On Optimal Sequential Prediction for General Processes

Andrew B. Nobel *

July 10, 2002

Abstract

This paper considers several aspects of the sequential prediction problem for unbounded, non-stationary processes under p-th power loss $\ell_p(u,v) = |u-v|^p$, $1 < p < \infty$. In the first part of the paper it is shown that Bayes prediction schemes are Cesaro optimal under general conditions, that Cesaro optimal prediction schemes are unique in a natural sense, and that Cesaro optimality is equivalent to a form of weak calibration. Connections between calibration and stronger forms of optimality are briefly considered. Extensions of the existence and uniqueness results to generalized prediction, and prediction from observations with additive noise, are established. For binary processes, it is shown that thresholding an optimal prediction scheme for the squared loss yields an optimal binary prediction scheme for the Hamming loss.

In the second part of the paper, it is shown how to construct, from a countable family of prediction schemes, a single composite scheme whose asymptotic performance on any suitable process dominates the performance of each member of the family. The construction is based on aggregating methods for individual binary sequences. Using the construction some results of Algoet on the existence of Cesaro optimal schemes for families of ergodic processes are rederived in a direct way and extended to unbounded processes.

Appears in the IEEE Transactions on Information Theory, vol. 49, pp.83-98, 2003.

Key words and phrases: Sequential prediction, stochastic process, non-stationary process, Bayes prediction scheme, p'th power loss.

*Andrew Nobel is with the Department of Statistics, University of North Carolina, Chapel Hill, NC 27599. Email: nobel@stat.unc.edu. His work was supported in part by NSF Grant DMS 9971964.

1 Introduction

The subject of this paper is stochastic sequential prediction. In this problem, the elements of a real-valued stochastic process $\mathbf{X} = X_1, X_2, \ldots$ are revealed to a forecaster, one at a time, beginning with X_1. At each time $t \geq 1$ the forecaster makes a real-valued prediction F_t of X_t, based on the observed values of $X^{t-1} = X_1, \ldots, X_{t-1}$. When X_t is revealed, the forecaster incurs a non-negative loss $\ell(F_t, X_t)$. We focus here on unbounded, non-stationary processes and restrict our attention throughout to p'th power loss of the form $\ell_p(u, v) = |u - v|^p$, with $1 < p < \infty$.

Let $\mathbb{R}$ denote the real line, and let $\mathbb{R}^* = \{\emptyset\} \cup \bigcup_{j=1}^{\infty} \mathbb{R}^j$ be the collection of all finite length sequences of real numbers, where a sequence of length zero is represented by the empty set. A prediction scheme is a map $F : \mathbb{R}^* \to \mathbb{R}$. It is assumed in what follows that, for each $j \geq 1$, the restriction of F to j-tuples $x^j = x_1, \ldots, x_j$ is a measurable function from $\mathbb{R}^j$ to $\mathbb{R}$. Each prediction scheme F represents a deterministic strategy for the prediction problem: having observed the past values X^{t-1} of a given process, the scheme makes a prediction $F(X^{t-1})$ of the next value X_t. The value of $F(X^{t-1})$ does not depend on side information or on auxiliary randomization.

Assume for the moment that a loss function $\ell = \ell_p$ has been fixed. If a scheme F is applied successively to the first n terms $X^n = X_1, \ldots, X_n$ of a process $\mathbf{X}$, its average cumulative loss is a random variable, denoted by

$$L_n(F) \;=\; L_n(F, \mathbf{X}) \;=\; \frac{1}{n} \sum_{t=1}^{n} \ell(F(X^{t-1}), X_t).$$

Of central interest here are prediction schemes having small long-run average cumulative loss. The following notion of optimality is considered, for example, in [2, 13, 18].

Definition: A prediction scheme G is Cesaro optimal, or optimal in the long run average sense, for a process $\mathbf{X}$ if

$$\liminf_{n \to \infty} \left[L_n(F, \mathbf{X}) \;-\; L_n(G, \mathbf{X}) \right] \;\geq\; 0 \;\; \text{wp1}$$

for every measurable prediction scheme F. A prediction scheme G is Cesaro optimal for a family $\mathcal{X}$ of processes if it is Cesaro optimal for every process in $\mathcal{X}$.

