The Annals of Probability
1992, Vol. 20, No. 2, 901–941

UNIVERSAL SCHEMES FOR PREDICTION, GAMBLING AND PORTFOLIO SELECTION[1]

BY PAUL ALGOET

Stanford University

We discuss universal schemes for portfolio selection. When such a scheme is used for investment in a stationary ergodic market with unknown distribution, the compounded capital will grow with the same limiting rate as could be achieved if the infinite past and hence of the distribution of the market were known to begin with. By specializing the market to a Kelly horse race, we obtain a universal scheme for gambling on a stationary ergodic process with values in a finite set. We point out the connection between universal gambling schemes and universal modeling schemes that are used in noiseless data compression. We also discuss a universal prediction scheme to learn, from past experience, the conditional distribution given the infinite past of the next outcome of a stationary ergodic process with values in a Polish space. This generalizes Ornstein's scheme for finite-valued processes. Although universal prediction schemes can be used to obtain universal gambling and portfolio schemes, they are not necessary.

1. Introduction. Let $\{J_t\}$ be a sequence of random variables taking values in the finite set $\{1, \dots, m\}$, and suppose a gambler is allowed to bet on these random variables according to a nonanticipating strategy. Thus the gambler may arbitrarily distribute his compounded wealth at the beginning of every round t with knowledge of the t-past $J^t = (J_0, \dots, J_{t-1})$ over the possible outcomes $j = 1, \dots, m$. The gambler will collect the return of his investment at the end of round t when the random outcome J_t is revealed, and start another round. We assume that bets are paid out at uniform odds, so, the return at the end of round t is m times the amount that was invested in the actual outcome J_t. Consequently, if the gambler starts with initial wealth $S_0 = 1$ and if he places a fraction $Q(j_t|J^t)$ on every possible outcome j_t during round t, then his compounded wealth after n rounds amounts to $S_n = [m^n Q(J^n)]$, where

$$(1) \qquad Q(J^n) = \prod_{0 \le t < n} Q(J_t|J^t).$$

Suppose the gambler knows the distribution P of the random process $\{J_t\}$. To maximize the growth rate of compounded wealth, he should place bets at the beginning of every round t proportional to the conditional probability $P(j_t|J^t)$ on every possible outcome j_t. Indeed, if $Q(j_t|J^t)$ is any alternative

Received June 1988; revised October 1990.

[1] An early version of this paper was presented at the IEEE International Symposium on Information Theory in Ann Arbor, Michigan, October 1986.

AMS 1980 *subject classifications*. 90A09, 60G10, 60G40, 47A35.

Key words and phrases. Portfolio selection, universal prediction schemes, universal gambling schemes, stationary ergodic process.

betting scheme, then $Q(J^n)/P(J^n)$ is a nonnegative supermartingale with respect to the information fields $\sigma(J^n)$. The initial value $Q(J^0)/P(J^0)$ is equal to 1 by convention, so that all the expectations $E\{Q(J^n)/P(J^n)\}$ are bounded by 1. It follows (cf. Lemma 2 below) that

$$\limsup_n \frac{1}{n} \log\left(\frac{[m^n Q(J^n)]}{[m^n P(J^n)]} \right) \le 0 \quad \text{a.s.,} \tag{2}$$

and consequently,

$$\limsup_n \frac{1}{n} \log[m^n Q(J^n)] \le \limsup_n \frac{1}{n} \log[m^n P(J^n)] \quad \text{a.s.,}$$

$$\liminf_n \frac{1}{n} \log[m^n Q(J^n)] \le \liminf_n \frac{1}{n} \log[m^n P(J^n)] \quad \text{a.s.}$$

If $\{J_t\}$ is stationary ergodic, then the maximum growth rate is well defined and almost surely equal to the constant $[\log m - H(J|J^-)]$, where $H(J|J^-) = E\{-\log P(J|J^-)\}$ is the entropy rate of $\{J_t\}$, that is the conditional entropy of $J = J_0$ given the infinite past $J^- = (\dots, J_{-2}, J_{-1})$. Indeed, the Shannon–McMillan–Breiman theorem implies that

$$\frac{1}{n} \log[m^n P(J^n)] \to [\log m - H(J|J^-)] \quad \text{a.s.} \tag{3}$$

This was observed for stationary ergodic processes with known distribution by Cover (1974), after Kelly (1956) and Breiman (1961) considered the independent identically distributed case. See also Cover and King (1978).

Cover (1974) also posed the challenging question how to gamble on a stationary ergodic sequence whose distribution is unknown. The conditional probability $P(j_t|J^t)$ can be estimated on the basis of the t-past J^t, and the question is whether wealth allocated according to these estimates will compound with the same maximum rate that would be achievable if the process distribution were known a priori. A gambling scheme is a computable function $Q(j_t|j^t)$ of sequences $j^t = (j_0, \dots, j_{t-1})$ and elements j_t in $\{1, \dots, m\}$, taking values in $[0, 1]$.

Problem 1 (Existence of a universal gambling scheme). Find a gambling scheme $\hat{P}(j_t|j^t)$ that is universal in the following sense. If $\{J_t\}$ is any stationary ergodic random process with distribution P on the m-ary sequence space, then a gambler who at time t with knowledge of the t-past J^t places bets according to the strategy $\hat{P}(j_t|J^t)$ on all possible outcomes j_t will earn money with the same limiting rate as a gambler who apportions his wealth according to the true but unknown conditional probabilities $P(j_t|J^t)$, that is,

$$\begin{aligned} \frac{1}{n} \log\left[m^n \hat{P}(J^n)\right] &\to [\log m - H(J|J^-)] \\ &= \lim_n \frac{1}{n} \log[m^n P(J^n)] \quad \text{a.s.} \end{aligned} \tag{4}$$

Let $K(j^n)$ denote the length of the shortest binary program which computes j^n, when no program is allowed to be a prefix of another. Thus $K(j^n)$ is the algorithmic entropy of Chaitin, which is nearly equivalent to the Kolmogorov complexity. Cover (1974) exhibited a scheme for gambling on individual binary sequences j^n such that the compounded wealth will grow by a factor $[2^n Q(j^n)] \geq 2^{n-K(j^n)}$. Consequently, if a data compression algorithm can reduce the description length of a sequence from n to K by removing $n - K$ bits of redundancy, then a gambler can double his wealth at least $n - K$ times when gambling on this sequence against even odds. Cover's scheme is universal in the sense of Problem 1, since Levin and Zhvonkin (1970) proved that $K(J^n)/n \to H(J|J^-)$ a.s. for stationary ergodic $\{J_t\}$. However, this scheme is not computable, since it requires evaluation of the Kolmogorov–Chaitin complexity of finite sequences. To be practical, a gambling scheme must be a recursive function with low complexity.

Gambling is related to prediction. Recall that the conditional distribution of the random variable $J = J_0$ given the t-past $J^{-t} = (J_{-t}, \dots, J_{-1})$ converges to the conditional distribution of J given the infinite past $J^- = (\dots, J_{-2}, J_{-1})$:

$$P(j|J^{-t}) \to P(j|J^-) \quad \text{a.s. for all } j,\ 1 \leq j \leq m. \tag{5}$$

A prediction scheme is a computable function $\hat{P}(j|j^{-t})$ of sequences $j^{-t} = (j_{-t}, \dots, j_{-1})$ and elements j in the finite set $\{1, \dots, m\}$. The following problem appears in a list of eight significant problems posed by Cover (1975) at the Moscow Information Theory Workshop.

Problem 2 (Existence of a universal prediction scheme). Find a prediction scheme that is universal in the sense that for any stationary ergodic random process $\{J_t\}$ with distribution P on the m-ary sequence space, we have

$$\hat{P}(j|J^{-t}) \to P(j|J^-) \quad \text{a.s. for all } j,\ 1 \leq j \leq m. \tag{6}$$

Thus a universal prediction scheme for stationary ergodic processes is an algorithm that generates estimates $\hat{P}(\cdot|J^{-t})$ of the conditional distribution of $J = J_0$ based on a finite but growing number of past observations, so that the estimates converge almost surely to the true conditional distribution of J given the infinite past.

A solution to Problem 2 has been obtained by Ornstein (1978). In Section 5 we review Ornstein's universal prediction scheme and formulate a generalized scheme to learn the conditional distribution of $X = X_0$ given the infinite past $X^- = (\dots, X_{-2}, X_{-1})$ of any stationary ergodic process $\{X_t\}$ with values in a Polish space $\mathscr{X}$.

If $\hat{P}(j|j^{-t})$ is a universal prediction scheme and $\{J_t\}$ is stationary ergodic, then Breiman's generalized ergodic theorem (cf. Lemma 1 below) implies that

$$\frac{1}{n} \sum_{0 \leq t < n} \left[\hat{P}(J_t|J^t) - P(J_t|J^t)\right] \to 0 \quad \text{a.s.}$$

A variant of Problem 2, also posed by Cover (1974), is to construct a scheme $\hat{P}(j_t|j^t)$ that satisfies the stronger property

$$\left[\hat{P}(J_t|J^t) - P(J_t|J^t)\right] \to 0 \quad \text{a.s.}$$

Bailey (1976) proved that no such scheme exists. However, empirical frequency counting can be used to learn the conditional distribution of future outcomes given the past if the process is known to be kth-order Markov, for some fixed k.

Bailey (1976) also used Ornstein's universal prediction scheme to formulate a computable gambling scheme which he claimed to be universal. But his proof is based on an invalid assumption, and the question whether his scheme is universal remains open. In Section 2 we discuss ways to avoid the difficulty with Bailey's approach. In fact, we shall construct a universal gambling scheme without the help of a universal prediction scheme.

In Section 3 we discuss the relation between universal gambling schemes and the universal modeling schemes of Rissanen and Langdon (1981). These authors were mainly interested in modeling for the purpose of compression of individual sequences of finite length, and they did not emphasize the asymptotic optimality of their schemes for stationary ergodic ensembles of random sequences. Langdon (1983) and Rissanen (1983) interpreted the universal data compression algorithm of Ziv and Lempel (1978) as a universal modeling scheme, and Feder (1991) pointed out that this scheme can also be regarded as a universal gambling scheme.

In Section 4 we discuss investment in the stock market. We follow Algoet and Cover (1988a) and describe the market by random vectors $X_t = (X_t^j)_{1 \le j \le m}$, where $X_t^j \ge 0$ is the factor by which capital invested in stock j will grow during period t. The investor is allowed to diversify his capital at the beginning of every period t according to a portfolio vector $b_t = (b_t^j)_{1 \le j \le m}$ in the unit simplex. The capital will grow by a factor $(b_t, X_t) = \sum_{1 \le j \le m} b_t^j X_t^j$ equal to the weighted average of the growth factors of the individual stocks. The objective is to select nonanticipating portfolios b_t depending on the t-past $X^t = (X_0, \dots, X_{t-1})$ so as to maximize the growth rate of the compounded capital

$$S_n = \prod_{0 \le t < n} (b_t, X_t). \tag{7}$$

The portfolio selection problem was first considered by Breiman (1961), after Kelly (1956) considered the special case of gambling. A Kelly horse race is a special type of market because exactly one horse will win and yield a nonzero return, so that the return vector X_t will be oriented along a coordinate axis of $\mathbb{R}_+^m$. For a general market, the return of several stocks may be nonzero. If the market distribution is known, then the maximum growth exponent is attained by maximizing the conditional expected growth exponent $E\{\log(b_t, X_t)|X^t\}$ at each step. A portfolio b_t^* which attains the maximum is called log-optimum.

Let the compounded capital after n periods of investment according to the log-optimum strategy $\{b_t^*\}$ and some alternative nonanticipating strategy $\{b_t\}$ be denoted by

$$S_n^* = \prod_{0 \le t < n} (b_t^*, X_t), \qquad S_n = \prod_{0 \le t < n} (b_t, X_t).$$

Then S_n/S_n^* is a nonnegative supermartingale with respect to the information fields $\sigma(X^t)$. Since $S_0 = 1 = S_0^*$ by convention, we see that $E\{S_n/S_n^*\} \le 1$ for all n and consequently (by Lemma 2 below) that

$$\limsup_n \frac{1}{n} \log\left(\frac{S_n}{S_n^*} \right) \le 0 \quad \text{a.s.} \tag{8}$$

If the market process $\{X_t\}$ is stationary ergodic, then S_n^* will grow exponentially fast almost surely with limiting rate equal to the maximum growth exponent given the infinite past:

$$\frac{1}{n} \log S_n^* \to W(X|X^-) = E\{\log(\bar{b}^*, X)\} \quad \text{a.s.,} \tag{9}$$

where $\bar{b}^*$ maximizes $E\{\log(b, X)|X^-\}$.

Note that b_t^* can be computed only if the conditional distribution $P(dx_t|X^t)$ of X_t given the t-past is known. Of course, it is not realistic to assume that we know the distribution of the market process. In Section 4 we solve the following problem.

Problem 3 (Existence of a universal portfolio selection scheme). Find a nonanticipating portfolio selection scheme $\{\hat{b}_t\}$ such that for any stationary ergodic market process $\{X_t\}$, the compounded capital $\hat{S}_n = \prod_{0 \le t < n}(\hat{b}_t, X_t)$ will grow exponentially fast almost surely with the same maximum rate as under the log-optimum strategy $\{b_t^*\}$, that is,

$$\lim_n \frac{1}{n} \log \hat{S}_n = W(X|X^-) = \lim_n \frac{1}{n} \log S_n^* \quad \text{a.s.} \tag{10}$$

We need the following result, which was used by Breiman (1957) to prove what he called the individual ergodic theorem of information theory.

Lemma 1 (Breiman's generalized ergodic theorem). *Let $\{g_t\}$ be a sequence of real-valued random variables defined on a stationary ergodic dynamical system $(\Omega, \mathscr{F}, P, T)$. If $g_t \to g$ a.s. and $\{g_t\}$ is L^1-dominated ($E\{\sup_t |g_t|\} < \infty$), then*

$$\frac{1}{n} \sum_{0 \le t < n} g_t(T^t \omega) \to E\{g\} \quad a.s. \tag{11}$$

In fact, if $E\{\inf_t g_t\} > -\infty$, then

$$\liminf_n \frac{1}{n} \sum_{0 \le t < n} g_t(T^t \omega) \ge E\left\{ \liminf_t g_t \right\} \quad a.s. \tag{12}$$

A different approach to prove the Shannon–McMillan–Breiman theorem was developed by Algoet and Cover (1988a, b). The foundation of our approach was the so-called asymptotic optimality principle (AOP) for log-optimum selections in convex families of random variables. This principle, which is also essential for our present purposes, can be abstractly formulated as follows. Suppose for all $n \geq 1$ that an element S_n must be selected in a convex family $\mathscr{S}_n$ of nonnegative random variables defined on a probability space $(\Omega, \mathscr{F}, P)$. Typically S_n will grow (or decrease) exponentially fast, and we wish to maximize the asymptotic growth rate $\liminf_n (1/n) \log S_n$. An element S_n^* is called log-optimum in $\mathscr{S}_n$ if S_n^* attains the maximum expected growth exponent $\sup_{S_n \in \mathscr{S}_n} E\{\log S_n\}$. Bell and Cover (1980, 1988) proved that log-optimum selections are uniquely characterized up to almost sure equivalence by the Kuhn–Tucker conditions

$$E\left\{\frac{S_n}{S_n^*}\right\} \leq 1, \quad \text{for all } S_n \in \mathscr{S}_n.$$

The log-optimum selections S_n^* in the convex families $\mathscr{S}_n$ are optimum in asymptotic growth rate by the following:

Lemma 2 (Algoet and Cover). *Let $\{S_n\}$ and $\{S_n^*\}$ be sequences of positive random variables such that $E\{S_n/S_n^*\} \leq 1$ for all n. Then*

$$\limsup_n \frac{1}{n} \log\left(\frac{S_n}{S_n^*}\right) \leq 0 \quad a.s., \tag{13}$$

and consequently

$$\limsup_n \frac{1}{n} \log S_n \leq \limsup_n \frac{1}{n} \log S_n^* \quad a.s., \tag{14}$$

$$\liminf_n \frac{1}{n} \log S_n \leq \liminf_n \frac{1}{n} \log S_n^* \quad a.s. \tag{15}$$

Proof. For any $\varepsilon > 0$, the Markov inequality asserts that

$$P\left\{\frac{1}{n} \log\left(\frac{S_n}{S_n^*}\right) \geq \varepsilon\right\} = P\left\{\frac{S_n}{S_n^*} \geq e^{n\varepsilon}\right\} \leq e^{-n\varepsilon} E\left\{\frac{S_n}{S_n^*}\right\} \leq e^{-n\varepsilon}$$

and the Borel–Cantelli lemma implies that

$$P\left\{\frac{1}{n} \log\left(\frac{S_n}{S_n^*}\right) \geq \varepsilon \text{ infinitely often}\right\} = 0.$$

This proves (13), and (14) follows by adding $\limsup_n (1/n) \log S_n^*$ to both sides. Replacing S_n and S_n^* by $1/S_n^*$ and $1/S_n$ proves (15). □

Lemma 2 holds in particular if S_n/S_n^* is a nonnegative supermartingale with initial value $S_0/S_n^* = 1$. In the case of gambling, $\mathscr{S}_n$ is the family of random variables of the form $S_n = [m^n Q(J^n)]$ where $Q(j^n)$ is an arbitrary probability or sub-probability mass function. The log-optimum selection in $\mathscr{S}_n$ is equal to $[m^n P(J^n)]$ where $P(j^n)$ is the true probability mass function of J^n.

