Compound Bayes Predictors for Sequences with Apparent Markov Structure

THOMAS M. COVER, FELLOW, IEEE, AND AARON SHENHAR

Abstract—**Sequential predictors for binary sequences with no assumptions upon the existence of an underlying process are discussed. The rule offered here induces an expected proportion of errors which differs by $0(n^{-1/2})$ from the Bayes envelope with respect to the observed kth order Markov structure. This extends the compound sequential Bayes work of Robbins, Hannan and Blackwell from sequences with perceived 0th order structure to sequences with perceived kth order structure. The proof follows immediately from applying the 0th order theory to 2^k separate subsequences. These results show the essential robustness of procedures which play Bayes with respect to (a perhaps randomized) version of an estimate of the distribution of the past. Such procedures still have asymptotically good properties even when the underlying assumptions for which they were originally developed no longer hold.**

I. Introduction

WE ARE interested in sequential prediction procedures that exploit any apparent order in the sequence. We do not assume the existence of any underlying distributions, but assume that the sequence is an outcome of a game against a malevolent intelligent nature. We shall show that if the sequence has any kth order Markov structure, the expected sequential prediction score will be as high as if the predcitor had known this empirical distribution at the beginning. Thus nature cannot "set up" the predictor for future disastrous predictions.

To build up motivation for the goal in the prediction problem, let us consider a sequence of questions. At first, let there be no dependence whatsoever. Suppose that a statistician is observing a Bernoulli process $x_1,x_2,\cdots$ with parameter p, where p is known, and wishes to minimize his proportion of errors. Clearly what he should do is assert that the ith term of the sequence will be one, for all i, if $p \geq \frac{1}{2}$, and assert that the ith term will be zero, for all i, if $p < \frac{1}{2}$. The proportion of correct guesses will then be asymptotically equal to $\max\{p, 1-p\}$.

Next, we suppose that the statistician observes the same Bernoulli process but that the parameter p is unknown. It then makes sense that the statistician should use the above Bayes rule based on

$$\hat{p}_n = \frac{1}{n}\sum_{1}^{n} x_j = \bar{x}_n,$$

an estimate of p based on the past. Clearly, since $\bar{x}_n \to p$, the statistician's decision rule is asymptotically Bayes, and for any p his average score will converge again to $\max\{p, 1-p\}$.

Robbins [1] and Blackwell [3] observed that the procedure of using the past to estimate the parameters of the distribution and then using the Bayes classification with respect to the estimated distribution has a curious robustness in the sense that it works even when there is no underlying distribution for the sequence $x_1x_2,\cdots$. To see this, suppose that the sequence at time n has k ones. If indeed the sequence were Bernoulli with unknown parameter, we would have liked to obtain a proportion of correct guesses equal to

$$\max\left\{\frac{k}{n}, 1-\frac{k}{n}\right\}.$$

In other words, a statistician would be quite upset if somehow he had used a sequential prediction procedure which ignored the fact that the sequence had a preponderence of ones or zeros. On the other hand, since he is predicting sequentially, how could he know ahead of time that the sequence was going to be biased in one direction or the other? Moreover, some sequences have all the k ones at the beginning, and others have all the ones at the end, thus leading the statistician to form a premature hypothesis about the overall preponderence of ones or zeros in the entire sequence. Nonetheless, Blackwell and Robbins have been able to show that if one slightly changes the past estimate of p by adding a small random perturbation to $\bar{x}_n$ and plays Bayes with respect to this randomized estimate of the past, then one can epsilon-achieve $\max\{\bar{x}_n, 1-\bar{x}_n\}$, for all n, for *all* sequences $(x_1,x_2,\cdots) \in \{0,1\}^\infty$. This is the robustness to which we referred. No matter how nature chooses a sequence of zeros and ones, the statistician can do as well in sequentially predicting the terms in this sequence as if he had known ahead of time the proportion of ones in the sequence. The natural generalization of this result was characterized by Neyman [9] in 1962 as one of three major breakthroughs in statistics. A scheme that seemingly would work only in an i.i.d. case works in the general non i.i.d. case and, in fact, without any statistical framework whatsoever.