By definition, a prediction scheme G is Cesaro optimal for a process $\mathbf{X}$ if its average cumulative loss is, asymptotically, less than or equal to the average cumulative loss of any competing scheme on the same process. Note that the definition does not require that

the quantities $L_n(G, \mathbf{X})$ or $L_n(F, \mathbf{X})$ converge as n tends to infinity. The notion of Cesaro optimality is somewhat weak, as it requires only that an optimal scheme perform well on the average. One may show, for example, that Cesaro optimal schemes exist for any countable family of processes $\mathcal{X}$ (see, e.g., Foster [18] or Proposition 6 below). In such cases, stronger criteria of predictive performance (*c.f.* [13, 38, 39]) may be more appropriate. Two such criteria, strong optimality and efficiency, are discussed briefly in Section 5. In seeking prediction schemes that perform well for an *uncountable* family of processes, Cesaro optimality provides a sensible measure of success. As noted in Section 9.2, no decision scheme is strongly optimal for the (uncountable) family of bounded ergodic processes; however, one may construct Cesaro optimal prediction schemes for this family in a variety of ways (see [2] and Theorem 6 below).

Numerous examples of sequential prediction problems for stationary and more general processes can be found in the literature; see for example [2, 38]. A good account of stochastic and non-stochastic sequential decision problems, and their relation to calibration and foundational questions in Statistics, can be found in the work of Dawid [11, 12, 13] and in the more recent work [38, 14, 39]. A thorough treatment of sequential decision and prediction problems for ergodic (and stationary) processes, and many references to related work on time series prediction, can be found in the work of Algoet [2]. Algoet studies general loss functions ℓ for which there exists an envelope $\Lambda(\cdot)$ such that $\ell(u, v) \leq \Lambda(v) < \infty$ for each u and v. The existence of a finite envelope for the p'th power loss ℓ_p requires that each process $\mathbf{X}$ under study take values in a bounded subset of $\mathbb{R}$, an assumption not made in this paper. Algoet's extension in [2] of the stability theorem for martingale differences (see Lemma 1.1 below) plays a central role in our results.

Our application in Section 9 of aggregating methods for individual sequences to stochastic prediction generalizes and extends recent work of Györfi, Lugosi, and Morvai [24], who used aggregating method to define randomized predictors for binary ergodic processes. Related methods were recently applied by Weissman and Merhav [42, 43] to the prediction of individual and ergodic binary sequences. Aggregating methods were applied in a different way by Foster [18] to the prediction of binary processes under the squared error. Generalizations of [24] to sequential prediction of bounded, real valued ergodic processes under the squared error have also been derived, independently, in recent work of Györfi and Lugosi [22].

1.1 Overview

Two preliminary results are presented in the next section. In Section 1.1 the existence and uniqueness of Cesaro optimal decision schemes for processes satisfying suitable population and sample moment conditions are established. In particular, it is shown that the Bayes decision scheme for $\mathbf{X}$ is Cesaro optimal, and that any two Cesaro optimal schemes are, in a natural sense, equivalent. Extensions of these results to generalized prediction and to prediction from observations with additive noise are established in Sections 4.1 and 4.2, respectively. Two alternative forms of optimality are described in Section 5. In Section 6 it is shown that Cesaro optimality under the squared loss is equivalent to a form of weak calibration, and that a stronger form of calibration, considered by Dawid and others, is equivalent to a stronger form of optimality. Existence, uniqueness, and several other properties of strongly optimal prediction schemes are briefly discussed. In Section 7 it is shown that, by suitably thresholding a prediction scheme that is Cesaro optimal under the squared loss for a binary process, one obtains an optimal prediction scheme under the Hamming loss.

The problem of aggregating prediction schemes is studied in Section 8. Given a countable family of prediction schemes, a composite scheme is constructed whose asymptotic performance dominates that of each member of the family on any suitable process. By appropriate choice of the countable family, some results of Algoet [2] on the existence of universal decision schemes for ergodic processes are rederived and extended to unbounded processes in a direct way in Section 9. In particular, it is shown that for each $p > 1$, there exists a single prediction scheme that is Cesaro optimal under the p'th power loss for any ergodic process $\{X_i\}$ such that $E|X_i|^q < \infty$ for some $q > p$.

2 Preliminary Results

Below we will make repeated use of the following stability result for martingale differences, due to Algoet [2]. A general account of such results can be found in [40]. For bounded Z_t the lemma may be deduced from standard exponential inequalities for martingale difference sequences [26, 5].

