The Annals of Probability
1979, Vol. 7, No. 1, 128–138

LOWER BOUNDS FOR NONLINEAR PREDICTION ERROR IN MOVING AVERAGE PROCESSES[1]

By Marek Kanter

Sir George Williams Campus, Concordia University, Montreal

As yet no efficiently computable algorithm for one step nonlinear prediction has been proposed for any general class of stationary processes which performs strictly better than the optimal linear predictor. In this paper it is shown that for the class of stationary moving average processes the improvement obtained by optimal nonlinear prediction versus optimal linear prediction is bounded by a constant which depends only on the distribution of the independent and identically distributed random variables Y_j used to form the moving average process $X_n = \Sigma a_j Y_{n-j}$.

1. Introduction. Let $(a_j; j = 0, \pm 1, \pm 2, \cdots)$ be a two-sided sequence of real numbers with $0 < \Sigma_{-\infty}^{+\infty} a_j^2 < \infty$ and let $(Y_j; j = 0, \pm 1, \pm 2, \cdots)$ be a two-sided sequence of independent identically distributed random variables. We consider in this paper moving average processes of the form

$$X_n = \Sigma_j a_j Y_{n-j}. \tag{1.1}$$

Subsidiary conditions are often needed for the sum in (1.1) to converge a.s., but the condition that $\Sigma_j a_j^2 < \infty$ is always necessary if $Y_j \not\equiv 0$. Processes of the form (1.1) have been often used as stochastic models and indeed the class of such processes seems sufficiently general to give insight into the general behavior of all ergodic stationary processes. (Processes of the above form are easily seen to be ergodic.) In this paper we shall see that the special linear nature of the construction of moving average processes makes possible a fairly painless derivation of some surprising results regarding nonlinear prediction for such processes.

To state our main result we define $\Phi(s) = |\Sigma_{j=-\infty}^{+\infty} a_j e^{ijs}|^2$ and we set $\Delta^2 = \exp(1/2\pi)\int_{-\pi}^{\pi}\log \Phi(s)ds)$. It is well known that in the case when $E(Y_j) = 0$ and $\mathrm{Var}(Y_j) = \sigma^2 < \infty$, then $\Delta^2\sigma^2$ is the mean square error of one step linear prediction, i.e.,

$$\Delta^2\sigma^2 = \inf E\big((X_{n+1} - \Sigma_{j=0}^r b_j X_{n-j})^2\big) \tag{1.2}$$

where the inf is taken over all finite sequences $b_0, \cdots, b_r$ of real numbers. This result is the basic ingredient in the work of Wiener and Kolmogorov on linear prediction theory. To describe our main result we let $\mathcal{Q}$ equal the set of all Borel measurable functions f from R^∞ into R. Then, letting $X^n = (\cdots, X_m, \cdots, X_{n-1}, X_n)$ we prove

$$\inf_{f \in \mathcal{Q}} E\big((X_{n+1} - f(X^n))^2\big) \geq Q(Y_0)\Delta^2 \tag{1.3}$$

Received October 3, 1977.
[1]Research supported in part by the National Research Council of Canada, Grant A8506.
AMS 1970 *subject classifications.* Primary 60G25; Secondary 60G10, 60G15, 62M20, 94A05, 94A15.
Key words and phrases. Moving average stationary process, differential entropy, mutual information, one step prediction error.

where the constant $Q(Y_0)$ is defined to be the variance of that Gaussian random variable whose differential entropy equals the differential entropy of Y_0. The assumption on the common distribution μ of the Y_j is only that we can express μ as an infinite convolution $u = *_{k=0}^{\infty}\mu_k$ where μ_0 is assumed to have bounded density and finite variance while for $k \geqslant 1$ we only assume that μ_k have finite variance. (Note that this allows for μ itself to have infinite variance.) Our result is sharp since $Q(Y_0) = \sigma^2$ if Y_i are Gaussian with variance σ^2. For sequences Y_j with $Q(Y_0) > 0$ and $\sigma^2 < \infty$ we see that nonlinear prediction can improve mean square error by at most a factor $Q(Y_0)/\sigma^2$. In particular if perfect nonlinear prediction is possible then perfect linear prediction is possible for moving average processes with finite variance and $Q(Y_0) > 0$! In this connection it is interesting to point out that there exists an example due to Moran (see [8], page 24) of a strictly stationary process which is perfectly predictable in a nonlinear fashion but not perfectly predictable by linear means (so our result is special to moving average processes).

Our main result raises many further questions. For instance, we do not yet have any example of a moving average process for which $Q(Y_0) < \sigma^2$ and equality holds in (1.3).

A second basic question is to produce constructively a sequence of nonlinear functions of the past which achieve the inf on the left-hand side of (1.3), assuming it is strictly less than $\sigma^2\Delta^2$ (otherwise linear functions of the past would suffice).

