Neville, Eric Harold
On the solution of numerical functional equations

Mr. E. H. NEVILLE.

HE SOLUTION OF NUMERICAL FUNCTIONAL EQUATIONS.

ON THE SOLUTION OF NUMERICAL FUNCTIONAL EQUATIONS

Illustrated by an Account of a Popular Puzzle and of its Solution.

By ERIC H. NEVILLE.

[Received September 22nd, 1914.—Read December 10th, 1914.—
Received, in revised form, February 16th, 1915.]

[*Extracted from the Proceedings of the London Mathematical Society, Ser.* 2, *Vol.* 14, *Part* 4.]

THE object of this paper is to introduce a method of solving simultaneous numerical functional equations which is original and is thought to be new. Any process of solution determines the value of each variable as the sum of a series whose terms are calculated in succession, but the practical value of a process depends not merely on the rate at which the several series converge, but also on the labour involved in evaluating the terms; in the method proposed this labour is almost beyond comparison less than in the classical method.

The problem is introduced and illustrated by a particular case, arising out of a showman's puzzle that has interested a large number of mathematicians, and when the general method has been explained and its application to two special problems of importance indicated, the equations arising from the puzzle are solved and some interesting consequences of the solution described.

I.

A familiar figure at fairs and shows is the sportsman with a cloth on which a large circle is painted and five smaller equal circular discs of thin metal, who offers the holiday-maker some considerable reward if he can lay the five discs on the cloth in such a way that no part of the large circle can be seen, the experimenter of course paying for each attempt. At a time when work of a more serious kind was for a few days impossible to me, I welcomed the problem of calculating the best arrangement of the discs, and the least value of the ratio of their radius to that of the painted circle, and the results may interest others. The equations on which the

solution of the problem depends were much easier to find, and much simpler in form, than I had expected, but the numerical resolution of these equations with sufficient accuracy seemed likely to prove intolerably tedious, until an original method of dealing with them was devised: what I am anxious to put into the hands of mathematicians is this method.

In deciding on the general features of the most efficient arrangement of the discs, I was helped by having a specimen of the apparatus actually used (belonging to Mr. J. H. Grace, whose kindness in lending it both then, and on the occasion of the reading of this paper to the Society, I gladly acknowledge); it is only to be expected that no great margin is left to the inaccurate speculator, and certain types of arrangement were seen unmistakably to be ineffective; such was, for example, the arrangement symmetrical about each of five diameters, the small circles all passing through the centre of the large circle. It is taken for granted that there is symmetry about one line, a common diameter of the large circle and of one of the small circles. If K is the centre of the large circle, D the centre of this small circle, B the end of the diameter DK of the large circle which is not covered by the small circle, C the point in which the small circle cuts DB, and G, H the points in which the small circle cuts the large circle, the arrangements between which decision must be made can be enumerated. Two circles must pass through B, and intersect in a point L in DB, which may be identical with C, or may be a distinct point in CB; let one of these circles cut the arc BG of the large circle in E, the other cut the arc BH in F. Of the remaining circles one covers E and G, the other covers F and H. If L is distinct from C, the circle covering E and G covers also L and C, and either passes through three of the four points E, G, L, C or has the line joining two of them for a diameter. If L coincides with C, the circle BCE cuts the circle whose centre is D in a point M distinct from C, and the circle covering E and G either is the circle through E, G, and M, or has one of the lines GM, ME, EG for its diameter. It would be possible to apply calculation to each case, but actual trial is sufficient to convince that the only arrangement which allows success with the apparatus used is of the last type; what remains for calculation is the discovery of the smallest ratio of the common radius of the discs to the radius of the painted circle which allows this most effective arrangement to succeed, and the determination of the corresponding position of the point we have denoted by C.

Let b be the radius of the large circle, a that of the small circles, let c be the distance KD, and let the angles BKE, DKG, BCE, KDG be 2θ, ϕ, ψ, $\pi-\chi$. It is a/b that it is our chief object to find, but $(a-c)/b$

derives considerable interest from being extremely small. The equations

$$a \sin \psi = b \sin \theta, \qquad (a-c) \sin \psi = b \sin (\psi - 2\theta),$$

$$c = b \cos \phi - a \cos \chi, \qquad a \sin \chi = b \sin \phi,$$

implicitly determine θ, ϕ, ψ, χ in terms of c. The best arrangement of the first three discs, a, b being supposed constant, is that in which c has the value which makes EG, FH as small as possible, that is, is found by making $2\theta+\phi$ a maximum subject to these equations; the condition for this is

$$\sin 2\psi \sin(\chi-\phi) = \cos \chi \{1-\cos 2(\psi-\theta)\},$$

and if, when the position is such that the five equations involving c, θ, ϕ, ψ, χ are all satisfied, EG and FH are less than $2a$, the last two discs can be set down completely to cover what is left exposed of the large circle. The least value of a/b which allows the covering is that in which EG, FH, found as before, are equal to $2a$, and this value of a/b is found by adding to the five equations already written down, the equation

$$a = b \cos (\theta + \tfrac{1}{2}\phi).$$

For purposes of calculation, the ratios $a : b : c$ are eliminated from the equations, and the resulting equations are taken in a form involving sums and differences, not products and quotients, of circular functions; the equations are

$$2 \sin \theta - \sin (\theta + \tfrac{1}{2}\phi + \psi) - \sin (\psi - \theta - \tfrac{1}{2}\phi) = 0,$$

$$2 \sin \phi - \sin (\theta + \tfrac{1}{2}\phi + \chi) - \sin (\chi - \theta - \tfrac{1}{2}\phi) = 0,$$

$$2 \sin \theta + \sin (\chi + \theta) - \sin (\chi - \theta) - \sin (\psi + \phi) - \sin (\psi - \phi) - 2 \sin (\psi - 2\theta) = 0,$$

$$\cos (2\psi - \chi + \phi) - \cos (2\psi + \chi - \phi) - 2 \cos \chi + \cos (2\psi + \chi - 2\theta) + \cos (2\psi - \chi - 2\theta) = 0,$$

and what we require is a numerical solution, which can be based on the fact that a crude approximation is given by the values $36°$, $36°$, $72°$, $72°$ for θ, ϕ, ψ, χ.