Lemma A *Let $X_1, X_2, \ldots$ be any stochastic process, and let $Z_1, Z_2, \ldots \in \mathbb{R}^d$ be random vectors such that, for each $t \geq 1$, Z_t is a measurable function of $X_1, \ldots, X_t$. If $\sup_{t\geq 1} E\psi(|Z_t|) <$*

∞ *where* $\psi(u) = u\log^2(1+u)$, *then*

$$\frac{1}{n}\sum_{t=1}^{n} Z_t \;-\; \frac{1}{n}\sum_{t=1}^{n} E(Z_t \,|\, X^{t-1}) \;\to\; 0 \quad wp1.$$

The following elementary lemma will also be useful.

Lemma 1 *Let* $\phi : [0,\infty) \to [0,\infty)$ *be any function such that* $\phi(x)/x \nearrow \infty$ *as* $x \nearrow \infty$. *If* $a_1, a_2, \ldots$ *are non-negative numbers such that* $n^{-1}\sum_{i=1}^{n}\phi(a_i) \le K < \infty$ *for each* $n \ge 1$, *then as* $c \to \infty$,

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} I\{a_i \ge c\} \;\to\; 0 \quad \textit{and} \quad \limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} a_i\, I\{a_i \ge c\} \;\to\; 0$$

Proof: Let $c > 0$ be so large that $\phi(c)/c \ge 1$. The first claim follows readily, as

$$\frac{1}{n}\sum_{i=1}^{n} I\{a_i \ge c\} \;\le\; \frac{1}{n}\sum_{i=1}^{n} \frac{\phi(a_i)}{a_i} I\{a_i \ge c\} \;\le\; \frac{1}{cn}\sum_{i=1}^{n}\phi(a_i).$$

The second claim is a consequence of the inequalities

$$\frac{1}{n}\sum_{i=1}^{n} a_i I\{a_i \ge c\} \;=\; \frac{1}{n}\sum_{i=1}^{n} \frac{a_i}{\phi(a_i)}\phi(a_i) I\{a_i \ge c\} \;\le\; \frac{c}{\phi(c)n}\sum_{i=1}^{n}\phi(a_i).$$

3 Existence and Uniqueness of Cesaro Optimal Schemes

In this section the existence and uniqueness of Cesaro optimal schemes for general, non-stationary stochastic processes is established. Let $\mathbf{X} = X_1, X_2, \ldots \in \mathbb{R}$ be any process satisfying the following population and sample moment conditions:

(A1) $\sup_{t\ge 1} E|X_t|^p \log^2(1+|X_t|^p) < \infty$

(A2) $\limsup_n n^{-1}\sum_{t=1}^{n}\phi(|X_t|^p) < \infty$ for some function ϕ such that $\phi(x)/x \nearrow \infty$ as $x \nearrow \infty$.

If $\mathbf{X}$ is ergodic, then (A2) follows immediately from (A1) and the ergodic theorem. In general, this implication need not hold.

Definition: The Bayes prediction scheme (c.f. Ferguson [20]) for a process $\mathbf{X}$ under the p'th power loss ℓ_p is defined by

$$B(X^{t-1}) \;=\; \arg\min_{a\in\mathbb{R}} E[\,|X_t - a|^p \,|\, X^{t-1}\,]. \tag{1}$$

At each time t, the Bayes scheme selects the unique prediction minimizing the conditional expected loss of the next outcome given the previous values of the process. It follows readily from (1) that $E|X_t - B(X^{t-1})|^p \leq E|X_t - f(X^{t-1})|^p$ for any measurable function $f : \mathbb{R}^{t-1} \to \mathbb{R}$. In particular, one may view $B(X^{t-1})$ as the projection of X_t onto the space of all random variables $f(X^{t-1})$ such that $E|f(X^{t-1})|^p < \infty$. Ando and Amemiya [1] have studied the general properties of such projections and shown that, for a general increasing sequence of sigma fields, they share the convergence and integrability properties of conditional expectations (see Section 9 for more details). We require a preliminary fact concerning integrability of the Bayes scheme (1); related results can be found in [1].

