Predictive Coding—Part I

PETER ELIAS†

Summary—**Predictive coding is a procedure for transmitting messages which are sequences of magnitudes. In this coding method, the transmitter and the receiver store past message terms, and from them estimate the value of the next message term. The transmitter transmits, not the message term, but the difference between it and its predicted value. At the receiver this error term is added to the receiver prediction to reproduce the message term. This procedure is defined and messages, prediction, entropy, and ideal coding are discussed to provide a basis for Part II, which will give the mathematical criterion for the best predictor for use in the predictive coding of particular messages, will give examples of such messages, and will show that the error term which is transmitted in predictive coding may always be coded efficiently.**

Introduction

TWO MAJOR contributions have been made within the past few years to the mathematical theory of communication. One of these is Wiener's work on the prediction and filtering of random, stationary time series, and the other is Shannon's work, defining the information content of a message which is such a time series, and relating this quantity to the bandwidth and time required for the transmission of the message.[1] This paper makes use of the point of view suggested by Wiener's work on prediction to attack a problem in Shannon's field: prediction is used to make possible the efficient coding of a class of messages of considerable physical interest.

Consider a message which is a time series, a function m_i which is defined for all integer i, positive or negative. Such a series might be derived from the sampling used in a pulse-code modulation system.[2] From a knowledge of the statistics of the set of messages to be transmitted, we may find a predictor which operates on all the past values of the function, m_j with j less than i, and produces a prediction p_i of the value which m will next assume. Now consider the error e_i, which is defined as the difference between the message and its predicted value:

$$e_i = m_i - p_i . \tag{1}$$

All of the information generated by the source in selecting the term m_i is given just as well by e_i; the error term may be transmitted, and will enable the receiver to reconstruct the original message, for the portion of the message that is not transmitted, p_i, may be considered as information about the *past* of the message and not about its present; indeed, since p_i is a quite determinate mathematical function, it contains no information at all by Shannon's definition of this quantity.[3]

The communications procedure which will be discussed is illustrated in Fig. 1. There is a message-generating source that feeds into a memory at the transmitter. The transmitter has a predictor, which operates on the past of the message as stored in the memory to produce an estimate of its future. The subtractor subtracts the prediction from the message term and produces an error term e_i, which is applied as an input to the coder. The coder codes the error term, and this coded term is sent to the receiver. In the receiver the transmitting process is reversed. The receiver also has a memory and an identical predictor, and has predicted the same value p_i for the message as did the predictor at the transmitter. When the coded correction term is received, it is decoded to reproduce the error term e_i. This is added to the predicted value p_i and the message term m_i is reproduced. The message term is then presented to the observer at the receiver, and is also stored in the receiver memory to permit the prediction of the following values of the message.

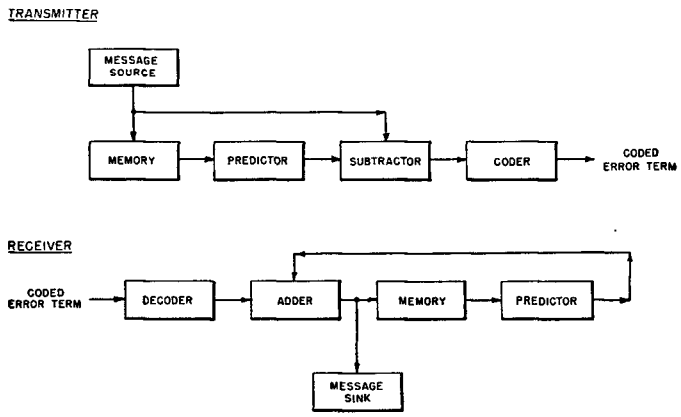

Fig. 1—Predicting coding and decoding procedure.

This procedure is essentially a coding scheme, and will be called *predictive coding*. The memory, predictor, subtractor, and coder at the transmitter, and the memory, predictor, adder, and decoder at the receiver may be considered as complex coding and decoding devices. Predictive coding may then be compared with the ideal coding methods given by Shannon and Fano.[4] In general,

† Elec. Engrg. Dept. and Res. Lab. Elec., Mass. Inst. Tech., Cambridge, Mass.

[1] For historical remarks on the origin of modern information theory see C. E. Shannon and W. Weaver, "The Mathematical Theory of Communication," Univ. of Illinois Press, Urbana, Ill., p. 52 (footnote) and p. 95 (footnote); 1949.

[2] B. M. Oliver, J. R. Pierce, and C. E. Shannon, "The philosophy of PCM," Proc. I.R.E., vol. 36, pp. 1324–1331; November, 1948; also, W. R. Bennett, "Spectra of quantized signals," *Bell Sys. Tech. Jour.*, vol. 27, pp. 446–472; July, 1948.

[3] Shannon and Weaver, *op. cit.*, p. 31.

[4] Shannon and Weaver, *op. cit.*, p. 30; also R. M. Fano, Tech. Rep. No. 65, Res. Lab. Elect., M.I.T., Cambridge, Mass.; 1949.

predictive coding cannot take less channel space for the transmission of a message at a given rate than does an ideal coding scheme, and it will often take more. However, there is a large class of message-generating processes which are at present coded in a highly inefficient way, and for which the use of large codebook memories, such as are required for the ideal coding methods, is impractical. Time series which are obtained by sampling a smoothly varying function of time are examples in this class. For many such processes predictive coding can give an efficient code, using a reasonable amount of apparatus at the transmitter and the receiver.