2. Universal gambling schemes. Let $\{J_t\}$ be a random process with distribution P on the m-ary sequence space. No gambling strategy $Q(j_t|J^t)$ can do better in the long run than proportional betting according to the true conditional distribution $P(j_t|J^t)$. Indeed, $Q(J^n)/P(J^n)$ is a nonnegative supermartingale with respect to $\sigma(J^n)$ since

$$\begin{aligned} E\left\{ \frac{Q(J_t|J^t)}{P(J_t|J^t)} \middle| J^t \right\} &= \sum_{\{j_t:\, P(j_t|J^t)>0\}} \frac{Q(j_t|J^t)}{P(j_t|J^t)} P(j_t|J^t) \\ &= \sum_{\{j_t:\, P(j_t|J^t)>0\}} Q(j_t|J^t) \le 1. \end{aligned}$$

The initial value of this supermartingale is equal to 1 since J^0 is the empty sequence and $Q(J^0) = 1 = P(J^0)$ by convention. Lemma 2 therefore implies that

$$\limsup_n \frac{1}{n} \log\left(\frac{[m^n Q(J^n)]}{[m^n P(J^n)]} \right) \le 0 \quad \text{a.s.}$$

If $\{J_t\}$ is stationary ergodic, then by the Shannon–McMillan–Breiman theorem,

$$\begin{aligned} \limsup_n \frac{1}{n} \log[m^n Q(J^n)] &\le [\log m - H(J|J^-)] \\ &= \lim_n \frac{1}{n} \log[m^n P(J^n)] \quad \text{a.s.} \end{aligned} \tag{16}$$

A gambling scheme $Q(j_t|j^t)$ is called universal if for any stationary ergodic process $\{J_t\}$,

$$\lim_n \frac{1}{n} [m^n Q(J^n)] = [\log m - H(J|J^-)] \quad \text{a.s.} \tag{17}$$

The Shannon–McMillan–Breiman theorem for stationary ergodic $\{J_t\}$ asserts that

$$\frac{1}{n} \sum_{0 \le t < n} g_t \circ T^t \to E\{g\} \quad \text{a.s.,}$$

where

$$g_t = -\log P(J|J^{-t}), \qquad g = -\log P(J|J^-).$$

Breiman (1957) derived this theorem from the generalized ergodic theorem that is stated in Lemma 1. Clearly $g_t \to g$ a.s. by the martingale convergence theorem for conditional probabilities. The difficult part was to verify the integrability condition $E\{\sup_t g_t\} < \infty$.

Bailey (1976) observed that a universal gambling scheme can be obtained from Ornstein's universal prediction scheme. In fact, if $\hat{P}(j|j^{-t})$ is any universal prediction scheme [such that $P(J|J^{-t}) > 0$ a.s.], then

$$\hat{g}_t = -\log \hat{P}(J|J^{-t}) \to g = -\log P(J|J^-) \quad \text{a.s.}$$

To prove that a universal prediction scheme $\hat{P}(j|j^{-t})$ yields a universal gambling scheme by shifting, one must argue for any stationary ergodic random sequence $\{J_t\}$ that

$$\frac{1}{n} \sum_{0 \le t < n} \hat{g}_t \circ T^t \to E\{g\} \quad \text{a.s.} \tag{18}$$

Since $\hat{g}_t \geq 0$ and $\hat{g}_t \to g$ a.s., Lemma 1 yields

$$\liminf_n \frac{1}{n} \sum_{0 \le t < n} \hat{g}_t \circ T^t \geq E\{g\} \quad \text{a.s.,}$$

which is equivalent to (16). To prove (18) and hence (17), we must verify that $\{\hat{g}_t\}$ is L^1-dominated. Bailey's (1976) proof is based on an invalid assumption, namely Lemma 5.3 on page 47. In fact, Breiman's proof of the integrability condition $E\{\sup_t g_t\} < \infty$ was quite delicate, and it remains a challenge to construct estimates $\hat{g}_t$ such that $E\{\sup_t \hat{g}_t\} < \infty$ and $\hat{g}_t \to g$ a.s. Empirical estimates of the probability of rare events are very unreliable, and taking logarithms amplifies the problem. We shall use a variation of Bailey's approach to avoid the integrability condition altogether.

For any betting scheme Q and $0 \leq \lambda < 1$, we define the betting scheme Q^λ by

$$Q^\lambda(j) = (1 - \lambda)\frac{1}{m} + \lambda Q(j), \qquad 1 \leq j \leq m. \tag{19}$$

If $Q(j_t|j^t)$ is a gambling scheme, then the gambling scheme $Q^\lambda(j_t|j^t)$ places at least a portion $(1 - \lambda)/m$ on every stock, so that the gambler will never go broke. The compounded growth factor of wealth over n rounds amounts to

$$\left[m^n Q^\lambda(J^n)\right] = \prod_{0 \le t < n} \left[m Q^\lambda(J_t|J^t)\right]. \tag{20}$$

THEOREM 1. *Let $\hat{P}(j|j^{-t})$ be a universal prediction scheme such as Ornstein's. If $\{J_t\}$ is stationary ergodic with distribution P on the m-ary sequence space, then the growth exponent of compounded wealth for the gambling scheme $\hat{P}^\lambda(j_t|j^t)$ is given by*

$$\lim_n \frac{1}{n} \log\left[m^n \hat{P}^\lambda(J^n)\right] = \left[\log m - H^\lambda(J|J^-)\right] \quad a.s., \tag{21}$$

where

$$H^\lambda(J|J^-) = E\{-\log P^\lambda(J|J^-)\}. \tag{22}$$

Proof. Consider the random variables

$$\hat{g}_t^\lambda = -\log \hat{P}^\lambda(J|J^{-t}), \qquad g^\lambda = -\log P^\lambda(J|J^-).$$

One must prove that

$$\frac{1}{n} \sum_{0 \le t < n} \hat{g}_t^\lambda \circ T^t \to E\{g^\lambda\} \quad \text{a.s.}$$

This follows from Breiman's extended ergodic theorem since g_t^λ is bounded (between 0 and $\log[m/(1-\lambda)]$) and $\hat{g}_t^\lambda \to g^\lambda$ a.s. □

Note that $H^\lambda(J|J^-)$ approaches $H(J|J^-)$ as λ approaches 1 since

$$H(J|J^-) - \log\lambda \ge H^\lambda(J|J^-) \ge H(J|J^-), \qquad 0 \le \lambda < 1. \tag{23}$$

Thus the maximum growth rate is asymptotically approached to within $\varepsilon = \log(1/\lambda)$ by the scheme $\hat{P}^\lambda(j_t|j^t)$. We now prove that the maximum growth rate can be asymptotically attained.

Theorem 2. *There exists a gambling scheme $\hat{P}(j_t|j^t)$ that is universal in the sense that for any stationary ergodic random process $\{J_t\}$, we have*

$$\begin{aligned} \frac{1}{n}\log\left[m^n\hat{P}(J^n)\right] &\to [\log m - H(J|J^-)] \\ &= \lim_n \frac{1}{n}\log[m^n P(J^n)] \quad a.s. \end{aligned} \tag{24}$$

Proof. We define a universal gambling scheme $\hat{P}(j_t|j^t)$ in terms of a universal prediction scheme $Q(j|j^{-t})$ as follows. The gambler divides his initial wealth into countably many piles indexed by k, and he manages the money in the kth pile according to the strategy $Q^{\lambda_k}(j_t|J^t)$. The money in the kth pile will grow exponentially fast almost surely with limiting rate $[\log m - H^{\lambda_k}(J|J^-)]$, by Theorem 1. The limiting growth rate of the total in all piles is no smaller than that of the kth pile, and is no larger than $[\log m - H(J|J^-)]$. We make sure that $\lambda_k \nearrow 1$, so that the total in all piles will grow with limiting rate

$$[\log m - H(J|J^-)] = \sup_k\left[\log m - H^{\lambda_k}(J|J^-)\right].$$

The return of all piles is pooled at the end of every round, but the gambler can keep track of the amount in every pile by keeping records of the past. If $\mu_k > 0$ denotes the initial allocation to the kth pile ($\sum_k \mu_k = 1$), then the total wealth after n rounds is given by

$$\left[m^n\hat{P}(J^n)\right] = \sum_k \mu_k\left[m^n Q^{\lambda_k}(J^n)\right].$$

This is achieved by the nonanticipating strategy

$$\hat{P}(j_t|J^t) = \frac{\Sigma_k \mu_k [m^t Q^{\lambda_k}(J^t)] Q^{\lambda_k}(j_t|J^t)}{\Sigma_k \mu_k [m^t Q^{\lambda_k}(J^t)]}. \qquad \square$$

REMARK 1. We can construct a universal gambling scheme that does not use a universal prediction scheme. We compute an empirical estimate of the conditional probability $P(j|J^{-k})$ on the basis of J^{-t-k}, namely

$$\hat{P}_t(j|J^{-k}) = \frac{\delta_{j_0}(j) + c_t(j|J^{-k})}{1 + c_t(J^{-k})},$$

where $\delta_{j_0}(\cdot)$ is the indicator function of some fixed $j_0 \in \{1, \dots, m\}$ and

$$\begin{aligned} c_t(j|J^{-k}) &= \#\{\tau: 1 \le \tau \le t, (J_{-\tau-k}, \dots, J_{-\tau-1}, J_{-\tau}) = (J_{-k}, \dots, J_{-1}, j)\}, \\ c_t(J^{-k}) &= \#\{\tau: 1 \le \tau \le t, (J_{-\tau-k}, \dots, J_{-\tau-1}) = (J_{-k}, \dots, J_{-1})\} \\ &= \sum_{1 \le j \le m} c_t(j|J^{-k}). \end{aligned}$$

Thus $\hat{P}_t(\cdot|J^{-k})$ is the empirical distribution of the symbols that follow past occurrences of the block J^{-k} within the longer block J^{-t-k}, when some symbol j_0 is given one tally a priori. The ergodic theorem implies that $\hat{P}_t(j|J^{-k})$ is a consistent estimate of the conditional probability $P(j|J^{-k})$, that is,

$$\hat{P}_t(j|J^{-k}) \to P(j|J^{-k}) \quad \text{a.s. as } t \to \infty.$$

Shifting $\hat{P}_t(j|J^{-k})$ yields an empirical estimate $\hat{P}_t(j_t|J_{t-k}, \dots, J_{t-1})$ of the conditional distribution $P(j_t|J_{t-k}, \dots, J_{t-1})$ that is computed on the basis of J^t and some arbitrary choice for J^{-k}. The growth rate of wealth when betting according to strategy $\hat{P}_t^{\lambda}(j_t|J_{t-k}, \dots, J_{t-1})$ is equal to

$$(25) \qquad \begin{aligned} &\lim_n \frac{1}{n} \log\left(\prod_{0 \le t < n} \left[m\hat{P}_t^{\lambda}(J_t|J_{t-k}, \dots, J_{t-1}) \right] \right) \\ &\quad = \left[\log m - H^{\lambda}(J|J^{-k}) \right] \quad \text{a.s.}, \end{aligned}$$

where

$$(26) \qquad H^{\lambda}(J|J^{-k}) = E\{-\log P^{\lambda}(J|J^{-k})\}.$$

Observe that $H^{\lambda}(J|J^{-k})$ decreases to $H(J|J^{-})$ as $k \to \infty$ and $\lambda \nearrow 1$. One obtains a universal gambling scheme by dividing the initial wealth into countably many piles indexed by k and using strategy $\hat{P}_t^{\lambda_k}(j_t|J^t)$ for the kth pile, where $\lambda_k \nearrow 1$.

REMARK 2. So far we have assumed that the gambler is paid out at uniform odds, so that the return is m times the amount wagered on the winning outcome. The results generalize if the odds are derived from an arbitrary reference measure μ on $\{1, \dots, m\}$, that is, if the return when outcome j is realized is μ_j times the amount wagered on j. The optimum nonanticipating strategy is to bet a portion on every possible outcome j proportional to its conditional win probability $P\{J_t = j|J^t\}$, as before. The maximum growth rate of compounded wealth is now given by the relative entropy rate or Kullback–Leibler information divergence rate

$$I_\mu(J|J^-) = E\left\{\log\left(\frac{P(J|J^-)}{\mu(J)}\right)\right\}. \tag{27}$$

If μ is the uniform distribution that assigns mass $1/m$ to every possible outcome j, then the odds are uniform and $I_\mu(J|J^-) = [\log m - H(J|J^-)]$.

3. Universal modeling and data compression. Gambling schemes are related to the modeling schemes of Rissanen and Langdon (1981). These authors were primarily interested in universal modeling of information sources for the purpose of data compression via arithmetic coding.

3.1. *Modeling schemes.* A modeling scheme is an assignment, for all $n \geq 1$, of numbers $Q(j^n)$ between 0 and 1 to all sequences $j^n = (j_0, \dots, j_{n-1})$ of elements in $\{1, \dots, m\}$. The number $Q(j^n)$ is called the code space assigned to the sequence j^n. The total amount of code space assigned to the sequences of a given length must fit in a unit interval, so that $Q(j^n)$ is a subprobability measure on n-tuples:

$$\sum_{j^n} Q(j^n) \leq 1. \tag{28}$$

Modeling is commonly done in the context of a computable structure function $f(j^t)$ that summarizes the past j^t. One poses that $Q(j^n) = \prod_{0 \leq t < n} Q(j_t|j^t)$, where the conditional probability $Q(j_t|j^t)$ depends on j^t only through the context or conditioning class $z_t = f(j^t)$. Thus

$$Q(j^n) = \prod_{0 \leq t < n} Q(j_t|z_t), \quad \text{where } z_t = f(j^t),\ 1 \leq j \leq m. \tag{29}$$

Many interesting structure functions are defined by finite automata with input set $\{1, \dots, m\}$. Given an automaton (S, δ, g, s_0) with finite state set S, initial state $s_0 \in S$, transition function $\delta(s, j)$ and output function $g(s)$, one defines the structure function

$$z_n = f(j^n) = g(s_n), \quad \text{where } s_{t+1} = \delta(s_t, j_t),\ 0 \leq t < n. \tag{30}$$

In particular, if z_t is the k-past, then the finite automaton can be implemented as a shift register of length k with state $s_t = z_t = (j_{t-k}, \dots, j_{t-1})$.

The optimum choice for the structure function f is an undecidable problem. However, the model defined by a particular structure function f expresses $Q(j^n)$ in terms of conditional probabilities $Q(j|z)$ as

$$Q(j^n) = \prod_{0 \leq t < n} Q(j_t|z_t) = \prod_{1 \leq j \leq m} \prod_z Q(j|z)^{c_n(j|z)}, \tag{31}$$

where $c_n(j|z)$ counts how often the symbol j has occurred within the context of the conditioning class z:

$$c_n(j|z) = \#\{t: 0 \le t < n, (j_t, z_t) = (j, z)\}. \tag{32}$$

Let the total number of occurrences of the conditioning class z be denoted by

$$c_n(z) = \sum_{1 \le j \le m} c_n(j|z) = \#\{t: 0 \le t < n, z_t = z\}. \tag{33}$$

The following result of Rissanen and Langdon (1981) generalizes Bartlett's (1951) characterization of the maximum likelihood estimates of the transition probabilities of a Markov chain by means of empirical distributions.

THEOREM 3. *Consider a structure function f and the modeling scheme $Q(j^n) = \prod_{0 \le t < n} Q(j_t|z_t)$, where $z_t = f(j^t)$. To minimize the codeword length $-\log Q(j^n)$ or equivalently to maximize the likelihood $Q(j^n)$, one must choose $Q(j|z)$ equal to the empirical conditional distribution*

$$\tilde{P}_n(j|z) = \frac{c_n(j|z)}{c_n(z)}, \qquad 1 \le j \le m. \tag{34}$$

The minimum per-symbol codeword length is equal to the conditional entropy of J given Z when the joint distribution of (J, Z) is the empirical distribution of the pairs $\{(j_t, z_t): 0 \le t < n\}$. Indeed, if $\tilde{P}_n(z) = c_n(z)/n$, then

$$-\frac{1}{n} \log \tilde{P}_n(J^n) = \tilde{H}^{(n)}(J|Z) = -\sum_z \tilde{P}_n(z) \sum_{1 \le j \le m} \tilde{P}_n(j|z) \log \tilde{P}_n(j|z). \tag{35}$$

The excess per-symbol codeword length for any alternative $Q(j|z)$ is equal to the conditional relative entropy

$$\begin{aligned} -\frac{1}{n} \log Q(J^n) + \frac{1}{n} \log \tilde{P}_n(J^n) &= \tilde{I}_Q^{(n)}(J|Z) \\ &= \sum_z \tilde{P}_n(z) \sum_{1 \le j \le m} \tilde{P}_n(j|z) \log\left(\frac{\tilde{P}_n(j|z)}{Q(j|z)}\right). \end{aligned} \tag{36}$$

Any modeling scheme defines a sequence of uniquely decipherable binary codes. The code for blocks of length n is characterized by the Shannon–Fano codeword length function

$$l(j^n) = \lceil -\log_2 Q(j^n) \rceil. \tag{37}$$

Conversely, if an assignment $l(j^n)$ of nonnegative integers to n-tuples j^n satisfies the Kraft inequality

$$\sum_{j^n} 2^{-l(j^n)} \le 1, \tag{38}$$

then there exists a uniquely decipherable binary code with length function $l(j^n)$, and $Q(j^n) = 2^{-l(j^n)}$ is a modeling scheme for n-tuples.