There might still be cause for disappointment for the statistician, however, if he noticed at time n that not only did the sequence have a preponderence of ones, but that it had a very orderly first-order structure. For example, it might be true that each zero was followed by a one. Such a sequence might be 0101101010110101101101101. He would be distressed that his sequential prediction scheme had not achieved an even higher proportion of correct

Manuscript received June 21, 1976; revised December 9, 1976. This work was supported by the Air Force Office of Scientific Research under Contract F44620-74-C-0068.

T. M. Cover is with the Department of Electrical Engineering and the Department of Statistics, Stanford University, Stanford, CA 94305.

A. Shenhar was with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305. He is now with the Scientific Department, Israeli Ministry of Defense, Haifa, Israel.

guesses taking into account the empirical first-order dependence.

The purpose of this paper is to show how the statistician can achieve the highest possible proportion of correct guesses up to kth-order dependence for any arbitrarily chosen k. We must first, of course, define the natural goal, which we refer to as the Bayes envelope for kth-order dependence, just as the natural goal for 0th-order dependence was max $\{\bar{x}_n, 1 - \bar{x}_n\}$.

For example, on the sequence in the previous paragraph, the second-order structure suggests that we should

i) never make an error after a 10 (a 1 always follows),
ii) make an error $\frac{1}{2}$ the time after a 01,
iii) never make an error after a 11 (a 0 always follows).

Since 10's occur with empirical probability 10/25, 01's with probability 10/25, and 11's with probability 5/25, the proportion of correct predictions "should" be (10/25) + (10/25)(1/2) + 5/25 = 4/5. To have achieved substantially less would be to have missed the obvious. We shall note the useful fact that the sequence can be decomposed into 2^k subsequences of "independent" variables and use the standard Blackwell, Robbins, Hannan predictor on each subsequence to achieve the natural Bayes envelope. Thus the apparently more difficult kth-order Markov problem will be seen by this trick to be no more than a direct application of existing theory.

II. Preliminaries

Let $x \in \{0,1\}^\infty$. At time n we observe the sequence $x(n) = (x_1,x_2,\cdots,x_n) \in \{0,1\}^n$. How do we predict x_{n+1}? We shall base our guess on the observed empirical distribution of sequences of length $k + 1$ which have the first k bits equal to $(x_{n-k+1},\cdots,x_n)$. As mentioned, this approach divides the question into 2^k parallel problems indexed by the immediate past.

Let $\delta(n) = (\delta_1,\delta_2,\cdots,\delta_n) \in \{0,1\}^n$ be a sequence of guesses for $x_1,x_2,\cdots,x_n$. Then the total prediction score (the fraction of correct guesses) will be given by

$$S_n = \frac{1}{n}\sum_{i=1}^{n}(1 - |\delta_i - x_i|). \tag{1}$$

Let a random sequential predictor be described by $p = (p_1,p_2,\cdots,p_n)$, with the interpretation that we predict $x_i = 1$ with probability $p_i = p_i(x_1,x_2,\cdots,x_{i-1}, \delta_1,\delta_2,\cdots,\delta_{i-1})$, $0 \le p_i \le 1$. The expected empirical average score will be given by

$$\bar{S}_n = \frac{1}{n}\sum_{i=1}^{n}(p_i x_i + (1 - p_i)(1 - x_i)). \tag{2}$$

The 0th-order prediction algorithms which were considered by Robbins [1], Hannan [2], and Blackwell [3] and [4] achieve $\bar{S}_n$ near the so-called "Bayes envelope"

$$\hat{S}_n = \max\left\{\frac{1}{n}\sum_{i=1}^{n} x_i,\ 1 - \frac{1}{n}\sum_{i=1}^{n} x_i\right\} \tag{3}$$

with an error bounded by $c/\sqrt{n}$. See also Cover [5] for the optimal solution when n is known in advance. In [8] Tainiter considers the problem of sequentially guessing a Markov sequence with an empirical Bayes formulation, but considers there to be a true underlying source of known order. Johns [7] considers a compound decision problem where it is assumed that each observation belongs to some distribution F in a specified family of distributions. The decision for each problem is based on a uniform mixture of the last k distributions. For a similar problem, Swain has shown in [6] that a Bayes risk function based on the kth-order empirical distribution of the past distributions is optimal if in fact the vector of distributions possesses a kth-order dependence. This demonstrates that the Bayes envelope is a natural goal. In the current paper, we demonstrate similar results hold even without an underlying probabilistic structure.