It should be noted that in the transmission scheme of Fig. 1 errors accumulate. That is, any noise which is introduced after the transmitter memory, or at the receiver, or in transmission, will be perpetuated as an error in all future values of the message, as will any discrepancy between the operation of the two memories, or the two predictors. This means that eventually errors will accumulate to such an extent that the message will disappear in the noise. If, therefore, continuous messages, i.e., time series each member of which is selected from a continuum of magnitudes, are to be transmitted, it will be necessary periodically to clear the memories of both the receiver and the transmitter and start afresh. This is undesirable, since after each such clearing there will be no remembered values on which to base a prediction, and more information transmission will be required for a period following each such clearing, until enough remembered values have accumulated to permit good prediction once more.

A more satisfactory alternative is the use of some pulse-code transmission system in which only quantized magnitudes of input are accepted. Such a system may be made virtually error-free.[5] A system of this kind has the further advantage that the only very reliable memory units now available or in immediate prospect are of a quantized nature, most of them being capable only of storing binary digits. The use of a quantized system requires that the predicted values be selected from the permissible quantized set of message values. Strictly interpreted, this severely limits the permissible predictors; if by a choice of scale the permissible quantized levels are made equal to the integers, then the restriction on $p(m_{i-1} \cdots m_{i-n})$ is that it take integer values for all sets of integer arguments. Actually the ordinary extrapolation formulas have this property, and may be used as predictors. But it is not necessary to limit the choice of predictors so severely. The problem may be evaded by using any function as a predictor and computing its value to a predetermined number of places by digital computing techniques, the prediction then being taken to be the function rounded off to the nearest integer. If the predictor as originally computed was optimum in some well-defined sense, then the rounded predictor will presumably be less good in that sense, but in cases where predictive coding may be expected to be useful the difference will usually be small.

[5] Oliver, Pierce, and Shannon, *loc. cit.*

It is necessary to define precisely what is meant by an optimum predictor for use in predictive coding—i.e., to define some quantity, which depends upon the choice of the predictor, and define as optimum a predictor which minimizes this quantity. Wiener's work uses as a criterion the minimization of the mean square error term $\overline{e^2}$. Wiener has pointed out that other criteria are possible, but that the mathematical work is made simpler by the mean square choice.[6] Minimizing the mean square error corresponds to minimizing the power of the error term, and if no further coding is to be done, this is a reasonable criterion for predictive coding purposes. However, in the system illustrated in Fig. 1, the error term is coded before it is transmitted, and its power may be radically altered in the coding process. What we are really interested in minimizing is the channel space which the system will require for the transmission of the error term. This leads to the following criterion which will be justified in Part II of this paper: *That predictor is best which leads to an average error-term distribution having minimum entropy.*

The coder of Fig. 1 also requires some consideration. Predictive coding eliminates the codebook requirement by using prediction. To take advantage of the resultant savings in equipment, it is necessary to show that the coder itself will not require a large codebook. This reduces to the problem of showing that a message whose terms are assumed independent of one another may always be coded efficiently by a process with a small memory requirement. It will be shown that this is true. It is necessary to use two kinds of coding processes: one for cases in which the entropy of the distribution from which the successive terms are chosen is large compared to unity, and another for cases in which the entropy is small compared to unity.

The following sections of the present paper are devoted to a discussion of messages, prediction, entropy, and ideal coding. Part II will discuss the predictor criterion given above, the classes of messages for which a predictor that is optimum by this criterion may be found, and other classes of messages for which predictive coding may be of use. Mathematically defined examples of message-generating processes which belong to these classes will be given, and the problem of coding the error term so as to take advantage of the minimal entropy of its average distribution will be examined.

Characterization of Messages

A necessary preliminary to a discussion of messages is a precise definition of what "message" is taken to mean.[7] Since a communication system is designed to transmit many messages, what is actually of interest is the

[6] N. Wiener, "The Extrapolation, Interpolation and Smoothing of Stationary Time Series with Engineering Applications," published in 1942 as an NDRC report, and in 1949 as a book, by the Mass. Inst. Tech. Press, Cambridge, Mass., and John Wiley & Sons, Inc., New York, N. Y., especially p. 13.

[7] Such definitions are given by Wiener, *ibid.*, and Wiener, "Cybernetics," Mass. Inst. Tech. Press, and John Wiley & Sons, Inc., 1948; also by Shannon and Weaver, *loc. cit.* Our discussion starts with a definition like Wiener's and ends with one like Shannon's.

characterization of the ensemble from which the transmitted messages are chosen, or the stochastic process by which they are generated. As a preliminary definition, we may say that a message is a single-valued real function of time, chosen from an ensemble of such functions. It will be denoted by $m(a, t)$, where a is a real number between zero and one which labels the particular message chosen from the ensemble, and $m(a, t)$ is defined, for each such a, for all values of t from $-\infty$ to ∞. This definition must be restricted in several respects, in part to take into account the physical requirements of transmitting systems and in part for mathematical convenience.

First, it is assumed that the ensemble from which the messages are chosen is ergodic. This means that any one message of the ensemble, except for a set whose measure in a is zero, is typical of the ensemble in the following sense: let $Q(a)$ be the probability distribution of the parameter of distribution a. Then with probability one, for any function $f[m(a, t)]$ and almost any a_1 ,

$$\lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} f[m(a_1 , t)]\, dt = \int_0^1 f[m(a, t)]\, dQ(a). \qquad (2)$$

I.e., any function of m has the same average value when averaged over time as a function of a single message, as when averaged over the ensemble of all possible messages. We can thus find out all possible statistical information about the ensemble by observing a single message over its entire history. The ergodic requirement implies that the ensemble is stationary: i.e., that the statistics do not change with time. Its practical importance is that it permits us to speak indifferently of the message or the ensemble, and makes it unnecessary to specify the sense in which we speak of an average. In particular, it permits the substitution of measurable time averages for experimentally awkward ensemble averages.