A third basic question is to interpret (1.3) in the case when Y_j have infinite variance. The situation is very unclear to us. For example, if $\Delta = 0$ then there is not a single moving average process X_n based on a sequence Y_j with infinite variance for which we can say if the left-hand side of (1.3) is 0, positive, or $+\infty$!

Possibly it is more natural to study mean absolute error in the case when $\mathrm{Var}(Y_j) = \infty$. We can prove

$$\inf_{f \in \mathcal{Q}} E(|X_{n+1} - f(X^n)|) \geqslant \Delta\big((2e)^{-1}\pi Q(Y_0)\big)^{\frac{1}{2}}. \tag{1.4}$$

We shall also indicate how to extend (1.4) so as to provide lower bounds for $E(|X_{n+1} - f(X^n)|^\alpha)$ for any $\alpha > 0$. We remark that if $\Delta = 0$ and $\mathrm{Var}(Y_j) = \infty$ then we do not know if the left-hand side of (1.4) is 0 or nonzero. We shall also show that (1.4) is sharp (in that it becomes an equality if X_n is an autoregressive process based on symmetric two-sided exponential random variables Y_j).

2. Definitions and preliminary lemmas. In this section we present some definitions, notations, and some known lemmas relating to information theory. The reader is referred to Berger [1] and Billingsley [2] for details.

Let $X = (X_j; j = 0, \pm 1, \cdots)$ stand for any stochastic process. For $m < n$ we let X_m^n stand for the vector $(X_m, X_{m+1}, \cdots, X_{n-1}, X_n)$. We set $X^n = X_{-\infty}^n$.

DEFINITION 2.1. If X_m^n assumes only finitely many or countably many values (which we shall label as x_j) then we define the entropy of X_m^n (denoted by $H(X_m^n)$) by setting $H(X_m^n)) = \Sigma - p_j \log p_j$, where $p_j = P[X_m^n = x_j]$. If X is also stationary then it is known that $(1/(n+1))H(X_0^n)$ is decreasing. We set $H(X) = \lim_{n\to\infty} 1/(n+1)H(X_0^n)$.

DEFINITION 2.2. If X is stationary but X_n^m assume uncountably many values then the definition of $H(X)$ is more complex (it is due to Kolmogorov). For any n we let $\tau_n \circ X$ be that process X' such that $X'_k = X_{k+n}$ a.s. We let $\hat{R}^\infty$ stand for the set of all two-sided sequences of real numbers and we let $\hat{\mathfrak{Q}}$ be the set of all Borel measurable functions from $\hat{R}^\infty$ into R which assume only finitely many values. We can then define

$$H(X) = \sup H(Z)$$

where the sup is taken over all processes Z of the form $Z_n = f(\tau_n \circ X)$ where $f \in \hat{\mathfrak{Q}}$.

DEFINITION 2.3. If X_m^n has a density function $p(x)$ we define the differential entropy $h(X_m^n)$ by setting

$$h(X_m^n) = \int_{R^{n-m+1}} -p(x)\log p(x)dx. \tag{2.1}$$

There are only three things that can go wrong in this definition. The right-hand side may be $+\infty$, $-\infty$, or undefined. To simplify things we shall work with bounded densities. This forces the right-hand side of (2.1) to be well defined, assuming either a finite value or $+\infty$. (If the right-hand side of (2.1) is well defined and is finite, then, of course, the function $-p(x)\log p(x)$ is integrable in the sense that $\int|p(x)\log p(x)|dx < \infty$). If X is stationary we define $h(X) = \lim_{n\to\infty}(1/(n+1))$ $h(X_0^n)$. (We shall see later that $1/(n+1)h(X_0^n)$ is decreasing as n increases.)

DEFINITION 2.4. Let $(Y_m, \cdots, Y_n)$ be a multivariate Gaussian random vector with $q_{ij} = E(Y_iY_j)$ for $m \leqslant i, j \leqslant n$ and $E(Y_i) = 0$. It is well known that

$$\frac{1}{n-m+1}h(Y_m^n) = \tfrac{1}{2}\log\left(2\pi e(\det Q)^{1/(n-m+1)}\right) \tag{2.2}$$

where $Q = (q_{ij})$. Furthermore it is well known that if $(X_m, \cdots, X_n)$ is a random vector with $q_{ij} = E(X_iX_j)$ then $h(X_m^n) \leqslant h(Y_m^n)$. We shall define $Q(X_m^n) = \sigma^2$ if the random vector $(Z_m, \cdots, Z_n)$ of independent Gaussian random variables Z_j with $E(Z_j) = 0$ and $E(Z_j^2) = \sigma^2$ satisfies $h(Z_m^n) = h(X_m^n)$. Clearly we can write

