STATISTICAL MECHANICS

COMPREHENSIVE VIEW OF PREDICTION THEORY[1]

NORBERT WIENER

Some years ago a paper appeared[2] by Kolmogoroff in C. R. Acad. Sci. Paris on extrapolation and interpolation. From the point of view in which we are interested, the main contribution of this paper was a discussion of the greatest lower bound of the error of mean square prediction when applied to time series, and when the method of prediction was to be linear in the past. This paper led to a couple of papers in Russia with which Kolmogoroff's own name and that of M. Krein are jointly associated.

About a year after the first Kolmogoroff paper, the present author independently started a series of investigations in the same direction. His motivation was the problem of predicting the future position of an airplane on the basis of a general statistical knowledge of its mode of flight and of a more concrete knowledge of its immediate past. Thus there were from the beginning two points in which his emphasis was different from that of Kolmogoroff. Also at least one of these was actually covered in Kolmogoroff's work. The first difference is that while Kolmogoroff's explicit work is primarily concerned with the irreducible minimum error of prediction, the author's own work is concerned with the actual method of securing a prediction with this irreducible error, or at least prediction with error as near to this as we wish.

Next, Kolmogoroff's work is primarily concerned with a time consisting of discrete instants, whereas my work is concerned with the continuous time in which the flight of an airplane takes place. Associated with this is the fact that my work, unlike the explicitly published work of Kolmogoroff, concerns the instrumentation which is necessary to realize the theory of prediction in automatic apparatus for shooting ahead of an airplane. This engineering bias leads me to emphasize more than does Kolmogoroff the problem of prediction in terms of linear operators in the scale of frequency, rather than in similar operators on the scale of time.

This same engineering standpoint led me from the prediction problem to the related problem of filtering. In this problem a message, a noise, and their relations to one another are known statistically as well as the sum of the message and the noise, from minus infinity to a given point in time. The problem is to disentangle from this combination the message alone, or more generally, the message under lead or lag in time.

Since the harmonic analysis of a signal is not complete within any finite time, this separation must always involve the prediction of the future of the message

[1] This address was listed on the printed program under the title *The statistical mechanics in communication.*

[2] A. Kolmogoroff, *Interpolation und Extrapolation von stationären zufälligen Folgen*, Bull. Acad. Sci. URSS Sér. Math. vol. 5 (1941) pp. 3–14.

and noise, and so cannot be isolated from prediction theory in general. It turns out that the technique for the design of filters with a minimum mean square error is precisely parallel to that for the design of predictors.

In one point both Kolmogoroff and his school and I myself have developed the theory of prediction along similar lines. This point is that of multiple prediction, in which not merely one quantity varies with time, but a number of quantitites—even an infinite number of quantities. This is an extremely important problem for the meterorologists and in general for the statisticians. Until recently its theory has not been implemented by a corresponding communicable tool.

Both the Russian school and my own have published what is a fairly extensive literature concerning prediction. What is missing is a definitive paper to take up all the threads of the argument and to close them off in a single comprehensive discussion in which as many theorems as possible are given necessary and sufficient conditions. It is the purpose of the present paper to fill exactly this gap.

Let me first take up the simplest prediction problem, which is that concerning prediction in a simple discrete time series. All that follows this will consist in an extension or development of methods here given.

In the first place, we shall consider a time series to be, not a single sequence of of numbers, but a sequence of numbers with a parameter of distribution. Thus α appears merely for purposes of integration, and α may therefore be mapped on the segment of a line between 0 and 1, since the main properties of integration are independent of dimensionality. Let us then consider a function $f(\alpha)$ defined on (0, 1) and belonging to the Lebesgue parameter class L_2. Since the particular time series with which we are concerned are not attached to any particular point in time, let us introduce the transformation $T\alpha$, which preserves measure or probability, and which moves each instant of time into the next one ahead. Then $f(T^\nu\alpha)$ represents our time series. Here ν runs between $-\infty$ and ∞, if ν is positive, T^ν represents the transformation T iterated ν times. $T^{-\nu}\alpha$ represents the inverse transformation T^{-1} iterated ν times; and again ν is positive. We then shall see that $f(T^\nu\alpha)$, where ν runs over all negative and positive integral values including 0, is a time series in statistical equilibrium.

The functions $f(T^{-\nu}\alpha)$, where $\nu > 0$, are a denumerable set of functions of class L_2, and as such have a linear extension, containing all functions of L_2 which can be approximated in the L_2 sense by polynomials in the given functions. This set of functions will be known as the *past*. The *present* consists only of linear multiples of $f(\alpha)$ itself; while the *future* is the linear extension of all functions of the form $f(T^\nu\alpha)$, where ν is positive.