Second, it is assumed that the average square of the message [in either sense of (2)] is finite. The message will be represented in physical systems by a voltage or a current, or the displacement of a membrane, or the pressure in a gas, or by several such physical variables, as it proceeds from its origin to its destination. All of these representations require power; in particular, representation as a voltage or a current between two points separated by a fixed impedance, which is a necessary intermediate representation in any presently used electrical communication method, requires a power proportional to the square of the message. Since only a finite amount of power may be supplied to a physical transmitter, it is obviously required that the average message power be bounded.

Third, it is assumed that the spectrum of the message vanishes for frequencies greater than some fixed frequency f_0 . This will not in general be true for the radio-frequency spectrum of the messages as they are generated by a source, and it has been shown that a function with an infinitely extended spectrum cannot be reduced to a function with a spectrum of finite range by any physically realizable filter; the transfer characteristic of a filter can be zero only for a set of frequencies of total measure zero.[8] However, this is no practical problem. For since the message has a finite total power distributed over the spectrum, there will always be an f_0 so high that a negligible fraction of the total power will be located beyond it in the power spectrum.

The reason for this assumption is that, as Shannon has pointed out, any function of time that is band-limited may be replaced by a time series, which gives the values of the function at times separated by an interval $1/2f_0$.[9] For any band-limited function we have the following identity:

$$m(t) = \sum_{i=-\infty}^{\infty} m(i/2f_0)\left\{\frac{\sin \pi(2f_0 t - i)}{\pi(2f_0 t - i)}\right\}. \qquad (3)$$

The values of the function at the sampling points $t = i/2f_0$, which are the coefficients of this series, thus completely determine the function. If the function is not initially band-limited, the expansion will give a function which passes through the same values at the sampling points, but which is band-limited. As we assume band-limited messages, for our purpose the series and the function are equivalent, and since the series is easier to deal with in the sequel, it is desirable to change the definition of the message. Henceforth the message will be defined as the series of coefficients in the expansion (3). By choice of the unit of time, the sampling interval is made unity, and the message is then $m_i(a)$, defined for all (positive and negative) integer values of the index i.

A message is thus a time series drawn from an ergodic ensemble of such series, and each term in any one message is drawn from a probability distribution whose form is determined by the preceding terms of that message. For the reasons indicated in the first section, we will be interested primarily in quantized messages, for which this probability distribution will be discrete. However, it will at times be more convenient in the analysis and the examples to deal with continuous distributions, it being understood that quantization will ultimately be used. In the discrete case, the message term m_i will be selected from a discrete probability distribution M_k , where $M_k(m_{i-1} \cdots m_{i-j} \cdots)$ is the conditional distribution giving the probability that, for a particular set of past values $m_{i-1} \cdots m_{i-j} \cdots$, the message term m_i will take the integer value k. In the continuous case, the message term m_i will be chosen from a continuous conditional distribution $M(m_i : m_{i-1} \cdots m_{i-j} \cdots)$. Both of these distributions are dependent on the set of values of the preceding message terms $m_{i-1} \cdots m_{i-j} \cdots$, but are of course independent of the value of the index i, by the stationary nature of the ensemble.

[8] Wiener, "The Extrapolation, Interpolation and Smoothing of Stationary Time Series with Engineering Applications," NDRC Report, Mass. Inst. Tech. Press, Cambridge, Mass., p. 37; 1942.

[9] C. E. Shannon, "Communication in the presence of noise," Proc. I.R.E., vol. 37, pp. 10–21; January, 1949.

Stochastic processes of this sort are known as Markoff processes and have an extensive mathematical literature.[10] An nth order Markoff process is one in which the distribution from which each term is chosen depends on the set of values of the n preceding terms only; a process in which each term is chosen from a single unconditional probability distribution may be called a Markoff process of order zero. It should be noted that, while any Markoff process yielding a message with a finite second moment is included in this definition, we will expect most of the messages to be Markoff processes of a rather special kind. The messages have been derived by the time-sampling of a continuously varying physical quantity. The sampling rate must be high enough so that the sampling does not suppress significant variations in the message—i.e., the f_0 must be above the bulk of the spectral power of the message. Now for most such messages, the average rate of variation with time is much lower than the highest rate that the system must be capable of transmitting. Consequently, it is to be expected that on the average, successive message values will be near to one another. This means, in particular, that in the discrete case the index k is not just an arbitrary labeling of a particular symbol—as it is, for example, in Shannon's finite-order Markoff approximations to English[11]—but may be expected to give a genuine metric: message values with indexes near to one another may be expected to have probabilities near to one another, and the conditional distributions mentioned above may be expected to be unimodal. This is not a restriction on what kinds of series will be considered to be messages, but is rather a specification of the class of messages for which predictive coding may be expected to be of use, as will be discussed in detail in Part II of this paper.

For a message ensemble for which the conditional distributions are not given *a priori*, it is necessary to determine them by the observation of a number of messages, or of a single message for a long time. It is obviously impossible to do this on the assumption that the distribution from which a particular message term is chosen depends on the infinite set of past message values. What can, in fact, be measured are the zeroth order approximation, in which each term is treated as if it were drawn from the same distribution, giving $M(m_i)$, an unconditional distribution; the first order conditional distribution $M(m_i : m_{i-1})$, and so on to the nth order conditional distribution for some finite n. A communications system which is designed to transmit this approximation will be inefficient: the approximating process itself would generate messages with a greater information content than the messages which are actually being transmitted, and a system designed for the approximation will waste time or power or bandwidth when transmitting the real message. This will be discussed more fully later.