A modeling scheme is considered attractive for the purpose of coding or compression of a finite sequence j^n if the assigned codeword length $-\log Q(j^n)$ is small. For a random sequence $\{J_t\}$ with distribution P, it makes sense to compare the assigned codeword length $-\log Q(J^n)$ to the ideal codeword length $-\log P(J^n)$.

THEOREM 4. *If $Q(j^n)$ is any modeling scheme and $\{J_t\}$ is any random process with distribution P, then*

$$\limsup_n \frac{1}{n} \log\left(\frac{Q(J^n)}{P(J^n)}\right) \leq 0 \quad a.s. \tag{39}$$

In particular, if $\{J_t\}$ is stationary ergodic, then

$$\liminf_n -\frac{1}{n}\log Q(J^n) \geq H(J|J^-) = \lim_n -\frac{1}{n}\log P(J^n) \quad a.s. \tag{40}$$

PROOF. This result is an immediate consequence of Lemma 2, since

$$E\left\{\frac{Q(J^n)}{P(J^n)}\right\} = \sum_{\{j^n:\, P(j^n)>0\}} \frac{Q(j^n)}{P(j^n)}P(j^n) = \sum_{\{j^n:\, P(j^n)>0\}} Q(j^n) \leq 1. \quad \square$$

Theorem 4 immediately implies that the entropy rate is almost surely a lower bound on the noiseless compressibility of a stationary ergodic random sequence. Barron (1985) was the first to derive this strong lower bound for noiseless coding from the Kraft inequality, with knowledge of the asymptotic optimality principle for log-optimum gambling of Algoet and Cover (1988a).

COROLLARY 1. *For $n \geq 1$ let $l(j^n)$ denote the length function of a uniquely decipherable block-to-variable-length binary code for blocks of length n. Then for any random process $\{J_t\}$ we have*

$$\liminf_n \left[\frac{1}{n}l(J^n) + \frac{1}{n}\log P(J^n)\right] \geq 0 \quad a.s. \tag{41}$$

In particular, if $\{J_t\}$ is stationary ergodic, then

$$\liminf_n \frac{1}{n}l(J^n) \geq H(J|J^-) = \lim_n -\frac{1}{n}\log P(J^n) \quad a.s. \tag{42}$$

PROOF. Set $Q(j^n) = 2^{-l(j^n)}$ and apply the preceding theorem. $\square$

A modeling scheme $Q(j^n)$ is called universal if for any stationary ergodic random sequence $\{J_t\}$, the assigned codeword length is equal to the ideal codeword length in the long run average or Cesàro mean sense, that is,

$$-\frac{1}{n}\log Q(J^n) \to H(J|J^-) = \lim_n -\frac{1}{n}\log P(J^n) \quad \text{a.s.} \tag{43}$$

THEOREM 5. *There exists a universal modeling scheme.*

PROOF. Let $Q_k(j^n)$ denote the scheme that models the sequence j^n as being kth-order Markov with stationary transition probabilities. Thus $Q_k(j^n)$ is the scheme whose structure function is defined by the finite automaton that keeps track of the k-past [and makes arbitrary assumptions about $j^{-k} = (j_{-k}, \dots, j_{-1})$]. If $\{J_t\}$ is ergodic, then clearly

$$-\frac{1}{n}\log Q_k(J^n) \to H(J|J^{-k}) \quad \text{a.s.}$$

By mixing the modeling schemes Q_k according to a priori weights $\mu_k > 0$ (where $\sum_{k \geq 0}\mu_k = 1$), one obtains the universal modeling scheme

$$Q(j^n) = \sum_{k\geq 0} [\mu_k Q_k(j^n)]. \qquad \square$$

Any universal gambling scheme $Q(j_t|j^t)$ yields a universal modeling scheme, and so does any universal data compression algorithm, including that of Ziv and Lempel (1978) and the recent scheme of Ornstein and Shields (1990). However, the proof that these schemes are universal is not as elementary as that for the scheme in Theorem 5.

3.2. *Gambling and arithmetic coding.* Any gambling scheme $Q(j_t|j^t)$ yields by compounding a modeling scheme

$$Q(j^n) = \prod_{0\leq t<n} Q(j_t|j^t). \tag{44}$$

This modeling scheme is special because the code space assignments to sequences of different lengths are compatible in the sense that

$$Q(j_0, \dots, j_{n-1}) = \sum_{1\leq j_n\leq m} Q(j_0, \dots, j_{n-1}, j_n). \tag{45}$$

In fact, any gambling scheme $Q(j_t|j^t)$ defines a probability measure Q on the m-ary sequence space, such that $Q(j^n)$ is the probability of a cylinder set. Conversely, any computable probability measure Q defines a gambling scheme

$$Q(j_t|j^t) = \frac{Q(j_0, \dots, j_{t-1}, j_t)}{Q(j_0, \dots, j_{t-1})}. \tag{46}$$

Note that $Q(j_t|j^t)$ is arbitrary if $Q(j^t) = 0$.

Observe that minimizing the description length $-\log Q(j^n)$ is equivalent to maximizing the likelihood $Q(j^n)$ or the compounded wealth $[m^n Q(j^n)]$ when gambling on the symbols j_t. Every bit of compression yields an extra factor 2 in capital growth.

Sometimes a gambler may wish to set aside a portion of his wealth, rather than bet everything. In this case $Q(j_t|j^t)$ is a conditional subprobability measure on the symbols j_t:

$$\sum_{1 \le j_t \le m} Q(j_t | j^t) \le 1. \tag{47}$$

In particular, a portion $[1 - \sum_{1 \le j_0 \le m} Q(j_0)]$ of the initial wealth $Q(j^0) = 1$ is never invested.

Let $\{J_t\}$ be a stationary ergodic process with values in $\{1, \dots, m\}$, and suppose modeling is done in the context of a structure function $z_n = f(J^n)$. The empirical conditional distribution $\tilde{P}(j | z_n)$ is not a satisfactory betting strategy since the gambler will go broke if the next symbol j_n happens to be the first occurrence of that symbol in the context $z = z_n$. A strategy that is asymptotically equivalent to $\tilde{P}(j | z_n)$ but that avoids going broke is defined by

$$\hat{P}(j | z_n) = \frac{c_n(j | z_n) + 1}{c_n(z_n) + m}, \qquad 1 \le j \le m. \tag{48}$$

The rationale for this strategy is as follows. One must select a distribution in the unit simplex, for betting on J_n within the context $z_n = f(J^n)$, having seen $c_n(j | z_n)$ occurrences of the symbol j in this context. If the prior distribution is uniform on the simplex of probability measures on $\{1, \dots, m\}$, then the posterior distribution after observing all these occurrences in the context z_n is a β-distribution with parameters $\{c_n(j | z_n),\ 1 \le j \le m\}$. The mean of this posterior β-distribution is exactly equal to the distribution $\hat{P}(\cdot | z_n)$ on $\{1, \dots, m\}$. The strategy $\hat{P}(\cdot | z_n)$ is identical to the strategy $\tilde{P}(\cdot | z_n)$, except that every symbol j is given an a priori count of 1 in every context z.

Any gambling scheme yields a modeling scheme that can be used for data compression. The asymptotic lower bound in Theorem 4 holds for any modeling scheme, and Corollary 1 is valid for any sequence of uniquely decipherable codes. In a realistic system, the codes must be compatible in the sense that the code for a prefix of a sequence must be a prefix of the code for that sequence. Also, the encoder must be able to generate the output bits eventually, given a sufficient amount of lookahead in the input sequence. If a modeling scheme is derived from a gambling scheme, then the encoding can be performed incrementally by arithmetic coding.

The original concept of arithmetic coding is due to Elias [cf. pages 476–489 of Jelinek (1968)]. The idealized encoding algorithm may require infinite precision arithmetic and unbounded lookahead. In practice, the computations are done with finite precision using a finite state automaton [cf. Pasco (1976) or Rissanen and Langdon (1979)].

An ideal arithmetic coding unit accumulates the total probability of sequences that are lexicographically smaller than or equal to the random sequence being encoded. Let $Q(j_t | j^t)$ be a universal gambling scheme and let

$$\underline{C}_n = \sum_{j^n < J^n} Q(j^n) = \sum_{0 \le t < n} \sum_{1 \le j < J_t} Q(J_0, \dots, J_{t-1}, j), \tag{49}$$

$$\overline{C}_n = \sum_{j^n \le J^n} Q(j^n) = \underline{C}_n + Q(J^n). \tag{50}$$

Note that $\underline{C}_n$ and $\overline{C}_n$ can be computed recursively as

$$\underline{C}_0 = 0, \qquad \underline{C}_{n+1} = \underline{C}_n + Q(J^n) \sum_{1 \le j < J_n} Q(j|J^n), \tag{51}$$

$$\overline{C}_0 = 1, \qquad \overline{C}_{n+1} = \overline{C}_n - Q(J^n) \sum_{J_n < j \le m} Q(j|J^n). \tag{52}$$

Clearly, $\underline{C}_n$ increases to the cumulative probability of infinite sequences that are lexicographically smaller than the actual sequence $\{J_t\}$, and $\overline{C}_n$ decreases to the cumulative probability of infinite sequences that are lexicographically smaller than or equal to $\{J_t\}$. The digits in the binary expansion of $\Gamma = \lim_n \underline{C}_n$ are taken as encoding of the sequence $\{J_t\}$. The limit Γ is always contained in the interval $[\underline{C}_n, \overline{C}_n)$, and the encoder can generate k output bits as soon as this interval is narrowed down to a dyadic interval of length 2^{-k}. This dyadic interval can then be scaled up by a factor 2^k to a unit interval. (An interval is called dyadic if its endpoints are dyadic numbers, i.e., rational numbers whose denominator is a power of 2.) This is done in the following *arithmetic encoding algorithm*:

```
begin
  $Q := 1$; $\underline{C} := 0$; $K := 0$;
  for $n := 0, 1, 2, \ldots$ do (* $Q = 2^K * Q(J^n)$ and $\underline{C} = (2^K * \underline{C}_n) \bmod 1$ *)
    begin
      $Q := Q * Q(J_n|J^n)$;
      $\underline{C} := \underline{C} + Q * \Sigma_{1 \le j < J_n} Q(j|J^n)$; $\overline{C} := \underline{C} + Q$;
      while $\overline{C} < \frac{1}{2}$ do begin output '0';
        (* replace dyadic interval by its lower half *)
        $K := K + 1$; $\overline{C} := 2 * \overline{C}$; $Q := 2 * Q$ end;
      while $\underline{C} \ge \frac{1}{2}$ do begin output '1';
        (* replace dyadic interval by its upper half *)
        $K := K + 1$; $\underline{C} := 2 * \underline{C} - 1$; $Q := 2 * Q$ end;
    end
end.
```

If $Q(j_t|j^t)$ is a universal gambling scheme, then the width $Q(J_n) = \overline{C}_n - \underline{C}_n$ of the interval $[\underline{C}_n, \overline{C}_n)$ decreases exponentially fast almost surely with limiting rate $H(J|J^-)$, so that roughly $nH(J|J^-)$ output bits are generated per n input bits. The encoder is delayed and unable to generate any output bits while the smallest dyadic interval containing $[\underline{C}_n, C_n)$ remains unchanged, or equivalently while the rescaled interval straddles the value $1/2$. The decoder can recover J_n as soon as the number Γ is known with sufficient precision, given enough lookahead.

If Q is replaced by the true distribution P, then the cumulative probability Γ is uniformly distributed over the unit interval $[0, 1]$, at least if the probability of any infinite sequence is 0. With probability 1, Γ will not be a dyadic rational and the encoder will not be delayed forever. But there is no such guarantee for an arbitrary gambling scheme Q.

3.3. *The Ziv–Lempel algorithm.* The incremental parsing algorithm of Ziv and Lempel (1978) is a universal data compression algorithm. Variations of this algorithm have been interpreted as modeling schemes by Langdon (1983) and Rissanen (1983).

The Ziv–Lempel algorithm parses individual sequences J^n into phrases by inserting commas. The first comma is placed in front of the first symbol J_0. Each phrase starts at a comma, and consists of a maximal length sequence that has occurred as an earlier phrase, followed by an innovation symbol and another comma. We denote by ν_n the number of complete phrases when parsing the finite sequence J^n. For example, the binary string $J^n = 0101000100$ with length $n = 10$ is parsed as ,0,1,01,00,010,0 and contains $\nu_n = 5$ complete phrases and an incomplete phrase ,0 at the end.

It is well known that the Ziv–Lempel parsing can be obtained by maintaining a dynamically growing data structure in the form of a tree. Initially this tree consists of a single node, the root. The search for a new phrase starts at the root and proceeds down the tree as directed by the input symbols. The search is complete when a branch is taken from an existing tree node to a new node that has not been visited before. The branch and the new node are added to the tree, a comma is inserted in the input stream and the search for the next node starts over again at the root.

Let T_n denote the tree, to the extent it has grown by the time the algorithm has finished reading J^n. The branches that diverge from every node in T_n are labeled by the symbols $1, \dots, m$, but some nodes may be incomplete in the sense that not all m outgoing branches are present. We define the completed tree $\overline{T}_n$ as the tree that contains all branches leaving all nodes of T_n. Thus every node of T_n is an interior node of $\overline{T}_n$. It is easily shown by induction that T_n contains $\nu_n + 1$ nodes, while $\overline{T}_n$ contains $1 + m(\nu_n + 1)$ nodes, namely $\nu_n + 1$ interior nodes and $1 + (m - 1)(\nu_n + 1)$ leaves. The search for a new phrase ends when a branch is taken from an interior node to a leaf in the completed tree. That leaf is then converted to an interior node and its m children are added to the completed tree.

Let $z_t = f(J^t)$ denote the node in the tree T_t where the search for the next phrase has arrived after J^t but before J_t is read. Langdon (1983) interpreted $f(j^t)$ as a structure function and modeled the next symbol J_t as being selected on the basis of the context or conditioning class $z_t = f(J^t)$. Let $c_n(j|z)$ and $c_n(z)$ be defined as in (32) and (33). Then $c_n(j|z)$ is the number of transitions from node z to the jth child of z. Also, $1 + c_n(z)$ is the number of nodes in the subtree of T_n that is rooted at z, unless z is an ancestor of z_n, in which case $1 + c_n(z)$ is one more than the number of nodes in that subtree. To maximize the likelihood

$$Q(J^n) = \prod_{0 \le t < n} Q(J_t|z_t) = \prod_{1 \le j \le m} \prod_{z \in T_n} Q(j|z)^{c_n(j|z)},$$

it suffices in view of Theorem 3 to choose

$$Q(j|z) = \frac{c_n(j|z)}{c_n(z)}, \qquad 1 \leq j \leq m.$$

Observe that $Q(j|z_n)$ is equal to the ratio of the number of nodes in the subtrees of T_n that are rooted at z_n and the jth child of z_n, when the roots of those subtrees are not counted. The assignment $Q(j|z_n)$ is not a good gambling strategy since the gambler will go broke at the end of a phrase, when the search proceeds from z_n to a new node and hence back to the root.

Rissanen (1983) described an incremental modeling scheme which ensures that all leaves of the completed tree $\overline{T}_n$ are equally likely outcomes of the search for a new phrase. The number of leaves in the subtrees rooted at node z_n and its jth child in the completed tree $\overline{T}_n$ are equal to

$$\begin{aligned} \gamma_n(z_n) &= 1 + (m-1)(c_n(z_n) + 1), \\ \gamma_n(j|z_n) &= 1 + (m-1)(c_n(j|z_n) + 1). \end{aligned} \tag{53}$$

To make all leaves of $\overline{T}_n$ equally likely outcomes, one must assign a probability to each child j of z_n that is proportional to the number of leaves of the completed subtree rooted at this child. Thus the proportion bet on outcome j must be equal to

$$\hat{P}(j|J^n) = \frac{\gamma_n(j|z_n)}{\gamma_n(z_n)} = \frac{m + (m-1)c_n(j|z_n)}{m + (m-1)c_n(z_n)}, \qquad 1 \leq j \leq m. \tag{54}$$

Feder (1991) observed that Rissanen's interpretation of the Ziv–Lempel parsing algorithm yields a universal gambling scheme.