In the next section we present formally the natural goal for a prediction process when a sequence seems to have some Markov structure. In Section IV we review Blackwell's algorithm for a prediction procedure for the zero-order problem based on a two-person zero-sum game. In Section V we then apply this algorithm to derive an expression for the asymptotic behavior of the expected score with respect to the Bayes envelope for the empirical statistics that arise from the kth-order Markov structure.

III. A Formal Description of the Natural Goal for the Prediction Score

Suppose that a sequence $x(n) \in \{0,1\}^n$ is observed that seems to have a Markov structure; i.e., the empirical Markov matrix of some order is far from Bernoulli. We now provide a formal description of the natural goal for the prediction score on sequences with empirically observed dependence. For a moment, suppose we have a true kth-order Markov process $\{x_i\}_{i=1}^{\infty}$, $x_i \in \{0,1\}$, with known distribution. Let a state of the process be denoted by $z \in \{0,1\}^k$ and let the transition probability matrix be $P(z'|z)$, $z, z' \in \{0,1\}^k$. For each z let the nonzero elements of the transition probability matrix be $P(1|z)$ and $P(0|z) = 1 - P(1|z)$, the respective probabilities that 1 and 0 follow state z. Let $\mu(z)$ be the stationary distribution on the state space. The Bayes decision, given state z, is to decide x to be that which maximizes $P(x|z)$, $x \in \{0,1\}$. Thus the Bayes predictor will induce the following steady-state probability of a correct guess:

$$P_c^*(\mu,P) = \sum_{z\in\{0,1\}^k} \mu(z) \max\{P(1|z), 1 - P(1|z)\}; \tag{4}$$

i.e., for each state we achieve the conditional Bayes score max $\{P(1|z), 1 - P(1|z)\}$, and the total score is a weighted sum of these scores according to the stationary distribution of each state.

Returning to the nonstatistical problem, it follows that a desirable score for our observed sequence will be $\bar{S}_n = P_c^*(\hat{\mu},\hat{P})$, where $\hat{\mu}$, $\hat{P}$ are the empirical statistics induced by $x(n)$ as follows. Let $z,1$ and $z,0$ be the sequence z followed by 1 and 0, respectively. Let $n(z,1)$ and $n(z,0)$ be respectively the number of times the sequences $z,1$ and $z,0$ were observed

in $x(n)$, and let $n'(z)$ be the number of times the sequence z was observed in $x(n-1) = (x_1, x_2, \cdots, x_{n-1})$. We note that

$$n'(z) = n(z,1) + n(z,0). \tag{5}$$

Now define

$$\hat{P}(1|z) = \frac{n(z,1)}{n'(z)}, \qquad \hat{P}(0|z) = \frac{n(z,0)}{n'(z)}$$

$$\hat{\mu}(z) = \frac{n'(z)}{\sum_{z' \in \{0,1\}^k} n'(z')}. \tag{6}$$

Then $\hat{P}(1|z)$ and $\hat{P}(0|z)$ define transition probabilities for a kth-order Markov process with a stationary distribution $\hat{\mu}(z)$.

IV. A Random Predictor with $k = 0$ (Blackwell's Procedure)

Let E^2 be two dimensional Euclidean space with coordinates $\bar{x}_n, S_n$, where

$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

and S_n is the empirical average score given by (1). Let the point $(\bar{x}_n, S_n) \in E^2$ be the current result of the game so far. On the next step, we predict $x_{n+1} = 1$ and $x_{n+1} = 0$ with probabilities p_{n+1} and $1 - p_{n+1}$, respectively. Following Blackwell's [3] definition of approachability of sets, the convex set in E^2 defined by the Bayes envelope $\bar{S}_n$ is achievable with the following procedure:

$$P_{n+1} = \begin{cases} 0, & \text{for } |\bar{x}_n - \tfrac{1}{2}| \geq |S_n - \tfrac{1}{2}| \text{ and } \bar{x}_n \leq \tfrac{1}{2} \\ 1, & \text{for } |\bar{x}_n - \tfrac{1}{2}| \geq |S_n - \tfrac{1}{2}| \text{ and } \bar{x}_n > \tfrac{1}{2} \\ 0.5, & \text{for } |\bar{x}_n - \tfrac{1}{2}| < |S_n - \tfrac{1}{2}| \text{ and } S_n > \tfrac{1}{2} \\ \dfrac{\bar{x}_n - S_n}{1 - 2S_n}, & \text{for } |\bar{x}_n - \tfrac{1}{2}| < |S_n - \tfrac{1}{2}| \text{ and } S_n \leq \tfrac{1}{2}. \end{cases} \tag{7}$$

