The Annals of Probability
1995, Vol. 23, No. 1, 474–478

CORRECTION

UNIVERSAL PREDICTION SCHEMES

BY PAUL H. ALGOET

The Annals of Probability (1992) **20** 901–941

In an earlier paper, the author proposed a universal scheme to learn the conditional distribution of the next outcome of a stationary process given the infinite past from past experience. An error in that paper is corrected.

1. Introduction. Let $\{X_t\}$ be a stationary random process with values in a Polish space $\mathscr{X}$. It follows from the martingale convergence theorem that the conditional distribution $P(dx \mid X^{-t})$ of $X = X_0$ given the t-past $X^{-t} = (X_{-t}, \dots, X_{-1})$ converges weakly almost surely to the conditional distribution $P(dx \mid X^-)$ of X given the infinite past $X^- = (\dots, X_{-2}, X_{-1})$. In Section 5 of Algoet (1992) (hereafter referred to as [A92]), we assumed that the process distribution P is unknown a priori and we constructed estimates $\hat{P}(dx \mid X^{-t})$ on the basis of the t-past X^{-t} such that

$$\hat{P}(dx \mid X^{-t}) \to P(dx \mid X^-) \quad \text{weakly almost surely as } t \to \infty. \tag{1}$$

A sequence of estimates $\hat{P}(dx \mid X^{-t})$ such that (1) holds under any stationary process distribution P was called a universal prediction scheme.

If $\mathscr{X}$ is a finite set, then the prediction scheme of [A92] reduces to that of Ornstein (1978) and is universal. The scheme of [A92] was claimed to be universal in general when $\mathscr{X}$ is a Polish space, but Gusztáv Morvai (1993; personal communication) has kindly informed me of a gap in the proof of Lemma 6. The purpose of this note is to show a way to avoid the gap. Lemma 6 may not hold as originally stated for all bounded continuous functions $h(x)$, but it does hold for certain simple functions $h(x)$, and this is all we need to prove the main result (Theorem 11) in Section 5 of [A92]. Lemma 4 was obtained by specializing Lemma 6 to the finite alphabet case, but remains valid as originally stated since Lemma 6 holds for the indicator functions $h = \delta_x$ of elements $x \in \mathscr{X}$ when $\mathscr{X}$ is a finite set. The results on gambling, investment, modeling and data compression in the earlier sections of [A92] did not depend on Section 5 and hence are not affected.

We reformulate Lemma 6 of [A92] and introduce an approximation argument to show that the prediction scheme of [A92] is universal. We use the same notation as in [A92] with some minor changes.

Received August 1993; revised May 1994.
AMS 1991 *subject classifications*. 62M20, 60G25.
Key words and phrases. Learning from experience, predictive modeling, stationary processes.

2. Reformulation of Lemma 6. Let $(\mathscr{X}, \mathscr{B}_{\mathscr{X}})$ be a separable metric space with its Borel σ-field and let $\{X_t\}$ be a stationary $\mathscr{X}$-valued random process. Let $\{\mathscr{G}_k\}_{k\geq 1}$ be an increasing sequence of finite subfields that asymptotically generate $\mathscr{B}_{\mathscr{X}}$ and let $\mathscr{F}^{-t}$ denote the finite subfield of $\sigma(X^{-t})$ that is generated by the events $\{X_{-t} \in B_t, \dots, X_{-1} \in B_1\}$, where $B_t, \dots, B_1$ are $\mathscr{G}_t$-measurable subsets of $\mathscr{X}$. Notice that $\mathscr{F}^{-t}$ is monotonically increasing and asymptotically generates the limiting σ-field $\sigma(X^-)$. For $s, t \geq 0$ we define the empirical estimate $\hat{P}_s(dx \mid \mathscr{F}^{-t})$ of the true conditional distribution $P(dx \mid \mathscr{F}^{-t})$ as in [A92].

Let $\varepsilon_k = 1/k$. We say that two distributions Q and R on the space $\mathscr{X}$ are k-close if

$$\sup_{B\in\mathscr{G}_k} |Q(B) - R(B)| \leq \varepsilon_k. \tag{2}$$

Thus the total variation distance between the restriction of Q and the restriction of R to the finite field $\mathscr{G}_k$ should be no larger than ε_k. If $h(x)$ is any $\mathscr{G}_k$-measurable function such that $0 \leq h(x) \leq 1$, then the expectations $Q\{h(X)\}$ and $R\{h(X)\}$ will differ by at most ε_k.