$$Q(X_m^n) = (2\pi e)^{-1}e^{(2/(n-m+1))h(X_m^n)}.$$

If X is stationary we define $Q(X) = \lim_{n\to\infty} Q(X_0^n)$. Clearly we have

$$Q(X) = (2\pi e)^{-1}e^{2h(X)}.$$

DEFINITION 2.5. If $X_m^n = (X_n, \cdots, X_m)$ and $Y_r^k = (Y_r, \cdots, Y_k)$ are jointly distributed random vectors we define $I(X_n^m, Y_r^k) = +\infty$ if $\mu_{X,Y}$ is not absolutely continuous with respect to $\mu_X \times \mu_Y$ where $\mu_{X,Y}$, μ_X and μ_Y stand for the joint distribution and marginal distributions of (X_m^n, Y_r^k). Otherwise we set $\lambda = (d\mu_{X,Y})/(d\mu_X \times d\mu_Y)$ and define

$$I(X_m^n, Y_r^k) = \int \log \lambda d\mu_{X,Y}.$$

The quantity $I(X_m^n, Y_r^k)$ is $\geqslant 0$ and is called the mutual information between X_m^n and Y_r^k. The following lemma is well known.

LEMMA 2.1. *Let X_m^n and Y_r^k be as above.*

(1) *If* $H(X_m^n) < \infty$ *then*

$$I(X_m^n, Y_r^k) = H(X_m^n) - H(X_m^n | Y_r^k).$$

(2) *If* X_m^n *has density* $p(x_n, \cdots, x_n)$ *with* $\int |p(x) \log p(x)| dx < \infty$ *then* $I(X_m^n, Y_r^k) = h(X_m^n) - h(X_m^n | Y_r^k)$.

(3) *Letting* $W = (X_m, \cdots, X_n, Y_r, \cdots, Y_k)$ *then* $h(W) = h(X_m^n) + h(Y_r^k | X_m^n)$.

In the above lemma $H(X_m^n | Y_r^k)$ is defined as the average conditional entropy of X_m^n given Y_r^k, i.e., if $P_{j|h} = P[X_m^n = x_j | Y_r^k = y_h]$ then $H(X_m^n | Y_r^k) = \Sigma_{j,h}(-p_{j|h} \log p_{j|h}) p_h$. The average conditional differential entropy is similarly defined.

REMARK. The above lemma is true (with the same proof) if $k = \infty$. It follows that

$$I(X_{n+1}, X^n) = h(X_{n+1}) - h(X_{n+1} | X^n). \tag{2.3}$$

We denote the left-hand side of (2.3) by $I(X)$.

If X is a stationary process then we can identify $h(X_{n+1} | X^n)$ with $h(X)$ by the following lemma.

LEMMA 2.2. *Suppose X is stationary and that X_0 has density $p(x_0)$ such that $-p(x_0) \log p(x_0)$ is integrable. Then*

$$h(X) = h(X_{n+1} | X^n) = h(X_0 | X^{-1}).$$

PROOF. We can write

$$h(X_0^n) = h(X_0) + \Sigma_{k=1}^n h(X_k | X_0, \cdots, X_{k-1}), \qquad \text{by Lemma 2.1.}$$

It is proved in Pinsker [7, page 11] that $\lim_{n\to\infty} I(X_0, X_{-n}^{-1}) = I(X_0, X^{-1})$ hence by stationarity we conclude that $\lim_{k\to\infty} h(X_k | X_0, \cdots, X_{k-1}) = h(X_0 | X^{-1})$. It follows that $h(X) = h(X_0 | X^{-1})$. □

LEMMA 2.3. *Let $p_k(x)$ be a sequence of probability density functions on R such that $\lim_{k\to\infty} \int |p_k(x) - p(x)| dx = 0$, where $p(x)$ is also a probability density function. Assume that $p_k(x)$ are uniformly bounded* a.e. *and that $p(x) \log p(x)$ is integrable. Then*

$$\liminf_{k\to\infty} \int - p_k(x) \log p_k(x) dx \geqslant \int - p(x) \log p(x) dx. \tag{2.4}$$

PROOF. Let $c > 1$ be chosen so that $p_k(x) \leqslant c$ a.e. for all k. We then see that $-p_k(x) \log p_k(x) \geqslant -c \log c$ a.e., hence by Fatou's lemma we get that

$$\liminf_{k\to\infty} \int_{-b}^{b} - p_k(x) \log p_k(x) dx \geqslant \int_{-b}^{b} - p(x) \log p(x) dx$$

for all finite b. Furthermore for all $\varepsilon > 0$, there exists a positive b such that

$$\int_{-b}^{b} p_k(x) dx \geqslant 1 - \varepsilon$$

for all k, and such that

$$\int_{|x|>b} |p(x) \log p(x)| dx < \varepsilon.$$

We conclude that

$$\liminf_{k\to\infty} \int - p_k(x) \log p(x)\, dx$$
$$\geqslant \int - p(x) \log p(x)\, dx - \varepsilon(\log c + 1). \qquad \square$$

