The Annals of Mathematical Statistics
1971, Vol. 42, No. 1, 75-80

AN ALGORITHM FOR COMPUTING THE NON-LINEAR PREDICTOR[1]

BY JAMES B. ROBERTSON

University of California, Santa Barbara

1. Introduction. Let $\{f_n, -\infty < n < \infty\}$ be a (strictly) stationary stochastic process and let v be a positive integer. Subject to certain restrictions on the stochastic process Masani and Wiener [6] have given an algorithm for calculating the non-linear predictor, $\hat{f}_v = E(f_v | \cdots, f_{-1}, f_0)$, of f_v given the past and present of the stochastic process. In [8] we showed that in theory $\hat{f}_v$ can be determined if $E(|f_v|) < \infty$. The main purpose of this paper is to give an algorithm for computing $\hat{f}_v$ which is perhaps more naive than the one given by Masani and Wiener but which is valid for the stochastic processes considered in [8].

In Section 2 we describe the algorithm in an informal way. In Section 3 we give precise definitions and prove that the algorithm converges to the non-linear predictor whenever the stochastic process is ergodic. In Section 4 we show that the algorithm is valid for non-ergodic stochastic processes. In Section 5 we extend our results to non-linear prediction in L_p, $1 < p < \infty$. In Section 6 we make some additional remarks. In particular we indicate that our methods apply to *multivariate*, non-linear prediction theory (cf. (6.3)).

The author wishes to express his appreciation to the referee, D. L. Hanson, and P. Masani for their many helpful comments on the original manuscript.

2. Description of the algorithm. The purpose of this section is to give a non-technical description of the algorithm, and the remainder of the paper is independent of this section. We suppose that we are given an infinite sequence of numbers $(\cdots, f_{-1}(\omega), f_0(\omega))$ which is the realization of the past and present of a stationary stochastic process $\{f_n, -\infty < n < \infty\}$. Also we are given a function g of infinitely many variables, $g = g(\cdots, f_{-1}, f_0)$, and a positive integer v. We interpret $g_v(\omega) = g(\cdots, f_{v-1}(\omega), f_v(\omega))$ as some numerical attribute of the stochastic process at time v. Our problem is to calculate the least squares predictor $\hat{g}_v(\omega)$ of $g_v(\omega)$, i.e. $\hat{g}_v$ is the function of the form $\hat{g}_v = h(\cdots, f_{-1}, f_0)$ which minimizes the expression $E\{(g_v - h)^2\}$. It is well known (cf. e.g. [6] Section 6) that $\hat{g}_v = E(g_v | \cdots, f_{-1}, f_0)$. The problem treated by Masani and Wiener is obtained on setting $g_v = f_v$.

We proceed by making three successive approximations. First by taking n sufficiently large, we approximate $E(g_v | \cdots, f_{-1}, f_0)$ by $E(g_v | f_{-n+1}, \cdots, f_0)$. Next the σ-field determined by $f_{-n+1}, \cdots, f_0$ is approximated by a discrete σ-field $\mathscr{F}_n$. For example if m is a sufficiently large fixed integer the atoms of $\mathscr{F}_n$ may be taken to be sets of the form $F = \bigcap_{i=0}^{n-1} \{\omega : l_i/2^m < f_{-i}(\omega) \leqq (l_i+1)/2^m\}$ where l_i is any integer. Then $E(g_v | \mathscr{F}_n)$ approximates $E(g_v | f_{-n+1}, \cdots, f_0)$. Suppose $l_i/2^m < f_{-i}(\omega) \leqq (l_i+1)/2^m$, $i = 0, \cdots, n-1$. Then there will exist infinitely many k_j,

Received October 7, 1969.

[1] Research partially supported by NSF Grant GP-8042.

$v \leqq k_1 < k_2 < \cdots$ such that $l_i/2^m < f_{-k_j-i}(\omega) \leqq (l_i+1)/2^m$, $i = 0, \cdots, n-1$. The average of the $g(\cdots, f_{-k_j+v}(\omega))$, i.e. $\lim_{k\to 0} k^{-1}\sum_{j=1}^{k} g(\cdots, f_{-k_j+v}(\omega))$, will then approximate $E(g_v \mid \mathscr{F}_n)$ and hence $\hat{g}_v$.

