region is included in that for which $\rho = |\rho|$, and further for $\rho \geq 0$ each region is included in that obtained by using P_1 and P_2, the maximum values of σ_1^2 and σ_2^2. Therefore C_{fb} is given by (3), as desired.

IV. Discussion

We have demonstrated a feedback coding scheme for the additive white Gaussian MAC which allows reliable communication at all points in the capacity region of the channel. The scheme is intrinsically interesting since it is deterministic and provides doubly exponential error decay, at least for the scheme described in Section II with $\alpha = 0$. In addition, along the way we have constructively shown what the capacity region is. The outer bound C_0 given by (23) is relatively simple to obtain, but in no other case has it been shown to be achievable. The largest achievable region for the discrete or Gaussian MAC with feedback to date is that of Cover and Leung [7], which for the Gaussian case is strictly smaller than the region obtained here. It remains to be seen when, for general discrete channels, C_0 is achievable. Willems [8] has shown that for a class of discrete MAC's, the region of Cover and Leung is optimal.

Acknowledgment

The author wishes to acknowledge the advice of Professors R. Gallager and P. Humblet of Massachusetts Institute of Technology, and Prof. C. Leung of University of British Columbia during the research and preparation of [9], from which this paper is largely drawn.

References

[1] T. Cover, "Some advances in broadcast channels," in *Advances in Communication Systems*, Vol. 4, A. Viterbi, Ed. San Francisco: Academic Press, 1975.
[2] A. D. Wyner, "Recent results in Shannon theory," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 2–10, Jan. 1974.
[3] J. P. M. Schalkwijk and T. Kailath, "A Coding scheme for additive noise channels with feedback—Part I: No bandwidth constraint," *IEEE Trans. Inform. Theory*, vol. IT-12, pp. 172–182, April 1966.
[4] J. P. M. Schalkwijk, "A coding scheme for additive noise channels with feedback—Part II: Band-limited signals," *IEEE Trans. Inform. Theory*, vol. IT-12, pp. 183–189, April 1966.
[5] R. Ahlswede, "Multi-way communication channels," in *Proc. 2nd Int. Symp. Inform. Theory*, Tsahkadsor, Armenian S.S.R., 1971, pp. 103–135.
[6] T. Berger, *Rate Distortion Theory*. Englewood Cliffs: Prentice Hall, 1971.
[7] T. Cover and S. K. Leung-Yan-Cheong, "A rate region for multiple access channels with feedback," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 292–298, May 1981.
[8] F. Willems, "The feedback capacity region of a class of discrete memoryless multiple-access channels," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 93–95, Jan. 1982.
[9] L. H. Ozarow, "Coding and capacity of additive white Gaussian noise multi-user channels with feedback," Ph.D. dissertation, Mass. Inst. Tech., Cambridge, MA, May 1979.
[10] N. T. Gaarder and J. K. Wolf, "The capacity region of a multiple access discrete memoryless channel can increase with feedback," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 100–102, Jan. 1975.

Universal Coding, Information, Prediction, and Estimation

JORMA RISSANEN

Abstract—**A connection between universal codes and the problems of prediction and statistical estimation is established. A known lower bound for the mean length of universal codes is sharpened and generalized, and optimum universal codes constructed. The bound is defined to give the information in strings relative to the considered class of processes. The earlier derived minimum description length criterion for estimation of parameters, including their number, is given a fundamental information theoretic justification by showing that its estimators achieve the information in the strings. It is also shown that one cannot do prediction in Gaussian autoregressive moving average (ARMA) processes below a bound, which is determined by the information in the data.**

Manuscript received July 13, 1983; revised January 16, 1984. This work was presented in part at the IEEE International Symposium on Information Theory, St. Jovite, Canada, September 26–30, 1983.

This work was done while the author was Visiting Professor at the Department of System Science, University of California, Los Angeles, while on leave from the IBM Research Laboratory, San Jose, CA 95193.

I. Introduction

THERE are three main problems in signal processing: prediction, data compression, and estimation. In the first, we are given a string of observed data points x_t, $t = 1, \cdots, n$, one after another, and the objective is to predict for each t the next outcome x_{t+1} from what we have seen so far. In the data compression problem we are given a similar sequence of observations, each truncated to some finite precision, and the objective is to redescribe the data with a suitably designed code as efficiently as possible, i.e., with a short code length.

0018-9448/84/0700-0629$01.00

Finally, in estimation—the most fundamental and important problem of them all—we seek an "explanation" of the observations, or, rather, of the underlying mechanism, which we believe has generated the data. More precisely, we select a parametrically defined statistical model $P_\theta(x)$ for the data x, and try to estimate the vector parameter $\theta = (\theta_1, \cdots, \theta_m)$ from the observations, where the number of the parameters m is also to be estimated.

We show that these problems are intimately related, with the common link being the information in the string. This information is defined as

$$\min_{m,\theta} \{ -\log P_\theta(x) + \tfrac{1}{2} m \log n \}, \qquad (1.1)$$

where n denotes the number of observed data points. (In this paper, we denote the binary logarithm by "log" and the natural logarithm by "ln".) This definition of information in x, relative to a class of parametrically defined processes, is justified by a theorem, which generalizes and strengthens the minimax bounds for universal codes derived in the series of important papers by Davisson [4], [5], Davisson, McEliece, Pursley, Wallace [6], Krichevskii and Trofimov [9], and others. In essence, this theorem states that no matter which universal code one uses, the mean code length is bounded from below by the mean information, given by the expression $-E_\theta \log P_\theta(x) + \frac{1}{2} k \log n$, for "practically all" processes defined by $\theta = (\theta_1, \cdots, \theta_k)$. To emphasize that the mean is taken relative to the indicated process, we used the subindex θ. Moreover, we demonstrate that optimum universal codes exist, for which the lower bound is reached for every process in the considered class.