It is evident that these particular equations are algebraic and even rational in the tangents of the angles $\frac{1}{2}\theta$, $\frac{1}{4}\phi$, $\frac{1}{2}\psi$, $\frac{1}{2}\chi$, so that an algebraic equation could be found for $\cos (\theta + \frac{1}{2}\phi)$, which is the number of greatest interest; to calculate by Horner's, or any other of the familiar methods, the root to which $\cos 54°$ is a rough approximation is in theory simplicity itself. The briefest effort will convince the reader that it is impracticable to solve the problem on these lines. The alternative is to regard the

equations as functional, rather than algebraic, and to accept suggestions from the differential calculus: to a consideration of numerical functional equations we now turn.

II.

Of all the common problems of mathematics whose solutions in theory are both simple and complete, probably in no other is the application of the theoretical solution to a numerical case as tedious, and in no other does this application find the mathematician who is not an accountant at as serious a disadvantage, as in the problem of solving a set of numerical functional equations. Let us outline the classical process in the case of three equations

$$F(x, y, z) = 0, \quad G(x, y, z) = 0, \quad H(x, y, z) = 0,$$

of which it is known that an approximate solution is given by

$$x = a_1, \quad y = b_1, \quad z = c_1.$$

If the corresponding accurate solution is

$$x = a_1+x_1, \quad y = b_1+y_1, \quad z = c_1+z_1,$$

an approximation to the values of x_1, y_1, z_1 is given by

$$x_1 = a_2', \quad y_1 = b_2', \quad z_1 = c_2',$$

where

$$a_2'F_x(a_1, b_1, c_1)+b_2'F_y(a_1, b_1, c_1)+c_2'F_z(a_1, b_1, c_1) = -F(a_1, b_1, c_1),$$

$$a_2'G_x(a_1, b_1, c_1)+b_2'G_y(a_1, b_1, c_1)+c_2'G_z(a_1, b_1, c_1) = -G(a_1, b_1, c_1),$$

$$a_2'H_x(a_1, b_1, c_1)+b_2'H_y(a_1, b_1, c_1)+c_2'H_z(a_1, b_1, c_1) = -H(a_1, b_1, c_1);$$

the second approximation to the values of x, y, z is given by

$$x = r_2', \quad y = s_2', \quad z = t_2',$$

where

$$r_2' = a_1+a_2', \quad s_2' = b_1+b_2', \quad t_2' = c_1+c_2'.$$

To obtain a third approximation to the values of x, y, z the process is repeated, terms a_3', b_3', c_3' being obtained from three equations of which the first is

$$a_3'F_x(r_2', s_2', t_2')+b_3'F_y(r_2', s_2', t_2')+c_3'F_z(r_2', s_2', t_2') = -F(r_2', s_2', t_2'),$$

and these terms being added to those already known. We may say that the exact roots r, s, t are the sums of series $\Sigma a_m'$, $\Sigma b_m'$, $\Sigma c_m'$ whose terms

are calculated in successive triads from three equations such as

$$a'_{m+1}F_x(r'_m, s'_m, t'_m)+b'_{m+1}F_y(r'_m, s'_m, t'_m)+c'_{m+1}F_z(r'_m, s'_m, t'_m) = -F(r'_m, s'_m, t'_m),$$

r'_m, s'_m, t'_m being the sums to m terms of the series whose m-th terms are a'_m, b'_m, c'_m and the first terms a'_1, b'_1, c'_1 being the given first approximations a_1, b_1, c_1.

This process is open to a twofold criticism. The solution of simultaneous linear algebraic equations is by no means the attractive process in practice that it is in theory: if there are only a few equations and the coefficients are small integers the solution is not prohibitively troublesome, especially if a multiplication table or a mechanical multiplier is accessible, but given coefficients with three or four significant digits, the operation is excessively tedious, and is one in which mistakes are easy to make, and if the coefficients are irrational it is troublesome to decide the degree of accuracy advisable at each stage. And in the classical process, although the coefficients of the variables, the first derivatives of the functions, change but slightly from step to step of the approximation, the labour of solving one set of linear equations is not in the least diminished in virtue of the work done in solving the earlier sets, and, of course, increases with the degree of accuracy maintained. In the alternative process which is to be described, use is made of the fact that the coefficients vary but little, and subsequently it is pointed out that the process is applicable if the coefficients are actually constant and can be used in the solution of a set of linear algebraic equations with effect if the coefficients are complicated or irrational.

If the coefficients and constant terms in one set of linear algebraic equations differ but little from the coefficients and constant terms in another set, the values of the variables which satisfy the one set differ from the values which satisfy the other by amounts which are small compared with the values themselves, it being assumed that the determinants of the coefficients are not small. Thus, if $\lambda_1, \mu_1, \ldots, \nu_3$ are any close approximations to

$$F_x(a_1, b_1, c_1), \quad F_y(a_1, b_1, c_1), \quad \ldots, \quad H_z(a_1, b_1, c_1),$$

the values of the nine first derivatives of three functions $F(x, y, z)$, $G(x, y, z)$, $H(x, y, z)$ for values a_1, b_1, c_1 of the variables for which the functions are known to be small simultaneously, and if a''_2, b''_2, c''_2 are such as to satisfy the equations

$$\begin{aligned}
\lambda_1 a''_2+\mu_1 b''_2+\nu_1 c''_2 &= -F(a_1, b_1, c_1),\\
\lambda_2 a''_2+\mu_2 b''_2+\nu_2 c''_2 &= -G(a_1, b_1, c_1),\\
\lambda_3 a''_2+\mu_3 b''_2+\nu_3 c''_2 &= -H(a_1, b_1, c_1),
\end{aligned}$$

then a_1+a_2'', b_1+b_2'', c_1+c_2'', which we may denote by r_2'', s_2'', t_2'', differ from the second approximations r_2', s_2', t_2' of the classical solution by amounts of the order $g_1''f_2''$, where f_2'' is the greatest of the moduli of the three terms a_2'', b_2'', c_2'' and g_1'' is an approximation factor* depending on the accuracy with which λ_1, μ_1, ..., ν_3 represent

$$F_x(a_1, b_1, c_1), \quad F_y(a_1, b_1, c_1), \quad \ldots, \quad H_z(a_1, b_1, c_1).$$

Since r_2', s_2', t_2' differ from the exact solutions r, s, t by amounts of order $f_2''^2$, r_2'', s_2'', t_2'' differ from r, s, t by amounts of order $(g_1''+f_2'')f_2''$; if f_2'' is of a higher order than g_1'', the approximation furnished by r_2'', s_2'', t_2'' is not as close as that furnished by r_2', s_2', t_2', but the somewhat arbitrary nature of λ_1, μ_1, ..., ν_3 enables us to take for them rational numbers which can all be comparatively simple without the value of g_1'' becoming unduly large, and so to render the calculation of r_2'', s_2'', t_2'' a much simpler matter than the calculation of r_2', s_2', t_2'.

