Use of Tschebyscheff-Polynomial Operators in the Numerical Solution of Boundary-Value Problems

GEORGE SHORTLEY
Operations Research Office, The Johns Hopkins University, Chevy Chase, Maryland
(Received August 11, 1952)

This paper concerns the numerical solution, in two or more dimensions, of boundary-value problems arising from linear partial differential equations, of which Laplace's and Poisson's equations furnish simple examples. Only techniques suitable for automatic computing machines are considered. The only method previously applicable to automatic machines is a straightforward iteration of an appropriate difference operator applied to an assumed trial solution on a network of points; as has been repeatedly pointed out, the rate of convergence of this iterative procedure is inordinately slow when the number of net points is large. The present paper shows how the employment of a Tschebyscheff polynomial of this same difference operator can cut the amount of work required in obtaining a solution by a very large factor—a factor of the order of magnitude of $\sqrt{N}$ in the two-dimensional case, $\sqrt[3]{N}$ in the three-dimensional case, where N is the number of net points. This method is an outgrowth of that developed by Flanders and Shortley [J. Appl. Phys. 21, 1326 (1950)] for employment of such Tschebyscheff-polynomial operators in the numerical solution of *eigenvalue* problems. The method is illustrated by a simple example involving Poisson's equation.

INTRODUCTION

THIS paper considers the numerical solution of boundary-value problems arising from elliptic linear partial differential equations, of which Laplace's and Poisson's equations furnish simple examples.

We are interested in problems in two- or three-dimensional regions of sufficient complexity so that hundreds or thousands of points are required to obtain adequate accuracy in an approximating solution on a network of points. In such cases, use of high speed computing machinery is justified or required, and some of the short cuts that have been developed for hand methods are inapplicable.[1]

The only method previously applicable to automatic-machine techniques is a straightforward iteration of an appropriate difference operator applied to an assumed trial solution on the net; as has been repeatedly pointed out, the rate of convergence of this iterative procedure is inordinately slow when the number of net points is large.

The present paper shows how the employment of a Tschebyscheff polynomial of this same difference operator can cut the amount of work required in obtaining a solution by a very large factor—a factor of the order of magnitude of $\sqrt{N}$ in the two-dimensional case, $\sqrt[3]{N}$ in the three-dimensional case, where N is the number of net points. This idea is an outgrowth of the employment, by Flanders and Shortley,[2] of such Tschebyscheff-polynomial operators in the numerical solution of *eigenvalue* problems.

[1] For a review of previous work on numerical methods applicable to boundary-value problems in two or three dimensions, see Thomas J. Higgins, *Numerical Methods of Analysis in Engineering* (The Macmillan Company, New York, 1949), Chapter 10. This chapter contains an extensive bibliography.

[2] Donald A. Flanders and George Shortley, J. Appl. Phys. 21, 1326 (1950), which see for many details not repeated in the present paper. Two minor errors have been detected in this reference: Equation (11) should have the factor 2^{p-1} inserted on the right-hand side, and the entry $m=6$, $r=0.9$ in Table I should read 0.03932.

THE DIFFERENCE EQUATION

While the method described below is applicable to boundary-value problems based on a wide variety of homogeneous and inhomogeneous elliptic linear partial differential equations, we can best illustrate the method by describing a simple example. Poisson's equation furnishes the simplest nonhomogeneous equation (of which Laplace's equation is a special, homogeneous case), and we shall use this equation for illustration.

For simplicity, we shall consider a rectangular two-dimensional region, although the method is readily generalized to more dimensions, and to irregular boundaries.[3] We shall write Poisson's equation as

$$\Delta u+\alpha(x, y)=0, \tag{1}$$

where $\alpha(x, y)$ is a given function of position. The boundary-value problem requires the solution of (1) that has given specified values of u along the boundary of the two-dimensional region.