[10] Shannon and Weaver, *op. cit.*, p. 15; also, M. Frechet, cited there, and P. Levy, "Processus Stochastique et Mouvement Brownien," Gauthier-Villars, 1948, which give further references.

[11] Shannon and Weaver, *op. cit.*, pp. 9–15.

Prediction

Norbert Wiener has developed a very general method for finding the linear predictor for a given ensemble of messages which minimizes the root mean square error of prediction. His method was developed for the difficult case of nonband-limited messages, i.e., continuous functions of time which cannot be reduced to time series. However, he has also solved the much simpler problem of the prediction of time series, such as the messages which were defined above. The details of this work are thoroughly covered in the literature,[12] and this section will merely define some terms, note some results, and discuss the prediction problem from a point of view which is weighted towards probability considerations and not towards Fourier transform considerations.

From a time series, a linear prediction p_i of the value of a message term m_i is a linear combination of the previous message values

$$p_i = \sum_{j=1}^{\infty} a_j m_{i-j} .$$

The error e_i is defined as

$$e_i = p_i - m_i .$$

The predictor itself may be considered to be the set of coefficients a_j . The best linear predictor, in the rms sense, is the set of coefficients which, on the average, minimizes e^2. Wiener has shown that this predictor is determined, not by the message ensemble directly, but by the autocorrelation function of the ensemble. In general, there will be many ensembles with the same autocorrelation function, and the same linear predictor will be the best in the rms sense for all of them.

The autocorrelation function for a time series is defined by

$$c_k = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{i=-N}^{N} m_i m_{i-k} .$$

Devices for rapidly obtaining approximate autocorrelation functions have been constructed.[13] These devices accept the message directly as an input, and graph or tabulate the function. By the use of such devices, or by a statistical examination of the message, or in some cases by an *a priori* knowledge of the message-generating process, it is possible to determine the autocorrelation function. The best linear predictor in the rms sense may then be determined. But it should be noted that there may be nonlinear predictors which are very much better.

Indeed, given a complete knowledge of the stochastic definition of the message, i.e., a complete knowledge of

[12] Wiener, *op. cit.* Also H. W. Bode and C. E. Shannon, "A simplified derivation of linear least square smoothing and prediction theory," Proc. IRE, vol. 38, pp. 417–425; April, 1950.

[13] T. P. Cheatham, Jr., Tech. Rep. No. 122, Res. Lab. Elect., M. I. T. (to be published). See also, Y. W. Lee, T. P. Cheatham, Jr., and J. B. Wiesner, "The Application of Correlation Functions in the Detection of Small Signals in Noise," Tech. Rep. No. 141, Res. Lab. Elect., M. I. T.; 1949.

the conditional probability distributions $M(m_i : m_{i-1} \cdots m_{i-j} \cdots)$ or $M_k(m_{i-1} \cdots m_{i-j} \cdots)$ the *best* rms predictor, with no restriction as to linearity, is directly available. Obviously the best rms predictor for a message term m_i defined in this way is the mean of the distribution from which it is chosen, which is determined by the past message history: i.e., the best rms predictor, p^*, is

$$p^* = \bar{m}_i = \sum_{k=-\infty}^{\infty} kM_k(m_{i-1} \cdots m_{i-j} \cdots)$$

or

$$= \int_{-\infty}^{\infty} m_i M(m_i : m_{i-1} \cdots m_{i-j} \cdots)\, dm_i$$

in the discrete and continuous cases respectively. For the mean of a distribution is that point about which its second moment is a minimum. Of course, the mean need not be a linear function of the past message values. However, it is some determinate function of these values unless the message values are completely uncorrelated—i.e., unless the Markoff process is of order zero. In this case, it is just the constant which is the mean of the zero-order distribution. We therefore have as the unconditionally best rms predictor the function $p^*(m_{i-1} \cdots m_{i-j} \cdots)$.

From this same general statistical viewpoint the best predictor on a mean-absolute error basis is the prediction of the median of the conditional distribution, since the median is that point about which the first absolute moment is a minimum. Like the mean, the median is defined by the conditional distribution M as a function of the past history of the message. This definition may not be unique: if there is a region of zero probability density between the two halves of a probability distribution, any point in the region is a median. However, the definition may be made unique by selecting a point within this range, for those sets of past message values for which the ambiguity arises. We will denote the best predictor in the mean-absolute sense by p^{**}, it being understood that the definition has been made unique in some suitable way if the ensemble is such as to require this.

Finally, it may be desired to predict in such a way that in the discrete case, the probability of no error is a maximum, and in the continuous case the probability density has the maximum possible value at zero error. This requires modal prediction. The mode of the conditional distribution will not be unique if there are several equal probabilities which are each larger than any other probability in the discrete case, or if the continuous distribution attains its maximum value at more than one point. The difficulty may again be removed by a suitable choice, and p^{***} will signify the best modal predictor.