The problem of prediction is that of the projection of a function belonging to the present or future on the past. Every function belonging to L_2 consists of the sum of a function belonging to the past and a function orthogonal to every function belonging to the past. The first is called the projection on the past of a given function; whereas the mean square of the second with respect to α is the mean square error of prediction. In order to carry out this process, it is useful to orthogonalize the set of functions $f(T^{-\nu}\alpha)$.

Let us first take $f(T^{-1}\alpha)$. There are two cases possible. Either this is equivalent to 0 or it is not. If it is equivalent to zero, $f(\alpha)$ may be expressed in terms of $f(T^{-1}\alpha)$, and hence of the past. Similarly, $f(T\alpha)$ may be expressed in terms of $f(\alpha)$ and thus in terms of $f(T^{-1}\alpha)$, or again in terms of the past. Thus the whole present and future may be expressed in terms of the past and the prediction problem may be solved with 0 mean square error.

On the other hand, let us suppose that $f(T^{-1}\alpha)$ is not equivalent to 0. If it is not equivalent to zero, it can be normalized by the multiplication of an appropriate factor, so that the integral of the square of its modulus is one. Let us call this normalized function $g_1(\alpha)$. Now let us consider $f(T^{-2}\alpha)$. Either this is equivalent to a multiple of $f(T^{-1}\alpha)$; in which case a perfect prediction is possible; or it is not so equivalent. In the second case, we shall have the formula

$$f(T^{-2}\alpha) = g_1(\alpha)\int_0^1 f(T^{-2}\beta)\overline{g_1(\beta)}\,d\beta + \left(f(T^{-2}\alpha) - g_1(\alpha)\int_0^1 f(T^{-2}\beta)\overline{g_1(\beta)}\,d\beta\right)$$

where the term

$$f(T^{-2}\alpha) - g_1(\alpha)\int_0^1 f(T^{-2}\beta)\overline{g_1(\beta)}\,d\beta$$

is not equivalent to 0. This term again may be normalized; and we shall call the result of this normalization $g_2(\alpha)$. We repeat the process and express $f(T^{-3}\alpha)$ in terms of $g_1(\alpha)$ and $g_2(\alpha)$, with the remainder

$$f(T^{-3}\alpha) - g_1(\alpha)\int_0^1 f(T^{-3}\beta)\overline{g_1(\beta)}\,d\beta - g_2(\alpha)\int_0^1 f(T^{-3}\beta)\overline{g_2(\beta)}\,d\beta.$$

Either this remainder is equivalent to 0 or it is not. If it is equivalent to zero, we again can express $f(T^{-3}\alpha)$ in terms of its past; and if we cannot, we may introduce a third function $g_3(\alpha)$ by normalizing the remainder. This process can be continued until we have either orthogonalized the entire past of $f(\alpha)$ or until there are no more terms left to orthogonalize. If the process terminates at any stage, a perfect prediction is possible. In all other cases, we have a normal or orthogonal set $g_1(\alpha), g_2(\alpha), \cdots$, in terms of which we can express the past of $f(\alpha)$.

Now let us form the function $h(\alpha)$, in accordance with the formula

$$h(\alpha) = f(\alpha) - \sum_1^\infty g_n(\alpha)\int_0^1 f(\beta)\overline{g_n(\beta)}\,d\beta.$$

If this function $h(\alpha)$ should prove equivalent to 0, this means that $f(\alpha)$ can be expressed linearly in terms of its past, without any error whatever of prediction. If not, then the function $h(\alpha)$ can itself be normalized, and we obtain a function $H(\alpha)$, in accordance with the formula

$$H(\alpha) = \frac{h(\alpha)}{\left(\int_0^1 |\,h(\beta)\,|^2\,d\beta\right)^{1/2}}.$$

The function $H(\alpha)$ is linearly dependent on the present and past of f and orthogonal to that past. It is accordingly orthogonal to every function $H(T^{-\nu}\alpha)$; and since the transformation T is measure preserving, then the set $H(T^{\nu}\alpha)$, where ν varies from $-\infty$ to ∞, is a normal orthogonal set. There are now two cases possible.

