based on the principle of interpreting what did happen in terms of what might have happened. When you think the situation over, I think that you will agree.

There are few problems, indeed, where it is sufficient and satisfactory to say, "Well, here are the numbers, and this is a sort of summary of them. Now, somebody ought to know enough to do something with this!" But many of my friends will try to do this, astronomers in war work or sociologists studying public opinion; they try to stop too soon. A reasonable quantitative model would take them much further.

In discussing new machines, it has been said that "statistics is counting." Many think so. But the sort of statistical procedures that I am going to discuss are basically procedures for analyzing measurements, not counts. They can also be used for counts which behave like measurements, and I believe Dr. Brandt will discuss this. There are other ways to bring counts into the picture, but we shall not go into them here.

Simple Models

When a physicist or engineer thinks of reduction of data, his first thought is of a set of points y, one or more for each of a set of values x, and a process of fitting a straight line. We are going to start two steps below this and work up slowly.

Let us suppose that we have a number of measurements of what is supposed to be the same quantity, perhaps the velocity of light, perhaps the conversion efficiency of a catalytic cracking plant, perhaps the response of guayule to a new fertilizer. These measurements will not all be the same (if they are, we are not carrying enough significant figures!). We must take account of their variability!

The simplest way to do this is to suppose that

1. they are composed of an invariable part and one that is fluctuating,
2. these parts are united by a simple plus sign, and
3. the fluctuating part behaves like the mathematician's ideal coin, or ideal set of dice, with the contribution to each measurement wholly unrelated to that in the last, or the next, or any others (nearby or not).

In terms of this situation, where these parts are things we shall never know, we want to get as good a hold on the underlying facts as we can—from three values, from thirty, or from three hundred. These three suppositions suppose a lot, and much experimental technique and much experimenter's knowledge of the subject matter go into making them good approximations by doing the experiment appropriately. We shall not go into these problems of design, as opposed to analysis, of experiment.

Of course, not all experiments can be conducted to fit such a simple model, and we shall also meet more complex ones. These will usually involve more parts, and when these parts are connected by plus signs, we shall usually use methods which grow naturally out of those used for the single sample.

Notation

We shall deal mainly with averages and variances. If we have, or we think about, N numbers, $w_1, w_2, \ldots, w_N$ then

$$w_\blacktriangle = \text{average } (w) = \frac{w_1 + w_2 + \ldots + w_N}{N},$$

$$\text{variance } (w) = \frac{w_1^2 + w_2^2 + \ldots + w_N^2 - \frac{1}{N}(w_1 + w_2 + \ldots + w_N)^2}{N-1}.$$

The variance, as defined here, differs from the mean-square-error of the physicists by just this denominator of $N-1$ instead of N. There is still a deep schism on whether you divide by N or $N-1$. I am far on the left, and divide everything of this kind by $N-1$ under all possible circumstances. It makes the formulas simpler; I think that is a good reason.

When you have a set of numbers $x_1, x_2, \ldots, x_n$ which we think of as a sample from a larger collection, as for example in our model just described, we calculate the same things but use different words and notation

$$x_\bullet = \text{mean } (x) = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{x_+}{n},$$

$$MSDx = \text{mean square deviation } (x) = \frac{x_1^2 + x_2^2 + \ldots + x_n^2 - \frac{1}{n}(x_1 + x_2 + \ldots + x_n)^2}{n-1}.$$

We shall steadily use these three conventions:

A $\bullet$ in place of a subscript means the mean (= average over the sample).

A $\blacktriangle$ in place of a subscript means the average (taken over the population).

A $+$ in place of a subscript means the sum over the sample.

Thus, for example, if y_{ij} has been observed for each combination of $i=1,2,\ldots,c$ with $j=1,2,\ldots,r$, then

$$y_{\bullet 3} = \frac{1}{c}\left\{y_{13} + y_{23} + \ldots + y_{c3}\right\} = \frac{1}{c}\, y_{+3}$$

$$y_{2\bullet} = \frac{1}{r}\left\{y_{21} + y_{22} + \ldots + y_{2r}\right\} = \frac{1}{r}\, y_{2+}\ .$$

This choice of notation and terminology is compact and convenient, and we shall use it systematically, but you are warned that it is not universal, not general, or, in some aspects, not even widespread.

It will allow us to state models and indicate possible computations in a relatively compact and clear way.

A Single Sample

The Model

We are now prepared to take the model for a single sample that we have already discussed and express it algebraically. It is

$$\begin{cases} y_i = \eta + \epsilon_i, & i=1,2,\ldots n, \\ \eta \text{ fixed}, \\ \epsilon_i \text{ a sample from a population of size } N, \text{ average } \epsilon_\blacktriangle \\ \text{and variance } \sigma^2. \end{cases} \qquad (1)$$

that is, we think of the fluctuating parts as a random selection of n values from N values. These N values will have an average, which we have already agreed to call $\epsilon_\blacktriangle$ and a variance which we now agree to call σ^2. We may think of N as large as we like, and can easily think of $N = \infty$ as a limiting case. We may confine our analysis to finite N, and always take the infinite case in the limiting sense. Any practical meaningful new features which might appear in a mathematical model for the infinite case would be statistically extraneous, and we should have to seek for ways to eliminate them by changing the model (this we might do by changing the mathematician).

Illustrations

Let us imagine ourselves in control of an army camp containing 50,000 men. We may be interested in their average height. Suppose that η is the average height for the whole army of 10,000,000 men, which we do not know; then we can define ϵ as the difference between an individual soldier's height and η. There will be 50,000 such values of ϵ (if we admit, for the sake of simplicity, that each soldier has a height which can be reliably measured). We may select 200 men at random (which is an operation requiring care) from the personnel files, and have their heights measured. We shall have available

$$y_\bullet = \eta + \epsilon_\bullet = \frac{1}{200}\left\{y_1 + y_2 + \ldots + y_{200}\right\},$$

and we are interested in

$$y_\blacktriangle = \eta + \epsilon_\blacktriangle = \text{average of all 50,000 heights.}$$

Here $N = 50{,}000$, $n = 200$, and the Greek letters, as usual, refer to quantities which we do not know, even after the experiment. We are, however, interested in learning as much as we can about some of these Greek quantities, or about some combinations of them. In this case, we wish to infer about $\eta + \epsilon_\blacktriangle$.

As another case, let us take the measurement of the velocity of light in a vacuum. Here η would naturally be the "true" velocity of light in a vacuum, if such exists, while the ϵ's are defined by difference, as the "errors" of single determinations. We have no obvious limit to the size of the population of errors; so we take $N=\infty$. The average velocity measured by this observer, under these conditions, with this apparatus is

$$y_\blacktriangle = \eta + \epsilon_\blacktriangle\ .$$

This will not be the "true" velocity because of systematic errors in theory, in instrument design, in instrument manufacture, and because of the personal equation of the observer, to name only a few reasons. These systematic effects are reflected in $\epsilon_\blacktriangle$. The statistical analysis can tell us nothing (directly) about $\epsilon_\blacktriangle$, since we cannot measure any ϵ, or anything related to an ϵ which does not involve η. We can learn about an individual $\eta+\epsilon$, since this combination is an observable value, and hence we can learn statistically, about

$$\eta + \epsilon_\blacktriangle\ .$$

The allowance for $\epsilon_\blacktriangle$ is a matter for the physicist, although the statistician may help a little.

We should hasten to say that in the models we use here, there is much flexibility. Another person might define η to be the average value obtainable by this observer under these conditions, with this apparatus. By doing so, he would define $\epsilon_\blacktriangle=0$. Since this would not change the experiment, it is very well that we shall find that it would not change our formulas or our conclusions.

The Identities—Simplest Case

We can certainly write that

$$y_i \equiv y_\bullet + (y_i - y_\bullet) \ ,$$

and many college freshmen would like to infer from this that

$$y_i^2 = y_\bullet^2 + (y_i - y_\bullet)^2 \ ,$$

which you know to be wrong unless $y_i = y_\bullet$ (or $y_\bullet = 0$)! However, the sum of the deviations $(y_i - y_\bullet)$ for all i is zero, so that

$$\begin{aligned} \Sigma y_i^2 &\equiv \Sigma y_\bullet^2 + \Sigma (y_i - y_\bullet)^2 \\ &\equiv n y_\bullet^2 + \Sigma (y_i - y_\bullet)^2 \\ &\equiv n y_\bullet^2 + (n-1)s^2 \quad \text{(defines } s^2\text{)} \qquad (1') \\ &\equiv \frac{y_+^2}{n} + \left(\Sigma y_i^2 - \frac{y_+^2}{n} \right), \ \text{(working form)}. \end{aligned}$$

The last line indicates how the two terms are ordinarily calculated by hand. It is written in terms of sums instead of averages. The divisions are postponed to the last. In general, people who do calculate this expression on hand machines makes a regular practice of such postponements. It saves miscellaneous rounding errors from arising, and makes it clear that certain decimal places are not needed after a division. Notice that there is one standard process. We shall see again and again that we have summed a certain number of entries, squared that sum, and then divided the square by the number of entries. This y_+^2/n is a standard sort of thing that arises again and again.

All this has been done as if we really expected y to be nearly zero. What if we had expected it to be nearly Y? There is an entirely analogous set of identities, namely

$$\left\{ \begin{aligned} \Sigma (y_i - Y)^2 &\equiv \Sigma (y_\bullet - Y)^2 + \Sigma (y_i - y_\bullet)^2 \\ &\equiv n(y_\bullet - Y)^2 + \Sigma (y_i - y_\bullet)^2 \qquad (1'') \\ &\equiv n(y_\bullet - Y)^2 + (n-1)s^2 \quad \text{(the same } s^2\text{!)} \\ &\equiv \frac{(y_+ - nY)^2}{n} + \left(\Sigma y_i^2 - \frac{y_+^2}{n} \right). \end{aligned} \right.$$

Notice that the term $\Sigma (y_i - y_\bullet)^2$, which appears here, is exactly the same term which appeared in the previous identity. Thus, it equals $(n-1)s^2$ as before, and can be calculated in the same simple way as before.

When the results are placed in a table, the standard form is that of Table IA, where the mystic letters along the top of the table refer to "degrees of freedom," "sums of squares," and "mean squares." The entries in Table IA show how the actual entries—the numbers found by computation which would be entered in such a table—are related to the actual observations.

TABLE IA

TABULAR ARRANGEMENT FOR MODEL (1)

Item	*DF*	*SS*	*MS*
mean	1	$n(y_\bullet - Y)^2$	$n(y_\bullet - Y)^2$
residue	$n-1$	$(n-1)s^2$	s^2

Average Values under the Model

The model under which we are working, specified precisely in (1), states that the $\binom{N}{n}$ kinds of samples of size n are equally probable. The various quantities we have been discussing vary from sample to sample. But we can determine their average values—averaged over the $\binom{N}{n}$ kinds of samples of size n—in terms of η and $\epsilon_\blacktriangle$ and σ^2. Together with some variances, these are given in Table IB, page 98.

The essential things to notice are the average values in the two bottom rows of Table IB; the average values of the two mean squares in Table IA. The average value of the "mean square for residue" is just σ^2, the variance of the fluctuating contribution. The average value of the "mean square for the mean" consists of two terms, $\left(1 - \frac{n}{N}\right)\sigma^2$, which is nearly σ^2 when the population is large (and is otherwise smaller), and $n(y_\blacktriangle - Y)^2$ which is zero when the contemplated value Y is equal to the population average $y_\blacktriangle = \eta + \epsilon_\blacktriangle$.

The first essential point to be gained from these average values of mean squares (which from now on we shall simply call "average mean square" and abbreviate by "AMS") is this. These average values really say that:

1. All the systematic contribution has been siphoned off into the mean square for the mean.
2. As much of the fluctuating contribution as possible has been siphoned into the mean square for residue.
3. We know how much, on the average, of the effect of the fluctuating contribution remains in the mean square for the mean; we know this in terms of σ^2 and so can judge its size from the mean square for residue.

This is the qualitative picture—one you need to understand!

Interpretation of Anova Table

Table IA is called an analysis of variance table (which I often like to abbreviate to "anova table"). How do we interpret specific tables containing numbers? We know that we must face sampling fluctuations from sample to sample, but if we neglect them momentarily, we can learn much.

The average value of the residue mean square is σ^2, bare and unadorned; it tells us about σ^2. The average value of the mean square for the mean is complex; it has two terms. We can avoid this complexity by forming

$$\frac{1}{n}\left\{ (\text{mean square for the mean}) - \left(1 - \frac{n}{N}\right)(\text{mean square for the residue}) \right\}$$

whose average value is

$$\frac{1}{n}\left\{ n(Y - y_\blacktriangle)^2 + \left(1 + \frac{n}{N}\right)\sigma^2 - \left(1 - \frac{n}{N}\right)\sigma^2 \right\} = (Y - y_\blacktriangle)^2 \ .$$

On the average this component of mean square for the mean tells us about $(Y - y_\blacktriangle)^2$. Thus, an observed large value suggests that $Y - y_\blacktriangle$ is not zero, while an observed value which is small, zero, or negative indicates that $Y - y_\blacktriangle$ could be zero (for all this sample tells us) and that $Y - y_\blacktriangle$ is surely small.

By analogy, we define the CMS (component of mean square) for the residue by

$$\text{CMS residue} = \text{MS residue},$$

TABLE Ib

AVERAGE VALUES AND VARIANCES FOR MODEL (1)

Quantity	*Average Value*	*Variance*
η	η	0
ϵ_i	$\epsilon_\blacktriangle$	$\left(1-\frac{1}{N}\right)\sigma^2$
y	$\eta+\epsilon_\blacktriangle$	$\left(1-\frac{1}{N}\right)\sigma^2$
$y_\bullet$	$\eta+\epsilon_\blacktriangle$	$\left(\frac{1}{n}-\frac{1}{N}\right)\sigma^2$
$y_\bullet - Y$	$\eta+\epsilon_\blacktriangle - Y$	$\left(\frac{1}{n}-\frac{1}{N}\right)\sigma^2$
$(y_\bullet - Y)^2$	$(\eta+\epsilon_\blacktriangle - Y)^2+\left(\frac{1}{n}-\frac{1}{N}\right)\sigma^2$	...
$\Sigma(y_i - y_\bullet)^2$	$(n-1)\sigma^2$	...
$\Sigma(y_\bullet - Y)^2 = n(y_\bullet - Y)^2$	$n(\eta+\epsilon_\blacktriangle - Y)^2+\left(1-\frac{n}{N}\right)\sigma^2$	...
$s^2=\frac{1}{n-1}\Sigma(y_i - y_\bullet)^2$	σ^2	...

since its average value is just σ^2, and notice that we may write

$$\text{CMS mean} = \frac{1}{n}\,\text{MS mean} - \left(\frac{1}{n}-\frac{1}{N}\right)\text{MS residue.}$$

Now a very usual case is $N=\infty$. Let us suppose that this is so, and that we find

$$\text{MS mean} = 1000\ ,$$
$$\text{MS residue} = 10\ .$$

We must conclude, if there were at least three observations, that the CMS mean, which came out to be

$$\frac{900}{n}\ ,$$

is quite sure not to be zero on the average. Thus, $Y-y_\blacktriangle$ is not zero, and we do not believe that Y can equal $y_\blacktriangle$.

On the other hand, if

$$\text{MS mean} = 9, 10 \text{ or } 11$$
$$\text{MS residue} = 10\ ,$$

we have

$$\text{CMS mean} = -\frac{1}{n}, 0, \frac{1}{n}\ ,$$

and we must conclude that its average value might be zero. Thus, $Y-y_\blacktriangle$ might be zero, and Y might be the unknown $y_\blacktriangle$. Moreover, $Y-y_\blacktriangle$ is probably small.

Thus, in extreme cases a look at the mean squares may tell us whether Y is quite sure not to be equal to (or near to) $y_\blacktriangle$ or whether it may equal $y_\blacktriangle$ (and is quite surely near it). In intermediate cases we should have to think hard, or use a carefully worked out procedure (such as we mention shortly).

Notice that the average values of the mean squares and, at least by implication, the components of mean square play an essential role in drawing such conclusions.

Test Ratios and Confidence Intervals

Now you may wish to make a critical test of the null hypothesis that some Y you have selected may be equal to $y_\blacktriangle$. This is ordinarily done with one of the two ratios indicated in Table Ic.

TABLE Ic

TEST RATIOS FOR MODEL (1)

$$F=\frac{n(y_\bullet - Y)^2}{\frac{1}{n-1}\Sigma(y_i - y_\bullet)^2}=\frac{(y_\bullet - Y)^2}{\frac{1}{n}s^2}=\left\{\frac{y_\bullet - Y}{\sqrt{s^2/n}}\right\}^2=t^2$$

$$t=\frac{y_\bullet - Y}{\sqrt{s^2/n}}$$

If the critical value of $t=t_a$, then confidence limits for

$$y_\blacktriangle = \eta+\epsilon_\blacktriangle$$

are given by

$$y_\bullet \pm t_a\left(\frac{s}{\sqrt{n}}\right)$$

F is used frequently for what is called a variance ratio. In this case, as usual, it is just the ratio of the two entries in the mean square column. In this particular case (as always when the numerator has only one degree of freedom) it is exactly the square of the very simple ratio denoted by t. (This is Student's t.)

If a given value of Y gives rise to a value of t (or F) near zero, we must, so far as this one set of observations is concerned, admit that $y_\blacktriangle$ (which we don't know) might equal this given (or contemplated) value Y. If some Y is far enough from the observed mean $y_\bullet$, then t (and also $F=t^2$) will be large, and we shall be more or less sure that this Y differs from $y_\blacktriangle$. Those values of Y, which might reasonably be the unknown value of $y_\blacktriangle$ form a confidence interval for $y_\blacktriangle$. The last line of Table Ic shows how easy it is to compute such an interval. The interpretation of the interval is merely "$y_\blacktriangle$ is likely to lie in here," with the strength of the conclusion depending on t_α and getting stronger as the interval increases in length.

The Diagram

Figure 1 is a very simple diagram of the sort that we can use again and again in more complex cases; at least it helps me in understanding what is going on. The vertical increments are sums of squares, while the horizontal increments are corresponding degrees of freedom. Thus, the slopes are mean squares. We obtain one slanting segment, which we shall call a trace, for each line in the analysis of variance table.

The possible allowance for fluctuations to be made in the sum of squares for mean is indicated by the small shaded triangle, whose altitude is s^2. In the right-hand portion of the figure, this possible allowance is divided into the ratio of $N-n$ to n to produce the actual allowance $(1-n/N)s^2$. This portion of the figure shows how the failure of Y to equal the $y_\blacktriangle$ seems to contribute to the total sum of squares of deviations from Y.

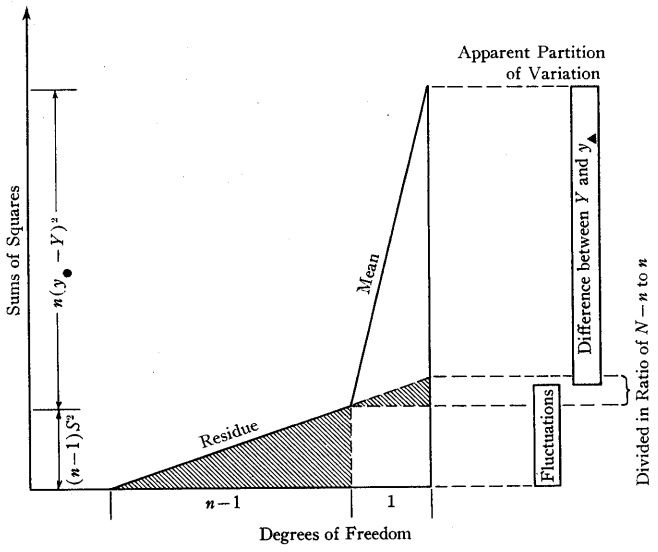


FIGURE 1. ANALYSIS OF VARIANCE DIAGRAM FOR MODEL (1), GENERAL

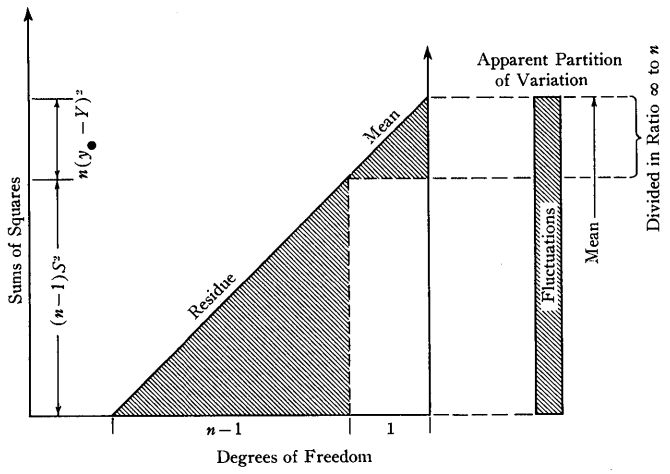


FIGURE 2. ANALYSIS OF VARIANCE DIAGRAM FOR MODEL (1), SPECIAL CASE ($n = \infty$ and Y near $y_\blacktriangle$)

If $N=\infty$, and the sum of squares for the mean came to exactly this height (as it does not in Figure 1, but does in Figure 2), the component of mean square would be zero, and we should have no reason to believe that $y_\blacktriangle$ should be different from Y. Thus, the fact that the mean trace is steeper than the residue trace indicates some evidence that Y is not equal to $y_\blacktriangle$.

To put it another way, when Y does happen to equal $y_\blacktriangle$ the slopes of the trace for the mean and the trace for the residue both average to σ^2. Thus, their failure to agree in slope—nicely measured when $N=\infty$ by the ratio, F, of the slopes—is an indication, not a proof, that Y does not happen to equal $y_\blacktriangle$.

The reason that we must use the ratio and not the difference is easy to see. The vertical scale of this diagram depends on the units in which we express our measurements. [If we change from feet to inches, $n(y_\bullet - Y)^2$ increases by a factor of 144!] So differences in slope would depend on the units used, which would clearly be wrong. The ratio does not depend on the units.

What sort of ratio of slopes you allow, before you feel that the Y contemplated cannot reasonably equal $y_\blacktriangle$, depends on n, among other things. If $n=1{,}000{,}000$, so that there are 999,999 degrees of freedom for the residue, then a ratio of 5 is suspect. But, when $n=2$ so that there is only 1 degree of freedom in the residue, a ratio of 150 may well escape suspicion. In any event, the diagram places the situation before you honestly and fully.

REGRESSION

We have gone further into the simplest case than we intend to go into the others. We shall take the next group of cases, which include fitting straight lines, curves and surfaces (in their simple cases), and multiple regression

analyses (sometimes miscalled correlation analyses) more rapidly. Here these are lumped together under the heading of regression, and can be treated together. We shall begin at the beginning, and build up to the general case.

Ultrasimple Regression

We begin with the simplest case of all, where the model is

$$\begin{cases} y_i = \beta x_i + \epsilon_i, & i = 1, 2, \ldots, n, \\ \beta \text{ fixed}, \\ \{x_i\} \text{ known without error}, \\ \{\epsilon_i\} \text{ a randomly arranged sample from a population of size } N, \text{ average zero, variance } \sigma^2. \end{cases} \quad (2)$$

The assumption "known without error" means, actually, that you have measured x so much better than y that you don't have to consider the errors in x. We have assumed $\epsilon_\blacktriangle = 0$ because we really believe the "true" line

$$y = \beta x$$

passes through the origin.

The identities are

$$\begin{cases} \Sigma y_i^2 \equiv \Sigma (bx_i)^2 + \Sigma (y_i - bx_i)^2 \\ \quad \equiv b^2 \Sigma x^2 + (n-1)s^2 \qquad \text{(defines } s^2\text{)} \quad (2') \\ \quad \equiv \dfrac{(\Sigma x_i y_i)^2}{\Sigma x_i^2} + \left\{ \Sigma y_i^2 - \dfrac{(\Sigma x_i y_i)^2}{\Sigma x_i^2} \right\}, \quad \text{(working form)} \end{cases}$$

where

$$b = \frac{\Sigma x_i y_i}{\Sigma x_i^2},$$

and

$$\begin{cases} \Sigma (y_i - Bx_i)^2 \equiv \Sigma (bx_i - Bx_i)^2 + \Sigma (y_i - bx_i)^2 \qquad (2'') \\ \quad \equiv (b-B)^2 \Sigma x_i^2 + (n-1)s^2 \quad \text{(same } s^2\text{)} \\ \quad \equiv \dfrac{(\Sigma x_i y_i - B \Sigma x_i^2)^2}{\Sigma x_i^2} + \left\{ \Sigma y_i^2 - \dfrac{(\Sigma x_i y_i)^2}{\Sigma x^2} \right\}, \\ \qquad \text{(working form)}. \end{cases}$$

Notice that the y_+^2/n term is back in its more general form

$$[\Sigma(\text{coefficient})\ y_i]^2 / \Sigma(\text{coefficient})^2 .$$

In treating the single sample all the coefficients were 1; here they differ from one i to another, being just x_i. It's the same principle in action.

The analysis of variance table, including a column for the average mean squares, is given in Table IIA. This table explains how DF, SS and MS and CMS are related to the observations, and how the average mean squares and components of mean square are related to the population.

Notice that n in $1 - (n/N)$ has been replaced by $x_+^2 \{\Sigma x_i^2\}^{-1}$, which reduces to n when the x_i are all equal.

An actual table would have numerical entries through the MS column or the CMS column and might or might not have some or all of the algebraic entries in the AMS and ACMS columns to guide the reader.

We can test the null hypothesis $\beta = B$, and set confidence limits for β just as we did with the mean. The formulas are set out in Table IIB.

TABLE IIB

TEST RATIOS AND CONFIDENCE LIMITS FOR MODEL (2)

$$F = \frac{(b-B)^2 \Sigma x_i^2}{s^2} = \left\{ \frac{b-B}{\sqrt{s^2/\Sigma x_i^2}} \right\}^2 = t^2 ,$$

$$t = \frac{b-B}{\sqrt{s^2/\Sigma x_i^2}},$$

If the critical value of $t = t_a$, then confidence limits for β are

$$b \pm t_a \frac{s}{(\Sigma x_i^2)^{1/2}} .$$

Thus, we can examine any contemplated value, B, for the slope (which is really β) and see if the observations object to it vigorously enough. The confidence limits will surround those values not objected to by the observations. The diagram for this case looks just like the one for the mean, and we shall omit it.

Casting Back

Suppose we take the special case where all x_i are the same—in fact, all equal to λ. The model becomes

$$y_i = \beta\lambda + \epsilon_i, \qquad i = 1, 2, \ldots, n,$$

and we see that we have a special case of model (1) where $\eta = \beta\lambda$. It is a special case only because we assumed $\epsilon_\blacktriangle = 0$

TABLE IIA

ANALYSIS OF VARIANCE TABLE FOR MODEL (2)

Item	*DF*	*SS*	*MS*	*CMS*
slope	1	$(b-B)^2\Sigma x_i^2$	$(b-B)^2\Sigma x_i^2$	$(b-B)^2 - \left(1 - \dfrac{x_+^2}{N\Sigma x_i^2}\right)\dfrac{s^2}{\Sigma x_i^2}$
residue	$n-1$	$(n-1)s^2$	s^2	s^2
		AMS		*ACMS*
slope		$(B-\beta)^2 \Sigma x_i^2 + \left(1 - \dfrac{x_+^2}{N\Sigma x_i^2}\right)\sigma^2$		$(B-\beta)^2$
residue		σ^2		σ^2

TABLE IIIA

ANALYSIS OF VARIANCE TABLE FOR MODEL (3)

Item	*DF*	*SS*	*MS*	*CMS*
mean	1	$n(y_\bullet - Y_\bullet)^2$	$n(y_\bullet - Y_\bullet)^2$	$(y_\bullet - Y_\bullet)^2 - \left(\frac{1}{n} - \frac{1}{N}\right)s^2$
slope	1	$(b-B)^2\Sigma(x_i - x_\bullet)^2$	$(b-B)^2\Sigma(x_i - x_\bullet)^2$	$(b-B)^2 - s^2\{\Sigma(x_i - x_\bullet)^2\}^{-1}$
residue	$n-2$	$(n-2)s^2$	s^2	s^2

	AMS	*ACMS*
mean	$n(y_\blacktriangle - Y_\bullet)^2 - \left(1 - \frac{n}{N}\right)\sigma^2$	$(y_\blacktriangle - Y_\bullet)^2$
slope	$(B-\beta)^2\Sigma(x_i - x_\bullet)^2 + \sigma^2$	$(B-\beta)^2$
residue	σ^2	σ^2

in model (2). But we can always force $\epsilon_\blacktriangle = 0$ in the first model by defining η properly. Thus, we see that fitting a mean is a special case of fitting a slope. This may surprise us a little, but when we think a while it becomes reasonable. We shall see more and more of this sort of thing.

The Linear Case

We are now going to branch out boldly, and let the line go anywhere—no longer must it go through the origin. The model is

$$y_i = \alpha + \beta x_i + \epsilon_i, \qquad i = 1, 2, \ldots, n$$
$$\{x_i\} \text{ known without error} \qquad (3)$$
$$\{\epsilon_i\} \text{ a sample from a population of size } N, \text{ average and variance } \sigma^2.$$

This is appropriate when we wish to fit a line to observations of x and y where

1. the errors and fluctuations in x are negligible compared to those in y, and
2. the size of the errors in y do not depend on whether x or y is large or small.

There are other procedures for more complicated cases, and for those you may start with references 3 to 11.

The identities are, briefly,

$$\Sigma y_i^2 \equiv \Sigma y_\bullet + \Sigma\{b(x_i - x_\bullet)\}^2 + \Sigma\{(y_i - y_\bullet) - b(x_i - x_\bullet)\}^2 \qquad (3')$$
$$\equiv n y_\bullet^2 - b^2\Sigma(x_i - x_\bullet)^2 + (n-2)s^2 \qquad (\text{defines } s^2)$$

where

$$b = \frac{\Sigma(x_i - x_\bullet)(y_i - y_\bullet)}{\Sigma(x_i - x_\bullet)^2},$$

and if $Y_i = A + Bx_i$ is the contemplated line

$$\Sigma(y_i - Y_i)^2 = n(y_\bullet - Y_\bullet)^2 + (b-B)^2\Sigma(x_i - x_\bullet)^2 + (n-2)s^2 . \qquad (3'')$$

The tabular arrangement is given in Table IIIA, and the test ratios and confidence limits are those of Table IIB where $y_\blacktriangle$ is defined by

$$y_\blacktriangle = \alpha + \beta x_\bullet + \epsilon_\blacktriangle$$

to be just the average y for the fixed set of x-values that we have observed.

TABLE IIIB

TEST RATIOS AND CONFIDENCE LIMITS FOR MODEL (3)

$$F = \frac{n(y_\bullet - Y_\bullet)^2}{s^2} = \left\{\frac{y_\bullet - Y_\bullet}{\sqrt{s^2/n}}\right\}^2 = t^2 , \text{ (for mean).}$$

$$F = \frac{(b-B)^2\Sigma(x_i - x_\bullet)^2}{s^2} = \left\{\frac{b-B}{\sqrt{s^2/\Sigma(x_i - x_\bullet)^2}}\right\}^2 = t^2 , \text{ (for slope).}$$

If the critical value of $t = t_a$, then confidence limits for $y_\blacktriangle$ are

$$y \pm t_a\frac{s}{\sqrt{n}}$$

and for β are

$$b = t_a\frac{s}{\sqrt{\Sigma(x_i - x_\bullet)^2}} .$$

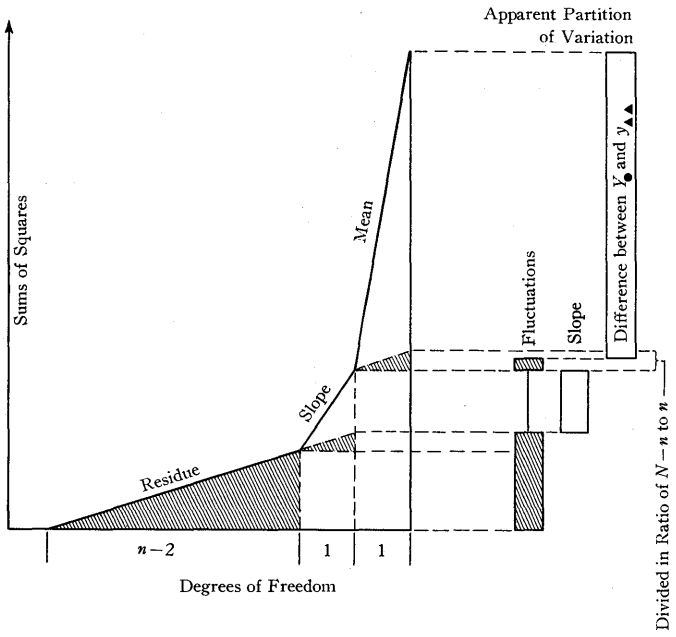


FIGURE 3. ANALYSIS OF VARIANCE DIAGRAM FOR MODEL (3)

We have broken up the sum of squares in a way entirely analogous to the one we used in the simple cases. The diagram is shown in Figure 3. Just as before, the test ratios (Table IIIB, page 101) are ratios of slopes of traces in this diagram. The shaded portions show how much allowance might carelessly (triangles) be made and is actually (blocks) made for the contribution of fluctuations to the sum of squares for the mean and for slope. Just as before, these allowances are correct on the average. The actual numerical values in the diagram correspond to a B which seems likely not to $= \beta$ and to a $Y_\bullet = A + Bx_\bullet$ which is quite clearly not $= y_\blacktriangle = \alpha + \beta x_\bullet$.

It may be interesting to write down the working forms of the three sums of squares. They are

$$\frac{\{y_+ - (nA+Bx_+)\}^2}{n}$$

$$\frac{\left\{\Sigma x_i y_i - \frac{1}{n}x_+y_+ - B\left(\Sigma x_i^2 - \frac{1}{n}x_+^2\right)\right\}^2}{\Sigma x_i^2 - \frac{1}{n}x_+^2} .$$

$$\Sigma y_i^2 - \frac{y_+^2}{n} - \frac{\left\{\Sigma x_i y_i - \frac{1}{n}x_+y_+\right\}^2}{\Sigma x_i^2 - \frac{1}{n}x_+^2} .$$

Orthogonality

Without explanation, the last model was taken apart into

1. a piece for the mean $\alpha + \beta x_\bullet$, and
2. a piece for the slope β.

It would be natural for you to ask why we didn't take out a piece for the intercept α and a piece for the slope β. The answer would be "this would be inconvenient, because the natural estimates of α and β are not orthogonal." And so you would wonder—what is orthogonality? Let us try to answer this question.

First, the change to $y_\blacktriangle = \alpha + \beta x_\bullet$ and β really corresponds to writing

$$\alpha + \beta x_i = (\alpha - \beta x_\bullet)1 + \beta(x_i - x_\bullet) .$$

It corresponds to using 1 and $(x_i - x_\bullet)$ as the quantities whose coefficients are to be found. The orthogonality of these quantities is easy to express algebraically. The condition is

$$\Sigma\, 1(x_i - x_\bullet) = 0 ,$$

which we know to be true. But why should we be interested in these quantities?

Let us write $y_\bullet$ and b in terms of b, β and the ϵ_i. We have

$$y_\bullet = \alpha + \beta x_\bullet + \epsilon_\bullet = y_\blacktriangle + \epsilon_\bullet - \epsilon_\blacktriangle$$

$$\begin{aligned} b\Sigma(x_i - x_\bullet)^2 &= \Sigma\{\alpha + \beta x_i + \epsilon_i - (\alpha + \beta x_\bullet + \epsilon_\bullet)\}(x_i - x_\bullet) \\ &= \beta\Sigma(x_i - x_\bullet)^2 + \Sigma(\epsilon_i - \epsilon_\bullet)(x_i - x_\bullet) , \\ &= \beta\Sigma(x_i - x_\bullet)^2 - \Sigma(x_i - x_\bullet)\epsilon_i \end{aligned}$$

so that

$$y_\bullet - Y_\blacktriangle = \epsilon_\bullet - \epsilon_\blacktriangle = \frac{1}{n}\Sigma\, 1(\epsilon_i - \epsilon_\blacktriangle)$$

$$\begin{aligned} b - \beta &= \frac{1}{\Sigma(x_i - x_\bullet)^2}\Sigma(x_i - x_\bullet)\epsilon_i \\ &= \frac{1}{\Sigma(x_i - x_\bullet)^2}\Sigma(x_i - x_\bullet)(\epsilon_i - \epsilon_\blacktriangle) . \end{aligned}$$

We are going to inquire about the tendency of these quantities to fluctuate together. To do this, we begin with

$$\text{average}\,\{(\epsilon_i - \epsilon_\blacktriangle)^2\} = \left(1 - \frac{1}{N}\right)\sigma^2 ,$$

$$\text{average}\,\{(\epsilon_i - \epsilon_\blacktriangle)(\epsilon_j - \epsilon_\blacktriangle)\} = -\frac{1}{N}\sigma^2 ,$$

hence

$$\begin{aligned} &\text{average}\,\{(y_\bullet - y_\blacktriangle)(b - \beta)\} \\ &= \frac{1}{n\,\Sigma(x_i - x_\bullet)^2}\{\sigma^2\Sigma 1(x_i - x_\bullet) + 0\} \end{aligned}$$

which vanishes because

$$\Sigma 1(x_i - x_\bullet)$$

does.

In general, the condition that

$$\text{average}\;\{\Sigma a_i(\epsilon_i - \epsilon_\blacktriangle)\,\Sigma b_j(\epsilon_j - \epsilon_\blacktriangle)\} \equiv 0$$

is that

$$\Sigma\, a_i b_i \equiv 0 .$$

Thus, one meaning of orthogonality is "no tendency to fluctuate together" (as measured by this particular sort of average). Since it is clearly convenient, but not essential, to work with quantities which do not tend to fluctuate together, this meaning will perhaps content you now. Another meaning is discussed on page 104.

Multiple Regression

Now we can look at the machinery for what is called multiple regression—a very general procedure. The model looks as follows:

$$\begin{cases} y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_m x_{mi} + \epsilon_i \\ \qquad\qquad\qquad\qquad i = 1, 2, \ldots, n \\ \{x_{ji}\} \text{ is known without error} \\ \{\epsilon_i\} \text{ is a randomly arranged sample from a population of size } N, \text{ average zero and variance } \sigma^2. \end{cases} \tag{4}$$

The independent variables, or carriers of regression, x_1, x_2, . . . , x_m can be anything. They can be related or unrelated, constant or not. Thus, the case

$$m = 2: \quad x_1 \equiv 1, \quad x_2 \equiv x \ ,$$

covers the general straight line, while

$$m = 3: \quad x_1 \equiv 1, \quad x_2 \equiv u, \quad x_3 \equiv v \ ,$$

covers fitting the general plane to the values of y in terms of $(1,u,v)$. In a medical case, for example, we might take

$m = 5$: $x_1 \equiv 1$, $x_2 \equiv$ log blood pressure, $x_3 \equiv$ age,
$x_4 \equiv$ weight as per cent of normal for height and age,
$x_5 \equiv$ excitability.

You can propose all sorts of examples in the fields that interest you.