THEOREM 6. *The gambling scheme $\hat{P}(j|J^n)$ is universal.*

PROOF. We argue that $\hat{P}(J^n) = \prod_{0 \leq t < n} \hat{P}(J_t|J^t)$ decreases exponentially fast almost surely with limiting rate $H(J|J^-)$ if $\{J_t\}$ is stationary ergodic. In Corollary 1, we used the Kraft inequality and the asymptotic optimality principle to prove that for any gambling strategy,

$$\liminf_n -\frac{1}{n} \log \hat{P}(J^n) \geq H(J|J^-) \quad \text{a.s.}$$

We prove that $H(J|J^-)$ is not only a lower bound but also an upper bound for the decay rate of $\hat{P}(J^n)$.

Let n_i denote the starting point of the ith phrase $(1 \leq i \leq \nu_n)$. Then

$$\prod_{n_i \leq t < n_{i+1}} \hat{P}(J_t|J^t) = \frac{1}{[1 + (m-1)i]}.$$

The product on the left-hand side telescopes, so that all numerators and denominators cancel except the denominator $[1 + (m-1)i]$ of the first factor $(t = n_i)$ and the numerator 1 of the last factor $(t = n_{i+1} - 1)$. Taking the product over all phrases proves that

$$\frac{1}{\prod_{1\le i\le \nu_n}[1+(m-1)i]} \ge \hat{P}(J^n) \ge \frac{1}{\prod_{1\le i\le \nu_n+1}[1+(m-1)i]}$$

and consequently

$$-\log \hat{P}(J^n) = \log \nu_n! + \mathscr{O}(\nu_n).$$

Wyner and Ziv (1990) have shown [their proof appears in Section 12.10 of Cover and Thomas (1991)] that $\log \nu_n! = \mathscr{O}(\nu_n \log \nu_n)$ and

$$\limsup_n -\frac{1}{n}\log \hat{P}(J^n) = \limsup_n \frac{1}{n}\log \nu_n! \le H(J|J^-) \quad \text{a.s.}$$

Thus we reach the desired conclusion:

$$-\frac{1}{n}\log \hat{P}(J^n) \to H(J|J^-) \quad \text{a.s.} \qquad \square$$

Theorem 6 was proved using different methods by Ornstein and Weiss (1990).

4. Universal portfolio selection schemes.

4.1. *Log-optimum investment in a market with known distribution.* The stock market is modeled by random vectors $X_t = (X_t^j)_{1\le j\le m}$, where $X_t^j \ge 0$ denotes the factor by which capital invested in stock j will grow during period t. We call X_t the return vector for period t. It must be emphasized that X_t^j is not the difference but rather the ratio of the price at the end of investment period t to the price at the beginning of that period. Often, X_t^j is called the price relative of stock j. We assume that the distribution P of market process $\{X_t\}$ is such that $P\{X_t = 0\} = 0$ and hence $X_t \ne 0$ a.s. for all t.

An investor must decide how to distribute his current capital at the beginning of every investment period t over the available opportunities $j = 1, \dots, m$. The allocation of capital is described by a vector of nonnegative weights that sum to 1, that is, by a portfolio vector $b_t = (b_t^j)_{1\le j\le m}$ in the unit simplex $\mathscr{B}$ of $\mathbb{R}_+^m$. Portfolio b_t must be nonanticipating, that is, b_t must be a measurable function of the t-past $X^t = (X_0, \dots, X_{t-1})$ [notation: $b_t \in \sigma(X^t)$]. The capital will grow by a factor $(b_t, X_t) = \Sigma_{1\le j\le m} b_t^j X_t^j$ equal to the weighted average of the returns of the individual stocks, and the total amassed at the end of the period is redistributed at the beginning of the next period. If the investor starts with initial wealth $S_0 = 1$, then the compounded capital after n investment periods amounts to

$$S_n = \prod_{0\le t<n} (b_t, X_t). \tag{55}$$

The objective is to invest according to nonanticipating portfolios that maximize the asymptotic growth rate of the compounded capital S_n.

First, we consider the simple case where the return vectors X_t are independent realizations of a random vector $X = (X^j)_{1 \le j \le m}$ with distribution P on $\mathbb{R}_+^m - \{0\}$. If the investor rebalances his capital at the beginning of every period according to some fixed portfolio b, then the compounded capital will grow exponentially fast almost surely with constant limiting rate $E\{\log(b, X)\}$, by the strong law of large numbers for products. (It must be assumed that the expected log return is well defined.) A portfolio b^* is called log-optimum for P if no alternative can improve upon it in growth exponent:

$$(56) \qquad E\left\{\log\left(\frac{(b, X)}{(b^*, X)}\right)\right\} \le 0 \quad \text{a.s. for all portfolios } b \in \mathscr{B}.$$

Any log-optimum portfolio attains the maximum growth exponent

$$(57) \qquad W(P) = \sup_{b \in \mathscr{B}} E\{\log(b, X)\} = E\{\log(b^*, X)\}.$$

Conversely, if $W(P)$ is well defined and finite, then any portfolio attaining the maximum is log-optimum. The maximum capital growth exponent may be informally denoted by $W(X)$, although strictly speaking it is a functional $W(P)$ of the distribution P of X. Bell and Cover (1980) proved that a portfolio b^* is log-optimum for P if and only if b^* satisfies the Kuhn–Tucker conditions

$$(58) \qquad E\left\{\frac{(b, X)}{(b^*, X)}\right\} \le 1 \quad \text{for all portfolios } b \in \mathscr{B}.$$

The growth factor (b^*, X) is uniquely defined almost surely, even if the log-optimum portfolio b^* is not unique. It is possible to select a log-optimum portfolio $b^*(P)$ that is Borel measurable in P when the space of probability distributions on $\mathbb{R}_+^m$ is equipped with the weak topology. See Algoet and Cover (1988a) for proofs of this as well as the following results.

The maximum capital growth rate for a general market with dependent return vectors is asymptotically attained by maximizing the conditional expected growth exponent given the currently available information at each step. Indeed, let b_t^* denote the result when the log-optimum portfolio selection function $b^*(\cdot)$ is applied to a regular conditional distribution $P(dx_t | X^t)$ of X_t given the t-past X^t. Then b_t^* is a nonanticipating portfolio attaining the maximum conditional expected log return

$$(59) \qquad E\{\log(b_t^*, X_t) | X^t\} = \sup_{b \in \sigma(X^t)} E\{\log(b, X_t) | X^t\}.$$

Let the capital growth over n periods of investment according to the log-optimum strategy $\{b_t^*\}_{0 \le t < \infty}$ and a competing nonanticipating strategy $\{b_t\}_{0 \le t < \infty}$ be denoted by

$$S_n^* = \prod_{0 \le t < n} (b_t^*, X_t), \qquad S_n = \prod_{0 \le t < n} (b_t, X_t).$$

Then $\{S_n / S_n^*, \sigma(X^n)\}$ is a nonnegative supermartingale with initial value 1 and

$$(60)\qquad \text{AOP:}\qquad \limsup_n \frac{1}{n}\log\left(\frac{S_n}{S_n^*}\right) \le 0 \quad \text{a.s.}$$

This asymptotic optimality principle holds with no restriction on the distribution of the market process $\{X_t\}$.

If the market is stationary, then its history can in principle be extended into the infinite past. Thus a stationary market can be modeled as a two-sided sequence of return vectors $\{X_t\}_{-\infty<t<\infty}$. Let $\bar{b}_t^*$ denote the portfolio that results when the log-optimum selection function $b^*(\cdot)$ is applied to a regular conditional distribution $P(dx|X^{-t})$ of $X = X_0$ given the t-past $X^{-t} = (X_{-t},\dots,X_{-1})$. Then $\bar{b}_t^*$ is conditionally log-optimum for period 0 given the t-past, that is, $\bar{b}_t^*$ attains the maximum conditional expected log return

$$(61)\qquad E\{\log(\bar{b}_t^*, X)|X^{-t}\} = \sup_{b\in\sigma(X^{-t})} E\{\log(b,X)|X^{-t}\}.$$

The maximum expected growth exponent given the t-past is given by

$$(62)\qquad W(X|X^{-t}) = \sup_{b\in\sigma(X^{-t})} E\{\log(b,X)\} = E\{\log(\bar{b}_t^*, X)\}.$$

The maximum is taken over larger sets of permissible portfolios as t increases, so that $W(X|X^{-t})$ is monotonically increasing with t.

Recall that $P(dx|X^{-t})$ converges weakly almost surely to the conditional distribution $P(dx|X^-)$ of X given the infinite past $X^- = (\dots,X_{-2},X_{-1})$. Application of $b^*(\cdot)$ to $P(dx|X^-)$ yields a portfolio $\bar{b}^*$ that is conditionally log-optimum given the infinite past. Thus $\bar{b}^*$ attains

$$(63)\qquad E\{\log(\bar{b}^*, X)|X^-\} = \sup_{b\in\sigma(X^-)} E\{\log(b,X)|X^-\}.$$

The maximum growth exponent given the infinite past is equal to

$$(64)\qquad W(X|X^-) = \sup_{b\in\sigma(X^-)} E\{\log(b,X)\} = E\{\log(\bar{b}^*, X)\}.$$

It can be shown that the maximum growth exponent given the t-past increases monotonically to the maximum growth exponent given the infinite past:

$$(65)\qquad W(X|X^{-t}) \nearrow W(X|X^-) \quad \text{as } t\to\infty.$$

If the market is stationary ergodic, then the maximum capital growth rate is well defined and almost surely equal to $W(X|X^-)$:

$$(66)\qquad S_n^* = \exp[nW(X|X^-) + o(n)], \quad \text{where } o(n)/n \to 0 \text{ a.s.}$$

This asymptotic equipartition property or AEP for log-optimum investment in a stationary ergodic market is a generalization of the Shannon–McMillan–Breiman theorem of information theory. To prove the AEP, Algoet and Cover (1988a, b) argued that an investor who at time t may look back at the t-past is sandwiched in asymptotic growth rate between an investor who may look back at the k-past and one who may look back at the infinite past. The maximum capital growth rate for an investor who may look back only at the k-past is equal to $W(X|X^{-k})$, and for an investor who may look back at the infinite past it is equal to $W(X|X^-)$. It follows from the asymptotic optimality principle that the growth rate $n^{-1}\log S_n^*$ is asymptotically sandwiched between the maximum rate given the k-past and the maximum rate given the infinite past:

$$(67)\quad W(X|X^{-k}) \le \liminf_n \frac{1}{n}\log S_n^* \le \limsup_n \frac{1}{n}\log S_n^* \le W(X|X^-) \quad \text{a.s.}$$

Since the lower bound $W(X|X^{-k})$ increases to the upper bound $W(X|X^-)$ as $k \to \infty$, we may conclude that the asymptotic equipartition property holds:

$$(68)\qquad \text{AEP:}\quad \frac{1}{n}\log S_n^* \to W(X|X^-) \quad \text{a.s.}$$

Thus an investor who at time t may only recall the t-past can attain the same limiting growth rate as an investor who can always remember the infinite past.

The maximum capital growth rate for a stationary ergodic market is attainable if the distribution of the market is unknown but randomly selected according to a known prior distribution on stationary ergodic modes. It suffices to select portfolios that are log-optimum under the stationary mixture distribution. Indeed, if the investor diversifies during each period t according to the portfolio b_t^* that is conditionally log-optimum given X^t under the stationary mixture distribution, then the compounded capital $S_n^* = \prod_{0\le t<n}(b_t^*, X_t)$ will grow exponentially fast almost surely with limiting rate equal to the maximum rate associated with the stationary ergodic mode that governs the actual realization of $\{X_t\}$. This is not surprising since the ergodic mode is uniquely identified by the infinite past.

Every stationary market is a mixture of stationary ergodic modes, but it is not clear how to construct a prior distribution that supports every stationary ergodic mode. Furthermore, the AEP for stationary markets only guarantees exponential growth with the maximum rate almost surely under the stationary mixture distribution. We want an AEP that holds for every stationary ergodic market, not one that only holds with probability 1 under some prior distribution on stationary ergodic modes. We shall argue that the maximum capital growth rate $W(X|X^-)$ can be achieved in the limit even if the distribution of the stationary ergodic market is unknown and must be learned from experience.

4.2. *A universal portfolio selection strategy.* Suppose the market process $\{X_t\}$ is stationary ergodic, but its distribution P is unknown. For $t \ge 0$ let $P(dx_t|X^t)$ denote a regular conditional distribution of X_t given X^t. If capital is allocated at the beginning of every investment period t according to the portfolio b_t^* that is conditionally log-optimum given the t-past X^t, then the compounded capital $S_n^* = \prod_{0\le t<n}(b_t^*, X_t)$ will grow exponentially fast almost surely with the maximum limiting rate $W(X|X^-)$. Portfolio b_t^* cannot be

computed directly since the conditional distribution $P(dx_t|X^t)$ is unknown. However, we can use the t-past X^t to compute an estimate $\hat{P}(dx_t|X^t)$ and diversify according to the portfolio $\hat{b}_t^*$ that is log-optimum with respect to this estimate. This results in the compounded capital

$$(69) \qquad \hat{S}_n^* = \prod_{0 \le t < n} (\hat{b}_t^*, X_t).$$

We shall prove that $\hat{S}_n^*$ grows with the same maximum rate $W(X|X^-)$ as S_n^*, provided the estimate $\hat{P}(dx_t|X^t)$ is carefully constructed and the market is safe in some formal sense. A general market is not safe, but we can approach the maximum growth exponent $W(X|X^-)$ to within any desired $\varepsilon > 0$.

The estimates $\hat{P}(dx_t|X^t)$ are constructed as follows. In Section 5.2 we shall describe a universal prediction scheme to learn the conditional distribution $P(dx|X^-)$ of $X = X_0$ given the infinite past X^-. The algorithm reads past outcomes and generates estimates based on what it has read so far. In particular, $\hat{P}(dx|X^{-t})$ will denote the estimate that is generated by the algorithm on the basis of the t-past $(X_{-1}, \dots, X_{-t})$, before it reads the next past outcome X_{-t-1}. The estimates $\hat{P}(dx|X^{-t})$ converge weakly almost surely to $P(dx|X^-)$. We define $\hat{P}(dx_t|X^t)$ as the output of the algorithm when it is applied to the shifted input sequence $X_{t-1}, \dots, X_0$. Thus

$$(70) \qquad \hat{P}(dx_t|X^t) = \hat{P}(dx|X^{-t}) \circ T^t.$$

The market is called *safe* if $E\{\log X^j\} > -\infty$ for $1 \le j \le m$ and hence

$$(71) \qquad E\{\log(b, X)\} > -\infty \quad \text{for } every \text{ portfolio selection } b \in \sigma(X^-).$$

The following theorem does not apply to a Kelly horse race, which is a most unsafe market.

THEOREM 7. *Let $\{X_t\}$ be a stationary ergodic market process, and let b_t^* and $\hat{b}_t^*$ denote the portfolios that result when the log-optimum portfolio selection function $b^*(\cdot)$ is applied to the true conditional distribution $P(dx_t|X^t)$ and an estimate $\hat{P}(dx_t|X^t)$, respectively. Let $\hat{P}(dx_t|X^t)$ be the shifted version of an estimate $\hat{P}(dx|X^{-t})$ such that*

$$(72) \qquad \hat{P}(dx|X^{-t}) \to P(dx|X^-) \quad weakly\ a.s.$$

If the market is safe, then diversification according to the portfolios b_t^* or $\hat{b}_t^*$ attains the maximum growth rate given the infinite past:

$$(73) \qquad \frac{1}{n} \log \hat{S}_n^* \to W(X|X^-) = \lim_n \frac{1}{n} \log S_n^* \quad a.s.$$

PROOF. The AEP for the log-optimum portfolios b_t^* and the AOP for the nonanticipating portfolios $\hat{b}_t^*$ assert that

$$\lim_n \frac{1}{n} \log S_n^* = W(X|X^-) \quad \text{a.s.},$$

$$\limsup_n \frac{1}{n} \log\left(\frac{\hat{S}_n^*}{S_n^*}\right) \le 0 \quad \text{a.s.}$$

This implies the asymptotic upper bound

$$\limsup_n \frac{1}{n} \log \hat{S}_n^* \leq W(X|X^-) = \lim_n \frac{1}{n} \log S_n^* \quad \text{a.s.}$$

The difficult part is proving the asymptotic lower bound

$$\liminf_n \frac{1}{n} \log \hat{S}_n^* \geq W(X|X^-) \quad \text{a.s.}$$

For this we use Breiman's extended ergodic theorem.