Either $f(\alpha)$ can be represented in terms of this set according to the formula

$$f(\alpha) = \sum_0^\infty H(T^{-\nu}\alpha) \int_0^1 f(\beta)\overline{H(T^{-\nu}\beta)}\, d\beta,$$

or there is a remainder

$$f_2(\alpha) = f(\alpha) - \sum_0^\infty H(T^{-\nu}\alpha) \int_0^1 f(\beta)\overline{H(T^{-\nu}\beta)}\, d\beta$$

not equivalent to 0. In the second case we may write

$$f(\alpha) = f_1(\alpha) + f_2(\alpha)$$

where $f_1(\alpha)$ may be shown to generate the same function $H(\alpha)$ as does $f(\alpha)$. On the other hand, $f_2(\alpha)$ will be completely determined by its own past from any period of time back. In other words, $f_2(\alpha)$ will be linearly dependent on the set $f(T^{-\nu}\alpha), f(T^{-\nu-1}\alpha), \cdots$ no matter how large ν may be. The complete present, past, and future of $f_1(\alpha)$ is orthogonal to the complete present, past, and future of $f_2(\alpha)$. f_1 and f_2 have already been given in terms of f alone. We have thus reduced every case of the prediction problem to the perfectly predictable case on the one hand, and the case where the function $H(\alpha)$ exists and f can be expressed in terms of H and its past, on the other.

It is the second case with which we are chiefly concerned. Let us notice that

$$\int_0^1 f(T^{\nu}\alpha)\overline{f(\alpha)}\, d\alpha = \sum_{\substack{\mu>0 \\ \mu-\nu\geqq 0}} \int_0^1 f(\beta)H(T^{\nu-1}\beta)\, d\beta \int_0^1 \overline{f(\beta)}H(T^{-\mu}\beta)\, d\beta.$$

Let us also notice that

$$\sum_0^\infty \left| \int_0^1 f(\beta)H(T^{-n}\beta)\, d\beta \right|^2 < \infty.$$

Thus the sum of the squares of the moduli of the coefficients of

$$\sum_0^\infty e^{in\omega} \int_0^1 f(\beta)H(T^{-n}\beta)\, d\beta = \Psi(\omega)$$

converges and the function belongs to L_2. $\Phi(\omega)$, the square of the modulus of this, will belong to L, and will have the Fourier coefficients

$$\int_0^1 H(T^n\alpha)\overline{H(\alpha)}\, d\alpha.$$

This suggests that with suitable hypotheses, we may proceed directly from the auto-correlation coefficients

$$\int_0^1 H(T^n\alpha)\overline{H(\alpha)}\, d\alpha$$

and their related harmonic analysis function $\Phi(\omega)$ to the coefficients

$$\int_0^1 f(\beta)\overline{H(T^{-\nu}\beta)}\, d\beta,$$

by means of which we express $f(\alpha)$ in a series in terms of the functions $H(T^{-\nu}\alpha)$. The conditions under which this is possible may be proved to be

$$\int_{-\pi}^{\pi} |\log \Phi(\omega)|\, d\omega < \infty. \tag{1}$$

To see this, we express $(1/2)\log\Phi(\omega)$ by the corresponding series

$$\sum A_n e^{in\omega}.$$

We then form the corresponding series

$$\sum A_n \operatorname{sgn} n\, e^{in\omega},$$

which may be shown to determine a pure imaginary function by Cesàro summation. Let this function be $F(\omega)$. Then if we put

$$(\Phi(\omega))^{1/2} e^{F(\omega)} = \Psi(\omega),$$

we shall find that the Fourier series of $\Psi(\omega)$ will contain no negative frequencies.

Another function closely related to $\Psi(\omega)$ is

$$\Psi_1(z) = \frac{1}{2\pi}\int_0^{2\pi} \Psi(\omega)\, \frac{e^{i\omega}\, d\omega}{e^{i\omega} - z}.$$

It can be shown that as r tends to 0 from 1,

$$\lim_{r\to 1} \int_0^1 |\Psi(\omega) - \Psi_1(re^{i\omega})|^2\, d\omega = 0;$$

that $\Psi(\omega)$ is analytic within the unit circle; and what is more, it can be proved to have no zeros within the unit circle.

Let us now suppose that $\Phi(\omega)$ is any real function of class L whose logarithm fulfills our critical condition (1). It is then possible, by the use of ideas from Brownian motion theory, to give time series $f(\alpha)$ and a measure preserving function T such that

$$\Phi(\omega) \sim \sum_{-\infty}^{\infty} e^{i\nu\omega} \int_0^1 f(T^\nu\alpha)\overline{f(\alpha)}\, d\alpha.$$

It is also possible to prove that $f(\alpha)$ cannot be expressed completely in terms of its own past, and that it is orthogonal to its remote past, in the sense that the

projection of $f(\alpha)$ on the set $f_2(T^{\nu-1}\alpha)$, $f(T^{-\nu-2}\alpha)$, $\cdots$ tends to 0 as ν becomes infinite. We thus have a complete set of conditions for the factoring of $\Phi(\omega)$, and we have the basis for a theory of simple discrete prediction.