We replace the continuum by a square net of points of spacing h, and the functions u and α by their values u_{ij} and α_{ij} at the net points. Equation (1) is then approximated by the difference equation

$$\omega u_{ij}+\tfrac{1}{4}h^2\alpha_{ij}=u_{ij}, \tag{2}$$

where ω is the operator that averages the four neighboring values:

$$\omega u_{ij}\equiv\tfrac{1}{4}(u_{i+1,\,j}+u_{i-1,\,j}+u_{i,\,j+1}+u_{i,\,j-1}). \tag{3}$$

The solution of (2) approximates that of (1) more and more closely as h is made smaller.[1]

THE ITERATION PROCEDURE

In the iteration procedure one starts with a "trial" function v_{ij} having the correct values at the boundary

[3] The "improvement formula" for a point near an irregular boundary is given in Weller, Shortley, and Fried, J. Appl. Phys. 11, 283 (1940).

points. Equation (2) is then used repeatedly as an "improvement formula" to give a series of functions that eventually converges to the true solution u_{ij}. Let us define the inhomogeneous operator O by the relation

$$Ov_{ij} \equiv \omega v_{ij} + \tfrac{1}{4}h^2 a_{ij}. \tag{4}$$

Then the sequence v_{ij}, Ov_{ij}, O^2v_{ij}, O^3v_{ij}, $\cdots$ converges to u_{ij}, as we can readily prove:

Let

$$v_{ij} = u_{ij} + \phi_{ij}, \tag{5}$$

where u is the true solution of (2) and ϕ is the "error" in the trial function v. The error ϕ will vanish at the boundary points. Then (omitting subscripts),

$$\begin{aligned} Ov &= \omega u + \omega\phi + \tfrac{1}{4}h^2\alpha \\ &= u + \omega\phi, \\ O^2v &= u + \omega^2\phi, \\ O^3v &= u + \omega^3\phi, \cdots. \end{aligned} \tag{6}$$

The sequence $\omega\phi$, $\omega^2\phi$, $\omega^3\phi$, $\cdots$ converges to zero because ϕ has zero boundary values and ω is of the nature of an averaging operator; hence the sequence in (6) converges to u. We can expand[2] ϕ in terms of the eigenvectors $\phi^{(1)}$, $\phi^{(2)}$, $\cdots$, $\phi^{(N)}$ of the operator ω. These eigenvectors satisfy the equation

$$\omega\phi^{(k)} = \lambda_k\phi^{(k)}, \tag{7}$$

where the eigenvalues λ_k have the property

$$1 > \lambda_1 > \lambda_2 \geqq \lambda_3 \geqq \cdots > \lambda_N > -1, \tag{8}$$

and N is the number of internal points in the net.

If we write this expansion as

$$\phi = \sum c_k\phi^{(k)}, \tag{9}$$

then

$$\omega^n\phi = \sum c_k\lambda_k{}^n\phi^{(k)}, \tag{10}$$

and $\omega^n\phi$ is seen to converge to zero because the λ's are all less than unity in absolute value. The rate of convergence is determined by the size of λ_1 and λ_N, which are generally close to unity in absolute value. In a rectangular region of $p \times q$ net points,

$$\begin{aligned} \lambda_1 = -\lambda_N &= \tfrac{1}{2}[\cos(\pi/p) + \cos(\pi/q)] \\ &\approx 1 - \tfrac{1}{4}(\pi^2/p^2) - \tfrac{1}{4}(\pi^2/q^2). \end{aligned} \tag{11}$$

If we desire the largest coefficient in (10) to be a small number ϵ, we must iterate a number of times n such that

$$\lambda_1{}^n = \epsilon \quad \text{or} \quad n = \log\epsilon/\log\lambda_1. \tag{12}$$

For a square net of $p \times p = N$ points, (11) becomes $1 - \lambda_1 = \frac{1}{2}\pi^2/N$, a very small number if N is large. Then $\log\lambda_1 \approx \lambda_1 - 1 = -\frac{1}{2}\pi^2/N$, and (12) becomes

$$n = 2(N/\pi^2)(-\log\epsilon), \tag{13}$$

a very large number.