3. The algorithm. Let $\{f_n, -\infty < n < \infty\}$ be a stationary stochastic process on a probability space $(\Omega, \mathscr{F}, P)$. Let R denote the real line with the usual topology and Borel field. Let $R^\infty = \cdots \times R \times R$ denote the countable product of R indexed by the non-positive integers. The Borel field, $\mathscr{B}$, (topology) on R^∞ is the smallest σ-field (topology) such that the coordinate functions, $X_k((x_i)_{-\infty}^{0}) = x_k$, are measurable (continuous) for $k \leqq 0$. Let g be a real-valued measurable function on R^∞ such that the random variables $g_n = g(\cdots, f_{n-1}, f_n)$ are integrable, i.e. $g_n \in L_1(\Omega, \mathscr{F}, P)$. Let v be a positive integer. The problem of non-linear prediction theory is to calculate $\hat{g}_v = E(g_v \mid \cdots, f_{-1}, f_0)$.

Now let $X = \cdots \times R \times R \times \cdots$ denote the countable product of R indexed by the integers. The Borel field, $\mathscr{B}$, (topology) on X is the smallest σ-field (topology) such that the coordinate functions, $X_k((x_i)_{-\infty}^{\infty}) = x_k$, are measurable (continuous) for all k. Let $\varphi: \Omega \to X$ be defined by $\varphi(\omega) = (f_i(\omega))_{-\infty}^{\infty}$. Then φ is measurable, and we shall denote by P the probability induced on $\mathscr{B}$ by φ, i.e. for all $A \in \mathscr{B}$, $P(A) = P(\varphi^{-1}(A))$. Let g_n' denote the random variable $g(\cdots, X_{n-1}, X_n)$ on $(X, \mathscr{B}, P)$. Then $g_n' \circ \varphi = g_n$, and $E(g_n \mid \cdots, f_{-1}, f_0) = E(g_n' \mid \cdots, X_{-1}, X_0) \circ \varphi$. Thus it is sufficient to consider the stationary process $\{X_n, -\infty < n < \infty\}$ on $(X, \mathscr{B}, P)$. On $(X, \mathscr{B}, P)$ the shift transformation, defined by $(T((x_i)_{-\infty}^{\infty}))_k = x_{k+1}$, is an invertible measure preserving transformation such that $X_k \circ T = X_{k+1}$ for all k. In this section we suppose that T is ergodic.

Now for each positive integer m let α_m be a countable measurable partition of R such that

(3.1) α_{m+1} is a refinement of α_m, i.e. every set in the partition α_{m+1} is a subset of some set in α_m, and

(3.2) $\bigvee_{m=1}^{\infty} \alpha_m = \mathscr{B}$, i.e. the smallest σ-field containing all the sets in α_m for all m is the Borel field of R. For each positive integer n let $\tilde{\alpha}_{m,n}$ be the countable, measurable partition of $(X, \mathscr{B}, P)$ consisting of all sets F of the form

$$F = \bigcap_{i=0}^{n-1} X_{-i}^{-1}(A_i) \quad \text{where } A_i \in \alpha_m. \tag{3.3}$$

Denote the σ-field generated by $\tilde{\alpha}_{m,n}$ by $\mathscr{F}_{m,n}$.

The recurrence theorem (cf. e.g. Halmos [5] page 10) states that for almost every point ω in a measurable set F, $\omega \in T^k(F)$ for infinitely many k. Thus for almost all $\omega \in F$, where F is as in (3.3), we may define inductively

$$n_1(\omega) = \min\{k : k \geqq v \text{ and } \omega \in F \cap T^k F\} \tag{3.4a}$$

and

$$n_i(\omega) = \min\{k : k > n_{i-1}(\omega) \text{ and } \omega \in F \cap T^k F\}, \quad i = 2, 3, \cdots. \tag{3.4b}$$

Finally for almost all ω we may define

(3.5) $$h_{i,m,n}(\omega) = g_{-n_i(\omega)+\nu}(\omega) = g(\cdots, X_{-n_i(\omega)+\nu-1}(\omega), X_{-n_i(\omega)+\nu}(\omega)).$$

THEOREM 3.6. $\lim_{n\to\infty}(\lim_{m\to\infty}(\lim_{k\to\infty}(1/k)\sum_{i=1}^{k} h_{i,m,n}(\omega))) = \hat{g}_\nu(\omega)$ *for almost all* ω.