Inspired by Akaike's pioneering work [1] on criteria that would allow estimation of parameters along with their number without a separate hypothesis testing, this author developed in [14] and [15] an estimation principle based on the purely information theoretic idea: pick the parameters so that the model they define permits the shortest possible representation of the observed sequence. Following a particular recipe of coding steps, involving a rather delicate way of dealing with the prior knowledge about the parameters, or, rather, the lack of it, we were able to derive a closed form expression for the criterion, which turns out to be identical with (1.1). Accordingly, our theorem stated above gives a fundamental justification for the resulting minimum description length (MDL) estimates; they allow the most efficient coding of the observed sequence among all universal codes.

Turning finally to the remaining problem, prediction, we show that in the class of Gaussian autoregressive moving average (ARMA) processes the mean per symbol prediction error of any measurable predictor of the past data is bounded from below by $\sigma^2(\theta)(1 + ((p+q)/n) \ln n)$, where the first factor gives the variance of the innovation sequence of a process with $p + q + 1$ parameters θ. As above, the inequality holds for practically all processes defined by θ. Because this variance is the reachable lower bound for predictors, provided that the process parameters are known, we see that something has to be paid for not knowing them. Perhaps not too surprisingly, the predictor bound is seen to be a function of the information in the string, and hence completely determined by it.

II. Main Theorem

We consider stationary random processes $X(\theta) = \{ X_t | t = 1, 2, \cdots \}$ depending on a vector $\theta = (\theta_1, \cdots, \theta_k)$ of real-valued parameters. When the range of the random variables X_t is a finite or a countable set, the process is assumed to be specified by a time invariant joint probability function $P_\theta(x^n)$, defined for all observations $x^n = x_1, \cdots, x_n$, $n = 1, 2, \cdots$, and satisfying the necessary compatibility condition, namely, that the sum of $P_\theta(xu)$ over the range of the symbol variable u immediately following the string x, is $P_\theta(x)$. When the range of X_t is the real line, the process is assumed to be defined by a density function $f_\theta(x^n)$, which, in turn, determines a probability $P_\theta(x^n)$ for the sequence of the observations x_t truncated to some number α of fractional digits in their binary representation. In this regard, we do not consider the problem of how to truncate a continuous signal optimally in the sense of rate-distortion theory. Instead, we begin with already truncated numbers. The process-defining function is assumed to be a twice continuously differentiable function of θ in a compact k-dimensional set Ω^k. Without repeating the above given description, we speak about the "process or source defined by θ", or even of the "process θ". Finally, the given parametric class generalizes in a straightforward manner to a class of conditional probabilities $P_\theta(x^n/y^m)$, where typically y^m results from an "input" sequence to a system and x^n from its "output" sequence.

The following two types of processes are particularly important.

Example 1: Let X_t take the values 0 and 1, and consider a stationary mth order Markov process. It is defined by the $k = 2^m$ conditional probabilities $P(0/x^m)$, which, in turn, determine the state probabilities $P(x^m)$. Upon the selection of an initial state, the probability $P_\theta(x)$ for every string gets defined in the usual fashion. We let each parameter range over a closed interval $[a, b]$, $a > 0$ and $b < 1$.

Example 2: Let the range of X_t be the real line, and consider a stationary ARMA(p, q) process

$$x_t + a_1 x_{t-1} + \cdots + a_p x_{t-p} = b_0 e_t + b_1 e_{t-1} + \cdots + b_q e_{t-q}, \qquad (2.1)$$

where $\{e_t\}$ is the Wold decomposition of the process x. It is an uncorrelated, zero-mean process. The process x is defined up to the second moments by the $p + q + 1$ parameters $\theta = (a_1, \cdots, a_p, b_0, \cdots, b_q)$, restricted such that the roots of the two polynomials $a(z) = 1 + a_1 z + \cdots + a_p z^p$ and $b(z) = b_0 + b_1 z + \cdots + b_q z^q$ are strictly outside the unit circle. For a minimal parameterization we must further require that the two polynomials have no common factors. An appropriate probability density function to go with such a process results if we specify e_t to be Gaussian of zero-mean and unit variance. This in turn defines a Gaussian density function for x^n. The compact

set Ω^k, $k = p + q + 1$, may be taken as any closed and bounded set with a nonempty interior such that the parameter combinations giving cancelable factors are excluded. Another, almost equally simple but more general, class results if we replace the Gaussian density by an exponential one of type $K\exp - |x|^{\mu}$, where μ is an additional parameter to be estimated.

We consider the problem of how well strings x as samples from a process of the considered type can be compressed, when the only thing known about the process is its type. By compression we mean a coding of a string such that it can be decoded from its code. Such problems have been studied by Davisson, [4], and others, who considered a countable sequence of codes, one for the set of strings of each length n, such that each code is a prefix code satisfying the Kraft inequality. Such a code is called (*weakly*) *universal* if the mean per symbol code length approaches the per symbol entropy for every source in the class, as the length of the string tends to infinity. Moreover, for the class of independent sources, (Davisson *et al.* [6]) and for the class of mth order Markov sources (Davisson [5]), a tight lower bound has been found for the minimax code redundancy

$$\min_{L} \max_{\theta} \left(EL(x^n) - H_n(\theta)\right),$$

where $H_n(\theta)$ denotes the entropy of the source for sequences of length n. In both cases the proof is based on a clever use of Shannon's rate distortion theory.