There are statisticians who prefer to deal with the original variables $x_1, x_2, \ldots, x_m$ all together and on an equal footing. This is permissible, but I don't like to do it, since it allows no place for using the extra knowledge that you almost always have. It also leads to matrix formulas that are too much trouble to remember for one like me who does such calculations infrequently. (If you do such computations regularly, and under circumstances where you are sure that you want complete analyses, then by all means use the matrix formulas.) Instead, I am going to describe a one-at-a-time procedure, where we start with the most promising variable first, the next most promising variable second, and so on. (We suppose the numbering changed so that x_1 is most promising, x_2 next, and so on.) Then we form a new set of variables

$$\begin{aligned} x_{1\cdot} &\\ x_{2\cdot 1} &= x_2 - b_{21}\, x_1 \\ x_{3\cdot 12} &= x_3 - b_{31}\, x_1 - b_{32\cdot 1}\, x_{2\cdot 1} \\ &\ldots \end{aligned}$$

which are orthogonal (the dots do not mean averages here). This is entirely like changing from 1, x to 1, $(x - x_\bullet)$ in fitting a straight line. The consequence is a simple procedure for fitting the constants in

$$b_1 x_1 + b_{2\cdot 1} x_{2\cdot 1} + b_{3\cdot 12} x_{3\cdot 12} + \ldots \ ;$$

a procedure of fitting them one at a time from left to right. Thus, we take out a piece for x_1 first (this is the biggest if we guessed well) then one for $x_{2\cdot 1}$ (this ought to be relatively big, too) and so on.

Notation and Identities

To make the wheels turn smoothly, we need a complex-seeming but simple operating notation for certain sums and the results of simple operations on them. We write

$$\begin{aligned} [j,k] &= \Sigma_i\, x_{ji} x_{ki} \ , \qquad 1 \leqq j,k \leqq m \ , \\ [y,j] &= \Sigma_i\, y_i x_{ji} \ , \qquad 1 \leqq j \leqq m \ , \end{aligned}$$

and then define dotted quantities according to the general symbolic principle

$$[j\cdot Lm, k\cdot Lm] = [j\cdot L, k\cdot L] - \frac{[j\cdot L, m\cdot L]\,[k\cdot L, m\cdot L]}{[m\cdot L, m\cdot L]} \ .$$

Thus, we have

$$[j\cdot 1, k\cdot 1] = [j,k] - \frac{[j,1]\,[k\cdot 1]}{[1,1]} \ ,$$

$$[j\cdot 12, k\cdot 12] = [j\cdot 1, k\cdot 1] - \frac{[j\cdot 1, 2\cdot 1]\,[k\cdot 1, 2\cdot 1]}{[2\cdot 1, 2\cdot 1]} \ ,$$

(and so on) ,

$$[y\cdot 1, j\cdot 1] = [y,j] - \frac{[y,1]\,[j,1]}{[1,1]} \ ,$$

(and so on) .

We also define the new variables by

$$x_{j\cdot 1} = x_j - \frac{[j,1]}{[1,1]}\, x_1 = x_j - b_{j1} x_1 \ ,$$

$$x_{j\cdot 12} = x_{j\cdot 1} - \frac{[j\cdot 1, 2\cdot 1]}{[2\cdot 1, 2\cdot 1]} = x_{j\cdot 1} - b_{j2\cdot 1}\, x_{2\cdot 1} \ ,$$

(and so on) ,

and the residuals by

$$y_{\cdot 1} = y - \frac{[y,1]}{[1,1]} x_1 = y - b_1 x_1$$

$$y_{\cdot 12} = y_{\cdot 1} - \frac{[y\cdot 1, 2\cdot 1]}{[2\cdot 1, 2\cdot 1]}\, x_{2\cdot 1} = y_{\cdot 1} - b_{2\cdot 1}\, x_{2\cdot 1}$$

(and so on) .

With all this machinery chosen for our convenience, it would be strange if the identities were not relatively simple. They are

$$\begin{aligned} \Sigma y_1^2 \equiv & \frac{[y,1]^2}{[1,1]} + \frac{[y\cdot 1, 2\cdot 1]^2}{[2\cdot 1, 2\cdot 1]} + \frac{[y\cdot 12, 3\cdot 12]^2}{[3\cdot 12, 3\cdot 12]} + (\text{to } m \text{ terms}) \\ & + \Sigma (y_{\cdot 123\ldots m})^2 \qquad (4') \\ \equiv & [1,1] b_1^2 + [2\cdot 1, 2\cdot 1]\,(b_{2\cdot 1})^2 \\ & + [3\cdot 12, 3\cdot 12]\,(b_{3\cdot 12})^2 + \ldots + (n-m) s^2 \end{aligned}$$

and

$$\begin{aligned} & \Sigma \{ y_i - (B_1 x_1 + B_{2\cdot 1} x_{2\cdot 1} + B_{3\cdot 12} x_{3\cdot 12} + (\text{to } m \text{ terms})) \}^2 \\ & \equiv [1,1]\,(b_1 - B_1) + [2\cdot 1, 2\cdot 1]\,(b_{2\cdot 1} - B_{2\cdot 1})^2 \qquad (4'') \\ & \qquad + (\text{to } m \text{ terms}) + (n-m) s^2. \end{aligned}$$

The working computations would go in terms of the brackets and the first form of (4'), and notation would be simplified by labeling $[j\cdot L, k\cdot L]$ just $[jk\cdot L]$.

Table and Diagram

The results would be given as in Table IVA, which follows the pattern of the previous analysis of variance tables. Usually such a table is given for $B_1 = B_{2\cdot 1} = \ldots = 0$, but this is habit rather than thought-out policy. Such a table should always be supplemented by a table of values of b_1, $b_{2\cdot 1}$, $b_{3\cdot 12}$, . . . and usually by a table of the expressions of $x_{2\cdot 1}$, $x_{3\cdot 12}$, and so on, in terms of the original variables $x_1, x_2, \ldots, x_m$.

TABLE IVA

ANALYSIS OF VARIANCE FOR MODEL (4)

Item	*DF*	*SS*	*MS*
x_1	1	$[1,1](b_1-B_1)^2$	$[1,1](b_1-B_1)^2$
$x_{2\cdot1}$	1	$[2\cdot1,2\cdot1](b_{2\cdot1}-B_{2\cdot1})^2$	$[2\cdot1,2\cdot1](b_{2\cdot1}-B_{2\cdot1})^2$
$x_{3\cdot12}$	1	$[3\cdot12,3\cdot12](b_{3\cdot12}-B_{3\cdot12})^2$	$[3\cdot12,3\cdot12](b_{3\cdot12}-B_{3\cdot12})^2$
.	...	...	...
.	...	...	...
.	...	...	...
residue	$n-m$	$(n-m)s^2$	s^2

	CMS	*AMS*
x_1	$(b_1-B_1)^2-\left(1-\frac{x_{1+}^2}{N[1,1]}\right)\frac{s^2}{[1,1]}$	$[1,1](\beta_1-B_1)^2+\left(1-\frac{x_{1+}^2}{N[1,1]}\right)\sigma^2$
$x_{2\cdot1}$	$(b_{2\cdot1}-B_{2\cdot1})^2-\frac{s^2}{[2\cdot1,2\cdot1]}$	$[2\cdot1,2\cdot1](\beta_{2\cdot1}-B_{2\cdot1})^2+\sigma^2$
$x_{3\cdot12}$	$(b_{3\cdot12}-B_{3\cdot12})^2-\frac{s^2}{[3\cdot12,3\cdot12]}$	$[3\cdot12,3\cdot12](\beta_{3\cdot12}-B_{3\cdot12})^2+\sigma^2$
.	...	...
.	...	...
.	...	...
residue	s^2	σ^2

	ACMS
x_1	$(\beta_1-B_1)^2$
$x_{2\cdot1}$	$(\beta_{2\cdot1}-B_{2\cdot1})^2$
$x_{3\cdot12}$	$(\beta_{3\cdot12}-B_{3\cdot12})^2$
.	...
.	...
.	...
residue	σ^2

The corresponding diagram follows along the usual pattern and is shown in Figure 4. This figure is drawn for the special case

$m = 4$: $x_1 \equiv 1$, $x_2 \equiv x$, $x_3 \equiv x^2$, $x_4 \equiv x^3$.

Thus, the various components, those for 1, x —$x_\bullet$, x^2 — . . . , x^3 — . . . , are called "mean," "slope," "curvature," and "twist."

Geometric Interpretation

All these mysterious notations and identities have a nice geometric interpretation. Just make a vector out of each of the m variables—m vectors in an n-dimensional space. The initial set of coordinates in the space is like the statistician's sample space; the coordinates of $\vec{x}_1$ in this system are the n observed values of x_1 for the n observations, and so on

$$\vec{x}_1 = \{x_{1i}\}$$
$$\vec{x}_2 = \{x_{2i}\}$$
$$\vec{x}_m = \{x_{mi}\}$$
$$\vec{y} = \{y_i\} .$$

Now, experimenters and quantities of interest being as they are, the m vectors $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_m$ that correspond to the m variables are unlikely to be at right angles to each other.

But the aim of our fitting process comes out to be just finding the component of $\vec{y}$ in the m-dimensional space determined by $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_m$. This would be easy if $\vec{x}_1, \ldots, \vec{x}_m$ were at right angles. So we set out to force them to right angles and calculate the projection all at once.

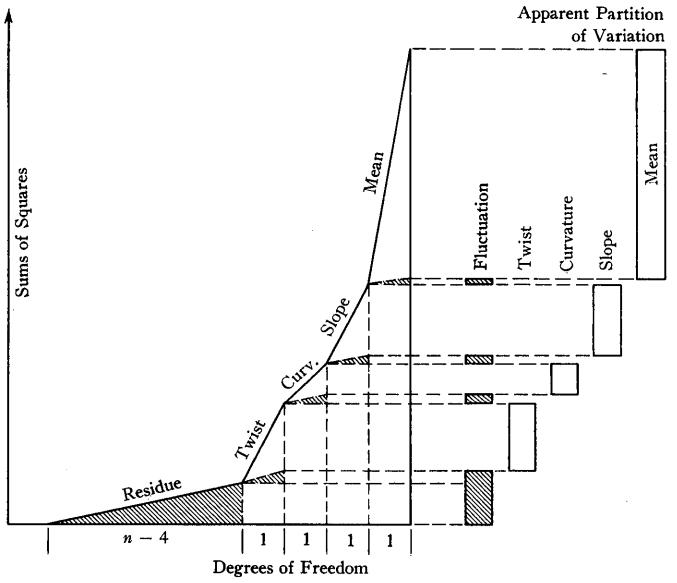


FIGURE 4. ANALYSIS OF VARIANCE DIAGRAM FOR MODEL (4) SPECIAL CASE OF POLYNOMIAL FITTING

It is easy to replace $\vec{x}_2, \ldots, \vec{x}_m, \vec{y}$ by their components perpendicular to $\vec{x}_1$. We have only to calculate a few dot products and proceed as follows:

$$\vec{x}_{2\cdot 1} = \vec{x}_2 - \frac{(\vec{x}_1 \cdot \vec{x}_2)}{(\vec{x}_1 \cdot \vec{x}_1)} x_1 ,$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$\vec{x}_{m\cdot 1} = \vec{x}_m - \frac{(\vec{x}_m \cdot \vec{x}_1)}{(\vec{x}_1 \cdot \vec{x}_1)} \vec{x}_1 ,$$

$$\vec{y}_{\cdot 1} = \vec{y} - \frac{(\vec{y} \cdot \vec{x}_1)}{(\vec{x}_1 \cdot \vec{x}_1)} \vec{x}_1 .$$

Then we can shift from $\vec{x}_{3\cdot 1}, \ldots, \vec{x}_{m\cdot 1}, \vec{y}_{\cdot 1}$ to their components perpendicular to $\vec{x}_{2\cdot 1}$. These new vectors will still be perpendicular to $\vec{x}_1$. And so on.

By comparison with the operations on brackets specified above, you can see that this orthogonalization procedure is just what we have been doing. From the vectors $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_m$, we have constructed new vectors $\vec{x}_1, \vec{x}_{2\cdot 1}, \vec{x}_{3\cdot 12}, \ldots,$ at right angles to each other, and we have found the components $b_1\vec{x}_1, b_{2\cdot 1}\vec{x}_{2\cdot 1}, \ldots,$ of y along the new vectors and the residual vector $\vec{y}_{\cdot 12\ldots m}$ which is the component of $\vec{y}$ perpendicular to all the $\vec{x}$'s.

All this still may seem complicated. So let us go back and fit a line to two points. We take $n = m = 2$, $x_1 \equiv 1$, $x_2 \equiv x$ and assume the observations

$$(x = 2, y = 1)$$
$$(x = 4, y = -2)$$

The vectors are shown in the original coordinate system in Figure 5.

The result of passing to $\vec{x}_{2\cdot 1}$ and $\vec{y}_{\cdot 1}$ is to replace (2,4) by $(-1,+1)$ and $(1,-2)$ by $\left(1\frac{1}{2}, -1\frac{1}{2}\right)$. This is done through

$$\vec{x}_{2\cdot 1} = \vec{x}_2 - 3\vec{x}_1 ,$$
$$\vec{y}_{\cdot 1} = \vec{y} - \frac{1}{2}\vec{x}_1 .$$

Now we notice that

$$\vec{y}_{\cdot 1} = -\frac{3}{2}\vec{x}_{2\cdot 1}$$

and hence

$$\vec{y} = -\frac{1}{2}\vec{x}_1 - \frac{3}{2}\vec{x}_{2\cdot 1} = 4\vec{x}_1 - \frac{3}{2}\vec{x}_2 .$$

Recalling that $x_1 = 1$, $x_2 = x$, we see that we have fitted the line

$$y = 4 - \frac{3}{2}x$$

to the points (2,1) and (4,—2). It is easy to see that this is a correct fit.

The geometric interpretation is the same for any regression problem. It merely requires more than two dimensions for the picture.

If we had not orthogonalized, then the problem of finding the projection of $\vec{y}$ on the m-dimensional plane of $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_m$ leads at once to m equations in m unknowns. As practical computers, how would we have solved these equations? By some method of elimination—whether we talk of Doolittle, Crout, or Dwyer's square root method. Geometrically

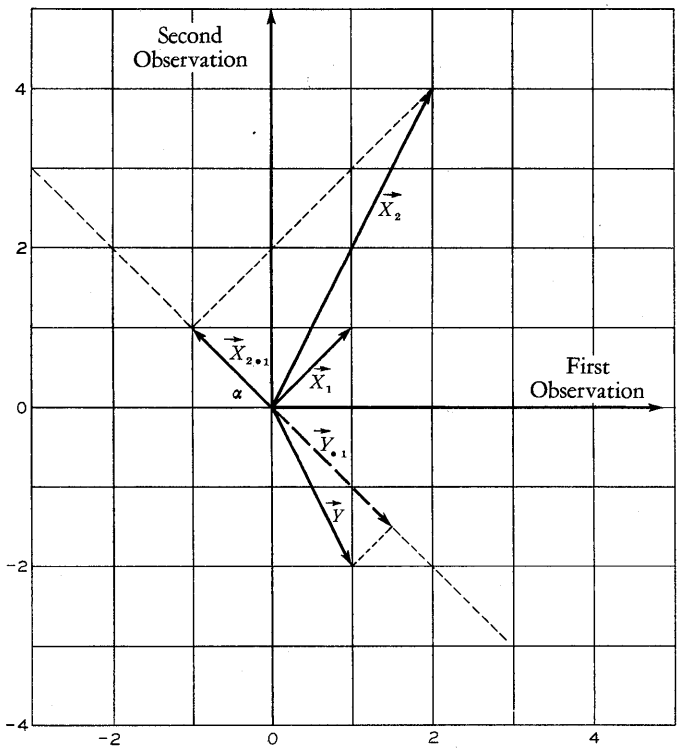


FIGURE 5

this means by orthogonalization, by something close to our actual procedure. What difference might there be!

Really the difference is only this: we have not abbreviated the method, we have put down all the steps. This leaves us with a chance to change our mind—if [7.123456, 7.123456] is very small, so small that the errors of measurement we have been neglecting account for most of it, we can drop variable 7 without loss of work and go on to 8.123456, and then to 9.1234568. We have written or typed or magnetized or punched or mercuried some extra numbers. With these we have bought insight at intermediate points and flexibility. I like to do it this way; you may like to do it another.

References

Now there are more complex problems in regression, and more complex ways to do simple problems. These we can only cover by reference.

For problems where the observations are of varying accuracy or where the coefficients appear nonlinearly, the classical methods are set out in reference 3.

For problems involving polynomials at equally spaced intervals, much time and trouble is saved by the use of already prepared orthogonal polynomials. These are available as follows:

To 5th degree, to 52 points, in reference 4.
To 5th degree, to 75 points, in reference 5.
To 5th degree, 53 to 104 points, in reference 6.
To 9th degree, to 52 or more points, in reference 7.
To 5th degree, to 21 points, in reference 8.

For problems involving error in more than one variable, the user should read references 9 and 10, and for further study the references given there. A very condensed summary of many theoretical results may be found in reference 11.

Application of regression ideas to more general problems can be found in reference 12, and computational suggestions can be found in some of the texts in references 14-18.

Single Classification

Several Groups—A Single Classification

Regression, in the sense that we have used it—curve-fitting, general mopping up with sums of terms, and the like—accounts for many physical and chemical applications. But, in many fields the type of analysis that we now enter is the standard. Particularly in agriculture, only slightly less in engineering and applied sciences, and to some extent everywhere, the comparative experiment is king.

Simple before-and-after experiments, or comparisons of two brands, two processes, two finishes or two raw materials, are easy; by taking differences you return to the type of case we have discussed. We need to consider comparisons of several categories. The simplest experiment is one in which you have n_i observations in category i, where i runs from 1 to c. That is, n_1 observations on the first brand, the product of the first process, material covered with the first finish or units made from the first raw material; n_2 observations on the second brand, the product of the second process, material covered with the second finish or units made from the second raw material; and so made on all categories. The model will exhibit each observation as the sum of a contribution depending on the category and a fluctuating contribution.

Here you can use quite complicated models for the fluctuating component with good sense and good results. For many of these models the part of the analysis which we are going to describe is the same. So I am not going to say just what I assume here. If you assume, for example, that all the fluctuations for all the categories come randomly out of one big population of fluctuations, then you will have a model that will fit a lot of circumstances. Everything that is going to be said here will apply to that model. If you want more ideas about possible models, read pages 69-75 of reference 13.

We need some sort of a model, however, so that we can describe average values and variances of things. We specify a simple one, namely,

$$\begin{cases} y_{ij} = \lambda + \eta_i + \epsilon_{ij},\ 1 \leqq j \leqq n_i, \\ \lambda \text{ fixed} \\ \{\eta_i\} \text{ a sample from a population (of categories) of size } M, \text{ average } \eta_{\blacktriangle} \text{ and variance } \sigma_\eta^2. \\ \{\epsilon_{ij}\} \text{ a sample from a population (of fluctuations) of size } N, \text{ average } \epsilon_{\blacktriangle\blacktriangle}, \text{ and variance } \sigma^2. \end{cases} \tag{5}$$

This is a good standard model, but not the most general for which our analysis is suitable.

The categories in such a comparative experiment may be anything. In a study of screws, they might be different automatic screw machines, where you had taken a handful of screws from each for measurement. Or they might be the different times at which you had taken a handful from one machine. Or they might be different months in which you had sampled the whole factory's production. In agriculture they may be different varieties receiving different fertilizers. The categories may be individual operators of a chemical unit process, or they may be different fatigue states (as measured by time on shift) of a single operator. You have a choice of a lot of things here.

Identities

Our identities follow the standard pattern; here there are three pieces, as we see in (5′) and (5″).

$$\begin{aligned} \Sigma y_{ij}^2 &\equiv \Sigma y_{\bullet\bullet}^2 + \Sigma (y_{i\bullet} - y_{\bullet\bullet})^2 + \Sigma (y_{ij} - y_{i\bullet})^2 && (5') \\ &\equiv n y_{\bullet\bullet}^2 + \Sigma_i\, n_i (y_{i\bullet} - y_{\bullet\bullet})^2 + (n-c)s^2 && (\text{defines } s^2) \\ &\equiv \frac{1}{n} y_{++}^2 + \left\{ \left(\Sigma_i \frac{1}{n_i} y_{i+}^2 \right) - \frac{1}{n} y_{++}^2 \right\} + \left\{ \Sigma y_{ij}^2 - \Sigma_i \frac{1}{n_i} y_{i+} \right\} \end{aligned}$$

$$\begin{aligned} \Sigma (y_{ij} - Y_i)^2 &\equiv \Sigma (y_{i\bullet} - Y_i)^2 - \Sigma (y_{ij} - y_{i\bullet})^2 \\ &\equiv \Sigma_i\, n_i (y_{i\bullet} - Y_i)^2 + (n-c)s^2 && (5'') \\ &\equiv n(y_{\bullet\bullet} - Y_\bullet)^2 + \\ &\qquad \Sigma_i\, n_i (y_{i\bullet} - y_{\bullet\bullet} - Y_i + Y_\bullet)^2 + (n-c)s^2. \end{aligned}$$

Where the dottings in place of the i's mean weighted averaging, that is,

$$y_{\bullet\bullet} = \frac{\Sigma y_{ij}}{n = \Sigma n_i} \qquad Y_{\bullet} = \frac{\Sigma n_i y_i}{n = \Sigma n_i}$$

and n is defined as in these denominators.

In terms of the second line of (5′) the three pieces stand out clearly.

1. A piece depending on $y_{\bullet\bullet}$ which expresses the fact that the sample grand mean is not zero.
2. A piece comparing the category means among themselves.
3. A piece expressing fluctuations within a category.

Just as in our first case, we have siphoned into the last term all that we possibly could of each of the fluctuations without getting category or grand mean effects. Likewise, we have siphoned as much as we can of the category-to-category differences into the second piece without getting grand mean effects. Our purification is only partial, but it is the best that we can do. It is the old method, applied at two levels instead of one, in two stages instead of one. We have isolated the fluctuations within categories from the category means as well as possible. Then we have isolated the differences in category means from the grand mean as well as we are able. On two levels at once, we use the same process which you use unconsciously when you take an average.

Tables

The elementary question that is going to be asked is: "Are these categories different?" This is only the first question, and those who stop with it are probably not getting what they should out of the observations. From the standpoint of the computing group, it doesn't make much difference, because if they have answered this they have, ready at hand, the other numbers which might be needed. *Whether asked for or not, always send the category means back upstairs with the analysis of variance table.* Don't let the statisticians forget the means for the sake of significance tests!

The form of the analysis of variance table is shown in Table Va. We have shown the one degree of freedom for the mean which many leave out. It has nothing to do with the comparison of categories, and since that is what such analyses are usually for, it is often omitted. But there are analyses that come in exactly this form where this line contains key information. If you attacked the problem which was mentioned previously at this Seminar—getting the average height of all the men in the United States—it would not be very practical to try to draw a random sample directly of a thousand men out of all the inhabitants of the United States. No one has a convenient card file that you can enter with random numbers and pull out names. You would want at the very least to break up the United States into pieces, and select randomly and measure two or three men in each of several randomly selected pieces. If you did this you would have a situation that comes under this model; because, if you broke up the United States into pieces in any reasonable way, the average heights of the men in the different pieces would be different, and these differences from piece to piece might be crucial in fixing the accuracy of your over-all mean.

There are approximately 3,000 counties in the United States. Some of them, like Manhattan, are a little large and inhomogeneous. Let us think in terms of 10,000 categories. These are to be geographical regions, each with about the same number of men. (What is a man, anyway?) If we

TABLE Va

ANALYSIS OF VARIANCE TABLE FOR MODEL (5)

Item	*DF*	*SS*	*MS*	*CMS*
mean	1	$ny_{\bullet\bullet}^2$	$ny_{\bullet\bullet}^2$	(*)
categories	$c-1$	$\Sigma_i n_i(y_{i\bullet}-y_{\bullet\bullet})^2$	$\frac{1}{c-1}\Sigma_i n_i(y_{i\bullet}-y_{\bullet\bullet})^2$	(*)
within	$n-c$	$(n-c)s^2$	s^2	s^2

	AMS	*ACMS*
mean	$\left(1-\frac{n}{N}\right)\sigma^2+\left(1-\frac{c}{M}\frac{\Sigma n_i^2}{n}\right)\sigma_\eta^2+ny_{\blacktriangle\blacktriangle}^2$	$y_{\blacktriangle\blacktriangle}^2$
categories	$\sigma^2+\frac{n^2-\Sigma n_i^2}{n(c-1)}\sigma_\eta^2$	σ_η^2
within	σ^2	σ^2

*Best computed from the numerical values of the coefficients in the *AMS* column.

selected 100 of these at random, and then selected three men for measurement randomly within each of the hundred, the grand average tells us a lot about the average height of U.S. men. The grand average is going to fluctuate for two reasons. One reason is that if you repeat the process you would not have the same three men in a given category. The other is likely to be more important; if you repeated the process you would have a different set of 100 categories.

You have here a situation where it makes sense to write for any individual

height = U.S. average + (category average − U.S. average)
+ (individual height − category average) .

Now, our grand average is the sum of the grand averages of the three contributions for each individual. If you redo the whole process, and use a new sample of 100 categories, then the average of the 100 category averages will be different; the grand average of the second contributions will be different. We must allow for this as well as for the fluctuations in the grand average of the third contribution.

Before we go on, we notice that Table VA is a little complicated, and conjecture that this is due to the possibility of having different numbers of observations in the different categories. So we treat the case (Table VIA)

like (5) except $n_i \equiv r$ for $i = 1, 2, \ldots, c$. (6)

Here things are quite simple in every line, except that for the mean.

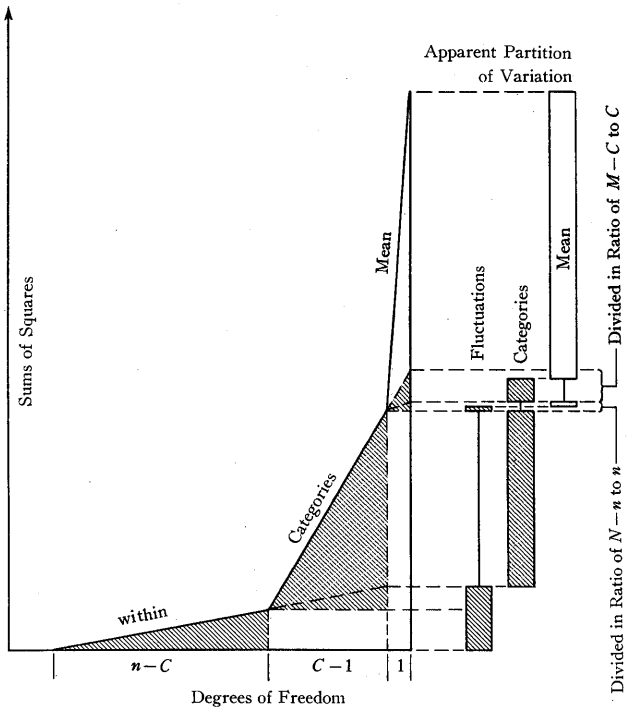


FIGURE 6. ANALYSIS OF VARIANCE DIAGRAM FOR MODEL (6)

Diagram

Having the tables, we can now set forth the diagrams, which we do in Figure 6 for model (6).

If we examine this diagram we see that it is much like the other diagrams. We have the traces, one for each line of the table. Clearly, one degree of freedom goes into the grand mean, and that is lost from among the c categories. So there must be $c-1$ degrees of freedom for categories. This disposes of a total of c degrees of freedom; there were n observations. Take c from n and you have $n-c$, the number of degrees of freedom within categories.

We should certainly call the differences between categories "apparently negligible" if the traces for "within" and

TABLE VIA

ANALYSIS OF VARIANCE TABLE FOR MODEL (6)

Item	*DF*	*SS*	*MS*	*AMS*
mean	1	$ny_{\bullet\bullet}^2$	$ny_{\bullet\bullet}^2$	$\left(1-\frac{n}{N}\right)\sigma^2+\left(1-\frac{c}{M}\right)r\sigma_\eta^2+ny_{\blacktriangle\blacktriangle}^2$
categories	$c-1$	$r\Sigma_i(y_{i\bullet}-y_{\bullet\bullet})^2$	$\frac{r}{c-1}\Sigma_i(y_{i\bullet}-y_{\bullet\bullet})^2$	$\sigma^2+r\sigma_\eta^2$
within	$n-c$	$(n-c)s^2$	s^2	σ^2

	ACMS	*CMS*
mean	$y_{\blacktriangle\blacktriangle}^2$	(*)
categories	σ_η^2	$\frac{1}{c-1}\Sigma(y_{i\bullet}-y_{\bullet\bullet})^2-\frac{1}{r}s^2$
within	σ^2	s^2

$$(*) = \frac{1}{n}(MS\ \text{mean}) - \frac{1}{n}\left(1-\frac{c}{m}\right)(MS\ \text{categories}) - \left(\frac{1}{N}-\frac{c}{nM}\right)(MS\ \text{within}).$$

categories have the same slope (lay along the same line). For this would mean that the component of mean square for categories was zero (it is zero on the average only when the η_i are the same), and this is a precise form of the rough statement that the categories are just alike. So we take the usual shaded triangle away from the triangle for categories.

When $N=M=\infty$, and the traces for categories and the mean fall in the same line, then the component of mean square for the mean is zero, and we conclude that the grand mean might be zero. Thus, we take both shaded and dotted triangles away from the triangle for the mean. When M is large, and many categories to be considered are not represented in the experiment, we compare the mean square for the mean with the mean square for categories.

Another extreme is $N=\infty$, $M=c$, when the traces for the mean and for "within" must lie on the same line for the component of mean square for the mean to vanish. Here only the shaded triangle is taken away from the triangle for the mean. When all categories to be considered were represented in the experiment, we compare the mean square for the mean with the mean square "within."

And some situations fall in between, as the diagram illustrates.

Test Ratios and Confidence Limits

We can again look for appropriate test ratios and confidence limits, with the results shown in Table VIB.

TABLE VIB

TEST RATIOS AND CONFIDENCE LIMITS FOR MODEL (6)

Are there differences between categories!

$$F=\frac{\frac{r}{c-1}\Sigma_i(y_{i\bullet}-y_{\bullet\bullet})^2}{s^2}=\frac{MS\text{ categories}}{MS\text{ within}}$$

Might the mean equal Y? (M large)

$$F=\frac{n(y_{\bullet\bullet}-Y)^2}{\frac{r}{c-1}\Sigma_i(y_{i\bullet}-y_{\bullet\bullet})^2}=\frac{MS\text{ mean}}{MS\text{ categories}}=t^2$$

Might the mean equal Y? ($M=c$)

$$F=\frac{n(y_{\bullet\bullet}-Y)^2}{s^2}=\frac{MS\text{ mean}}{MS\text{ within}}=t^2$$

Confidence limits for $y_{\blacktriangle\blacktriangle}$? ($M$ large)

$$y_{\bullet\bullet}\pm t_{a_1}\left\{\sqrt{\frac{r}{n(c-1)}\Sigma_i(y_{i\bullet}-y_{\bullet\bullet})^2}\right\}$$

Confidence limits for $y_{\blacktriangle\blacktriangle}$? ($M=c$)

$$y_{\bullet\bullet}\pm t_{a_2}\frac{s}{\sqrt{n}}$$

(For $c<M<\infty$ combine MS within and MS categories as suggested by AMS's of Table VIA.)

The great difference in testing the mean—the great dependence on whether

sampled categories = considered categories

or

sampled categories << considered categories

shows up in more complex designs with avidity, subtlety and frequency. There it affects comparisons and is worthy of the analyst's best attention.

DOUBLE CLASSIFICATION

Basis

Fifty years ago it was claimed that the way to run an experiment was to vary one thing at a time. If the nature of the subject is such that the results are not going to make sense, this is still the way to run an experiment. But, if in your subject the results make some kind of sense, it is usually much better to vary two things at once, or three things at once, or more. One of my friends is faced with an engineering problem where he is planning to vary 22 things at once. I don't advise you to start with that many, but he will learn more per dollar than if he varied one at a time.

If it makes sense to set up a model like this,

$$y_{ij}=\lambda+\eta_i+\phi_j+\epsilon_{ij}$$

where i refers to the level or nature of one thing and j to the level or nature of the other, where the plus signs are really plus signs, and the ϵ_{ij} are really random fluctuations, then it is much more efficient and useful to vary both things in a single experiment.

If there is no semblance of a plus sign—if, for example, y increases when i increases for one value of j, but decreases when i increases for another value of j—then there may be little profit in varying two at once. There is not likely to be loss in varying only two at once, but more complex experiments (such as Latin squares, which we will not discuss) may burn the hand that planned them.

But, fortunately, life is reasonably simple in most subjects. The plus sign will be a good approximation often enough for the use of such experiments to pay. It may not be gratis; you may have to work for it. For example, the y's that you finally analyze may not be those with which you started. If you happened to be working on blood pressure, you may have to use the logarithm of the measured blood pressure; it is unlikely to be satisfactory to use the raw data in millimeters. For reasons that make sense when you think about them, factors that affect blood pressure tend to multiply together rather than add together in their effects, and then the logarithms are additive.

Again the statistician ought to think hard about such matters. He ought to see the need for transformations. But sometimes the computing people may see something going on that will make clear to them that there ought to be a transformation. If the plot of the effect of one variable for

different values of another looks like Figure 7, for example, if the effect seems faster at higher levels, then we are a long way from a plus sign. The cure for this particular sort of deviation is to squeeze things closer together at the higher values than at the lower ones. You can do this by changing to the square root of the observed values, or to their logarithms; one or the other may work.

Things of this sort need to be kept in mind. The honesty of the plus sign controls the extent to which the observations are adequately squeezed dry by this procedure.

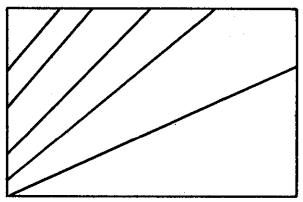


FIGURE 7

Model, Identity, Table and Diagram

We shall treat this case briefly. A reasonable model for many uses is

$$\left\{\begin{array}{l} y_{ij} = \lambda + \eta_i + \phi + \epsilon_{ij}, \quad 1 \leqq i \leqq c, \quad 1 \leqq j \leqq r, \\ \lambda \text{ fixed}, \\ \{\eta_i\} \text{ a sample from a population of size } N_\eta, \\ \quad \text{average } \eta_\blacktriangle \text{ and variance } \sigma_\eta^2, \\ \{\phi_j\} \text{ a sample from a population of size } N_\phi, \\ \quad \text{average } \phi_\blacktriangle, \text{ and variance } \sigma_\phi^2, \\ \{\epsilon_{ij}\} \text{ a sample from a population of size } N, \\ \quad \text{average } \epsilon_{\blacktriangle\blacktriangle}, \text{ and variance } \sigma^2. \end{array}\right. \tag{6}$$

This is a case where the number of observations with $i=i_0$, $j=j_0$ is $g(i_0)h(j_0)$—in fact $g(i) \equiv 1 \equiv h(j)$—so that the i-classification and the j-classification are orthogonal. This simplifies matters considerably.

One identity is

$$\begin{aligned} \Sigma y_{ij}^2 &\equiv \Sigma y_{\bullet\bullet}^2 + \Sigma (y_{i\bullet} - y_{\bullet\bullet})^2 + \Sigma (y_{\bullet j} - y_{\bullet\bullet})^2 \\ &\quad + \Sigma (y_{ij} - y_{i\bullet} - y_{\bullet j} + y_{\bullet\bullet})^2 \\ &\equiv rc y_{\bullet\bullet}^2 + r\Sigma_i (y_{i\bullet} - y_{\bullet\bullet})^2 - c\Sigma_j (y_{\bullet j} - y_{\bullet\bullet})^2 \\ &\quad + (r-1)(c-1)s^2 \qquad (6') \\ &\equiv \frac{1}{rc} y_{++}^2 + \left\{ \frac{1}{r} \Sigma_i y_{i+}^2 - \frac{1}{rc} y_{++}^2 \right\} + \left\{ \frac{1}{c} \Sigma_j y_{+j}^2 - \frac{1}{rc} y_{++}^2 \right\} \\ &\quad + \left\{ \Sigma y_{ij}^2 - \frac{1}{c} \Sigma_j y_{+j}^2 - \frac{1}{r} \Sigma_i y_{i+}^2 + \frac{1}{rc} y_{++}^2 \right\}. \end{aligned}$$

The analysis of variance table is given as Table VIIa, and the diagram as Figure 8. The two classifications are conveniently referred to as "columns" and "rows."

The details of this model and those of many more complicated ones we must leave to the reader's thought and study. No two books will give him the same account, but a few of interest are given in references 14-18.

TABLE VIIa

ANALYSIS OF VARIANCE TABLE FOR MODEL (7)

Item	*DF*	*SS*	*MS*	*CMS*
mean	1	$rc(y_{\bullet\bullet}-Y)^2$	$rc(y_{\bullet\bullet}-Y)^2$	(*)
columns	$c-1$	$r\Sigma_i(y_{i\bullet}-y_{\bullet\bullet})^2$	$\frac{r}{c-1}\Sigma_i(y_{i\bullet}-y_{\bullet\bullet})^2$	$\frac{1}{r}(MS\text{ columns}) - \frac{1}{r}s^2$
rows	$r-1$	$c\Sigma_j(y_{\bullet j}-y_{\bullet\bullet})^2$	$\frac{c}{r-1}\Sigma_j(y_{\bullet j}-y_{\bullet\bullet})^2$	$\frac{1}{c}(MS\text{ rows}) - \frac{1}{c}s^2$
residue	$(c-1)(r-1)$	$(c-1)(r-1)s^2$	s^2	s^2

	AMS	*ACMS*
mean	$\left(1-\frac{rc}{N}\right)\sigma_\eta^2 + \left(1-\frac{r}{N_\phi}\right)\sigma_\phi^2 - \left(1-\frac{c}{N_\eta}\right)\sigma_\eta^2 - rc(y_{\bullet\bullet}-Y)^2$	$(y_{\bullet\bullet}-Y)^2$
columns	$\sigma^2 + r\sigma_\eta^2$	σ_η^2
rows	$\sigma^2 - c\sigma_\phi^2$	σ_ϕ^2
residue	σ^2	σ_ϕ^2

$$(*) = \frac{1}{rc}(MS\text{ mean})\frac{1}{r^2}\left(\frac{1}{c}-\frac{1}{N_\eta}\right)(MS\text{ columns} - MS\text{ residue}) - \frac{1}{c^2}\left(\frac{1}{r}-\frac{1}{N_\phi}\right)(MS\text{ rows} - MS\text{ residue}) - \left(\frac{1}{rc}-\frac{1}{N}\right)(MS\text{ rows} - MS\text{ residue})$$

*Is most easily found numerically from *AMS* mean, *CMS* columns, *CMS* rows, and *CMS* residue.