EMPLOYMENT OF THE TSCHEBYSCHEFF-POLYNOMIAL OPERATOR

Going back to (6), we see that if we apply to the trial function v a polynomial $P(O)$ in the operator O, we obtain

$$P(O)v = P(1)u + P(\omega)\phi, \tag{14}$$

where $P(1)$ is the value of the polynomial at argument unity. If we require that

$$P(1) = 1, \tag{15}$$

and use the result (10), we obtain

$$P(O)v = u + \sum c_k P(\lambda_k)\phi^{(k)}. \tag{16}$$

We now see that there is a "best" polynomial of order m to apply to v in order to obtain u, namely, that polynomial satisfying (15) that has the least maximum absolute value at all eigenvalues λ_k. Since the number N of eigenvalues is large compared to the order m of polynomial that it is feasible to use, we must apply this criterion in the following form: *The best polynomial $S_m(O)$ of order m to use in* (16) *is that satisfying* (15) *and having the least maximum absolute value throughout the range from $\lambda_N = -\lambda_1$ to $+\lambda_1$.* Flanders and Shortley[2] show that this best polynomial is

$$S_m(O) = T_m(O/\lambda_1)/T_m(1/\lambda_1), \tag{17}$$

where $T_m(\mu)$ is the Tschebyscheff polynomial obtained by expanding

$$\begin{aligned} T_m(\mu) &= \cos(m \arccos\mu), \quad (\mu \leqslant 1) \\ T_m(\mu) &= \cosh(m \operatorname{arccosh}\mu), \quad (\mu \geqslant 1) \end{aligned} \tag{18}$$

in powers of μ. The denominator in (17) is designed to make $S_m(\lambda) = 1$ at $\lambda = 1$.

The technique of employing such a polynomial is then, in principle, the following: Compute the series of functions v, Ov, O^2v, $\cdots$, O^mv, and then, with an estimated value of λ_1 (preferably slightly greater rather than slightly less than the correct value), form the linear combination $S_m(O)v$. This linear combination will have all coefficients of $\phi^{(k)}$ in (16) reduced to a small (and computable) fraction of their original values, since

$$S_m(O)v = u + \sum c_k S_m(\lambda_k)\phi^{(k)}. \tag{19}$$

COMPARISON OF THE METHODS

Because the Tschebyscheff polynomial $T_m(\mu)$ has unity (repeated $m+1$ times) as its greatest absolute value in the range $-1 \leqslant \mu \leqslant +1$, the function $S_m(\lambda)$ has $1/T_m(1/\lambda_1)$ as its greatest absolute value in the range $-\lambda_1 \leqslant \lambda \leqslant +\lambda_1$. The value $1/T_m(1/\lambda_1)$ can be computed accurately from (18), and is tabulated in Table I of the reference in footnote 2. All coefficients c_k in the last term of (19) are multiplied by a quantity at least as small as $1/T_m(1/\lambda_1)$.

Reference 2 did not estimate the saving in labor resulting from the use of the Tschebyscheff polynomial. In order to make a comparison with the iteration pro-

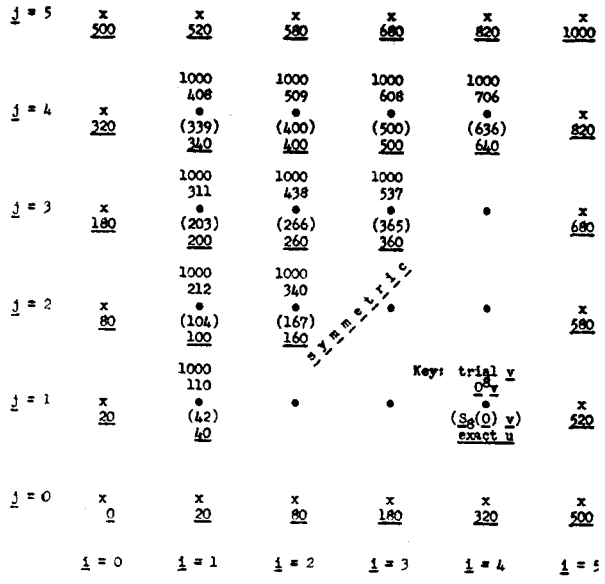

FIG. 1. Solution of Poisson's difference equation $\omega u-20=u$ on a 5×5 net.