To obtain a closer approximation than that given by r_2'', s_2'', t_2'', we have to solve linear equations which either are equations such as

$$a_3''F_x(r_2'', s_2'', t_2'')+b_3''F_y(r_2'', s_2'', t_2'')+c_3''F_z(r_2'', s_2'', t_2'') = -F(r_2'', s_2'', t_2''),$$

or are equations whose coefficients and constant terms differ little from those of these equations. Now it is assumed throughout that none of the second derivatives of the functions F, G, H are large compared with the largest of the first derivatives, and it follows that numbers λ_1, μ_1, ..., ν_3, which are approximations to

$$F_x(a_1, b_1, c_1), \quad F_y(a_1, b_1, c_1), \quad \ldots, \quad H_z(a_1, b_1, c_1),$$

are approximations also to

$$F_x(r_2'', s_2'', t_2''), \quad F_y(r_2'', s_2'', t_2''), \quad \ldots, \quad H_z(r_2'', s_2'', t_2''),$$

and indeed to the values of the first derivatives for any values of the arguments not differing greatly from the values r, s, t, which we are endeavouring to calculate. Hence, instead of solving equations such as that last written, we may solve equations such as

$$\lambda_1a_3''+\mu_1b_3''+\nu_1c_3'' = -F(r_2'', s_2'', t_2''),$$

which differ only in the constant terms from the equations of the set solved to find a_2'', b_2'', c_2''; the sums r_3'', s_3'', t_3'', that is, $r_2''+a_3''$, $s_2''+b_3''$, $t_2''+c_3''$,

* On the nature of this factor, see Section III below.

differ from r, s, t by amounts of order $(g''_2+f''_3)f''_3$, where f''_3 is the greatest of the moduli of a''_3, b''_3, c''_3 and g''_2 is an approximation factor differing but slightly from g''_1.

The process may be continued indefinitely: the solutions are found as the sums of series $\Sigma a''_m$, $\Sigma b''_m$, $\Sigma c''_m$ whose terms are calculated in successive triads from sets of equations such as

$$\lambda_1 a''_{m+1}+\mu_1 b''_{m+1}+\nu_1 c''_{m+1} = -F(r''_m, s''_m, t''_m),$$

$$\lambda_2 a''_{m+1}+\mu_2 b''_{m+1}+\nu_2 c''_{m+1} = -G(r''_m, s''_m, t''_m),$$

$$\lambda_3 a''_{m+1}+\mu_3 b''_{m+1}+\nu_3 c''_{m+1} = -H(r''_m, s''_m, t''_m),$$

the arguments r''_m, s''_m, t''_m being the m-th partial sums of the series $\Sigma a''_m$, $\Sigma b''_m$, $\Sigma c''_m$, and the coefficients λ_1, μ_1, ..., ν_3 being the same at every step: the remainder of each series after m terms is of order $(g''_{m-1}+f''_m)f''_m$, where f''_m is the greatest of the moduli of a''_m, b''_m, c''_m, and g''_{m-1} depends on the accuracy with which λ_1, μ_1, ..., ν_3 represent

$$F_x(r''_{m-1}, s''_{m-1}, t''_{m-1}),\ F_y(r''_{m-1}, s''_{m-1}, t''_{m-1}),\ \ldots,\ H_z(r''_{m-1}, s''_{m-1}, t''_{m-1}),$$

and so tends as m increases to a definite limit g'' dependent on the nearness of the coefficients to $F_x(r, s, t)$, $F_y(r, s, t)$, ..., $H_z(r, s, t)$. Since f''_m tends to zero, but g''_{m-1} as a rule does not, we may say that in general the remainder of each series is of order $g''f''_m$, where g'' is a fractional approximation factor, though in exceptional cases the remainder may be of order f''^2_m.

When we have to solve a number of sets of linear equations with common rational coefficients but different constants, the most tedious part of the work, and, what is equally important in practice, the part of the work in which the greatest care is needed if mistakes are to be avoided, can be performed once for all. From the set of coefficients λ_1, μ_1, ..., ν_3 a set ρ''_1, σ''_1, ..., τ''_3, the reciprocal set with the sign of every member changed, can be found such that the set of relations

$$\lambda_1 a+\mu_1 b+\nu_1 c = -u,\quad \lambda_2 a+\mu_2 b+\nu_2 c = -v,\quad \lambda_3 a+\mu_3 b+\nu_3 c = -w,$$

between three variables a, b, c and three variables u, v, w is equivalent to the set of relations

$$a = \rho''_1 u+\sigma''_1 v+\tau''_1 w,\quad b = \rho''_2 u+\sigma''_2 v+\tau''_2 w,\quad c = \rho''_3 u+\sigma''_3 v+\tau''_3 w,$$

and the nine coefficients ρ''_1, σ''_1, ..., τ''_3 having once been calculated, the

successive triads of which we are in search are obtained from equations such as

$$a''_{m+1} = \rho''_1 F(r''_m, s''_m, t''_m) + \sigma''_1 G(r''_m, s''_m, t''_m) + \tau''_1 H(r''_m, s''_m, t''_m).$$