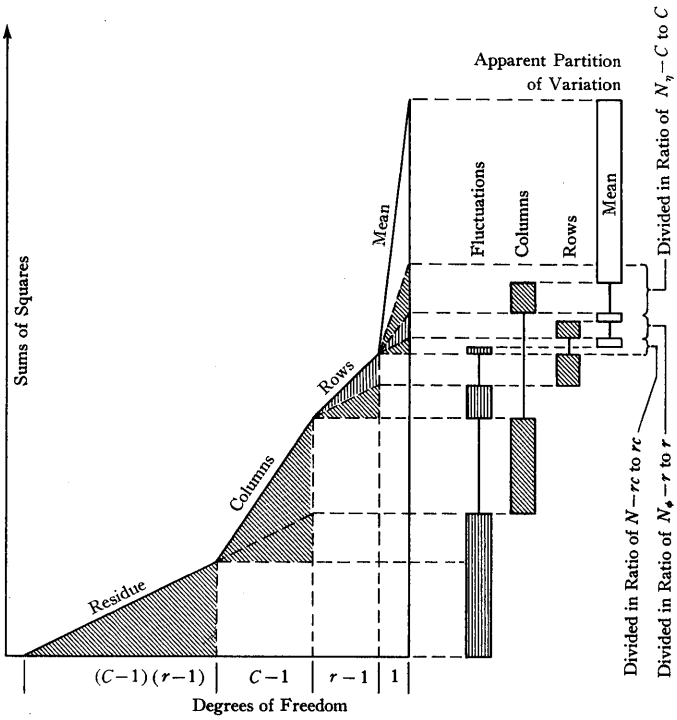


FIGURE 8. ANALYSIS OF VARIANCE DIAGRAM FOR MODEL (7)

REFERENCES

1. R. J. MONROE, "The Applications of Machine Methods to Analysis of Variance and Multiple Regression," pp. 113-116.
2. A. E. BRANDT, "Forms of Analysis for Either Measurement or Enumeration Data Amenable to Machine Methods," pp. 149-153.
3. W. E. DEMING, *Statistical Adjustment of Data* (John Wiley, New York, 1943).
4. R. A. FISHER and F. YATES, *Statistical Tables for Biological, Agricultural and Medical Research* (Oliver and Boyd, Edinburgh; 1st ed. 1938; 2nd ed. 1943).
5. *Ibid.*, 3rd ed. (1948).
6. R. L. ANDERSON and E. E. HOUSEMAN, "Tables of Orthogonal Polynomial Values Extended to $N = 104$," *Research Bulletin 297* (Agr. Expt. Sta., Ames, Iowa, 1942).
7. D. VAN DER REYDEN, "Curve Fitting by the Orthogonal Polynomials of Least Squares," *Onderstepoort Journal of Veterinary Science and Animal Industry, 18* (1943), pp. 355-404.
8. W. E. MILNE, *Numerical Calculus* (Princeton University Press, 1949), Table VI.
9. C. P. WINSOR, "Which Regression," *Biometrics* (Bulletin) 2 (1946), pp. 101-109.
10. J. BERKSON, "Are There Two Regressions?", *Jour. Amer. Stat. Assn. 45* (1950), pp. 164-180.
11. D. V. LINDLEY, "Regression Lines and Linear Functional Relationship," *Supp. Jour. Roy. Stat. Soc. 9* (1947), pp. 218-244.
12. S. S. WILKS, *Mathematical Statistics* (Princeton University Press, 1943, 1946).
13. JOHN W. TUKEY, "Dyadic Anova, an Analysis of Variance for Vectors," *Human Biology, 21* (1949), pp. 65-110.
14. R. A. FISHER, *Statistical Methods for Research Workers* (Oliver and Boyd, Edinburgh, 11th ed.).
15. G. W. SNEDECOR, *Statistical Methods* (Collegiate Press, Ames, Iowa, 4th ed.).
16. L. H. C. TIPPETT, *The Methods of Statistics* (Williams and Norgate, London, 1937, 2nd ed.).
17. A. M. MOOD, *Introduction to the Theory of Statistics* (McGraw Hill, 1950).
18. A. HALD, *Statistics* (John Wiley, New York; in the process of publication).

DISCUSSION

Mr. Keller: In our steam turbine testing work we have found, by running duplicate tests, or tests on two units that are duplicates, that we obtain from one-half to one per cent unexplained variation. It is quite important, in our design work, that we take advantage of differences that may change the performance to the extent of one or two-tenths of a per cent.

In our talks with statisticians they usually point out that, if we use Latin squares and change several variables at once, we could obtain the required design information at smaller cost and with fewer tests. I should like to ask whether the question of how you should plan your experiment is affected by the expected difference that a change in design would cause, relative to the unexplained difference between two tests on the same unit. In other words, the standard deviation is one per cent, and you want to test for three or four items, each of which might amount to two-tenths of a per cent; do the standard methods still apply?

Professor Tukey: Yes. This is entirely typical of agriculture, where these methods were first developed, because that is where they first realized they were in serious trouble. An alteration of one per cent in the yield of barley in Ireland means a considerable number of pounds, shillings, and pence to the Irish. It is awfully hard to get any one field experiment to have a fluctuation as low as 10 per cent, and these methods were developed just to get at that sort of situation. If you are interested in only ten per cent, and your experimental error is one per cent, it doesn't matter how you do it; you will find out. But if you have to work for it, things of this sort are indicated. Whether you want to use Latin squares or not is another matter. You have to know a lot about the situation; and I don't know about steam turbines. So I can't tell you whether you are to use a Latin square or not. But I think that you would find some design, more complicated than the one you are probably using, is likely to help.

Mr. Keast: In the example that you have shown, where your function y_{ij} is affected by i and j, and the diagram underneath, which shows greater variation at the right-hand side than at the left-hand side—what is the essential point of the diagram?

Professor Tukey: The essential point which I tried to make in this diagram, is that there is greater variation at higher levels, rather than that it happens to be at the side. As the level of the original y increased, the differences became larger. If that is the case, you have some hope of controlling it by going to the square root of y or the logarithm of y.

Mr. Keast: My problem is, if you don't know the variation in the first place, is it not that you are considering the variation with each factor to be linear over the range with which you are working? That is, in planning an experiment where you had everything vary at once, are you assuming a linearity there?

Professor Tukey: No, very definitely not, because the remark I made would apply perfectly well if the situation went as follows: in the regions where y is low, the differences are small; and in the region where y is large the differences are large. You are talking about a plus sign in the way that different things interact; but you let η_i be any function of i it chooses, and you have allowed ϕ_j to be any function of j it chooses. The problem is to make the interaction behave, and you can let the individual variables run as they choose to make things go.

Dr. Lotkin: I would like to ask two questions pertaining to some of the work we are doing at the moment, dealing with measurements of angles such as you come across when you contend with theodolite data.

In smoothing such data we have a choice of selecting successive groups of data. The question arises: how large should you take such groups in order to obtain feasible fits? Because we have found that, depending on the size of the groups you take, you get slight variations in the fit.

Second, in doing this smoothing by means of orthogonal polynomials, the degree of the polynomial will vary on your significant answer. In planning this for the machine, we have a choice, then, of either varying your degree of the polynomial—which can become quite involved—or adhering to a certain fixed prescribed degree.

Now, we are aware of the fact that, if we take a fixed degree for this polynomial, we might run into some danger of over-smoothing the data. What I would like to know is if this danger is not, possibly, over-emphasized.

Professor Tukey: I don't want to try to answer this question in detail, due to time limitations, but I can say some things about it. Basically, you have a problem where you are getting data out of a process. You have some theodolites, and you hope they run about the same from day to day; and what you are going to do with this ought to depend on a whole backlog of experience, and not what you obtained on this particular run, generally speaking, unless you have evidence that this run is out of line in some way.

What is needed here, then, is to find out the essential characteristics of the situation, and make up your mind what smoothing you want to do on the basis of that—not just to apply some sort of test to this strip of data and work with it accordingly.

You are raising the question, really, of how should this kind of data be analyzed. How should one analyze the data on soybeans compared to data on potatoes? That requires going back and looking at the essentials of the data.

I think that trying to get at the power spectrum is the way to find out what you want to do in this case.

The Applications of Machine Methods to Analysis of Variance and Multiple Regression

ROBERT J. MONROE

Institute of Statistics, North Carolina State College

THE generally-recognized machine methods which have been adapted to statistical calculations were first outlined by A. E. Brandt[1] in 1935, although some of the methods are known to have been in use before then. Dr. Brandt described the use of progressive digiting methods to obtain sums of squares and sums of products which were required in the statistical analyses. Since that time little has been added to the methods, save some improvements in details as a result of the steadily increasing efficiency of the newer models of machines. The following is essentially a description of the applications of the progressive digiting methods.

The methods of analysis of variance and multiple regression are a part of what have been called "standard methods of analyzing data." The two methods are closely related mathematically, i.e., the analysis of variance can be regarded as a special case of multiple regression in which the independent variables are arbitrary or dummy variates. It is usual, however, to think of the analysis of variance as concerning itself with a single variable, while the purpose of multiple regression is to relate one or more dependent variables with two or more independent (or "causative") variables. In either case it is conceptually easy to regard either problem simply as a breakdown of the "total variation" in a single variable into several component parts, regardless of how this breakdown is accomplished.

The example chosen for this paper came from a problem where both of the above-mentioned techniques were found useful.

In a continuing project at the North Carolina Agricultural Experiment Station the attempts to improve on the present varieties of peanuts involve experiments embracing large numbers of measurements on individual plants. Consider, for example, an experiment made up of four different crosses from which were selected seven different strains. From each of the 28 seed stocks plantings were made to allow the measurement of ten plants, and each seed stock was planted in five different replications (locations). This kind of an experimental design is called, in statistical parlance, a "randomized block."

The model for this design may be written as

$$y_{ijk} = \mu + \rho_i + \gamma_j + \xi_{jk} + (\rho\gamma)_{ij} + (\rho\xi)_{ijk} + \epsilon_{ijkl}.$$

μ = unspecified mean parameter

ρ_i = effect of ith replication ($i = 1, \ldots, 5$)

γ_j = effect of jth cross ($j = 1, \ldots, 4$)

ξ_{jk} = effect of kth strain in the jth cross
$k = (1, \ldots, 7)$ for each j

$(\rho\gamma)_{ij}$ = effect of interaction of jth cross with ith replication

$(\rho\xi)_{ijk}$ = effect of interaction of kth strain with ith replication for each of j crosses

ϵ_{ijkl} = a random error NID$(0, \sigma^2)$ associated with each plant ($l = 1, \ldots, 10$).

The analysis of variance, with associated degrees of freedom, is derived from the model, each line in the analysis being associated with the indicated parameters of the model.

ANALYSIS OF VARIANCE

Source of Variation	*Degrees of Freedom*	
general mean	1	C.F.
replications	4	SS(R)
crosses	3	SS(C)
replications × crosses	12	SS(RC)
strains within crosses	24	SS(SC)
replications × strain within crosses	96	SS(RSC)
individual plants within plots	1260	SS(IP)
Total	1400	SS(T)

The sums of squares for each of the above effects may then be segregated.

1. General mean: $(1/ijkl)\left(\sum_{ijkl} y\right)^2 = \text{C.F.}$

2. Replications: $\left(\frac{1}{jkl}\right)\sum_i\left(\sum_{jkl} y\right)^2 - \text{C.F.} = \text{SS(R)}$

3. Crosses: $\left(\frac{1}{ikl}\right)\sum_{j}\left(\sum_{ikl}y\right)^2 - \text{C.F.} = \text{SS(C)}$

4. R × C $\left(\frac{1}{kl}\right)\sum_{ij}\left(\sum_{kl}y\right)^2 - \text{C.F.} - \text{SS(R)} - \text{SS(C)} = \text{SS(RC)}$

5. Strains in crosses:
$$\frac{1}{il}\sum_{jk}\left(\sum_{il}y\right)^2 - \text{C.F.} - \text{SS(C)} = \text{SS(SC)}$$

6. Replication by strains in crosses:
$$\frac{1}{l}\sum_{ijk}\left(\sum_{l}y\right)^2 - \frac{1}{kl}\sum_{ij}\left(\sum_{kl}y\right)^2 - \text{SS(SC)} = \text{SS(RSC)}$$

7. Individual plants within plots:
$$\sum_{ijkl}y^2 - \frac{1}{l}\sum_{ijk}\left(\sum_{l}y\right)^2 = \text{SS(IP)}$$

8. Total: $\sum_{ijkl}y^2 = \text{SS(T)}$

[NOTE: The abbreviated notation used here avoids the bulky multiple summations, and the term C.F. refers to the "correction factor" for the general mean. The term SS(R) is read "sum of squares for replications," etc.]

The card used for the analysis is indicated in Figure 1. Each classification is punched in a separate field, as is each variable to be considered in either the analysis of variance or the regression analysis.

The analysis is produced by successive sorting, summing, and progressive digiting operations, interspersed with occasional use of a desk calculating machine. Detailed steps follow:

1. Sort on "strains," "crosses," and "replications."
2. Tabulate and summary punch with controls on the three sorts.
3. Compute $\sum_{ijk}\left(\sum_{l}y\right)^2$ by progressive digiting with the summary deck.
4. Remove the controls on "strains" from original deck and tabulate again.
5. From the tabulation in (4) compute $\sum_{ij}\left(\sum_{kl}y\right)^2$ with desk calculator.
6. From the same tabulation in (4) the totals
$$\sum_{j}\left(\sum_{ikl}y\right) \text{ and } \sum_{i}\left(\sum_{jkl}y\right)$$
are obtained with which to compute SS(C) and SS(R), respectively.
7. The correction factor C.F. is computed from a final total of the tabulation (4).
8. The total sum of squares, SS(T), is then computed with the original deck, using the progressive digiting technique as before.
9. Each line of the analysis of variance is then derived from the above quantities by applying the proper divisors and making the subtractions indicated in the formulas already given.

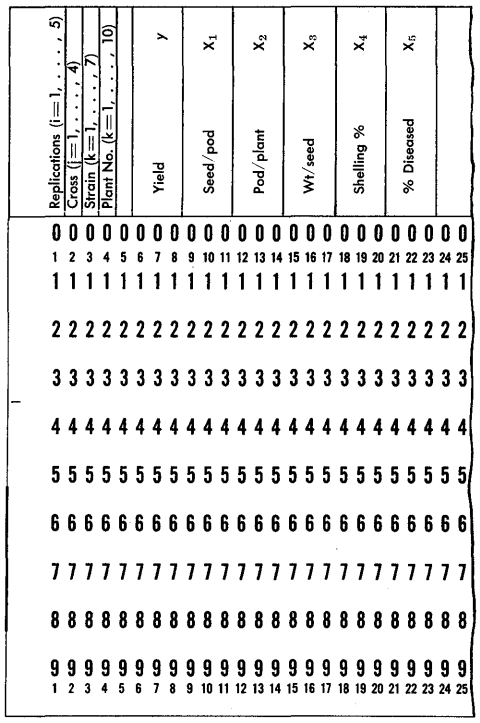


FIGURE 1

While the above procedure may not be the shortest possible method for any one instance, it seems clear that such a process is admirably adapted to routine computing and requires a minimum of supervision.

THE MULTIPLE REGRESSION ANALYSIS

In the same study it was necessary to compare the inherent (genetic) relationships between several of the plant characteristics with the observed (phenotypic) correlations. For these purposes are required the sums of squares and sums of products of all variables under consideration.

The phenotypic correlations are defined, simply, as the product-moment correlations between the variables concerned. That is, if x_1 and x_2 are the variables, then the phenotypic correlation r_p is given as:

$$r_p = \frac{\sum(x_1-\bar{x}_1)(x_2-\bar{x}_2)}{\sqrt{\sum(x_1-\bar{x}_1)^2\sum(x_2-\bar{x}_2)^2}}.$$

The same quantities appear in a multiple regression analysis of the regression of a dependent variable, y_i, on several, p, independent variables x_j. Suppose the regression of yield on the variables x_1 = seed/pod, x_2 = pods/plant, and x_3 = wt./seed were desired. The model, $p = 3$,

$$(y_i-\bar{y}) = b_1(x_{1i}-\bar{x}_1) + b_2(x_{2i}-\bar{x}_2) + b_3(x_{3i}-\bar{x}_3) + e_i[i=1,\ldots,n]$$

represents the observed data where e_i is an estimate of the measurement errors involved. In ordinary least squares theory, the solution for the coefficients b_j proceeds as follows:

In matrix notation if

$$a_{jk} = \sum_{i=1}^{n} (x_{ji} - \bar{x}_j)(x_{ki} - \bar{x}_k), \qquad (j,k = 1,2,3)$$

and

$$g_k = \sum_{i=1}^{n} (x_{ki} - \bar{x}_k)(y_i - \bar{y}),$$

then

$$b_j = [c_{jk}]\ [g_k] \text{ where } [c_{jk}] = [a_{jk}]^{-1}.$$

$[c_{jk}]$ is also called the covariance matrix because it contains the coefficients for computing the variances and covariances of the b_j's, e.g., $V(b_j) = c_{jj}s^2$. Then the reduction due to regression is given by $\sum b_j g_j$, and the residual sum of squares to estimate the measurement error is

$$\sum (y_i - \bar{y})^2 - \sum b_j g_j = (n - p - 1)s^2,$$

where $(n - p - 1)$ represent the degrees of freedom in s^2.

It is clear, then, that both types of analyses require the computation of a number of quantities of the form

$$\sum (x_{ji} - \bar{x}_j)(x_{ki} - \bar{x}_k),$$

and these are easily obtained from a progressive digiting scheme using only a sorter and accounting machine, although use of a summary punch will shorten the time of calculating somewhat.

In usual practice it will be desirable to set up the entire group of variables in the counters and successively multiply the entire group by each variable in the group. This procedure gives a check on the machine work, since the $\sum x_j x_k = \sum x_k x_j$, the two operations being accomplished independently.

The matrix inversion required in the regression analysis is a more difficult thing to accomplish without special equipment. It is sufficient here to mention that this can be done with the IBM Type 602 Calculating Punch, but the process is quite involved. This job is, perhaps, one best done on the newer models which have been demonstrated at this seminar.

The foregoing discussion was intended as a brief summary of the application of IBM equipment to statistical analysis. Only a few of the basic operations were described. To anyone familiar with both the analytical and computational problems many short cuts and improvements will suggest themselves—especially in particular problems.[2]

REFERENCES

1. A. E. BRANDT, *The Punched Card Method in Colleges and Universities* (Edited by G. W. Baehne, pp. 423-436, Columbia University Press, 1935).
2. P. G. HOMEYER, MARY A. CLEMM, and W. T. FEDERER, *Research Bulletin 347* (April, 1947), Iowa State College Agricultural Experiment Station.

DISCUSSION

Mr. Dye: There are several other ways, in which you might be interested, by which you can derive the same results that have just been described, by use of standard IBM equipment. I think they might be of passing interest, if nothing else. You can, with small volumes, on a 604, actually square the individual components and obtain the cross-products of the various variables. Then, by running the cards through an accounting machine, add them and obtain the summations in printed and punched card form. By running the cards through the 604 again, actually the means are derived, the standard deviation squared, and in some cases the actual r squared.

As far as progressive digiting is concerned, there are, also, two methods which are available, depending upon what kind of equipment is present. If the machine contains the card cycle total transfer, you can do up to 40 digits by sorting on the complete item, transferring at the end of each control break enough items so that, when you are finished, you will have on your minor breaks, a summation of the terms themselves, and on the major, a summation of their squares.

Chairman Hurd: I might make a note here of something that Mr. Bailey told me. He indicated that he had prepared a control panel for the 604 which would handle all analysis of variance procedures up to a reasonable three classifications, with one pass of the machine.

Mr. Bailey: The technique that we generally use is to prepare one card for each item in the table. We prepare one card for each of those ten items through all the replications. Ordinarily, we subtract a base from the items and code them to reduce the number to a reasonable size. We have it set up so that, on one card, we can handle numbers with as many as four or five digits, subtracting a base from each of those numbers. In one pass through the 604, the base is subtracted from each of the five-digit numbers, then the differences and the squares are summed and punched in the detail card. We sort the detail cards on our first classification, summary punch, repeat the process for the second classification and the third classification, then sort on two classifications at a time, summary punch, and finally summary punch with all controls off. Next, the summary cards and the detail cards are put together and fed into the 604, using the same control panel, with a switch where each of the sums is divided by the number of items represented on the summary card and obtain the means. Multiply the mean by the sum again to obtain the correction factors.

All these correction factors and sums of squares are listed both for summary cards and detail cards, and it is a simple matter, just by a process of subtraction, to obtain all the sums of squares in the analysis of variance table.

Mr. Clark: There is a method of obtaining sums of squares and cross products, called digiting without sorting, in which the multiples are represented as combinations of 1, 3, 5, the same way that you would a binary code 1, 2, 4, 8. In other words, 9 is 5 + 3 + 1, 8 is 5 + 3, and so on. If there is enough capacity you can, at the same time, digit for 10, 30, 50, and then 100, 300, 500. Really, there are four methods: the multiplication, the summary punching, the card total transfers, and the digiting without sorting. And one should always weigh the economics of which way to do it. However, you can make a general statement to the effect that, if you have a large number of variables and a large amount of data per dollar of computing, your best method is summary punching with the method that you outlined.

Mr. Belzer: The first method you described requires three counters for each bank?

Mr. Clark: That is right. It is an extravagant method. The beauty of it is that when you have an enormous number of cards, a small number of variables, you don't want to sort.

Examples of Enumeration Statistics

W. WAYNE COULTER

International Chiropractors Association

MATHEMATICIANS are working with numerical constants. Once a specific problem is solved, the same formula will apply to similar problems. This is not true when multiple variables are introduced from problem to problem. We, in the healing arts, deal with such a great number of variables within the human body, as well as from individual to individual, that it has not been possible to apply mathematics to the human race as a mass; trends, indications or approximations are the best we can hope to obtain.

Our IBM installation consists of a Type 16 Motor Drive Duplicating Punch. The problems we have encountered have been merely to standardize methods and procedures of collecting data, proper coding and punching, so that the IBM Service Bureau can compute averages or percentages on several pertinent items.

Our field research program is now in its third year and is a cumulative study. With each passing year, the study will become more useful as the number of cases in each diagnosis increases. The field research data booklet contains the information concerning the cases compiled. The first year, 1947-1948, this program was in operation we studied 700 cases with 16 diagnosed conditions. By longhand methods it took us 800 man hours to calculate our data. During the year 1948-1949, we processed 3,400 cases on 38 diagnosed conditions in 400 man hours by IBM methods. This is exclusive of the two hours required by the Service Bureau to tabulate the data. What this amounts to, roughly, is 4½ times the work load in one-half the time required by longhand methods.

Since this program was started previous to switch-over to IBM cards, and since it was a cumulative study, it became necessary to have our codes made up into rubber-stamp form in order to bring our previous cases up to date.

The case history of each patient studied in the research program is recorded on a form such as Figure 1, page 118. This wealth of information may be placed on one IBM card, as indicated in the right-hand side of the figure. Each case is coded as to:

1. *Industry.* There are 13 classifications which indicate the field of work in which the patient is engaged. Thus, percentages in each different type of industry may be determined. (At a later date if we should require data on specific types of industry, special studies may be conducted.)
2. *Occupation.* The type of activity the patient pursues is indicated by 11 categories. The occupation code enables the determination of percentagewise distribution.
3. *Injury.* Ten classifications indicate the nature of the injury, while 16 other classifications give the way in which the injury was incurred.
4. *Chiropractic Analysis.* The analysis of the patient's condition after spinal analysis is coded into one of 16 categories.
5. *Diagnosis.* The coding of diagnosis of the patient's condition consists of merely assigning 1 to anemia, 2 to angina pectoris, 3 to arthritis, 4 to asthma, etc.
6. *Patient's Condition.* This is coded for before chiropractic care, with chiropractic care, and after chiropractic care.
7. *Insurance.* Information as to whether claims were paid, the compensation involved, and the type of insurance policy.

Other necessary information for research analysis such as case number, age, sex, days under medical care, number of office visits, number of X-rays, and number of days (working) lost before chiropractic care and while under it is in actual numbers.

The Service Bureau sorts like numbers together according to diagnosis, and from there on it is a simple matter of tabulating like data in the same columns, with totals, so that we can obtain various data and averages or percentages such as:

1. Average age.
2. % females—% males.
3. Average number of days under chiropractic care.
4. Patients' condition at end of chiropractic care.
5. % well.
   % much improved.
   % slightly improved.
   % same.
   % worse.
6. Average number of years the diagnosed condition had existed previous to chiropractic service.

INTERNATIONAL CHIROPRACTORS ASSOCIATION

INDUSTRIAL RESEARCH (Form IR4) Revised

1. Name or Case No. ..... Date ...../...../ 19.....
2. Employer .....
3. Age..... Sex..... Occupation .....
4. Nature of Injury .....
5. How Was Injury Incurred .....

.....

6. Type of Work Being Performed When Injured .....
7. Date of Injury . . . ./...../19..... Time of Day .....
8. Date Reporting to Chiropractor ...../...../19..... Time of Day .....
9. Date of Discharge . . . ./...../19.....
10. Number of Days Under Chiropractic Care ..... Number of Office Visits .....
11. Patient's Condition After Chiropractic Service .....

    (Well, improved, temporary partial disability, temporary total disability, permanent partial disability, permanent total disability, or other.)

12. Chiropractic Analysis .....
13. Number of X-rays for Chiropractic Analysis .....
14. Number of Days Lost From Work on This Case:
    (a) Before Chiropractic Care .....
    (b) Under Chiropractic Care .....
15. Number of Days Under Medical Care, If Any, In Connection With This Injury:
    (a) Before Chiropractic .....
    (b) With Chiropractic .....
    (c) After Chiropractic .....
16. X-ray Costs . . . . . $.....
17. Service or Adjustment Costs . . . $.....

    Total Costs $.....

18. Name of Insurance Company .....
19. Was Claim Paid In Full .....
20. If Additional Information Given on Back of This Sheet, Check This Space .....

☐ Industrial Case
Claim Paid by Insurance Company ☐
Claim Denied by Insurance Company ☐

☐ Industrial Case
Claim Not Presented for Legal Reasons

☐ Health & Accident Policy—Claim Paid ☐
(Not Industrial) Claim Denied ☐

WHEN COMPLETED MAIL TO:
INDUSTRIAL RESEARCH,
ICA, 838 Brady Street,
Davenport, Iowa

Chiropractor .....

Address ..... City ..... State .....

PLEASE CHECK ALL INFORMATION FOR ACCURACY

DO NOT WRITE IN THIS SPACE

1	
2	
3	
4	
5	Case No.
6	
7	
8	
9	
10	Indust.
11	
12	Age
13	Sex
14	
15	
16	
17	
18	Occupation
19	
20	
21	
22	
23	Injury
24	
25	Inj. Inc.
26	
27	Days
28	Under
29	Chiro.
30	
31	Off. Visits
32	
33	Pat. Cond.
34	
35	Chiro. Anal.
36	No. X-Rays
37	Before
38	Chiro. DAYS LOST
39	
40	Under
41	Chiro.
42	
43	Before
44	Chiro.
45	Care
46	With
47	Chiro.
48	Care
49	After
50	Chiro.
51	Care
52	DAYS UNDER MEDICAL CARE
53	
54	
55	X-Ray Cost
56	
57	
58	
59	Adj. Cost
60	
61	
62	
63	
64	Total Cost
65	
66	
67	
68	Insur. Code
69	Ins. Cl. Pd.
70	Ins. Cl. Denied
71	Legal Reas.
72	H & A Pd.
73	H & A Denied
74	State
75	

FIGURE 1

*Transforming Theodolite Data**

HENRY SCHUTZBERGER

Sandia Corporation

A THEODOLITE is a precise instrument for measuring the azimuth and elevation angles of an object. It is desired to reduce these angular measurements, arising from two or more theodolites, to the rectangular Cartesian coordinates of the object. The quantity of computation for several hundred observations from more than two theodolites becomes overwhelming when these calculations are performed on hand calculating machines. However, the computation for these same observations becomes quite feasible with automatic computing machines. The method discussed here has been used with as many as five theodolites with much more satisfactory results than previously used two-theodolite procedures.

The instruments most generally used to obtain the observations are the Askania theodolites. They have proved to be the most valuable of the ciné-theodolites available. Captured from the Germans after the war, they are used extensively on test ranges in this country. Attempts have been made to duplicate them, but with little success up to the present time.

Theodolite Instrumental Errors

The Askania ciné-theodolite, like any other high-precision instrument, is subject to many errors. Frequently these errors arise, not from any inherent defects in the instrument, but from the fact that the instrument can be read more accurately than the adjustments of the instruments can practically be made.

The errors to which the Askania is subject are:

1. *Tracking Errors.* These are not really errors in the ordinary sense but arise from the fact that the operators are not able to keep a moving object exactly on the center of each frame of the film. Thus, it is necessary to correct for this displacement on each frame.
2. *Orientation Error.* This error occurs because the instrument is not oriented to read the proper predetermined elevation and azimuth angles when set on a given target.
3. *Leveling Error.* This error occurs when the base plate (azimuth circle) of the instrument is not exactly level. The base plate error consists of two parts: the angle of inclination of the base plate with the true horizontal plane, and the azimuth direction of the intersection of the base plate and this horizontal plane.
4. *Collimation Error.* This error occurs when the line of sight down the instrument telescope is not exactly perpendicular to the horizontal axis of the instrument.
5. *Standard Error.* This error occurs when the horizontal axis of the instrument does not lie exactly parallel to the base plate of the instrument.
6. *Tilt Correction.* Because the local zenith at the instrument is not parallel to the zenith at a removed origin, owing to the curvature of the earth, this correction must be applied.
7. *Refraction Error.* This error is due to the bending of light rays when passing through media of changing density.
8. *Scale Error.* The Askania ciné-theodolite has an extremely precise scale, but the optical system of several prisms, transmitting the scale to the film, may be out of adjustment and so introduce an error in the scale reading.
9. *Bearing Error.* As the instrument is rotated through the azimuth, its weight is supported by a main thrust bearing. Any irregularities in this bearing, or in the ways in which it rises, introduces an error in the elevation angle.

For the accuracy of measurements desired, each of these corrections must be taken into account. At present these corrections are made by hand computations, as it was not considered efficient to perform them on mechanical equipment. However, a device built by the Telecomputing Corporation, known as an Askania Reader, has been ordered. This machine, which is connected to an IBM Type 517 Summary Punch, enables an operator to make the necessary measurements on the film, and records these measurements and the instrumental constants automatically on an IBM card. With these data on cards, it will be possible on the IBM Card-Programmed Electronic Calculator to make all necessary corrections.

A Solution Used in the Past to the Two-theodolite Problem

Let S = azimuth angle measured from the positive X direction

Let H = elevation angle

$0 - XYZ$ = right-handed reference frame in which Z is vertical

X, Y, Z = space coordinates of object

x, y, z = space coordinates of observation point

Subscripts 1 and 2 = quantities pertaining to theodolite 1 and theodolite 2, respectively.

*This paper was presented by Kenneth C. Rich.

The usual relations yielding X, Y, and Z from a pair of theodolite observations are set down for reference:

$$X = x_1 + \frac{(x_2-x_1)\tan S_2 - (y_2-y_1)}{\tan S_2 - \tan S_1} \quad (1)$$

$$Y = y_1 + (X-x_1)\tan S_1 \quad (2)$$

$$Z = z_1 + (Y-y_1)\frac{\tan H_1}{\sin S_1} = Z_1 + (X-x_1)\frac{\tan H_1}{\cos S_1} \quad (3)$$

or

$$Z = z_2 + (Y-y_2)\frac{\tan H_2}{\sin S_2} = Z_2 + (X-x_2)\frac{\tan H_2}{\cos S_2}. \quad (4)$$

The order in which the preceding relations are given is convenient for computations. Under certain conditions, it is necessary to change the form of these relations,[a] but the significance of the method remains unchanged. Relations (3) and (4) give the same value of Z only when the lines of sight make a true intersection.

It will be noted that a redundancy exists, in that four quantities (S_1, H_1, S_2, H_2) are given from which three quantities (X, Y, Z) are to be determined. Except when the lines of sight make a true intersection, this problem has no proper solution.

Derivation of a Least-squares Method of Data Reduction

The system and angles are defined in exactly the same manner as before. The direction cosines of the line of sight of each theodolite may be determined from the H and S angles and are:

$$l_i = \cos H_i \cos S_i$$
$$m_i = \cos H_i \sin S_i$$
$$n_i = \sin H_i$$

where the subscript i denotes the number of each theodolite. For convenience in notation, the space coordinates of the ith theodolite shall be denoted as (x_i, y_i, z_i).

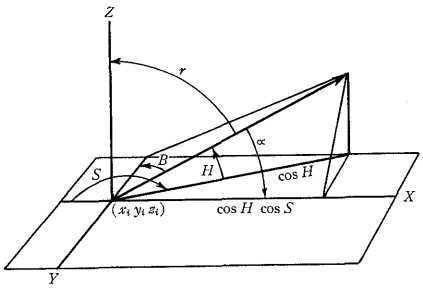
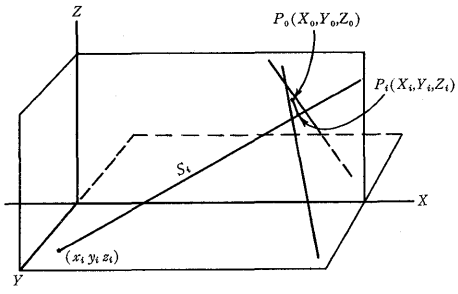


FIGURE 1. DIAGRAM ILLUSTRATING GEOMETRY FOR DERIVATION OF DIRECTION COSINES

[a]Should S_1 approach 90 degrees, i.e., X—x_1 be very small, it is better to compute $Y-y_1$ from a relation similar to Equation 1 but involving cotangents of the angles, and then to compute $X-x_1$ from a relation similar to Equation 2.



The coordinates of any point lying on the ith line of sight may be expressed as:

$$X_i = x_i + l_i s_i$$
$$Y_i = y_i + m_i s_i$$
$$Z_i = z_i + n_i s_i$$

where s_i is the distance along the line of sight from this point to the theodolite at (x_i, y_i, z_i).

FIGURE 2. DIAGRAM SHOWING LINES OF SIGHT AND LOCATION OF DESIRED POINT

Now, if the several theodolites are pointing at a fast moving object in space at a considerable distance, the lines of sight in general will be skew with respect to each other. Let the coordinates, to be determined, of the object be denoted by $P_0(X_0, Y_0, Z_0)$. From the point P_0, consider the construction of lines perpendicular to each line of sight. Denote the intersection of each line of sight with its perpendicular by $P_i(X_i, Y_i, Z_i)$. By this notation, then, the distance from the point P_0 to each line of sight may be expressed by

$$d_i^2 = (X_0-x_i-l_i s_i)^2 + (Y_0-y_i-m_i s_i)^2 + (Z_0-z_i-n_i s_i)^2 \quad (5)$$

For the determination of the point P_0 which best fits the lines of sight, a least-squares approach is believed to give the closest representation of the actual condition; i.e., the sum of the squares of the distances from this point to each of the lines of sights is to be minimized. Thus, since in equation (5) the values of (x_i, y_i, z_i), the ith theodolite coordinates, and l_i, m_i, n_i, the direction cosines of the line of sight from the ith theodolite, are determined by the position of the theodolite and the direction in which it is pointing; the values of (X_0, Y_0, Z_0) and s_i are the variables which may be changed to minimize the sum of the squares of the distances from point P_0 to the lines of sight. Thus,

$$d_i^2 = f(X_0, Y_0, Z_0, s_i).$$

A set of direction numbers for the ith line of sight is l_i, m_i, n_i, and a set of direction numbers of the line joining the point in space (X_0, Y_0, Z_0) to any point (X_i, Y_i, Z_i) on the ith line of sight is X_0-X_i, Y_0-Y_i, Z_0-Z_i, or $X_0-x_i-l_i s_i$, $Y_0-y_i-m_i s_i$, $Z_0-z_i-n_i s_i$.

A necessary and sufficient condition that these lines be perpendicular to each other is that the sum of the products of corresponding direction numbers be zero; i.e.,

$$l_i(X_0-x_i-l_i s_i) + m_i(Y_0-y_i-m_i s_i) + n_i(Z_0-z_i-n_i s_i) = 0 \quad (6)$$

Solving for s_i

$$s_i = l_i(X_0-x_i) + m_i(Y_0-y_i) + n_i(Z_0-z_i). \quad (7)$$

Thus, the parameter s_i may be eliminated by making use of the condition of perpendicularity.

Substituting this value of s_i in (5), letting $l_ix_i + m_iy_i + n_iz_i = p_i$, and simplifying

$$d_i^2 = [(1-l_i^2)\ X_0 - l_im_iY_0 - l_in_iZ_0 - x_i + l_ip_i]^2 \quad (8)$$
$$+ [-l_im_iX_0 + (1-m_i^2)\ Y_0 - m_in_iZ_0 - y_i + m_ip_i]^2$$
$$+ [-l_in_iX_0 - m_in_iY_0 + (1-n_i^2)\ Z_0 - z_i + n_ip_i]^2 .$$

In order to minimize the sum of the squares of the distances to each line of sight, i.e.,

$$\sum_{i=1}^{n} d_i^2 = F\ (X_0, Y_0, Z_0) = \text{min.} , \quad (9)$$

the following condition is necessary:

$$\frac{\partial F}{\partial X_0} = \frac{\partial F}{\partial Y_0} = \frac{\partial F}{\partial Z_0} = 0 . \quad (10)$$

Summing d_i for all theodolites

$$\sum_{i=1}^{n} d = F(X_0, Y_0, Z_0) = \quad (11)$$

$$\sum_{i=1}^{n} \left\{ \left[(1-l_i^2)\ X_0 - l_im_iY_0 - l_in_iZ_0 - x_i + l_ip_i \right]^2 \right.$$
$$+ \left[- l_im_iX_0 + (1-m_i^2)\ Y_0 - m_in_iZ_0 - y_i + m_ip_i \right]^2$$
$$\left. + \left[- l_in_iX_0 - m_in_iY_0 + (1-n_i^2)\ Z_0 - z_i + n_ip_i \right]^2 \right\}$$

Taking a partial derivative with respect to X_0 and simplifying

$$\frac{\partial F}{\partial X_0} = \sum_{i=1}^{n}(1-l_i^2)\ X_0 + \sum_{i=1}^{n}(-l_im_i)\ Y_0 + \sum_{i=1}^{n}(-l_in_i)\ Z_0$$
$$- \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} l_ip_i = 0 . \quad (12)$$

Similarly:

$$\frac{\partial F}{\partial Y_0} = \sum_{i=1}^{n}(-l_im_i)\ X_0 + \sum_{i=1}^{n}(1-m_i^2)\ Y_0 + \sum_{i=1}^{n}(-m_in_i)\ Z_0$$
$$- \sum_{i=1}^{n} y_i + \sum_{i=1}^{n} m_ip_i = 0 . \quad (13)$$

$$\frac{\partial F}{\partial Z_0} = \sum_{i=1}^{n}(-l_in_i)\ X_0 + \sum_{i=1}^{n}(-m_in_i)\ Y_0 + \sum_{i=1}^{n}(1-n_i^2)\ Z_0$$
$$- \sum_{i=1}^{n} z_i + \sum_{i=1}^{n} n_ip_i = 0 . \quad (14)$$

Rewriting (12), (13), and (14)

$$\sum_{i=1}^{n}(1-l_i^2)\ X_0 + \sum_{i=1}^{n}(-l_im_i)\ Y_0 + \sum_{i=1}^{n}(-l_in_i)\ Z_0$$
$$= \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} l_ip_i \quad (15)$$

$$\sum_{i=1}^{n}(-l_im_i)\ X_0 + \sum_{i=1}^{n}(1-m_i^2)\ Y_0 + \sum_{i=1}^{n}(-m_in_i)\ Z_0$$
$$= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} m_ip_i$$

$$\sum_{i=1}^{n}(-l_in_i)\ X_0 + \sum_{i=1}^{n}(-m_in_i)\ Y_0 + \sum_{i=1}^{n}(1-n_i^2)\ Z_0$$
$$= \sum_{i=1}^{n} z_i - \sum_{i=1}^{n} n_ip_i$$

Thus, an array of three symmetric simultaneous linear equations in the variables X_0, Y_0, Z_0 is obtained. All other values are obtained from theodolite data.

An Outline of the Abbreviated Doolittle Method of Solution of a System of Symmetric Linear Equations

There are many methods for solving systems of linear equations such as those given in equation (15), and it was felt, after several methods were investigated—inasmuch as automatic calculating equipment would be available—that the abbreviated Doolittle method was the most economical in time for the desired accuracy. Dwyer's method, known as the Abbreviated Method of Multiplication and Subtraction-Symmetric or, as he calls it, the Compact Method, is somewhat shorter than the abbreviated Doolittle method, but it involves more difficult 602-A control panel wiring.[1]

This method is applicable to any number of theodolites, greater than one, with no changes. Because, occasionally, tracking operators lose the object, this is a very necessary condition.

Approximately 45,000 arithmetical steps, depending upon the number of observations, are involved in the data reduction for one test.