cedure, we can derive a satisfactory approximation to the value of $1/T_m(1/\lambda_1)$. Since λ_1 is only slightly less than 1, write

$$1/\lambda_1=1+\zeta. \tag{20}$$

Now

$$\operatorname{arccosh}(1+\zeta)=(2\zeta)^{\frac{1}{2}}(1-\tfrac{1}{12}\zeta+\tfrac{3}{160}\zeta^2+\cdots).$$

If $\zeta\ll 1$, we can write [compare (18)],

$$T_m(1+\zeta)\approx\cosh m(2\zeta)^{\frac{1}{2}}=\tfrac{1}{2}e^{m\sqrt{(2\zeta)}}+\tfrac{1}{2}e^{-m\sqrt{(2\zeta)}}.$$

The last term here will be negligible whenever $1/T_m(1+\zeta)$ is small; it is negligible to 4-significant-figure accuracy whenever $1/T_m(1+\zeta)<0.02$, and all cases of interest fall within this range. Neglecting this last term, we find

$$1/T_m(1/\lambda_1)=1/T_m(1+\zeta)\approx 2e^{-m\sqrt{(2\zeta)}}. \tag{21}$$

If we want this to be a small number ϵ, we must choose an order m of polynomial given by

$$m=(-\log\tfrac{1}{2}\epsilon)/(2\zeta)^{\frac{1}{2}}. \tag{22}$$

For a square net of $p\times p=N$ points, $\zeta\approx 1-\lambda_1=\frac{1}{2}\pi^2/N$, and

$$m=(N/\pi^2)^{\frac{1}{2}}(-\log\tfrac{1}{2}\epsilon). \tag{23}$$

This polynomial order may be directly compared with the required power (13) of the operator O^n used in the iteration procedure. The ratio

$$\frac{n}{m}=\frac{2}{\pi}\cdot N^{\frac{1}{2}}\cdot\frac{-\log\epsilon}{-\log\frac{1}{2}\epsilon}. \tag{24}$$

Of the three factors, the first and third are of the order of unity, and the large factor is $N^{\frac{1}{2}}$. The ratio (24) is essentially the ratio of the amount of work involved in the two methods, since building up the polynomial $S_m(O)$ is less labor than any single application of O. Hence we conclude that *the use of the Tschebyscheff operator is capable of cutting machine time by a large factor of the order of $N^{\frac{1}{2}}$, where N is the total number of points in the net.* For a three-dimensional cube, the formula corresponding to (24) contains $N^{\frac{1}{3}}$ in place of $N^{\frac{1}{2}}$.

EXAMPLE

As a short example that will illustrate this method, let us consider Poisson's equation on the 5×5 network of Fig. 1, in which the boundary points are indicated by crosses, the internal points by dots. With $h=1$, the underlined values, given by

$$u=20(i^2+j^2),$$

are an exact solution of both the differential equation

$$\Delta u-80=0, \tag{25}$$

and the corresponding difference equation [compare Eq. (2)]

$$\omega u_{ij}-20=u_{ij}. \tag{26}$$

Let us now solve the difference equation (26) numerically by using the boundary values given in Fig. 1 but starting with a grossly erroneous trial function v having the value 1000 at all interior points. The operator O is [compare Eq. (4)]

$$Ov=\omega v-20. \tag{27}$$

In this case the 16 eigenvectors of ω are given by

$$\sin(ki\pi/5)\sin(lj\pi/5),\quad (k,l=1,2,3,4) \tag{28}$$

with eigenvalues

$$\begin{aligned}\lambda_1,\lambda_{16}&=\pm 0.809,\\ \lambda_2,\lambda_3,\lambda_{14},\lambda_{15}&=\pm 0.559,\\ \lambda_4,\lambda_{13}&=\pm 0.309,\\ \lambda_5,\lambda_6,\lambda_{11},\lambda_{12}&=\pm 0.250,\\ \lambda_7,\lambda_8,\lambda_9,\lambda_{10}&=0.\end{aligned} \tag{29}$$