For our purposes it is more natural to replace a sequence of codes by one code with the length $L(x)$ of the codeword for string x satisfying the Kraft inequality

$$\sum_{x^n} 2^{-L(x^n)} \leq 1, \qquad \text{for all } n. \tag{2.2}$$

We call a code *regular*, if in addition to (2.2) also $L(xu) \geq L(x)$, for all strings $x = x_1, \cdots, x_n$, where $xu = x_1, \cdots, x_n, u$. Although one can construct a nonregular code, we are not aware of any code actually discussed in the literature which would not satisfy this additional requirement. We can then associate with any regular code, satisfying (2.2) with equality, a statistical *model*, which we take to be a probability function $Q(x)$, such that $Q(\lambda) = 1$, where λ denotes the empty string and

$$\sum_{u} Q(xu) = Q(x). \tag{2.3}$$

Indeed, put $Q(x) = 2^{-L(x)}$ and define $Q(u|x) = Q(xu)/Q(x) \leq 1$. Then from

$$\sum_{xu} Q(xu) = 1 = \sum_{x} Q(x) \sum_{u} Q(u|x)$$

and the fact that $\Sigma Q(x) = 1$ we observe by subtraction that the mean of a nonnegative quantity is zero, which gives

$$\sum_{u} Q(u|x) = 1, \qquad \text{for all } x.$$

This immediately implies (2.3).

The relationship between the code length of a regular code and a model becomes one-to-one if we let the length be a positive, real-valued function. Indeed, for any model $Q(x)$ we can put $L(x) = -\log Q(x)$, and the Kraft inequality (2.2) holds with equality. To distinguish this abstract length function from the integer-valued length function, we call it the *ideal code length*, as was done in [13]. The difference is quite irrelevant, because from any ideal code length a regular prefix code can be designed in a routine manner such that the mean per symbol length deviates insignificantly from the mean ideal per symbol length. The justification, then, for the use of the ideal code length is its simplicity. For example, in these terms the code redundancy turns out to be Kullback's information $EL(x^n) - H_n(\theta) = E\log(P_\theta(x^n)/Q(x^n))$. We might say that the notions of a regular code and its ideal length satisfying (2.2) with equality is a coding theoretic equivalent of a random process. But although they are logically equivalent, the code length interpretation is preferable for the reason that it is valid even when the objects we are coding are "deterministic" parameters, admitting no traditional probabilistic interpretation. In fact, the code length *defines* a probability, which, therefore, always can be interpreted in the same coding theoretic manner.

There are two particularly important types of parametrically defined models and the associated code length functions. The first is a model $P_\theta(x^n)$, where the parameter vector $\theta = \theta(x^n)$ is estimated from the observed sequence x^n. When such a model is used to compress strings, a description of the parameters must be attached to the code string as a preamble. This type of model may be called "nonadaptive", because the entire string must be available before the parameters can be estimated, and once they have been estimated they will not change as the string is being encoded. The other basic type of model, called "adaptive", is defined by a conditional probability function $P_{\theta(x^t)}(x_{t+1}|x^t)$, thus

$$L(x) = -\sum_{t=0}^{n-1} \log P_{\theta(x^t)}\left(x_{t+1}|x^t\right). \tag{2.4}$$

Here we set x^0 arbitrarily to some constant, say 0. This length is seen to be regular. Observe that there is no preamble in the code string to include the coding of the parameters. No such preamble is needed, because the decoder is thought to know the rule for calculation of the estimated parameters.

Using quite different arguments from the above cited authors on universal codes, we prove a basic inequality for classes of parametrically defined processes, which generalizes and in a crucial way strengthens the earlier minimax results. In order to shorten the statement of the theorem, we first list the background definitions and conventions, and include in the theorem only the more restricting conditions. We consider a stationary random process $\{X_t|t = 1, 2, \cdots\}$ defined by a probability function $P_\theta(x^n)$, where the parameter θ ranges over a compact subset Ω^k of R^k with a nonempty interior for every $k = 1, 2, \cdots$. For Part

b) we also need the smoothness condition that $P_\theta(x^n)$ be twice continuously differentiable in Ω^k, for each k. This condition is satisfied for many important classes of processes.

Theorem 1: Let the central limit theorem hold for the maximum likelihood (ML) estimators $\theta^*(x^n)$ of each θ in the interior of Ω^k, so that the distribution of $(\theta^*(x^n) - \theta)\sqrt{n}$ converges to the normal distribution with zero-mean and a covariance matrix $\Sigma(\theta)$.

a) If $L(x^n)$ is a length function satisfying the Kraft inequality for each n, then for all k and all positive ϵ

$$n^{-1}E_\theta L(x^n) \geq n^{-1}H_n(\theta) + (\tfrac{1}{2} - \epsilon)(k/n)\log n, \quad (2.5)$$

for all points θ except in a set $A_\epsilon(n)$ whose volume goes to zero as $n \to \infty$. $H_n(\theta)$ denotes the entropy of strings of length n.

b) There exist *optimum* length functions for which the opposite inequality " $<$ " holds for all negative ϵ when $n > n_\epsilon$, and for all θ in Ω, the union of Ω^k over k.

Proof: Part a) is proved in Appendix A, while part b) is shown in Section III.

Remarks: The central limit theorem requirement holds for many important classes of processes. For Markov processes the ML estimates are the count ratios of the symbols at each state, and they are efficient estimators of the transition probabilities. For them the central limit theorem due to De Moivre is a classical result. Also, for Gaussian ARMA processes all the requirements hold, and even in the non-Gaussian case only mild additional regularity conditions are needed (Ljung and Caines [11], Hannan, [8]).