In any of these cases, and indeed for any other prediction criterion which yields a determinate value of the prediction as a function of the past history of the message, the error term e_i is drawn from a distribution $E(e_i : m_{i-1} \cdots m_{i-j} \cdots)$ or $E_k(m_{i-1} \cdots m_{i-j} \cdots)$ which is of exactly the same form as the original distribution of the message term, but which has been shifted along the axis by the amount of the prediction. If it is desired to limit predictions to the possible quantized values of a discrete probability distribution, it is only necessary to make p^{**} and p^{***} unique in a way which does this in the cases of ambiguity; where the median and mode are uniquely defined, they will always coincide with one of the possible values of the message. For rms prediction it is necessary to take the quantized value that is nearest to the computed mean of the distribution as the value of p^*.

As an example of a predictable function, consider

$$M(m_i : m_{i-1}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-(m_i - am_{i-1})^2/2\sigma^2\right]. \quad (4)$$

The unconditional distribution of m_i may be found by using the reproductive property of the normal distribution. $\overline{M}(m_i)$ will be normal, with a standard deviation σ', and am_{i-1} will have a normal distribution with standard deviation $a\sigma'$: then,

$$\sigma^2 + a^2\sigma'^2 = \sigma'^2; \qquad \sigma' = \frac{\sigma}{\sqrt{1-a^2}} \quad (5)$$

and

$$\overline{M}(m_i) = \frac{1}{\sigma'\sqrt{2\pi}} \exp\left[-m_i^2/2\sigma'^2\right]. \quad (6)$$

The zero-order approximation to this first-order Markoff process has, then, a message term distribution of the same form as the original conditional distribution, but a standard deviation which is larger by a factor $1/\sqrt{1-a^2}$. By our definition in a previous section the process will generate messages only if $a < 1$: otherwise the standard deviation will be infinite, and the message will require infinite power for transmission. A more general example in complete analogy to (4) is:

$$M(m_i : m_{i-1} \cdots m_{i-j} \cdots) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\left(m_i - \sum_{j=1}^{\infty} m_{i-j}a_j\right)^2 \Big/ 2^2\right]. \quad (7)$$

Wiener's prediction procedure is designed for functions of the form (7), in which each term of the time series is drawn from a normal distribution with constant σ, with a mean which is a linear combination of past values, the permissible combinations being limited by the requirement that the resultant average distribution have a finite second moment. The linear combination of past values which is the mean of the conditional distribution is also the best linear rms predictor, and is indeed *the* best rms predictor p^*, as noted above. Wiener's method is then a procedure for finding this linear combination in terms of the autocorrelation function of the message.

The combination of past terms in the exponent may be rewritten as a sum of differences, less a constant times the message value m_i . The stochastic function determined by the conditional distribution will then be as approximation to the solution of the difference equation obtained by setting the exponent in (7) equal to zero. In the limit $\sigma \to 0$, the stochastic function will become precisely the

function which is a solution to this equation, as determined by the set of past message values (initial conditions): as σ grows, the function will wander about in the neighborhood of this solution, diverging from it more and more as i increases. In (4) above, the equation obtained is just $m_i - am_{i-1} = 0$, and the solution, $m_i = am_{i-1}$, gives a geometric approach to the origin.

In the case of continuous functions of time, taking appropriate limits gives a normal distribution about a linear function of the past which may include integral or differential operators on the past. The bulk of Wiener's analysis is devoted to this case. Although the method was designed with functions like (7) in mind, it is clearly not limited to them. In the case of time series it is possible to use a distribution which is not normal, with a standard deviation (or other parameter or parameters) which is not constant, but is also determined by the past values of the message. So long as the *mean* of the distribution is still a linear combination of past values, the predictor derived from the autocorrelation function will still give the best rms predictor. If the mean is a nonlinear function of the past values, the predictor obtained from the autocorrelation function will be the best linear approximation to this nonlinear function in the rms sense.

Where the best predictor is indeed linear, or is well approximated by a linear combination of past values, the great practical superiority of Wiener's method over the use of the conditional distribution should be clear. For in this method only the autocorrelation function, a function of a single variable, needs to be measured; the predictor can then be computed no matter what the order of the Markoff process may be. Using the conditional probability distribution directly, an nth order Markoff process will require the observational determination of a function of $n + 1$ variables. This becomes a task of fantastic proportions when n is as large as four or five: it is practical for small n only for a quantized system with very few possible quantized levels.

The direct use of the conditional distribution may, however, be quite valuable if the best rms predictor is a highly nonlinear function of only a few past values, particularly in a quantized system. Nonlinearity is no more difficult to treat than is the linear case as far as analysis by this method is concerned. For the synthesis problem the lack of suitable nonlinear elements for the physical construction of nonlinear operators on the past is confined to the case of continuous functions of time; in the case of time series with quantized terms, digital computer techniques can provide any desired nonlinear function of any number of variables—at, of course, an expense in equipment which may become very large for large n.

When the conditional distribution always has a point of symmetry, we may note that the best rms predictor p^* is equal to the best mean absolute predictor p^{**}. If the distribution is also always unimodal, then the best modal predictor p^{***} will also be the same as p^*. In particular, this will be the case for the examples (4) and (7), but it does not, of course, depend on the linearity of the predictor.

Entropy, Averaging, and Ideal Coding

The entropy H of a probability distribution M has been defined as[14]

$$H = -\sum_{k=-\infty}^{\infty} M_k \log M_k$$

and

$$H = -\int_{-\infty}^{\infty} M(m_i) \log M(m_i)\, dm_i \qquad (8)$$

in the discrete and continuous cases, respectively. The entropy of a probability distribution may be used as a measure of the information content of a symbol or message value m_i chosen from this distribution. The choice of the logarithmic base corresponds to the choice of a unit of entropy: when logarithms are taken to the base two, as is convenient in many discrete cases, the unit of entropy is the "bit," a contraction for binary digit, since in a two-symbol system with the two symbols equiprobable, the entropy per symbol is one bit for this choice of base. In the continuous case computations are often made simpler by the use of natural logarithms. The resultant unit of entropy is called by Shannon the natural unit. We have one natural unit $= \log_2 e$ bits.