CHART I

OUTLINE OF ABBREVIATED DOOLITTLE METHOD OF SOLUTION

$$l = \cos H \cos S$$
$$m = \cos H \sin S$$
$$n = \sin H$$
$$p = lx + my + nz$$

	A	B	C	D
1	$(1A) = \Sigma(1-l^2)$*	$(1B) = -\Sigma lm$	$(1C) = -\Sigma ln$	$(1D) = \Sigma x - \Sigma lp$
2	same as (1B)	$(2B) = \Sigma(1-m^2)$*	$(2C) = -\Sigma mn$	$(2D) = \Sigma y - \Sigma mp$
3	same as (1C)	same as (2C)	$(3C) = \Sigma(1-n^2)$*	$(3D) = \Sigma z - \Sigma np$
4	(4A) = (1A)	(4B) = (1B)	(4C) = (1C)	(4D) = (1D)
5	(5A) = 1	$(5B) = \frac{(1B)}{(1A)}$	$(5C) = \frac{(1C)}{(1A)}$	$(5D) = \frac{(1D)}{(1A)}$
6		(6B) = (2B) − (5B)(1B)*	(6C) = (2C) − (5C)(1B)	(6D) = (2D) − (5D)(1B)
7		(7B) = 1	$(7C) = \frac{(6C)}{(6B)}$	$(7D) = \frac{(6D)}{(6B)}$
8			(8C) = (3C) − (1C)(5C) − (6C)(7C)*	(8D) = (3D) − (1C)(5D) − (6C)(7D)
9			(9C) = 1	$(9D) = \frac{(8D)}{(8C)}$
10	$Z = (9D)$ $Y = (7D) - (7C)Z$ $X = (5D) - (5C)Z - (5B)Y$			

*ALWAYS POSITIVE.

The time required to compute by hand a reduction, of all the data obtained from a complete test, by the abbreviated Doolittle method is prohibitive. It is roughly estimated that two experienced persons might be able to compute the coordinates for one complete test in about a month. By use of the IBM equipment now at hand, which includes two type 602-A calculating punches, this process, starting with the film, can be completed in approximately three days. However, by use of the Telecomputing Askania Reader and the IBM card-programmed electronic calculator, now on order, it is estimated that a reduction of one complete test from data of five theodolites may be completed in a matter of hours.

In regard to a two-theodolite solution versus a five-theodolite solution, a study has been made to answer the questions as to which method produces the better solution, and how much better is this solution. In using a comparison of third differences as a measure of the random errors present, it was found that a five-theodolite solution was considerably smoother: in fact, it had only about 50 per cent as much random error as did the two-theodolite solution.

REFERENCE

1. PAUL S. DWYER, "The Solution of Simultaneous Equations," *Psychometrika*, Vol. 6, No. 2 (April, 1941).

DISCUSSION

Mr. Rich: Mr. Schutzberger's problems are very similar to those of the Naval Ordnance Test Station, and I'm sure that similar methods are used by many groups in the country. The concern of groups involved in data reduction is the speed with which results may be obtained after the tests and then placed in the hands of the interested parties.

Professor Tukey: A number of years ago we were trying to do things like this with two theodolites, using Mitchell

CHART II

OUTLINE OF COMPACT METHOD OF SOLUTION

(Abbreviated Method of Multiplication and Subtraction—Symmetric)

$$l = \cos H \cos S$$
$$m = \cos H \sin S$$
$$n = \sin H$$
$$p = lx + my + nz$$

	A	B	C	D
1	1A $\Sigma(1-l^2)$	1B $-\Sigma lm$	1C $-\Sigma ln$	1D $\Sigma x-\Sigma lp$
2		2B $\Sigma(1-m^2)$	2C $-\Sigma mn$	2D $\Sigma y-\Sigma mp$
3			3C $\Sigma(1-n^2)$	3D $\Sigma z-\Sigma np$
4	4A = 1A	4B = 1B	4C = 1C	4D = 1D
5		5B = (1A)(2B) − (1B)^2	5C = (1A)(2C) − (1B)(1C)	5D = (1A)(2D) − (1B)(1D)
6		6C = [(1A)(3C) − (1C)^2] 5B − (5C)^2		6D = [(1A)(3D) − (1C)(1D)] 5B − (5D)(5C)
7	$Z = \frac{6D}{6C}$ $Y = \frac{1}{5B}[5D-(5C)(Z)]$ $X = \frac{1}{1A}[1D-(1C)(Z)-(1B)(Y)]$			

cameras. With the Mitchell, if your film is read on a Recordak with about three special curves on it, the refraction corrections could not be entered, but most of the instrumental corrections could be entered. I have never handled an Askania, but I would think there might be some possibility of this.

In some circumstances might not it be desirable to have a computing procedure that would reject the worst of the five theodolites on each point and take the least squares solution of the other four, or reject the worst of three and keep the other two?

Mr. Rich: We have the problem at Inyokern of having a three-station reduction system. Essentially, we obtain checks on the accuracy of the three stations used before we even place it into a data reduction scheme.

Dr. Lotkin: At Aberdeen we are doing exactly the same type of problem, among others. As far as the method of reduction is concerned, we have found, by comparison, that it is better to use the method of minimizing the squares of the sides of the triangle involved, rather than the squares of the distances from the lines of sight—one reason being that the normal equations become more simple, and we are able to put in more checks on the computing procedure as we go along. We have been able to mechanize the whole procedure on the IBM relay calculators where we start with the smoothing operation on the angles, then, by means of least squares, compute average position, smooth the position by means of least squares again, and obtain velocities and acceleration. It takes about two minutes per point for this whole process; and trajectories containing as many as 300 points may be reduced within an hour and a half to two hours.

Minimum Volume Calculations with Many Operations on the IBM Type 604 Electronic Calculating Punch

WILLIAM D. BELL

Telecomputing Corporation

PROBLEMS involving a small card volume and many mathematical operations or steps are particularly troublesome to handle on standard IBM machines. Usually the job will not justify spending any considerable amount of time on procedure. If simple machine operations are to be used, the number of control panels and associated wiring time become large. A more elegant procedure may consolidate the number of control panels, but each panel becomes correspondingly more difficult, and the total elapsed time may be increased rather than reduced. When IBM machines are used for general purpose computing, it becomes mandatory that some efficient method of handling this type of problem be devised. This paper describes one such method.

Previous attempts to develop an adequate answer to the problem discussed here did not meet with much success. The advent of the IBM Type 604 Electronic Calculating Punch has changed this picture materially. The most efficient method of solution would be with a card-programmed electronic calculator. The method described is admittedly crude and slow in comparison with results that can be achieved easily with the sequence machine. However, for persons having a 604, but lacking more elaborate machines, the method to be described will make feasible machine solution of many problems which would otherwise be rejected as not amenable to punched card techniques.

Ideally, a procedure utilizing only a single 604 setup and card form was desired, which would be capable of handling any problem regardless of the number of terms, size of amounts and the formulas involved. Within practical limits this has been achieved.

A master deck containing one card for every term, either given or to be computed, is used. This deck is expanded by reproduction for the number of particular solutions desired. Each card contains instructions for the 604. Since a single card (or term) may be used in several calculating steps, more than one set of instruction fields may be provided in each card.

The 604 is completely under the control of each card as it is read. A multiplicity of operations can be called out by the card.

The procedure is simple. A check list instructs the operator as to the terms to be used, the sequence of the cards and the control field to be used for selecting instructions. All the cards are filed by term under index cards. The operator selects the proper cards, sorts and runs them through the 604. The cards are then separated, refiled, and those cards to be used for the next step selected. The process is explained in more detail below.

604 Operation

The functional control of the 604 is accomplished by means of coded punching in a five-column field in each card. There can be a maximum of five such control fields in a single card. They are labeled *A*, *B*, *C*, *D* and *E* for identification. Field selection for the proper control field is accomplished by picking up punch selectors with a rotary switch connected to the control panel. The field is selected by the operator setting the selector switch at the beginning of each operation.

The operations which can be called out by the card are:

1. Multiply 2 eight-digit numbers.
2. Multiply a five-digit constant by an eight-digit multiplicand.
3. Divide 2 eight-digit numbers for an eight-digit ratio.
4. Compute sums involving either addition or subtraction.
5. Compute first differences or first sums.
6. Selective read in.
7. Selective punching.
8. A shifting arrangement to select 8 digits to be punched from the 13 digits of the electronic counter.

The versatility which can be attained with these operations is surprising. The 604 used had 40 program steps. A 60-program machine would make even more flexibility possible.

Card Form

The card form is shown in Figure 1. Each card is identified by the necessary parameters, and by a term number. There is a single eight-digit amount field of three whole

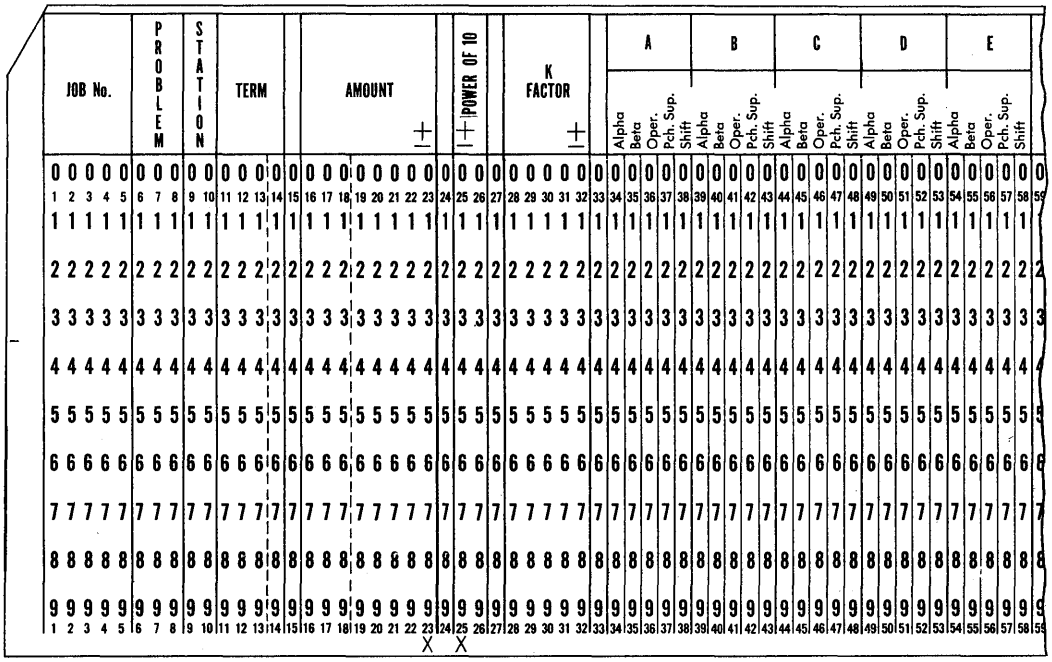


FIGURE 1

and five-decimal numbers. Associated with it is a power of 10 field which locates the decimal point. This power of 10 is not calculated, but is punched into the initial setup cards in advance, as determined by the range which a particular term may have. (It is desirable to keep all amounts as close to one whole number as possible.)

The K field permits a five-digit multiplying or adding constant to be set up.

The operational control of five columns has the following functional operation:

Column 1. Alpha

An X will cause the amount to enter Factor Storage 1 and 2.

Column 2. Beta

An X will cause the amount to enter Factor Storage 3 and 4.

Column 3. Operation

This is coded. Punching one or more digits will call out the associated function.

Code	*Function*
1.	$\alpha \cdot \beta$
2.	$K\alpha$
3.	$\alpha \div \beta$
4.	$+\alpha$
5.	$-\alpha$
6.	$\alpha_N - \alpha_{N-1}$
7.	$\alpha_N + \alpha_{N-1}$

Column 4. Punch Suppress

An X will prevent punching and readout and reset of the electronic counter. The results are always punched from the electronic counter on NX cards, which also clears the counter.

Column 5. Shift

A coded punch of 1 through 5 selects eight positions from the 13-place electronic counter for punching.

Program Sheet

Figure 2 shows the form of the layout sheet for setting up problems. This can be filled out quickly and simply. A card is key punched for each line. This becomes a master deck which is expanded by the number of solutions. The initial data are put in the proper term cards. Then the solution is attained by repeatedly running selected groups of cards through the 604 as explained above.

As an example of the relative efficiency, a single problem, involving approximately 100 operations for each variable, required one week for a manual computer. Five similar problems were solved using the described method in eight hours.

DISCUSSION

Mr. Clark: As I understand it, you must have what is called a semifloating decimal. You float it when necessary, and the rest of the time you leave it fixed. Do I understand, then, that when you expect a crossfoot, you must arrange your powers of 10 so that alpha and beta must have the same power of 10? Is that part of your planning?

Constant	Term	Power	A					B					C					D					E				
			α	β	Oper.	Pch. S.	Shift	α	β	Oper.	Pch. S.	Shift	α	β	Oper.	Pch. S.	Shift	α	β	Oper.	Pch. S.	Shift	α	β	Oper.	Pch. S.	Shift
	1.0	0	X	0	0	X	0	X	0	0	X	0	X	0	0	X	0	0	X	0	X	0	0	X	0	X	0
	2.0	0	0	X	0	X	0	X	0	0	X	0	X	0	0	X	0	X	0	5	X	0					
	3.0	+2						X	0	0	X	0															
	4.0	0	X	0	0	X	0	X	0	4	X	0															
	5.0	0	0	X	0	X	0	0	X	0	X	0	0	X	0	X	0										
	6.0	0	X	0	0	X	0																				
	7.0	+2	0	X	0	X	0																				
	8.0	+1	0	0	1	0	2	0	X	0	X	0	X	0	0	X	0	X	0	0	X	0	X	0	5	X	0
	9.1	0	X	0	0	X	0	0	0	7	0	3															
000.50	9.9	0	0	0	2	0	3	0	X	0	X	0	X	0	0	X	0	X	0	0	X	0	X	0	0	X	0
	10.0	0	X	0	0	X	0	0	0	1	0	3	X	0	0	X	0										
	11.1	0											0	0	6	0	3	X	0	4	X	0					
	11.2	0											0	X	0	X	0	0	0	0	0	3					
	11.9	0	X	0	0	X	0						0	0	1	0	3						X	0	0	X	0
	12.1	+3						0	0	1	0	3	0	X	0	X	0										
	12.9	+3	X	0	0	X	0						0	0	1	0	3										
	13.0	+3	0	0	6	0	3											0	X	0	X	0					
	14.0	+3																0	0	1	0	3	X	0	0	X	0
	15.0	−1	X	0	0	X	0	X	0	4	X	0	0	0	3	0	4	X	0	5	X	0					
	16.0	−1	0	0	3	0	4	X	0	5	X	0	X	0	4	X	0	X	0	0	X	0	X	0	5	X	0
	17.1	0						0	0	0	0	3						X	0	0	X	0					
	17.2	0						X	0	0	X	0						0	0	3	0	3					
	17.9	−1	X	0	0	X	0	0	0	3	0	4	0	X	1	X	0	X	0	5	X	0					
	20.0	−1	0	X	1	X	0	X	0	4	X	0	0	0	0	0	3	0	X	0	X	0	0	X	0	X	0

FIGURE 2

Mr. Bell: The process used in the method described is very simple. We have a field which, in size, is three whole numbers and five decimals. We multiply that by another number with three numbers and five decimals. That is the size of the field punched in the card with the decimal point indicated, and we obtain a product consisting of six whole numbers and seven decimals. That is the way in which the 604 is programmed to obtain a product. Now we are going to keep eight positions out of the possible 13, and shift the decimal point as previously determined when we laid out the problem. Since the decimal point has been shifted, the corresponding power of 10 must be indicated. This indication is actually punched into the card when the problem is set up, and it is this punching which controls the decimal shift.

Mr. Clark: Don't you look ahead and fix your shift instructions so that the power of 10 will align with the two numbers which you expect to add?

Mr. Bell: Addition is the problem, then? Suppose I want to add two numbers together and obtain a sum. If the power of the two is the same, there is no problem; but assume they are different. Suppose the power of the first number is 1, and the second is zero. There are several possibilities. I could punch a 10 into the zero power card. Then I could multiply the first factor by this term and add it to the other number whose power was 1. This gives a result whose power is 1.

Mr. Clark: If you are never going to use that factor for anything else, you would be just as well off to store it originally the way you expect to use it. You might use it otherwise; therefore, it might be an advantage to have it stored this way.

Mr. Bell: There are a lot of modifications with this method. The more you use it, the more you see easy ways of doing something which was done the hard way first.

Mr. Belzer: Do you keep eight places at all times, no matter how many figures?

Mr. Bell: No, we don't. There is nothing in this method which guarantees that you will have eight significant figures. We designed it primarily for engineering problems where a high degree of accuracy and number of significant figures were unnecessary. So, in laying out the problem, it is necessary to keep track of the size of the numbers, and, when the number of digits of the result decreases, shift the numbers over and increase the number of digits in the answer.

Dr. Grosch: In other words, this is not a floating decimal system, and it is exactly identical to the ordinary work that we do in shifting wires on a control panel, except that it is useful for small-scale work, since only one control panel is used.

Mr. Bell: Right.

Mr. Ferber: It is really not limited to just small jobs, however. You can do a job with n operations horizontally.

Mr. Keast: There are many problems where sines, cosines, and logarithms are involved, very frequently. How do you feel about that, please?

Mr. Bell: Well, in this particular problem a point was reached at which square roots were needed. We have a 604 control panel that evaluates square roots by the iterative approach. The 604 actually computes the first guess that is used. So we are not forced to use the method discussed. We simply take the cards of the values for which we want square roots, run them through the 604 with the square-root control panel. Then reproduce these cards into the field which is to contain the value to be used in repeating the calculations. Practically, you could do this iteration process on the described setup, but you would use a great many cards, and the process would take a long time.

Transition from Problem to Card Program

GREGORY J. TOBEN

Northrop Aircraft, Incorporated

I WANT to talk about card programming as applied to engineering problems. The IBM Card-Programmed Electronic Calculator is only now being delivered; so I should explain how we have been able to do some card programming on other combinations. I shall describe a simple two-card program for generating sine and cosine, a 13-card program for wing loads on an airplane, and a five-card program for a Monte Carlo calculation. Then I should like to mention briefly a number of setups made by members of our group and interesting program variations in each. Then I shall summarize some of the results gained in solving these problems.

In the days prior to the IBM Type 602 Calculating Punch, we were presented with a problem involving the step-by-step solution of a set of 14 simultaneous nonlinear differential equations. No analytical solution was possible because of step functions. The two large-scale digital machines capable of doing the problem were busy on higher priority work, and the problem was not suited for analogue computers such as the differential analyzer because of a large spread in the size of the numbers. Shifts of 10^4 sometimes meant that a gear would have to turn for years to effect the result. Moreover, it was a design problem, and the whole course of investigation depended upon the early results. The solution was important to the guided missile program, and we were able to persuade IBM to convert our IBM Type 405 Alphabetical Accounting Machine into something suitable for the job. They made available an IBM Type 603 Calculating Punch, which was then out of production, and connected it, via cable, to the 405. Forty class selectors and 40 x distributors were added to complete the job. The elapsed time from preliminary design to delivery of the machine was only six weeks, and it was so well done that after two years of use we have been able to think of only minor improvements.

When the two machines are attached, the 405 provides storage space for constants and factors which may be transferred from counter to counter to obtain desired combinations and from counter to multiplicand or multiplier of the 603 for multiplication. Products may be sent to multiplier and multiplicand for remultiplication, raising to power, etc., to the 405 counters for accumulation, or to type bars for printing of final results. The sequence of operation is controlled by a set of program cards fed through the 405. These cards contain x or digit punchings to pickup selectors to call for the required transfers, etc. These cards may also contain factors to be used in computations. The results may be listed step by step until the problem checks and then tabulated and summary punched as usual. When so used, of course, the multiplication rate is 150 cards per minute. Counter transfers are accomplished by wiring counter exits of the transmitting counter to counter entry of the receiving counters, with instructions for either add or subtract. No special wiring is necessary to take care of negative numbers. If a negative or complement number is standing in a counter, it appears in the multiplier as a positive number, and the sign of the counter receiving the product is reversed. A six-place column shift is placed ahead of the multiplier output and controlled by x's. Round off may be wired on or off. A negative balance test on the counters is available. While only one multiplication can be done on each card cycle, other parallel operations are always used. Table look-ups can be made, additions, counter-to-counter transfers and comparing operations are all possible.

The 603 is not able to divide, and multiplications larger than 6 by 6 had to be done in four passes. It was soon apparent that even the routine everyday work could be done more economically on the combination than any other way. This one machine keeps six men busy; we have no other multipliers and no regular key punch operator.

Card programming is merely the use of punched cards to control selectors and other functions of the machine. Actually, it accomplishes a great deal more than this. It is an attempt to replace all of the programming that would otherwise be done by the operator, when he handles cards and changes control panels, as he gathers together the separate parts of a computing routine. When all outside handling is eliminated, spectacular time and cost savings can be made. Ten-to-one ratios over other tabulating procedures are not uncommon. Card programming is used to supplement the machine programming available on machines such as the IBM Type 604 Electronic Calculating Punch.

Illustration of Card Programming

Sine and cosine functions may be generated together using two counters for the functions and a third for x in radians. The formulae are:

$$\sin x = \int \cos x \, dx$$

$$\cos x = \int -\sin x \, dx$$

If the interval is $dx = 1/1000$ the products of dx can be obtained by shifting the numbers three places (Figure 1).

Card No.	8A cos		8B sin	4A x
1 Read in initial values	1.000000	+	.000000 1000	.000
2	1	−	.001000	.001
3	.999999	+	999	
4	1	−	.001999	.002
5	.999998	+	999	
6		−	.002998	.003
.				
.				
.	etc.			

FIGURE 1

By using a selector as an alternator the shift may be accomplished without card punching. Card count adds in 4A on alternate cards. This procedure will generate a three-place sine and cosine, and it is often used as a part of a program. Accuracy can be improved by making dx smaller or by adding a second order term, or both.

Consider the problem of determining wing loads for an airplane (Figure 2).

It is necessary to know the total shearing forces on the wing. These forces will be the weight and lift. From weight measurements and wind tunnel tests it is known just what these loads will be for level flight. To obtain the loads for other conditions such as pullout, dive or landing, they can be multiplied by a factor. The bending moment on the wing

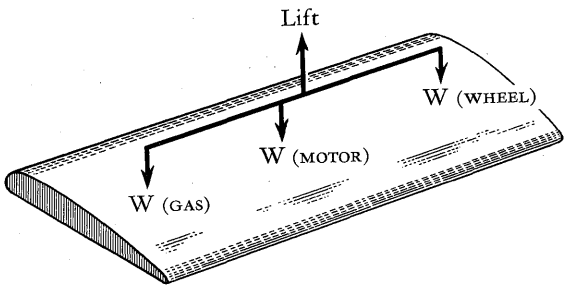


FIGURE 2

will be the shears times their distances. The torque will be the weights times the distance from the reference line in the wing. In addition to these forces, there are chord shear and moment due to accelerations in turning, etc. The formulae have the following form:

shear-beam = [4] ④ + [9] ①
moment-beam = [4] ⑤ + [9] ②
torque = [7] ⑥ + [8] ⑦ − [9] ③
shear-chord = [5] ④ + [6] ⑧ + [10] ①
moment-chord = [5] ⑤ + [6] ⑨ + [10] ②

The labels on the factors need not be listed, except to say that, in general, the circles represent weights times distances, and the boxes are load factors. The circles will remain constant for a good part of the run, while the boxes change very often. There are 200 different combinations of load factors necessary to cover all reasonable attitudes of the plane. The five forces listed above are needed at perhaps twenty different points or stations from the tip of the wing to the root. The problem consists, therefore, of 13 products, grouped into five sums, for 20 stations and 200 conditions ($13 \times 20 \times 200 = 52{,}000$ multiplications). The job will take 14 program steps, all alike in that two factors will be read in and a product read back. For 20 stations there will be 280 program cards. These 280 cards must go through the machine once, at the rate of 150 per minute, for each set of load factors, i.e., 200 times. The seven load factors are contained in a single read-in card, and they remain in the machine while one run of 20 stations is being computed.

If the operator runs the program, 50,000 pairs of factors should be punched into cards. This will require handling 50,000 cards, or 10,000 cards five times. Fifty thousand multiplication cycles at 100 per minute must be performed. In addition, there will be more than a little gang punching, sorting, and collating, plus a run in the accounting machine to print the answers. The wear and tear on the operator is, obviously, much less where card programming is used. It is easy for him to watch the answer columns for abnormally high stresses, and a recheck of factors can be started immediately. There have been occasions where wrong signs were picked up on load factors and a few erroneous numbers resulted. Re-runs can be made in two minutes.

MONTE CARLO CALCULATIONS

In this type of problem, part of the program is determined by a random choice from a file of random digits. Our source is the RAND Corporation's 20,000 card file. Each card contains 50 random digits and a serial number. The problems we perform usually involve the life history of a neutron as it goes through a slab of material. Several groups are working these problems with a variety of approaches. Many here have worked with a mass sample, i.e., a thousand neutrons are taken from impact to impact and

when the sample is pretty well depleted, statistics are gathered. With card programming it is possible to follow the course of a single fictitious neutron until it is either captured or goes outside the limits of the experiment. When this happens, the calculator may be restored to its initial conditions and the same routine continued with another neutron. When stable curves have been produced, the material is ready to be changed or another experiment tried.

To set up this problem it is necessary to make a numerical analogy of each of the physical events to be considered. Counters may be used to store the numbers representing the probability that a certain event will occur. For example, if the probability of capture was 5%, a 95 might be stored in a counter. Then, if two-digit random numbers were subtracted from this counter, it would indicate a negative balance 5 times out of 100. If there were four possibilities stored in four counters and the probabilities were made accumulative from left to right, the leftmost negative sign would indicate the process to be used. Similarly, the dimensions of the slab may be stored in a counter and continuously compared with the position of the neutron. When the progress exceeds the slab thickness, the machine is cleared.

When neutrons travel through a shielding material, there are four possibilities:

Type of Collision	*Operation*
capture	start over
elastic light	energy change
elastic heavy	no energy change
inelastic	variable energy change

Energy levels are represented by discrete numbers, using 0 for the highest energy and dropping to 9 for the lowest. The intervals and probabilities are chosen so that they confirm closely to the physical problem. Needless to say, our mathematicians do most of the work on a problem of this kind. Not only must they contrive experiments with numbers to fit the actual physical conditions, but they must also be alert to see that a truly random sample has been used and none of the rules of the game has been inadvertently violated (Figure 3).

When the calculator is started, a set of random digits and a mean-free path are picked up. At the start the direction can only be forward. A direction at random (cos) is selected. A multiplication factor is needed, depending upon the energy level (L) and a set of process probabilities which also depend on the energy level (00). In the meantime, a product of PL cos has been formed and accumulation of the progress of the neutron (Z) is begun. Now take a total, printing everything, but clearing only PL cos and the random digits. The process, as shown by the status of the probability counters, is determined for the succeeding impact which is now considered. Card 1 reads the previous direction to a digit selector so as to make a look-up in a table punched in card 2. A new pair of digits and a new mean free path, both random, are stored. On card 2 there is a choice; if the probability counters indicate three minus signs, the number from the table look-up on card 2 is taken and used to select a cosine on card 3. The first table look-up also provides a change for the energy counter. In case the probability counters indicate one or two minus signs, then a random digit is used to select the cosine, and a corresponding energy change is made. Then the new energy is read while the progress is computed; and the process continues until either the forward progress exceeds 10 centimeters or turns negative, which means that the neutron has come out of the front face, or until the process indicates a capture.

The five card decks are made in sets of several thousand, and the 20,000 random digits are collated in as the runs are made. Frequent re-randomizing is done by re-sorting and by the use of new random digit columns.

Stress work has been assigned job numbers since the first of the year. They are now up to number 120. Work has been done for such groups as aerodynamics, thermodynamics, guided missile section, wind tunnel, quality control, NEPA, etc. We estimate an average of four jobs per week. The usual procedure is for someone to bring in both mathematics and data. The author stands by while the operator wires the control panel and runs through a set of check calculations. Then the operator stands in front of the machine while the answers are being printed. In a week the operator doesn't really become familiar with a problem, but the large variety makes the work interesting.

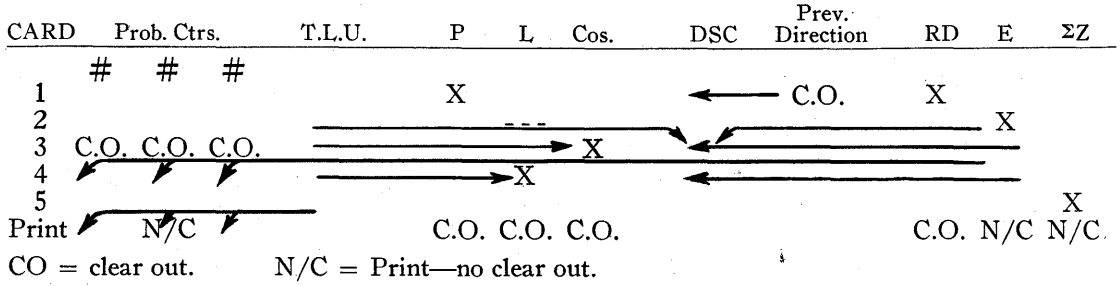


FIGURE 3

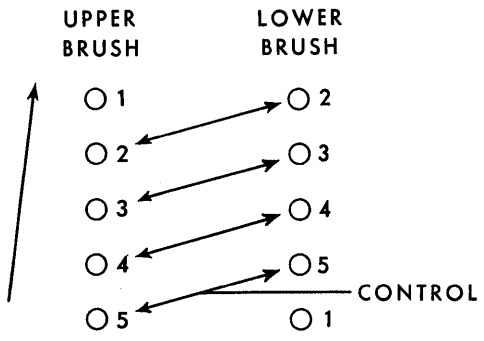


FIGURE 4

STRESS EXAMPLES

Fuselage Balance-Tail Loads. First a weight is assumed. Then a table look-up is made for the drag and airload, each producing torque. From the known lever arm, the up force on the tail to balance the torque is computed. Subtract this uplift from the wing lift and repeat three times. After the third try, a refined table look-up is used which gives fuselage and wing components separately. Thus, there are eight table look-ups and three iteration cycles.

Wing Loads. Over 1,000 runs were done, averaging 200 flight conditions on each plane analyzed. Considerable saving on cards is possible on each setup.

Fuselage Loads. Some 800 runs have been made, 442 on a cargo plane which had a large number of possible cargo distributions.

Section Properties. Both wing and fuselage section properties have been run, using a newly developed tapered beam analysis. Thirty stations with three types of bending per station are shown, 94 items of area, 94 programs maximum. Except for the tail loads, all stress jobs were run on one set of control panels, and by operators having no previous experience.

Wind Tunnel Balance Corrections. Six counters were used for storage and six for answers. The program involved 65 program steps giving first and second order corrections to six balance readings. Two lines of six components are computed each minute, which is nearly 20 times as fast as any previous runs.

Missile Flight Paths. Most of these jobs are classified, but they do involve generation of sine, cosine and arc sine, by power series to nine places, which is better than can be had from any reasonable sized tables.

Performance. This involves six table look-ups. It was necessary to go in with two values and come out with a third. For coarse lattices, it was necessary to use second differences. This job took twice as long to do with card programming as by other methods.

Conclusion on Programming. A planning sheet should be made for each job. A good planning sheet will account for 75 per cent of the work on most jobs. The form doesn't seem to be too important as long as it is consistent. Standard or fixed control panels can be used if the work is not too varied. Thus, it is possible to program a large variety of jobs by changing only the program deck.

Each program step should do a multiplication; parallel operations should be used whenever possible.

A sequence control on each program deck should be used to insure that no card gets out of step. This is provided by punching the number of each card and number of the card following (Figure 4).

DISCUSSION

Dr. Polachek: This question is not necessarily directed to the speaker but to any of these numerous random mathematicians. The question is about the table of random digits. Apparently they have serial numbers. My question is: Suppose you sort the table on the random numbers; would the serial numbers then become a random table?

Dr. Hamming: We were required one time to randomize the numbers 1 to 1,000; so we took the first thousand cards, sorted on ten random columns, one after another, and listed the original column. This gives us the numbers 1 to 1,000—and fairly random.

Professor Tukey: Maybe it needs to be pointed out that these are pasteurized random numbers and not certified, if you do them this way, because they are going to come up without repetition, and in a thousand random numbers there ought to be some repetitions; but for many purposes this might be better.

Dr. Hamming: As you recall, it was to get distinct the numbers 1 to 1,000 random without any duplications.

Professor Tukey: That is what you wanted for that purpose; but an ordinary random number table is supposed to have some repetitions. This is a different kind of table.

*Best Starting Values for an Iterative Process of Taking Roots**

PRESTON C. HAMMER

University of California
Los Alamos Scientific Laboratory

WITH THE ADVENT of computing machines such as IBM type 602, 602-A, and 604 calculating punches capable of repeating calculation sub-programs making use only of temporary or internal storage capacity, iterative processes of calculating certain functions have become feasible. The most important such function, in terms of its frequency of occurrence, is the square root. The question of how to pick the best starting values for one type of iteration for square roots is answered completely in this note when one is concerned with limits on absolute or relative errors for a minimum fixed number of iterations.

Errors in an Iterative Method for Square Roots

The iterative method here discussed is the classical one given by the formula:

$$a_k = .5(x^2/a_{k-1} + a_{k-1})$$

where x^2 is the radicand, a_k, $k = 1,2, \ldots$, is the kth approximation to the square root x, and a_0 is the starting value. The error $a_k - x$ is readily calculated[a] by observing:

$$\begin{cases} a_k - x = 1/2\dfrac{(a_{k-1} - x)^2}{a_{k-1}} \\ a_k + x = 1/2\dfrac{(a_{k-1} + x)^2}{a_{k-1}} . \end{cases} \tag{1}$$

From which, by division and extension, we have

$$\frac{a_k - x}{a_k + x} = \left(\frac{a_{k-1} - x}{a_{k-1} + x}\right) = \ldots = \left(\frac{a_0 - x}{a_0 + x}\right)^{2^k}. \tag{2}$$

Solving (2) for a_k and then for $a_k - x$ in terms of a_0 and x we have

$$a_k = \frac{x[(a_0+x)^{2^k} + (a_0-x)^{2^k}]}{(a_0+x)^{2^k} - (a_0-x)^{2^k}} \tag{3}$$

$$a_k - x = \frac{2x(a_0-x)^{2^k}}{(a_0+x)^{2^k} - (a_0-x)^{2^k}} \tag{4}$$

$$\frac{a_k - x}{x} = \frac{2(a_0-x)^{2^k}}{(a_0+x)^{2^k} - (a_0-x)^{2^k}} \qquad k = 1,2, \ldots . \tag{5}$$

Now, we observe that for $k = 1,2, \ldots$. the error always has the positive sign. It is also readily proved that the error and the relative error as given by (5) are monotonic decreasing functions of a_0 for fixed x for $a_0 < x$ and monotonic increasing functions of a_0 for $a_0 > x$. Similar monotonic properties hold for the errors as functions of x when a_0 is fixed.

The Best Starting Values

We now state the problems: Given a radicand in the interval (N_1,N_2) $x_1^2 = N_1$, $x_2^2 = N_2$; to find a single starting value a_0 which for a will minimize the maximum absolute error in the kth approximation and find likewise a starting value which will minimize the maximum relative error for all choices of radicands in the interval.

To solve these problems, we first define an auxiliary function

$$P(a,x,k) = (a-x)^{2^k}/(a+x)^{2^k}, \text{ then}$$

$$a_k - x = \frac{2x\,P(a_0,x,k)}{1 - P(a_0,x,k)} \quad \text{and} \tag{6}$$

$$\frac{a_k - x}{x} = \frac{2\,P(a_0,x,k)}{1 - P(a_0,x,k)} . \tag{7}$$

From the monotonic properties of the errors as functions of the starting values we may state that for any starting value a_0 the largest absolute error will occur for $x = x_1$ or $x = x_2$. Hence, the largest error will be minimized by equating the errors at the upper and lower extremes of the range of the radicand. The same statement will apply to relative errors. Hence, we have the theorems:

THEOREM 1. *The starting value a_0 which minimizes the maximum absolute error in the nth approximation for a range of radicands between $N_1 = x_1^2$ and $N_2 = x_2^2$ is the solution of the equation*

$$\frac{2x_1\,P(a_0,x_1,n)}{1 - P(a_0,x_1,n)} = \frac{2x_2\,P(a_0,x_2,n)}{1 - P(a_0,x_2,n)} . \tag{8}$$

In general, these solutions a_0 depend on the number of iterations n, and a_0 is a decreasing function of n. The quantity on the left in equation (8) gives the actual maximum error when the solution a_0 is substituted.

THEOREM 2. *The starting value a_0 which minimizes the maximum relative error for a range of radicands between*

*This paper was presented by title. Work was done under U. S. government contract.

[a]This method appears in Whittaker and Robinson, *The Calculus of Observations.*

$N_1 = x_1^2$ and $N_2 = x_2^2$ is $a_0 = \sqrt[4]{N_1 N_2} = \sqrt{x_1 x_2}$, independent of the number of iterations k. The maximum relative error then is

$$\frac{2P(\sqrt{x_1x_2},x_1,k)}{1-P(\sqrt{x_1x_2},x_1,k)} = \frac{2P(\sqrt{x_1x_2},x_2,k)}{1-P(\sqrt{x_1x_2},x_2,k)}. \quad (9)$$

It is easily seen that the ratio a_0/x_1 depends only on the ratio x_2/x_1 in both theorems; hence, one may multiply a_0, x_1, x_2 by a positive number and still have theorems 1 and 2 valid for the new quantities. The maximum relative error does not change under a scalar change.

To give an idea of the maximum relative errors in using $\sqrt{10}$ as a starting value for radicands between 1 and 100, we have that the third approximation gives a relative error less than 1.08×10^{-2}, the fourth approximation a relative error less than 5.7×10^{-5}, the fifth approximation a relative error less than 1.5×10^{-9}, and the sixth approximation a relative error less than 1.26×10^{-18}. The best integer starting value for radicands between 1 and 100 for six iterations is 3 for either absolute error or relative error. If the absolute error is to be minimized, the best starting value for one iteration is $(x_1 + x_2)/2$. As the number of iterations increase, the best starting value decreases toward $\sqrt{x_1x_2}$, although we have not succeeded in showing actual convergence to $\sqrt{x_1x_2}$.

If the machine under consideration can discriminate among several classes of radicands, then one can use the method here proposed to determine the starting value associated with each class.

High Order Roots

If one uses the iteration for nth roots

$$a_k = \frac{1}{n}\left(\frac{x^n}{a_{k-1}^{n-1}} + (n-1)\, a_{k-1}\right), n = 2,3,\ldots (10)$$

the formulas for errors are not simply derived. However, to minimize the maximum relative error one has the following theorem:

THEOREM 3. *The single starting value for taking nth roots which minimize the maximum relative error for all radicands between $N_1 = x_1^n$ and $N_2 = x_2^n$ is*

$$a_0 = \sqrt[n]{\frac{x^1x^2\,(x_2^n - x_1^n)}{(x_2-x_1)(n-1)}} \quad (11)$$

regardless of the number of iterations used.

Proof: Let a be any positive number. Consider the relative error in the first approximation a_1 resulting from using a as a "guess" for the nth root of $N = x^n$. This relative error is

$$\frac{a_1-x}{x} = \frac{x^n + (n-1)a^n - n\,a^{n-1}\,x}{n\,a^{n-1}\,x}$$

$$= \frac{1}{n}\left[\left(\frac{x}{a}\right)^{n-1} + (n-1)\left(\frac{a}{x}\right) - n\right]. \quad (12)$$

We want to show that this relative error is always positive unless $a = x$, when it is zero. Let $y = x/a$ and denote the relative error by z, then (12) becomes

$$n\,z = \frac{y^n + (n-1) - n\,y}{y}. \quad (13)$$

This relative error z is positive, zero, or negative according as the numerator, $y^n + (n-1) - n\,y$, is positive, zero, or negative. Setting the first derivative of this function equal to zero, we have

$$n\,y^{n-1} - n = 0 \quad \text{whence} \quad |y| = 1\ . \quad (14)$$

The second derivative is

$$n(n-1)\,y^{n-2} \quad (15)$$

which is positive for $y > 0$.