A last simplification suggests itself immediately. If $\lambda_1, \mu_1, \ldots, \nu_3$ are rational, the coefficients $\rho''_1, \sigma''_1, \ldots, \tau''_3$ are rational, but they may be complicated fractions. It is pointless that the relations such as

$$a = \rho''_1 u + \sigma''_1 v + \tau''_1 w$$

should represent the relations such as

$$\lambda_1 a + \mu_1 b + \nu_1 c = -u,$$

with greater accuracy than that with which the coefficients $\lambda_1, \mu_1, \ldots, \nu_3$ represent the derivatives $F_x, F_y, \ldots, H_z$, and therefore we may facilitate calculation by substituting for the coefficients $\rho''_1, \sigma''_1, \ldots, \tau''_3$ any coefficients $\rho_1, \sigma_1, \ldots, \tau_3$, which do not differ greatly from them, the effect being to substitute for the set of approximation factors $g''_1, g''_2, \ldots$, a set of approximation factors $g_1, g_2, \ldots$ tending to a limit g which may be either larger or smaller than the limit g''. So, finally, the solutions r, s, t are the sums of series $\Sigma a_m, \Sigma b_m, \Sigma c_m$, whose terms are calculated in successive triads from the equations

$$a_{m+1} = \rho_1 F(r_m, s_m, t_m) + \sigma_1 G(r_m, s_m, t_m) + \tau_1 H(r_m, s_m, t_m),$$

$$b_{m+1} = \rho_2 F(r_m, s_m, t_m) + \sigma_2 G(r_m, s_m, t_m) + \tau_2 H(r_m, s_m, t_m),$$

$$c_{m+1} = \rho_3 F(r_m, s_m, t_m) + \sigma_3 G(r_m, s_m, t_m) + \tau_3 H(r_m, s_m, t_m),$$

the coefficients $\rho_1, \sigma_1, \ldots, \tau_3$ being constant throughout and simple in form, and the arguments r_m, s_m, t_m being the m-th partial sums of the series $\Sigma a_m, \Sigma b_m, \Sigma c_m$ themselves. The restriction to three equations in three variables has been purely a matter of convenience, and, in general, we have the theorem :

If it is known that an approximate simultaneous solution of any number n of independent functional equations

$$F_p(x_1, x_2, \ldots, x_n) = 0 \quad (p = 1, 2, \ldots, n),$$

in the same number of variables is given by

$$x_p = a_{p1} \quad (p = 1, 2, \ldots, n),$$

then rational coefficients ρ_{st} can readily be obtained, such that, if sets of

numbers a_{pm} are calculated in succession from the formulæ

$$a_{p,\,m+1} = \sum_{q=1}^{n} \rho_{qp} F_q(r_{1m},\ r_{2m},\ \ldots,\ r_{nm}) \quad (p = 1,\ 2,\ \ldots,\ n),$$

where

$$r_{lm} = \sum_{k=1}^{m} a_{lk} \quad (l = 1,\ 2,\ \ldots,\ n),$$

then the n series $\sum_{m=1}^{\infty} a_{pm}$ are convergent, and if the sum of the series Σa_{pm} is r_p, the functional equations are all satisfied for the set of values $r_1, r_2, \ldots, r_n$ of the variables; the series are ultimately dominated by geometric series with a common ratio dependent on the differences between the values chosen for the coefficients and the values of the quotients

$$(-)^{s+t+1} \frac{\partial(F_1, F_1, \ldots, F_{t-1}, F_{t+1}, \ldots, F_{n-1}, F_n)}{\partial(r_1, r_2, \ldots, r_{s-1}, r_{s+1}, \ldots, r_{n-1}, r_n)} \Big/ \frac{\partial(F_1, F_2, \ldots, F_n)}{\partial(r_1, r_2, \ldots, r_n)},$$

unless these differences all happen to vanish, in which case the dominating series is as in the classical method of approximation a series of the form

$$c\,(1 + k + k^2 + k^4 + k^8 + \ldots).$$

Theoretically, the classical method is more powerful than the method described, since its dominating series ultimately converges more rapidly than any geometric series, but in practice each step of the classical method is incomparably more troublesome than the whole group of steps recommended here for advancing the approximation by the same amount.

If in any branch of applied mathematics the solution of sets of numerical functional equations became a daily necessity, it might be worth while to have tabulated sets of rational coefficients ρ_{st} corresponding approximately to standard sets of coefficients λ_{ts}; the undertaking would be weighty, for even with only three equations the number of entries would be very large if the number of different values, positive, zero, and negative, which each coefficient was allowed to assume was adequate.

Before returning to the particular set of functional equations for whose solution the method of this paper was devised, let us refer to the application of the method to the two simplest cases; to find that there is still something to be said on familiar problems is always an encouragement to research.

III.

We have already remarked that if the coefficients of a set of simultaneous linear equations are irrational or are complicated, the solution, to a preassigned degree of accuracy, is troublesome: we can apply our general method to deduce the solution by successive steps from equations with simple coefficients. Thus to solve three equations

$$f_1x+g_1y+h_1z=p,\quad f_2x+g_2y+h_2z=q,\quad f_3x+g_3y+h_3z=r,$$

we take any numbers $\lambda_1, \mu_1, \ldots, \nu_3$ which do not differ greatly from $f_1, g_1, \ldots, h_3$, and find simple rational numbers $\rho_1, \sigma_1, \ldots, \tau_3$, such that the equations

$$\lambda_1x+\mu_1y+\nu_1z=-u,\quad \lambda_2x+\mu_2y+\nu_2z=-v,\quad \lambda_3x+\mu_3y+\nu_3z=-w,$$

between two sets of variables x, y, z and u, v, w are approximately equivalent to the equations

$$x=\rho_1u+\sigma_1v+\tau_1w,\quad y=\rho_2u+\sigma_2v+\tau_2w,\quad z=\rho_3u+\sigma_3v+\tau_3w;$$

then, if

$$\left.\begin{aligned} x_1&=-(\rho_1p+\sigma_1q+\tau_1r)\\ y_1&=-(\rho_2p+\sigma_2q+\tau_2r)\\ z_1&=-(\rho_3p+\sigma_3q+\tau_3r)\end{aligned}\right\}$$

$$\left.\begin{aligned} p_2&=p-(f_1x_1+g_1y_1+h_1z_1)\\ q_2&=q-(f_2x_1+g_2y_1+h_2z_1)\\ r_2&=r-(f_3x_1+g_3y_1+h_3z_1)\end{aligned}\right\}$$

$$\left.\begin{aligned} x_2&=-(\rho_1p_2+\sigma_1q_2+\tau_1r_2)\\ y_2&=-(\rho_2p_2+\sigma_2q_2+\tau_2r_2)\\ z_2&=-(\rho_3p_2+\sigma_3q_2+\tau_3r_2)\end{aligned}\right\}$$

$$\left.\begin{aligned} p_3&=p_2-(f_1x_2+g_1y_2+h_1z_2)\\ q_3&=q_2-(f_2x_2+g_2y_2+h_2z_2)\\ r_3&=r_2-(f_3x_2+g_3y_2+h_3z_2)\end{aligned}\right\}$$

and so on, the series $x_1+x_2+\ldots$, $y_1+y_2+\ldots$, $z_1+z_2+\ldots$ tend to the actual values of x, y, z satisfying the proposed equations. The method is particularly useful if the original equations have irrational coefficients, since it renders it unnecessary to determine in advance how far accuracy in approximating to the values of the coefficients is significant in the attainment of any required degree of accuracy in the solutions, and it renders it possible to make use of any given approximate solution in searching for an approximation still closer. So true is it that when coefficients are complicated an economy is effected by this process, that the

simplest way of conducting an approximation to a solution of a set of functional equations by the classical method is to solve by this method the various sets of linear equations that arise: this mixture of methods is naturally less satisfactory than a frank desertion of the classical method, but when in any case a point is reached from which linear equations would complete the solution to the required degree of accuracy, to write down these linear equations enables the calculator to avoid further reference to tables.