If we subtract the first term in the right-hand side of (2.5) from the left-hand side, we get Kullback's measure for the difference between two distributions (assuming of course that the Kraft inequality holds with equality). Hence, we may interpret (2.5) as a generalization of the basic inequality stating the nonnegativity of Kullback's measure, the generalization being the inclusion of the term measuring the complexity of the system that generates the signal. Because of this, one use of the inequality (2.5) is to provide a measure of the distance between two signals, whose generating mechanisms are not known but must be estimated. But (2.5) has other uses as well and, in fact, it turns out to play a central role in modeling and related contexts.

The statement of the inequality (2.5) is, unfortunately, somewhat complex, and we would clearly like to select the various qualifications so as to have the maximum strength for it. That the inequality cannot hold for all parameters θ (except in the degenerate case of Kullback's where k, the number of free parameters subject to estimation, is zero), is clear, because a length function defined by $-\log P_\theta(x)$ for some fixed θ will, of course, have the entropy as the mean for that single process, determined by the same θ. We may view the earlier minimax bounds for universal codes as weak forms of the inequality (2.5) in that the strict inequality " $<$ " (with $\epsilon = 0$) could hold for the optimum codes for all but one value of θ. The current form strengthens this, because parts a) and b) together imply that for the optimum length functions the right-hand side bound with $\epsilon = 0$ is asymptotically reachable for every θ; it cannot be beaten in the sense that the strict inequality " $<$ " holds except for relatively rare points θ. (Note, however, that Davisson's [5] minimax bound is not a corollary of ours for the reason that his bound is exact rather than asymptotic.)

One may wonder whether the statement in part a) could be further strengthened to the form that not only the volume of $A_\epsilon(n)$ tends to zero but even the union of these sets over $n \geq N$ tends to zero as N tends to infinity. In other words, the measure of limsup of the relevant sets is zero. This turns out to be tied to the rate at which the distribution of the ML estimates approaches the limit. And in the case of Markov sources such a strengthening, indeed, can be achieved. Because such sources are of primary interest only in coding theory rather than in general prediction and estimation and, because the proof for them is difficult, we leave the study of such an ultimate statement of the theorem to another paper.

Another, perhaps more desirable, form of the theorem might result if we add the requirement that the length function $L(x, \theta^*(x))$ be defined by an unbiased estimator θ^*. Then the inequality (2.5) even with $\epsilon = 0$ might hold for all values of θ. However, we have been unable to carry out the proof for the general case.

III. Optimum Codes And Models

In accordance with an optimum code we call a model *optimum*, when its ideal code length is optimum in the sense of part b), Theorem 1. We show now that an optimum model exists, first for the case with fixed k and then for the general case.

Proof: Let $\theta_i''(x^n)$ be the ML estimate $\theta_i^*(x^n)$, truncated to $q(n) = \text{floor}(\log\sqrt{n})$ fractional binary digits, where floor(x) denotes the greatest integer equal to or smaller than x. Now define

$$L(x^n) = -\log P_{\theta''(x^n)}(x^n) + \tfrac{1}{2}k\log n + C(n), \quad (3.1)$$

where $C(n)$ is a normalizing constant, determined so that the Kraft equality holds. In this section, we show that this length is optimum.

When $\theta^* = \theta^*(x)$, where x refers to a string of length n, falls within the interior of Ω^k, we can expand

$$\begin{aligned} -\log P_{\theta''}(x) &= -\log P_{\theta^*}(x) + \tfrac{1}{2}(\theta'' - \theta^*)' \\ &\quad \cdot M(\theta^{**})(\theta'' - \theta^*) \\ &\leq -\log P_\theta(x) + (k/n)K, \end{aligned}$$

where θ^{**} is a point between θ^* and θ'', and K is the maximum norm of the Hessian matrix $M(\theta)$ in Ω^k. Observe that $-\log P_{\theta^*}(x) \leq -\log P_\theta(x)$ by the definition of θ^*; we also used the upper bound $2/\sqrt{n}$ for the components of the difference between $\theta'' = \theta''(x^n)$ and $\theta^* =$

$\theta^*(x^n)$. When again θ^* falls on the boundary, we can write with the same substitutions

$$-\log P_{\theta''}(x) = -\log P_{\theta^*}(x) + D(\theta^{**})(\theta'' - \theta^*) \leq -\log P_\theta(x) + 2D\sqrt{(k/n)},$$

where D denotes the maximum norm of $D(\theta)$ in Ω^k, and again θ^{**} is a point between θ^* and θ''. Using (3.1) we then get in any case

$$(1/n)EL(x^n) \leq H_n(\theta)/n + (k/2n)\log n + C(n)/n + O(1/n^2).$$

We must show that $C(n)/\log n \to 0$. We do it by demonstrating that there is a prefix code for the truncated parameters θ'' with length $L(\theta'') \leq \frac{1}{2}k\log n + O(\log\log n)$. The code is obtained as follows. First, each θ_i'' is converted to an integer $m_i = \theta_i'' 2^{q(n)}$, which is encoded with Elias' universal representation of the integers [7], for the reason that the codeword for the parameters could be separated from the rest of the code. This assigns to integer m the length $c + \log^* m$, where the function $\log^* m$ is defined as the sum $\log m + \log\log m + \cdots$ until the last nonnegative term, and c is a constant. In Leung-Yan-Cheung and Cover [10] it was shown that $2^{-\log^* m}$ is summable; for other interesting properties of this function we refer also to Bentley and Yao [2], and Rissanen [15].