Wiener, Shannon, and Fano[14] give a number of reasons for the use of this function as a measure of information per symbol, and the arguments are plausible and satisfying, but as Shannon remarks, the ultimate justification of the definition is in the implications and applications of entropy as a measure of information.[15] For the analysis of communications systems, the definition is completely justified by theorems which prove that it is possible to code any message with entropy H bits per symbol in a binary code which uses an average of $H + \epsilon$ binary digits per message symbol, where ϵ is a positive quantity which may be made as small as desired, and by equivalent theorems in the case of the discrete channel with noise —i.e., where there is a finite probability that a symbol

[14] This is the definition given by Shannon (Shannon and Weaver, *op. cit.*) and Fano (R. M. Fano, "The Transmission of Information", Tech. Rep. No. 65, Res. Lab. Elec., M. I. T.; 1949) Wienes ("Cybernetics", *op. cit.*, p. 76) gives a definition with the opposite sign. There is no real conflict here, however, for Wiener is talking about a different measure. Wiener asks, how much information we are given about a message term, whose exact value will never be known, when we are given the probability distribution from which it is chosen. The answer is that we know a good deal when the distribution is narrow, and very little when the distribution is broad. Correspondingly, entropy as Wiener defines it has a large positive value for very narrow distributions and a large negative value for very broad distributions. This measure is useful in determining how much information has been transmitted when a message term which is contaminated by noise with a known distribution is received; we can use Bayes' theorem and find the probability distribution of the original message, and measure information transmitted by measuring the entropy of this distribution. Shannon, on the other hand, asks how much information is transmitted by the precise transmission of a message symbol, when we know *a priori* the probability distribution from which it was selected. In this case, if the distribution is very narrow, the message term tells us very little when it arrives; we knew what it would be before we received it. If the distribution is broad, however, then the arrival of the term tells us a good deal. This requires the use of the opposite sign for entropy. Shannon's definition will be used through this paper; it is the more appropriate one for the kind of problem with which we are concerned.

[15] Shannon and Weaver, *op. cit.*, p. 19.

transmitted at one quantized level will be received at a different level, and in the case of the continuous channel with noise—in which the message term is chosen from a continuous distribution, and is received mixed with noise, so that each received term is the sum of a signal term and a noise term, and reception is always approximate.

For messages as defined in a previous section, we have, in general, that the entropy of the distribution from which any single message term is chosen is a function of the message history: in both the continuous and discrete cases we are concerned with conditional distributions, whose form depends on the set of values of the terms $m_{i-1} \cdots m_{i-j} \cdots$ which precede the message term m_i whose entropy is defined in (8). For such cases—i.e., Markoff processes of order one or greater—the entropy is defined in terms of the probability, not of each message term, but of a sequence of N message terms, the limit being taken as N approaches infinity. Following Shannon,[16] we define G_N in the discrete and the continuous cases as

$$G_N = -(1/N) \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} M(m_i , \cdots m_{i-N}) \cdot \log M(m_i , \cdots m_{i-N})\, dm_i \cdots dm_{i-N}$$

$$= -(1/N) \sum_{m_i=-\infty}^{\infty} \cdots \sum_{m_{i-N}=-\infty}^{\infty} M(m_i , \cdots m_{i-N}) \cdot \log M(m_i , \cdots m_{i-N}). \quad (9)$$

Then the entropy per symbol of the process is defined as

$$H = \lim_{N \to \infty} G_N . \quad (10)$$

The distribution $M(m_i , \cdots m_{i-N})$ in (9) is not a conditional but a joint distribution: the distribution which determines the probability of getting a given set of N values for the $N + 1$ message terms m_{i-N} to m_i . Now the joint probability distribution of order $N + 1$ is related to the conditional probability distribution and the joint distribution of order N by

$$M(m_i , \cdots m_{i-N}) = M(m_i : m_{i-1} \cdots m_{i-N}) M(m_{i-1} , \cdots m_{i-N}). \quad (11)$$

Using the relation (11) in the expression (9), for a message generating process which is a Markoff process of finite order k, and taking the limit (10), we have

$$H = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} M(m_{i-1} , \cdots m_{i-k})\, dm_{i-1} \cdots dm_{i-k} \cdot \left\{ \int_{-\infty}^{\infty} M(m_i : m_{i-1} \cdots m_{i-k}) \cdot \log M(m_i : m_{i-1} \cdots m_{i-k})\, dm_i \right\} \quad (12)$$

with a similar relation for the discrete case, in which the integrals are replaced by sums. In words, what (12) states is that the entropy for the process as a whole is just the average over-all past histories of the entropy of the conditional distribution of order k which defines the process: the information content per symbol of a message generated by such a stochastic process is the average of the entropies of the distribution from which the successive message terms are chosen.

It was noted that only a finite order Markoff process can, in general, be used as a model of a message source, and that, in general, the use of such an approximation is inefficient. We may now state this more exactly. If a kth order Markoff process is approximated by a process of order less than k, then the entropy of the approximating process will be greater than or equal to the entropy of the original process, with the equality holding only if the original process is actually of order less than k: i.e., only if the kth order conditional distribution can be expressed in terms of conditional distributions of lower order. The result holds also for suitably convergent processes of infinite order. It is a direct consequence of the following more general theorem.