In the case of linear algebraic equations it is easy to shew the validity of the process employed. With the three equations just discussed, we have

$$x_{m+1} =$$
$$\{1+(\rho_1 f_1+\sigma_1 f_2+\tau_1 f_3)\}\, x_m+(\rho_1 g_1+\sigma_1 g_2+\tau_1 g_3)\, y_m+(\rho_1 h_1+\sigma_1 h_2+\tau_1 h_3)\, z_m,$$

and similar expressions for y_{m+1}, z_{m+1}. The nine coefficients

$$1+(\rho_1 f_1+\sigma_1 f_2+\tau_1 f_3),\ (\rho_1 g_1+\sigma_1 g_2+\tau_1 g_3),\ \ldots,\ 1+(\rho_3 h_1+\sigma_3 h_2+\tau_3 h_3)$$

can be made as small as necessary by proper choice of the nine numbers $\rho_1, \sigma_1, \ldots, \tau_3$, and it is evident that if the greatest of their moduli is not greater than $\frac{1}{3}g$, and g is a proper fraction, the series

$$x_1+x_2+\ldots,\ y_1+y_2+\ldots,\ z_1+z_2+\ldots$$

all converge not less rapidly than a geometric series with ratio g. In practice the method may often be found to succeed even if the greatest modulus is not less than $\frac{1}{3}$, for the signs of the various terms do not as a rule combine in the most unfavourable manner conceivable. The method being proved valid for linear algebraic equations, its validity for equations of any form is a consequence of the validity of the classical process.

The other simple case of which we wish to speak is that of a single functional equation. What we say is that if a_1 is a first approximation to a root of a functional equation

$$F(x) = 0,$$

and $-\rho$ is any number not very different from the value of the reciprocal of dF/dx when x is equal to a_1, then the root itself is the sum of the series Σa_m whose partial sums are calculated in succession from the formulæ

$$\left.\begin{array}{llll} a_2 = \rho F(a_1) & a_3 = \rho F(r_2) & a_4 = \rho F(r_3) & a_5 = \rho F(r_4) \\ r_2 = a_1+a_2 & r_3 = r_2+a_2 & r_4 = r_3+a_3 & r_5 = r_4+a_4 \end{array}\right.$$

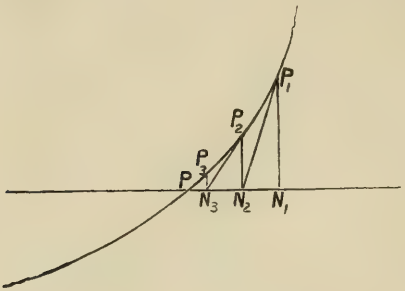
and so on.

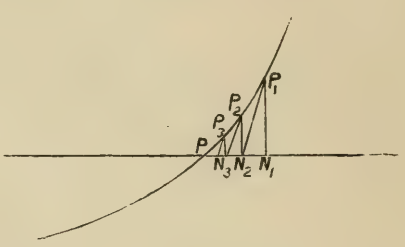
For example, taking $2-x^2$ for $F(x)$ and $1\frac{1}{2}$ as a first approximation, we naturally take ρ to be $\frac{1}{3}$, and we have

$$a_2 = -\ \cdot 083, \quad a_3 = -\ \cdot 0026, \quad a_4 = -\ \cdot 00018, \quad a_5 = -\ \cdot 000006,$$

$$r_2 = 1\cdot 417, \quad r_3 = 1\cdot 4144, \quad r_4 = 1\cdot 41422, \quad r_5 = 1\cdot 414214.$$

*The process of solving a single functional equation may be illustrated graphically, the relation of the method of this paper to the classical method being made apparent. The problem is, given a point N_1 on the axis of x known to be not far from the point P in which a curve $y = F(x)$

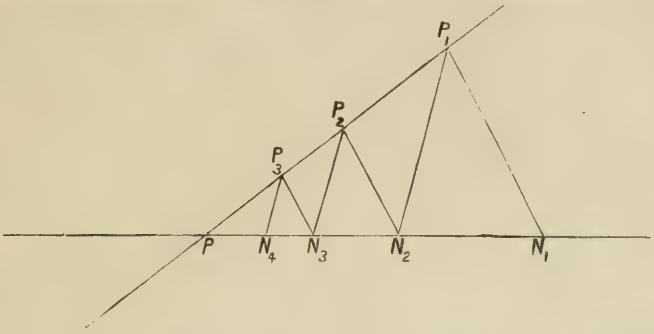




meets that axis, to construct a sequence of points N_1, N_2, N_3, ..., having P for limiting point. In the older method, from N_1 is drawn the ordinate N_1P_1 to the curve (the line parallel to the axis of y, not necessarily the line perpendicular to the axis of x), and the tangent P_1N_2 to the curve at P_1 cuts the axis of x in the second point N_2 of the sequence; N_3 lies on the tangent at the point P_2 in which the line N_2P_2 parallel to N_1P_1 cuts the curve, and so on. In the present method, P_1 is found from N_1, P_2 from N_2, and so on, as before, but the lines P_1N_2, P_2N_3, ... by which N_2 is found from P_1, N_3 from P_2, and so on, are not tangents to the curve, but are lines all parallel to some direction not differing greatly from that of the tangent at P_1. If actual tangents are used, tables must be consulted for the determination not only of the lengths of the ordinates N_1P_1, N_2P_2, ..., but also of the slopes of the tangents P_1N_2, P_2N_3, ...; if the lines P_1N_2, P_2N_3, ... have a common direction, the triangles $N_2N_1P_1$, $N_3N_2P_2$, ... are all similar, and when the ratio of N_2N_1 to N_1P_1 has been found it is only for the values of the ordinates N_1P_1, N_2P_2, ... that reference need be made to tables. This graphical consideration brings out clearly another point. The rapidity of the approximation depends on the closeness of the