For each $k \geq 1$ we define the empirical estimate $\hat{P}_k(dx \mid X^-)$ of the true conditional distribution $P(dx \mid X^-)$ as follows. Find the least integer n for which there exists an integer K and a sequence $(s_i)_{0\leq i\leq K}$ such that $k \leq K = s_0 < s_1 < \cdots < s_K = n$ and all empirical estimates $\hat{P}_{s_i}(dx \mid \mathscr{F}^{-t})$, $s_0 \leq t \leq s_{i-1}$, $1 \leq i \leq K$ are well defined and k-close to each other. Choose K smallest possible, choose the sequence $(s_i)_{0\leq i\leq K}$ smallest in lexicographic order when read in reverse and set

$$\hat{P}_k(dx \mid X^-) = \hat{P}_{s_K}(dx \mid \mathscr{F}^{-s_{K-1}}). \tag{3}$$

The estimate $\hat{P}_k(dx \mid X^-)$ was denoted by $\hat{P}(dx \mid X^{-\sigma_k})$ in [A92] and is well defined almost surely for all $k \geq 1$ since the search for K and $(s_i)_{0\leq i\leq K}$ must terminate by the ergodic theorem and the martingale convergence theorem (see Lemma 5 of [A92]).

The bad event $B^{\ell}_{\alpha,K}(h)$ is defined as in [A92] for any bounded measurable function $h(x)$, any integers $K \geq 1$ and $\ell \geq 0$ and any real number $0 < \alpha < 1$. Lemma 6 of [A92] should be replaced by the following result.

LEMMA 6′. *Suppose the function $h(x)$ is $\mathscr{G}_K$-measurable and $0 \leq h(x) \leq 1$. Then for $\ell \geq 0$ and $0 < \alpha < 1$ we have*

$$P\{B^{\ell+1}_{\alpha,K}(h) \mid B^{\ell}_{\alpha,K}(h)\} \leq (1-\alpha) \quad \text{almost surely} \tag{4}$$

and consequently, by induction on ℓ,

$$P\{B^{\ell}_{\alpha,K}(h)\} \leq (1-\alpha)^{\ell}. \tag{5}$$

The proof of Lemma 6 in [A92] breaks down at the foot of page 938, where it is asserted that the atom $W = W(\omega)$ is a cylinder set in $\mathscr{F}^{-\sigma_l}$ since all evidence proving that $\omega \in B^{l}_{\alpha K}(h)$ is contained in $\mathscr{F}^{-\sigma_l}$. Morvai (1993) observed that

W is not $\mathscr{F}^{-\sigma_l}$-measurable, while the assertion on page 940 that $A_E(N) \to P\{h(X) \mid \mathscr{F}^{-\sigma_l}\}$ requires that W be an atom of $\mathscr{F}^{-\sigma_l}$. [In general we have $A_E(N) \to P\{h(X) \mid W\}$.] Thus the proof of Lemma 6 is not valid for every bounded continuous or bounded measurable function $h(x)$, as was claimed. However, the lemma is valid if $h(x)$ is constant on atoms of the finite field $\mathscr{G}_K$, and this turns out to be sufficient for our purposes.

3. Consistency of the estimates $\hat{P}_k(dx \mid X^-)$. Theorem 11 of [A92] is valid, but the proof needs adjustment because we cannot rely on the old Lemma 6. We proceed in two steps: first we infer some conclusions from the revised Lemma 6′, and then we prove universality of the estimates $\hat{P}_k(dx \mid X^-)$.

THEOREM 11A. *For any measurable set B in the generating field $\bigcup_k \mathscr{G}_k$, we have*

$$\lim_k \hat{P}_k\{X \in B \mid X^-\} = P\{X \in B \mid X^-\} \quad P\text{-almost surely.} \tag{6}$$

If a function $h(x)$ is $\mathscr{G}_\kappa$-measurable for some $\kappa \geq 1$, then the estimate $\hat{P}_k\{h(X) \mid X^-\}$ converges almost surely to the true conditional expectation $P\{h(X) \mid X^-\}$ as $k \to \infty$.

PROOF. Suppose $\kappa \geq 1$, $h(x)$ is a $\mathscr{G}_\kappa$-measurable function such that $0 \leq h(x) \leq 1$ and $0 < \alpha < 1$. We consider the event

$$\hat{P}_k\{h(X) \mid X^-\} \geq 2\alpha + P\{h(X) \mid X^-\} \quad \text{i.o.} \tag{7}$$

If this event occurs, then the event $B^K_{\alpha,K}(h)$ occurs for infinitely many K. However, if $K \geq \kappa$, then $h(x)$ is $\mathscr{G}_K$-measurable and Lemma 6′ implies that

$$P\{B^K_{\alpha,K}(h)\} \leq (1-\alpha)^K.$$

The event $B^K_{\alpha,K}(h)$ will occur for only finitely many K with probability 1 by the Borel–Cantelli lemma. Thus the event (7) has vanishing probability and Theorem 11A follows. □

We now define the increasing subfields $\mathscr{G}_k$ in a special way and prove that the estimates $\hat{P}_k(dx \mid X^-)$ converge weakly (in law, in distribution) almost surely to $P(dx \mid X^-)$ as $k \to \infty$.