Because Ω^k is compact, there is a uniform upper bound for the m_is, and the code length for all the parameters is $L'(\theta'') = kq(n) + O(\log\log n)$, which also includes the length of encoding the integer $q(n)$. We have shown that $2^{-L'(\theta'')}$, when summed over all values of θ'', does not exceed 1, By putting $L^*(x) = -\log P_{\theta''(x)}(x) + L'(\theta'')$, we have

$$\sum_x 2^{-L^*(x)} \leq \sum_\theta 2^{-L'(\theta)} \sum_{x\in X_\theta} P_\theta(x) \leq \sum_\theta 2^{-L'(\theta)} \leq 1,$$

where x runs through all strings of length n in the first sum, but only through strings such that $\theta''(x) = \theta$, defined as the set X_θ, in the third sum. The parameter θ, in turn, runs through all values such that each component is truncated to $q(n)$ fractional binary digits. By comparing $L^*(x)$ with $L(x)$ in (3.1), we see that $C(n)$ is of the order of $\log\log n$, which in turn implies the optimality of (3.1).

It turns out to be a simple matter to modify the length definition (3.1), which we now denote by $L_k(x)$ to indicate its dependence on the number of parameters k, such that the result is optimum over the entire class Ω of sources. Indeed, put

$$L(x) = \min_m \{L_m(x) + \log^* m + c\}, \qquad (3.2)$$

where $\log^* m = \log m + O(\log\log m)$ and c is a constant. We then have

$$\sum_x 2^{-L(x)} \leq \sum_m 2^{-c-\log^* m} \sum_x 2^{-L_m(x)} \leq 1,$$

where x runs through all the sequences of length n, and m runs through all positive integers.

By the definition of $L(x)$, $L(x) \leq L_k(x) + \log^* k + c$, where k denotes the number of the components in the process-generating parameter θ. Hence with the optimality of $L_k(x)$,

$$EL(x^n) \leq H_n(\theta) + \tfrac{1}{2}k\log n + nr(n), \qquad (3.3)$$

where $nr(n)/\log n \to 0$ as $n \to \infty$. This completes the proof of part b) of Theorem 1.

The optimum model we constructed is clearly nonadaptive, which leaves the question whether adaptive optimum models exist. Such models are required in prediction, while both types are of use in data compression. At the moment we can construct optimum adaptive models for Markov processes and autoregressive (AR) processes only. In Section V, where we discuss prediction, we demonstrate an optimum adaptive model for the AR class while conjecturing its existence for the larger ARMA class.

IV. Information

In view of Theorem 1 we define the quantity

$$I(x^n) = \min_{\theta,k} \{-\log P_\theta(x^n) + \tfrac{1}{2}k\log n\} = -\log P_{\theta^*(x^n)}(x^n) + \tfrac{1}{2}k^*(x^n)\log n \qquad (4.1)$$

to be the *information* in the string x^n, relative to the class Ω of processes. Indeed, our definition is justified, because by Theorem 1 the mean information, $EI(x^n) = H_n(\theta) + \frac{1}{2}k\log n$, with the mean taken relative to the process defined by θ in Ω^k for every k, then provides an asymptotically reachable lower bound for virtually all processes. This notion of information is seen to be a mixture of Shannon's probabilistic and Kolmogorov's algorithmic notions of information, the former referring to the observations x generated by the modeled source, and the latter belonging to the nonrandom selection of the models or parameters.

The mean of the second part $\frac{1}{2}k^*(x^n)\log n$ in the information, taken relative to a θ in Ω^k, is seen to represent the length needed to describe the parameters in the optimum nonadaptive model as constructed in Section III. In the adaptive models no such length is needed, but nevertheless the same length penalizing cost gets added to any optimum model, and possibly more with others. Intuitively speaking, the source of that cost term is the estimation errors that accrue when the model parameters are estimated from the past string. It is still a perplexing fact that the cost is the same in both kinds of models, and we call the term $\frac{1}{2}k^*(x^n)\log n$ *optimum model cost*. This cost term is seen to be a refinement of the measure of model complexity in Rissanen and Langdon [13], which was taken just as the number of parameters. In effect, the refinement takes into account the fact that the optimum precision in writing the parameters grows with the number of observations.

The derived expression for the optimum model cost is of central importance in applications, where it often is replaced by various ad hoc expressions. We illustrate one of its less obvious uses, obtaining an adaptive universal data compression system; in Rissanen [16], where some of these results were anticipated. Suppose that we wish to fit Markov

models, but we do not want to restrict the states to be of any fixed order. It seems first that in order to find the optimum state space and the associated model complexity, we must according to (3.2) perform a multiple parallel encoding process for every $t = 1, \cdots, n$, one encoding for each possible choice of a state space. However, a much simpler way is to collect in a binary tree each conceivable state of any order found in the past string, and gather the occurrence counts of the "next" symbol in each. The actual state, among the many possible ones, which is to be used to encode the next symbol, can then be determined as that state which permits the shortest incremental ideal code length for the next symbol. This state is not simply the state with the shortest ideal code length $\log(c/c_0)$, where c denotes the state count and c_0 the number of times the symbol has been 0 at that state, but rather is the state with the least *predicted* length, the information $\log(c/c_0) + (\log c)/2c$, in which the effect of the parameter estimates is included. The result of this procedure was shown to be successful in [16].