* The remainder of this section has been added since the reading of the paper; the graphical illustration was suggested by one of the referees.

common direction of P_1N_2, P_2N_3, ... to the direction of the tangent at P, the point to be found, but the approximation may be valid even if the two directions differ considerably: if PP_1 is actually a straight line, and N_2 is



any point between N_1 and the image of N_1 in P, then by drawing N_2P_2 parallel to N_1P_1, P_2N_3 parallel to P_1N_2, N_3P_3 parallel to N_1P_1, P_3N_4 parallel to P_1N_2, and so on, we can construct a sequence of points converging to P, the convergence being the faster the nearer N_2 is taken to P. Similarly in the more general case, there may be much latitude in the choice of the common direction of P_1N_2, P_2N_3, and the rest, and because of this latitude a choice may be made which gives a simple value to the ratio of N_2N_1 to N_1P_1, of N_3N_2 to N_2P_2, and so on.

Thus we see a connection between the method of this paper and a method used* in a number of problems in applied mathematics. The simplest ratio which N_2N_1 can bear to N_1P_1 is unity, and to draw P_1N_2 in such a direction as to make N_2N_1 equal to N_1P_1 is to evaluate the root of the equation $F(x) = 0$ by the steps

$$a_2 = -F(a_1), \quad a_3 = -F(a_1+a_2), \quad a_4 = -F(a_1+a_2+a_3), \quad \ldots,$$

that is,

$$r_2 = r_1 - F(r_1), \quad r_3 = r_2 - F(r_2), \quad r_4 = r_3 - F(r_3), \quad \ldots.$$

This process is specially useful when the equation to be solved has the particular form

$$x = f(x),$$

for if we write

$$F(x) = x - f(x)$$

we see that the approximation then advances by the steps

$$r_2 = f(r_1), \quad r_3 = f(r_2), \quad r_4 = f(r_3), \quad \ldots,$$

* I am indebted to Dr. Bromwich for drawing my attention to this method,

which are so simple as to justify considerable slowness in the convergence. It can be shewn that convergence is certain (though it may be slow), if $f'(x)$ is positive and less than unity in the neighbourhood of the root, and since if $f^{-1}(x)$ is the function inverse to $f(x)$, the equation $x = f(x)$ is identical with the equation $x = f^{-1}(x)$ and $df^{-1}(x)/dx$ is the reciprocal of $df(x)/dx$, it follows that the root of the equation $x = f(x)$ can be found either by the sequence

$$r_2 = f(r_1), \quad r_3 = f(r_2), \quad r_4 = f(r_3), \quad \ldots,$$

or by the sequence

$$r_2 = f^{-1}(r_1), \quad r_3 = f^{-1}(r_2), \quad r_4 = f^{-1}(r_3), \quad \ldots,$$

provided only that $f'(x)$ is positive and different from unity near the required root.

To more equations than one, even to a set given in such a form as

$$x = f(x, y, z), \quad y = g(x, y, z), \quad z = h(x, y, z),$$

this process is not necessarily adaptable, but the knowledge of the process may well be an encouragement to the use of very crude approximations in choosing the rational coefficients which we have denoted in general by

$$\rho_{11}, \rho_{21}, \ldots, \rho_{nn}.$$

IV.

In illustration of the general method we have described, we give some details of the calculation in the case of the equations connected with the covering puzzle. First, a rough calculation, of which no account need be given, shows that $36°$, $36°$, $72°$, $66°$ is a better approximation than $36°$, $36°$, $72°$, $72°$ to the solution of the equations. To have only acute angles to consider, we substitute $36°-\alpha$, $36°-2\beta$, $72°+\gamma$, $66°-\eta$ for θ, ϕ, ψ, χ, and the equations solved are

$$-2\sin(36°-\alpha) + \cos(36°-\alpha-\beta+\gamma)+\sin(18°+\alpha+\beta+\gamma) = 0,$$

$$-2\sin(36°-2\beta)+\cos(30°-\alpha-\beta-\eta)+\sin(12°+\alpha+\beta-\eta) = 0,$$

$$-2\sin(36°-\alpha)-\cos(12°-\alpha-\eta)+2\sin(2\alpha+\gamma)+\sin(30°+\alpha-\eta)$$
$$+\cos(18°-2\beta+\gamma)+\sin(36°+2\beta+\gamma) = 0,$$

$$-\cos(6°+2\alpha+2\gamma+\eta)-\cos(6°-2\beta-2\gamma+\eta)+\cos(42°-2\alpha-2\gamma+\eta)$$
$$+\sin(24°-2\beta+2\gamma+\eta)+2\sin(24°+\eta) = 0.$$

The equations which the classical process requires us first to solve are

$$\begin{aligned}
3{\cdot}157\alpha' + 1{\cdot}539\beta' + {\cdot}363\gamma' &= {\cdot}058,\\
1{\cdot}478\alpha' + 4{\cdot}714\beta' - {\cdot}478\eta' &= {\cdot}102,\\
6{\cdot}276\alpha' + 2{\cdot}236\beta' + 2{\cdot}500\gamma' - 1{\cdot}074\eta' &= {\cdot}115,\\
1{\cdot}547\alpha' - 2{\cdot}036\beta' + 3{\cdot}165\gamma' + 2{\cdot}280\eta' &= {\cdot}026,
\end{aligned}$$