All the essential ideas in this theory of prediction may be extended to multiple prediction. Let us first take up the case of finite multiple prediction. Here we start with a set of functions $f_n(T^\nu\alpha)$, where n ranges from 1 to N. The past consists of the linear extension of the functions $f_n(T^\nu\alpha)$, where n ranges from 1 to N and ν is negative. T is, of course, as before, a measure preserving transformation. We now can proceed as before to determine the parts of $f_1(\alpha)$ to $f_n(\alpha)$ which are orthogonal to the complete past of all the f's. We carry this out by a procedure of orthogonalization exactly like that which we have already used. If this procedure of orthogonalization terminates before it gives N functions $H_n(\alpha)$ which are normal and orthogonal to one another and to the past, but linearly dependent on the past and present, then some one at least of the functions $f_n(\alpha)$ is completely determined by its own past and the past of the other functions. If that is not the case, we obtain a set of normal and orthogonal functions $H_n(T^\nu\alpha)$.

Either all the functions $f_k(\alpha)$ may be completely developed in terms of these, or there are remainders in the development which are not equivalent to 0. In the second case, just as in the corresponding simple case, we can separate a multiple time series into the sum of two multiple time series, such that the past, present, and future of one will be completely orthogonal to the past, present, and future of the other. One of these time series will be perfectly predictible, and the other will be expressible in terms of its own H functions. We now go through procedures exactly analogous to those through which we have gone in the simple case. Let us notice that

$$f_i(\alpha) \sim \sum_{\nu=0}^{\infty} \sum_{n=1}^{N} H_n(T^{-\nu}\alpha) \int_0^1 f_i(\beta)\overline{H_n(T^{-\nu}\beta)}\, d\beta. \tag{2}$$

Let us also notice that

$$\int_0^1 f_i(T^\nu\alpha) f_j(\alpha)\, d\alpha = \sum_{\substack{\mu>0 \\ \mu-\nu\geqq 0}} \sum_n \int_0^1 f_i(\beta)\overline{H_n(T^{\nu-\mu}\beta)}\, d\beta \int_0^1 \overline{f_j(\beta)} H_n(T^{-\mu}\beta)\, d\beta.$$

Then we obtain a matrix of functions

$$\Psi_{ij}(\omega) = \sum_{\nu=0}^{\infty} e^{i\nu\omega} \int_0^1 f_i(\beta)\overline{H_j(T^{-\nu}\beta)}\, d\beta,$$

belonging to L_2. We shall have

$$\mathbf{\Phi}(\omega) = \mathbf{\Psi}(\omega)\cdot\overline{\mathbf{\Psi}(\omega)},$$

where the matrix $\mathbf{\Phi}(\omega)$ has the Fourier coefficients

$$\int_0^1 f_i(T^\nu\beta)\overline{f_j(\beta)}\, d\beta.$$

As before, it is interesting to know what condition on the matrix

$$\Phi(\omega) \qquad (-\pi \leqq \omega \leqq \pi)$$

will make the existence of the set of functions $H_h(T\alpha)$ and their closure with respect to the corresponding f functions possible. Without going into proofs, it may be said that the condition is that $\Phi_{ij}(\omega)$ all belong to L_2, and that

$$\int_0^1 |\log|\text{Determinant } \Phi(\omega)\,||\, d\omega < \infty.$$

If this condition is fulfilled for an Hermitian matrix of positive definite Hermitian type, then this matrix may always be obtained as indicated from a set of functions $f_n(T^r\alpha)$.

In the simple prediction case, we have already given an algorithm which will enable us to carry out the computational work of the resolution of $\Phi(\omega)$ in the form

$$\Phi(\omega) = |\Psi(\omega)|^2.$$

The similar resolution of the matrix $\mathbf{\Phi}(\omega)$ in the form

$$\mathbf{\Phi}(\omega) = \mathbf{\Psi}(\omega)\cdot\overline{\mathbf{\Psi}(\omega)}$$

is complicated by the fact that matrix multiplication is not commutative, and therefore there is no easy use of the logarithm. However, there is a computational process which is not too difficult. This is a generalized form of the alternating process known in potential theory.

Let us suppose that we have two linear subspaces of Hilbert space, say S_1 and S_2, and that we have a vector in that space. This vector is to be projected on the smallest linear space containing these two subspaces S_1 and S_2. We then project the vector on S_1. The remainder we project on S_2. The remainder after this second projection we project on the space S_1, etc. We then add all the projections that we have obtained on the two spaces. It may be shown that this process is a convergent one, at least in the mean sense, which is the only relevant sense here, and that it ultimately yields the projection of the vector on the smallest linear extension of the two spaces.

For the moment let us consider only a prediction process of multiplicity 2. The problem which we are facing is to take some vector not necessarily exclusively belonging to the past and to project it on a past which represents the smallest linear extension of the space combining the past of one component with the past of the other. It is then possible to do this by a procedure of successive projections, which will turn out to have a computable algorithm.