Averaging Theorem I

Let $P(x: y)$ be a probability density distribution of x, for each value of the parameter y: i.e., for all y,

$$\int_{-\infty}^{\infty} P(x: y)\, dx = 1,$$

and

$$P(x: y) \geq 0$$

for all x and y. Let $Q(y)$ be a probability density distribution of y:

$$\int_{-\infty}^{\infty} Q(y)\, dy = 1$$

$$Q(y) \geq 0.$$

Let $R(x)$ be the distribution $P(x: y)$ averaged over the parameter y, and let H' be its entropy:

$$R(x) = \int_{-\infty}^{\infty} Q(y) P(x: y)\, dy$$

$$H' = -\int_{-\infty}^{\infty} R(x) \log R(x)\, dx. \quad (13)$$

Let $H(y)$ be the entropy of the distribution $P(x: y)$ as a function of the parameter y, and let H be its average value:

$$H(y) = -\int_{-\infty}^{\infty} P(x: y) \log P(x: y)\, dx$$

$$H = \int_{-\infty}^{\infty} Q(y) H(y)\, dy.$$

Then we always have $H' \geq H$, and the equality holds only when the y dependence of $P(x: y)$ is fictitious. In words, the entropy of the average distribution is always greater than the average of the entropy of the distribution.

[16] Shannon and Weaver, *op. cit.*, p. 25.

The proof is given in the appendix.[17] The theorem remains true for discrete distributions, and the statement is unchanged except for the uniform substitution of the summation indexes i and j for the continuous variables x and y and the replacement of integrations by sums. By successive application of the proof it is also obvious that the result holds for a distribution which is a function of n parameters y_1 to y_n . The application to Markoff processes is direct, for a conditional distribution of order $k - 1$ may be expressed as an integral of the form $R(x)$ in (13), where $P(x: y)$ is the conditional distribution of order k and y is the term m_{i-k} .

The theorem is also applicable to cases in which the dependence of the distribution on past history is not explicit. If the dependence of the distribution $M(m_i : m_{i-1} \cdots m_{i-k})$ on the set of past message values is through a dependence on one or several parameters (e.g., the mean and the standard deviation of a distribution are functions of the set of past message values but the distribution is always normal), the conclusion still holds: the entropy of the average distribution, averaged over the distribution of the parameters, is always greater than the average over the parameters of the entropy. This is illustrated by the example of (4). The average message term distribution of the process is a normal distribution with a standard deviation $\sigma/\sqrt{1 - a^2}$, with an entropy which may easily be computed[18] as

$$H_0 = \log \sigma\sqrt{2\pi e} + \log (1/\sqrt{1 - a^2}),$$

but each message term has a normal distribution with standard deviation, with entropy just

$$H = \log \sigma\sqrt{2\pi e},$$

which is thus the average entropy of the process as a whole. The difference between these two entropies may be made as large as we like by letting a approach one.

A second averaging theorem which will be useful later deals with averages over a single distribution.

Averaging Theorem II

Let $P(x)$ be a probability distribution with entropy H:

$$\int_{-\infty}^{\infty} P(x)\, dx = 1, \qquad P(x) \geq 0 \qquad \text{for all } x,$$

$$H = -\int_{-\infty}^{\infty} P(x) \log P(x)\, dx.$$

Let $Q(x, y)$ be a weighting function:

$$\int_{-\infty}^{\infty} Q(x, y)\, dx = \int_{-\infty}^{\infty} Q(x, y)\, dy = 1,$$

$$Q(x, y) \geq 0 \qquad \text{for all } x \text{ and } y.$$

Let $R(x)$ be the averaged distribution with entropy H':

$$R(x) = \int_{-\infty}^{\infty} P(y)Q(x, y)\, dy$$

$$H' = -\int_{-\infty}^{\infty} R(x) \log R(x)\, dx.$$

Then we always have $H' \geq H$, and the equality holds only when the weighting function is a Dirac delta function.

This theorem is given by Shannon.[19] It is also true in the discrete case: the equality then holds only if the average distribution $R(x)$, or R_i in the discrete case, is a mere permutation of the distribution $P(x)$, or P_i .

At the beginning of this section it was stated that it is possible to code a message with entropy H bits per symbol by a coding method which uses $H + \epsilon$ binary output symbols per input symbol, on an average. Such a coding scheme will be called an *ideal code*. Shannon has given two such coding procedures, and Fano has given one which is quite similar to one of Shannon's.[20] We will call coding by means of Shannon's second procedure, or by means of Fano's method, *Shannon-Fano coding*. Both are procedures for giving short codes to common messages and long codes to rare messages. They are given in the references. We will here only note the important result. Coding a group of N message terms at once, the average number H_1 of output binary symbols per input message symbol is bounded:

$$G_N \leq H_1 \leq G_N + 1/N. \tag{14}$$

Here G_N is the quantity defined in (9). As N increases G_N approaches H, the true entropy of the process, so H_1 also approaches H. For a discrete process, an *efficient* code may be defined as one for which the ratio H/H_1 is near one. It is clear that there are two reasons why a Shannon-Fano code for small N may be inefficient: first, if G_N is small, the ratio G_N/H_1 may be small, if H_1 is near its upper bound in (14). Second, for small N, G_N may be a poor approximation to H.