α, β, γ, η being expressed in radians; no attempt is made to solve these equations, but the set

$$\begin{aligned}
3\tfrac{1}{6}\alpha + 1\tfrac{1}{2}\beta + \tfrac{1}{3}\gamma &= -t,\\
1\tfrac{1}{2}\alpha + 4\tfrac{3}{4}\beta - \tfrac{1}{2}\eta &= -u,\\
6\tfrac{1}{4}\alpha + 2\tfrac{1}{4}\beta + 2\tfrac{1}{2}\gamma - \eta &= -v,\\
1\tfrac{1}{2}\alpha - 2\beta + 3\tfrac{1}{6}\gamma + 2\tfrac{1}{4}\eta &= -w,
\end{aligned}$$

that is to say

$$\begin{aligned}
19\alpha + 9\beta + 2\gamma &= -6t,\\
6\alpha + 19\beta - 2\eta &= -4u,\\
25\alpha + 9\beta + 10\gamma - 4\eta &= -4v,\\
18\alpha - 24\beta + 38\gamma + 27\eta &= -12w,
\end{aligned}$$

is inverted, not accurately, but to the approximate form

$$\begin{aligned}
\alpha &= -(\tfrac{1}{2}-\tfrac{1}{12})t + \tfrac{1}{7}u + \tfrac{1}{30}w,\\
\beta &= \tfrac{1}{16}t - (\tfrac{1}{4}+\tfrac{1}{60})u + \tfrac{1}{30}v - \tfrac{1}{30}w,\\
\gamma &= (1-\tfrac{1}{4}-\tfrac{1}{40})t - \tfrac{1}{6}u - (\tfrac{1}{4}+\tfrac{1}{30})v - \tfrac{1}{6}w,\\
\eta &= -(1-\tfrac{1}{3}-\tfrac{1}{60})t - \tfrac{1}{9}u + (\tfrac{1}{2}-\tfrac{1}{16})v - (\tfrac{1}{4}+\tfrac{1}{40})w,
\end{aligned}$$

the form* of coefficient used here being the most convenient. Substituting $-{\cdot}058$, $-{\cdot}102$, $-{\cdot}115$, $-{\cdot}026$ for t, u, v, w, for a second approximation to the values of α, β, γ, η (simultaneous zeroes being the first

* If a set of equations with integral coefficients is inverted accurately, the coefficients in the reciprocal set are fractions with a common denominator Δ, the determinant of the original coefficients, and if the set of numbers Δ, $\Delta/2$, $\Delta/3$, ..., $\Delta/(k-1)$ is written down, k being any integer greater than 9, an approximation of any desired accuracy to each coefficient in the convenient form

$$e_0 s_0 + \frac{e_1}{s_1} + \frac{e_2}{10 s_2} + \frac{e_3}{100 s_3} + \frac{e_4}{1000 s_4} + \ldots,$$

where each of the letters e_0, e_1, e_2, ... stands for one of the three numbers 1, 0, -1, and each of the letters s_0, s_1, s_2, ... for a positive integer less than k, can be written down at sight.

approximation), we have*

		70	
		4	84
	29	4	
	1	29	1
	2	17	7
	26	1	7
		15	11
$\alpha_2 =$ ·029	$\beta_2 = -$ ·004	$\gamma_2 = -$ ·058	$\eta_2 =$ ·058
− 5	− 4		− 19
− 15			− 1
− 1	− 8		− 58
− 21			− 78

that is,

$$\alpha_2 = \cdot 008 = 28', \quad \beta_2 = \cdot 021 = 1^\circ 12', \quad \gamma_2 = \cdot 012 = 41', \quad \eta_2 = \cdot 006 = 21'.$$

To find a third approximation, the values of the functions on the left-hand sides of the equations for the values $\alpha_2, \beta_2, \gamma_2, \eta_2$ of the variables are found, this time to five places of decimals; the calculation requires only addition and subtraction of numbers read from tables, and the values found are ·00439, ·00663, ·00516, −·00170. By substitution of these values for t, u, v, w in the formulæ already used, the third terms in the approximation are found to be

$$\alpha_3 = -\cdot 00094 = -3', \qquad \beta_3 = -\cdot 00127 = -4',$$

$$\gamma_3 = \cdot 00090 = 3', \qquad \eta_3 = -\cdot 00087 = -3',$$

and these are, in fact, true to a single minute. For continuing the approximation, the linear equations to which the functional equations are, to seven places of decimals, now equivalent, are written down, the variables being expressed in seconds and denoted by a, b, c, e, and the coefficients being found from the difference columns in the tables. For example, the term $\cos(36^\circ - \alpha - \beta + \gamma)$ in the first function is replaced by

$$\cos\{36^\circ - (\alpha_2 + \beta_2 - \gamma_2) - (\alpha_3 + \beta_3 - \gamma_3) - (a + b - c)\},$$

* This arrangement of the figures, though not elegant in appearance, is the most convenient in practice: the terms are calculated in order, and each one is placed above or below those already written according as it is positive or negative. Separate addition of the positive and negative components is desirable on account of the uncertainty of the sign which is to prevail. After this we give only the results at the various stages, but nothing is actually omitted except such columns as these.

that is, by $\cos\{35°11'-(a+b-c)\}$, and this in turn by

$$\{8173125+(1676/60)(a+b-c)\}\times 10^{-7}.$$

The equations so found, multiplied by 60×10^7, are

$$\begin{aligned}
9137a+\ 4405b+1053c &= -121500,\\
4203a+13879b\ \ -1459e &= -155580,\\
18311a+\ 6168b+7252c-3085e &= -182040,\\
4610a-\ 5597b+9683c+6666e &= -\ 41460,
\end{aligned}$$

equations in which the coefficients of the variables, divided by 2909, that is, by 10^7 times the number of radians in a minute of arc, necessarily differ but little from the rational coefficients in the earlier equations connecting t, u, v, w with $\alpha, \beta, \gamma, \eta$. Four equations such as

$$9137a+4405b+1053c=-k,$$

connecting four variables k, l, m, n with the four variables a, b, c, e are therefore approximately equivalent to four equations of which the first is

$$2909a=-(\tfrac{1}{2}-\tfrac{1}{12})k+\tfrac{1}{7}l+\tfrac{1}{30}n.$$

First approximations to a, b, c, e, or fourth approximations to $\alpha, \beta, \gamma, \eta$, are immediately found to be $-9''$, $-10''$, $1''$, $-10''$; these cannot be trusted to a second, but to proceed one stage further it is necessary only to find four numbers k_2, l_2, m_2, n_2 by substituting these last terms in such equations as

$$k_2=9137a_1+4405b_1+1053c_1+121500,$$

and then to find the final terms in the approximation from such equations as

$$2909a_2=-(\tfrac{1}{2}-\tfrac{1}{12})k_2+\tfrac{1}{7}l_2+\tfrac{1}{30}n_2.$$

It is found that, to two significant figures,

$$k_2=-3700,\quad l_2=-6400,\quad m_2=-6300,\quad n_2=-1000,$$

and that b_2 is slightly less than $\frac{1}{2}''$, while a_2, c_2, e_2 are about $\frac{1}{5}''$, $\frac{1}{9}''$, $\frac{1}{4}''$. The conclusion is that the required solution of the given equations, correct to the nearest second, is

$$\theta=35°35'9'',\quad \phi=33°44'19'',\quad \psi=72°44'1'',\quad \chi=65°42'10''.$$

The value of $\theta+\frac{1}{2}\phi$, the angle whose cosine is the ratio of the radii of the circles, is $52°27'18\frac{1}{3}''$, accurate in point of fact to at least $\frac{1}{20}''$, and the cosine is ·6094183, with an error less than 2 in the last place: we may

say confidently that covering is possible if the ratio exceeds ·6094185, impossible if the ratio is less than ·6094180. The ratio of $a-c$ to b is $\sin 1°33'43\frac{1}{3}''/\sin 72°44'1''$, that is, ·028545.