The error $\phi=v-u$ [compare Eq. (5)] in the trial function v, when expanded as in (9) in terms of the eigenvectors (28), has as its leading term

$$c_1\phi^{(1)}=1076\sin(i\pi/5)\sin(j\pi/5). \tag{30}$$

This term is the troublesome one to remove by numeri-

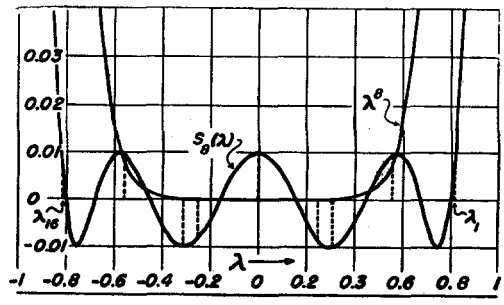

FIG. 2. Plot showing λ^8, $S_8(\lambda)$ and the eigenvalues (broken lines) for the net of Fig. 1.

cal procedures. The coefficient of $\phi^{(16)}$, on the other hand, happens to be zero, so this eigenvector will not prove troublesome.

If we apply the operator O eight times to v, we get the function O^8v shown in the second line of Fig. 1. (The computations were actually made by carrying two more figures to avoid rounding-off errors.) Figure 2 shows a plot of λ^8 with the eigenvalues (29) indicated by the broken lines. This plot shows that the coefficients of all eigenvectors in ϕ will be reduced in O^8v to less than one percent of their initial values, *except* c_1, which will be multiplied by the factor $\lambda_1{}^8=(0.809)^8=0.183$. The maximum error in O^8v in Fig. 1 is 180, which is consistent with (30) multiplied by $\lambda_1{}^8$. To achieve a value $\lambda_1{}^n=0.01$ would require $n=22$, i.e., 22 iterations.

Figure 3 is a plot along the diagonal $i=j$ of the trial function v and the results of the eight iterations, showing the manner of convergence to the true solution u.

Now let us apply the Tschebyscheff scheme. We see from Table I of reference 2 that $1/T_8(1/0.818)=1/T_8(1/9/11)=0.0106$; hence the operator [compare Eq. (17)]

$$S_8(O)=T_8(1\tfrac{1}{9}O)/T_8(1\tfrac{1}{9}) \tag{31}$$

should reduce all coefficients to one percent or less from $\lambda=-0.818$ to $+0.818$. Since this region includes λ_1 (compare Fig. 2), the operator (31) should reduce the coefficient in (30) to less than one percent.

We easily compute that

$$S_8(O)v=6.7620\,O^8v-9.0532\,O^6v+3.7878\,O^4v \\ -0.5071\,O^2v+0.0106\,v.$$

Formation of this linear combination gives the values in parentheses in Fig. 1. The maximum error in the function $S_8(O)v$ is 7, so that the expected reduction of (30) to less than one percent has been achieved. The difference between $S_8(O)v$ and u is too small to show on the plot of Fig. 3.

COMMENTS

The techniques for machine application of the Tschebyscheff polynomial suggested in reference 2, including the factoring of the polynomial into polynomials of order six or eight, are applicable also to boundary-value problems.

In order to cut the order of polynomial required [compare Eq. (22)], it may be desirable not to try to remove $\phi^{(1)}$, or even $\phi^{(1)}$ and $\phi^{(2)}$, from the error, but to concentrate on removing $\phi^{(3)}, \cdots, \phi^{(N)}$. One can do this by using an appropriate shift of scale and origin of the Tschebyscheff polynomial as described in reference 2. The following process similar to the "harmonic analysis" of reference 2, can then be used to remove $\phi^{(1)}$ and $\phi^{(2)}$.