PROOF. First fix m and n and consider a set F of the form (3.3). Let

(3.7) $$G_k = \bigcap_{l=\nu}^{k-1} T^l F' \cap T^k F \qquad \text{for } k = \nu, \nu+1, \cdots \text{and}$$

$$D_k = \bigcap_{l=1}^{k-1} T^l F' \cap T^k F \qquad \text{for } k = 1, 2, \cdots.$$

Since G_ν, $G_{\nu+1}, \cdots$ are disjoint measurable subsets of F and since almost all $\omega \in F$ are in $\bigcup_{k=\nu}^{\infty} G_k$, we may define for almost all $\omega \in F$

(3.8) $$Q(\omega) = T^{-k}(\omega) \qquad \text{for } \omega \in G_k, k = \nu, \nu+1, \cdots.$$

Clearly Q is defined for almost all $\omega \in F$ and the range of Q is contained in F. We shall now show that Q is nonsingular. Let A be a subset of F. Then $Q(A) \supseteq Q(\bigcup_{k=\nu}^{\infty}(A \cap G_k)) = \bigcup_{k=\nu}^{\infty} Q(A \cap G_k) = \bigcup_{k=\nu}^{\infty} T^{-k}(A \cap G_k)$. Hence if $Q(A)$ is contained in a set of probability zero, then $P^*(A \cap G_k) = P^*(T^{-k}(A \cap G_k)) = 0$ for all k ($P^*(A)$ is the outer measure of the set A). Thus $P^*(A) \leqq \sum_{k=\nu}^{\infty} P^*(A \cap G_k) = 0$.

Next for almost all $\omega \in F$ we define

(3.9) $$S(\omega) = T^{-k}(\omega) \qquad \text{for } \omega \in D_k, k = 1, 2, \cdots.$$

Moy (cf. [6] Lemma 1 and Corollary 1 applied to T^{-1} instead of T) has shown that S is an ergodic, invertible measure preserving transformation on the probability space $(F, F \cap \mathscr{B}, P(\cdot \mid F))$. Next we shall show that

(3.10) $$T^{-n_i(\omega)}(\omega) = S^{i-1} \circ Q(\omega) \qquad \text{for almost all } \omega \in F.$$

From (3.4a) and (3.7) we see that $n_1(\omega) = k$ if and only if $\omega \in G_k$. Thus using (3.8) we obtain for $\omega \in G_k$, $T^{-n_1(\omega)}(\omega) = T^{-k}(\omega) = Q(\omega)$. Proceeding by induction, consider ω such that $n_{i-1}(\omega) = k-1$ and $n_i(\omega) = l$. By induction $T^{-n_{i-1}(\omega)}(\omega) = S^{i-2} \circ Q(\omega)$ a.e. Using (3.4b) we see that $T^{-n_{i-1}(\omega)}(\omega) \in T^{-k+1}(\bigcap_{j=k}^{l-1} T^j F' \cap T^l F) = \bigcap_{j=1}^{l-k} T^j F' \cap T^{l-k+1} F = D_{l-k+1}$. Thus using (3.9), we obtain $S^{i-1} \circ Q(\omega) = S(S^{i-2} \circ Q(\omega)) = T^{-l+k-1}(T^{-k+1}(\omega)) = T^{-l}(\omega) = T^{-n_i(\omega)}(\omega)$ which proves (3.10).

Therefore we obtain

$$(1/k)\sum_{i=1}^{k} h_{i,m,n}(\omega) = (1/k)\sum_{i=1}^{k} g_\nu \circ T^{-n_i(\omega)}(\omega)$$
$$= (1/k)\sum_{i=1}^{k} g_\nu \circ S^{i-1}(Q(\omega)).$$