Lemma 2 *Let B be the Bayes scheme under ℓ_p for a process $\mathbf{X}$ satisfying (A1). Set $B_t = B(X^{t-1})$ and let $\psi(u) = u \log^2(1+u)$. Then for each t,*

(a) $|B_t|^p \leq 2^p \, E[|X_t|^p \mid X^{t-1}]$ *wp1*

(b) $\psi(\, |B_t|^p \,/\, 2^p \,) \leq E[\, \psi(|X_t|^p) \mid X^{t-1}]$ *wp1*

(c) $\sup_{t \geq 1} E\psi(|B_t|^p) < \infty$

Proof: Let $\mathbf{X}$ be a process satisfying (A1). Then for fixed $t \geq 1$,

$$\begin{aligned} |B_t| &= E[\,|(B_t - X_t) + X_t| \mid X^{t-1}\,] \leq E[\,|X_t| \mid X^{t-1}\,] + E[\,|X_t - B_t| \mid X^{t-1}\,] \\ &\leq E[\,|X_t| \mid X^{t-1}\,] + \big(E[\,|X_t - B_t|^p \mid X^{t-1}\,]\big)^{1/p} \\ &\leq E[\,|X_t| \mid X^{t-1}\,] + \big(E[\,|X_t|^p \mid X^{t-1}\,]\big)^{1/p} \end{aligned}$$

The second inequality above is a consequence of the monotonicity of L_p-norms, and the third follows directly from the definition of $B(X^{t-1})$. The elementary inequality $(a+b)^p \leq 2^{p-1}\,(a^p + b^p)$ implies that

$$|B_t|^p \leq 2^{p-1}\,\big(\,E[\,|X_t| \mid X^{t-1}\,]\,\big)^p + 2^{p-1} E[\,|X_t|^p \mid X^{t-1}\,].$$

By Jensen's inequality, the first term on the right is at most $2^{p-1}\,E[\,|X_t|^p \mid X^{t-1}\,]$, and conclusion (a) follows. Inequality (b) follows directly from (a) and the convexity of ψ. Inequality (c) is an immediate consequence of (b) and assumption (A1).

For processes $\mathbf{X} = X_1, X_2, \ldots$ taking values in a bounded interval of the reals, the Cesaro optimality of the Bayes scheme B follows directly from Theorem 3 in [2]. The next theorem shows that the Bayes scheme is Cesaro optimal under the more general conditions (A1) and (A2). In many cases, the Bayes scheme is optimal in much stronger senses (see Section 5 below).

Theorem 1 *(***Existence***) Let* $\mathbf{X}$ *be a stochastic process taking values in* $\mathbb{R}$ *and let* $p > 1$. *If conditions (A1) and (A2) hold then the Bayes prediction scheme*

$$B(X^{t-1}) \;=\; \arg\min_{a\in\mathbb{R}} E[\,|X_t - a|^p \,|\, X^{t-1}\,]. \tag{2}$$

is Cesaro optimal for $\mathbf{X}$ *under the p'th power loss.*

Proof: Let F be any prediction scheme. To simplify notation, let $F_t = F(X^{t-1})$ and $B_t = B(X^{t-1})$. Fix $c > 0$, and define auxiliary schemes $F'_t = F_t\, I\{|F_t| \le c\}$ and $F''_t = F_t\, I\{|F_t| > c\}$ for $t \ge 1$. A routine calculation shows that

$$L_n(F) \;=\; L_n(F') \;+\; L_n(F'') \;-\; \frac{1}{n}\sum_{t=1}^{n} |X_t|^p.$$

Observe that if $|X_t| \le c/3$ and $|F_t| > c$, then $|X_t - F_t| > |X_t|$. This yields the lower bound

$$L_n(F''_t) \;\ge\; \frac{1}{n}\sum_{t=1}^{n} |X_t - F_t\, I\{|F_t| > c\}|^p \cdot I\{|X_t| \le c/3\} \;\ge\; \frac{1}{n}\sum_{t=1}^{n} |X_t|^p\, I\{|X_t| \le c/3\}.$$

It follows from the last two displays that

$$\begin{aligned}
&\liminf_{n\to\infty} [\, L_n(F) - L_n(B)\,] \\
&\quad\ge\; \liminf_{n\to\infty} [\, L_n(F') - L_n(B)\,] \;-\; \limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} |X_t|^p\, I\{|X_t| > c/3\} \quad \text{wp1.}
\end{aligned} \tag{3}$$