Hence as $z = 0$ for $y = 1$ is a minimum, and the first derivative of the numerator is negative for $0 < y < 1$ $(x<a)$, but positive for $y > 1$, we conclude that $z = 0$ if, and only if, $a = x$, and it is otherwise positive. Furthermore, it is easily established that z is a monotonic increasing function of $|a - x|$ for either x fixed and a varying, or a fixed and x varying.

In view of the above, one may proceed as with the square roots. For radicands between $N_1 = x_1^n$ and $N_2 = x_2^n$ any starting value a will give a maximum relative error in the first approximation for either N_1 or N_2 by the monotonic property mentioned above. Hence, the minimum largest relative error in the first approximation will occur when the error using N_1 is the same as the error using N_2. Equating these relative errors, and using the fact proved above that the relative error is positive or zero, we have from (12)

$$\left(\frac{x_1}{a}\right)^{n-1} + (n-1)\left(\frac{a}{x_1}\right) - n = \left(\frac{x_2}{a}\right)^{n-1} + (n-1)\left(\frac{a}{x_2}\right) - n\ . (16)$$

From (16), one finds the unique solution

$$a_0 = \sqrt[n]{\frac{x_1x_2\,(x_2^{n-1} - x_1^{n-1})}{(n-1)\,(x_2-x_1)}}. \quad (17)$$

Now we have still to prove that this same value a_0 is the best starting value, regardless of the number of iterations. Before proving that, we remark that the best starting value a_0 to minimize the maximum absolute error with one approximation is

$$a_0 = \sqrt[n-1]{\frac{x_2^n - x_1^n}{n(x_2-x_1)}}. \quad (18)$$

It may be demonstrated, although we here omit the details of proof, that the best starting value a_0, to minimize the maximum relative error in k iterations, is that value of a_0 which yields approximations to $\sqrt[n]{x_1^n}$ and $\sqrt[n]{x_2^n}$ at the kth iterate such that the relative errors in both are equal. The following lemma then shows that a_0 of (17) is the best starting value to minimize the maximum relative errors independently of the number of iterations.

Lemma 1. If the relative errors in using a_1 as an approximation of $\sqrt[n]{x^n}$ and b_1 as an approximation of $\sqrt[n]{y^n}$ are equal, then the relative errors in a_2 and b_2 are equal, where

a_2 and b_2 are the iterates, obtained respectively by substituting a_1 for a_{k-1} in formula (10) and y for x, b_1 for a_{k-1}.

Proof: We have given that $(a_1-x)/x = (b_1-y)/y$ or what is equivalent, $a_1/x = b_1/y$. Now

$$\begin{cases} a_2 = \dfrac{x^n + (n-1)\,a_1^n}{n\,a_1^{n-1}} \text{ and } b_2 = \dfrac{y^n + (n-1)\,b_1^n}{n\,b_1^{n-1}} \\ \text{hence,} \\ \dfrac{a_2}{x} = \dfrac{1}{n}\left[\left(\dfrac{x}{a_1}\right)^{n-1} + (n-1)\left(\dfrac{a_1}{x}\right)\right] \text{ and} \\ \dfrac{b_2}{y} = \dfrac{1}{n}\left[\left(\dfrac{y}{b_1}\right)^{n-1} + (n-1)\left(\dfrac{b_1}{y}\right)\right] \end{cases} \tag{19}$$

as $a_1/x = b_1/y$ we have $a_2/x = b_2/y$, which is equivalent to the conclusion of the lemma.

This proves the lemma. Now, as a_0 was chosen so that the relative errors of the first approximations to $\sqrt[n]{N_1} = x_1$ and to $\sqrt[n]{N_2} = x_2$ were equal, it follows from Lemma 1 that all successive approximations to x_1 and x_2 will have the same relative error at a given iteration. This concludes the proof of theorem 3.

Concluding Remarks

We have called attention to formulas for the errors in the classical iterative method of taking square roots, and applied these formulas to determination of the best starting values to use, if one wishes to obtain a certifiable accuracy in the smallest number of steps. While the determination of errors for higher order roots was not given algebraically, this may be done numerically for particular circumstances. For example, in taking cube roots with $N_1 = 1$, $N_2 = 1000$, the best value for minimizing maximum relative errors for a fixed number of iterations is

$$a_0 = \sqrt[3]{\frac{10(10^2-1)}{2(10-1)}} = \sqrt[3]{55}\,.$$

If one chose to use 4 for a starting value, then the maximum relative error would occur at $x^3 = 1$ as $\sqrt[3]{55}$ is a dividing point and $4 > \sqrt[3]{55}$.

It should be pointed out that the methods here discussed do not apply if one has a machine for which the number of iterations need not be fixed, but which will test the accuracy at each iteration. In that case, a choice of a best starting value would be better determined by minimizing the average length of time of calculations for the distribution of radicands at hand and for the accuracy desired. However, the results stated here will be of some use in proceeding with the minimum calculation time determination.

*Improvement in the Convergence of Methods of Successive Approximation**

L. RICHARD TURNER

Lewis Flight Propulsion Laboratory, NACA

THE METHOD of successive approximation is frequently used in numerical mathematics, but in some cases the rate of convergence is discouragingly slow. Professor Southwell has shown that the rate of convergence of Liepmann's method of solving partial differential equations may be substantially improved by examining the distribution of "residuals" (increments in Liepmann's method) and applying local corrections.

Southwell's "relaxation" technique is not readily adaptable to machine computing methods. It is possible, however, by examining the whole solution to determine the rate of disappearance of the currently dominant error terms and then to remove such dominant terms in a single step of calculation.

Theory of the Method

Let the ultimate solution of a given problem be the K-dimensional vector x:

$$X = (x_1, x_2, x_3, \ldots, x_k),$$

which is obtained as the limit of the convergent sequence

$$X = \lim_{n\to\infty} \{X^{(0)}, X^{(1)}, X^{(2)}, \ldots, X^{(n)}, \ldots\} . \tag{1}$$

(It will be assumed in the analysis that the components $x_i^{(m)}$ of the mth iteration are all real numbers, although this restriction can be removed.)

We now suppose that, at the nth step, that $X^{(n)}$ is composed principally of the solution X and two error terms $E^{(n)}$ and $F^{(n)}$ of a form such that

$$E^{(n+1)} = \bar{\lambda} E^{(n)}$$
$$\text{and} \qquad F^{(n+1)} = -\bar{\lambda} F^{(n)} . \tag{2}$$

Then it is found that

$$\begin{aligned} X^{(n)} &= X + E^{(n)} + F^{(n)} \\ X^{(n+1)} &= X + \bar{\lambda} E^{(n)} - \bar{\lambda} F^{(n)} \\ X^{(n+2)} &= X + \bar{\lambda}^2 E^{(n)} + \bar{\lambda}^2 F^{(n)} \\ X^{(n+3)} &= X + \bar{\lambda}^3 E^{(n)} - \bar{\lambda}^3 F^{(n)} \end{aligned} \tag{3}$$

so that

$$\bar{\lambda}^2 = \frac{X^{(n+3)} - X^{(n+2)}}{X^{(n+1)} - X^{(n)}} \tag{4}$$

and

$$X' = \frac{X^{(n+3)} - \bar{\lambda}^2 X^{(n+1)}}{1 - \bar{\lambda}^2} . \tag{5}$$

Now, in general, the operation indicated in equation (4) is not even defined. Therefore, some adequate working approximation must be substituted for equation (4). Two of these appear to be worthwhile. We define

$$\delta_j^{(K+1)} = x_j^{(K+1)} - x_j^{(K)} \tag{6}$$

and in terms of these δ's which are defined for each point for which a calculation is made

$$\bar{\lambda}_1^2 \equiv \frac{\sum_{j=1}^{K} \delta_j^{(n+3)} \delta_j^{(n+1)} / |\delta_j^{(n+1)}|}{\sum_{j=1}^{K} |\delta_j^{(n+1)}|} \tag{7}$$

or

$$\bar{\lambda}_2^2 \equiv \frac{\sum_{j=1}^{K} \delta_j^{(n+3)} \overline{\delta_j^{(n+1)}}}{\sum_{j=1}^{K} \delta_j^{(n+1)} \overline{\delta_j^{(n+1)}}} \tag{8}$$

Equation (7) is meaningful only if the δ's are real numbers. Equation (8) makes sense for any definition of δ_j for which a complex conjugate $\bar{\delta}_j$ and the operation ab are defined. Equation (8), which corresponds to taking a first moment in statistics, is more elegant than equation (7) but involves much more effort [and is really not much better, because it is only on rare occasions that the initial hypothesis, equation (2), is sufficiently near to the truth to justify the use of great precision in the adjustment indicated in equation (5)]. For this reason it is recommended that, where at all possible, equation (7) be used. This rule should not be

*This paper was presented by title.

applied if K is a small number. In this case equation (8) is a much safer rule.

When λ^2 has been found, equation (5) is applied to each of the elements of $X^{(n+3)}$ and $X^{(n+1)}$ to obtain an improved iterant X'.

Application Notes

Strictly speaking, the basic hypothesis of the method can be met only for linear algorithms, that is, algorithms in which the $n+1$st iterant is the result of linear operations on the elements of the nth iterant. In practice the method is found to apply satisfactorily to various nonlinear processes such as the calculation of the latent roots of matrices by repeated multiplication and normalization of an arbitrary vector.

It is suggested that if λ^2 is a small number, say 0.1, the correction technique should not be applied unless it is found that λ^2 is substantially constant for several iterations.

In the use of the method by the author, no case has occurred in which λ^2 fell outside the range 0 to 1. Such cases will form a fresh field for the experimentalist.

Illustrations

Three charts illustrate the method for the solution of Laplace's equation, with the boundary conditions shown around the edge of the L-shaped domain of the figures. The initial approximation (Figure 1A) was taken to be zero at all interior points. The $(n+1)$st iteration $n = 0$ to 2 was computed as

$$x_{ij}^{(n+1)} = \frac{1}{4}\left[x_{i-1,j}^{(n)} + x_{i+1,j}^{(n)} + x_{i,j-1}^{(n)} + x_{i,j+1}^{(n)}\right].$$

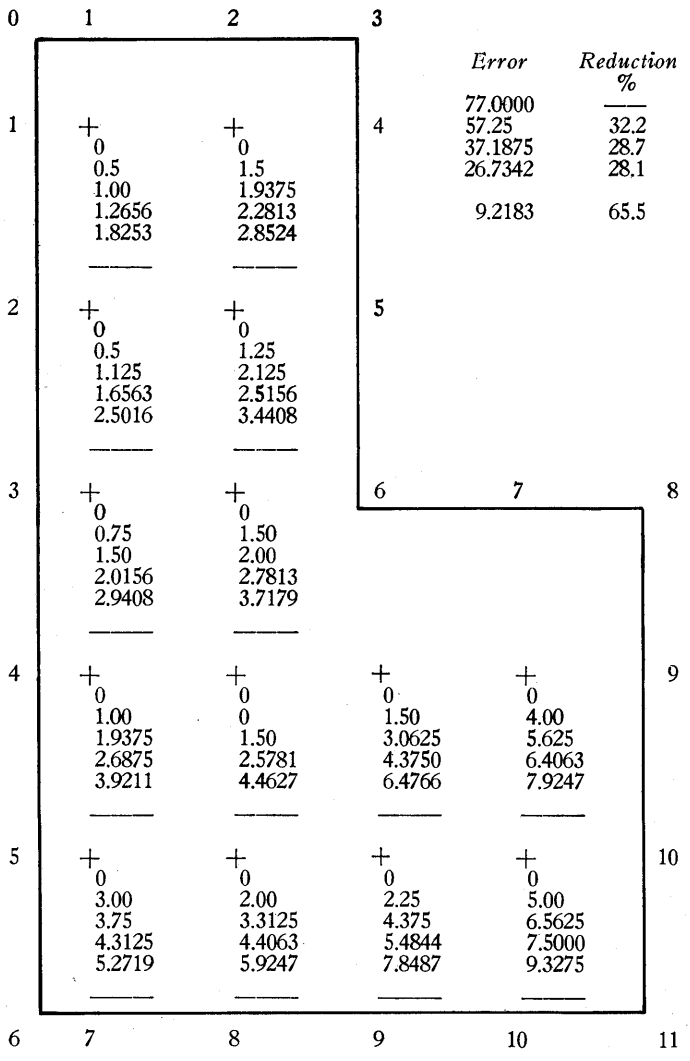


FIGURE 1A. RESULTS OF THE FIRST THREE NORMAL ITERATIONS AND FIRST ACCELERATION

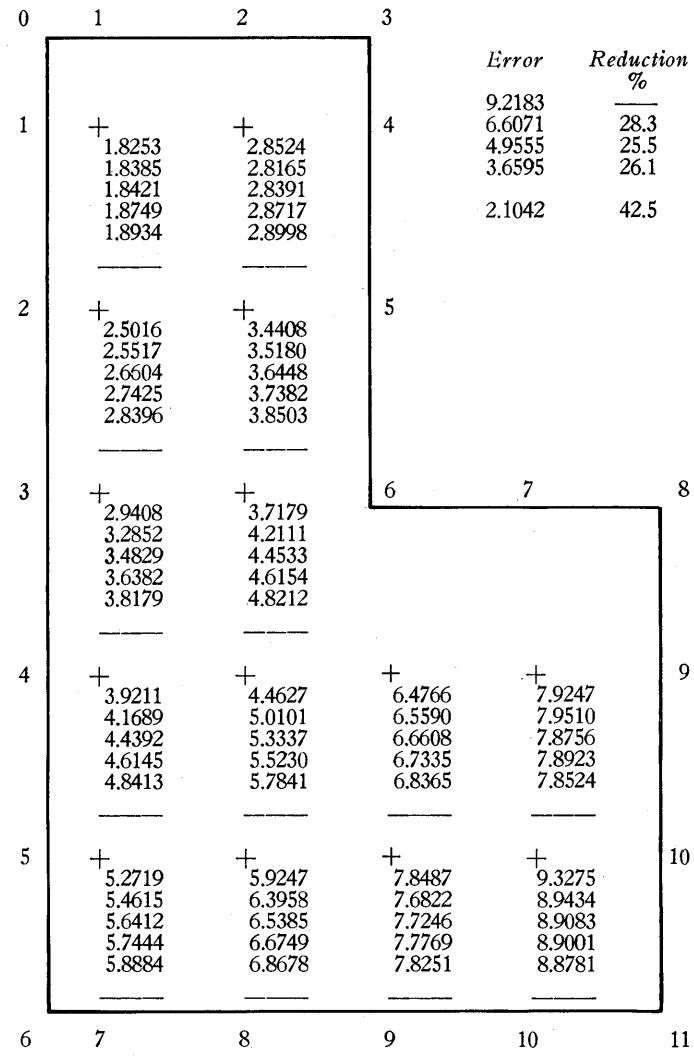


FIGURE 1B. RESULTS OF THE SECOND THREE NORMAL ITERATIONS AND SECOND ACCELERATION

Three iterations were carried out on each sheet, after which $\bar{\lambda}^2$ was computed from equation (7) and the underscored result x'_{ij} computed. The corresponding values of $\bar{\lambda}^2$ and $\bar{\lambda}^2/(1-\bar{\lambda}^2)$ are:

Figure	$\bar{\lambda}^2$	$\bar{\lambda}^2/(1-\bar{\lambda}^2)$
1A	0.422	0.731
1B	0.337	0.509
1C	0.478	0.915

The "errors" are the sum for all points of the absolute value of the deviation of each $x_{ij}^{(n)}$ from the true solution x_{ij}. The "reduction" is the reduction in the error from its value in the last preceding iteration.

The final error (Figure 1c) after nine normal iterations and three acceleration adjustments is approximately one-tenth of the error that would result from 12 normal iterations. A similar reduction of error without the use of the acceleration technique would have required roughly 20-25 iterations.

The acceleration technique has recently been applied successfully to the calculation of the latent roots of matrices. In this case the adjustment is made on the value of the elements in the characteristic vector or modal column. Since, as a rule, only a few elements occur in the modal column, it is recommended that equation (8) be used to compute $\bar{\lambda}^2$ in such cases.

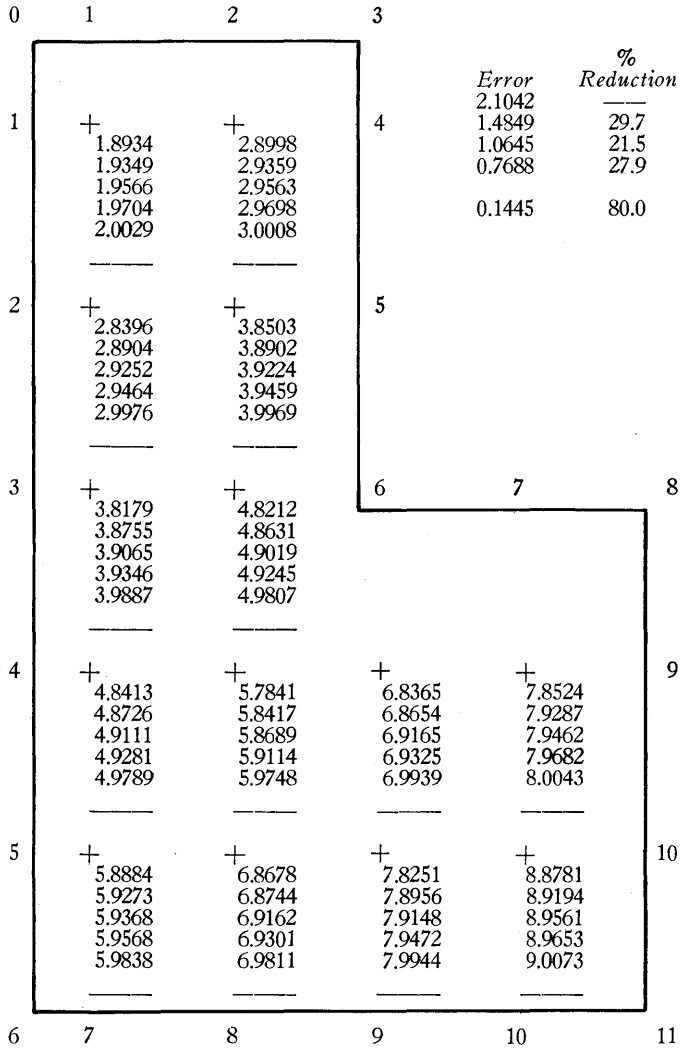


FIGURE 1C. RESULTS OF THE THIRD THREE NORMAL ITERATIONS AND THIRD ACCELERATION

*Single Order Reduction of a Complex Matrix**

RANDALL E. PORTER

Boeing Airplane Company

THE REDUCTION of a matrix is a systematic application of the rule that the value of a matrix is not altered if each element of a row (or column) is multiplied by the same number, and the products subtracted from (or added to) the corresponding elements of another row (or column). This rule is used to reduce all but one element in a column (or row) of a matrix to zero.

Repeated reductions may be employed to alter the form of the matrix to one from which a desired result may be readily obtained. The determinant of a square matrix, for example, may be obtained from the "*n* fold" product of the main diagonal of the equivalent triangular matrix (Figure 1).

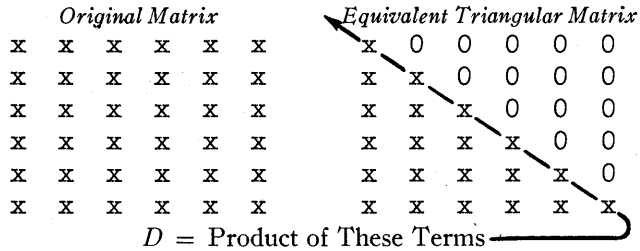


FIGURE 1

The solution of linear simultaneous equations, the computation of a reciprocal matrix, etc., may be accomplished by using a reduction procedure to alter the form of the matrix.

The following reduction procedure is for use with matrices having complex elements. No expansion of the matrix elements is required. A standard IBM Type 602-A Calculating Punch may be used for the required calculations.

Key Punch Instructions

Punch one card for each element of the matrix as indicated below:

Item	*Max. Size*	*Card Cols.*
problem no.	xxxxx	1-5
row no.	xx	6-7
col. no.	xx	8-9
*real sign	x	10
real amount	xxx.xxxxx	11-18
*imag. sign	x	19
imag. amount	xxx.xxxxx	20-27
common x-28		

*NOTE: Punch "0" for plus and "x" for minus.

Check key punched cards by listing and auditing, or by verifying.

Card Preparation Instructions

1. Reproduce working cards from key punched cards as follows:

Item	*Max. Size*	*From*	*Into*
problem no.	xxxxx	cols. 1-5	cols. 1-5
*		and cols. 80-76	
row no.	xx	cols. 6-7	cols. 41-42
*		and cols. 40-39	
col. no.	xx	cols. 8-9	cols. 43-44
*	±	and cols. 38-37	
**real amount	0xxx.xxxxx	cols. 10-18	cols. 45-53
	±		
**imag. amount	0xxx.xxxxx	cols. 19-27	cols. 54-62
common "x"	x	col. 28	col. 7

*NOTE: Punch field in reverse.
**NOTE: Punch the algebraic sign over the last position and gang punch a "0" in the first position.

2. Hold the original key punched cards aside.
3. Select all cards where the row (columns 41-42) equals the column (columns 43-44) and reproduce 80-80, gang punching a common x-44 into the extra deck.
4. Combine all reproduced cards and sort to row order within the problem.

Item	*Card Cols.*	*Type*
problem no.	1-5	major
row	41-42	minor

5. Summarize the cards by row on an accounting machine. Group indicate the problem number into columns 1-5 and columns 80-76, and the row into col-

*This paper was presented by title.

umns 41-42 and columns 40-39 of the common x-7 summary cards. Punch the summations of real and imaginary terms into the summary cards (real amount in columns 45-53; imaginary amount in columns 54-62). Hold the detail cards. Summary cards are used in step 6.

6. The summary cards are to be used as check column cards. Gang punch a column number "0 0" into columns 43-44 and 38-37 of all summary cards.
7. Combine the detail cards from step 5 with the summary cards from step 6 to form a complete deck ready for reduction.

Reduction Procedure

1. Sort all cards in reverse order by row (columns 41-42).
2. Select the last row cards from the front of the pack. Hold the unselected rows until step 4.
3. Pull all x-44 cards from selected row cards, and place the x-44 cards with the unselected cards of step 2.
4. Place the nx-44 cards (last row) from step 3 in front of and face to face with the cards from step 2 and the x-44 cards from step 3 (top edges up).
5. Sort all cards to column order within the problem number.

Item	*Card Cols.*	*Type*
problem no.	1-5	major
column	43-44	minor

6. Detail x-7 gang punch and master x-74 compare (master card is x-7 card which has been reversed by step 4). Gang punch:

Item	*From*	*Into*
*imag. amount	cols. 19-27	cols. 19-27
*real amount	cols. 28-36	cols. 28-36

compare

Item	*From*	*Against*
problem no.	cols. 1-5	cols. 1-5
column	cols. 43-44	cols. 43-44
imag. amount	cols. 19-27	cols. 19-27
real amount	cols. 28-36	cols. 28-36

*NOTE: These fields are read in reverse order.

7. Select all x-74 gang punch masters and hold them aside.
8. Select the last column cards from each problem by sorting on column. Hold the unselected column cards until step 15.
9. Multiply the selected last column cards on a type 602-A calculating punch wired to compute:

$M = ge + hf$ ± xxxx.xxxxx cols. 10——18

$Ni = [gf - he]\ i$ ± xxxx.xxxxx cols. 63——71

from the cards with the fields as indicated in Figure 2.

10. Reproduce the multiplied cards as follows:

Item	*From*	*Into*
problem no.	cols. 1-5	cols. 1-5
M	cols. 10-18	cols. 45-53
working spots	cols. 37-44	cols. 37-44
Ni	cols. 63-71	cols. 54-62
problem no.	cols. 76-80	cols. 76-80
common "x"	col. 7	col. 9

Hold the common x-7 cards aside. The common x-9 cards are now in the original card form except for the common "x".

11. Select all x-44 cards from the common x-9 cards.
12. Place the x-44 cards in front of and face to face (top edges up) with the nx-44 cards and sort on problem number (columns 1-5).
13. Multiply the step 12 cards on a type 602-A calculating punch wired to compute:

$S = \dfrac{M}{g^2 + h^2}$ ± xxxx.xxxxx cols. 10——18

$Ti = \dfrac{Ni}{g^2 + h^2}$ ± xxxx.xxxxx cols. 63——71

from the cards with the fields as indicated in Figure 3, page 140.

14. Select the x-72 masters and hold them aside.
15. Place the x-9 cards ahead of the x-7 detail cards from step 8 and sort to row order within the problem.

Item	*Card Cols.*	*Type*
problem no.	1-5	major
row	41-42	minor

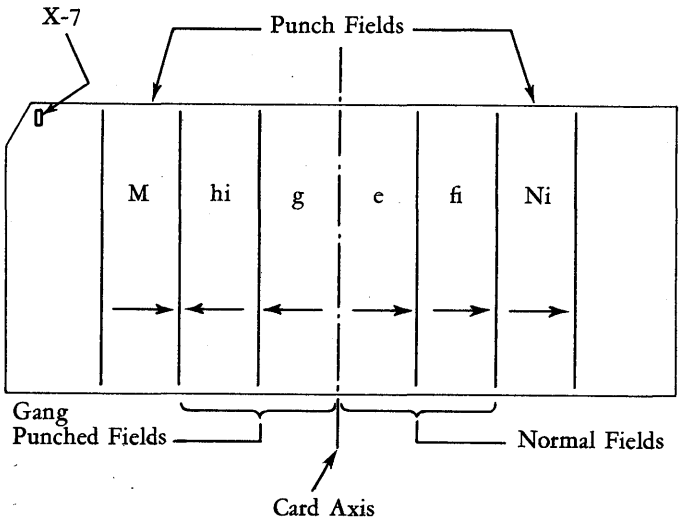


FIGURE 2

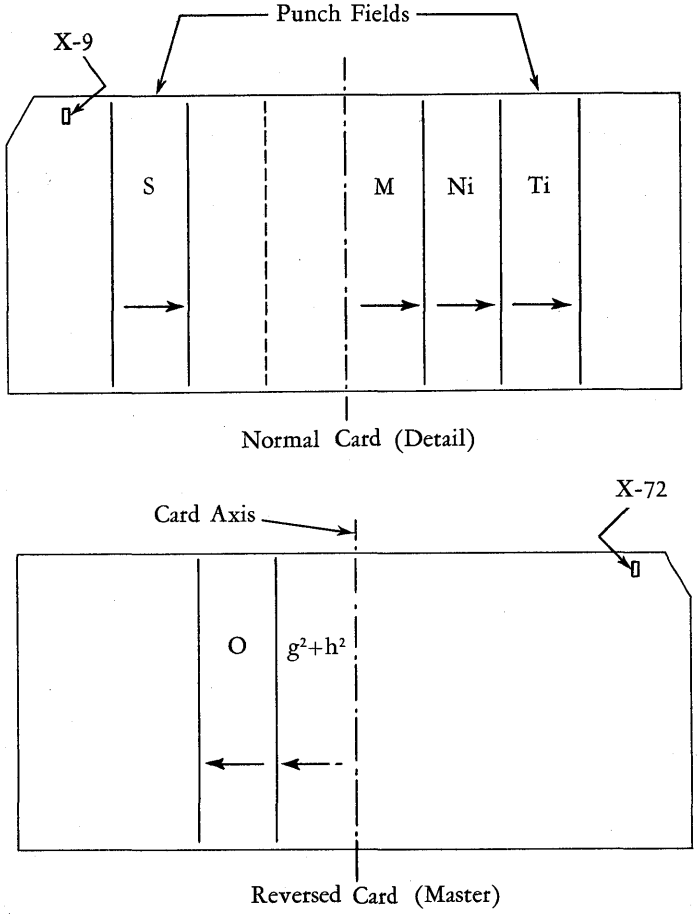


FIGURE 3

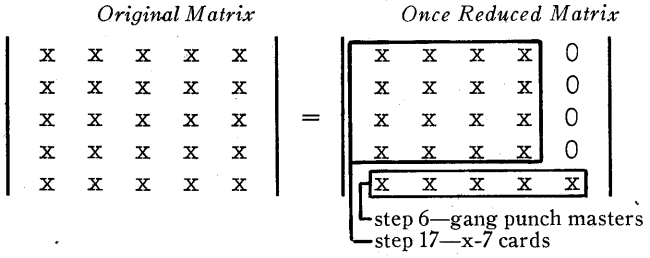
16. Multiply the step 15 cards on a type 602-A calculating punch wired to compute:

$$a' = a - Sc + Td \qquad \begin{matrix}\pm \\ \texttt{xxxx.xxxxx} \\ \text{cols. } 10\text{——}18\end{matrix}$$

$$b'i = [b - Sd - Tc]\ i \qquad \begin{matrix}\pm \\ \texttt{xxxx.xxxxx} \\ \text{cols. } 63\text{——}71\end{matrix}$$

from the cards with the fields as indicated in Figure 4.

17. Select the x-9 row masters and hold them aside. The remaining x-7 detail cards comprise a matrix with all values in the last column reduced to zero, except the value in the last row when combined with the column gang punch masters used in step 6, in the table below.



Check Procedure

The arithmetical operations may be checked as follows:

1. Select all the column "0 0" check cards from the step 17, x-7 detail cards.
2. Tabulate the unselected x-7 cards by row, summing the a' and $b'i$ fields separately.
3. Compare the totals of each row with the a' and $b'i$ values in the column "0 0" cards for that same row. They must compare.
4. Errors may be corrected by using the general formula given below:

$$[a' + b'i] = [a + bi] - [c + di]\,\frac{[e + fi]}{[g + hi]}$$

where: $[a + bi]$ is any complex element in the matrix outside of the last row or last column.

$[c + di]$ is the complex element in the last row and the same column as $[a + bi]$.

$[e + fi]$ is the complex element in the last column and the same row as $[a + bi]$.

$[g + hi]$ is the complex element in the last row and the last column.

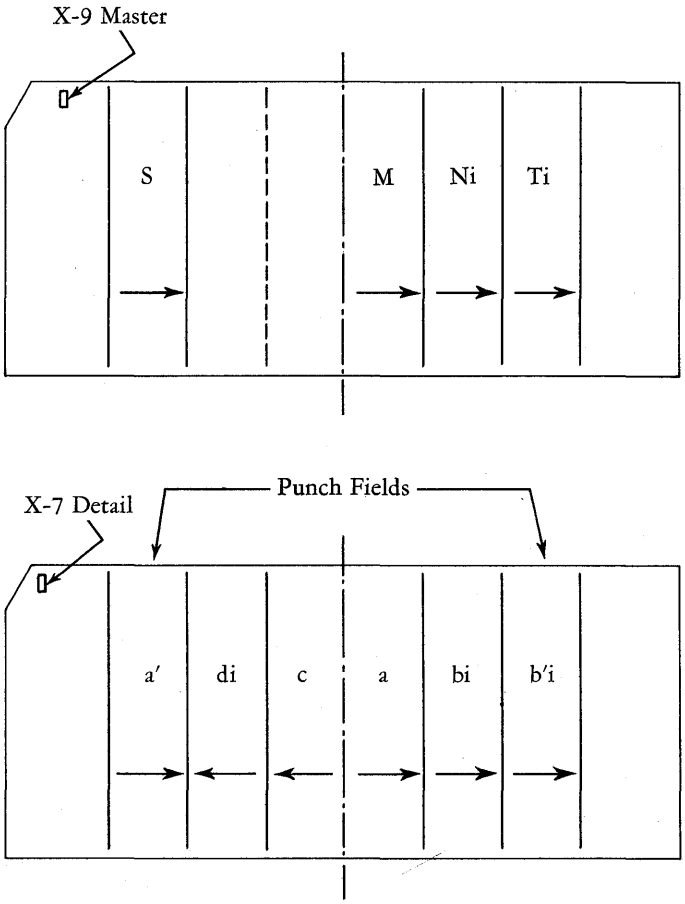


FIGURE 4

*Simplification of Statistical Computations as Adapted to a Punched Card Service Bureau**

W. T. SOUTHWORTH J. E. BACHELDER

State College of Washington

THE punched card service bureau at the State College of Washington will be used, in this paper, as an example of how the statistical computational needs of a college or university may be obtained as a by-product of an installation established primarily for servicing the needs of the administrative offices.

By standardizing and simplifying punched card statistical methods, training of instructors and students, and organizing time schedules, many statistical computations may be processed mechanically by the person who will analyze the results.

In almost all of the educational institutions, installations are either leased by an individual division of the institution or leased as a self-supporting service bureau. In the former, it would be a rare case when one division alone could reconcile the cost in maintaining enough of the right type of equipment and personnel to service them efficiently. In the latter, too much time is wasted in accounting for the time spent by the operators on each type of equipment used for a scheduled operation in order that appropriate charges may be made. This, in institutional work, is "taking money out of one pocket and putting it into the other." Most important in either case is the loss of the flexibility in using the punched card records of one department to develop, assist or maintain the records of another department.

The example installation experimented by giving the director a working budget to cover rental, salaries, supplies and all incurred expenditures. This budget is reconciled, in statement form each year, by the director to the business manager, in the form of tangible savings in salaries directly due to punched card services and intangibles such as faster service, accuracy of records, and services performed that could not have been accomplished by manual methods.

The important point is that the director has complete control of all proposed applications, and can use one department's records to aid another's. As a result of this organization the volume of the punched card application has expanded to the extent where up-to-date models of almost all types of punched card equipment and experienced personnel can be maintained and their costs reconciled. Thus, there are few limits, if any, to the types of application they can undertake.

The bureau has the following IBM equipment:

1 completely equipped alphabetic accounting machine
80 counters, 88 type bars, automatic carriage, 6 class selectors, 2 digit selectors, progressive totals, card cycle transfer device, and net balance.
1 summary punching reproducer (80-column comparing)
1 alphabetic interpreter
1 collator
1 card counting sorter
1 electronic sorter (high speed)
2 alphabetic duplicating key punches
1 numerical verifier

Personnel of the bureau is as follows:

A director to survey new applications, keep procedures up-to-date, direct administrative details, instruct new personnel, and instruct in courses teaching punched card technique.
An operator-supervisor to operate equipment and supervise all machine work, and furnish guidance to students and instructors processing their own studies.
One additional operator
Two key punch operators
One part-time employee to fill in "peak loads" in either machine operation or key punching.

Work load of the bureau is divided about as follows:

President's office, for personnel accounting, 5%
Accounting office, 40%
All phases of payroll for 3,500 part- and full-time employees
Budget reports (encumbered) to each department of the college and associated financial reports and statements.
Fee distribution
Billing of veterans' fees
Annuity reports
Student damage claim and deposit statements and checks
Withholding tax annual statements
Registrar's office, 25%
All phases of student records completely converted to punched cards including class lists, drop and add control, grade reports, posting of permanent records, academic standings, and innumerable student statistical reports.
Statistical work—administrative requirements, 5%
This covers many varied types of operations which are processed completely by the staff of the service bureau.
Housing and food service, 10%
4,200 income accounts
Dormitory listings
Expense accounting

*This paper was presented by title.

Statistical computations for professors, instructors and graduate student research, 10%

This covers guidance time expended, key punching of cards, and machine time allowed. The instructors or graduate students operate the equipment and process their own studies. To accomplish the above, methods had to be established to simplify and standardize the more common statistical requirements. The balance of this paper is devoted to presenting some of these methods, many of which are generally known but seldom used on account of lack of cooperation between research and punched card technicians within the institution.

The most important simplification of the bureau is the instruction given to staff members and graduate students who anticipate processing their research project or thesis with the aid of punched cards.

A two-credit-hour course is offered each semester, including one hour of lecture and three hours of evening laboratory in the service bureau. This course is offered to all staff members and students with a prerequisite of one year of statistics. It is the purpose of this course, however, to attract the persons who will be most likely to have need of punched card services in the future. The outline of the course is as follows:

A. The Theory of the Punched Card
   1. The card
   2. The purpose of each machine
   3. Multiple punching

B. Statistical Applications
   1. Sampling techniques
   2. Construction of schedule or questionnaire
   3. Coding
   4. Data checking
   5. Cross-tabulations

C. Operation and Wiring of Punched Card Equipment
   1. Key punch and verifier
   2. Sorters
   3. Interpreter
   4. Reproducer
   5. Collator
   6. Alphabetic accounting machine
   7. Brief explanations of the functions of punched card calculators, i.e., multiplier, electronic calculator and statistical machine.

D. Statistical Applications and Their Related Machine Procedures
   1. Frequency distribution
   2. Matrix distribution
   3. Correlations
   4. Analysis of variants
   5. Chi square
   6. T ratios
   7. Standard deviation

E. Planning Research Projects to Utilize Punched Card Equipment
   1. Hypothesis
   2. Measurement
   3. Format preparation
   4. Coding problems
   5. Punching problems
   6. Analysis

A professional library of many of the latest books and publications on punched card applications is maintained by the service bureau and is available to all interested persons at the institution.

The Washington Public Opinion Laboratory at the State College has opened its doors to all persons entering into the organizational stages of a statistical research study to offer advice on sample design interviewing methods, or schedule design. This office is proficient in the art of adapting the forms used to an efficient document, for transfer to punched cards by key punching. As completed data are returned from the field they are transferred to punched cards by the key punch operators of the bureau. The director or supervisor of the bureau then furnishes the necessary technical guidance for the student to process his own study mechanically.

The Sum of Products and/or Squares

The bureau's method in establishing a standard procedure is to compute all data entering into the many types of statistical formulas using ΣX, ΣX_1^2, $\Sigma X_1 X_2$, $\Sigma X_1 X_3$, etc. The equipment required is:

1. A permanently wired control panel as illustrated in Figure 1.
2. A standard sorter.
3. An 80-counter alphabetical accounting machine.

For the operating theory see Figure 2, page 144.

Sample Procedure

Operating procedure for multiple correlation of 9 three-digit variables X_1, X_2, through X_{10} using 1000 samples (raw data formulas will be used).

1. Inasmuch as there are many methods of creating the cards to be used (transferring from another punched card file by reproducer, key punching, etc.), we shall start with the completed detail cards illustrated in Figure 3, page 145, assuming that the data have been checked for a straight line of regression.
2. Sort all detail cards on the units position of each variable to reject all missing variables. These must either be adjusted to the average of the others or deleted from the study.
3. Assuming that none of the variables will exceed 450, digit cards, consecutively numbered from 000 to 450 in each of the 9 variable fields, will be required. Almost every punched card installation maintains a consecutively numbered set of punched cards for many varied applications. Reproduce the consecutive number file to blank cards, transferring the one consecutive number field to the 9 variable fields by split wiring. A designating x punch is gang punched into column 80 of the card. This operation is handled by the staff of the bureau and will consume no more than fifteen minutes.
4. Placing the digit cards in front of the detail cards, sort the X_1 variable field in descending sequence. To sort cards in descending sequence, the cards are sorted on the units, tens, and hundreds positions the same as for ascending sequence, but the cards are removed from the pockets of the sorter in reverse sequence (9's in front of 8's, 8's in front of 7's, etc.).
5. The permanently wired control panel shown in Figure 1 is maintained at all times in the bureau. To this control panel, wire variable fields X_1, X_2, X_3, X_4, X_5 from the add brushes to the X_1, X_2, X_3, X_4, X_5 entry fields of the right-hand panel.