COROLLARY 2.1. *Let X_0 be a random variable with bounded density $p_0(x)$. Let Y_k be a sequence of independent random variables such that $Z_\infty = \Sigma_{j=1}^{\infty} Y_j$ converges* a.s. *and set $Z_k = \Sigma_{j=1}^{k} Y_j$. Assume that X_0 is independent of the sequence Y_j and that $X_0 + Z_\infty$ has density $p(x)$ such that $p(x) \log p(x)$ is integrable. Then* $\lim_{k\to\infty} h(X_0 + Z_k) = h(X_0 + Z_\infty)$.

PROOF. Let p_k be the density of $X_0 + Z_k$. By Lemma 4.1 we know that p_k tends to p in L^1 norm. Furthermore p_k is uniformly bounded by a simple application of Fubini's theorem. We conclude from Lemma 2.3 that

$$\liminf_{k\to\infty} h(X_0 + Z_k) \geqslant h(X_0 + Z_\infty). \tag{2.5}$$

The opposite inequality

$$h(X_0 + Z_\infty) \geqslant h(X_0 + Z_k).$$

(which is valid for all k), is well known and follows from Lemma 2.1. $\square$

We now present some background material on the subject of "rate distortion," which turned out to play an essential role in our work. In fact, it is surprising that the connection between rate distortion and nonlinear prediction has not been brought out before. We shall need only the one-dimensional version of this theory.

DEFINITION 2.6. Let $\rho(s)$ be any strictly increasing continuous mapping of $[0, \infty)$ onto itself. If X is a real valued random variable we define the rate distortion function $R_X(d)$ by setting

$$R_X(d) = \inf_{(X, Y)} I(X, Y)$$

where the inf is taken over all bivariate distributions (X, Y) with $E(\rho(|X - Y|) \leqslant d$ and $d \geqslant 0$.

We define the distortion rate function $D_X(r)$ by setting

$$D_X(r) = \inf_{(X, Y)} E(\rho(|X - Y|))$$

where the inf is taken over all bivariate distributions (X, Y) with $I(X, Y) \leqslant r$ and $r \geqslant 0$.

The functions R_X and D_X are decreasing wherever they are not zero and in fact they are inverse functions, i.e., $R_X(D_X(r)) = r$ and $D_X(R_X(d)) = d$. (The reader is referred to Berger [1] as a general reference on this area.)

If $\rho(s) = s^\alpha$ we shall write $D_X(r) = D_X^{(\alpha)}(r)$ and $R_X(d) = R_X^{(\alpha)}(d)$. Shannon has derived an interesting lower bound for $R_X(d)$ which specializes as follows:

$$R_X^{(2)}(d) \geqslant h(X) - \tfrac{1}{2}\log(2\pi e d) \tag{2.6}$$

$$R_X^{(1)}(d) \geqslant h(X) - \log(2ed). \tag{2.7}$$

(see Berger [1] for a derivation of these inequalities).

We can rewrite (2.6) as

$$e^{2R_X^{(2)}(d)} \geqslant \frac{1}{(2\pi ed)} e^{2h(X)},$$

whence we conclude that

$$D_X^{(2)}(r) \geqslant Q(X)e^{-2r}. \tag{2.8}$$

Similarly we get from (2.7) that

$$D_X^{(1)}(r) \geqslant (\pi Q(X)/2e)^{\frac{1}{2}} e^{-r}. \tag{2.9}$$

The use of these lower bounds in nonlinear prediction theory follows from the following simple lemma which is basic in our work.

LEMMA 2.4. *Let* $X = (X_k;\ k = 0, \pm 1, \cdots)$ *be a stationary real valued stochastic process. Let* f *be a Borel measurable function from* R^∞ *to* R. *We then have*

$$E(\rho(|X_{n+1} - f(X^n)|)) \geqslant D_{X_0}(I(X))$$

where $I(X) = I(X_0, X^{-1}) = I(X_{n+1}, X^n)$.

PROOF. Note that $I(X_{n+1}, f(X^n)) \leqslant I(X_{n+1}, X^n)$ by Pinsker [7, page 11] and also that $E(\rho(|X_{n+1} - Z|)) \geqslant D_{X_{n+1}}(r)$ for any (X_{n+1}, Z) with $I(X_{n+1}, Z) \leqslant r$. □

3. Main results.

LEMMA 3.1. *Let* $(a_j;\ j = 0, \pm 1, \cdots)$ *be a sequence of constants with* $a_j = 0$ *for* $|j| > k$, *where* k *is a fixed positive integer. (Assume that not all the* a_j *are* 0). *Let* Y_j *be a sequence of independent identically distributed random variables with* $-\infty < h(Y_j) < \infty$. *Let* $X_n = \Sigma_j a_{n-j} Y_j$ *and let* Δ^2 *be defined as in the introduction. Assume also that* $-\infty < h(X_1^m) < \infty$ *for all* n.