Let $\bar{b}^*$ and $\hat{\bar{b}}_t^*$ denote the portfolios that result when the log-optimum portfolio selection function $b^*(\cdot)$ is applied to $P(dx|X^-)$ and the estimate $\hat{P}(dx|X^{-t})$. Since $\hat{P}(dx|X^{-t})$ converges weakly almost surely to $P(dx|X^-)$, it follows [cf. Theorem 4 of Algoet and Cover (1988a)] that any accumulation point of $\{\hat{\bar{b}}_t^*\}$ is log-optimum for $P(dx|X^-)$. Thus

$$\left(\hat{\bar{b}}_t^*, X\right) \to (\bar{b}^*, X) \quad \text{a.s.}$$

To prove that $n^{-1} \log \hat{S}_n^* \to W(X|X^-)$, we must argue that

$$\frac{1}{n} \sum_{0 \leq t < n} \hat{g}_t \circ T^t \to E\{g\} \quad \text{a.s.},$$

where $\hat{g}_t = \log(\hat{\bar{b}}_t^*, X)$ and $g = \log(\bar{b}^*, X)$. Clearly $\hat{g}_t \to g$ a.s., and we may assume without loss of generality (by rescaling X) that X is bounded and hence that $\hat{g}_t$ is bounded above by a fixed constant. To prove that $\{\hat{g}_t\}$ is L^1-dominated, it therefore suffices to verify the integrability condition $E\{\inf_t \hat{g}_t\} > -\infty$. But $\inf_t \hat{g}_t \geq \min_j \log X^j$ since $(b, X) \geq \min_j X^j$ for any portfolio b, and $E\{\min_j \log X^j\} > -\infty$ since the market is safe. Thus the integrability condition is satisfied, and the theorem follows. □

The theorem holds more generally if the scaled return vector $U = X/(\beta, X)$ is safe, for some constant portfolio $\beta = (\beta^j)_{1 \leq j \leq m}$ that places a positive amount $\beta^j > 0$ on every stock j. The most obvious choice for β is the portfolio $(1/m)_{1 \leq j \leq m}$ that allocates equal amounts to all stocks.

If the market is not safe, then some stocks will occasionally yield zero return or perform much worse than other stocks. The investor should prevent the possibility of going broke by bounding portfolios away from the boundary of the simplex, that is, by placing some amount on all stocks, including the least promising ones. Let β be a fixed portfolio such that $\beta^j > 0$ for all j, and for $0 \leq \lambda < 1$ let

$$\hat{S}_n^\lambda = \prod_{0 \leq t < n} \left(\hat{b}_t^{*\lambda}, X_t\right), \quad \text{where } \hat{b}_t^{*\lambda} = (1 - \lambda)\beta + \lambda \hat{b}_t^*. \tag{74}$$

The jth component of portfolio $\hat{b}_t^{*\lambda}$ is bounded below by the positive constant $(1 - \lambda)\beta^j$.

For any information field $\mathscr{G}$ and $0 \leq \lambda < 1$, let

$$W^\lambda(X|\mathscr{G}) = E\{\log(b^{*\lambda}, X)\}, \tag{75}$$

where $b^{*\lambda} = (1-\lambda)\beta + \lambda b^*$ and b^* attains the maximum conditional expected growth exponent $\sup E\{\log(b, X)|\mathscr{S}\}$. Since $(b^{*\lambda}, X) \geq \lambda(b^*, X)$, we see that

$$W(X|\mathscr{S}) + \log\lambda \leq W^\lambda(X|\mathscr{S}) \leq W(X|\mathscr{S}). \tag{76}$$

[Alternatively, one may define $W^\lambda(X|\mathscr{S}) = E\{\log(b^{\lambda *}, X)\}$, where $b^{\lambda *}$ attains $\sup E\{\log(b^\lambda, X)|\mathscr{S}\}$ and the supremum is taken over the shrunken simplex $\mathscr{B}^\lambda$ of portfolios $b^\lambda = (1-\lambda)\beta + \lambda b$. The portfolio $b^{\lambda *}$ is log-optimum in $\mathscr{B}^\lambda$ and is at least as good as the portfolio $b^{*\lambda}$ which is obtained by shrinking the log-optimum portfolio b^*.]

THEOREM 8. *Consider a stationary ergodic market with unknown distribution. If capital is diversified according to the nonanticipating portfolio $\hat{b}_t^{*\lambda}$ at the beginning of every investment period t, then the compounded capital $\hat{S}_n^\lambda$ will grow exponentially fast almost surely with a well-defined limiting rate. In fact,*

$$\frac{1}{n}\log\hat{S}_n^\lambda \to W^\lambda(X|X^-) \quad a.s., \tag{77}$$

where [*writing* $\bar{b}^{*\lambda} = (1-\lambda)\beta + \lambda\bar{b}^*$]

$$W^\lambda(X|X^-) = E\{\log(\bar{b}^{*\lambda}, X)\}. \tag{78}$$

PROOF. We use Breiman's extended ergodic theorem for the random variables

$$\hat{g}_t^\lambda = \log\left(\frac{((1-\lambda)\beta + \lambda\hat{b}_t^*, X)}{(\beta, X)}\right),$$

$$g^\lambda = \log\left(\frac{((1-\lambda)\beta + \lambda\bar{b}^*, X)}{(\beta, X)}\right).$$

Since $\hat{g}_t^\lambda \to g^\lambda$ a.s. and $\{\hat{g}_t^\lambda\}$ is bounded between $\log(1-\lambda)$ and $\max_j \log(1/\beta^j)$, we have

$$\frac{1}{n}\sum_{0\leq t<n} \hat{g}_t^\lambda \circ T^t \to E\{g^\lambda\} \quad \text{a.s.}$$

The theorem follows since

$$\frac{1}{n}\log\hat{S}_n^\lambda = \frac{1}{n}\sum_{0\leq t<n}\hat{g}_t^\lambda\circ T^t + \frac{1}{n}\sum_{0\leq t<n}\log(\beta, X_t)$$
$$\to E\{g^\lambda\} + E\{\log(\beta, X)\} = W^\lambda(X|X^-) \quad \text{a.s.} \qquad \square$$

The nonanticipating strategy $\{\hat{b}_t^{*\lambda}\}$ achieves capital growth exponent $W^\lambda(X|X^-)$ that approaches $W(X|X^-)$ to within any desired ε as $\lambda \nearrow 1$ (namely, to within $\varepsilon = -\log\lambda$). We show that the maximum growth exponent given the infinite past can be asymptotically achieved.

Theorem 9 (Universal portfolio selection). *Suppose the stock market process $\{X_t\}$ is stationary ergodic with unknown distribution. There exists a universal nonanticipating portfolio selection strategy $\{\hat{b}_t\}$ such that the compounded capital $\hat{S}_n = \prod_{0 \le t < n}(\hat{b}_t, X_t)$ grows exponentially fast almost surely with the maximum rate $W(X|X^-)$*:

$$\frac{1}{n} \log \hat{S}_n \to W(X|X^-) \quad a.s. \tag{79}$$

Proof. The AEP, in combination with the AOP for any nonanticipating portfolio selection strategy $\{\hat{b}_t\}$, asserts that

$$\limsup_n \frac{1}{n} \log \hat{S}_n \le W(X|X^-) = \lim_n \frac{1}{n} \log S_n^* \quad \text{a.s.}$$

We use the bookkeeping technique of Theorem 2 to construct a particular nonanticipating strategy such that

$$\liminf_n \frac{1}{n} \log \hat{S}_n \ge W(X|X^-) \quad \text{a.s.}$$

The initial capital $S_0 = 1$ is distributed over a countable number of separate accounts, indexed by k. Let $\mu_k > 0$ denote the initial value of the kth account, so that $\sum_k \mu_k = S_0 = 1$. The kth account is managed using the strategy $\{\hat{b}_t^{*\lambda_k}\}$ so that its value after n investment periods is given by $[\mu_k \hat{S}_n^{\lambda_k}]$. By Theorem 8, the capital in the kth account will compound with limiting rate

$$\lim_n \frac{1}{n} \log\left[\mu_k \hat{S}_n^{\lambda_k}\right] = W^{\lambda_k}(X|X^-) \quad \text{a.s.}$$

The total capital in all accounts is given by

$$\hat{S}_n = \sum_k \left[\mu_k \hat{S}_n^{\lambda_k}\right].$$

We assume that $\lambda_k \nearrow 1$ as $k \to \infty$. Since $\hat{S}_n \ge [\mu_k \hat{S}_n^{\lambda_k}]$ for all k, we may conclude that

$$\liminf_n \frac{1}{n} \log \hat{S}_n \ge W(X|X^-) = \sup_k W^{\lambda_k}(X|X^-) \quad \text{a.s.}$$

The accounts form a bookkeeping device, that is, they are separate only on paper. The total capital is pooled at the end of every investment period t, and is then reinvested according to the nonanticipating portfolio

$$\hat{b}_t = \frac{\sum_k \left[\mu_k \hat{S}_t^{\lambda_k}\right] b_t^{*\lambda_k}}{\sum_k \left[\mu_k \hat{S}_t^{\lambda_k}\right]}. \qquad \square$$

The universal prediction scheme $\hat{P}(dx|X^{-t})$ is complicated, but it allows construction of a strategy $\{\hat{b}_t^{*\lambda}\}$ whose growth exponent $W^\lambda(X|X^-)$ approaches the maximum $W(X|X^-)$ uniformly to within any desired $\varepsilon = \log(1/\lambda)$.

4.3. *A simpler approach*. We now describe a universal portfolio selection scheme that does not require a universal prediction scheme as a subroutine. For any $k \geq 0$, any finite subfield $\mathscr{F}^{-k}$ of $\sigma(X^{-k})$, and any $0 \leq \lambda < 1$ we construct a strategy $\{\hat{b}_t^{k,\lambda}\}$ for which the compounded capital $\hat{S}_n^{k,\lambda}$ grows exponentially fast almost surely with rate $W^\lambda(X|\mathscr{F}^{-k})$. The maximum rate $W(X|X^-)$ can be asymptotically attained by combining such strategies.

Fix some $k \geq 0$ and some finite subfield $\mathscr{F}^{-k}$ of the k-past $\sigma(X^{-k})$. The empirical estimate of the conditional distribution $P(dx|\mathscr{F}^{-k})$ based on X^{-k-s} is defined for $s \geq 0$ as

$$(80) \qquad \hat{P}_s(dx|\mathscr{F}^{-k}) = \frac{\delta_{\xi_0}(dx) + \sum_{\tau \in I_s^{-k}} \delta_{X_{-\tau}}(dx)}{1 + \|I_s^{-k}\|},$$

where $\xi_0 \in \mathscr{X}$ is arbitrary, $\delta_\xi(dx)$ is the unit mass at $\xi \in \mathscr{X}$ and

$$(81) \qquad I_s^{-k} = \{\tau\colon 1 \leq \tau \leq s, (X_{-\tau-k}, \dots, X_{-\tau-1}) \text{ and } (X_{-k}, \dots, X_{-1}) \text{ belong to the same atom of } \mathscr{F}^{-k}\}.$$

It follows from the ergodic theorem that

$$(82) \qquad \hat{P}_s(dx|\mathscr{F}^{-k}) \to P(dx|\mathscr{F}^{-k}) \quad \text{weakly a.s. as } s \to \infty.$$

Let $\mathscr{F}_t^k$ denote the finite subfield of $\sigma(X_{t-k}, \dots, X_{t-1})$ that is obtained by shifting $\mathscr{F}^{-k}$:

$$(83) \qquad \mathscr{F}_t^k = T^t \mathscr{F}^{-k} = \{T^t F\colon F \in \mathscr{F}^{-k}\}.$$

Shifting $\hat{P}_t(dx|\mathscr{F}^{-k})$ yields an empirical estimate $\hat{P}_t(dx_t|\mathscr{F}_t^k)$ of $P(dx_t|\mathscr{F}_t^k)$. This empirical estimate is a function of X^t and X^{-k}. [To compute the estimate, one must make some arbitrary choice for X^{-k}, since X^{-k} is unknown. It is perhaps more convenient to work with $\hat{P}_{s-k}(dx|\mathscr{F}^{-k})$ which is a function of X^{-s}. Shifting $\hat{P}_{t-k}(dx|\mathscr{F}^{-k})$ yields an estimate for $P(dx_t|\mathscr{F}_t^k)$ that is based on X^t.]

We consider the log-optimum portfolios $\bar{b}^k$ and $\hat{b}_t^k$ for $P(dx|\mathscr{F}^{-k})$ and $\hat{P}_t(dx_t|\mathscr{F}_t^k)$. Let β be a fixed portfolio with $\beta^j > 0$ for all j, and for $0 \leq \lambda < 1$ let

$$(84) \qquad \hat{S}_n^{k,\lambda} = \prod_{0 \leq t < n} \left(\hat{b}_t^{k,\lambda}, X_t\right) \quad \text{where } \hat{b}_t^{k,\lambda} = (1-\lambda)\beta + \lambda \hat{b}_t^k.$$

It is easily verified using Breiman's generalized ergodic theorem that $\hat{S}_n^{k,\lambda}$ grows exponentially fast almost surely with limiting rate

$$(85) \qquad \lim_n \frac{1}{n} \log \hat{S}_n^{k,\lambda} = W^\lambda(X|\mathscr{F}^{-k}) = E\{\log(\bar{b}^{k,\lambda}, X)\} \quad \text{a.s.},$$

where $\bar{b}^{k,\lambda} = (1-\lambda)\beta + \lambda\bar{b}^k$. Note that $W^\lambda(X|\mathscr{F}^{-k})$ increases to $W(X|\mathscr{F}^{-k})$ as $\lambda \nearrow 1$ since

$$(86) \qquad W(X|\mathscr{F}^{-k}) + \log\lambda \leq W^\lambda(X|\mathscr{F}^{-k}) \leq W(X|\mathscr{F}^{-k}).$$

To attain the maximum growth exponent $W(X|X^-)$, we divide the capital into countably many accounts, indexed by $k \geq 0$. Let $\mu_k > 0$ denote the initial amount in the kth account ($\Sigma_k \mu_k = 1$), and suppose the kth account is managed according to the strategy $\{\hat{b}_t^{k,\lambda_k}\}$, where $\lambda_k \nearrow 1$ and $\mathscr{F}^{-k}$ is a finite subfield of $\sigma(X^{-k})$ such that $\mathscr{F}^{-k}$ increases to $\sigma(X^-)$ as $k \to \infty$. The total capital, obtained by summing all accounts, will grow exponentially fast with limiting rate

$$W(X|X^-) = \sup_k W^{\lambda_k}(X|\mathscr{F}^{-k}). \tag{87}$$

4.4. *Some observations.*

Remark 1. Markets for which the return vector X_t is always oriented along a coordinate axis of $\mathbb{R}_+^m$ deserve special attention. Log-optimum investment for such a market reduces to log-optimum gambling on the identity J_t of the stock that will yield the nonzero return. A Kelly horse race is the canonical example of such a market, because exactly one horse will win and yield a nonzero return. Placing bets on every horse j proportional to its conditional win probability is log-optimum. If the winning horse J_t returns m times the amount invested in it, then X_t is m times the unit vector in direction J_t, and $W(X|X^-) = [\log m - H(J|J^-)]$, where $H(J|J^-)$ is the entropy rate of the stationary ergodic process $\{J_t\}$. If a universal portfolio selection strategy is applied to a stationary ergodic market such that exactly one stock will yield a nonzero return, then the result is a universal gambling scheme.

Remark 2. A stationary ergodic market with side information is described by a stationary ergodic pair process $\{(X_t, Y_t)\}$, where $X_t \in \mathbb{R}_+^m$ is the return vector for period t and Y_t is side information taking values in some Polish space $\mathscr{Y}$. If the nonanticipating portfolio b_t may depend on the side information Y_t and the t-past Y^tX^t, then the maximum capital growth exponent is given by

$$W(X|YY^-X^-) = E\{\log(\bar{b}^*, X)\}, \tag{88}$$

where portfolio $\bar{b}^*$ is conditionally log-optimum for period 0 given the side information $Y = Y_0$ and the infinite past Y^-X^-. The maximum growth rate can be asymptotically attained even if the pair process distribution P is unknown. It suffices to use estimates $\hat{P}(dx_t|Y_tY^tX^t)$ of $P(dx_t|Y_tY^tX^t)$ rather than estimates $\hat{P}(dx_t|X^t)$ of $P(dx|X^t)$ in the preceding strategies. In Section 5.2 we show how to generate estimates $\hat{P}(dx|YY^{-t}X^{-t})$ such that

$$\hat{P}(dx|YY^{-t}X^{-t}) \to P(dx|YY^-X^-) \quad \text{weakly a.s.} \tag{89}$$

Remark 3. Móri (1984) carried out a refined analysis for markets with independent identically distributed return vectors such that the log-optimum portfolio b^* is unique and contained in the interior of the unit simplex. His results imply that the capital $S_n = \Pi_{0 \le t < n}(b_t, X_t)$ will grow with the maximum rate $W(X)$ if $b_t = (1 - \lambda_t)\beta + \lambda_t\hat{b}_t^*$, where $\lambda_t \nearrow 1$ and $\hat{b}_n^*$ is log-optimum for the empirical measure $\hat{P}_n = (1/n)\Sigma_{0 \le t < n}\delta_{X_t}$. In fact he proves the stronger result that $n^{(m-1)/2}S_n/S_n^*$ has a nondegenerate limit distribution. Thus the capital of an investor who must learn the market distribution from experience grows only polynomially less fast than the capital of an investor who already knows the distribution to begin with. However, this result is valid only if the return vectors are independent and identically distributed.