We now show that $\{|X_t - B_t|^p : t \ge 1\}$ and $\{|X_t - F'_t|^p : t \ge 1\}$ satisfy the moment condition of Lemma A. As F' is uniformly bounded, the finiteness of $\sup_{t\ge1} E\psi(|X_t - F'_t|^p)$ follows directly from assumption (A1). Moreover,

$$\psi(|X_t - B_t|^p) \;\le\; \psi(2^{p-1}\,|X_t|^p \;+\; 2^{p-1}\,|B_t|^p) \;\le\; \psi(2^p\,|X_t|^p) \;+\; \psi(2^p\,|B_t|^p),$$

and therefore $\sup_{t\ge1} E\psi(|X_t - B_t|^p)$ is finite by (A1) and Lemma 2. Applying Lemma A to $L_n(F')$ and $L_n(B)$ yields the equation

$$\liminf_{n\to\infty} [\, L_n(F'_t) - L_n(B)\,] \;=\; \liminf_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} E[\,|X_t - F'_t|^p - |X_t - B_t|^p \,|\, X^{t-1}] \quad \text{wp1.}$$

The definition of B_t ensures that each term in the last sum is positive with probability one. It then follows from inequality (3) that

$$\liminf_{n\to\infty} [\, L_n(F) - L_n(B)\,] \;\ge\; -\limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} X_t^2 I\{|X_t| > c/3\} \quad \text{wp1.}$$

Letting c tend to infinity, condition (A2) and Lemma 1 imply that the limit supremum tends to zero. As F was arbitrary, B is Cesaro optimal for $\mathbf{X}$.

The next result shows that Cesaro optimal schemes are essentially unique, the form of uniqueness depending on the value of p. Taken together, Theorems 1 and 2 show that every Cesaro optimal scheme for $\mathbf{X}$ under ℓ_p is equivalent to the Bayes scheme B.

Theorem 2 *(***Uniqueness***) Let* $\mathbf{X}$ *satisfy (A1) and (A2). Suppose that* B *is the Bayes scheme defined in (2), and that* F *is any other Cesaro optimal scheme for* $\mathbf{X}$ *under* ℓ_p*. If* $p \geq 2$ *then*

$$\frac{1}{n}\sum_{t=1}^{n} |F_t(X^{t-1}) - B(X^{t-1})|^p \;\to\; 0 \quad \textit{wp1.}$$

If $1 < p < 2$ *then*

$$\frac{1}{n}\sum_{t=1}^{n} |F_t(X^{t-1}) - B(X^{t-1})|^q \;\to\; 0 \quad \textit{wp1}$$

for each $1 \leq q < p$.

Proof: As both F and the B are Cesaro optimal for $\mathbf{X}$,

$$\begin{aligned} 0 \;&\leq\; \liminf_{n\to\infty} [L_n(F) - L_n(B)] \;\leq\; \limsup_{n\to\infty} [L_n(F) - L_n(B)] \\ &=\; -\liminf_{n\to\infty} [L_n(B) - L_n(F)] \;\leq\; 0 \quad \text{wp1.} \end{aligned} \tag{4}$$

Thus $L_n(F) - L_n(B) \to 0$. Define the compound decision scheme $H(X^{t-1}) = (F(X^{t-1}) + B(X^{t-1}))/2$, and write

$$L_n(H) - L_n(B) \;=\; \left[L_n(H) - \frac{1}{2}L_n(F) - \frac{1}{2}L_n(B)\right] \;+\; \frac{1}{2}\left(L_n(F) - L_n(B)\right).$$

It follows from the last equation and (4) that

$$\begin{aligned} \liminf_{n\to\infty} [L_n(H) - L_n(B)] \;&=\; \liminf_{n\to\infty} \left[L_n(H) - \frac{1}{2}L_n(F) - \frac{1}{2}L_n(B)\right] \\ &=\; \liminf_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} -\Gamma(F_t - X_t, B_t - X_t), \end{aligned} \tag{5}$$

where

$$\Gamma(u,v) \;:=\; \frac{|u|^p}{2} \;+\; \frac{|v|^p}{2} \;-\; \left|\frac{(u+v)}{2}\right|^p. \tag{6}$$

If $p \geq 2$ then $|a+b|^p \;+\; |a-b|^p \;\geq\; 2(|a|^p + |b|^p)$ for each $a, b \in \mathbb{R}$ (c.f. Royden [36], Lemma 22). Setting $a = (u+v)/2$, $b = (u-v)/2$, and rearranging terms shows that $\Gamma(u,v) \geq 2^{-p}\,|u-v|^p$. It then follows from (5) that

$$\begin{aligned} \liminf_{n\to\infty} [L_n(H) - L_n(B)] \;&\leq\; \liminf_{n\to\infty} \frac{-1}{n}\sum_{t=1}^{n} 2^{-p}\,|F_t - B_t|^p \\ &=\; -2^{-p}\,\limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} |F_t - B_t|^p. \end{aligned}$$

If $\limsup_n n^{-1} \sum_{t=1}^n |F_t - B_t|^p$ is positive with positive probability then, by the above inequality, B fails to be Cesaro consistent, which contradicts Theorem 1. The case $1 < p < 2$ is considered in Section 10.1.