We then have $Q(X) \geqslant \Delta^2 Q(Y_0)$.

PROOF. For any $m > 2k$ let G_m stand for the group of integers with addition mod $2m$ and let $\{-m + 1, \cdots, -1, 0, 1, 2, \cdots, m\}$ be a list of the elements of G_m.

Let $\tilde{X}_n = \Sigma_{j \in G_m} a_{n \ominus j} Y_j$ where $n \in G_m$, $n \ominus j$ stands for subtraction mod $2m$, and Y_j are as before. ($\tilde{X}_n$ is a "circulant" process, i.e., $\tilde{X}_n$ is a stationary process on the group G_m.) We let $\hat{Y}_j$ be a sequence of independent mean zero Gaussian random variables with $h(\hat{Y}_j) = h(Y_j)$ and we define $\hat{X}_n = \Sigma_{j \in G_m} a_{n \ominus j} \hat{Y}_j$ for $n \in G_m$.

It is clear that

$$h(\hat{X}_{-m+1}^m) = h(\tilde{X}_{-m+1}^m) \tag{3.1}$$

because both sides of (3.1) are equal to $2mh(Y_0) + \log(\det(\hat{A}_m))$ where $\hat{A}_m$ is the matrix whose (i, j)th entry is $a_{i \ominus j}$. We now note that

$$h(X_1^m) = h(\tilde{X}_1^m)$$

for $m > 2k$, and that

$$I(\tilde{X}_1^m, \tilde{X}_{-m+1}^0) = h(\tilde{X}_1^m) + h(\tilde{X}_{-m+1}^0) - h(\tilde{X}_{-m+1}^m)$$
$$I(\hat{X}_1^m, \hat{X}_{-m+1}^0) = h(\hat{X}_1^m) + h(\hat{X}_{-m+1}^0) - h(\hat{X}_{-m+1}^m)$$

by Lemma 2.1.

We now write

$$\begin{aligned} h(\tilde{X}_1^m) &= h(\hat{X}_1^m) + h(\tilde{X}_1^m) - h(\hat{X}_1^m) \\ &= h(\hat{X}_1^m) + \tfrac{1}{2}\big(I(\tilde{X}_1^m, \tilde{X}_{-m+1}^0) - I(\hat{X}_1^m, \hat{X}_{-m+1}^0)\big) \\ &\geqslant h(\hat{X}_1^m) - \tfrac{1}{2}I(\hat{X}_1^m, \hat{X}_{-m+1}^0) \\ &= \tfrac{1}{2}h(\hat{X}_{-m+1}^m). \end{aligned}$$

We conclude that

$$h(X_1^m) \geqslant \tfrac{1}{2}h(\hat{X}_{-m+1}^m). \tag{3.2}$$

We now let C_m be the matrix with $C_m(i,j) = \Sigma_r a_{i-r}a_{j-r}$ for $1 \leqslant i, j \leqslant 2m$. Let $\hat{C}_m$ be the matrix with $\hat{C}_m(i,j) = \Sigma_{r\in G_m} a_{i\ominus r}a_{j\ominus r}$ for $1 \leqslant i, j \leqslant 2m$, i.e., $\hat{C}_m$ is the circulant approximation to C_m as defined in Gray [4, page 728] (remembering again that $m > 2k$). We now note that the eigenvalues of $\hat{C}_m$ are simply $\{\Phi(\pi j/m); j = 1, 2m\}$ for $m > 2k$ (by [4], page 728), and that

$$\lim_{m\to\infty}(2m)^{-1}\Sigma_{j=1}^{2m}\log \Phi(\pi j/m) = (2\pi)^{-1}\textstyle\int_{-\pi}^{\pi}\log \Phi(s)ds.$$

We use (2.2) to write

$$\begin{aligned} (2m)^{-1}h(\hat{X}_{-m+1}^m) &= \left(\tfrac{1}{2}\right)(2m)^{-1}\log\left((2\pi e Q(Y_0))^{2m}\det(\hat{C}_m)\right) \\ &= h(Y_0) + \tfrac{1}{2}(2m)^{-1}\Sigma_{j=1}^{2m}\log \Phi(\pi j/m). \end{aligned}$$

We conclude that

$$\lim_{m\to\infty}(2m)^{-1}h(\hat{X}_{-m+1}^m) = h(Y_0) + \tfrac{1}{2}\log \Delta^2.$$