Remark 4. Cover (1991) considers a different type of universal portfolio selection, for markets with arbitrary return vectors X_t that need not form a stationary random process. If the capital is rebalanced at the beginning of every period t according to some fixed portfolio b, then the capital growth over n investment periods amounts to

$$S_n(b) = \prod_{0\le t<n} (b, X_t). \tag{90}$$

Let $b_n^* = b^*(\hat{P}_n)$ denote the portfolio that is log-optimum for the empirical distribution

$$\hat{P}_n = \frac{1}{n} \sum_{0\le t<n} \delta_{X_t}. \tag{91}$$

Then

$$S_n(b_n^*) = \max_{b\in\sigma(X^n)} S_n(b). \tag{92}$$

It is possible to generate nonanticipating portfolios $\hat{b}_t$ such that the resulting capital $\hat{S}_n = \prod_{0\le t<n}(\hat{b}_t, X_t)$ grows only polynomially less fast than S_n^*. To prove this, Cover considers a continuum of accounts, indexed by all possible portfolio vectors $b \in \mathscr{B}$. The initial capital $S_0 = 1$ is uniformly distributed over all possible accounts (according to the distribution that is proportional to the Lebesgue measure db on the simplex $\mathscr{B}$), and the strategy for the bth account is to rebalance according to the constant portfolio b. Thus the compounded capital after n period amounts to

$$\hat{S}_n = \frac{\int_{\mathscr{B}} S_n(b)\, db}{\int_{\mathscr{B}} db} = \prod_{0\le t<n} (\hat{b}_t, X_t), \tag{93}$$

where

$$\hat{b}_t = \frac{\int_{\mathscr{B}} b S_t(b)\, db}{\int_{\mathscr{B}} S_t(b)\, db}. \tag{94}$$

Note that $\hat{b}_t$ is an average over the simplex $\mathscr{B}$, when each portfolio $b \in \mathscr{B}$ is weighted according to how well it would have performed in the past. Similarly, $\hat{S}_n$ is the average growth over the continuum of accounts indexed by portfolios in the simplex.

Cover's result is stronger than ours because he makes no statistical assumptions about the return vectors X_t (which need not be random variables). In fact, the empirical distribution $\hat{P}_n$ need not converge, and the growth rate of $S_n(b_n^*)$ or $\hat{S}_n$ need not exist. But Cover's universal strategy is also weaker because it only attains the maximum limiting rate achievable by rebalancing according to constant portfolios. Our universal strategy for stationary ergodic markets attains the maximum capital growth rate achievable by nonanticipating strategies that may be time-varying. The maximum achievable by portfolios that may depend on the infinite past is higher than the maximum achievable by constant portfolios, except when the return vectors are independent and identically distributed.

5. Universal prediction schemes.

We consider a discrete-time stationary ergodic process $\{X_t\}$ with values in a Polish space $\mathscr{X}$. Thus $X_t(\omega) = X(T^t\omega)$, where $X:\Omega \to \mathscr{X}$ is a random variable defined on a stationary ergodic dynamical system $(\Omega, \mathscr{F}, P, T)$. We assume that $(\Omega, \mathscr{F})$ is the two-sided sequence space $\mathscr{X}^{\mathbb{Z}}$ with its Borel σ-field and T is the left shift on Ω. The distribution P is stationary ergodic but unknown, and we wish to learn the conditional distribution $P(dx|X^-)$ of the next outcome $X = X_0$ given the infinite past $X^- = (X_{-1}, X_{-2}, \dots)$. An algorithm that achieves this goal is called a universal prediction scheme.

The infinite past uniquely determines the ergodic mode of the process as well as the conditional distribution $P(dx|X^-)$, but no finite number of past outcomes is sufficient to identify $P(dx|X^-)$ exactly. However, it is possible to generate estimates $\hat{P}_k = \hat{P}_k(dx|X^{-\sigma_k})$ that converge weakly almost surely to $P(dx|X^-)$ no matter what the unknown distribution P of $\{X_t\}$ happens to be, as long as P is stationary ergodic. The estimates will depend on a finite but growing number of past observations, that is, $\hat{P}_k$ will be a function of $X^{-\sigma_k} = (X_{-1}, \dots, X_{-\sigma_k})$, where $\sigma_k = \sigma_k(\omega)$ is a stopping time (when going backwards into the past) such that $\sigma_k(\omega) \to \infty$ more or less fast depending on the particular realization ω of $\{X_t\}$.

The estimates $\hat{P}_k$ are easily transformed into estimates $\hat{P}(dx|X^{-t})$ which are functions of the t-past $X^{-t} = (X_{-1}, \dots, X_{-t})$ and converge weakly almost surely to $P(dx|X^-)$. It suffices to set

$$\hat{P}(dx|X^{-t}) := \hat{P}_k(dx|X^{-\sigma_k}) \quad \text{if } \sigma_k \le t < \sigma_{k+1}. \tag{95}$$

Thus $\hat{P}(dx|X^{-t})$ is the most recent estimate $\hat{P}_k$ that is generated before X_{-t-1} is input. The initial choice of $\hat{P}(dx|X^{-0})$ is arbitrary during the startup phrase when $0 \le t < \sigma_1$.

Ornstein (1978) formulated an algorithm which uses past experience to learn the conditional distribution of the next outcome given the infinite past of a stationary ergodic process with values in a finite [e.g., binary] set. This universal prediction scheme is also described in the thesis of Bailey (1976). We shall review Ornstein's algorithm for finite-valued processes and then generalize it when the space of possible outcomes is Polish. A random process with values in the finite set $\{1, \dots, m\}$ will be denoted by $\{J_t\}$, to be consistent with previous sections.

5.1. *Ornstein's universal prediction scheme.*

Let $\{J_t\}$ be a stationary ergodic process with values in the finite set $\{1, \dots, m\}$. Let $P(\cdot|J^-)$ and $P(\cdot|J^{-t})$ denote the conditional distributions of $J = J_0$ given the infinite past $J^- =$

$(J_{-1}, J_{-2}, \dots)$ and the t-past $J^{-t} = (J_{-1}, \dots, J_{-t})$. Lévy's martingale convergence theorem for conditional probabilities asserts that

$$P(j|J^{-t}) \to P(j|J^{-}) \quad \text{a.s. as } t \to \infty.$$

The conditional probability $P(j|J^{-t})$ itself is the limit of empirical estimates. Indeed, for finite $s > 0$ let the set of occurrences of the block J^{-t} within the longer block J^{-s-t} be denoted by

$$I_s^{-t} = \{\tau \colon 1 \le \tau \le s, (J_{-t-\tau}, \dots, J_{-1-\tau}) = (J_{-t}, \dots, J_{-1})\}. \tag{96}$$

(Note that the occurrence of J^{-t} at $\tau = 0$ is not counted.) The empirical estimate of the conditional probability $P(j|J^{-t})$ based on J^{-s-t} and some fixed $j_0 \in \{1, \dots, m\}$ is defined as

$$\hat{P}_s(j|J^{-t}) = \frac{\delta_j(j_0) + \Sigma_{\tau \in I_s^{-t}} \delta_j(J_{-\tau})}{1 + \|I_s^{-t}\|}, \tag{97}$$

where $\delta_j(\cdot)$ is the indicator function of $j \in \{1, \dots, m\}$. To prove that $\hat{P}_s(j|J^{-t})$ converges to $P(j|J^{-t})$ as $s \to \infty$, observe that $\|I_s^{-t}\|/s$ is the relative frequency of past occurrences of the block J^{-t} within the longer block J^{-s-t}, and that $[\Sigma_{\tau \in I_s^{-t}} \delta_j(J_{-\tau})]/s$ is the relative frequency with which such occurrences are followed by j. By the ergodic theorem,

$$\begin{aligned} \hat{P}_s(j|J^{-t}) &= \frac{\left[\delta_j(j_0) + \Sigma_{\tau \in I_s^{-t}} \delta_j(J_{-\tau})\right]/s}{\left[1 + \|I_s^{-t}\|\right]/s} \\ &\to \frac{P\{J^{-t}, J = j\}}{P\{J^{-t}\}} = P(j|J^{-t}) \quad \text{a.s. as } s \to \infty. \end{aligned} \tag{98}$$

Recall that $\hat{P}_s(\cdot|J^{-t}) \to P(\cdot|J^{-t})$ a.s. as $s \to \infty$ and $P(\cdot|J^{-t}) \to P(\cdot|J^{-})$ a.s. as $t \to \infty$. If $\{J_t\}$ is νth-order Markov, then $\hat{P}_s(\cdot|J^{-\nu})$ converges to $P(\cdot|J^{-\nu}) = P(\cdot|J^{-})$ as $s \to \infty$. However, no fixed (nonrandom) sequences $\{s_k\}$ and $\{t_k\}$ exist such that $\hat{P}_{s_k}(\cdot|J^{-t_k}) \to P(\cdot|J^{-})$ a.s. in a universal sense, irrespective of the stationary ergodic distribution P. One must rely on the observations themselves as a source of evidence to determine how deep one must go back into the past. Ornstein (1978) proved that $\hat{P}_{s_k}(\cdot|J^{-t_k}) \to P(\cdot|J^{-})$ a.s. for some carefully constructed $s_k = s_k(\omega)$ and $t_k = t_k(\omega)$ which depend on the realization ω of the process $\{J_t\}$. The number $\sigma_k(\omega) = s_k(\omega) + t_k(\omega)$ is a stopping time and the estimate $\hat{P}_{s_k}(\cdot|J^{-t_k})$ depends on ω only through $J^{-\sigma_k} = (J_{-1}, \dots, J_{-\sigma_k})$.

Ornstein's construction of estimates for $P(\cdot|J^{-})$ is supported by the following rationale. We say that two probability distributions Q and R on the set $\{1, \dots, m\}$ are within ε from each other if $\sup_{1 \le j \le m} |Q(j) - R(j)| \le \varepsilon$. Given $\varepsilon > 0$ and $k, l \ge 1$, suppose a sequence $(s_i)_{0 \le i \le l}$ with $k \le s_0 < s_1 < \cdots < s_l < \infty$ exists such that all empirical estimates $\hat{P}_{s_i}(\cdot|J^{-t})$, $s_0 \le t \le s_{i-1}$, $1 \le i \le l$, are within ε from each other. We may regard such $(s_i)_{0 \le i \le l}$ as a certificate of the quality of these empirical estimates. If ε is small and k, l are large, then the estimates are likely to be close to $P(j|J^{-})$. Indeed, $\hat{P}_{s_i}(\cdot|J^{-t})$ is an estimate for $P(\cdot|J^{-t})$, which itself is close to $P(\cdot|J^{-})$ if k and hence $t \ge k$

is large. Furthermore, it appears that we have gone through a number l of rounds to generate, validate and possibly revise estimates. After generating the first estimate $\hat{P}_{s_1}(\cdot|J^{-s_0})$ during the first round $i = 1$, we went through $(l-1)$ confirmation stages $i = 2, 3, \dots, l$, each of which conferred an additional degree of confidence. Round i $(2 \le i \le l)$ certified that all the estimates $\hat{P}_{s_i}(\cdot|J^{-t})$, $s_0 \le t \le s_{i-1}$, are within ε from one another and from the previous estimates $\hat{P}_{s_\iota}(\cdot|J^{-t})$, $s_0 \le t \le s_{\iota-1}$, $1 \le \iota < i$. During the whole process there may have arisen a need on several occasions to revise previous estimates in order to make them consistent with new evidence.

We claim that a certificate $(s_i)_{0 \le i \le l}$ with these properties exists almost surely.

Lemma 3. *Let $\{J_t\}$ be a stationary ergodic process with values in the finite set $\{1, \dots, m\}$. Then for any $\varepsilon > 0$ and $k \ge 1$, there almost surely exists a sequence $(s_i)_{0 \le i < \infty}$ with $k \le s_0 < s_1 < \cdots$ such that all the empirical estimates $\hat{P}_{s_i}(\cdot|J^{-t})$, $s_0 \le t \le s_{i-1}$, $1 \le i < \infty$, are well defined, within $\varepsilon/2$ from $P(\cdot|J^-)$, and hence within ε from each other.*

Lemma 3 will follow from the more general Lemma 5 that is stated and proved in the next section.

Once we have found a certificate $(s_i)_{0 \le i \le l}$ as above, we may be confident that all the estimates $\hat{P}_{s_i}(\cdot|J^{-t})$, $s_0 \le t \le s_{i-1}$, $1 \le i \le l$, are close to $P(\cdot|J^-)$. How confident we may be will now be quantified.

For real $\alpha > 0$, integer $K \ge 0$ and $j \in \{1, \dots, m\}$, let $B^0_{\alpha K}(j)$ denote the entire sample space Ω and let the "bad" event $B^l_{\alpha K}(j)$ be defined for integer $l \ge 1$ as follows. Event $B^l_{\alpha K}(j)$ is said to occur if there exists a certificate $(s_i)_{0 \le i \le l}$ with $K = s_0 < s_1 < \cdots < s_l < \infty$ such that $\|I^{-t}_{s_i}\| > 0$ and

$$\hat{P}_{s_i}(j|J^{-t}) \ge P(j|J^{-t}) + \alpha \quad \text{for } s_0 \le t \le s_{i-1}, 1 \le i \le l. \tag{99}$$

Among all sequences $(s_i)_{0 \le i \le l}$ that witness membership in $B^l_{\alpha K}(j)$, we always select the one that is minimal in the following sense. A sequence $(r_i)_{0 \le i \le l}$ precedes a different sequence $(s_i)_{0 \le i \le l}$ if there exists some λ in the range $0 \le \lambda \le l$ such that $r_\lambda < s_\lambda$ and $r_i = s_i$ for $\lambda < i \le l$. This is the standard lexicographic ordering of sequences read in reverse.

Notice that $B^{l+1}_{\alpha K}(j) \subseteq B^l_{\alpha K}(j)$ for all $l \ge 0$. The event $B^{l+1}_{\alpha K}(j)$ is "worse" than the "bad" event $B^l_{\alpha K}(j)$ but occurs less often. We prove that the probability of $B^l_{\alpha K}(j)$ decreases exponentially fast with l.

Lemma 4. *Let the "bad" event $B^l_{\alpha K}(j)$ be defined as above for $0 < \alpha < 1$, $K \ge 0$, $l \ge 0$ and $j \in \{1, \dots, m\}$. Then $P\{B^{l+1}_{\alpha K}(j)|B^l_{\alpha K}(j)\} \le (1-\alpha)$, and consequently*

$$P\{B^l_{\alpha K}(j)\} \le (1-\alpha)^l. \tag{100}$$

Proof. See the Appendix.

Armed with this insight, we now formulate an algorithm to generate estimates $\hat{P}_k = \hat{P}_k(\cdot|J^{-\sigma_k})$ that converge almost surely to $P(\cdot|J^-)$. We fix some sequence $\{\varepsilon_k\}_{1 \le k < \infty}$ such that $\varepsilon_k \searrow 0$. Two measures Q, R on the finite set $\{1, \dots, m\}$ are said to be k-close if $\sup_{1 \le j \le m} |Q(j) - R(j)| \le \varepsilon_k$.

Algorithm to generate the kth estimate $\hat{P}_k = \hat{P}_k(\cdot|J^{-\sigma_k})$.

1. Find the least integer n for which there exists an integer K and a certificate $(s_i)_{0 \le i \le K}$ such that $k \le K = s_0 < s_1 < \cdots < s_K = n$ and all empirical estimates $\hat{P}_{s_i}(\cdot|J^{-t})$, $s_0 \le t \le s_{i-1}$, $1 \le i \le K$, are well defined and k-close to each other.
2. Choose K smallest possible and choose the certificate $(s_i)_{0 \le i \le K}$ so that it is lexicographically smallest when read in reverse.
3. Set $\sigma_k := s_K + s_{K-1}$ and $\hat{P}_k := \hat{P}_{s_K}(\cdot|J^{-s_{K-1}})$.

The estimate $\hat{P}_k$ is well defined for all $k \ge 1$, since the search for the integers n and K and for the certificate $(s_i)_{0 \le i \le K}$ must terminate by Lemma 3. Also, $\hat{P}_k$ is a function of $J^{-\sigma_k}$, where $\sigma_k(\omega) = s_{K(k,\omega)}(\omega) + s_{K(k,\omega)-1}(\omega)$ is a stopping time for each fixed k.

Theorem 10 (Ornstein). *Let $\{J_t\}$ be a stationary ergodic process with values in the finite set $\{1, \dots, m\}$, and let $P(\cdot|J^-)$ denote the conditional distribution of $J = J_0$ given the infinite past $J^- = (J_{-1}, J_{-2}, \dots)$. The estimate $\hat{P}_k$ generated during phase k of the algorithm converges to $P(\cdot|J^-)$ with probability* 1.

Theorem 10 will follow by specializing Theorem 11 which is stated and proved in the next section.