While this method of carrying out the alternating process for purposes of prediction is available in the perfectly general multiple case, we shall illustrate it here in the case of multiplicity 2. Actually the best computational procedure for a case of higher multiplicity consists in a step by step use of a very similar

process, to embrace each time one new variable, after the process has already been completed for a number of variables.

To go to the case $n = 2$, let us start with two functions, $f_1(\alpha)$ and $f_2(\alpha)$. Using each of these functions alone we form the orthogonal set of functions $H_1(T^\nu\alpha)$ and $H_2(T^\nu\alpha)$ as before. We shall assume that f_1 can be completely expressed in terms of the orthogonal set belonging to H_1 and that f_2 can be completely expressed in terms of the orthogonal set belonging to H_2. This, as we have seen, is no great restriction for practical computation. Now we take $H_1(\alpha)$, and develop it in terms of the past of the orthogonal set belonging to the H_2's in the following form.

$$H_1(\alpha) = \sum_0^\infty H_2(T^{-\nu}\alpha) \int_0^1 H_1(\beta)\overline{H_2(T^{-\nu}\beta)}\, d\beta + r_1(\alpha).$$

We now develop $r_1(\alpha)$ in the form,

$$\begin{aligned} r_1(\alpha) &= \sum_0^\infty H_1(T^{-\nu}\alpha) \int_0^1 r_1(\beta)\overline{H_1(T^{-\nu}\beta)}\, d\beta + r_2(\alpha) \\ &= r_2(\alpha) - \sum_{\nu=0}^\infty H_1(T^{-\nu}\alpha) \sum \int_0^\infty H_2(T^{-\mu}\beta)\overline{H_1(T^{-\nu}\beta)}\, d\beta \\ &\qquad \times \int_0^\infty H_1(\beta)\overline{H_2(T^{-\mu}\beta)}\, d\beta. \end{aligned}$$

This process can be continued indefinitely, and $r_n(\alpha)$ will converge in the mean to a function orthogonal both to the past of $H_1(\alpha)$ and to that of $H_2(\alpha)$. What will be left apart from this remainder will be

$$\begin{aligned} \eta_1(\alpha) &= \sum_0^\infty H_2(T^{-\nu}\alpha) \int_0^1 H_1(\beta)\overline{H_2(T^{-\nu}\beta)}\, d\beta \\ &\quad - \sum_{\nu=0}^\infty \sum_{\mu=0}^\infty H_1(T^{-\nu}\alpha) \int_0^1 H_2(T^{-\mu}\beta)\overline{H_1(T^{-\nu}\beta)}\, d\beta \\ &\quad \times \int_0^1 H_1(\beta)\overline{H_2(T^{-\mu}\beta)}\, d\beta + \sum_{\nu=0}^\infty \sum_{\mu=0}^\infty \sum_{\lambda=0}^\infty H_2(T^{-\nu}\alpha) \int_0^1 H_1(T^{-\mu}\beta)\overline{H_2(T^{-\nu}\beta)}\, d\beta \\ &\quad \times \int_0^1 H_2(T^{-\lambda}\beta)\overline{H_1(T^{-\mu}\beta)}\, d\beta \int_0^1 H_1(\beta)\overline{H_2(T^{-\lambda}\beta)}\, d\beta - \cdots. \end{aligned}$$

It has been shown that this sum converges in the mean, and it is perfectly possible to show that it is expressible in terms of $H_1(\alpha)$ and the pasts of $f_1(\alpha)$ and $f_2(\alpha)$.

Similarly,

$$\begin{aligned} \eta_2(\alpha) &= \sum_0^\infty H_1(T^{-\nu}\alpha) \int_0^1 H_2(\beta)\overline{H_1(T^{-\nu}\beta)}\, d\beta \\ &\quad - \sum_{\nu=0}^\infty \sum_{\mu=0}^\infty H_2(T^{-\nu}\alpha) \int_0^1 H_1(T^{-\mu}\beta)\overline{H_2(T^{-\nu}\beta)}\, d\beta \int_0^1 H_2(\beta)\overline{H_1(T^{-\mu}\beta)}\, d\beta + \cdots \end{aligned}$$

converges in the mean, and may be expressed in terms of $H_2(\alpha)$ and the past of the f's. If $\eta_1(\alpha)$ and $\eta_2(\alpha)$ are not linearly independent, then either $H_1(\alpha)$ may be expressed in terms of $H_2(\alpha)$ and its past, or $H_2(\alpha)$ may be expressed in terms of $H_1(\alpha)$ and its past. In other words, only one of the two variables $H_1(\alpha)$ and $H_2(\alpha)$ really occurs in the fundamental prediction problem. Otherwise, we may orthogonalize $\eta_1(\alpha)$ and $\eta_2(\alpha)$ by the formulae:

$$X_1(\alpha) = \frac{\eta_1(\alpha)}{\left(\int_0^1 |\eta_1(\beta)|^2 d\beta\right)^{1/2}};$$

$$X_2(\alpha) = \frac{\eta_2(\alpha) - X_1(\alpha)\int_0^1 \eta_2(\beta)\overline{X_1(\beta)}\, d\beta}{\left(\int_0^1 |\eta_2(\beta)|^2 d\beta - \left|\int_0^1 \eta_2(\beta)\overline{X_1(\beta)}\, d\beta\right|^2\right)^{1/2}}.$$