We conclude this section with a generalization of the just-defined information, which was done by Wax and Rissanen [18]. Let x^n and y^m be two strings such that the components are truncated to a finite precision. Define the relative information in the string x^n given the string y^m to be

$$I(x^n|y^m) = \min_{\theta, k} \left\{ -\log P_\theta(x^n|y^m) + \tfrac{1}{2} k \log n \right\}, \tag{4.2}$$

where $P_\theta(x^n|y^m) = P_\theta(x^n, y^m)/P_\theta(y^m)$ denotes the conditional probability, parameterized by θ. Further, define the *information in the string y^m about the string x^n* to be

$$I(x^n, y^m) = I(x^n) - I(x^n|y^m). \tag{4.3}$$

To justify these definitions we prove that the mean of the generalized mutual information (4.3) is nonnegative with the same qualifications as in Theorem 1. If we substitute in the expression of the relative information the "true" process-generating parameters θ^0 and k^0 for θ and k, we get a value $I^0(x^n|y^m)$ which is not smaller than the minimum, i.e., the relative information. by Theorem 1 and the nonnegativity of Shannon's mutual information,

$$EI(x^n, y^m) \geq E\left[I(x^n) - I^0(x^n|y^m)\right] \geq 0,$$

for virtually all points θ^0 in Ω as described in Theorem 1. An application of these notions in measuring the amount of feedback between two processes, where again, the cost of the required models is included, is given in Wax and Rissanen [18].

V. Estimation

The right-hand side expression in (4.1) was derived in Rissanen [14], [15] as the so-called MDL criterion for estimation purposes. The criterion extends the classical maximum likelihood criterion by the second model cost term, and it permits estimation of the number of parameters without the need to use a separate hypothesis test. Unlike the estimates obtained by Akaike's AIC criterion, the MDL estimates $\theta^*(x^n)$ and $k^*(x^n)$ are (strongly) consistent for a wide class of cases, including the ARMA processes [8]. This last case is hard to prove, because the Fisher-information matrix becomes singular with overparameterization. For a thorough study of Akaike's criterion we refer to Shibata [17]. Although consistency is an asymptotic property and does not in itself guarantee good estimation results for small samples, it certainly is a desirable one, and without it the soundness of any criterion appears to us to be suspect.

The MDL criterion, as originally derived, produces estimates which minimize the total code length for x^n when a *particular* coding process is used. Theorem 1, however, removes this arbitrariness and gives the criterion a fundamental justification: *no parameter estimates exist which would permit a shorter mean code for the string regardless of the coding technique used, except for processes defined by parameters in a set of small volume.* In other words, its estimates produce the information in the strings relative to the class Ω of processes.

Theorem 1 provides a more general way to assess the goodness of estimators than the Cramer–Rao inequality in that the number of the parameters is included. Indeed, given any estimator of θ, we may calculate the length function (2.4) and compare it with the optimum length, that is, the information. It seems to us that ultimately such a comparison is all we can meaningfully do in order to get an idea of how "valid" and reliable our estimates are.

One might argue that there is no reason why a criterion should minimize information, except when the resulting model is used for data compression. In fact, if there is one single measure of a model's performance, it ought to be in terms of its capability to predict, since this is what most models ultimately are used for. But we show in Section VI that information and prediction error are virtually measured by the same expression, at least in the Gaussian case, and there is no conflict between the two.

VI. Prediction Bound

In this section we study the question of how well one can predict the numbers x_{t+1} from the past sequence $x^t = x_1, \cdots, x_t$, $t = 0, 1, \cdots$, with $x^0 = 0$, when the only additional information is that some Gaussian ARMA process (2.1) has generated the data. In the past, Davisson [3], has studied a related problem, where he fitted an autoregressive model with p parameters to N observations, and he derived the asymptotic expression $\sigma^2(1 + p/N)$ for the mean-square prediction error of the associated optimum predictor. Moreover, he also proposed to find the optimum value for p, which amounts to the first statement of the "final prediction error" criterion and the closely related AIC criterion for order estimation. As we shall see, this problem formulation is almost the "right" one, but not quite, because it can be shown that the resulting order estimates of p are not consistent.

We propose to measure the prediction error in the accumulated mean square errors along the length of the

data sequence by

$$V_\theta(n) = \sum_{t=0}^{n-1} E\left(x_{t+1} - x^*_{t+1|t}\right)^2, \qquad (6.1)$$

where the predictor $x^*_{t+1|t}$ is a measurable function $g(t, x^t)$ of the past data. We indicate with a subindex θ that the expectation is to be taken relative to an ARMA process with the $k = p + q + 1$ parameters $\theta = (a_1, \cdots, a_p, b_0, \cdots, b_q)$. This measure is meant to incorporate not only the best prediction error at the end of the data, but also the errors along the way, which necessarily are larger because they result from predictions calculated from fewer data points.

Theorem 2: Consider a set of Gaussian ARMA processes, where the $k = p + q + 1$ parameters $\theta = (a_1, \cdots, a_p, b_0, \cdots, b_q)$ range over a compact set Ω^k with nonempty interior for each k. Then for all p, q, positive ϵ, and any predictor $x^*_{t+1|t} = g(t, x^t)$ which is measurable function of the past observations x^t,

$$V_\theta(n)/n \geq \sigma^2(\theta)\left[1 + ((p + 1 - \epsilon)/n)\ln n\right], \qquad (6.2)$$

for all points θ except in a set $A_\epsilon(n)$ whose volume goes to zero as $n \to \infty$. Here $\sigma^2(\theta)$ denotes the variance of the stationary Wold decomposition $\{e_t\}$ of the process defined by θ.