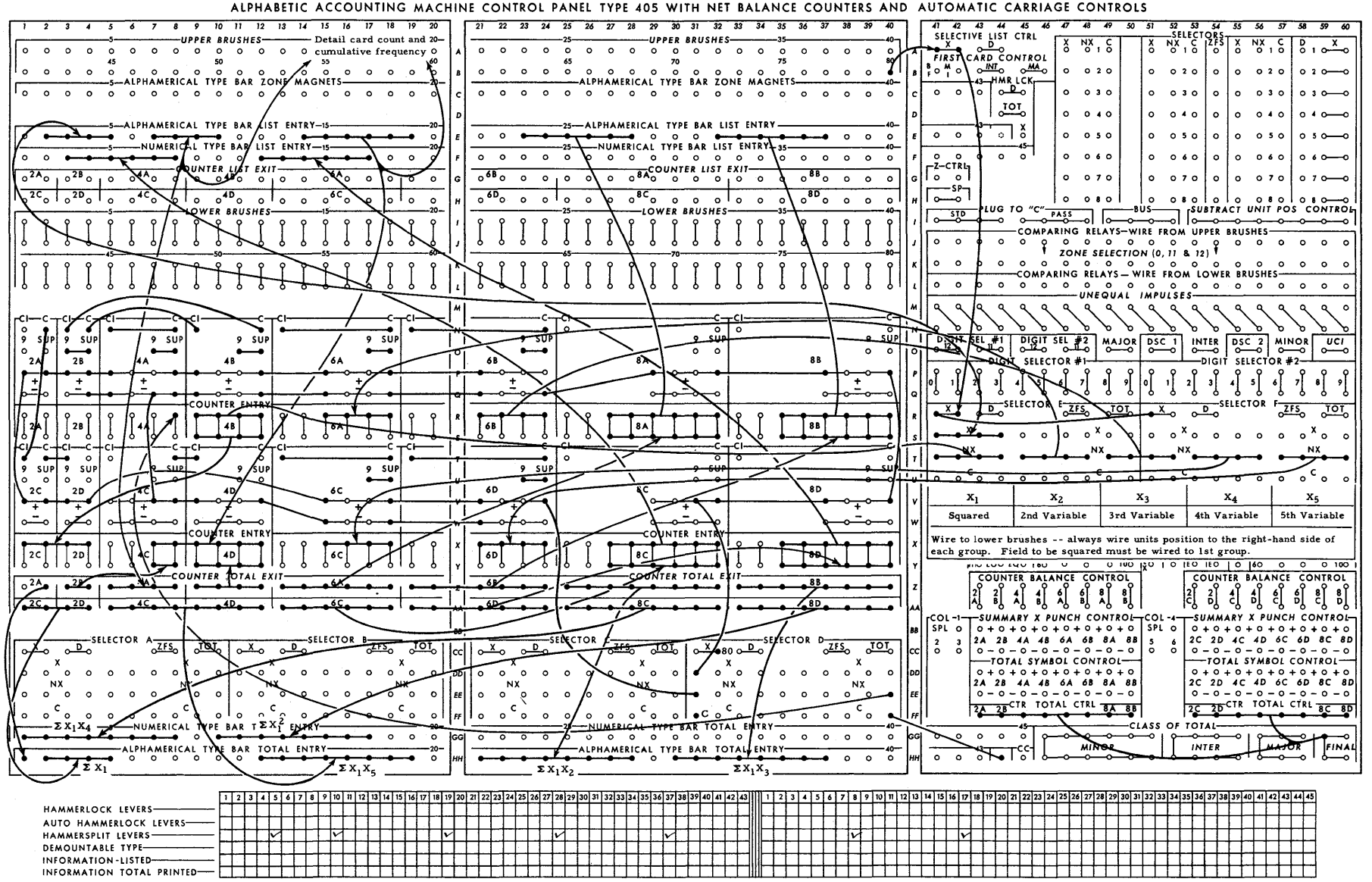


FIGURE 1. SUMS OF SQUARES AND PRODUCTS (5 VARIABLES)

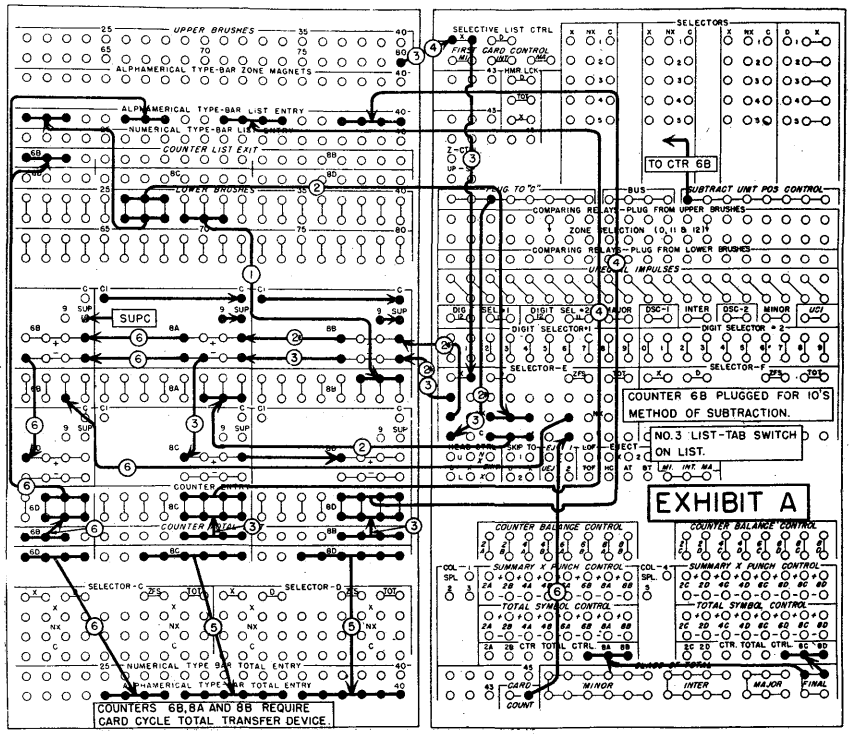
SUM OF PRODUCTS AND/OR SQUARES BY CARD CYCLE TOTAL TRANSFER

In several types of mathematical and statistical calculations it is desired to obtain the sum of products or squares. Normally for such calculations first the individual multiplications are carried out or the squares are calculated and then their sum is obtained by adding up the individual products, although actually only the sum of the products is desired. For example when an inventory is taken the quantity of each item is multiplied by its unit price and the sum of the products is performed to arrive at the total value of the items. In psychological tests, when calculating the standard deviation, only the sum of squares has to be obtained. For correlation analysis the sum of the squares and products of scores (rating of individuals) are needed as ΣX_1^2, $\Sigma X_1 X_2$, $\Sigma X_1 X_3$, $\Sigma X_1 X_4$, etc.. Several IBM methods are known by which the total of products or squares is obtained on Electric Accounting Machines without carrying out the individual calculations. Such methods, as digiting, progressive digiting and digiting without sorting, are also discussed in Pointers. Generally one run through the Accounting Machine is required for each position of the multiplier for calculating the sum of products or squares. The partial totals obtained have to be added up in order to secure the final total.

The sum of squares and or products can be obtained by a single run of the IBM cards through the Type 405 Accounting Machine. The cards are sorted together with a set of X-punched "digit cards" in descending order (from highest to lowest) on the field of the multiplier (or field to be squared). If the multiplier field consists of one column, 9 digit cards (9-1), two columns 99 digit cards, three columns 999 digit cards, etc., must be sorted in. If there is no detail card for a number, at least a digit card for that number will be present. After the file of cards is sorted, all digit cards preceding the first detail card are removed. The last card of the file will be the digit card number "1". The cards are tabulated on a Type 405 Accounting Machine and totals accumulated by multiplier group are card cycle total transferred from one Counter-Group into another to obtain the final total of squares and/or products. By this procedure different sums of squares and/or products may be obtained simultaneously. This is especially important for correlation analysis.

Exhibit A shows a plugging for obtaining the sum of squares of variable X_1 and the sum of products of variables X_1 X_2. Two Counter-Groups are necessary to obtain each sum. One of each pair of Counter-Groups must be equipped with Card Cycle Total Transfer Device. The multiplicand X_2, (Columns 29-31) is plugged directly to Counter-Group 8B (plugging 1). Columns 26-28 (as multiplicand) are plugged to Counter-Group 8A through the NO-X side of Selector E in order to prevent accumulating from digit cards (plugging 2). The add and subtract impulses are also under control of this Selector (plugging 2*). When a digit card passes the upper brushes the X in column 80 sets up Selector E so that when this card passes the lower brushes Selector E is in controlled position and accumulation from this digit card is eliminated. When a digit card passes the lower brushes Counter-Groups 8A and 8B subtract and the totals standing in them are transferred to Counter-Groups 8C and 8D respectively which add these totals (plugging 3). Simultaneously, these totals may be listed (plugging 4). After the transfer, Counter-Groups 8A and 8B (equipped with Card Cycle Total Transfer Device) will contain the same figures that they did before the transfer. If there are no detail cards but only digit cards present for a group, the totals transferred for the previous group will be transferred again for this group. After all cards have passed through the machine, Counter-Group 8C contains the sum of squares (ΣX_1^2), and Counter Group 8D the sum of products ($\Sigma X_1 X_2$). These totals are printed as final totals (plugging 5). The final figures standing in Counter-Groups 8A and 8B are the totals of the single items (ΣX_1 and ΣX_2) accumulated from the fields 26-28 and 29-31 respectively. These totals are the last items listed by plugging 4. Counts by multiplier group will be obtained in Counter-Group 6B and total counts in Counter-Group 6D (plugging 6). Counter-Group 6B must also be equipped with Card Cycle Total Transfer Device for restoring itself after each transfer. Exhibits B and C illustrate an example for this application.

STUDENT SERIAL NO.	INTELLIGENCE X_1	SOCIAL STUDY X_2	NATURAL SCIENCE X_3	LANGUAGES X_4	MATH X_5	HISTORY X_6	CHECKING TOTALS S
001	39	38	39	40	17	39	212
002	15	31	21	18	15	29	129
003	08	32	29	38	20	41	168
EXHIBIT B							
077	16	36	18	24	13	35	142
078	04	35	20	35	11	53	158



INTELLIGENCE	NO. STUDENTS IN GROUP	ΣX_1^2	PRODUCTS OF VARIABLE ONE SCORE WITH SCORES OF THE REMAINING FIVE VARIABLES AND WITH THE SUMS OF ALL SIX VARIABLES. $\Sigma X_1 X_2$	$\Sigma X_1 X_3$	$\Sigma X_1 X_4$	$\Sigma X_1 X_5$	$\Sigma X_1 X_6$	$\Sigma X_1 S$
39	2	78	74	65	74	28	86	405
38	2	154	152	122	144	59	165	796
37		154	152	122	144	59	165	796
36	1	190	186	142	178	70	213	979
35	1	225	224	167	209	82	262	1169
34		225	224	167	209	82	262	1169
33	5	390	394	278	364	150	475	2051
32	4	518	528	347	472	212	586	2663
31	3	611	621	432	581	261	720	3226
30	7	821	854	578	818	345	1001	4417
29	1	850	887	595	855	356	1029	4572
28		850	887	595	855	356	1029	4572
		850	887	595	855	356	1029	4572
EXHIBIT C								
8	2	1619	2347	1479	2218	958	2648	11269
7		1619	2347	1479	2218	958	2648	11269
6		1619	2347	1479	2218	958	2648	11269
5		1619	2347	1479	2218	958	2648	11269
4	6	1643	2545	1623	2424	1037	2880	12152
3	2	1649	2615	1663	2478	1067	2948	12420
2		1649	2615	1663	2478	1067	2948	12420
1		1649	2615	1663	2478	1067	2948	12420
	78	43163	55537	35854	52913	22398	63277	273142

FIGURE 2

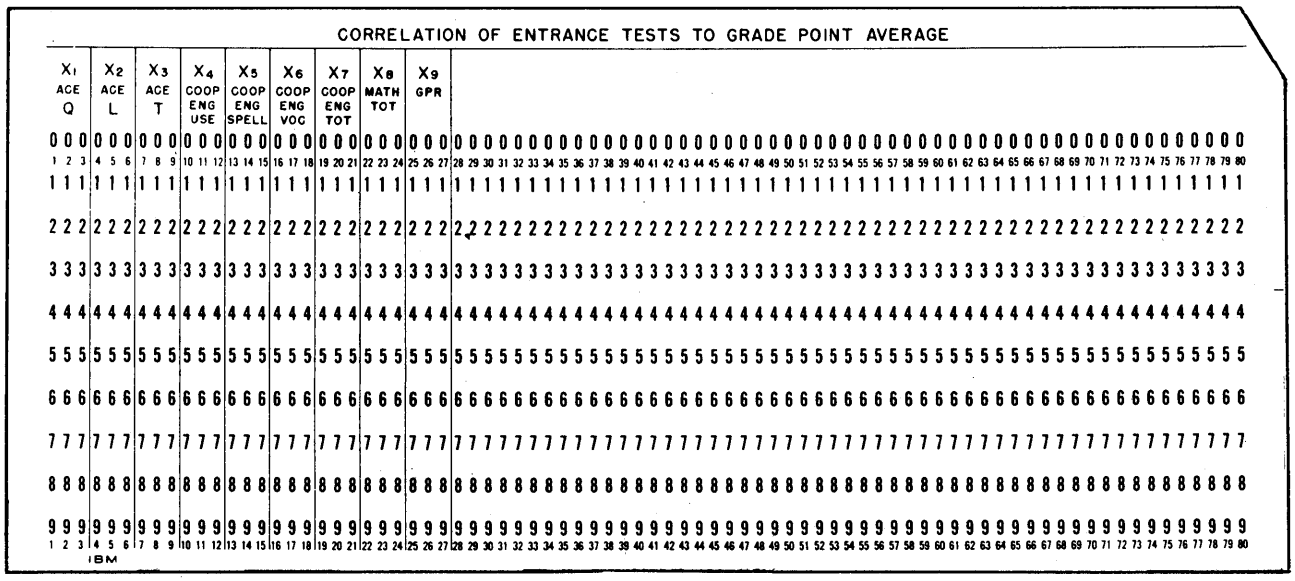


FIGURE 3

Remove all digit cards preceding the first detail card and tabulate on an IBM Type 405 Alphabetical Accounting Machine. The following results are obtained:

a. Count of detail cards
b. ΣX_1
c. ΣX_1^2
d. ΣX_1X_2
e. ΣX_1X_3
f. ΣX_1X_4
g. ΣX_1X_5

Post these results to the first horizontal line opposite X in the correlation chart illustrated in Figure 4, page 146.

6. Change wires in accounting machine control panel as follows:

Variable field 6 to control field 2
" " 7 " " " 3
" " 8 " " " 4
" " 9 " " " 5

Tabulate cards the second time in the same sequence. The following results are obtained:

a. C.C. of detail cards
b. ΣX_1
c. ΣX_1^2
d. ΣX_1X_6
e. ΣX_1X_7
f. ΣX_1X_8
g. ΣX_1X_9

7. Sort column 80, separate digit and detail cards. Place digit cards in front of detail cards and sort the variable X_2 field in descending sequence, the same as the variable X_1 field was sorted.

8. Change wires in control panel as follows:

Variable field 2 to control field 1
" " 3 " " " 2
" " 4 " " " 3
" " 5 " " " 4
" " 6 " " " 5

Remove all digit cards preceding the first detail cards and tabulate. The following results are obtained:

a. Card count of detail cards
b. ΣX_2
c. ΣX_2^2
d. ΣX_2X_3
e. ΣX_2X_4
f. ΣX_2X_5
g. ΣX_2X_6

Post these results to the second line of correlation chart.

9. Change wires as follows:

Variable field 7 to control field 2
" " 8 " " " 3
" " 9 " " " 4
" " 1 " " " 5

Tabulate with cards in same sequence. The following results are obtained:

a. C.C. of detail cards
b. ΣX_2
c. ΣX_2^2
d. ΣX_2X_7
e. ΣX_2X_8
f. ΣX_2X_9
g. ΣX_2X_1

Check items 1, 2, 3 to the posting of the second line of correlation chart. Complete posting of the second line with items 4, 5, 6, 7. Check posting of X_2X_1 on the second line with posting of X_1X_2 on first line for verification.

10. Sort, tabulate, and post the results of each variable X_3 through X_9 the same as X_1 and X_2 to complete table shown in Figure 4. A study of this table will readily reveal the ease with which the data may be substituted in the various multiple correlation formulas by the calculator. No attempt, at this institution, has been made to compute the formulas themselves by punched cards.

Advantages of the foregoing methods are:

1. Accuracy
   a. All sums of cross products are totaled twice, in separate machine operations, with the cards in a different sequence and through separate machine circuits. There can be no doubt of the accuracy if all cross checks (X_1X_3 against X_3X_1, etc.) are in balance.
   b. Card counts for each tabulation prove cards are all present.
   c. The sums of each variable, which are wired to a control field, are produced at the end of each run.

CORRELATION TABLE

	Card Count	ΣX	ΣX^2	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
X_1	1000	463451	56321543	—	74132158	10568215	46321789	18342586	42798546	14358261	72143688	56182153
X_2	1000	786543	74342156	74132158	—	46385215	46315784	70560985	43780542	95865021	98532142	13478532
X_3	1000	256852	40350985	10568215	46385215	—	78942153	43851278	15790432	47832157	40305890	78952178
X_4	1000	678178	78543218	46321789	46315784	78942153	—	74321586	47851235	49521587	10325178	78543102
X_5	1000	465785	25678154	18342586	70560985	43851278	74321586	—	47351251	38742105	40321578	42035421
X_6	1000	135781	21578521	42798546	43780542	15790432	47851235	47351251	—	20531542	36152578	46321425
X_7	1000	89432	36421587	14358261	95865021	47832157	49521587	38742105	20531542	—	36421503	10678154
X_8	1000	76587	10320567	72143688	98532142	40305890	10325178	40321578	36152578	36421503	—	37421578
X_9	1000	94378	43215872	56182153	13478532	78952178	78543102	42035421	46321425	10678154	37421578	—

FIGURE 4

This affords many checks for the sum of each variable.

d. The sum of the squares of each variable is totaled twice, but each time in the same machine circuit. To check the machine accuracy (thoroughly), one variable is split-wired to each of the five control fields, and a test deck is tabulated. If all results are the same, it is a good proof that the machine is operating properly. It is recommended that a test deck of approximately 200 cards be maintained with the control panel to test the machine, before tabulations are started and again when completed.

2. Speed of Operation. The illustrated problem can be completely processed to the completion of the correlation table in a little less than eight hours.

3. The procedure is simple enough in structure that persons not trained in punched card techniques can process their own studies with a minimum of guidance.

4. Each digit card, as it passes through the machine, forces the totals standing in each counter to list. This automatically produces a cumulative frequency for each variable.

4- and 5-Digit Variables

Since one digit card is needed for each progression, obviously it is not feasible to process 4- or 5-digit variables in one run through the machine. A 4-digit field is split into two 2-digit fields, a 5-digit field into 2 and 3. The resulting squares for each portion are then offset two places and added together.

This method of computing the sums of squares and the sums of cross-products has been used successfully at Washington State College for over a year. The wiring of the permanent control panel and the establishment of standard procedures have saved many hours of work.

Cross Tabulations

I am sure that most persons operating punched card installations in institutions are familiar with the methods of obtaining cross-tabulations through digit selection on the alphabetical accounting machine. It seems, however, that very few of them are using this method extensively because of the amount of time consumed in control panel wiring for each individual project or study. It has been proved that one permanent control panel could be wired in such a way that the changing of five wires would produce 90 per cent of all cross-tabulations requested. (Refer to Figure 5 for sample of a cross-tabulation.)

The wiring is very simple, but many wires are needed, thus making it profitable to maintain a permanent control panel to save the tedium and danger of rewiring for each study. A standard alphabetical accounting machine with one digit selector and 12 single-position X-distributors is required (10 single position X-distributors may be installed at no additional rental charge to replace the five standard three-position distributors.)

Wiring of Control Panel

The 12 outlet hubs of the digit selector are wired to the D inlet hubs of the 12 X-distributors. A card count impulse is wired to each of the C inlet hubs of the 12 X-distributors. The top row or A and B counters are coupled together to form one 40-position counter. The lower row or C and D counters are coupled to form a second 40-position counter. The controlled side or X hubs of each of the 12 X-distributors are wired to the top counter row, allowing three adding positions for each. This leaves four positions for the total. Each of the 12 counter entries is coupled to the corresponding entries of the lower counter row. The counter exits are then "total transferred" through the four standard class selectors and wired to the type bars for the printing of both minor and final totals. All of the above wiring is done with permanent wires. Now it is only necessary to add one temporary wire from one of the 80 upper brushes to the digit selector entry hub. This wire will be changed as required for the column to be distributed. If controls are required, two temporary wires are used for each column to be controlled.

Advantages of the above method include:

1. The few temporary wires to be changed to suit the particular study can be explained to a person, not familiar with the equipment, in a very few words. Again, this leads us to our goal of allowing novices in the field of punched cards, but experts in statistics, to process their own studies economically and safely.
2. Automatic totals are created for both horizontal and vertical breakdowns. This is especially important for analysis by chi square.
3. Increased accuracy through the saving of hand posting from sorter counters.
4. In almost all cases where many breakdowns are required, the accounting machine method is a tremendous saving in operator-time.

Cross Tabulations by Sorter

In cases where volume is large and breakdowns are few, the use of a card counting sorter and hand posting is more

CROSS TABULATION BY ACCOUNTING MACHINE

	1	2	3	4	5	6	7	8	9	0	X	Y	Totals
1	5	8	22	45	66	14	22	5	3	5	1	2	200
2	52	46	20	32	15	18	5	4	9	7	0	0	213
4	175	5	6	8	9	4	10	20	50	45	5	8	247
7	121	276	3	5	10	100	80	70	42	12	21	18	759
8	5	8	9	4	3	5	7	9	12	420	5	2	489
11	9	10	15	4	20	40	10	6	8	2	0	0	124
95	12	8	4	100	5	21	6	8	4	0	5	2	177
96	5	14	12	15	80	20	4	3	8	1	4	3	171
97	120	105	40	21	82	10	9	8	2	5	1	2	407
99	5	3	0	0	5	4	1	2	0	1	0	1	28
	682	405	200	195	325	425	275	207	152	725	95	82	3768*

Figure 5

economical. For this type of study the service bureau supplies standard forms on which to post cross-tabulations from sorter counts.

Random Sample Selection

A good many times in the last two years this bureau has been requested to select a random sample or every *N*th card from a punched card file. This can be accomplished easily on the collator, with a card counting device, by wiring to select every third, fifth, or tenth card as the case requires.

The bureau, not needing the card counting device on the collator, maintains a deck of 400 cards in which column 2 is *X*-punched in every other card, column 3 is *X*-punched in every third, column 4 in every fourth, column 5 in every fifth, etc., to every 40th card being *X*-punched in column 40. A consecutive number from 1-400 is also punched to guarantee sequence. By placing this deck of cards in the secondary feed and the cards from which the sample is to be selected in the primary feed, any combination of every other, through every 40th card, may be selected by wiring from the column required in the secondary brushes to the secondary *X*-pickup. Once prepared, this file may be used over and over again and will serve this purpose as well as the card counting device.

It is not the purpose of this paper to bring new discoveries to light, but to illustrate that at least 75 per cent of the statistical computations being processed every day in the modern institutions of higher learning can be simplified and standardized by punched card methods with the cooperation of the individual punched card technician.

*Forms of Analysis for Either Measurement or Enumeration Data Amenable to Machine Methods**

A. E. BRANDT

Atomic Energy Commission

A COMMON PROBLEM in statistics is to estimate, from a sample, the parameters of a population. If the population is normally or not exceedingly anormally distributed, the mean and the standard deviation or the first and second moments will adequately describe or specify the population. For example, suppose we have a sample of 12 observations drawn at random from a given population. The mean and the standard deviation, or what is simpler to use, the variance of the population can be estimated from this sample.

A convenient method of securing these estimates is based on a matrix-vector product. To illustrate, we shall use the random sample of 12 observations previously mentioned. Since the sample contains 12 observations, a 12 by 12 matrix will be required. In order to utilize all of the information in the 12 observations in estimating the population mean and variance, an orthogonal matrix must be used. Such a matrix can always be written by following a few simple rules. First, the terms of the first row of the matrix may always be written as a series of plus ones. Second, the sum of the terms in each row after the first must be zero. Third, the sums of products of corresponding terms in all possible pairs of rows, omitting the first, must be zero. It will be noted that, under the conditions just stated, if the terms of each row of the table are divided by the square root of the sum of squares of the terms in that row, the sum of the squares of the resulting terms in each row and column will be unity.

The number of such matrices that can be written is very large. For example, two such matrices are presented in Table I, page 150.

It is especially important to notice that the first row of the matrix provides an estimate of the mean or first moment, and that the remaining 11, or 12 minus 1, rows are available for estimating the variance or the second moment. In general, an n by n matrix of this sort and a single column vector of n rows furnish $n - 1$ comparisons on which to base an estimate of variance. I have found this presentation very useful for convincing students and research workers that the use of degrees of freedom in estimating a variance is not a matter of choice or what school of statistics one follows but of mathematical rigor. They readily recognize the absurdity of introducing an arithmetic error by dividing the sum of $n - 1$ quantities by n instead of $n - 1$ to obtain their mean. This use of $n - 1$ in the divisor is independent of the magnitude of n, so long as n is finite, although the difference between the approximate value obtained by dividing by n and the precise value obtained by dividing by $n - 1$, decreases as n increases.

The set of values used in the above example was designated a random sample. This means that no restrictions were imposed on the drawing process or, in other words, that the probability that any value in the group or population from which these 12 were drawn had exactly the same probability in a given drawing of being drawn in that drawing, as did every other value in the group. The example above is of slight interest to workers conducting critical experiments, because in the design of such experiments, restrictions upon randomness are deliberately imposed.

For instance, let us suppose that the 12 values used above resulted from an experiment on the elimination of weeds from a crop such as flax by spraying with chemicals. In designing the experiment we knew that two chemicals showed definite value as differential sprays, that is, they would kill weeds but not injure the crop when sprayed on a field. Reports on which was the better were conflicting, and considerable doubt remained as to the best period in the growing season for spraying. Let us suppose that our chief interests center in which is the better spray and whether this superiority is constant for the different spraying periods. Thus, if we spray when the crop is one-quarter grown, is one-half grown, and is three-quarters grown, and use both chemicals at each spraying, six plots will be required. If two replicates are used, 12 plots will be needed. If 12 plots of appropriate size and shape are located in a field of the crop to be studied, the combination of time and chemical to be assigned to each may be determined without restriction by some scheme of randomization such as the use of random numbers. Experience has shown that the

*This paper was presented by title.

TABLE I

MATRIX-VECTOR PRODUCT METHOD FOR CALCULATING MEAN AND VARIANCE

A

12 × 12 Matrix												Vector	Divisor	Matrix Vector Product	Square of M-V Prod. Divided by Divisor	M-V Prod. Divided by Divisor
+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	1	12	51		4.25
+1	−1	0	0	0	0	0	0	0	0	0	0	2	2	−1	0.50000	
+1	+1	−2	0	0	0	0	0	0	0	0	0	3	6	−3	1.50000	
+1	+1	+1	−3	0	0	0	0	0	0	0	0	4	12	−6	3.00000	
+1	+1	+1	+1	−4	0	0	0	0	0	0	0	5	20	−10	5.00000	
+1	+1	+1	+1	+1	−5	0	0	0	0	0	0	4	30	−5	0.83333	
+1	+1	+1	+1	+1	+1	−6	0	0	0	0	0	5	42	−11	2.88095	
+1	+1	+1	+1	+1	+1	+1	−7	0	0	0	0	6	56	−18	5.78571	
+1	+1	+1	+1	+1	+1	+1	+1	−8	0	0	0	7	72	−26	9.38889	
+1	+1	+1	+1	+1	+1	+1	+1	+1	−9	0	0	6	90	−17	3.21111	
+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	−10	0	5	110	−7	0.44545	
+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	−11	3	132	15	1.70455	
															34.24999	

$$\text{Variance} = \frac{\text{sum of quotients}}{\text{one less than number of rows in vector}} = \frac{34.24999}{11} = 3.1136$$

B

+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	1	12	51		4.25
+1	−1	0	0	0	0	0	0	0	0	0	0	2	2	−1	0.5	
0	0	+1	−1	0	0	0	0	0	0	0	0	3	2	−1	0.5	
0	0	0	0	+1	−1	0	0	0	0	0	0	4	2	+1	0.5	
0	0	0	0	0	0	+1	−1	0	0	0	0	5	2	−1	0.5	
0	0	0	0	0	0	0	0	+1	−1	0	0	4	2	+1	0.5	
0	0	0	0	0	0	0	0	0	0	+1	−1	5	2	+2	2.0	
+1	+1	−1	−1	0	0	0	0	0	0	0	0	6	4	−4	4.0	
0	0	0	0	+1	+1	−1	−1	0	0	0	0	7	4	−2	1.0	
0	0	0	0	0	0	0	0	+1	+1	−1	−1	6	4	+5	6.25	
+1	+1	+1	+1	0	0	0	0	−1	−1	−1	−1	5	8	−11	15.125	
+1	+1	+1	+1	−2	−2	−2	−2	+1	+1	+1	+1	3	24	−9	3.375	
															34.250	

$$\text{Variance} = \frac{34.25}{11} = 3.1136$$

material, time, and money required for this experiment can generally be more efficiently used if certain restrictions are placed on the assignment of treatment to plots.

A restriction on randomness might be imposed by dividing the 12 plots into two compact groups or blocks of six plots each. In general the plots should be comparatively long and narrow, and the blocks should be square or as nearly so as practical. The six combinations of two chemicals and three spraying times can now be assigned by a scheme of randomization to the six plots in each block, but the two assignments must be separate and independent. Since we are chiefly interested in the comparison between the two chemicals, we should further restrict the method of assignment so that the two chemicals will always appear next to each other. This is done by assigning the three times of spraying at random to three main plots, of two subplots each, in the two blocks, and then assigning the two chemicals to the two subplots in each main plot independently by some scheme of randomization, such as flipping a coin.

The effect of these restrictions on randomness is to reduce the number of matrices that can be used, compared to the vast number of matrices possible when no restrictions are imposed. One of this limited number of possible matrices is presented in Table II. In this table the observation vector, consisting of values representing the efficacy of treatment on each plot, is placed at the top so that its rows correspond to appropriate columns of the matrix; further, the design of the experiment—but not the field layout—and hence the analysis of the data is specified. Terms of +1 and −1 are represented by + and − alone, and those of zero by blanks.

It is at once evident from the examples presented that in this form of analysis of research data, the number of computing steps is large and the number of data small. I under-

TABLE II

MATRIX-VECTOR PRODUCT METHOD OF DOING ANALYSIS OF VARIANCE CALCULATIONS

Block	I						II								
Time	early		middle		late		early		middle		late				
Chemical	a	b	a	b	a	b	a	b	a	b	a	b			
Observation Vector	1	3	4	2	5	4	6	5	7	6	5	3	*Divisors*	*Matrix-Vector Product*	*Square of M-V Prod. Divided by Divisor*	*M-V Prod. Divided by Divisor*
comparisons																
first moment																
mean	+	+	+	+	+	+	+	+	+	+	+	+	12	51		4.25
second moment																
main plots																
block B	−	−	−	−	−	−	+	+	+	+	+	+	12	13	14.0833	
time																
linear T_1	+	+			−	−	+	+			−	−	8	−2	0.5000	
quadratic T_2	+	+	−2	−2	+	+	+	+	−2	−2	+	+	24	−6	1.5000	
error (a)																
$B \times T_1$	−	−			+	+	+	+			−	−	8	8	8.0000	
$B \times T_2$	−	−	+2	+2	−	−	+	+	−2	−2	+	+	24	−8	2.6667	
sub-plots																
chemical																
C	−	+	−	+	−	+	−	+	−	+	−	+	12	−5	2.0833	
interaction																
$C \times T_1$	−	+			+	−	−	+			+	−	8	4	2.0000	
$C \times T_2$	−	+	+2	−2	−	+	−	+	+2	−2	−	+	24	4	0.6667	
error (b)																
$B \times C$	+	−	+	−	+	−	−	+	−	+	−	+	12	+3	0.7500	
$B \times C \times T_1$	+	−			−	+	−	+			+	−	8	−2	0.5000	
$B \times C \times T_2$	+	−	−2	+2	+	−	−	+	+2	−2	−	+	24	−6	1.5000	
															34.25	

stand that the card-programmed electronic calculator has been built to fit this situation. The number of observations, which directly reflects the size of an experiment, that can be handled by the CPC is determined by the capacity of the internal storage.

Frequently, the number of observations resulting from an experiment is much larger than the internal storage of the CPC can accommodate, but a less detailed analysis than that yielded by the above method will suffice. In such cases, whether they be counts or measurements, the data can be reduced by machine methods by employing sorting and tabulating techniques more fully. The same set of 12 observations will be used to illustrate the method. Only a few of the necessary machine details will be given in Table III (A, B, C). The number of runs or subtables required for this scheme is equal to the number of major divi-

TABLE III

STEP METHOD OF DOING ANALYSIS OF VARIANCE CALCULATIONS

(PART A)

		Chemical	*Time*	*Block*	*Total*
		1			
		3	4		
		4			
		2	6		
		5			
		4	9	19	
		6			
		5	11		
		7			
		6	13		
		5			
		3	8	32	51
number of entries	n	12	6	2	1
number of values per entry	k	1	2	6	12
number of degrees of freedom	d/f	12−6=6	6−2=4	2−1=1	
sum of squares of entries		251	487	1385	2601
sum of squares of entries over k		251.0	243.5000	230.8333	216.75
correction terms		243.5	230.8333	216.7500	
difference		7.5	12.6667	14.0833	

$C + C \times T + C \times B + C \times T \times B = 7.5$

$T + T \times B = 12.6667$

$B = 14.0833$

TABLE III [continued]

(PART B)

		Time	*Block*	*Chemical*	*Total*
		1			
		4			
		5	10		
		6			
		7			
		5	18	28	
		3			
		2			
		4	9		
		5			
		6			
		3	14	23	51
number of entries	n	12	4	2	1
number of values per entry	k	1	3	6	12
number of degrees of freedom	d/f	12−4=8	4−2=2	2−1=1	
sum of squares of entries		251	701	1313	2601
sum of squares of entries over k		251.000	233.6667	218.8333	216.75
correction terms		233.6667	218.8333	216.7500	
difference		17.3333	14.8334	2.0833	

$$T + T \times B + T \times C + T \times B \times C = 17.3333$$
$$B + B \times C = 14.8334$$
$$C = 2.0833$$

(PART C)

		Block	*Chemical*	*Time*	*Total*
		1			
		6	7		
		3			
		5	8	15	
		4			
		7	11		
		2			
		6	8	19	
		5			
		5	10		
		4			
		3	7	17	51
number of entries	n	12	6	3	1
number of values per entry	k	1	2	4	12
number of degrees of freedom	d/f	12−6=6	6−3=3	3−1=2	
sum of squares of entries		251	447	875	2601
sum of squares of entries over k		251.0	223.50	218.75	216.75
correction terms		223.5	218.75	216.75	
difference		27.5	4.75	2.00	

$$B + B \times C + B \times T + B \times C + T = 27.5$$
$$C + C \times T = 4.75$$
$$T = 2.00$$

C = sum of squares attributable to differences between chemicals
T = sum of squares attributable to effects of time
B = sum of squares attributable to block differences
$C \times T$ = sum of squares attributable to interaction between chemicals and times
$T \times B$ = sum of squares attributable to interaction between times and blocks
$C \times T \times B$ = sum of squares attributable to interaction between chemicals, times, and blocks

sions of the variance to be made which is three in this case. All but the last column in each table will be headed by a variable or major division, and the order of sorting will be determined by reading these headings from left to right. After the first subtable has been made, only one sort is required for each of the others. In each subtable the first

TABLE IV
SUMMARY OF ANALYSIS

Source of Variability			*D/F*	*Sum of Squares*
major plots				
block	B		1	14.0833
time	T		2	2.0000
error (a)	$B \times T$		2	10.6667
minor plots				
chemical		C	1	2.0833
interaction		$C \times T$	2	2.6667
error (b)	$B \times C + B \times C \times T$		3	2.7500
Total			11	34.2500

column is formed by listing all observations, the second consists of subtotals secured by minor controlling on the variable heading the column, the third consists of subtotals secured by major controlling on the variable at its head, and the fourth is the grand total or sum of all observations.

From a combination of the three equations at the bottom of each subtable, the values of the components of variability defined above can be determined. To clarify the notation, it might be well to point out that $C \times T \times B = T \times B \times C = B \times C \times T$, etc. The solution of this set of equations may be summarized as shown in Table IV.

The sum of squares for the divisions of the total sum of squares in Table IV can also be secured from Table II.

The details on how to adapt machines to this step method will depend, of course, on the machines available. With only a sorter and accounting machine, the procedure would follow that indicated in IIIA, IIIB, IIIC, with considerable use of a table model calculator being required. With higher types of machines at hand, the alert operator will readily adapt them to the problem.

DISCUSSION

Chairman Hurd: One idea, that has appealed to me particularly, was the first idea described. One has this matrix, Table I. In order to compute the sums of squares, one needs to multiply the matrix by the vector, essentially. So if one had a number of experiments which were like this in form —a number of factors—one would punch a deck of program cards in which each card has an operation which is either add or subtract. That is clearly all that is required; and, given an experiment, you punch one card for each observation, drop these in the calculator, followed by this program deck which performs the matrix-by-vector multiplication, and you have sensibly analyzed it.

Dr. Brandt: If you wish to do a multiple covariance correlation or multiple regression with this method, all you have to use is a multiple-column vector, each column of the vector representing one of those variables. As it was here, I had merely the one variable, and so I used a one-column vector. But I can have a vector with a good many columns, if I desire.

*Remarks on the IBM Relay Calculator**

MARK LOTKIN

Ballistic Research Laboratories, Aberdeen Proving Ground

ASIDE from its large-scale digital computers, such as the ENIAC and the Bell Relay Computer, the Computing Laboratory of the Ballistic Research Laboratories, Aberdeen Proving Ground, has at its disposal some special high-speed computing devices built by IBM for the Ordnance Department, the IBM Relay Calculators. Two of these machines, identical in every respect, were installed in December, 1944, and they have been in continuous operation ever since, except for short periods of time when the incorporation of certain improvements necessitated their shutdown.

While there are presently in existence three other IBM Relay Calculators—one at the Naval Proving Ground at Dahlgren, Virginia, and two at the Watson Scientific Computing Laboratory, Columbia University—the Aberdeen relay calculators have acquired additional interest owing to certain modifications that have been made lately on these machines. I am referring here to the hooking up of the twin machines, thus effectively transforming them into a single device of greatly increased computing power.

To arrive at a better understanding of the "coupled" calculator, as it is presently constituted, it seems best to acquaint the reader first with the principal features of these sequence-controlled, digital relay calculators, and then discuss some of the modifications resulting from the hookup. For more detailed descriptions of the relay calculators the reader is advised to study the recently published report of J. Lynch and C. E. Johnson, "Programming Principles for the IBM Relay Calculators," *Ballistic Research Laboratories Report No. 705,* October, 1949.

The Relay Calculator

Data, to be introduced into the storage or computing registers of the machine, must be punched in decimal form on standard IBM cards; each number may have up to 12 digits with a plus or minus sign. Two separate card feeding units, the reproduce feed and the punch feed, each operating normally at the rate of 100 cards a minute, are associated with five reading stations and one punching station. Four of these reading stations permit as many as four cards to be read simultaneously, while results are punched on a fifth card. The fifth reading station provides a means of reading from the cards certain punches which may be used to control the sequencing of operations to be performed. This possibility of programming by means of control cards, in addition to the usual wiring of control panels, will be discussed in greater detail later on.

A comparing unit, pluggable to reading brushes as well as counters, permits the comparison of two sets of data each having as many as 80 columns. Thus, data punched in two decks may be compared, and the machine may be instructed to stop when such a comparison reveals a disagreement.