Let $\{h_\kappa\}_{\kappa \geq 1}$ be a separating sequence of bounded continuous functions on $\mathscr{X}$. For any probability distributions Q_n and Q on $\mathscr{X}$, we have weak convergence $Q_n \to Q$ iff

$$Q_n\{h_\kappa(X)\} \to Q\{h_\kappa(X)\} \quad \text{as } n \to \infty \text{ for each fixed } \kappa \geq 1.$$

We assume without loss of generality that all $h_\kappa(x)$ are bounded between 0 and 1 and we define $\mathscr{G}_k$ as the finite field that is generated by atoms of the form

$$B_{i_1 i_2 \cdots i_k} = \{x \in \mathscr{X}\colon\ i_\kappa 2^{-k} \le h_\kappa(x) < (i_\kappa + 1)2^{-k}, 1 \le \kappa \le k\}, \tag{8}$$

where $0 \le i_\kappa \le 2^k$, $1 \le \kappa \le k$. The functions $g_1, g_2, \dots, g_k$ oscillate by no more than 2^{-k} on each atom $B_{i_1 i_2 \cdots i_k}$ of $\mathscr{G}_k$. The finite fields $\mathscr{G}_k$ are monotonically increasing and asymptotically generate the Borel σ-field $\mathscr{B}_{\mathscr{X}}$.

THEOREM 11B. *Suppose the fields $\mathscr{G}_k$ and $\mathscr{F}^{-t}$ are defined as above. If P is a stationary distribution, then for any bounded continuous function $h(x)$ on $\mathscr{X}$ we have*

$$\lim_{k\to\infty} \hat{P}_k\{h(X) \mid X^-\} = P\{h(X) \mid X^-\} \quad P\text{-almost surely.} \tag{9}$$

If a regular conditional distribution $P(dx \mid X^-)$ exists, then

$$\hat{P}_k(dx \mid X^-) \to P(dx \mid X^-) \quad \text{weakly } P\text{-almost surely as } k \to \infty. \tag{10}$$

PROOF. It suffices to prove that for any fixed $\kappa \ge 1$ and $\varepsilon > 0$,

$$\limsup_k |\hat{P}_k\{h_\kappa(X) \mid X^-\} - P\{h_\kappa(X) \mid X^-\}| \le \varepsilon \quad \text{almost surely.} \tag{11}$$

For each $K \ge 1$ we select a representative point $\xi_B^{(K)}$ in each atom B of $\mathscr{G}_K$ and for any $\kappa \ge 1$ we consider the $\mathscr{G}_K$-measurable function

$$h_\kappa^{(K)}(x) = h_\kappa(\xi_B^{(K)}) \quad \text{if } x \in B, B \in \text{Atoms}(\mathscr{G}_K).$$

If $K \ge \kappa$, then by construction

$$|h_\kappa(x) - h_\kappa^{(K)}(x)| \le 2^{-K}. \tag{12}$$

To prove (11), choose some K such that $K \ge \kappa$ and $2^{-K+1} \le \varepsilon$. Obviously

$$\hat{P}_k\{h_\kappa(X) \mid X^-\} - P\{h_\kappa(X) \mid X^-\} = U_{k,\kappa}^{(K)} + V_{k,\kappa}^{(K)},$$

where

$$U_{k,\kappa}^{(K)} = \hat{P}_k\{h_\kappa^{(K)}(X) \mid X^-\} - P\{h_\kappa^{(K)}(X) \mid X^-\},$$
$$V_{k,\kappa}^{(K)} = \hat{P}_k\{h_\kappa(X) - h_\kappa^{(K)}(X) \mid X^-\} - P\{h_\kappa(X) - h_\kappa^{(K)}(X) \mid X^-\}.$$

The function $h_\kappa^{(K)}(x)$ is $\mathscr{G}_K$-measurable and bounded between 0 and 1; hence, $U_{k,\kappa}^{(K)} \to 0$ almost surely as $k \to \infty$ by Theorem 11A. On the other hand, it follows from (12) that $|V_{k,\kappa}^{(K)}| \le 2^{-K+1} \le \varepsilon$. The desired conclusion (11) and the theorem follow. □

In general $X_t(\omega) = X(T^t\omega)$ for some random variable X with values in a separable metric space $\mathscr{X}$ and some invertible measure-preserving transformation T of the underlying probability space $(\Omega, \mathscr{F}, P)$. A regular conditional distribution $P(dx \mid X^-)$ exists if $\mathscr{X}$ is a Polish space or a universally measurable subset of such a space, or alternatively if the measure P is perfect.

REFERENCES

ALGOET, P. H. (1992). Universal schemes for prediction, gambling and portfolio selection. *Ann. Probab.* **20** 901–941.

MORVAI, G. (1993). Personal communication.

ORNSTEIN, D. S. (1978). Guessing the next output of a stationary process. *Israel J. Math.* **30** 292–296.

2613 ELMDALE
PALO ALTO, CALIFORNIA 94303