Proof: Consider the sum

$$s^2(x^n) = n^{-1} \sum_{t=1}^{n} u_t^2,$$

where $u_t = x_t - g(t-1, x^{t-1})$ is the prediction error corresponding to a predictor g. Next, let $f(x_{t+1}|x^t)$ denote the conditional Gaussian density function with mean $g(t, x^t)$ and the variance parameter defined as $s^2(x^n)$. If the observations and the predictions are written with precision $\delta = 2^{-\alpha}$, this density defines for the truncated observation x_{t+1} a conditional probability

$$P_{s(x^n)}(x_{t+1}|x^t) = K_t(\delta) f(x_{t+1}|x^t)\delta.$$

The first factor on the right is a normalizing constant such that the sum of these probabilities over all values of x_{t+1} is unity. Clearly, $K_t(\delta) \to 1$ as $\delta \to 0$. When we multiply these conditional probabilities over $t = 0, \cdots, n-1$, we get a function $P_{s(x^n)}(x^n)$, which does not add to unity when summed over all the truncated strings x^n, because the parameter $s(x^n)$ is not constant. However, if we add the optimum code length $\frac{1}{2}\log n$ needed to encode this parameter, to the quantity $-\log P_{s(x^n)}(x^n)$, we get a length function

$$L(x^n) = n\left[\log\left((\sqrt{2\pi})s(x^n)\right) + \tfrac{1}{2}\log e\right] + \tfrac{1}{2}\log n - n\log\delta - \sum_{t=0}^{n-1}\log K_t(\delta),$$

which does satisfy the Kraft inequality to within terms of order $O(\log n)$. It is clearly also regular. Further

$$\mathrm{EL}(x^n) = nE\log\left[(\sqrt{2\pi e})s(x^n)\right] + \tfrac{1}{2}\log n - n\log\delta - \sum_{t=0}^{n-1}\log K_t(\delta),$$

while the per symbol entropy in bits of the truncated x-process generated by θ is given by $\log[(\sqrt{2\pi e})\sigma(\theta)] - \log\delta - C(\delta)$, where $C(\delta)$ accounts for the adjustment due to the truncation. Clearly, $C(\delta) \to 0$ as $\delta \to 0$.

By Theorem 1, on the nonexceptional points mentioned in the theorem,

$$E\log\left[s(x^n)/\sigma(\theta)\right] \geq \left(\tfrac{1}{2} - \epsilon\right)(p + q)n^{-1}\log n - n^{-1}\sum_t \log K_t(\delta) + C(\delta).$$

By Jensen's inequality, applied to (6.1),

$$\log\left(V_\theta(n)/n\right) = \log Es^2(x^n) \geq E\log s^2(x^n).$$

Use this in the previous inequality, and let δ go to zero, to get

$$\log\left(V_\theta(n)/n\sigma^2(\theta)\right) \geq (p + q - \epsilon)n^{-1}\log n.$$

The right-hand side, call it y, is an upper bound for $\ln(1 + y) = \log(1 + y)^{\ln 2}$. By expanding the binomial, and by retaining the two first terms, we deduce the inequality (6.2). □

The issue remains whether the right-hand side bound in (6.2), this time with $\epsilon < 0$, can be reached by some estimator, for example, by the ML estimator. It is readily seen from the above cited result of Davisson that this is true at least for the AR processes. Indeed, let $\sigma^2(\theta)$ denote the variance of the innovation sequence of an AR process defined by the p-component parameter θ. When a pth order AR model is fitted to a sequence of length N and the result is applied to predicting the next symbol generated by the same process, then the mean prediction error is given by a particular instance of the Davisson estimate [3] as $\sigma^2(\theta)(1 + p/N) + o(1/N)$, where $o(1/N)N \to 0$. Now replace N by t and add the result from $t = 0$ to $t = n - 1$ in accordance with (6.1), which gives the sum of a harmonic series

$$V_\theta(n)/n = \sigma^2(\theta)\left[1 + (p/n)\ln n\right] + o(n^{-1}\ln n).$$

This shows first that the bound in (6.2) can be reached, and, further, that the model so obtained is optimum.

As a final remark, the mean square prediction error measure is universally accepted for Gaussian processes, but for others its justification is less convincing. Because of the similarity of the information and the prediction bounds, the former being just the logarithm of the latter, we are tempted to propose the information as the primary measure of prediction. In the Gaussian case this will give virtually the old measure, and the difference will become relevant only in other cases. Perhaps the greatest advantage of the proposed measure is that it makes sense even for discrete random variables, such as those occurring in weather prediction, where as a matter of fact, prediction error is measured in terms of probabilities.

APPENDIX A

Proof of Theorem 1, Part a): For each parameter θ in Ω^k let $E_n(\theta)$ denote a neighborhood of radius $r_n = c/\sqrt{n}$ with θ as its center. Define for the process determined by θ the set of its "θ-typical" strings of length n,

$$X_n(\theta) = \{x^n | \theta^*(x^n) \in E_n(\theta)\}.$$

Let $P_n(\theta)$ denote the sum of $P_\theta(x)$ over x in $X_n(\theta)$. Observe that $P_n(\theta)$ also gives the probability that $(\theta^*(x^n) - \theta)\sqrt{n}$ falls within a neighborhood with radius c centered at the origin. By the assumption, this probability converges to the probability mass of the normal distribution $N(0, \Sigma(\theta))$ that falls within such a neighborhood. This mass therefore exceeds a number $1 - \delta(c)$, $\delta(c)$ tending to zero as c grows, uniformly in θ in the compact set Ω^k. Hence the probability of the θ-typical sequences is bounded by

$$P_n(\theta) \geq 1 - \delta(c)$$

for all n greater than some number n_c, uniformly in θ.