Hannan [2, p. 139] has indicated that the Blackwell procedure satisfies

$$\hat{S}_n - \bar{S}_n \leq \frac{3}{\sqrt{n}}, \qquad \forall n, \forall x \in \{0,1\}^\infty; \tag{8}$$

i.e., one can achieve a score arbitrarily close to the Bayes envelope uniformly in $x(n)$.

The extension of these results to kth-order dependence is found by applying Blackwell's procedure to a series of parallel problems determined by the current state of the observed sequence.

V. kth-Order Markov Predictor

At time n, $n = k, k+1, \cdots$, let $z = (x_{n-k+1}, \cdots x_n)$ be the current state of the observed sequence. We now predict the next bit according to the distribution of the sequences $z,1$ and $z,0$ in $x_1, x_2, \cdots, x_n$, using the procedure described in Section IV. Therefore, for each z we have a separate prediction game, and each game is played exactly $n'(z)$ times.

Let $\bar{S}_{n'(z)}$ be the expected score of the game indexed by state z, and let the partial Bayes envelope for the z game be denoted by

$$\hat{S}_{n'(z)} = \max\left\{\frac{n(z,1)}{n'(z)}, \frac{n(z,0)}{n'(z)}\right\}$$

$$= \max\{\hat{P}(1|z), 1 - \hat{P}(1|z)\}. \tag{9}$$

We can immediately apply Blackwell's bound to each game. Thus $\bar{S}_{n'(z)}$ will satisfy

$$\hat{S}_{n'(z)} - \bar{S}_{n'(z)} \leq \frac{3}{\sqrt{n'(z)}},$$

$$\forall z \in \{0,1\}^k, \forall n'(z), \forall x \in \{0,1\}^\infty. \tag{10}$$

Let m be the total number of prediction plays. Using (5) for $n'(z)$ we obtain

$$m = \sum_{z \in \{0,1\}^k} n'(z) = n - k. \tag{11}$$

(The predictor is not defined on the first k terms of $x(n)$.) Now the average expected score over $m = n - k$ plays will be

$$\bar{S}_m = \frac{1}{m} \sum_{z \in \{0,1\}^k} n'(z) \bar{S}_{n'(z)}. \tag{12}$$

However, we are interested in the total score achieved over the full length of $x(n)$. Therefore, counting no score for the first k plays,

$$\bar{S}_n = \frac{1}{n} \sum_{z \in \{0,1\}^k} n'(z) \bar{S}_{n'(z)} = \frac{n-k}{n} \bar{S}_m. \tag{13}$$

But the total kth-order Bayes envelope is

$$P_c^*(n, \hat{\mu}, \hat{P}) = \sum_{z \in \{0,1\}^k} \hat{\mu}(z) \max\{\hat{P}(1|z), 1 - \hat{P}(1|z)\}$$

$$= \frac{1}{n-k} \sum_{z \in \{0,1\}^k} n'(z) \hat{S}_{n'(z)}, \tag{14}$$

where $\hat{\mu}(z)$ and $\hat{P}(1|z)$ are the empirical statistics given by (6). We can therefore establish the following.

Theorem 1: The prediction procedure in (7), performed conditionally on each $z \in \{0,1\}^k$, achieves

$$P_c^*(n, \hat{\mu}, \hat{P}) - \bar{S}_n \leq \frac{2^k 3}{\sqrt{n}} + \frac{k}{n}, \quad \forall x \in \{0,1\}^\infty, \forall n. \tag{15}$$

Thus the expected score $\bar{S}_n$ is essentially as high as if we had initially known the empirical kth-order Markov structure.