Applying the ergodic theorem (cf. e.g. Halmos [5] page 18) to S on $(F, F \cap \mathscr{B}, P(\cdot \mid F))$ we see that $\lim_{k\to\infty}(1/k)\sum_{i=1}^{k} g_\nu \circ S^{i-1}(\omega') = E(g_\nu \mid F)$ for almost all $\omega' \in F$. Since Q is nonsingular, $\lim_{k\to\infty}(1/k)\sum_{i=1}^{k} h_{i,m,n}(\omega) = E(g_\nu \mid F)$ for almost all $\omega \in F$. Hence, since F was an arbitrary atom of $\mathscr{F}_{m,n}$, $\lim_{k\to\infty}(1/k)\sum_{i=1}^{k} h_{i,m,n}(\omega) = E(g_\nu \mid \mathscr{F}_{m,n})(\omega)$ a.e. Using (3.1) and (3.2) and a martingale convergence theorem (cf. Doob [3] page 331) we see that $\lim_{m\to\infty} E(g_\nu \mid \mathscr{F}_{m,n}) = E(g_\nu \mid X_{-n+1}, \cdots, X_0)$ a.e.

and $\lim_{n\to\infty} E(g_\nu \mid X_{-n+1}, \cdots, X_0) = E(g_\nu \mid \cdots, X_{-1}, X_0) = \hat{g}_\nu$ a.e. This completes the proof of the theorem.

4. Ergodicity. In this section we shall relax the assumption of Section 3 that T be ergodic. The function $h_{i,m,n}$, cf. (3.5), is defined on a certain measurable set and is measurable. Thus

$$h_\nu(\omega) = \lim_{n\to\infty} (\lim_{m\to\infty} (\lim_{i\to\infty} (1/k) \textstyle\sum_{i=1}^{k} h_{i,m,n}(\omega))) \tag{4.1}$$

is defined on some measurable set and is measurable with respect to $\cdots, X_{-1}, X_0$. We may suppose that h_ν is a measurable function on $(X, \mathscr{B})$ by extending it to be zero wherever (4.1) is not defined. If P is a probability measure on $(X, \mathscr{B})$, we shall denote expectations and conditional expectations with respect to P by E_P. Then Theorem 3.6 states that if P is ergodic, then $h_\nu = E_P(g_\nu \mid \cdots, X_{-1}, X_0)$ a.e.

Now let P be an invariant probability measure on $(X, \mathscr{B})$. Let $\mathscr{E}$ be the class of all ergodic, invariant probability measures on $(X, \mathscr{B})$. The smallest σ-field on $\mathscr{E}$ such that the function $\pi \to \pi(A)$ is measurable for all $A \in \mathscr{B}$ will be denoted by $\mathscr{A}$. Farrell has established (cf. [4] Theorem 5, page 461) that under our condition, i.e. $\{T_n, -\infty < n < \infty\}$ is a countable group of measure preserving transformations and X is a complete separable metric space, there exists a probability measure, λ_P, on $(\mathscr{E}, \mathscr{A})$ such that for all $A \in \mathscr{B}$, $P(A) = \int_{\mathscr{E}} \pi(A)\, d\lambda_p(\pi)$. It is then easy to establish, using linearity and the monotone convergence theorem, that if f is a nonnegative measurable function on $(X, \mathscr{B})$, then $\int_X f(x)\, d\pi(x)$ is a nonnegative measurable function of π on $(\mathscr{E}, \mathscr{A})$ and that

$$\int_X f(x)\, dP(x) = \int_{\mathscr{E}} \{\int_X f(x)\, d\pi(x)\}\, d\lambda_P(\pi). \tag{4.2}$$

Now if $f \in L_1(X, \mathscr{B}, P)$, it can be seen by standard arguments that $\int_X f(x)\, d\pi(x)$ is defined for almost all π with respect to λ_P, that as a function of π it is in $L_1(\mathscr{E}, \mathscr{A}, \lambda_P)$, and that (4.2) holds. Thus if $g_\nu \in L_1(X, \mathscr{B}, P)$ and if A is measurable with respect to $\cdots, X_{-1}, X_0$, then

$$\begin{aligned}\int_A g_\nu(x)\, dP(x) &= \int_{\mathscr{E}} (\int_A g_\nu(x)\, d\pi(x))\, d\lambda_P(\pi) \\ &= \int_{\mathscr{E}} (\int_A h_\nu(x)\, d\pi(x))\, d\lambda_P(\pi) \\ &= \int_A h_\nu(x)\, dP(x).\end{aligned}$$

This together with the fact that h_ν is measurable with respect to $\cdots, X_{-1}, X_0$ implies that $h_\nu = E_P(g_\nu \mid \cdots, X_{-1}, X_0)$ a.e. Thus the algorithm is valid for arbitrary stationary stochastic processes and integrable g_ν.