Suppose we have arrived at the function

$$v=u+a_1\phi^{(1)}+a_2\phi^{(2)}. \tag{32a}$$

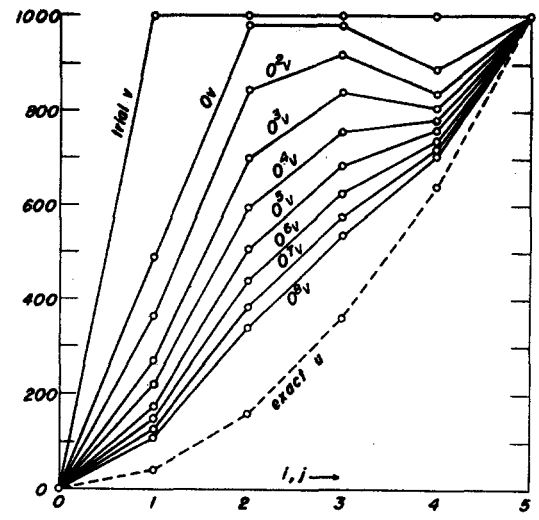

FIG. 3. Plot along the diagonal of symmetry, $i=j$, of Fig. 1.

Then [compare Eqs. (6) and (10)], we can compute

$$Ov=u+\lambda_1a_1\phi^{(1)}+\lambda_2a_2\phi^{(2)}, \tag{32b}$$

$$v-Ov=a_1(1-\lambda_1)\phi^{(1)}+a_2(1-\lambda_2)\phi^{(2)}. \tag{33a}$$

Write this last equation for the moment as

$$w=c_1\phi^{(1)}+c_2\phi^{(2)}. \tag{33b}$$

This function w is of the form of Eq. (12a) of reference 2. We can now operate on w twice by ω and follow the procedure on p. 1330 of reference 2 to determine λ_1, λ_2, $c_1\phi^{(1)}$, and $c_2\phi^{(2)}$. Since $a_1(1-\lambda_1)=c_1$ and $a_2(1-\lambda_2)=c_2$, we can eliminate the error from (32) and determine u.

The procedure just mentioned can also be used to obtain estimates of λ_1 and λ_2 from a coarse net, as discussed in reference 2, for guidance in choosing the order of polynomial required on a refined net.

The device used in reference 3 of defining the operator ω in O so that, in traversing the net with the improvement formula, the improved values Ov are used whenever they are available, doubles the value of $1-\lambda_1$ and hence, according to (12), cuts the number of necessary iterations by a factor of two in the iteration procedure. The same device can be used in the Tschebyscheff method, in which case, according to (22), it cuts the order of polynomial required by $\sqrt{2}$. However in machine work (as is the case with IBM machines) this use of improved values may not be very convenient.

We might briefly discuss one other type of example. In the case of axisymmetrical boundary-value problems,[4] the equation

$$\frac{\partial^2u}{\partial z^2}+\frac{\partial^2u}{\partial\rho^2}+\frac{K}{\rho}\frac{\partial u}{\partial\rho}=0 \tag{34}$$

arises with $K=1$, 3, and 5 in various cases. Here ρ is the radial coordinate and z the coordinate parallel to

[4] Shortley, Weller, Darby, and Gamble, J. Appl. Phys. 18, 116 (1947).

the axis. If a square net is laid down with points at $\rho=ih$, $z=jh$, the differential equation (34) is approximated by the difference equation

$$\Omega u=u,$$

where Ω is defined by

$$\Omega v_{ij}=\tfrac{1}{4}v_{i,j+1}+\tfrac{1}{4}v_{i,j-1}+[\tfrac{1}{4}+(K/8i)]v_{i+1,j}+[\tfrac{1}{4}-(K/8i)]v_{i-1,j}.$$

This operator (with certain modifications near the axis[4]) can be used as an improvement operator like O above. It can be represented by a real matrix that satisfies the conditions (see footnote 3 of reference 2) for having real eigenvalues. Hence the employment of a Tschebyscheff polynomial in Ω can speed up the convergence in the same way as for Poisson's equation.