The proof of the lemma is completed by applying (3.2). □

THEOREM 3.1. *Let a_j be a sequence of constants with $0 < \Sigma_j a_j^2 < \infty$. Let Y_j be a sequence of independent random variables with a common bounded density and $\sigma^2 = Var(Y_j) < \infty$. Defining X_n as before we conclude that*

$$Q(X) \geqslant \Delta^2 Q(Y_0).$$

PROOF. For any $k > 0$ let $X_n(k) = \Sigma_{j=-k}^{k} a_j Y_{n-j}$ and let $\Delta_k^2 = \exp((1/2\pi))\int_{-\pi}^{\pi}\log \Phi_k(s)ds)$ where $\Phi_k(s) = |\Sigma_{j=-k}^{k} a_j e^{ijs}|^2$. We claim that

$$\Delta^2 \leqslant \liminf_{k\to\infty}\Delta_k^2. \tag{3.3}$$

To see this note that

$$\Delta_k^2\sigma^2 = \inf_V E\left((X_{n+1}(k) - V)^2\right)$$

where the inf is taken over the set of all $V = b + \Sigma_{j=0}^{r} b_j X_{n-j}(k)$. Clearly

$$\lim_{k\to\infty} E\big((X_{n+1}(k) - b - \Sigma_{j=0}^{r} b_j X_{n-j}(k))^2\big) = E\big((X_{n+1} - b - \Sigma_{j=0}^{r} b_j X_{n-j})^2\big)$$

and (3.3) follows. We now use Corollary 2.1 to get

$$\lim_{k\to\infty} h(X_0(k)) = h(X_0). \tag{3.4}$$

Furthermore we have

$$I(X) \leqslant \liminf_{k\to\infty} I(X(k)) \tag{3.5}$$

by [7, page 20]. We note that $I(X) = h(X_0) - h(X)$ and $I(X(k)) = h(X_0(k)) - h(X(k))$ by Lemma 2.2, so we get

$$h(X) \geqslant \limsup_{k\to\infty} h(X(k)). \tag{3.6}$$

We can rewrite (3.6) as

$$Q(X) \geqslant \limsup_{k\to\infty} \Delta_k^2 Q(Y_0)$$

by using Lemma 3.1. We now apply (3.3) to finish the proof of the theorem. ▯

COROLLARY 3.1. *Let Y_j and X_n be as in Theorem 3.1. Let $f: R^\infty \to R$ be Borel measurable. Then for any n we have $E((X_{n+1} - f(X^n))^2) \geqslant Q(Y_0)\Delta^2$.*

PROOF. We have $E((X_{n+1} - f(X^n))^2) \geqslant D_{X_0}^{(2)}(I(X))$ by Lemma 2.4. Remember now the relations $I(X) = h(X_0) - h(X)$, $Q(X_0) = (2\pi e)^{-1} e^{2h(X_0)}$, and $Q(X) = (2\pi e)^{-1} e^{2h(X)}$; apply the Shannon lower bound (2.7) and get $D_{X_0}^{(2)}(I(X)) \geqslant Q(X)$. Use Theorem 3.1 to complete the proof. ▯

We now extend Corollary 3.1 to processes with infinite variance.

THEOREM 3.2. *Let $(Y_j; j = 0, \pm 1, \cdots)$ be a sequence of random variables such that for some double array of independent random variables $(Y_{kj}; k \geqslant 0, j = 0, \pm 1, \cdots)$ we have $Y_j = \Sigma_{k=0}^{\infty} Y_{kj}$ a.s. Assume that for each fixed k the sequence $(Y_{kj}; j = 0, \pm 1, \cdots)$ is identically distributed with finite variance σ_k^2. Assume also that the common distribution of $(Y_{0j}; j = 0, \pm 1, \cdots)$ has bounded density. Let $(a_j; j = 0, \pm 1, \cdots)$ with $X_n = \Sigma_j \Sigma_k a_j Y_{k, n-j}$ a.s. convergent (and such that the sum does not depend on the order of summation).*

Then for any Borel measurable function f from R^∞ to R we have

$$E\big((X_{n+1} - f(X^n))^2\big) \geqslant Q(Y_0)\Delta^2 \tag{3.7}$$

for all n.

PROOF. Let $X_n(k) = \Sigma_{r=0}^{k} \Sigma_j a_j Y_{r, n-j}$. We see that

$$E\big((X_{n+1}(k)) - f(X^n(k))^2\big) \geqslant Q(Y_0(k))\Delta^2 \tag{3.8}$$

by Corollary 3.1, where $Y_j(k) = \Sigma_{r=0}^{k} Y_{rj}$. Using Corollary 2.1 we see that $\lim_{k\to\infty} Q(Y_0(k)) = Q(Y_0)$. Now let $Z_n(k) = \Sigma_{r=k+1}^{\infty} \Sigma_j a_j X_{r, n-j}$. We have

$$E\big((X_{n+1} - f(X^n))^2\big) = E\big((X_{n+1}(k) + Z_{n+1}(k) - f(X^n(k) + Z^n(k))^2\big) \tag{3.9}$$