5. Extension to $L_p(1 < p < \infty)$. If $g_\nu \in L_2(\Omega, \mathscr{F}, P)$, then the non-linear predictor $\hat{g}_\nu$ given in Section 3 is that function h which minimizes $\|g_\nu - h\|_2$ among all functions which are measurable with respect to $\cdots, f_{-1}, f_0$. In this section we shall show that our method can be extended so as to minimize the $L_p(1 < p < \infty)$ error. We first prove an ergodic theorem (Lemma 5.4) for certain generalized means studied by Brøns, Brunk, Franck, and Hanson [2].

Let φ be a real-valued function on $R \times R$ such that:

(5.1) $\varphi(x, \theta)$ is strictly increasing in θ for each $x \in R$,

(5.2) $\varphi(x, \theta) \geqq 0$ for $\theta > x$ and $\varphi(x, \theta) \leqq 0$ for $\theta < x$,

(5.3) $\varphi(x, \theta)$ is a Borel measurable function of x for each fixed θ in R.

Notice that (5.1) is stronger than [2, (2.1)]. For a random variable X on $(\Omega, \mathscr{F}, P)$ let $\lambda(\theta) = E(\varphi(X, \theta))$ (assuming that $\varphi(X, \theta)$ is integrable for all θ). It follows (use [2, (2.8) and (2.11)] and the monotone convergence theorem) that there exists exactly one real number μ such that for $\theta > \mu$, $\lambda(\theta) > 0$ and for $\theta < \mu$, $\lambda(\theta) < 0$. This μ is called the *φ-mean* of X.

If $\{x_1, \cdots, x_n\}$ are real numbers, their *sample φ-mean* is that μ such that $\lambda(\theta) \leqq 0$ for $\theta < \mu$ and $\lambda(\theta) \geqq 0$ for $\theta > \mu$ where $\lambda(\theta) = \sum_{i=1}^{n} \varphi(x_i, \theta)$.

LEMMA 5.4. *Let T be an ergodic measure preserving transformation on $(\Omega, \mathscr{F}, P)$. Let X be a random variable such that $\varphi(X, \theta)$ is integrable for all θ. Finally let $M_n(\omega)$ be the sample φ-mean of $\{X \circ T^i(\omega), i = 0, \cdots, n-1\}$ and let M be the φ-mean of X. Then* $\lim_{n\to\infty} M_n = M$ a.e.

PROOF. Let $X_n = X \circ T^n$, $n = 0, 1, \cdots$. Then by the ergodic theorem $(1/n)\sum_{k=0}^{n-1} \varphi(X_k, \theta)$ converges to $E(\varphi(X, \theta))$ almost everywhere. If θ is such that $E(\varphi(X, \theta)) > 0$, then for almost all ω, $\sum_{k=0}^{n} \varphi(X_k(\omega), \theta) > 0$ for all but finitely many n. Hence $M_n(\omega) \leqq \theta$ for all but finitely many n which implies that $\limsup_{n\to\infty} M_n(\omega) \leqq \theta$. Thus $\limsup_{n\to\infty} M_n(\omega) \leqq M$. In the same way by considering those θ for which $E(\varphi(X, \theta)) < 0$ one finds that $\liminf_{n\to\infty} M_n(\omega) \geqq M$ for almost all ω. Hence $\lim_{n\to\infty} M_n = M$ a.e.

Now fix p such that $1 < p < \infty$. When $\varphi(x, \theta) = |\theta - x|^{p-1} \operatorname{sgn}(\theta - x)$ and X is in L_p, the conditions of Lemma 5.4 are satisfied and the φ-mean is the best approximation in L_p to X by a constant (cf. [2] Theorem 2.2).