Let $L(x)$ be any length function satisfying the Kraft inequality (2.2) and denote by $Q_n(\theta)$ the sum of $q(x) = 2^{-L(x)}$ over x in $X_n(\theta)$. Then by the basic inequality (McEliece [12, p. 278]),

$$\sum_{x \in X_n(\theta)} P_\theta(x) \log [P_\theta(x)/q(x)] \geq P_n(\theta) \log [P_n(\theta)/Q_n(\theta)]. \tag{A2}$$

Pick a positive number ϵ, and let $A_\epsilon(n)$ be the set of θ such that the left-hand side of (A2), denoted $T_n(\theta)$, satisfies the inequality

$$T_n(\theta) < (1 - \epsilon) \log n^{k/2}. \tag{A3}$$

We wish to calculate an upper bound for the volume of $A_\epsilon(n)$. For this purpose let N_n denote the maximal number of disjoint neighborhoods $E_n(\theta)$ that can be constructed such that the centers θ lie in $A_\epsilon(n)$. Let C_n denote the set of the center points. Of course, these neighborhoods may not cover $A_\epsilon(n)$, because there may be points in $A_\epsilon(n)$ that are too close to some of the constructed neighborhoods, without being covered by any. However, if we expand each neighborhood in the maximal collection by doubling the radius, we get a cover for $A_\epsilon(n)$. Hence the volume V_n of $A_\epsilon(n)$ is bound by

$$V_n \leq K N_n r_n^k, \tag{A4}$$

where K is a constant.

From (A2) and (A3) we conclude that

$$-\log Q_n(\theta) < [(1 - \epsilon)/P_n(\theta) - (\log P_n(\theta))/\log n^{(1/2)k}] \log n^{(1/2)k}$$

for θ in $A_\epsilon(n)$. Pick c so large that $\delta(c)$ in (A1) gets small enough to make the expression within the brackets less than a number α, such that $0 < \alpha < 1$, for all sufficiently large n. Hence

$$Q_n(\theta) > n^{-(1/2)k\alpha} \qquad \text{for } \theta \text{ in } A_\epsilon(n) \tag{A5}$$

for n larger than some number.

The neighborhoods $E_n(\theta)$ for θ in C_n are disjoint by construction, which makes the sets $X_n(\theta)$ disjoint. By the Kraft inequality (2.2)

$$1 \geq \sum_{\theta \in C_n} Q_n(\theta). \tag{A6}$$

From (A4), (A5), and (A6) we get

$$V_n < K c^k n^{(1/2)k(\alpha - 1)},$$

which holds for all sufficiently large values for n. Clearly, $V_n \to 0$. To finish the proof, let θ be in $\Omega^k - A_\epsilon(n)$. Then the opposite inequality " $\geq$ " in (A3) holds. By letting x in (A2) range over the set of all strings of length n, which does not reduce the left-hand side, we see that the claim holds. □

REFERENCES

[1] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Automat. Contr.*, vol. AC-19, pp. 716–723, 1974.

[2] J. L. Bentley and A. C. Yao, "An almost optimal algorithm for unbounded searching," *Inform. Processing Letters*, vol. 5, pp. 82–87, 1976.

[3] L. D. Davisson, "The prediction error of stationary Gaussian time series of unknown covariance," *IEEE Trans. Inform. Theory*, vol. IT-11, pp. 527–532, Oct. 1965.

[4] ——, "Universal noiseless coding," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 783–795, Nov. 1973.

[5] ——, "Minimax noiseless universal coding for Markov sources," *IEEE Trans. Inform. Theory*, vol. IT-29, no. 2, pp. 211–215, 1983.

[6] L. D. Davisson, R. J. McEliece, M. B. Pursley, and M. S. Wallace, "Efficient universal noiseless source codes," *IEEE Trans. Inform. Theory*, vol. IT-27, no. 3, pp. 269–279, May 1981.

[7] P. Elias, "Universal codeword sets and representations of the integers," *IEEE Trans. Inform. Theory*, vol. IT-21, no. 2, pp. 194–203, 1975.

[8] E. J. Hannan, "The estimation of the order of an ARMA process," *Ann. Statist.*, vol. 8, no. 5, pp. 1071–1081, 1980.

[9] R. E. Krichevskii and V. K. Trofimov, "The performance of universal encoding," *IEEE Trans. Inform. Theory*, vol. IT-27, no. 2, pp. 199–207, 1981.

[10] S. K. Leung-Yan-Cheong and T. Cover, "Some equivalences between Shannon entropy and Kolmogorov complexity," *IEEE Trans. Inform. Theory*, vol. IT-24, no. 3, pp. 331–338, 1978.

[11] L. Ljung and P. Caines, "Asymptotic normality of prediction error estimators for approximate system models," *Stochastics*, vol. 3, pp. 29–46, 1979.

[12] R. J. McEliece, *The Theory of Information and Coding.* Reading, MA: Addison-Wesley, 1977.

[13] J. Rissanen and G. G. Langdon, Jr., "Universal modeling and coding," *IEEE Trans. Inform. Theory*, vol. IT-27, no. 1, pp. 12–23, Jan. 1981.

[14] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[15] ——, "A universal prior for integers and estimation by minimum description length," *Ann. Statist.*, vol. 11, no. 2, pp. 416–431, June 1983.

[16] ——, "A universal data compression system," *IEEE Trans. Inform. Theory*, vol. IT-29, no. 5, pp. 656–664, Sept. 1983.

[17] R. Shibata, "Asymptotically efficient selection of the order of the model for estimating parameters of a linear process," *Ann. Statist.*, vol. 8, pp. 147–164, 1980.

[18] M. Wax and J. Rissanen, "Information theoretic measures for feedback between time series," (to appear in *J. Amer. Statist. Assn.*).