Proof:

$$P_c^*(n, \hat{\mu}, \hat{P}) - \bar{S}_n = \frac{1}{n-k} \sum_{z \in \{0,1\}^k} n'(z) \hat{S}_{n'(z)}$$

$$- \frac{1}{n} \sum_{z \in \{0,1\}^k} n'(z) \bar{S}_{n'(z)}$$

$$= \frac{1}{n} \sum_{z \in \{0,1\}^k} n'(z) (\hat{S}_{n'(z)} - \bar{S}_{n'(z)})$$

$$+ \frac{k}{n(n-k)} \sum_{z \in \{0,1\}^k} n'(z) \hat{S}_{n'(z)}. \tag{16}$$

From (10) and the fact

$$\frac{1}{n-k}\sum_{z\in\{0,1\}^k} n'(z)\hat{S}_{n'(z)} \le 1,$$

we obtain

$$P_c^*(n,\hat{\mu},\hat{P}) - \bar{S}_n \le \frac{1}{n}\sum_{z\in\{0,1\}^k} n'(z)\frac{3}{\sqrt{n'(z)}} + \frac{k}{n}$$

$$\le \frac{2^k 3}{n}\sqrt{\max_{z\in\{0,1\}^k} n'(z)} + \frac{k}{n} \le \frac{2^k 3}{\sqrt{n}} + \frac{k}{n}. \tag{17}$$

VI. Remarks

Although our motivation is nonstatistical and not tied to the existence of a true underlying process, we remark that if $\{X_i\}_{i=1}^{\infty}$ is indeed a kth-order Markov process, the predictor given in the theorem is asymptotically Bayes in the sense that

$$\bar{S}_n \to P_c^*(\mu, P), \text{ wp } 1. \tag{18}$$

If $\{x_i\}_{i=1}^{\infty}$ is a sample sequence from a kth-order Markov process with unknown statistics, it can be seen that there exists a deterministic sequential predictor which learns the statistics and asymptotically achieves the Bayes risk. One may wonder why randomization is required for the sequential predictor studied here. The answer is essentially game theoretic. In fact, it can be easily seen [5] that for any deterministic sequential predictor there exists a sequence for which $S(x(n)) = 0$ and $P_c^*(n,\hat{\mu},\hat{P}) = \frac{1}{2}$.

The Blackwell predictor could be replaced by other compound sequential Bayes predictors. For example, using Hannan's procedure [2, Section 7] for a Bayes predictor, where uniformly distributed random variables weighted by $cn^{-1/2}$ are added to the empirical transition probabilities, we can achieve the Bayes envelope (14), where the bound in (15) is replaced by $(12/n)^{1/2}2^k + k/n$. Finally, the above convergence rates hold for arbitrary finite alphabet size X.

References

[1] H. Robbins, "Asymptotically subminimax solutions of compound statistical decision problems," *Proc. Second Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, pp. 131–148, 1951.

[2] J. Hannan, "Approximation to Bayes risk in repeated play," Contributions to the Theory of Games, vol. III, Annals of Mathematics Studies No. 39, pp. 97–139, Princeton 1957.

[3] D. Blackwell, "An analog of the minimax theorem for vector payoffs," *Pacific Journal of Mathematics*, vol. 6, pp. 1–8, 1956.

[4] ——, "Controlled random walks," *Proc. International Congress of Mathematicians 1954*, vol. III, Amsterdam, North Holland, pp. 336–338, 1956.

[5] T. M. Cover, "Behavior of sequential predictors of binary sequences," *Trans. Fourth Prague Conference on Information Theory Statistical Decision Functions, Random Processes*, 1965, Publishing House of the Czechoslovak Academy of Sciences, Prague 1967, pp. 263–272.

[6] D. D. Swain, "Bounds and rates of convergence for the extended compound estimation problem in the sequence case," Technical Report No. 81, Department of Statistics, Stanford University, 1965.

[7] M. Johns, "Two-action compound decision problems," *Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, vol. 1, pp. 463–478, 1967.

[8] M. Tainiter, "Sequential hypothesis test for r-dependent marginally stationary processes," *The Annals of Mathematical Statistics*, vol. 37, no. 1, pp. 90–97, Feb. 1966.

[9] J. Neyman, "Two breakthroughs in the history of statistical decision making," *Review of the International Statistical Institute*, vol. 30, pp. 11–27, 1962.