It should be noted that it is *not* reasonable to define an efficiency measure for continuous distributions as a ratio of entropies. For a process which is ultimately to be quantized, the entropy of a continuous distribution does not approximate the entropy of the discrete distribution which is obtained by quantization, unless the scale of the variable in the continuous distribution is so chosen as to make the interval between quantized levels unity. Using a different choice of scale adds a constant to the entropy of the distribution, so that the ratio which defines efficiency is changed. For this reason, until a quantizing level spacing is chosen, it is possible to speak only of the differences between the entropies of continuous distributions, and not of their ratios.

[17] The content of this theorem is implied by the derivations leading up to Shannon's fundamental theorem, Shannon and Weaver, *op. cit.*, p. 28. However, the theorem can be stated and proved as a property of entropy as a functional of a probability distribution, with no reference to sequences of message terms, and the proof is so straightforward and simple that the theorem deserves an independent statement.

[18] Shannon and Weaver, *op. cit.*, p. 56.

[19] Shannon and Weaver, *op. cit.*, p. 21, property 4 for the discrete case; p. 55, property 3 for the continuous case.

[20] Shannon and Weaver, *op. cit.*, p. 29; Fano, *op. cit.* Shannon's procedure is simpler to handle mathematically; Fano's is perhaps somewhat simpler to grasp. Fano's method is not quite completely determinate. In cases in which the two methods do not agree, Fano's provides a more efficient code than Shannon's.

Appendix

Proof of Averaging Theorem I

Expanding H' by the definitions given, we have

$$H' = -\int_{-\infty}^{\infty} dx \left\{ \int_{-\infty}^{\infty} dy\, Q(y)P(x:y) \log \int_{-\infty}^{\infty} dz\, Q(z)P(x:z) \right\}.$$

Adding and subtracting a term gives

$$H' = -\int_{-\infty}^{\infty} dx \left\{ \int_{-\infty}^{\infty} dy\, Q(y)P(x:y) \log P(x:y) + \int_{-\infty}^{\infty} dy\, Q(y)P(x:y) \log \frac{\int_{-\infty}^{\infty} dz\, Q(z)P(x:z)}{P(x:y)} \right\}.$$

The quotient in the last integral cannot cause trouble, since the integrand as a whole approaches zero with $P(x:y)$. Interchanging the order of integration in the first integral and using the definition of H gives

$$H' = H - \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy\, Q(y)P(x:y) \cdot \log \frac{\int_{-\infty}^{\infty} dz\, Q(z)P(x:z)}{P(x:y)}. \tag{14}$$

Changing the logarithmic base will multiply both sides of (14) by the same constant, so we are free to use natural logarithms and measure entropy in natural units. Using the inequality $\log u \leq u - 1$ in the integral in (14) gives

$$H' \geq H - \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy\, Q(y)P(x:y) \cdot \left\{ -1 + \frac{\int_{-\infty}^{\infty} dz\, Q(z)P(x:z)}{P(x:y)} \right\}$$

$$\geq H + \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy\, Q(y)P(x:y) - \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy \int_{-\infty}^{\infty} dz\, Q(y)Q(z)P(x:z).$$

Integrating first with respect to x, we have by the normalization requirements on $P(x:y)$ and $Q(y)$ that

$$H' \geq H + 1 - 1 \geq H.$$

The equality can be realized only when $\log u \equiv 1$, or in this case when

$$\int_{-\infty}^{\infty} P(x:z)Q(z)\, dz \equiv P(x:y). \tag{15}$$

For this to hold, $P(x:y)$ must have no dependence on variable y, since y does not appear on the left of (15), Q.E.D. In the discrete case, the precise same proof holds when summations are uniformly substituted for integrations.

Predictive Coding—Part II

Summary—**In Part I predictive coding was defined and messages, prediction, entropy, and ideal coding were discussed. In the present paper the criterion to be used for predictors for the purpose of predictive coding is defined: that predictor is optimum in the information theory (IT) sense which minimizes the entropy of the average error-term distribution. Ordered averages of distributions are defined and it is shown that if a predictor gives an ordered average error term distribution it will be a best IT predictor. Special classes of messages are considered for which a best IT predictor can easily be found, and some examples are given.**

The error terms which are transmitted in predictive coding are treated as if they were statistically independent. If this is indeed the case, or a good approximation, then it is still necessary to show that sequences of message terms which are statistically independent may always be coded efficiently, without impractically large memory requirements, in order to show that predictive coding may be practical and efficient in such cases. This is done in the final section of this paper.

Definition of Information-Theory Criterion for Predictors

We have now a sufficient vocabulary and collection of results to define and discuss a criterion of prediction that is appropriate for the kind of communications scheme outlined in the Introduction. An obvious definition is: that predictor is best, in the sense of information theory, which requires the minimum channel space for the transmission of its error term. But this specification is not yet sufficient. It is necessary to define to some extent the way in which the error term is to be coded, in order to define a predictor uniquely for a given message-generating process.

One procedure is to use Shannon-Fano coding for the transmission of the error term. This means that the predictor $p(m_{i-1} \cdots m_{i-j} \cdots)$ should be chosen to minimize the ensemble average of the entropy of the error distribution. The average of

$$-\int_{-\infty}^{\infty} E(m_i : m_{i-1} \cdots m_{i-J} \cdots) \cdot \log E(m_i : m_{i-1} \cdots m_{i-j} \cdots)\, dm_i \tag{16}$$

or of

$$-\sum_{-\infty}^{\infty} E_k(m_{i-1} \cdots m_{i-j} \cdots) \log E_k(m_{i-1} \cdots m_{i-j} \cdots)$$

is to be minimized, the averaging being done over the