We shall then have as a normal and orthogonal set:

$$X_1(T^\nu\alpha); \qquad X_2(T^\nu\alpha).$$

These will take the place of the $H_1(T^\nu\alpha)$ and $H_2(T^\nu\alpha)$ of formula (2).

In the problem of continuous prediction, we are now up against the fact that the set of orthogonal functions $H(T^\nu\alpha)$ which occurs in the problem of discrete prediction, and the similar set which occurs in the set of discrete multiple prediction, are replaced by functions which are no longer of the Lebesgue class L_2. This is not a finally forbidding difficulty, as it is possible to introduce the Hilbert theory of spectra to take the place of a theory of orthogonal functions. Still, the theory of spectra is much more detailed and inconvenient than that of orthogonal functions, and we must consider ourselves fortunate that there is a method to avoid introducing it directly. This depends on the fact that in the prediction theory which we have already developed, which makes use of measure-preserving point transformations, we may completely replace these measure-preserving point transformations by a unitary functional transformation. That is, wherever $f(T^\nu\alpha)$ appears, we may introduce an expression $T^\nu f(\alpha)$, where $T^\nu f(\alpha)$ is a linear transformation of Hilbert space into itself, and preserves all lengths and distances in Hilbert space.

Now, although there is a continuous group of measure-preserving functional transformations which plays the same role in continuous prediction theory that the discrete group of powers of a single measure-preserving transformation does in discrete prediction theory, and it is impossible to map out a continuous group on any such discrete group, there is a discrete group of functional transformations whose future is the same as the future of the continuous group. In order to introduce these functional transformations, let me introduce the Laguerre polynomials.

If I consider the expression $e^{i\omega\nu}$ which occurs in prediction theory, there is closely related to it the expression

$$\left(\frac{1 + i\vartheta/2}{1 - i\vartheta/2}\right)^\nu.$$

Similarly, if I replace $e^{i\omega\nu}(d\omega)^{1/2}$ by

$$\frac{(1 + i\vartheta/2)^n}{(1 - i\vartheta/2)^{n+1}} (d\vartheta)^{1/2},$$

I shall have found a way to transform the interval from $-\pi$ to π into the interval from $-\infty$ to ∞, if only I put $\vartheta = 2 \tan (\omega/2)$. If now I examine the functions

$$\frac{1}{(2\pi)^{1/2}} \frac{(1 + i\vartheta/2)^n}{(1 - i\vartheta/2)^{n+1}},$$

they will clearly prove to be a normal and orthogonal set if n runs through all integral values from $-\infty$ to ∞. We shall have

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{\left(1 + \frac{i\vartheta}{2}\right)^n}{\left(1 - \frac{i\vartheta}{2}\right)^{n+1}} e^{-i\vartheta t}\, d\vartheta$$

$$= \int_{-\infty}^{\infty} \sum_{0}^{n-1} A_k \frac{1}{\left(1 - \frac{i\vartheta}{2}\right)^{k+1}} e^{-i\vartheta t}\, d\vartheta \begin{cases} = \sum_{0}^{n} B t^k e^{-2t} \\ = 0 \ (t < 0), \end{cases} = p(t)e^{-2t} \ (t > 0)$$

where $p(t)$ is an appropriate polynomial of the nth degree. Similarly if n is 0 or negative, we shall have

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{\left(1 + \frac{i\vartheta}{2}\right)^n}{\left(1 - \frac{i\vartheta}{2}\right)^{n+1}} e^{-i\vartheta t} d\vartheta$$

$$= \int_{-\infty}^{\infty} \sum_{0}^{n-1} A_k \frac{1}{\left(1 - \frac{i\vartheta}{2}\right)^{k+1}} e^{-i\vartheta t}\, d\vartheta \begin{cases} = \sum_{0}^{-n} C_k t^k e^{2t} \\ = 0 \ (t > 0). \end{cases} = g(t)e^{2t} \ (t < 0)$$