There is, further, available in the machine a collating unit, of importance when it is desired to compare six-digit or 12-digit numbers punched in cards or stored in counters. Depending on the result of this comparison, the machine may then be instructed, by means of class selectors, to start certain operations such as card feeding, transfers, etc.

The timing of all operations performed within the machine is regulated by a shaft in the card feeding unit rotating at the rate of 100 RPM. During one revolution of this shaft, a "machine cycle," exactly one card may be read or punched. During the same revolution a cam timing circuit produces 48 impulses that may be used to activate relays. These sequence impulses, emitted at regular time intervals of 1/80 of a second, form the basic sequence of "points" that governs the timing of all machine operations. A point represents the approximate length of time required for the proper operation of the relays.

While the timing unit, just mentioned, produces the 48 relay impulses, a second unit emits 12 digit impulses; these are of importance for the reading, comparing and punching of cards, and in the coupled calculator, also, for the control of operations. For the proper time sequencing of operations it is important to keep in mind the time of occurrence of these digit pulses within the card cycle. This basic relationship is shown in Figure 1.

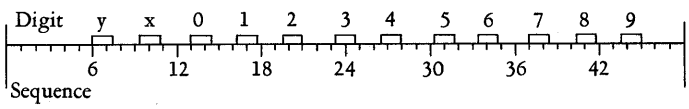


FIGURE 1. SEQUENCE PULSES AND DIGIT PULSES

*This paper was presented by title.

While the basic mode of operation is of the single-cycle type, it is possible, in cases where the normal sequence of 48 points is insufficient to permit the performance of series of operations required in the solution of a problem, to run the machines on the double-cycle basis. Only 50 cards then are read per minute, but the machine carries out approximately twice as many operations as it does during the single cycle setup. The storing and computing of numbers in the machine is effected entirely by means of approximately 2,500 electro-mechanical relays arranged in 31 counters, of which five can have 12-digit signed numbers, and the other 26 can accommodate six-digit signed numbers. Numbers are stored in the counters by a "pentad" system; in this type of arrangement each decimal digit is represented by a certain position in a "column" consisting of five relays, as shown in Figure 2. There the configurations *a, b, . . . , f* represent, respectively, the decimal digits 0, 1, 2, 5, 6, 9. Thus, a number of six digits requires the services of 30 relays, aside from the sign relay.

1	2	3	4	5	
O	O	O	O	O	a
●	O	O	O	O	b
O	●	O	O	O	c
O	O	O	O	●	d
●	O	O	O	●	e
O	O	O	●	●	f

FIGURE 2. THE "PENTAD" SYSTEM

In addition to the 31 counters there are 24 dial switches, each capable of storing one digit. The machines, therefore, have an internal storage capacity of 240 digits each, or twice that number when operated in tandem, as described below.

The different counters can perform some, but not all, of a variety of functions. Such functions have to do with the receiving of data from the switches, the cards, or other counters; transmitting data to other counters or the punch magnets; cooperating with other counters in the formation of new quantities such as sum, product, etc.; shifting numbers to the right or left; rounding and comparing of results. All these operations are carried out at relatively high speed. Thus, the time required for the addition or subtraction of two signed numbers, having as many as 12 digits, an operation which requires three counters, is two sequence points. For multiplication of two numbers of six digits each, resulting in a 12-digit product, 16 points are needed; this ties up 10 counters. The division of a 12-digit numerator by a six-digit denominator, resulting in a quotient having as many as 12 digits, takes, on the average, 0.2 seconds per quotient digit. This operation engages six counters. It takes about the same amount of time, but only four counters, to obtain one of the six digits that form the square root of a 12-digit number.

Now, a few words about the two control panels of the relay calculator. As mentioned previously, all operations carried out within the machine are synchronized with a timing unit which emits 48 sequence impulses during one revolution of a rotating shaft, or 96 impulses when double-cycle operation is called for. These impulses, coming in succession, are routed to 96 sequencing relays, each of which is associated with a number of outlets located on the right-hand section of a two-section control panel. Sequence impulses through these outlets, thus, can be made to control the sequencing of operations. The left half of the two-section control panel, on the other hand, serves to decide the type of operation to be performed on each sequence point.

The general purpose of the other control panel, consisting of three sections, is to give instructions pertaining to the reading, comparing, and punching of data, and to assign the channels through which the data can be routed.

Coupling of the Relay Calculators

While, in general, very fast and efficient, the machines, when operated singly, suffer from two shortcomings that quickly become apparent when problems of a more complex nature are to be solved: they are rather limited in both storage and programming facilities.

Now these limitations have been greatly mitigated. C. E. Johnson, in cooperation with C. B. Smith of IBM, has worked out a method that utilizes both machines simultaneously in a manner designed to minimize the deficiencies cited above. Coincidentally, work is progressing on an all-purpose control panel to be used when the machines are operated in tandem. The two machines are connected by the installation of a single connector cable between the main units of each machine. The changes in the internal wiring of both machines necessary for this type of coupling were, however, made in such a way that by simply disconnecting the cable, both machines may be used independently as done prior to the connection. When working in tandem, one of the calculators acts as a storage organ only, while the other one takes over the computing functions. Merely by a switch of the ends of the connector cable, the roles of storage unit and computing unit also become interchanged. Obviously, this type of arrangement has much to recommend it, since it may permit continuation of a problem without interruption in the event certain types of machine failure occur.

In tandem operation, the machine that supplies the sequence impulses is the computing organ, while the machine receiving these impulses becomes the storage organ.

Operational control may be exercised entirely by punches on cards, through the use of class selectors, which are emi-

nently suited for variable sequencing of operations. Each machine has 25 such selectors, of which six have three positions, and 19 have six positions, with four of the latter category being of limited applicability. Twenty-one of these selectors are presently being used when problems are solved on the coupled calculator. A further enlargement of programming facilities has been achieved by splitting up each six-position class selector into two class selectors of three positions, thus obtaining a total of 72 three-position selectors, a number which seems to be quite adequate for most of the jobs usually considered for the calculator.

Also, aside from code punches located in pre-assigned columns, the control cards are punched to govern the operation of the feed and punch units of both individual units; they may, moreover, contain constants that are needed in the course of the computation. Any of the symbols Y, X, 0, 1, . . . , 9, in any of the 80 columns of a standard IBM card, may be used as code in the card control. Thus, a Y in column 3 may run the punch feed of the storage unit, an X in column 59 may instruct counter 10 to transfer out, a 0 in column 64 may initiate the performance of a certain shift, etc. The choice of control punches is quite arbitrary and may be made to suit the requirements of each particular problem, in conformity with the wiring of the control panel.

The sensing of the code punches by the reading contacts of the calculator initiates a signal which, properly routed through the class selectors, starts a certain routine previously wired on the control panel. A deck of control cards, then, placed in one of the four feeds of the computer, will govern the sequences of operations into which the problem has been broken previously.

In a typical setup the calculator will operate as follows: first, input data are punched on cards and placed into the reproduce (R_s) and punch (P_s) feeds of the storage unit. Reproduce (R_c) and punch (P_c) feeds of the computer, on the other hand, contain the deck of control cards and blank cards, respectively. The reading of the first control card causes R_s to feed a card, thus introducing initial data into the machine. The following control cards cause the operations to be carried out in the desired sequential manner. Ultimately, the last control card instructs P_c to punch the final result. Now the sequence of the operations is started again by the reading of the first control card, which causes a new set of initial data to be introduced through R_s, etc. To preserve the continuity of operations, it has been found advisable to produce not one, but a sufficient number of such control decks, thus avoiding a delay, otherwise encountered, because of the necessary handling of the control cards.

Examples of Problems

The coupled calculator has been in operation now for over a year, and the results have been extremely satisfactory. Many of the problems, whose complexity made their solution on the relay calculators impractical, are now handled speedily and efficiently. Thus, the calculation of the instantaneous position in space of a moving body by means of theodolite observations has been achieved in slightly over two minutes.

In another problem it was necessary to solve a system of 32 linear equations arising from a triangulation problem. To give an idea of the variety of types of computations now possible with the relay calculators, perhaps, it may be worthwhile to present a short account of some of the phases into which the solution of a problem of this latter type may be divided conveniently.

After carrying out the measurements previously decided upon, and inserting the conditions inherent in the geometry of the problem, there will result certain conditional relationships represented by p linear equations in n unknowns v_k:

$$f_i \equiv \sum_{k=1}^{n} a_{ik} v_k + r_i = 0, \quad i = 1, 2, \ldots, p. \tag{1}$$

The solution of (1) may be achieved by the minimization of $\sum v_k^2$, subject to constraints imposed by (1). Introducing p Lagrange multipliers, c_i, we then must make

$$F \equiv \sum_{k} (v_k^2/2) - \sum_{i} c_i f_i = \min. \tag{2}$$

Now the equations $\partial F/\partial v_k = 0$ necessary for this minimization lead to

$$v_k = \sum_{j=1}^{p} a_{jk} c_j. \tag{3}$$

These expressions, in turn, after being introduced into (1), lead to the normal equations

$$\sum_{k} A_{ij} c_j + r_j = 0, \quad j = 1, 2, \ldots, p, \tag{4}$$

with

$$A_{ij} = \sum_{k} a_{ik} a_{jk}. \tag{5}$$

Once equations (4) are solved, then the v_k, as computed by (3), will satisfy (1).

In the carrying out of actual calculations, the following remark is of importance for the checking of the normal coefficients: if we put

$$s_k = \sum_{j} a_{jk}, \quad k = 1, 2, \ldots, n, \tag{6}$$

and

$$S_i = \sum_{j} A_{ij}, \quad i = 1, 2, \ldots, p,$$

then, obviously,

$$S_i = \sum_{k} a_{ik} s_k. \tag{7}$$

The first phase of the problem, then, deals with the calculations of the normal matrix, a job which may be handled on the relay calculators in straightforward fashion. The

second phase, i.e., the solution of equations (4), however, may be done in a variety of ways. We have found Jordan's method, which is a modification of Gauss' method of elimination, most suitable for treatment on the machines. Applied to our matrix (A_{ij},r_i), i.e., (A_{ij}) augmented by the column r_i, this method proceeds as follows:

Step 1. Find the A_{i_1} of largest absolute value. Without loss of generality this coefficient may be labeled A_{11}. Then by using the factors

$$m_i^{(1)} = A_{i1}/A_{11}$$

it is permissible to transform the matrix (A_{ij},r_i) into the equivalent matrix $(A_{ij}^{(1)}, r_i^{(1)})$ by means of the relationships

$$\begin{aligned} A_{ij}^{(1)} &= A_{ij} - m_i^{(1)}A_{1j}, \\ r_i^{(1)} &= r_i - m_i^{(1)}r_1, \quad i = 2, 3, \ldots, p. \end{aligned} \tag{8}$$

Accordingly,

$$A_{i1}^{(1)} = 0, \text{ while } A_{11}^{(1)} = A_{11}.$$

Note also that if $A_{ji} = A_{ij}$ then $A_{ij}^{(1)} = A_{ji}^{(1)}$ for $i, j > 1$.

Step 2. Find the $A_{i2}^{(1)}$ of the largest absolute value, for $i = 2, 3, \ldots, p$. This coefficient may be labeled $A_{22}^{(1)}$, without loss of generality. Form the factors

$$m_i^{(2)} = A_{i2}^{(1)}/A_{22}^{(1)}, \tag{9}$$

and put

$$\begin{aligned} A_{ij}^{(2)} &= A_{ij}^{(1)} - m_i^{(2)}A_{2j}^{(1)}, \\ r_i^{(2)} &= r_i^{(1)} - m_i^{(2)}r_2^{(1)}, \quad i = 1, 3, 4, \ldots, p. \end{aligned}$$

Then

$$A_{i2}^{(2)} = 0, \text{ while } A_{22}^{(2)} = A_{22}^{(1)}.$$

Also, $A_{11}^{(2)} = A_{11}^{(1)} = A_{11}$, but $A_{i1}^{(2)} = 0$ for $i \neq 1$, since $A_{i1}^{(1)} = 0$ for these i.

Thus, the transformed matrix has the form

$$\begin{pmatrix} A_{11}^{(1)} & 0 & A_{13}^{(2)} & \ldots. & A_{1p}^{(2)} & r_1^{(2)} \\ 0 & A_{22}^{(2)} & A_{23}^{(2)} & \ldots. & A_{2p}^{(2)} & r_2^{(2)} \\ 0 & 0 & A_{33}^{(2)} & \ldots. & A_{3p}^{(2)} & r_3^{(2)} \\ . & . & . & & . & . \\ . & . & . & & . & . \\ . & . & . & & . & . \\ 0 & 0 & A_{p3}^{(2)} & \ldots. & A_{pp}^{(2)} & r_p^{(2)} \end{pmatrix}.$$

Again, if the original matrix A_{ij} is symmetric, then $A_{ij}^{(2)} = A_{ji}^{(2)}$ for $i, j > 2$.

Generally, let us assume that the first e columns have been so transformed. Among the $A_{i,e+1}^{(e)}$, $i = e+1, e+2, \ldots, p$, let $A_{e+1,e+1}^{(e)}$ be the one of largest absolute value. Then the relations

$$\begin{aligned} m_i^{(e+1)} &= A_{i,e+1}^{(e)}/A_{e+1,e+1}^{(e)} \\ A_{ij}^{(e+1)} &= A_{ij}^{(e)} - m_i^{(e+1)}A_{e+1,j}^{(e)}, \quad i \neq e+1, \end{aligned} \tag{10}$$

will carry the diagonalization process one step further.

After p such steps the matrix (A_{ij}, r_i) will look like this:

$$\begin{pmatrix} A_{11}^{(1)} & 0 & 0 & 0 & r_1^{(p)} \\ 0 & A_{22}^{(2)} & 0 & 0 & r_2^{(p)} \\ . & . & . & . & \\ . & . & . & . & \\ . & . & . & . & \\ 0 & 0 & 0 & A_{pp}^{(p)} & r_p^{(p)} \end{pmatrix}. \tag{11}$$

The desired solutions are, therefore,

$$c_j = -r_j^{(p)}//A_{jj}^{(j)}.$$

For purposes of checking calculations it is useful to carry along at each step the sums of all coefficients in each row. If, namely, one computes at the start

$$T_i = -\left(\sum_j A_{ij} + r_i\right), \tag{12}$$

and, during step 1, transforms the T_i together with the $r_i^{(1)}$:

$$T_i^{(1)} = T_i - m_i^{(1)}T_1, \tag{13}$$

then, obviously,

$$\sum_j A_{ij}^{(1)} + r_i^{(1)} + T_i^{(1)} = 0. \tag{14}$$

A similar relationship holds for each step.

Once the first solution $c^{(1)} \equiv (c_1^{(1)}, c_2^{(1)}, \ldots, c_p^{(1)})$ has been obtained, it takes considerably less time to get an improved solution $c = c^{(1)} + d$. If, namely, equations (4) are written in matrix form:

$$Ac + r = 0,$$

and the residuals due to $c^{(1)}$ are $t \equiv (t_1, t_2, \ldots, t_p)$:

$$Ac^{(1)} + r = t,$$

then

$$Ad + t = 0,$$

and since A has been diagonalized previously, it is necessary only to transform the column t.

The relay calculators, completing a whole step with each machine run, took about 20 hours to solve the problem with the desired degree of accuracy. From the experience gained thus far, it may be stated that the coupling of the IBM Relay Calculators has resulted in a decidedly superior computing machine.

*An Improved Punched Card Method for Crystal Structure Factor Calculations**

MANDALAY D. GREMS

General Electric Company

THOSE of you who have worked with calculations dealing with the structure of complex crystals, are reminded, probably, of the long monotonous operations involved. For this reason, a few persons here and there have attempted to find methods for simplifying the tremendous amount of hand calculations. Shaffer, Schomaker and Pauling, at the California Institute of Technology, were the first to report a method using the IBM equipment for this purpose. However, at our own Research Laboratory at the General Electric Company in Schenectady, there is a group of scientists who have spent considerable time and effort on this work, both analytically and theoretically.

After a few discussions of their problem, it seemed more efficient and better suited to the IBM equipment to begin with the general expression,

$$F_{hkl} = \sum_{j=1}^{N} f_j \cos 2\pi\,(hx_j+ky_j+lz_j) + i\sum_{j=1}^{N} f_j \sin 2\pi\,(hx_j+ky_j+lz_j)$$

rather than to use a specific and modified expression for each type of structure factor calculation.

This expression doesn't look difficult until you consider that it involves many combinations of the reflections h, k, l with the trial parameters x, y, z to find the best sets of x, y, z.

At the beginning, three separate decks of cards are key punched:

1. *Table cards.* For $\sin 2\pi\alpha$ and $\cos 2\pi\alpha$, where α ranges from 0.001 to 1.000, in intervals of 0.001. This pack is used for all crystal structure calculations.
2. *Reflection cards.* One card for each reflection h, k, l. This card also contains the scattering factor for each kind of atom, the temperature factor, and the absorption factor (if known) for that particular reflection. These reflection cards are used for all trials for a specific crystal structure factor.
3. *Parameter cards.* One card for each set of trial parameters x, y, z. The number of cards depends upon the unit of structure. These cards are used for a specific calculation.

First, reproduce the set of reflection cards as many times as there are sets of parameter cards, gang punching a set of x, y, z values on each reflection deck. If there are 400 reflections and eight sets of parameters, then there are 3,200 detail cards each containing an h, k, l, x, y, and z.

There are four main machine operations in the solution of this problem. The two important steps, or the two contributing the most to a more compact and general procedure, are steps I and III.

I. Forming the quantities $a_j = (hx_j + ky_j + lz_j)$
II. Obtaining the cosines and sines of a_j
III. Multiplying the trigonometric functions by the scattering factors, f_j
IV. Summing the previous products

Step I indicates the formation of the quantities $a_j = (hx_j+ky_j+lz_j)$. Using the IBM Type 602 Calculating Punch with the above detail cards, it is possible to find a_j, b_j, c_j, and d_j at the same time—that is, with only one passage of the cards through the machine, where

$$a_j = hx_j + ky_j + lz_j\,,$$
$$b_j = hx_j + ky_j - lz_j\,,$$
$$c_j = hx_j - ky_j + lz_j\,,$$
$$d_j = -hx_j + ky_j + lz_j\,.$$

As the next step involves looking up the sine or cosine of the quantities a, b, c, and d, it is sufficient to carry only the decimal places in the product and sums. Therefore, multiply h by x, and carry the three decimal places to the four summary counters, adding the product in counters 14, 15, and 16 and subtracting the product in counter 13. Then multiply k by y, add the three decimal places of the product into counters 13, 15, and 16, and subtract it into counter 14. In the same manner, multiply l by z, add the three decimal places of the product into counters 13, 14, 16, and subtract it into counter 15. To eliminate the possibility of cal-

*This method was presented at the American Society for X-ray and Electron Diffraction in Columbus, Ohio, on December 16, 1948. It also appeared in the December issue of *Acta Crystallographica*, Vol. 2, Part 6.

culating a negative value, add 1.000 in each of the counters 13, 14, 15, and 16 on the reading cycle. Now it is unnecessary to include negative α's in the sine and cosine table. When these four sums are punched, each card contains a positive number for a, b, c, and d.

For certain symmetry elements the structure factor will contain any or all of the terms b, c, d, as well as a; so there is a considerable reduction in the number of atomic parameters necessary when all can be found at one time.

This, also, makes the procedure general for most structures. The effect of symmetry is illustrated by the space group P_{mmm},

atoms at $x\ y\ z$, $x\ y\ \bar{z}$, $x\ \bar{y}\ z$, $\bar{x}\ y\ z$
$\bar{x}\ \bar{y}\ \bar{z}$, $\bar{x}\ \bar{y}\ z$, $\bar{x}\ y\ \bar{z}$, $x\ \bar{y}\ \bar{z}$,

$$F_{hkl} = 8\sum_{j=1}^{N/8} f_j(\cos 2\pi a_j + \cos 2\pi b_j + \cos 2\pi c_j + \cos 2\pi d_j).$$

As you notice, the structure factor can be written in terms of a, b, c, d. Therefore, the parameters of only 1/8 of the atoms in the unit cell need be considered.

Now look at another space group P_{nnm}, for example:
For $(h + k + l) = 2n$,

$$F_{hkl} = 8\sum_{j=1}^{N/8} f_j(\cos 2\pi a_j + \cos 2\pi b_j + \cos 2\pi c_j + \cos 2\pi d_j).$$

For $(h + k + l) = 2n + 1$,

$$F_{hkl} = 8\sum_{j=1}^{N/8} f_j(\cos 2\pi a_j + \cos 2\pi b_j - \cos 2\pi c_j - \cos 2\pi d_j).$$

For this space group, there are two different expressions for F_{hkl}, depending upon whether $(h+k+l)$ is odd or even. They both contain a, b, c, and d, but the algebraic combinations of the cosines differ. This does not change our general procedure, however, and it is a simple matter for the machines to separate the two groups at the proper time.

At step II, use the previously prepared sine or cosine table deck; sort, collate, and intersperse gang punch the table values on the detail cards for the four factors a, b, c, d where

$$\cos 2\pi a_j = A_j,$$
$$\cos 2\pi b_j = B_j,$$
$$\cos 2\pi c_j = C_j,$$
$$\cos 2\pi d_j = D_j.$$

Each detail card should now contain the following: $h,k,l,x,y,z,a,b,c,d,A,B,C,D$ as well as a code for the type and number of the atom.

Until step III there was no need for a particular arrangement of the cards. At this point, the cards must be sorted on the column indicating the kind of atom. The different kinds are kept separate, and these packs are sorted according to reflection h, k, l.

Using the accounting machine with the summary punch attached, punch on blank cards the (A, B, C, D) for each reflection. Repeat this operation for each kind of atom separately, where

$$\sum \cos 2\pi a_1 = \sum A_1 = A_{m_1},$$

and

$$\sum \cos 2\pi a_2 = \sum A_2 = A_{m_2}, \text{ etc.}$$

Therefore, each summary card for atom (1) now contains the code for kind of atom, the reflection, and $(A,B,C,D)\ m1$; summary card for atom (2) contains $(A, B, C, D)\ m2$ substituted for $(A, B, C, D)\ m1$.

At this time, it is necessary to refer to the expression for the particular structure being studied, in order to determine how to combine $(A, B, C, D)\ m1$ or $m2$.

Referring to space group P_{mmm}, $(A+B+C+D)\ m_1$ or m_2 is necessary.

Referring to space group P_{nnm}, $(A+B+C+D)\ m_1$ or m_2 for $2n$; and
$(A+B) - (C+D)\ m_1$ or m_2 for $2n + 1$.

This is only a minor change on the IBM type 602 control panel to perform either operation. Another variation for a complex group can be done simply and easily at this time, if both F_{hkl} and $F_{\bar{h}kl}$ are required.

The sums of the A, B, C, and D's can be calculated for both at the same time. Only one set of reflection cards (hkl) are required until the final stages of the work.

After the A, B, C, and D's are combined properly, the sum is multiplied by the proper scattering factor, f_{m_1} or f_{m_2}.

$$f_{m_1}(A+B+C+D)\ m_1 = Rm_1, \text{ etc.}$$

The final step consists of simply adding these products together for the proper reflection and multiplying by a factor for that reflection,

$$T_{hkl}(Rm_1 + Rm_2) = F_{hkl}.$$

It is usually of interest to note the contributions of each kind of atom to the final result; so it is advisable to list the factors Rm_1 and Rm_2, as well as F_{hkl}, on the record sheets.

DISCUSSION

Mr. Smith: How long did it take you to calculate, say, for the order of 600 reflexes for your space group P_{nnn} or P_{mmm}, or that order between 100 and 600?

Miss Grems: It first took me twice as long as it did later, because I checked, and after I had done quite a number of these I found there was no point in checking the calculation of the quantities $hx + ky + lz$ or a, b, c, and d, because if there was an error it wouldn't make much difference. I would say roughly about three and a half days—perhaps four or four and a half. It added about an extra day to carry on the checking, although a good part of the checking could go on at the same time. I always did check the last part of calculations after I found the A's. Of course, it was a simple check.

Mr. Smith: That was for roughly about 500 reflexes?

Miss Grems: That is right; and breaking it down to about eight y, x, z's.

Mr. Smith: That would be roughly about a fifth of the time it would take you with a hand calculator, maybe less?

Miss Grems: For the particular case about which I was talking, we found for both the F_{hkl} plus and F_{hkl} minus, it took only a half-hour longer to get the F_{hkl} minus.

Chairman Hurd: Is the method which you have used, Mr. Smith, roughly analogous to this?

Mr. Smith: Unfortunately, no. I have been using a method somewhat similar to the one they use at California Tech., which differs somewhat from this; and, unfortunately, most of it has been done on the hand calculator. Also, unfortunately, the last case, instead of having, say, eight terms, had twenty terms in the general space group. It was a little more involved than that, but I was able, by using some Fourier transforms, to eliminate the necessity of calculating those two longer terms.

Mr. Thompson: Regarding the layout cards for master cards, which most of us use, our local IBM man made a very good suggestion of which some of us may not have thought. He suggested that we punch every hole in the card. When you want to read a detailed card, you put the layout card right over the detail card as a mask, and this makes it very quick to read. A couple of warnings, however: When you do this, don't punch every hole at once. If you punch them all simultaneously, two things might happen. The punches may stick in the dies or, as a matter of fact may come out of the left-hand side. It is advisable to send them through about eight times.

The Calculation of Complex Hypergeometric Functions with the IBM Type 602-A Calculating Punch

HARVEY GELLMAN
University of Toronto

THE hypergeometric function $F(a,b;c;z)$ is usually defined, for purposes of calculation, by the hypergeometric series:

$$F(a,b;c;z) = 1 + \frac{a\cdot b}{1\cdot c}z + \frac{a(a+1)\,b(b+1)}{1\cdot 2\cdot c(c+1)}z^2 + \frac{a(a+1)(a+2)\,b(b+1)(b+2)}{1\cdot 2\cdot 3\cdot c(c+1)(c+2)}z^3 + \cdots . \quad (1)$$

this series being convergent for $|z| < 1$.

Many physical problems lead to integrals which can be expressed in terms of hypergeometric functions, and many important functions employed in analysis are merely special forms of $F(a,b;c;z)$. Thus:

$$(1+z)^n = F(-n,\beta;\beta;-z)$$
$$\log(1+z) = zF(1,1;2;-z)$$
$$J_\nu(z) = \lim_{\substack{\mu\to\infty\\ \lambda\to\infty}} \frac{\frac{1}{2}(z)^\nu}{\Gamma(\nu+1)} F\left(\lambda,\mu;\nu+1;-\frac{z^2}{4\lambda\mu}\right).$$

The purpose of this paper is to describe a method for computing $F(a,b;c;z)$ from (1) when a, b, c, and z are each of the form $x+iy$; x,y real, $i^2 = -1$. We were confronted with these functions through the problem of internal conversion of γ-rays emitted by a radioactive nucleus. The radial integrals arising from the calculation of transition probabilities were expressible in terms of hypergeometric functions. Our problem involved 90 hypergeometric functions, and on the basis of a preliminary estimate, 99 terms of the series were used. Such complex hypergeometric functions cannot be conveniently tabulated since they involve eight variables, and so a method is required which will readily yield $F(a,b;c;z)$ for special values of a, b, c, and z.

Calculations

We begin by defining

$$f_n = g_n + ih_n = \frac{(a+n)(b+n)}{(n+1)(c+n)}z \quad (2)$$

where $a = A_1+iA_2$; $b = B_1+iB_2$; $c = C_1+iC_2$; $z = z_1+iz_2$.

Then $F(a,b;c;z) = 1+f_0f_1+f_0f_1f_2+f_0f_1f_2f_3+\cdots$ (3)

The expanded form of f_n can be written as:

$$f_n = (z_1+iz_2)\left[\frac{F_1}{F_3}+i\frac{F_2}{F_3}\right] = \left[z_1\left(\frac{F_1}{F_3}\right)-z_2\left(\frac{F_2}{F_3}\right)\right]+i\left[z_2\left(\frac{F_1}{F_3}\right)+z_1\left(\frac{F_2}{F_3}\right)\right] \quad (4)$$

where $F_1 = F_1(n) = n^3+n^2(A_1+B_1+C_1) + n[A_1(B_1+C_1) - A_2(B_2-C_2) + B_1C_1 + B_2C_2] + [A_1(B_1C_1+B_2C_2) + A_2(B_1C_2-B_2C_1)]$
$= n^3 + a_2n^2 + a_1n + a_0$ (5)

$F_2 = F_2(n) = n^2(A_2+B_2-C_2) + n[A_2(B_1+C_1) + A_1(B_2-C_2) + B_2C_1 - B_1C_2] + [A_2(B_1C_1+B_2C_2) - A_1(B_1C_2-B_2C_1)]$
$= b_2n^2 + b_1n + b_0$ (6)

$F_3 = F_3(n) = n^3 + n^2(2C_1+1) + n(C_1^2+C_2^2+2C_1) + (C_1^2+C_2^2)$
$= n^3 + d_2n^2 + d_1n + d_0$ (7)

Thus, our object is to compute (7), (6), (5), (4) and (3) in that order.

Machine Procedure

The six numbers representing the real and imaginary parts of a, b and c are key punched on cards, and the coefficients $a_2, a_1, \ldots, d_0$ are computed by basic multiplications and additions. This computation requires eight sets of cards which are later combined on the reproducing punch to yield a set of master cards for the coefficients of the polynomials in n. The layout for this master set is given below:

MASTER SET 9

Card Columns	*Data*	*Remarks*
1-2	group number	each $F(a,b;c;z)$ defines a group; we required 90 values of the hypergeometric function
3-10	a_2	
11-18	a_1	
19-26	a_0	
27-34	b_2	
35-42	b_1	
43-50	b_0	
51-58	d_2	
59-66	d_1	
67-74	d_0	
79	'X'	
80	set number 9	

Computation of Polynomials

A set of detail cards (set 10) containing n, n^2, and n^3 for $n = 0(1)k$ is generated on the reproducing punch. This set contains $(k+1)$ cards for each hypergeometric function to be evaluated, k being the highest value of n used. In our calculation each group contained the same number of cards (i.e., the same value of k was used throughout) to minimize the handling of cards by the operator.

The master set 9 cards are inserted in front of their appropriate groups in detail set 10, and three separate group multiplications are performed on the 602-A calculating punch to yield F_1, F_2 and F_3. The planning chart and control panel wiring for F_3 is shown in Figure 1.

Sign Control in Group Multiplication

Since the coefficients on a master set 9 card may be positive or negative, their sign must be retained in the machine for the complete group of detail set 10 cards following the master card. This is achieved by dropping out pilot selectors 4 to 7, which control the sign of multiplication, through the transfer points of pilot selector 1. Pilot selector 1 is picked up at the control brushes by the master card and is dropped out in the normal manner.

Computing F_1/F_3 and F_2/F_3

Since the polynomials in n can have from 6 to 12 significant digits, 12 columns are assigned to them. Detail set 10 cards are sorted on F_3 into separate groups which have 12, 11, 10, 9 and 8 significant digits, respectively. Treating each of these groups separately, the above quotients are easily formed through basic division. The layout of detail set 10 is given below:

DETAIL SET 10

Card Columns	*Data*	*Remarks*
1-2	group number	n, n^2 and n^3 reproduced
3-4	n	from card table of n^x
5-8	n^2	
9-14	n^3	
15-26	F_1	
27-38	F_2	
39-50	F_3	
51-58	F_1/F_3	
59-66	F_2/F_3	
79-80	set number 10	

Computation of f_n

From equation (4) it is seen that the computation of f_n requires a complex multiplication of $(F_1/F_3) + i\,(F_2/F_3)$ by $(z_1 + i\,z_2)$, and this is equivalent to four real multiplications grouped together as shown in (4). The quantities: group number, n, F_1/F_3 and F_2/F_3 are reproduced from set 10 into a new set, 11, of detail cards. The values of z_1 and z_2 are key punched into a new set, 12, of master cards. By performing a complex group multiplication from set 12 to set 11 as shown in Figure 2, the values of f_n are generated. In our case z_1 was positive, and z_2 negative for all the groups, so that sign control on group multiplication as shown in Figure 1 was unnecessary in this operation.

Consecutive Complex Multiplication

Having obtained the f_n in the previous operation, we now require the products f_0, f_0f_1, $f_0f_1f_2$, etc. The method used for this computation is given below in schematic form:

Card No.	*Quantity Read from Card*	*Operation*
1	$f_0 = g_0+ih_0$	$1\cdot f_0 = R_0+iI_0$
2	$f_1 = g_1+ih_1$	$f_0f_1 = R_1+iI_1$
3	$f_2 = g_2+ih_2$	$f_0f_1f_2 = R_2+iI_2$
.	.	.
.	.	.
.	.	.
.	.	.

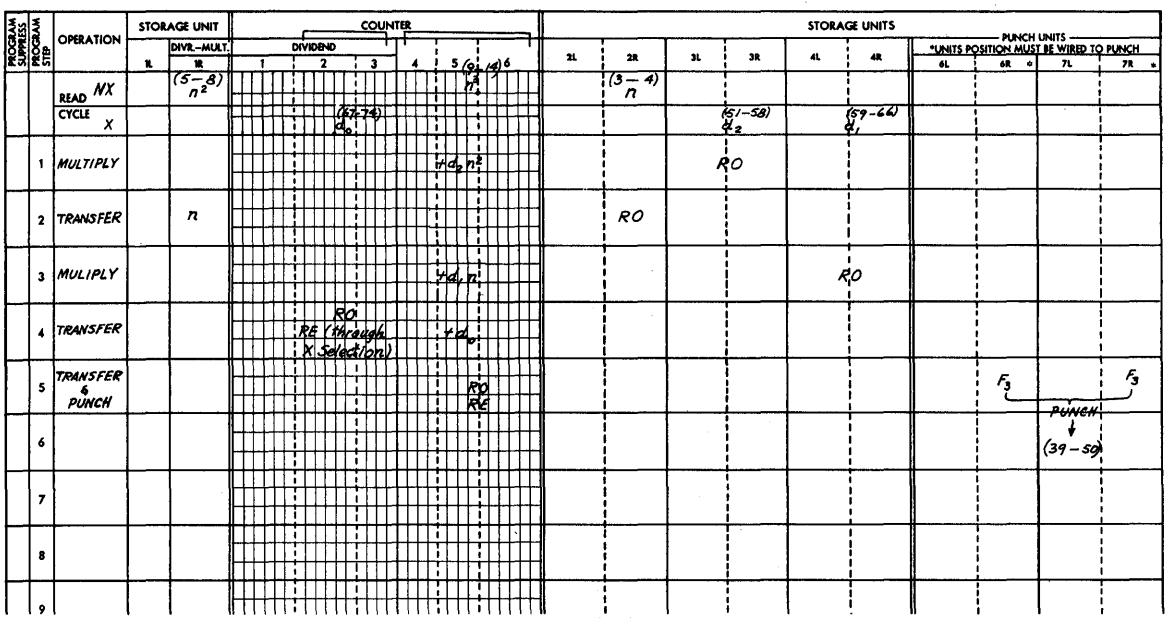
The planning charts and control panel wiring for this operation are shown in Figures 3 and 4, pages 165, 166. The essential features here are the retention of the calculated result from one card to act as a multiplier for the following card, and the conversion to true form of this multiplier if it should turn out originally as a complement figure in the counter. (The machine ordinarily converts complement numbers during the punching operation only.) In addition to this we must "remember" the sign of, say f_0f_1, when we multiply it by f_2 to form $f_0f_1f_2$. The scheme is started by feeding a blank card under the direction of a separate control panel which reads a 1 into storage $4R$ and resets all other storage units, counters and pilot selectors to normal.

The panel of Figure 4 is then used with one group of set 11 cards. The first card of this group reads into the machine the numbers $f_0 = g_0+ih_0$ and has punched on it $f_0 = R_0+iI_0$. The quantities g_0 and h_0 are retained in the machine and are used to multiply $f_1 = g_1+ih_1$ from the second card which has punched on it $f_0f_1 = R_1+iI_1$, and so on.

At the end of program step 6, counters 1, 2 and 3 contain R_k, the real part of $(f_0f_1f_2 \ldots . f_k)$. If R_k is negative, the counter group will contain the complement of R_k. On program 7, pilot selector 3 is picked up by an NB impulse, is transferred on program 8 and is dropped out on program 7 of the following card. Thus, the sign of R_k is remembered both during the conversion of R_k on program 8 and the multiplication by R_k on programs 2 and 6. A similar procedure is used for I_k.