JOURNAL OF APPLIED PHYSICS VOLUME 24, NUMBER 4 APRIL, 1953

Optimum Nonlinear Filters

L. A. ZADEH
Department of Electrical Engineering, Columbia University, New York, New York
(Received September 11, 1952)

The theory of optimum nonlinear filters outlined in this paper is based on the consideration of a sequence of classes of nonlinear filters, designated as $\mathfrak{N}_1$, $\mathfrak{N}_2$, $\mathfrak{N}_3,\cdots$, such that each class in the sequence includes all the preceding classes and, furthermore, the class of linear filters is a subclass of every class in the sequence. A filter of class $\mathfrak{N}_m$ is described in terms of a characteristic function which involves m age variables and m values of the input time-function. The input-output relationship for a filter of class $\mathfrak{N}_m$ has the form of an m-fold integral of the characteristic function with respect to the m age variables. It is shown that the characteristic function of the optimum filter (in the least squares sense) within the class $\mathfrak{N}_m$ satisfies a linear integral equation of $2m$th order. The optimization of filters of class $\mathfrak{N}_1$ is treated in detail, and methods of approximate realization of such filters in the form of nonlinear delay line filters and power series filters are indicated. The results are extended to the case of nonstationary time series.

1. INTRODUCTION

THE theory of optimum linear filters has been the subject of numerous investigations since the publication of Wiener's classic monograph[1] on the prediction and smoothing of stationary time series. By contrast, relatively little work has been done on the optimization of nonlinear filters, although it has long been recognized that, in principle, better results can be achieved with filters of this type than with linear filters. An important deterrent to the study of nonlinear filters has been the fact that, in general, nonlinear systems are not susceptible of a strictly analytical treatment and therefore do not provide the system theorist with a fruitful field for purely theoretical investigation. In recent years, however, the advent of large scale digital computers and other mechanized means of computation has profoundly influenced the basic philosophy of system design and analysis. Thus, in virtue of the availability of machine computers, it has become sufficient to carry the analytical treatment only to a point where the problem is reduced to mathematical operations which can be handled by such computers. This development has made it practicable to study nonlinear systems which cannot be completely analyzed by purely analytical means.

The use of machine computers is implicit in the important contributions by Singleton[2] and White[3] to the theory of optimum nonlinear filters. Essentially, Singleton and White consider the class of nonlinear filters which are characterized by input-output relationships of the form

$$v(t)=f[u(t), u(t-T),\cdots, u(t-(n-1)T)], \qquad (1)$$

where $u(t)$ denotes the input, $v(t)$ is the output, T is a constant, and f is an arbitrary real function, and determine the function f in such a way as to minimize the mean square error (or the probability of error). In this method, the determination of the minimizing function f requires, in general, the knowledge of $2n$th order joint probability distribution function for the signal and noise components of the input.

The purpose of this paper is to outline a more flexible approach which is based on the consideration of a certain system of classes of nonlinear filters.[4] Specifically, we consider a sequence of classes, designated as $\mathfrak{N}_1$, $\mathfrak{N}_2$, $\mathfrak{N}_3,\cdots$, such that each class in the sequence includes all the preceding classes and, furthermore, each

[1] N. Wiener, "The extrapolation, interpolation and smoothing of stationary time series," Report to the Services 19, Research Project DIC-6037, M.I.T. (February, 1942). Published in book form by John Wiley and Sons, Inc., New York, 1949.

[2] H. E. Singleton, "Theory of nonlinear transducers," Tech. Report No. 160, M.I.T. Research Laboratory of Electronics (1950).

[3] W. D. White, Proc. Inst. Radio Engrs. 39, 303 (1951).

[4] This and other systems of classes of nonlinear two-poles are discussed in greater detail in a paper by the writer, "A contribution to the theory of nonlinear systems," to be published in J. Franklin Inst.