In other words the Fourier transforms of

$$\frac{\left(1 + \frac{i\vartheta}{2}\right)^n}{\left(1 - \frac{i\vartheta}{2}\right)^{n+1}} \qquad (n > 0)$$

differ from 0 only on the positive half line, and those of

$$\frac{\left(1 + \frac{i\vartheta}{2}\right)^n}{\left(1 - \frac{i\vartheta}{2}\right)^{n+1}} \qquad (n \leqq 0)$$

differ from 0 only on the negative half line. The functions

$$p_n(t)e^{-2t} \ (t > 0); \qquad 0 \ (t < 0)$$

$$g_n(t)e^{2t} \ (t < 0); \qquad 0 \ (t > 0)$$

are known, with proper normalization with respect to dependent and independent variables, as the Laguerre functions. The transformation of each Laguerre function into the next later function is a unitary transformation whose powers, positive and negative, constitute a discrete group of all translations along the line. I repeat that up to the present we have used in our prediction theory no properties of our transformations which involve their being used as point transformations rather than general unitary valuations. Thus on the frequency scale, the change from $e^{in\omega}$ to

$$\frac{\left(1 + \frac{i\vartheta}{2}\right)^n}{\left(1 - \frac{i\vartheta}{2}\right)^n}$$

is one which involves no difficulty.

In view of this transformation we are now in a position to factor functions of frequency running from $-\infty$ to ∞, whether they are scalar or matrix functions, into a product in which one term is the transform of functions vanishing only for past time and the other is the transform of functions vanishing only for future time. The previous conditions of factorizability

$$\Phi(\omega) \in L \qquad (-\pi \leqq \omega \leqq \pi)$$

and

$$\int_{-\pi}^{\pi} |\log \Phi(\omega)| \, d\omega < \infty$$

are clearly replaced by the equivalent conditions

$$\Phi^*(\varphi) \in L, \qquad -\infty < \varphi < \infty$$

and

$$\int_{-\infty}^{\infty} \frac{1}{1 + \vartheta^2} |\log \Phi^*(\vartheta)| \, d\vartheta < \infty,$$

if only

$$\Phi(\omega) = \Phi(\vartheta);$$

$$\omega = 2 \tan^{-1} \frac{\vartheta}{2}.$$

Similar results and equivalences hold in the case of matrix factorization. If

$$\Psi^*(\vartheta) = \Psi(\omega),$$

we have

$$\Phi^*(\vartheta) = |\Psi^*(\omega)|^2,$$

and the Fourier transforms of the L_2 functions $\Psi^*(\vartheta)$ will contain no negative frequencies.

We now come to the actual mechanism of prediction. We have not time to take this up in more than the simplest case; namely, that of a one-dimensional discrete prediction, but the methods are valid with the most obvious changes both for the continuous and the multiple case. Let then

$$\sum A_n f(T^{-n}\alpha)$$

be any polynomial in the past which we desire to study as an approximation to $f(T^\nu\alpha)$. The mean square error of prediction will then be

$$\int_0^1 |\sum A_n f(T^{-n}\alpha) - f(T^\nu\alpha)|^2 d\alpha.$$

As we now wish to write this in terms of frequency rather than time, it will become

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} |\sum A_n \Psi(\omega) e^{in\omega} - \Psi(\omega)e^{-i\nu\omega}|^2 d\omega = \sum_{m=-\infty}^{\infty} \left|\sum_n \psi_{m-n-\nu} - \psi_m\right|^2.$$

It will then be seen that we have to reduce at the same time

$$\sum_n A_n \psi_{m-n-\nu}$$

as near to 0 as we can for $m < \nu$ and as near to ψ_m as we can for $n \geqq \nu$. If we have at our disposal not merely polynomials, but arbitrary combintations of the past, this will give us

$$\sum_{m=0}^{\infty} e^{im}(\sum_n A_n \psi_{m-n-\nu}) = \sum_p^{\infty} \psi_n e^{in\omega}$$

or, where ψ_m represents the Fourier coefficients of $\Psi(\omega)$,

$$\sum_0^{\infty} A_n e^{in\omega} = e^{i\nu\omega} \frac{\sum_p^{\infty} \psi_n e^{in\omega}}{\sum_0^{\infty} \psi_n e^{in\omega}}.$$

However, it is not difficult to prove that even if we have only polynomials at our disposal, we may reduce

$$\int_0^1 |\sum A_n f(T^{-n}\alpha) - f(T^\nu\alpha)|^2 d\alpha = \sum_0^p |A_n|^2$$

as near its absolute minimum as we wish.

We thus have solved the problem of prediction as nearly as we wish to optimum prediction in the case where the prediction is not perfect. Where the least mean

square error of prediction is 0, while there is in general no single optimum prediction, it is possible to approximate as nearly as we wish to perfect prediction by the following means:

In the first place, we blur the spectrum of the function which we wish to predict by taking its convolution with a narrow Gaussian distribution of unit area. Then we obtain the optimum prediction on the strength of this blurred spectrum. It may be shown that as the Gaussian distribution gets narrower and narrower the optimum prediction becomes more and more perfect.