where $X^n(k) = (\cdots, X_{n-1}(k), X_n(k))$ and $Z^n(k) = (\cdots, Z_{n-1}(k), Z_n(k))$. By Corollary 3.1 we know that for any sequence $(z_j, j = 0, \pm 1, \cdots)$ of real numbers we have

$$E\big((X_{n+1}(k) + z_{n+1} - f((X^n(k) + z^n)))^2\big) \geqslant Q(Y_0(k))\Delta^2$$

where $z^n = (\cdots, z_{n-1}, z_n)$. If we now apply Fubini's theorem to (3.9) we get

$$E\big((X_{n+1} - f(X^n))^2\big) \geqslant Q(Y_0(k))\Delta^2$$

valid for any k. Letting $k \to \infty$, the theorem follows. □

As an example of the applicability of Theorem 3.2 we can show that if Y_j have the distribution of an infinitely divisible random variable with bounded density then Y_j can be expressed by a sum $\Sigma_{r=0}^{\infty} Y_{rj}$ as in Theorem 3.2. Indeed let μ be the common distribution of Y_j. Since μ is infinitely divisible we have

$$\int_{-\infty}^{+\infty} e^{isx} d\mu(x) = \exp\left(\int_{-\infty}^{\infty}\left(e^{ist} - 1 - \frac{ist}{1+t^2}\right)\frac{1+t^2}{t^2}\right)d\nu(t) \tag{3.10}$$

where ν is the Lévy-Khintchine measure for μ, and where we are assuming for simplicity that the centering constant in (3.10) is 0. Let ν_k stand for the measure ν cut down to the set $\{k + 1 > |x| \geqslant k\}$. We can write $\nu = \Sigma_{k=0}^{\infty} \nu_k$. Let μ_k stand for the infinitely divisible distribution with Lévy-Khintchine measure ν_k. We can write $\mu = *_{k=0}^{\infty} \mu_k$. Noting that the measure $*_{k=1}^{\infty} \mu_k$ has nonzero mass at the origin, we conclude that μ_0 has bounded density if μ does. Finally it is easy to see that all the μ_k have finite variance.

We now turn our attention towards getting a lower bound for mean absolute error of one step prediction.

THEOREM 3.3. *Let Y_{kj}, Y_j, X_n, and $X_n(k)$ be as in Theorem 3.2. Then for any Borel function $f : R^{\infty} \to R$ we have*

$$E(|X_{n+1} - f(X^n)|) \geqslant \Delta\big((2e)^{-1}\pi Q(Y_0)\big)^{\frac{1}{2}}. \tag{3.11}$$

PROOF. We have $E(|X_{n+1} - f(X^n)|) \geqslant D_{X_0}^{(1)}(I(X))$ by Lemma 2.4. Now use the Shannon lower bound (2.8) and argue as in Corollary 3.1 to get that

$$D_{X_0(k)}^{(1)}(I(X)) \geqslant \Delta\big((2e)^{-1}\pi Q(Y_0(k))\big)^{\frac{1}{2}}$$

for any k, and conclude that

$$E(|X_{n+1}(k) - f(X^n(k)|) \geqslant \Delta\big((2e)^{-1}\pi Q(Y_0(k))\big)^{\frac{1}{2}}.$$

Conclude the proof by arguing as in Theorem 3.2. □

Using Shannon lower bounds for the distortion rate functions $D_{X_0}^{\alpha}(r)$, $0 < \alpha < \infty$ (see Linkov [5] and Pinkston [6]) we get results analogous to Theorem 3.3, establishing lower bounds for $E(|X_{n+1} - f(X^n)|^{\alpha})$. We do not go into further details but turn our attention to showing (3.11) is sharp.

To see that (3.11) is sharp we produce a class of examples for which (3.11) is an equality. Let Y_j be independent random variables with common density $p(x) = \frac{1}{2}e^{-|x|}$. Let $b_0, \cdots, b_r$ be constants with $b_0 = 1$ and $\Sigma_0^r b_j z^j$ having all its complex roots outside the unit circle. Let X_n be the unique stationary process which satisfies the autoregressive equation

$$\Sigma_{j=0}^r b_j X_{n-j} = Y_n$$

for all integers n. (Assume $b_r \neq 0$.) It is clear that $E(X_{n+1}|X^n) = \Sigma_{j=1}^r - b_j X_{n-j}$; hence $E(|X_{n+1} - f(X^n)|) \geqslant E(|X_{n+1} - (\Sigma_{j=1}^r - b_j X_{n-j})|) = E(|Y_{n+1}|)$ for any Borel function f. We now note that $E(|Y_{n+1}|) = 1$, $\Delta = 1$, and $Q(Y_0) = 2e/\pi$; hence equality is achieved in (3.11).