The final step in the calculation of $F(a,b;c;z)$ consists in summing R_n and I_n separately on the accounting machine and recording the value:

$$F(a,b;c;z) = \left(1 + \sum_{n=0}^{k} R_n\right) + i\left(\sum_{n=0}^{k} I_n\right).$$



CALCULATING PUNCH – TYPE 602A – CONTROL PANEL

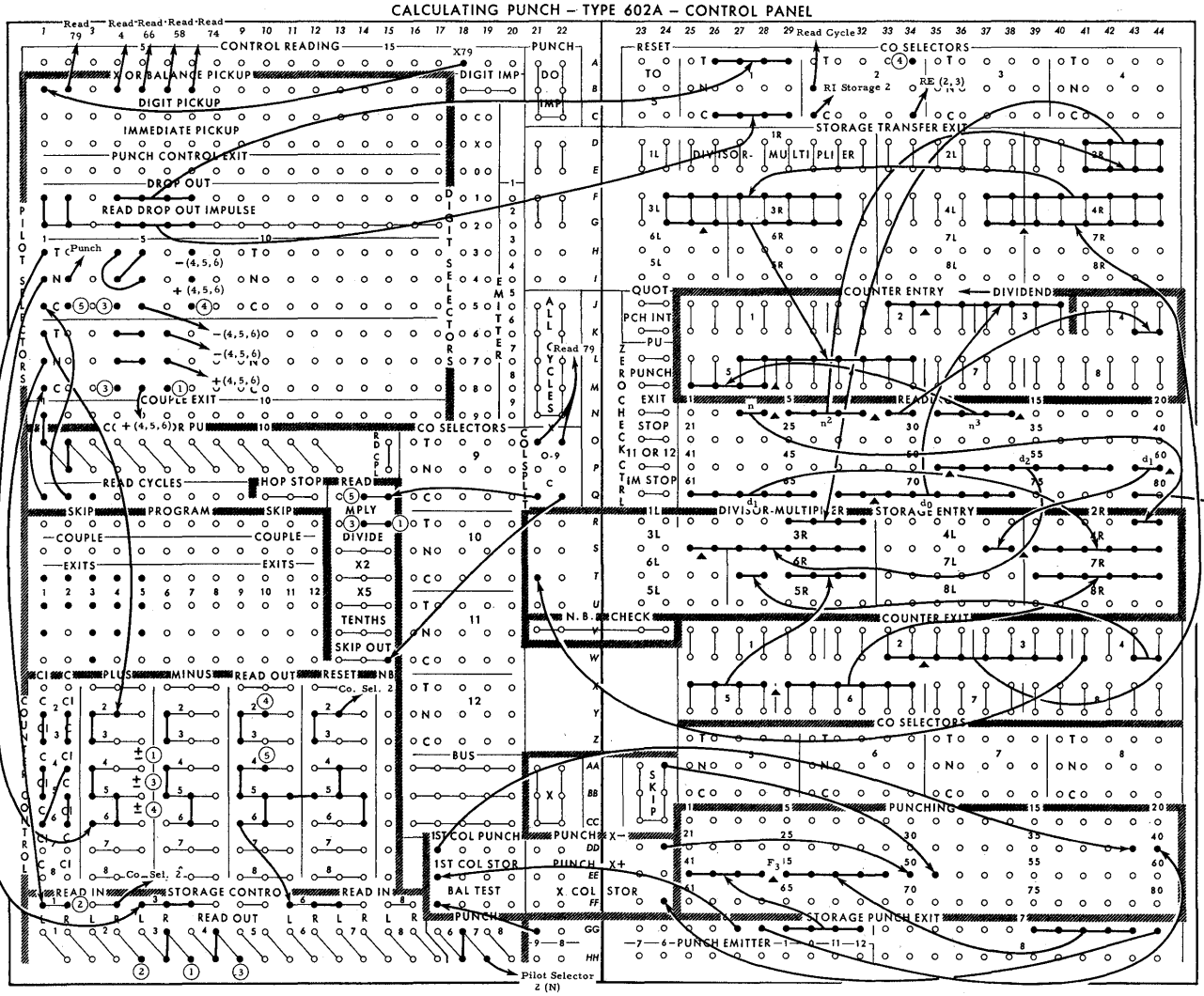
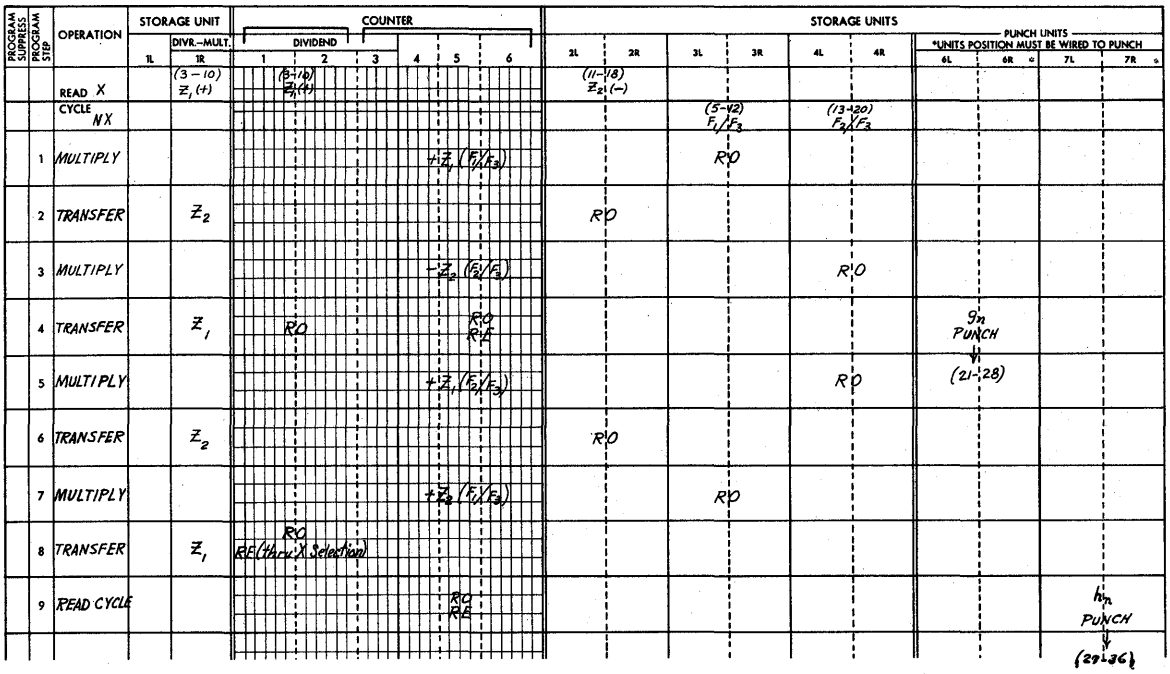


FIGURE 1. CUBIC POLYNOMIAL (SIGN CONTROL IN GROUP MULTIPLICATION)

$$n^3 + d_2n^2 + d_1n + d_0 = F_3(n)$$



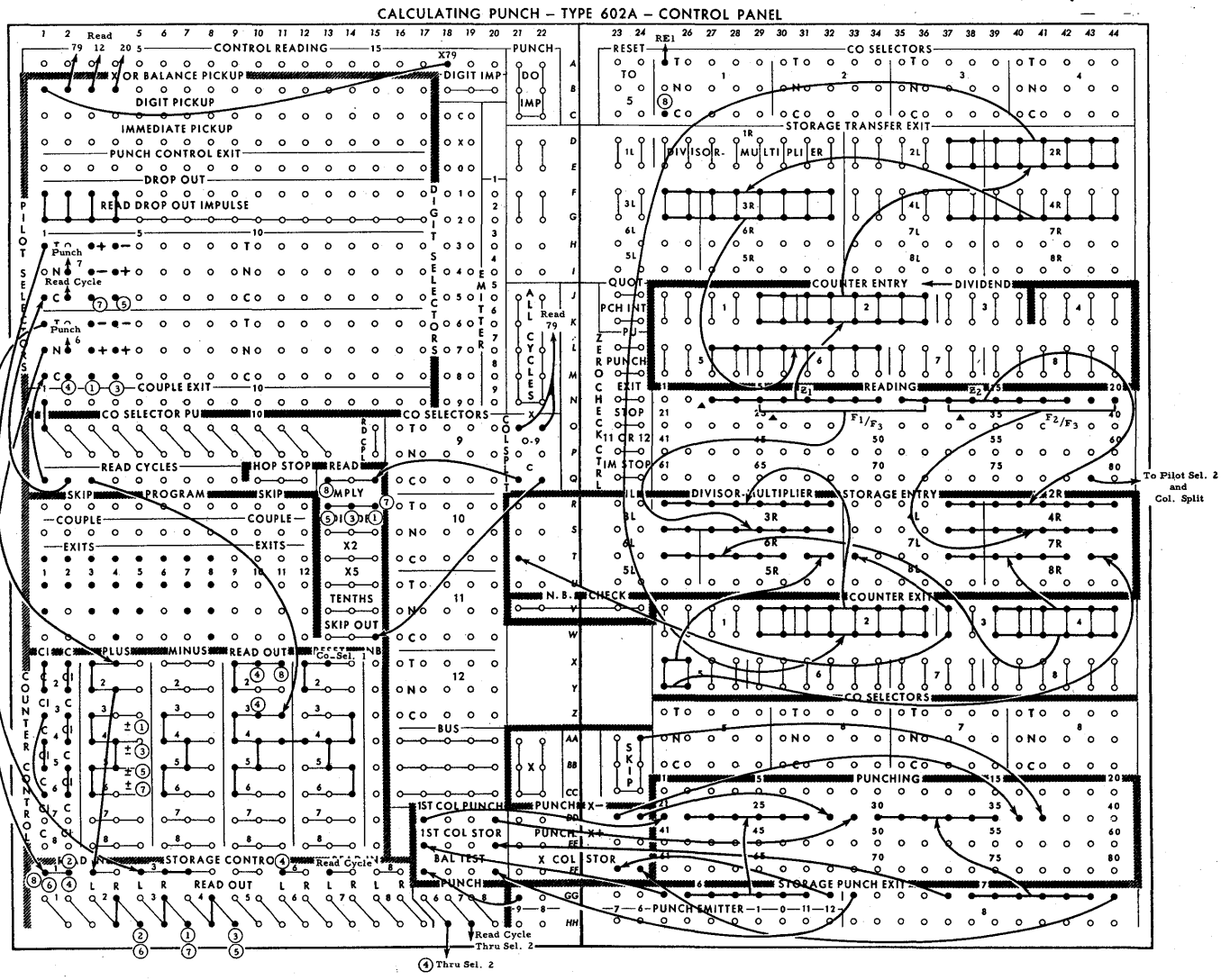


FIGURE 2. COMPLEX GROUP MULTIPLICATION

$$z_1\left(\frac{F_1}{F_3}\right) - z_2\left(\frac{F_2}{F_3}\right) = g_n\,;\, z_2\left(\frac{F_1}{F_3}\right) + z_1\left(\frac{F_2}{F_3}\right) = h_n$$

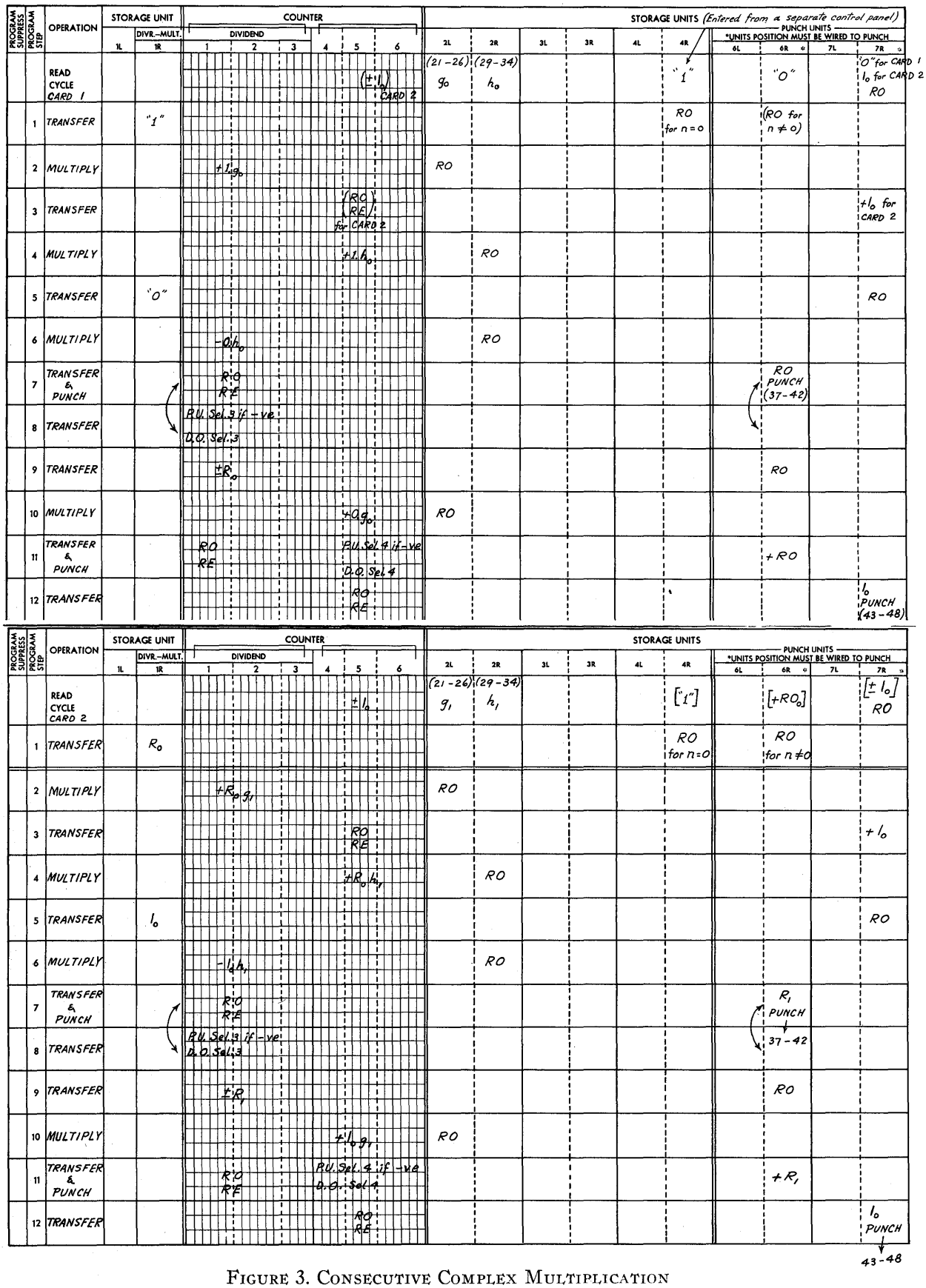


FIGURE 3. CONSECUTIVE COMPLEX MULTIPLICATION

Card 1: $f_0 = g_0 + ih_0 \qquad = R_0 + iI_0$

Card 2: $f_1 = g_1 + ih_1\,;\ f_0 f_1 \ = R_1 + iI_1$

Card 3: $f_2 = g_2 + ih_2\,;\ f_0 f_1 f_2 = R_2 + iI_2$

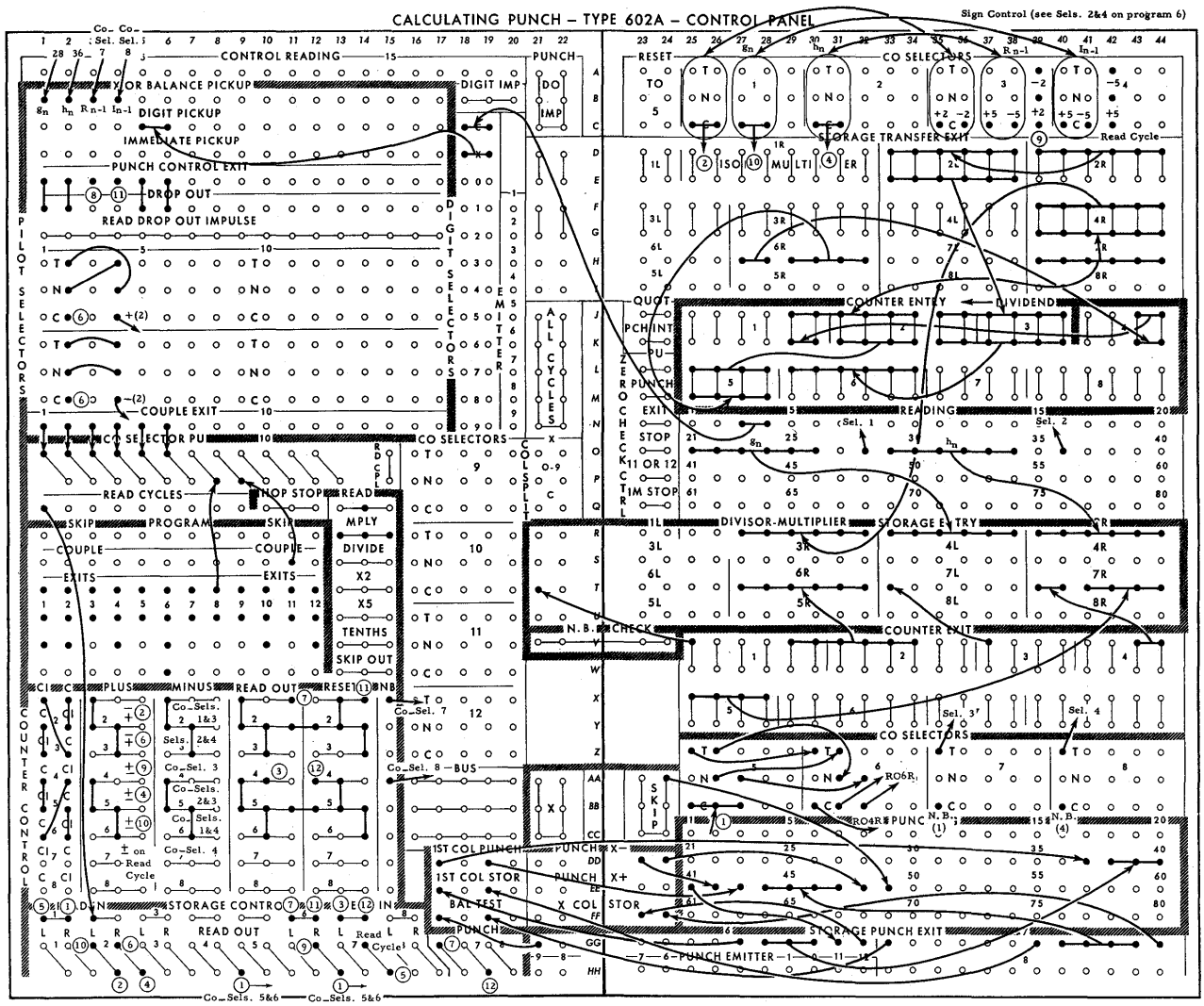


FIGURE 4. CONSECUTIVE COMPLEX MULTIPLICATION

Checking of Computations

The coefficients $a_2, a_1, \ldots, d_0$ of the polynomials in n are checked by manually checking one or two cards, performing the machine operation twice, and testing the resulting punches for double punching.

The polynomials F_1, F_2 and F_3 were checked separately by summation on the accounting machine according to the following formula:

$$\sum_{n=0}^{k} F_1(n) = \sum_{n=0}^{k} (n^3) + a_2 \sum_{n=0}^{k} (n^2) + a_1 \sum_{n=0}^{k} (n) + a_0(k+1)$$

$$= \left[\frac{k(k+1)}{2}\right]^2 + a_2\left[\frac{k(k+1)(2k+1)}{6}\right] + a_1\left[\frac{k(k+1)}{2}\right] + a_0(k+1)$$

where k is the last value of n in the series. F_2 and F_3 are checked in a similar manner.

Since the third differences of a cubic polynomial are constant, an alternative check consists of finding

$$\triangle F_1(n) = F_1(n+1) - F_1(n)$$

$$\triangle^2 F_1(n) = \triangle F_1(n+1) - \triangle F_1(n)$$

$$\triangle^3 F_1(n) = \triangle^2 F_1(n+1) - \triangle^2 F_1(n) = \text{constant}.$$

Generating Differences

Generally, functions which are tabulated at equal intervals of the argument can be conveniently checked by taking differences up to an order which is sufficient to show a "jump" indicating an error.

For this reason a planning chart and control panel wiring scheme is shown in Figure 5 for finding first differences of

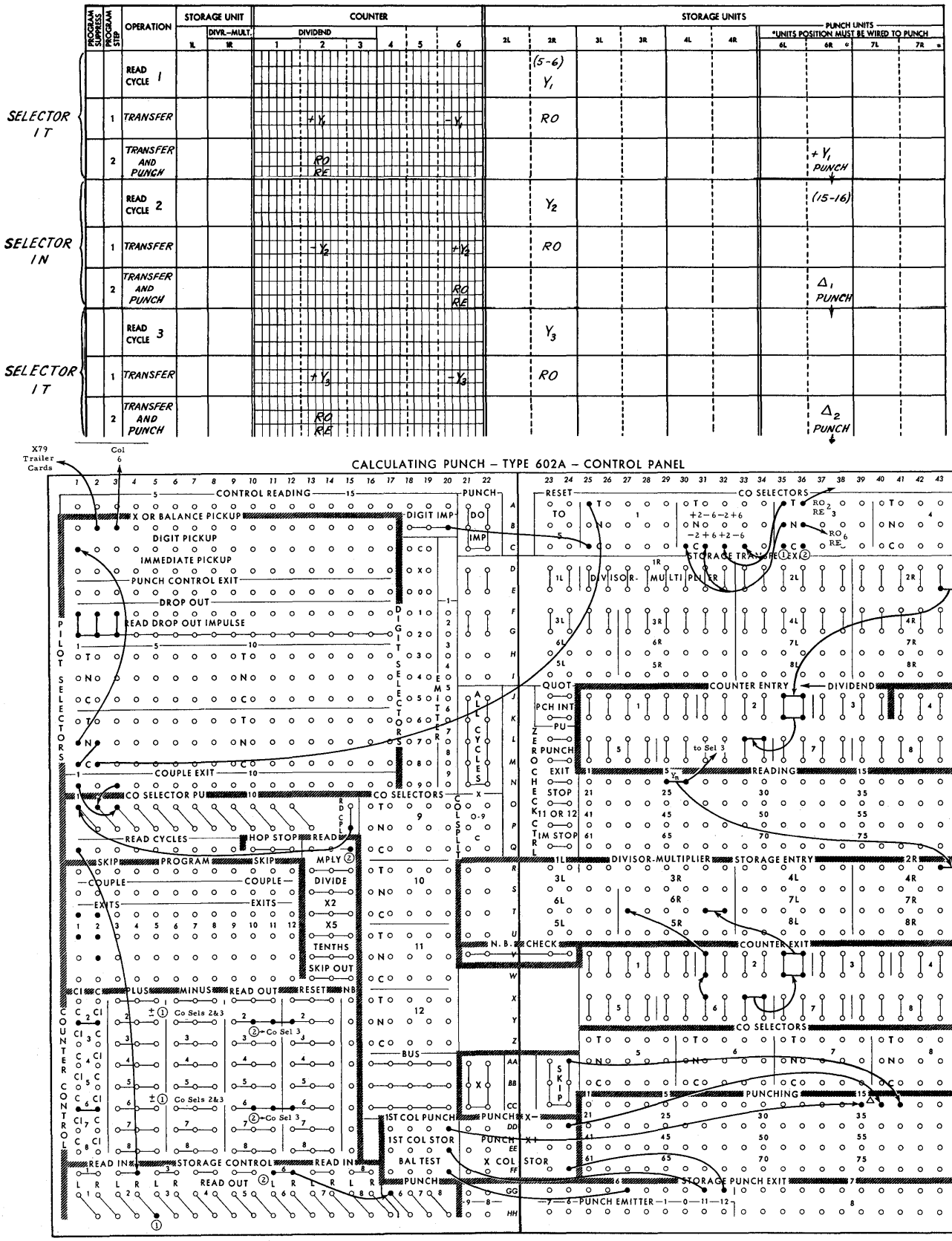


FIGURE 5. FIRST DIFFERENCES

$y_2 - y_1 = \Delta_1$ Card 1: y_1

$y_3 - y_2 = \Delta_2$ Card 2: y_2 punch Δ_1

$y_4 - y_3 = \Delta_3$ Card 3: y_3 punch Δ_2

etc. etc.

any tabulated function. Second differences are formed by operating on the first differences, and so on for the higher orders. The key to the scheme shown in Figure 5 consists in "alternating" a pilot selector so that it is transferred for the first card cycle, normal for the second card cycle, transferred for the third card cycle, and so on.

The complex group multiplication shown in Figure 2 was checked by summations on the accounting machine, using the normal check for group multiplication. The succeeding operations are each performed twice after one or two cards are checked manually, and the punched results are then tested for double punching.

The time required to calculate 90 hypergeometric functions containing 99 terms each is about 140 hours.

ACKNOWLEDGEMENT

The author wishes to thank Dr. C. C. Gotlieb, director of the Computation Centre, and Mr. George K. Copeland of the International Business Machines Company, Toronto, for their assistance in solving the problems discussed in this paper.

REFERENCES

1. E. T. WHITTAKER and G. N. WATSON, *Modern Analysis* (Cambridge University Press, London. 1946).
2. G. N. WATSON, *Theory of Bessel Functions* (Cambridge University Press, London. 1944).

DISCUSSION

Mr. Lowe: This problem of the presence of complement numbers in storage which are needed in multiplication seems to be of quite general interest, and I thought I might mention something of which seemingly not too many people are aware.

There is a device available which solves that problem fairly neatly. A relay is attached to whatever storage unit is specified, which picks up if a complement number is read into the storage unit. When the unit is read out, the complement number is converted so that the true number is available either for multiplying or for transferring back to the counter.

Mr. Gellman: I should have mentioned a little more clearly, I suppose, but I am sure everyone realizes that the 602-A converts complements only during punching. It punches a true figure, but the complement figure remains in the storage.

Mr. Bisch: I think Mr. Gellman's example is a very good illustration of a point which would interest engineers. He has completely systemized the problem to the point where the work can be done by someone who doesn't understand anything about it. Thus, computers can carry out the machine calculation, and the only work they have to do is to execute one particular step with care before they go to the next. In this way he is able to free himself of the routine calculations and can devote his time to the more important task of planning new problems.

*The Calculation of Roots of Complex Polynomials Using the IBM Type 602-A Calculating Punch**

JOHN LOWE

Douglas Aircraft Company, Incorporated

COMPLEX polynomials, or real polynomials with complex roots, of the form $\sum_{i=0}^{k} a_{k-i}\, x^i = 0$ are common computing problems. In aircraft engineering they arise frequently in studies of airplane flutter and vibration.

A successful method of solving these polynomials with accounting machines should meet certain requirements:

1. It must be completely automatic and reasonably fast and easy to operate.
2. It must permit simultaneous attack on all roots.
3. It must be independent of the order of the equations to be solved.
4. It must be able to accommodate a wide decimal range to permit solution of equations whose roots vary in order of magnitude.

The system presented here satisfies requirements 1, 2, and 3, and it is satisfactory for many purposes with regard to requirement 4.

Theory

Newton's method of successive approximations where

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \qquad (1)$$

is the basis of this system. $f(x_n)$ and $f'(x_n)$ are calculated by synthetic division.

The synthetic division process may be illustrated using a third degree equation:

$$f(x) = a_0x^3 + a_1x^2 + a_2x + a_3 = 0\ .$$

Then, writing only the coefficients and letting x_n be some complex number for which $f(x_n)$ and $f'(x_n)$ are desired,

$$\begin{array}{cccc} & a_0 & a_1 & a_2 \\ x_n & b_0 = a_0, & b_1 = a_1 + b_0x_n, & b_2 = a_2 + b_1x_n, \\ & c_0 = a_0, & c_1 = b_1 + c_0x_n, & f'(x_n) = b_2 + c_1x_n \end{array}$$

$$a_3$$
$$f(x_n) = a_3 + b_2x_n\ . \qquad (2)$$

As discussed later, $f(x_n)$ *and* $f'(x_n)$ are calculated in one run through the IBM Type 602-A Calculating Punch, and x_{n+1} in a second run.

Using a standard, full-capacity 602-A, it seems that the real and imaginary parts of x_n can be expressed to six digits each and the complex quantities a_i, b_i and c_i to eleven digits (twelve if the storage unit conversion device is available).

Since all numbers must be handled according to some fixed decimal point, it is frequently desirable to effect a coordinate transformation, $x' = cx$, such that the largest root of the transformed equation is near unity. Such a transformation permits groups of dissimilar equations to be handled simultaneously with a fixed decimal point. If, as is the case in airplane flutter analysis, the equations are characteristic equations of complex matrices, it is advantageous to make this transformation on the matrix. Incidentally, such a transformation on the matrix greatly facilitates the process of deriving the characteristic equation.

An automatic application of Newton's method, as described here, will not always find all roots from first approximations. The difficulties are discussed in many numerical analysis texts. On the other hand, it is always possible to obtain some of the roots and, thus, by manual analysis, arrive at better approximations to the missing roots. These can be used as the basis for further machine iteration.

Another approach to the problem of missing roots is to factor known roots from the original equation (this can be done by the synthetic division process), and operate on the reduced equation $\sum_{i=0}^{k-1} b_{k-i}\, x^i = 0$. In fact, if nothing is known about the roots, it may be advisable to start with an initial approximation of $x_0 = 0$, find a root, factor it out and repeat on the reduced equation. However, caution should be exercised in factoring out large roots, as it may be necessary to know such roots to a relatively large number of significant figures to obtain an accurate reduced equation.

*This paper was presented by title.

Procedure

One detail card is punched for each complex coefficient, a_i, of the equation. These are arranged in order of subscript, constant term last. A master card containing x_n is filed in front and a blank trailer behind. The group is run through the 602-A, the master and detail cards skipped out without punching, and complex $f(x_n)$ and $f'(x_n)$ punched in the trailer. The detail cards are sorted out, and the master and trailer run through the 602-A again to calculate x_{n+1} by equation (1). Also, in the second run, an 11 is punched in the trailer to make it a master. It may then be filed in front of the detail, a new trailer placed behind, and the cycle repeated.

If a number of equations are being handled simultaneously, the cards may be collated after x_{n+1} is calculated and both cards pulled if $x_n = x_{n+1}$. By match-merging the detail behind the remaining new masters, roots which have converged will automatically be eliminated.

Obviously, any number of groups for one equation can be going through the procedure simultaneously. Thus, one or more approximations to each root can be handled at the same time.

A good check on the work is provided by listing both masters after x_{n+1} is formed. If $f(x)$ is decreasing and x is not changing rapidly, it can be assumed that the process is working correctly.

Thus, the operations are seen to be extremely simple and can, in fact, be handled by an operator with minimal knowledge of machine accounting. This leaves a discussion of the two 602-A control panels.

The synthetic division control panel[a] can be explained by reference to (2)

Master	Read-in new x and clear old $f(x)$ and $f'(x)$	
a_0 detail	$x0 + a_0 = b_0$	$x0 + 0 = 0$
a_1 detail	$xb_0 + a_1 = b_1$	$x0 + b_0 = c_0$
a_2 detail	$xb_1 + a_2 = b_2$	$xc_0 + b_1 = c_1$
a_3 detail	$xb_2 + a_3 = f(x)$	$xc_1 + b_2 = f'(x)$
Trailer	Punch $f(x)$ and $f'(x)$	

Note that the formation of both b_i and c_i follows the same pattern. If b_i is calculated and the program unit allowed to repeat itself, c_i can be formed with little selection of program exits.

a_0 is often equal to unity. If this is true for a particular group to be handled, $b_0 = 1$ can be emitted into storage on the master card, and the a_0 detail card omitted.

The quantities b_i and c_i, which are often negative, must be placed in storage units and used as multiplicands. Hence, the wiring can be simplified, time saved, and one more digit carried if the 602-A is equipped with storage unit conversion devices.

The control panel to form x_{n+1} by equation (1) is straightforward. The square of the modulus of $f'(x)$ is formed, $f(x)$ is multiplied by the conjugate of $f'(x)$ and the two indicated divisions and subtractions performed. x_{n+1} can be calculated to eight decimal digits, although only six can be used in the synthetic division. This may be desirable as when the root has converged to six digits, the next approximation may be accurate to seven or eight digits.

If both $f(x)$ and $f'(x)$ are small, the products of these by the conjugate of $f'(x)$ may be very small, and the quotients may have insufficient significant figures. In this case, the trailers may be reproduced before the calculation of x_{n+1} multiplying both $f(x)$ and $f'(x)$ by a power of 10. Obviously, the same device will work if $f(x)$ and $f'(x)$ are too large.

Factoring out a root and punching the reduced equation may be accomplished simply. The coefficient cards are reproduced, omitting amounts and common punching. These new cards are filed behind the original coefficient cards by i. The master containing the root is filed in front, and the entire deck is run through the 602-A. The blanks serve as trailers and receive the coefficients of the reduced equation, the amount in the last one being the residual and providing a check on the work. The 602-A control panel for this operation accommodates eight digits of x, uses the same synthetic division process described above, and is quite easy to wire.

As an example of time required to obtain roots by this method, consider a group of sixth degree complex equations with coefficients given to not more than eleven (or twelve) decimal digits. Evaluation of $f(x_n)$ and $f'(x_n)$ for one root requires about one minute. Thus, if the roots converge after an average of five iterations, it requires about one-half hour per equation to obtain all the roots.

[a]The writer will supply copies of the layouts for the 602-A control panels on request. Address John Lowe, Douglas Aircraft Company, Inc., Engineering Department, Santa Monica, California.

*Practical Inversion of Matrices of High Order**

WILLIAM D. GUTSHALL

North American Aviation, Incorporated

THE PURPOSE of this paper is to give an account of experience in the inverting of matrices of high order, obtained in connection with the engineering problems of North American Aviation. Because information on this subject for really large matrices (i.e., for $n >> 10$) is notoriously sparse in the literature of the subject, and because of the great importance of the problem in many fields, such a paper should be of interest to this seminar. In addition, it is hoped that research organizations will be stimulated to make investigations, which we, not being a research organization, are neither qualified for, nor encouraged to carry out.

When simultaneous linear algebraic equations are met in engineering work, as they frequently are, it is usually necessary to solve a number of systems having the same coefficients, but different constants on the right-hand sides. It is shown in reference 1 that if the number of systems is at least four (i.e., if there are at least four columns of constants), it is economical to compute the inverse matrix. In our work this is generally the case, and for this reason matrix inversion is important.

It is not practical to perform the inversion process manually (with desk machines) unless $n < 10$, and even for these small matrices, if it is necessary to invert a large number of them at any one time, the work is done much more quickly, and accurately with the help of IBM. Therefore, both systems of high order ($n > 10$) and large groups of small systems are handled by the IBM group.

The small systems are handled quite successfully by a variant of the well-known Aitken method. To give an idea of the efficiency with which this work is done, 48 fifth-order matrices were recently inverted and checked in 16 hours—an average of 20 minutes per inversion. This work was done using the IBM Type 602 Calculating Punch; use of the IBM Type 604 Electronic Calculating Punch would cut down the time somewhat, but not greatly so because of the large amount of card handling involved. Although this is far from being the most efficient use of the machines, those with experience in numerical inversion will recognize it as being amply justified.

The large systems present special difficulties which remain to be solved. Not only does the number of operations increase enormously with the order, making the process very slow, but also such systems rapidly tend toward instability as the order increases—i.e., the rounding errors, which are inevitable—accumulate in a serious way. In this connection, see references 1 and 2.

Because of the inherent instability of the direct methods, several well-known iterative methods (classical iteration, and the method of steepest descents) were tried. For large systems convergence of these methods is much too slow. Convergence is theoretically assured for positive definite matrices. Positive definiteness was guaranteed by the expedient of taking $A'A$, where A' is the transpose of A, the matrix in question, and

$$A^{-1} = (A'A)^{-1} A'.$$

Despite the theoretical assurance of convergence, in this case, numerous iterations showed no evidence of convergence at all! We probably did not give sufficient trial to this approach, but our negative results with the methods tried are confirmed by other investigators.[3] Iteration has its value in the improvement of an approximate solution obtained by some other means.

In choosing among the many direct methods, several considerations are important: (1) The number of elementary operations should be a minimum. (2) The arrangement of work and order of operations should be convenient with respect to the peculiarities of the machines.

Concerning the first requirement, it is asserted in reference 2 that the elimination method requires fewer operations than other known methods. There are numerous methods, however, which can be considered to be but slight variations of the elimination method. The method of partitioning of matrices, for example, is a generalization of the elimination method, and the various methods which involve pivotal reduction—those of Aitken, Crout, Doolittle, Gauss, etc.—are closely similar and require about the same number of operations. The method of determinants is an example of a method definitely inferior to those mentioned above.

With respect to the second requirement, suitability for the machines, methods which include such things as repeated cross multiplication are to be avoided.

*This paper was presented by title by Paul E. Bisch.

A direct method which fits these requirements is a new variant (as yet unpublished) of the elimination method developed by Mr. Charles F. Davis of our IBM group. Although this method has several features to commend it for use with IBM machines (and with desk machines as well) it is not claimed that the successful inversion of several large matrices could not have been achieved by other methods. The method used was simply the elimination method with some new twists. Since the details are still being written up they cannot be given here.

The point of most interest is that inversion of a matrix as large as 88 by 88 has actually been carried out satisfactorily, using standard IBM equipment. The prevailing opinion among the authorities is that the inversion of a matrix of this order is a practical impossibility (cf. reference 3, page 2 and pages 6-8). In a sense this may be true, for this first attempt at inverting an 88th order matrix took about nine weeks and involved between 60 and 70 thousand cards. The better performance which should come with experience might still be prohibitively long and expensive from the engineering point of view. Nevertheless, the degree of success we have had seems hopeful.

A word of explanation should be given about what we have considered "satisfactory" in the way of accuracy. Unfortunately this is a difficult question if one demands precise limits. The question might be phrased this way: If a solution of a linear system is substituted in the original equations and all the remainders are small, is the error in the solution small? How small? If the system is

$$AX - B = 0$$

and a solution X_1 is substituted there results a column of remainders R_1

$$AX_1 - B = R_1 \ .$$

Elementary matrix theory gives the answers to the above questions in terms of the norms of the column vectors R_1 and $E_1 = X - X_1$, and the quantities λ and μ, the upper and lower bounds, respectively, of the matrix A.

$$1/\lambda \, |R_1| \leqq |E_1| \leqq 1/\mu \, |R_1| \ .$$

From this we see that if λ is large and μ small, the limits of error can be very wide; in particular if μ is sufficiently small, $|E_1|$ may be large, although $|R_1|$ is small. Thus, what is often considered to be good check may conceal large errors. The main difficulty, however, is that the quantities λ and μ are not known, and the work required to find good estimates of them is usually prohibitive.

We have been obliged to get around this difficulty in a manner rather unsatisfying to the mathematician but wholly acceptable to the engineer. The engineer looks at the numerical results, and, with physical intuition as a guide, decides whether they are reasonable. To take an example, the solution of a set of 66 equations gave us the stress distribution of a sheet stringer combination. The regularity of the results and confirmation of what should be expected on the basis of experimental evidence convinced engineers that this was the "right" answer. Whether a given numerical value found in this way is correct to two, three, or more significant figures is not known.

There is still another rough indication of accuracy. The experienced IBM operator working in a fixed field (8 digits in our case) can fairly well tell when things are behaving nicely or not. Common to all variants of the elimination method is a division at each reduction. The continued product of these divisors is the determinant, and although the determinant is large, some of these divisors may be small. The occurrence of small divisors means the loss of significant figures, i.e., the process blows up. In the reduction process of the 88 equations, as well as the 66, this difficulty was not apparent. Furthermore, iteration of solutions obtained showed convincing evidence of rapid convergence.

Strictly speaking, the inverses of these large matrices were only partially determined. The earlier statement that it is economical to compute the inverse when there are more than four columns of constants needs qualification. It is not strictly true for the matrices of quasi-diagonal character (explained later) dealt with by us. However, the process effectively leads to a decomposition into diagonal and semi-diagonal factors, from which with some additional work the inverse can be found explicitly.

It is true that no general conclusions can be drawn from such limited experience as ours with matrices of a particular type. But matrices of this type occur frequently in structural analysis and elsewhere. The matrices spoken of may be considered to be made up of the finite difference approximants of linear partial differential equations. Each stepwise approximant has the important feature that coefficients of all but a few of the unknowns are zero. Thus, these coefficients can be arranged in such a way that the matrix has large triangles of zeros in the upper right and lower left corners. Such a matrix might be called *quasi-diagonal* and is certainly one of the most important types. Since the limit case, a diagonal, is stable (unless some diagonal element is zero) and trivially easy to invert, it is plausible to suppose that quasi-diagonal matrices are particularly stable. It is suggested that our experience lends weight to this supposition and that a study of quasi-diagonal matrices will yield much more optimistic estimates of precision than are found in the literature.

On page 1023 of reference 2, von Neumann and Goldstine conclude that, "for example, matrices of the orders 15, 50, 150 can usually be inverted with a (relative) precision of 8, 10, 12 decimal digits less, respectively, than the number of digits carried throughout." By "usually" is meant that "if a plausible statistic of matrices is assumed, then these estimates hold with the exception of a low probability minority." These conclusions, which are the result of a thorough analysis, are valuable to those who are anticipat-

ing the automatic computing machines of the future, but to those who think it might be practical to invert certain types of large matrices, using standard IBM equipment here and now, they seem unduly pessimistic.

The critical question is that of "a plausible statistic of matrices." The estimates of von Neumann and Goldstine are made in terms of λ and μ, the upper and lower bounds, respectively, of the matrix—quantities not known in advance and very difficult to determine. The numerical estimates quoted above are the results of introducing statistical hypotheses concerning the sizes of these quantities. This is done in the form of the results of V. Bargmann concerning the statistical distribution of proper values of a "random" matrix. It is possible that a similar study of the *quasi-diagonal* type, defined in this paper, might lead to less discouraging conclusions. It is stated in reference 1, page 59, that these estimates "can also be used in the case when the matrix is not given at random but arises from the attempt to solve approximately an elliptic partial differential equation by replacing it by n linear equations." But no reason is given, and perhaps this manner of stating the point indicates a certain lack of conviction. However, the original source of these estimates is not accessible to this writer.

To summarize: On the basis of limited experience inverting matrices, it appears to us at North American that, contrary to prevailing opinion, it might be practical to invert certain important types of matrices of high order, using standard IBM equipment. What is needed is further statistical study of these types, and, if the estimates of precision so obtained are favorable, a comparative study of known methods of inversion.

REFERENCES

1. V. Bargmann, D. Montgomery, and J. von Neumann, "Solution of Linear Systems of High Order," U. S. Navy Bureau of Ordnance Contract NORD 9596, 25 (October, 1946).
2. J. von Neumann and H. H. Goldstine, "Numerical Inverting of Matrices of High Order," *Bulletin of the American Mathematical Society,* Vol. 53 (November 11, 1947), pp. 1021-1099.
3. H. Hotelling, "Some New Methods in Matrix Calculation," *Annals of Mathematical Statistics,* Vol. 14 (1943), pp. 1-34.