To return to the case where there is not a perfect prediction, and in which the function

$$\sum_0^\infty e^{in\omega} \int_0^1 f(T^\nu\alpha)\overline{f(\alpha)}\, d\alpha,$$

can be factored. Let us note that formally, from the point of view of frequency rather than time, the optimum prediction operator for a lead μ amounts to multiplication by

$$\frac{1}{2\pi}\sum_{\nu=0}^\infty e^{-i\nu\omega}\int_{-\pi}^{\pi}\Psi(u)e^{-i\nu u}\, e^{i\mu u}\, du,$$

and that the mean square error of prediction is

$$\sum_{\nu=0}^{\mu}\left|\frac{1}{2\pi}\int_{-\pi}^{\pi}\Psi(u)e^{-i\nu u}\, du\right|^2.$$

From the original Kolmororoff point of view, this frequency treatment is not important, but from our point of view the frequency representation of operators is important just because it is the standard representation of alternating current engineering. It is in this form that we strive to realize operators through a network of coils, resistances, and condensers, and in fact the prediction operators which we have just obtained are very generally suitable for realization. This leads us to the problem of filtering.

I shall describe this in terms of the continuous case rather than the discrete, because filtering is commonly an electrical engineering operation, although indeed its precise analogue is useful in the statistical laboratory.

We start then with a message $f_1(\alpha)$ and a noise $f_2(\alpha)$ and we put formally

$$\phi_{11}(t) = \int_0^1 f_1(T^t\alpha)\,\overline{f_1(\alpha)}\, d\alpha; \qquad f(\alpha) = f_1(\alpha) + f_2(\alpha);$$

$$q(t) = \int_0^1 f_1(T^t\alpha)\,\overline{f(\alpha)}\, d\alpha; \qquad \phi(t) = \int_0^1 f(T^t\alpha)\,\overline{f(\alpha)}\, d\alpha;$$

$$\Phi_{11}(\omega) \sim \int_{-\infty}^{\infty}\phi_{11}(t)e^{it\omega}\, dt; \qquad \Phi(\omega) \sim \int_{-\infty}^{\infty}\phi(t)e^{i\omega t}\, dt;$$

$$k(\omega) \sim \int_0^\infty K(t)e^{i\omega t}\, dt; \qquad \Phi(\omega) = |\Psi(\omega)|^2$$

as before. The problem which we have is to minimize

$$\int_0^1 | f_1(T^{-\lambda}\alpha) - \int_0^\infty f(T^{-\tau}\alpha)K(\tau)\, d\tau\, |^2\, d\alpha.$$

Formally, this leads to the process of minimizing

$$\phi_{11}(0) - 2\mathrm{Re}\int_0^\infty q(\tau - \lambda)K(\tau)\, d\tau + \int_0^\infty K(\tau)\, d\tau \int_0^\infty \overline{K(\sigma)}\phi(\sigma - \tau)\, d\sigma$$

$$= \int_{-\infty}^\infty \left| \frac{\Phi(\omega)e^{-i\lambda\omega}}{\overline{\Psi(\omega)}} - k(\omega)\Psi(\omega) \right|^2 + \text{const.}$$

Thus again, formally, the optimum prediction is given by the frequency operator

$$k(\omega) = \frac{1}{2\pi\Psi(\omega)} \int_0^\infty e^{-i\omega t}\, dt \int_{-\infty}^\infty e^{itu} \frac{\Psi_1(u)}{\overline{\Psi(u)}} e^{-i\lambda u}\, du.$$

Let us notice that the technique of prediction may be carried over to multiple time series. From an engineering point of view this means that we have a number of messages linearly jumbled, but we can put them through an apparatus so that each message will come out of it in as pure a form as possible.

As I have said before this allows us to use interference to eliminate a message as well as simple attenuation available in the ordinary filter.

There are a number of other topics for which I have no more time available than enough simply to mention them. In the first place, the methods of multiple prediction make it possible to analyze the direction of causality in complicated situations. In the second place, the whole theory of prediction as given up to this point involves a perfect knowledge of statistical parameters of the past. This knowledge is in fact never available. It must be supplemented by some theory of extenuation from which we can obtain not merely the most probably values of our spectra, but also our distribution. We have made some headway in the problem of extenuation of parameters in the case where time series represents the impact on a resonator of a large number of randomly distributed phenomena. However, not even in this case have we brought the estimated theory to a point where it is yet suitable for practical computation, and secondly this is by no means the only significant case of linear time series.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY,
CAMBRIDGE, MASS., U. S. A.