5.2. *Generalization when $\mathscr{X}$ is Polish.* Let $\{X_t\}$ be a stationary ergodic process with values in a Polish space $\mathscr{X}$. We generalize Ornstein's algorithm and generate estimates $\hat{P}_k = \hat{P}_k(dx|X^{-\sigma_k})$ that converge weakly almost surely to the conditional distribution $P(dx|X^-)$ of $X = X_0$ given the infinite past X^-. The same algorithm as before is used to generate the estimates $\hat{P}_k$, but we need new definitions of empirical estimates and what it means for two distributions on $\mathscr{X}$ to be k-close.

We consider the conditional distribution $P(dx|\mathscr{F}^{-t})$ of $X = X_0$ given a finite field $\mathscr{F}^{-t}$ that approximates $\sigma(X^{-t})$. Let $\{\mathscr{G}_p\}_{0 \le p < \infty}$ be an increasing sequence of finite subfields that asymptotically generate the Borel σ-field on $\mathscr{X}$, and define $\mathscr{F}^{-t}$ as the finite subfield of $\sigma(X^{-t})$ that is generated by cylinder sets of the form

$$F = \{X_{-1} \in B_1, \dots, X_{-t} \in B_t\},$$

where $B_1, \dots, B_t$ are atoms of the finite field $\mathscr{G}_t$. If $\mathscr{X}$ is finite, then we may assume that $\mathscr{F}^{-t} = \sigma(X^{-t})$ for all t. In general $\mathscr{F}^{-t}$ is a finite subfield approximating $\sigma(X^{-t})$, and the approximations get better as t increases.

Indeed, $\sigma(X^{-t})$ and the finite approximating subfield $\mathscr{F}^{-t}$ both increase to the limiting σ-field $\sigma(X^{-})$.

It is well known if $\mathscr{F}^{-t}$ increases to $\sigma(X^{-})$ that

$$P(dx|\mathscr{F}^{-t}) \to P(dx|X^{-}) \quad \text{weakly a.s. as } t \to \infty.$$

Indeed, if $h(x)$ is a bounded continuous (or even a nonnegative measurable) function on $\mathscr{X}$, then by the martingale convergence theorem for conditional expectations,

$$P\{h(X)|\mathscr{F}^{-t}\} \to P\{h(X)|X^{-}\} \quad \text{a.s.}$$

We may approximate the distribution $P(dx|\mathscr{F}^{-t})$ by the empirical estimate

$$\hat{P}_s(dx|\mathscr{F}^{-t}) = \frac{\delta_{\xi_0}(dx) + \Sigma_{\tau \in I_s^{-t}} \delta_{X_{-\tau}}(dx)}{1 + \|I_s^{-t}\|}, \tag{101}$$

where $\xi_0 \in \mathscr{X}$ is arbitrary, $\delta_\xi(dx)$ is the Dirac distribution that places unit mass at $\xi \in \mathscr{X}$, and

$$I_s^{-t} = \{\tau: 1 \le \tau \le s,\ T^{-\tau}\omega \text{ and } \omega \text{ belong to the same atom of } \mathscr{F}^{-t}\}. \tag{102}$$

Thus $\hat{P}_s(dx|\mathscr{F}^{-t})$ is an empirical estimate of $P(dx|\mathscr{F}^{-t})$ based on X^{-s-t}. If $h(X)$ is integrable, then the empirical estimate

$$\hat{P}_s\{h(X)|\mathscr{F}^{-t}\} = \int_{\mathscr{X}} h(x)\hat{P}_s(dx|\mathscr{F}^{-t}) = \frac{h(\xi_0) + \Sigma_{\tau \in I_s^{-t}} h(X_{-\tau})}{1 + \|I_s^{-t}\|} \tag{103}$$

converges almost surely by the ergodic theorem to the conditional expectation

$$P\{h(X)|\mathscr{F}^{-t}\} = \int_{\mathscr{X}} h(x)P(dx|\mathscr{F}^{-t}). \tag{104}$$

Thus $\hat{P}_s(dx|\mathscr{F}^{-t}) \to P(dx|\mathscr{F}^{-t})$ weakly almost surely as $s \to \infty$. We now generalize Lemma 3.

LEMMA 5. *Let $\{X_t\}$ be a stationary ergodic process with values in a Polish space $\mathscr{X}$ and let $\mathscr{H}$ be a finite collection of measurable functions on $\mathscr{X}$ such that $h(X)$ is integrable for all $h \in \mathscr{H}$. Then for any $\varepsilon > 0$ and $k \ge 0$, there exist integers $(s_i)_{0 \le i < \infty}$ with $k \le s_0 < s_1 < \cdots$ such that for all $h \in \mathscr{H}$, the empirical expectations $\hat{P}_{s_i}\{h(X)|\mathscr{F}^{-t}\}$, $s_0 \le t \le s_{i-1}$, $1 \le i < \infty$, are within $\varepsilon/2$ from the conditional expectation $P\{h(X)|X^{-}\}$ and hence within ε from each other.*

PROOF. For s_0 we can take the least integer $s \ge k$ such that the conditional expectation $P\{h(X)|\mathscr{F}^{-t}\}$ is within $\varepsilon/4$ from $P\{h(X)|X^{-}\}$ for all $t \ge s$ and $h \in \mathscr{H}$. Such s exists since $P\{h(X)|\mathscr{F}^{-t}\}$ converges almost surely to $P\{h(X)|X^{-}\}$ by the martingale convergence theorem for conditional expectations. Given $s_0, s_1, \dots, s_{i-1}$, we can inductively define s_i as the least integer

$s > s_{i-1}$ such that the empirical estimate $\hat{P}_s\{h(X)|\mathscr{F}^{-t}\}$ is within $\varepsilon/4$ from $P\{h(X)|\mathscr{F}^{-t}\}$ for all $h \in \mathscr{H}$ and $s_0 \le t \le s_{i-1}$. Such s exists since $\hat{P}_s(dx|\mathscr{F}^{-t})$ converges weakly almost surely to $P(dx|\mathscr{F}^{-t})$ as $s \to \infty$ by the ergodic theorem. The estimates $\hat{P}_{s_i}\{h(X)|\mathscr{F}^{-t}\}$, $s_0 \le t \le s_{i-1}$, $1 \le i < \infty$, are all within $\varepsilon/2$ from $P\{h(X)|X^-\}$ and therefore within ε from each other. □

Lemma 5 reduces to Lemma 3 when $\mathscr{X}$ is a finite set equipped with the discrete topology and $\mathscr{H}$ is the family of indicator functions $\{\delta_x\}_{x \in \mathscr{X}}$. Note that weak convergence of conditional distributions reduces to ordinary convergence when $\mathscr{X}$ is finite.

To generalize Lemma 4, we define events $B^l_{\alpha K}(h)$ for $\alpha > 0$, $K \ge 0$, $l \ge 0$, and bounded continuous $h(x)$ on $\mathscr{X}$, as follows. First, $B^0_{\alpha K}(h)$ will denote the entire sample space Ω as before. For $l \ge 1$, we say that $B^l_{\alpha K}(h)$ occurs if a certificate $(s_i)_{0 \le i \le l}$ with $K = s_0 < s_1 < \cdots < s_l < \infty$ exists such that $\hat{P}_{s_i}\{h(X)|\mathscr{F}^{-t}\} \ge P\{h(X)|\mathscr{F}^{-t}\} + \alpha$ for all $s_0 \le t \le s_{i-1}$, $1 \le i \le l$.

LEMMA 6. *Suppose the continuous function $h(x)$ is bounded between min and max. Then for $K \ge 0$, $l \ge 0$ and $0 < \alpha < (max - min)$, we have*

$$P\{B^{l+1}_{\alpha K}(h)|B^l_{\alpha K}(h)\} \le \left(1 - \frac{\alpha}{max - min}\right) \tag{105}$$

and hence, by induction on l,

$$P\{B^l_{\alpha K}(h)\} \le \left(1 - \frac{\alpha}{max - min}\right)^l. \tag{106}$$

PROOF. See the Appendix.

The algorithm of Section 5.1 can be modified to generate estimates $\hat{P}_k = \hat{P}_k(dx|X^{-\sigma_k})$ that converge weakly almost surely to $P(dx|X^-)$. It suffices to use the empirical estimates $\hat{P}_s(dx|\mathscr{F}^{-t})$ rather than $\hat{P}_s(\cdot|J^{-t})$. Also, a new definition of what it means for two distributions on $\mathscr{X}$ to be k-close is needed.

A sequence of probability measures Q_k converges weakly to a distribution Q on $\mathscr{X}$ if $Q_k\{h\} \to Q\{h\}$ for all bounded continuous functions h on $\mathscr{X}$. Since $\mathscr{X}$ is a Polish space, there exists a convergence-determining sequence $\{h_\kappa\}_{0 \le \kappa < \infty}$ of bounded continuous functions on $\mathscr{X}$ such that Q_k converges weakly to Q if and only if $Q_k\{h_\kappa\} \to Q\{h_\kappa\}$ as $k \to \infty$ for all $0 \le \kappa < \infty$. Given a convergence-determining sequence $\{h_\kappa\}_{0 \le \kappa < \infty}$ and a real sequence $\{\varepsilon_k\}_{0 \le k \le \infty}$ such that $\varepsilon_k \searrow 0$, we say that two distributions Q and R on $\mathscr{X}$ are k-close if

$$|Q\{h_\kappa\} - R\{h_\kappa\}| \le \varepsilon_k, \qquad 0 \le \kappa < k.$$

THEOREM 11. *Let $\{X_t\}$ be a stationary ergodic process with values in a Polish space $\mathscr{X}$, and let $P(dx|X^-)$ denote a regular conditional distribution of $X = X_0$ given the infinite past X^-. The estimate $\hat{P}_k = \hat{P}_k(dx|X^{-\sigma_k})$ that is generated during phase k of the modified algorithm will converge weakly almost surely to $P(dx|X^-)$.*

Proof. Let $(s_i)_{0 \le i \le K(k)}$ denote the certificate found and $\hat{P}_k$ the estimate of $P(dx|X^-)$ that is generated during phase k of the algorithm. [Note: The sequence $(s_i)_{0 \le i \le K(k)}$ depends on k, but this dependence is left implicit to keep the notation simple.] It is clear that $\hat{P}_k$ is well defined for all $k \ge 1$, since the search for the integers n and K and for the certificate $(s_i)_{0 \le i \le K}$ must terminate by Lemma 5. Furthermore, $\hat{P}_k$ is a function of $X^{-\sigma_k}$, where $\sigma_k(\omega) = s_{K(k,\omega)}(\omega) + s_{K(k,\omega)-1}(\omega)$ is a stopping time.

To prove the theorem, we argue by contradiction. If $\hat{P}_k$ does not converge weakly almost surely to $P(dx|X^-)$, then there exists a bounded continuous function $h(x)$, equal to either $h_\kappa(x)$ or $-h_\kappa(x)$ for some $0 \le \kappa < \infty$, such that $\hat{P}_k\{h(X)\}$ does not converge to $P\{h(X)|X^-\}$. In fact there exist constants min, max and α such that $min \le h(x) \le max$. $0 < \alpha < (max - min)$, and

$$\hat{P}_k\{h(X)\} - P\{h(X)|X^-\} \ge 2\alpha \quad \text{i.o.}$$

If $k > \kappa$, then all the empirical estimates $\hat{P}_{s_i}\{h(X)|\mathscr{F}^{-t}\}$, $s_0 \le t \le s_{i-1}$, $1 \le i \le K(k)$, are within ε_k from $\hat{P}_k\{h(X)\}$. Consequently, if k is sufficiently large so that $k > \kappa$ and $\varepsilon_k < \alpha/2$ and $P\{h(X)|\mathscr{F}^{-t}\}$ is within $\alpha/2$ from $P\{h(X)|X^-\}$ for all $t \ge k$, then

$$\hat{P}_{s_i}\{h(X)|\mathscr{F}^{-t}\} - P\{h(X)|\mathscr{F}^{-t}\} \ge \alpha, \qquad s_0 \le t \le s_{i-1}, 1 \le i \le K(k).$$

It follows that the bad event $B^{K(k)}_{\alpha K(k)}(h)$ must occur for infinitely many values of k. If K is fixed, then $K(k) \le K$ for only finitely many rounds k, since $K(k) \ge k$. Thus $B^K_{\alpha K}(h)$ must occur for infinitely many values of K. This can only happen with zero probability, in view of the Borel–Cantelli lemma and Lemma 6 which implies that

$$\sum_{K \ge 0} P\{B^K_{\alpha K}(h)\} \le \sum_{K \ge 0} \left(1 - \frac{\alpha}{max - min}\right)^K$$
$$= \frac{(max - min)}{\alpha} < \infty. \qquad \square$$

Remark 1. For any fixed $l \ge 1$ it is possible to learn the conditional distribution $P(dx^l|X^-)$ of the next l-block $X^l = (X_0, \dots, X_{l-1})$ given the infinite past X^-. In fact, it is possible to learn these distributions simultaneously for different values of l. Indeed, let $P(dx^l|\mathscr{F}^{-t})$ denote the conditional distribution of X^l given $\mathscr{F}^{-t}$, and let $\hat{P}_s(dx^l|\mathscr{F}^{-t})$ denote the empirical estimate that is computed on the basis of X^{-s-t} as

$$\hat{P}_s(dx^l|\mathscr{F}^{-t}) = \frac{\delta_{\xi^l_{-l+1}}(dx^l) + \sum_{\tau \in I_s^{-t}(l)} \delta_{X^l_{-\tau}}(dx^l)}{1 + \|I_s^{-t}(l)\|}. \tag{107}$$

Here ξ^l_{-l+1} denotes the l-tuple $(\xi_{-l+1}, \dots, \xi_0)$ at the end of a fixed sequence $(\dots, \xi_{-1}, \xi_0)$, $X^l_{-\tau} = (X_{-\tau}, \dots, X_{-\tau+l-1})$, and $I_s^{-t}(l) = I_s^{-t} \cap \{\tau\colon l \le \tau\}$. The following steps for fixed $k \ge 1$ yield estimates $\hat{P}_k(dx^l|X^{-\sigma_k})$ of the conditional distribution $P(dx^l|X^-)$, for all values l in the range $1 \le l \le k$. We assume that $\varepsilon_k \searrow 0$, and for each $k \geq 1$ that $h_0^{(k)}, h_1^{(k)}, \dots$ is a convergence-determining sequence of bounded continuous functions on $\mathscr{X}^k$.

ALGORITHM TO GENERATE $\hat{P}_k(dx^l | X^{-\sigma_k})$ FOR $1 \leq l \leq k$.

1. Find the least integer n for which there exists an integer K and a certificate $(s_i)_{0 \leq i \leq K}$ such that $k \leq K = s_0 < s_1 < \cdots < s_K = n$ and, for $1 \leq l \leq k$ and $0 \leq \lambda \leq k - l$, all empirical estimates $\hat{P}_{s_i}\{h_\lambda^{(l)}(X^l) | \mathscr{F}^{-t}\} = \int_{\mathscr{X}^l} h_\lambda^{(l)}(x^l) \hat{P}_{s_i}(dx^l | \mathscr{F}^{-t})$, $s_0 \leq t \leq s_{i-1}$, $1 \leq i \leq K$, are well defined and within ε_k from each other.
2. Choose K smallest possible and choose the certificate $(s_i)_{0 \leq i \leq K}$ such that it is lexicographically smallest when read in reverse.
3. For $1 \leq l \leq k$, set $\sigma_k := s_K + s_{K-1}$ and $\hat{P}_k(dx^l | X^{-\sigma_k}) := \hat{P}_{s_K}(dx^l | X^{-s_{K-1}})$.

If l is fixed, then the estimate $\hat{P}_k(dx^l | X^{-\sigma_k})$ will converge weakly almost surely as $k \to \infty$ to the true conditional distribution $P(dx^l | X^-)$. Thus it is possible to learn the conditional distribution of the entire future given the infinite past.

REMARK 2. Suppose $\{(X_t, Y_t)\}$ is a stationary ergodic pair process with values $\mathscr{X} \times \mathscr{Y}$, where $\mathscr{X}$ and $\mathscr{Y}$ are Polish spaces. It is possible to generate estimates $\hat{P}_k$ that converge weakly almost surely to the conditional distribution $P(dx | YY^-X^-)$ of $X = X_0$ given the side information $Y = Y_0$ and the infinite past Y^-X^-. The same method works as before, provided $\mathscr{F}^{-t}$ is now defined as a finite approximating field for the σ-field $\sigma(YY^{-t}X^{-t})$ with limit $\sigma(YY^-X^-)$.

APPENDIX

In this appendix we shall prove Lemma 6. Lemma 4 for finite $\mathscr{X}$ will follow as a special case, by setting h equal to the indicator function δ_x of elements $x \in \mathscr{X}$. First, we make some preliminary observations.