We end this section with a remark on the conjecture that $\Delta = 0$ implies the process $X_n = \Sigma_j a_j Y_{n-j}$ is perfectly nonlinearly predictable when $\mathrm{Var}(Y_j) = \infty$. (If $\mathrm{var}(Y_j) < \infty$ then X_n are perfectly linearly predictable.)

We shall show, by generalizing a result of Pinsker, that $H(X) > 0$; hence the above conjecture, if true, has a rather subtle proof. Recalling Definition 2.2 we see that $H(X) = 0$ if for all $f \in \hat{\mathcal{Q}}$ we have $H(X') = 0$ where $X'_n = f(\tau_n \circ X)$. Note now that for any $f \in \hat{\mathcal{Q}}$ there exists $g \in \hat{\mathcal{Q}}$ such that $f(\tau_n \circ X) = g(\tau_n \circ Y)$ a.s. We shall prove the following theorem (due to Pinsker if Y_j assume only finitely or countably many values).

THEOREM 3.4. *Let $(Y_j; j = 0, \pm 1, \cdots)$ be any stationary real valued process with trivial remote past. Let $g \in \hat{\mathcal{Q}}$ and $X'_n = g(\tau_n \circ Y)$. Then $H(X') = 0$ implies X'_n are constant.*

PROOF. Let h be any piecewise constant function from R to R which assumes only finitely many rational values and has discontinuities at only finitely many rational points. Let $Y'_n = h(Y_n)$. If $H(X') = 0$ then $\vec{I}(X', Y') = 0$, where $\vec{I}$ is defined in [7], page 76. Also $\vec{I}(X', Y') = \vec{I}(Y', X')$ by [7], page 80. Finally Y' has trivial remote past so $\vec{I}(Y', X') = 0$ implies X' is independent of the σ-field $\mathfrak{B}_h$ generated by Y'. We conclude $H(X') = 0$ implies X' is independent of $\bigvee_h \mathfrak{B}_h$, hence X'_n is constant a.s. □

4. Appendix. In this section we present the measure theoretic result which was used in the proof of Corollary 2.1. In the following we shall write $\mu_n \to_d \mu$ to stand for weak convergence of measures as defined in Feller [3, page 248]. We shall write $\|\mu_n - \mu\|$ to stand for the total variation norm of the signed measure $\mu_n - \mu$. If μ_n and μ have density f_n and f we shall use the fact that $\|\mu_n - \mu\| = \|f_n - f\|_1$, where the latter expression stands for the L^1 norm of the difference $f_n - f$.

LEMMA 4.1. *Let μ_n be a sequence of probability measures on R such that $\mu_n \to_d \mu$ where μ is also a probability measure. Then for any probability measure η with density, we have $\|\mu_n * \eta - \mu * \eta\| \to 0$.*

PROOF. For any $\varepsilon > 0$ it is well known that there exists a compact interval $[a_\varepsilon, b_\varepsilon]$ such that $\mu_n([a_\varepsilon, b_\varepsilon]) \geqslant 1 - \varepsilon$ for all n. This fact shows us that we can assume without loss of generality that the measures μ_n and μ all have support in a fixed compact set $[a, b]$. Letting $d\eta = fdx$ we know that f can be approximated in the L^1 norm by continuous functions with compact support so we can also assume without loss of generality that f has support in $[a, b]$. It follows by Theorem 1 in [3, page 255] that $(\mu_n * f)(x) \to (\mu * f)(x)$ uniformly on R. Since all these functions have support in a fixed compact set it then follows that $\| \mu_n * f - \mu * f \|_1 \to 0$. □

REFERENCES

[1] Berger, T. (1971), *Rate Distortion Theory*. Prentice-Hall, New Jersey.

[2] Billingsley, P. (1965). *Ergodic Theory and Information*. Wiley, New York.

[3] Feller, W. (1966). *An Introduction to Probability Theory and its Applications*, **2**, 2nd ed. Wiley, New York.

[4] Gray, R. M. (1972). On the asymptotic eigenvalue distribution of Toeplitz matrices. *IEEE Trans. Information Theory* **IT-18** 725–730.

[5] Linkov, Y. N. (1965). Evaluation of ε-entropy of random variables for small ε. *Problems of Information Transmission* **1** 12–18.

[6] Pinkston, J. T. (1966). Information rates of independent sample sources. M.S. Thesis, Dept. of Elec. Engrg., M.I.T.

[7] Pinsker, M. S. (1960). *Information and Information Stability of Random Variables and Processes*. Izol. Akad. Nauk. SSR, Moscow. (Transl. Holden-Day, San Francisco (1964).)

[8] Whittle, P. (1963). *Prediction and Regulation*. English Universities Press, Suffolk.

Department of Mathematics
Concordia University
1455 de Maisonneuve Blvd. West
Montreal, Quebec
Canada H3G1M8