Recalling the notation of Section 3, let $M_{k,m,n}$ be the sample φ-mean of $h_{1,m,n}(\omega), \cdots, h_{k,m,n}(\omega)$, i.e. of $g_v(Q(\omega)), \cdots, g_v \circ S^{k-1}(Q(\omega))$ (cf. (3.5)). By Lemma 5.4 and the nonsingularity of Q we find that for almost all ω in F, $M_{k,m,n}(\omega)$ converges to the φ-mean of g_v with respect to the conditional probability given F as $k \to \infty$. Now it is easy to show that the $\mathscr{F}_{m,n}$ measurable function h which minimizes $\|g_v - h\|_p$ takes the value of the φ-mean of g_v on the set F. Thus $\lim_{k\to\infty} M_{k,m,n}$ exists almost everywhere and is the best L_p approximation to g_v which is $\mathscr{F}_{m,n}$ measurable. The analogue of the martingale convergence theorem used in Theorem 3.6 is the martingale type theorem of Ando and Amemiya (cf. [1] Theorem 3). Hence

THEOREM 5.5. *Let g_v be in $L_p(\Omega, \mathscr{F}, P)$, $(1 < p < \infty)$. Then*

$$\lim_{n\to\infty}(\lim_{m\to\infty}(\lim_{k\to\infty} M_{k,m,n}(\omega))) = \hat{g}_v$$

almost everywhere where $\hat{g}_v$ is the best L_p approximation to g_v by a function measurable with respect to $\cdots, f_{-1}, f_0$.

REMARK 5.6. The sample means $M_{k,m,n}(\omega)$ do not seem to lend themselves to direct calculation except in the case $p = 2$.

6. Remarks.

REMARK 6.1. There are a number of advantages which our method seems to have over the method of Masani and Wiener.

(A) They assume their process to be nondeterministic whereas our procedure applies also to deterministic processes.

(B) They assume that their process is bounded and they use the multivariate moments to calculate the predictor. We do not assume the existence of any moments other than the first. In particular, our algorithm does not use multiplication which means the random variable may be of a more general nature (cf. Section 5 and (6.3)).

(C) They obtain the predictor as an L_2 limit whereas our limit is almost everywhere. At least for bounded stochastic processes our conclusion is therefore stronger.

REMARK 6.2. In the case of ergodic stochastic processes the error $\{E(|g_\nu - \hat{g}_\nu|^p)\}^{1/p}$ can be calculated by $\{\lim_{n\to\infty}(1/n)\sum_{k=1}^{n}|g_0 - \hat{g}_\nu \circ T^{-\nu}|^p \circ T^{-k}(\omega)\}^{1/p}$. For a more general stochastic process it does not seem possible to calculate the error, although the above limit can be used as an estimate.

REMARK 6.3. We have not made use of the multiplicative properties of the reals. In particular it is clear that our procedure and proof are valid for n-dimensional real or complex space. Thus our theory is actually concerned with *multivariate* non-linear prediction theory.

REMARK 6.4. Of course, $\hat{g}_\nu$ can be interpreted as a least squares predictor only when g_ν is in L_2. However $\hat{g}_\nu$ as a conditional expectation can be calculated when g_ν is only integrable. Indeed using the positivity of the conditional expectation operator and truncation methods our results will extend to the case when $E(g_\nu) = +\infty$ or $E(g_\nu) = -\infty$.

REFERENCES

[1] ANDO, T. and AMEMIYA, I. (1965). Almost everywhere convergence of prediction sequences in $L_p(1 < p < \infty)$. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **4** 113–120.

[2] BRØNS, H. K., BRUNK, H. D., FRANCK, W. E., and HANSON, D. L. (1969). Generalized means and associated families of distributions. *Ann. Math. Statist.* **40** 339–355.

[3] DOOB, J. L. (1953). *Stochastic Processes*. Wiley, New York.

[4] FARRELL, R. H. (1962). Representation of invariant measures. *Illinois J. Math.* **6** 447–467.

[5] HALMOS, P. R. (1956). *Ergodic Theory*. Chelsea, New York.

[6] MASANI, P. and WIENER, N. (1959). Non-linear prediction. *Probability and Statistics*, The Harold Cramer Volume. Wiley, New York. 190–212.

[7] MOY, S. C. (1960). Equalities for stationary processes similar to an equality of Wald. *Ann. Math. Statist.* **31** 995–1000.

[8] ROBERTSON, J. B. (1965). On a problem in non-linear prediction theory. *Ann. Math. Statist.* **36** 1554–1555.