The smallness of $(a-c)/b$ suggests a supplementary question, that of the least value of a/b for which covering is possible while three of the small circles actually pass through the centre of the large circle. To solve this, we have only to take instead of the equation expressing that $2\theta+\phi$ is a maximum, the geometrical condition that C and K coincide, a condition expressed by the equations

$$\chi = \psi = 2\phi = 2\theta.$$

The angle θ must satisfy the equation

$$2\cos\theta\cos\tfrac{3}{2}\theta = 1,$$

that is,

$$1-\cos\tfrac{5}{2}\theta-\cos\tfrac{1}{2}\theta = 0,$$

so that $\frac{1}{2}\theta$ satisfies the equation

$$F(\xi) \equiv 1-\cos 5\xi-\cos\xi = 0,$$

a first approximation to the required root of this equation being 18°. A good approximation to $1/(5\sin 90°+\sin 18°)$ is $(\frac{1}{5}-\frac{1}{90})$, and the root is the sum of a series $\xi_1+\xi_2+\ldots$, whose terms are found in succession from the formulæ

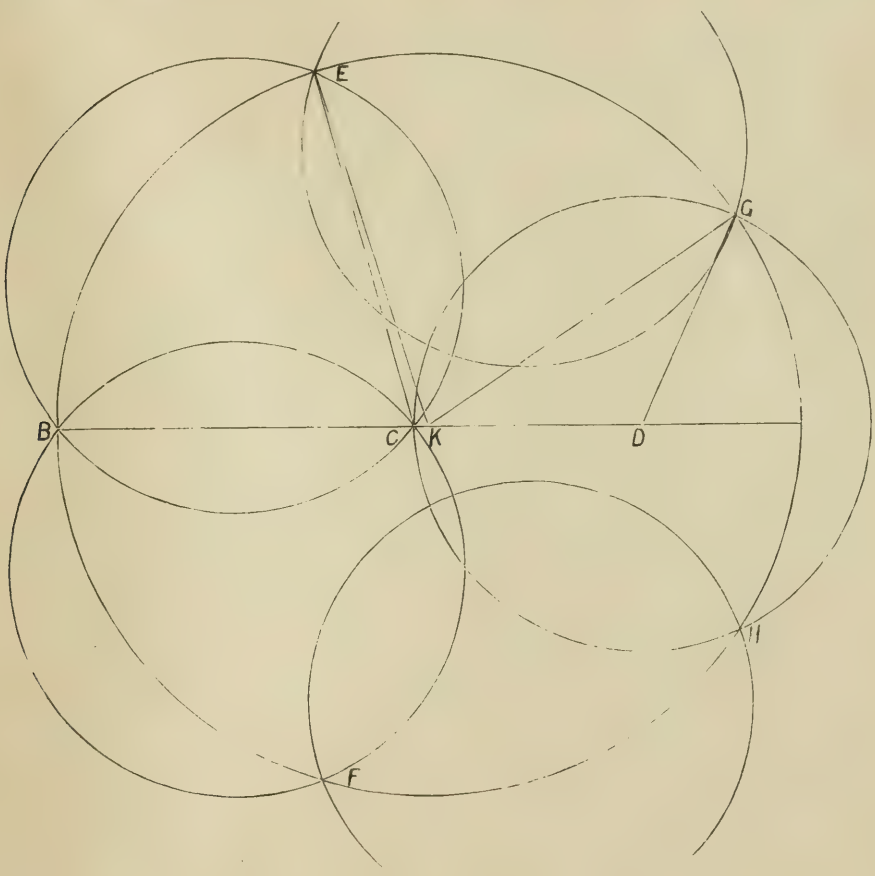
$$\xi_1 = 18°, \quad \xi_2 = -(\tfrac{1}{5}-\tfrac{1}{90})F(\xi_1), \quad \xi_3 = -(\tfrac{1}{5}-\tfrac{1}{90})F(\xi_1+\xi_2), \quad \ldots.$$

It is found that in fact $F(\xi_1+\xi_2+\xi_3)$ vanishes to seven places of decimals, $\xi_1+\xi_2+\xi_3$ being equal to $17°28'16\frac{1}{4}''$, so that the ratio of the radii, being the cosine of $\frac{3}{2}\theta$, that is, of $52°24'48\frac{3}{4}''$, is ·6099579.

Perhaps the most curious feature of the whole problem is the nearness of this ratio, on the one hand to the smallest ratio permitting covering, and on the other hand to the smallest ratio which allows the five small circles all to pass through the centre of the large circle, this last ratio being $\frac{1}{2}\sec 36°$, that is, ·6180340. The difference between the smallest ratio allowing covering and the last ratio found is just large enough to take effect in practice, but the lack of precision in the painted circle and the thickness of the discs prevent the accuracy which would be necessary if a distinction was to be made between the first two arrangements discussed, and if my readers can perceive the centre of the large circle they may proceed to pocket as many of the showman's rewards as they feel themselves to have earned by reading these pages.

[My brother, Mr. B. M. Neville, to whom I am indebted for the drawing

to scale which accompanies this paper, finds that unless the large circle has a diameter of about a metre the possibility of the best construction, when the construction through the centre is inadequate, cannot be made evident. To him is due the discovery of a most convenient and accessible form of covering disc—the pieces of parchment sold as jam covers: provided with five of these, one has only to draw on a sheet of paper a large circle of appropriate radius; the transparency of the small circles is a great advantage.]



London: Printed by C. F. Hodgson & Son, 2 Newton Street, Kingsway, W.C.