Recall that $B_{\alpha K}^0(h) = \Omega$, and for $l \geq 1$ that $B_{\alpha K}^l(h)$ occurs if there exists a witnessing sequence $(s_i)_{0 \leq i \leq l}$ with $K = s_0 < s_1 < \cdots < s_l$ such that $\|I_{s_i}^{-t}\| > 0$ and

$$\hat{P}_{s_i}\{h(X) | \mathscr{F}^{-t}\} \geq P\{h(X) | \mathscr{F}^{-t}\} + \alpha, \qquad s_0 \leq t \leq s_{i-1}, 1 \leq i \leq l.$$

Here

$$\hat{P}_s\{h(X) | \mathscr{F}^{-t}\} = \int_{\mathscr{X}} h(x) \hat{P}_s(dx | \mathscr{F}^{-t}) = \frac{h(\xi_0) + \sum_{\tau \in I_s^{-t}} h(X_{-\tau})}{1 + \|I_s^{-t}\|}$$

is an empirical estimate for $P\{h(X) | \mathscr{F}^{-t}\} = \int_{\mathscr{X}} h(x) P(dx | \mathscr{F}^{-t})$, and

$$I_s^{-t} = \{\tau \colon 1 \leq \tau \leq s;\ \omega \text{ and } T^{-\tau}\omega \text{ belong to the same atom of } \mathscr{F}^{-t}\}.$$

If $(s_i)_{0 \leq i \leq l}$ is a witness of $B_{\alpha K}^l(h)$, then the prefix $(s_i)_{0 \leq i \leq \lambda}$ is a witness of $B_{\alpha K}^\lambda(h)$, for any λ in the range $1 \leq \lambda \leq l$. Among all witnesses of $B_{\alpha K}^l(h)$, we always select the sequence $(s_i)_{0 \leq i \leq l}$ that is lexicographically smallest when read in reverse. The prefix $(s_i)_{0 \le i \le \lambda}$ is then automatically lexicographically smallest, when read in reverse, among witnesses of $B^{\lambda}_{\alpha K}(h)$. In fact, s_λ is the smallest integer s for which a sequence $(r_i)_{0 \le i \le \lambda}$ with $K = r_0 < r_1 < \cdots < r_\lambda = s$ exists such that

$$\hat{P}_{r_i}\{h(X)|\mathscr{F}^{-t}\} \ge P\{h(X)|\mathscr{F}^{-t}\} + \alpha \quad \text{for } r_0 \le t \le r_{i-1}, 1 \le i \le \lambda.$$

[We set $s_\lambda = \infty$ if $\omega \notin B^{\lambda}_{\alpha K}(h)$.] Indeed, $s \le s_\lambda$ since the lexicographically smallest $(s_i)_{0 \le i \le l}$ yields by truncation a sequence $(s_i)_{0 \le i \le \lambda}$ that is a candidate for $(r_i)_{0 \le i \le \lambda}$, and $s \le r_\lambda$ for all candidate sequences $(r_i)_{0 \le i \le \lambda}$. The inequality $s < s_\lambda$ cannot be strict, since otherwise there would exist a candidate sequence $(r_i)_{0 \le i \le \lambda}$ with $s = r_\lambda < s_\lambda$ that could be padded to a sequence $(r_0, r_1, \dots, r_\lambda = s, s_{\lambda+1}, \dots, s_l)$ witnessing occurrence of $B^{l}_{\alpha K}(h)$. Since we may assume without loss of generality that λ is smallest possible and hence that $r_i = s_i$ for $0 \le i < \lambda$, this would contradict the minimality of $(s_i)_{0 \le i \le l}$ in reverse lexicographic order.

Proof of Lemma 6. Given a continuous function h on $\mathscr{X}$ that is bounded between min and max, a number α such that $0 < \alpha < (max - min)$, and integers $K \ge 0$ and $l \ge 0$, we must prove that

$$P\{B^{l+1}_{\alpha K}(h)|B^{l}_{\alpha K}(h)\} \le \left(1 - \frac{\alpha}{max - min}\right).$$

We consider a countable partition $\mathscr{W}^{l}_{\alpha K}(h)$ of $B^{l}_{\alpha K}(h)$ and prove for every atom W of $\mathscr{W}^{l}_{\alpha K}(h)$ that

$$P\{B^{l+1}_{\alpha K}(h)|W\} \le \left(1 - \frac{\alpha}{max - min}\right).$$

That will establish the desired result since the inequality for each term will imply the inequality for the weighted average

$$P\{B^{l+1}_{\alpha K}(h)|B^{l}_{\alpha K}(h)\} = \sum_{W \in \mathscr{W}^{l}_{\alpha K}(h)} P\{B^{l+1}_{\alpha K}(h)|W\}P\{W|B^{l}_{\alpha K}(h)\}.$$

The countable partition $\mathscr{W}^{l}_{\alpha K}(h)$ of $B^{l}_{\alpha K}(h)$ is defined as follows. We say that two realizations ω and ω' in $B^{l}_{\alpha K}(h)$ belong to the same atom of $\mathscr{W}^{l}_{\alpha K}(h)$ if:

(i) the same certificate $(s_i)_{0 \le i \le l}$ witnesses membership of ω and ω' in $B^{l}_{\alpha K}(h)$;

(ii) ω and ω' belong to the same atom of the finite field $\mathscr{F}^{-\sigma_l}$, where $\sigma_l = s_l + s_{l-1}$.

Let $W = W(\omega)$ denote the atom of partition $\mathscr{W}^{l}_{\alpha K}(h)$ that contains the actual realization ω, assuming $\omega \in B^{l}_{\alpha K}(h)$. Clearly, W is a cylinder set in $\mathscr{F}^{-\sigma_l}$, since all evidence proving that $\omega \in B^{l}_{\alpha K}(h)$ is contained in $\mathscr{F}^{-\sigma_l}$. The atom $W(\omega)$ is uniquely determined by the certificate $(s_i)_{0 \le i \le l}$ and the atoms of the $\mathscr{X}$-partition $\mathscr{I}_{\sigma_l}$ in which the successive outcomes $X_{-\sigma_l}, \dots, X_{-1}$ happen to fall. In particular, if $\mathscr{X}$ is finite and $\mathscr{I}_j = \sigma(X)$ for all j, then $W(\omega)$ is uniquely determined by the sequence $X^{-\sigma_l}$.

An integer $\tau \geq 1$ will be called an occurrence of $W = W(\omega)$ if $T^{-\tau}\omega \in W$, that is, if ω and $T^{-\tau}\omega$ belong to the same atom W of the partition $\mathscr{W}^{l}_{\alpha K}(h)$. An occurrence τ of W is said to be extensible to an occurrence of $B^{l+1}_{\alpha K}(h)$ if $T^{-\tau}\omega \in W \cap B^{l+1}_{\alpha K}(h)$. In this case we designate by $\sigma^{(\tau)}_{l+1}$ the length of the extension. Thus $\sigma^{(\tau)}_{l+1} = s^{(\tau)}_{l+1} + s_l$, where $s^{(\tau)}_{l+1}$ is the smallest integer that must be appended to the certificate $(s_i)_{1 \leq i \leq l}$ to get a sequence that witnesses membership of $T^{-\tau}\omega$ in $W \cap B^{l+1}_{\alpha K}(h)$.

We say that an occurrence τ of W is covered by an occurrence t of $W \cap B^{l+1}_{\alpha K}(h)$ if the interval $[\tau + 1, \tau + \sigma_l]$ is wholly contained in the interval $[t + 1, t + \sigma^{(t)}_{l+1}]$.

If $\omega \in W \cap B^{l+1}_{\alpha K}(h)$, then the empirical estimate $\hat{P}_{s_{l+1}}\{h(X)|\mathscr{F}^{-s_l}\}$ exceeds the conditional expectation $P\{h(X)|\mathscr{F}^{-s_l}\}$ by at least α. Thus the empirical estimate for the shifted sequence $T^{-t}\omega$, computed for each occurrence t of $W \cap B^{l+1}_{\alpha K}(h)$, exceeds by at least α the long run average $P\{h(X)|\mathscr{F}^{-s_l}\}$. Since h is bounded between *min* and *max*, it follows that occurrences of $W \cap B^{l+1}_{\alpha K}(h)$ must occur with limiting frequency less than $(1 - \alpha/(max - min))$ among occurrences of W. We give a formal justification of this fact, proving that $P\{B^{l+1}_{\alpha K}(h)|W\} \leq (1 - \alpha/max - min))$.

Let E denote the set of occurrences of W and $F \subseteq E$ the subset of those occurrences of W that are extensible to occurrences of $B^{l+1}_{\alpha K}(h)$. Thus

$$E = \{\tau \geq 1 \colon T^{-\tau}\omega \in W\},$$

$$F = \{t \geq 1 \colon T^{-t}\omega \in W \cap B^{l+1}_{\alpha K}(h)\}.$$

We use a greedy strategy to extract from F a sequence $t_0 < t_1 < t_2 < \cdots$ such that no instance τ of W is covered by more than one occurrence t_i of $W \cap B^{l+1}_{\alpha K}(h)$. Thus $t_0 = \inf\{t\colon t \in F\}$ and t_{i+1} is the first occurrence t of $W \cap B^{l+1}_{\alpha K}(h)$ such that $t > t_i$ and the occurrence of W at t is not covered by the occurrence of $W \cap B^{l+1}_{\alpha K}(h)$ at t_i:

$$t_{i+1} = \inf\{t \in F \colon t > t_i, t + \sigma_l > t_i + \sigma^{(t_i)}_{l+1}\}.$$

Let G denote the set of occurrences of W that are completely covered by one of the occurrences t_i of $W \cap B^{l+1}_{\alpha K}(h)$ and let H denote the remaining set:

$$G = \{\tau \in E \colon \exists\, i \geq 0 \colon \tau \geq t_i \text{ and } \tau + \sigma_l \leq t_i + \sigma^{(t_i)}_{l+1}\},$$

$$H = \{\tau \in E \colon \tau \notin G\}.$$

Note that $F \subseteq G$ since every occurrence of $W \cap B^{l+1}_{\alpha K}(h)$ is an occurrence of W that is covered by exactly one of the occurrences t_i of $W \cap B^{l+1}_{\alpha K}(h)$.

Let $E(N) = E \cap \{\tau\colon 1 \leq \tau \leq N\}$ and similarly define $G(N)$ and $H(N)$ for $1 \leq N < \infty$. Let $A_E(N)$, $A_G(N)$ and $A_H(N)$ denote the average of $h(X_{-\tau})$ as τ ranges over $E(N)$, $G(N)$ and $H(N)$, respectively. Thus

$$A_E(N) = \frac{1}{\|E(N)\|} \sum_{\tau \in E(N)} h(X_{-\tau})$$

and similarly for $A_G(N)$ and $A_H(N)$. Obviously,

$$\|E(N)\| = \|G(N)\| + \|H(N)\|,$$
$$\|E(N)\|A_E(N) = \|G(N)\|A_G(N) + \|H(N)\|A_H(N).$$

If $\omega \in W \cap B^{l+1}_{\alpha K}(h)$, then $\hat{P}_{s_{l+1}}\{h(X)|\mathscr{F}^{-s_l}\} \geq P\{h(X)|\mathscr{F}^{-s_l}\} + \alpha$. When this last inequality is applied within the occurrences t_i of $W \cap B^{l+1}_{\alpha K}(h)$, we obtain (after summing the numerators and the denominators of the empirical estimates)

$$A_G(N) \geq P\{h(X)|\mathscr{F}^{-s_l}\} + \alpha.$$

But $A_E(N) \to P\{h(X)|\mathscr{F}^{-s_l}\}$ by the ergodic theorem, so $\liminf_N (A_G(N) - A_E(N)) \geq \alpha$. Since $A_G(N)$, $A_H(N)$ and $A_E(N)$ are bounded between *min* and *max*, we obtain

$$\begin{aligned}
\limsup_N \frac{\|G(N)\|}{\|E(N)\|} &= \limsup_N \frac{(A_E - A_H)}{(A_G - A_H)} \\
&= \limsup_N \frac{(A_E - A_H)}{(A_G - A_E) + (A_E - A_H)} \\
&\leq \limsup_N \frac{(A_E - A_H)}{\alpha + (A_E - A_H)} \\
&\leq \frac{(max - min - \alpha)}{\alpha + (max - min - \alpha)} = \left(1 - \frac{\alpha}{max - min}\right).
\end{aligned}$$

The first inequality holds since $\liminf_N (A_G - A_E) \geq \alpha$, and the second inequality holds since $Z/(\alpha + Z)$ is monotonically increasing in $Z = (A_E - A_H)$ and

$$\begin{aligned}
\limsup_N (A_E - A_H) &\leq \limsup_N (A_G - A_H) - \liminf_N (A_G - A_E) \\
&\leq (max - min) - \alpha.
\end{aligned}$$

Finally, $\|F(N)\| \leq \|G(N)\|$ so that, again by the ergodic theorem,

$$P\{B^{l+1}_{\alpha K}(h)|W\} = \lim_N \frac{\|F(N)\|}{\|E(N)\|} \leq \limsup_N \frac{\|G(N)\|}{\|E(N)\|} \leq \left(1 - \frac{\alpha}{max - min}\right).$$

This concludes the proof of Lemma 6. □

Acknowledgment. I thank Tom Cover for many discussions on log-optimum investment and for bringing the thesis of David Bailey to my attention.

REFERENCES

Algoet, P. H. and Cover, T. M. (1988a). Asymptotic optimality and asymptotic equipartition properties of log-optimum investment. *Ann. Probab.* **16** 876–898.

Algoet, P. H. and Cover, T. M. (1988b). A sandwich proof of the Shannon–McMillan–Breiman theorem. *Ann. Probab.* **16** 899–909.

Bailey, D. H. (1976). Sequential schemes for classifying and predicting ergodic processes. Thesis, Dept. Mathematics, Stanford Univ.

Barron, A. R. (1985). Logically smooth density estimation. Ph.D. dissertation, Dept. Electrical Engineering, Stanford Univ.

Bartlett, M. S. (1951). The frequency goodness of fit test for probability chains. *Proc. Cambridge Philos. Soc.* **47** 86–95.

Bell, R. and Cover, T. M. (1980). Competitive optimality of logarithmic investment. *Math. Oper. Res.* **5** 161–166.

Bell, R. and Cover, T. M. (1988). Game-theoretic optimal portfolios. *Management Sci.* **34** 724–733.

Breiman, L. (1957). The individual ergodic theorem of information theory. *Ann. Math. Statist.* **28** 809–811. [Correction (1960) **31** 809–810.]

Breiman, L. (1961). Optimal gambling systems for favorable games. *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* **1** 65–78. Univ. California Press, Berkeley.

Cover, T. M. (1974). Universal gambling schemes and the complexity measures of Kolmogorov and Chaitin. Technical Report 12, Dept. Statistics, Stanford Univ.

Cover, T. M. (1975). Open problems in information theory. In *Moscow Information Theory Workshop*. IEEE Press, New York.

Cover, T. M. (1991). Universal portfolios. *Math. Finance* **1** 1–29.

Cover, T. M. and King, R. C. (1978). A convergent gambling estimate of the entropy of English. *IEEE Trans. Inform. Theory* **24** 413–421.

Cover, T. M. and Thomas, J. A. (1991). *The Elements of Information Theory*. Wiley, New York.

Feder, M. (1991). Gambling using a finite state machine. *IEEE Trans. Inform. Theory* **37** 1459–1465.

Jelinek, F. (1968). *Probabilistic Information Theory*. McGraw-Hill, New York.

Kelly, J. L., Jr. (1956). A new interpretation of information rate. *Bell System Tech. J.* **35** 917–926.

Langdon, G. G., Jr. (1983). A note on the Ziv–Lempel model for compressing individual sequences. *IEEE Trans. Inform. Theory* **29** 284–287.

Levin, L. A. and Zhvonkin, A. K. (1970). The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms. *Russian Math. Surveys* **25** 83–124.

Móri, T. F. (1984). Asymptotic properties of the empirical strategy in favourable stochastic games. In *Limit Theorems in Probability and Statistics*. (P. Révész, ed.) *Colloq. Math. Soc. János Bolyai* **36** 777–790. North-Holland, Amsterdam.

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

Ornstein, D. S. and Shields, P. C. (1990). Universal almost sure data compression. *Ann. Probab.* **18** 441–452.

Ornstein, D. S. and Weiss, B. (1990). Entropy and data compression. Preprint.

Pasco, R. C. (1976). Source coding algorithms for fast data compression. Ph.D. dissertation, Dept. Electrical Engineering, Stanford Univ.

Rissanen, J. (1983). A universal data compression system. *IEEE Trans. Inform. Theory* **29** 656–664.

Rissanen, J. and Langdon, G. G. (1979). Arithmetic coding. *IBM J. Res. Dev.* **23** 149–162.

Rissanen, J. and Langdon, G. G. (1981). Universal modeling and coding. *IEEE Trans. Inform. Theory* **27** 12–23.

Wyner, A. D. and Ziv, J. (1990). On entropy and data compression. Unpublished manuscript.

Ziv, J. and Lempel, A. (1978). Compression of individual sequences via variable rate coding. *IEEE Trans. Inform. Theory* **24** 530–536.

IBM Almaden Research Center
Department K65/802
650 Harry Road
San Jose, California 95120-6099