Example: We present a simple example here to illustrate Theorems 1 and 2. Let $\phi : \mathbb{R} \to \mathbb{R}$ be a measurable, nonlinear function, and suppose for simplicity that the range of ϕ is bounded. Let $\{\varepsilon_i\}$ be i.i.d. with $E\varepsilon_i = 0$ and $E|\varepsilon_i|^2 \log^2(1 + |\varepsilon_i|^2) < \infty$. Let X_0 be any random variable independent of $\{\varepsilon_i\}$ and for $t \geq 1$ define X_t via the recursion

$$X_t \;=\; \phi(X_{t-1}) \,+\, \varepsilon_{t-1}$$

Then $\mathbf{X} = \{X_t\}$ is a (possibly non-stationary) nonlinear AR(1) process. Under the squared loss, the Bayes prediction scheme for $\mathbf{X}$ is given by $B(X^{t-1}) = \phi(X_{t-1})$. By Theorem 1, B is Cesaro optimal for $\mathbf{X}$ and, as expected, the limiting average cumulative loss of any scheme is bounded below by $\lim_n L_n(B, \mathbf{X}) = E\varepsilon_i^2$. Theorem 2 implies that if F is any Cesaro optimal prediction scheme for $\mathbf{X}$, then $n^{-1} \sum_{t=1}^n (F(X^{t-1}) - \phi(X_{t-1}))^2 \to 0$ with probability one.

4 Two Extensions

4.1 Generalized Prediction

In the generalized prediction problem, the goal is to determine from past observations the value of a known function of the next observation, rather than the next observation itself. Let $g : \mathbb{R} \to \mathbb{R}$, and let $\mathbf{X} = X_1, X_2, \ldots \in \mathbb{R}$ be a given stochastic process. Suppose that, having observed $X_1, \ldots, X_{t-1}$, we wish to predict the value of $g(X_t)$ in such a way as to minimize the long run average p'th power loss. The prediction problem considered above corresponds to the special case where $g(x) = x$. In the generalized prediction problem, the average performance of a prediction scheme $F : \mathbb{R}^* \to \mathbb{R}$ over n time units is given by

$$L_n^g(F, \mathbf{X}) \;=\; \frac{1}{n} \sum_{t=1}^n |g(X_t) - F(X^{t-1})|^p.$$

A prediction scheme G is Cesaro optimal for $(\mathbf{X}, g)$ if

$$\liminf_{n\to\infty} \left[L_n^g(F, X^n) \;-\; L_n^g(G, X^n)\right] \;\geq\; 0 \;\; \text{wp1}$$

for every measurable prediction scheme F. Fix $g : \mathbb{R} \to \mathbb{R}$ and $p > 1$, and let $\mathbf{X}$ be a real-valued process such that

(A1') $\sup_{t\geq 1} E\psi(|g(X_t)|^p) < \infty$ where $\psi(u) = u\log^2(1+u)$.

(A2') $\limsup_n n^{-1}\sum_{t=1}^n \phi(|g(X_t)|^p) < \infty$ for some function ϕ such that $\phi(x)/x \nearrow \infty$ as $x \nearrow \infty$.

The generalized Bayes prediction scheme for $(\mathbf{X}, g)$ under ℓ_p is defined by

$$B^g(X^{t-1}) \;=\; \arg\min_{a\in\mathbb{R}} E[\,|g(X_t)-a|^p \,|\, X^{t-1}\,]. \tag{7}$$

The next result is an extension of Theorems 1 and 2 to the problem of generalized prediction. Its proof is vitually the same, so we omit the details.

Theorem 3 *If (A1') and (A2') hold then the generalized Bayes scheme (7) is Cesaro optimal for $(\mathbf{X}, g)$ under ℓ_p. Let F be any Cesaro optimal prediction scheme for $(\mathbf{X}, g)$ under ℓ_p. If $p \geq 2$ then $n^{-1}\sum_{t=1}^n |F_t - B_t^g|^p \to 0$ with probability one. If $1 < p < 2$ then $n^{-1}\sum_{t=1}^n |F_t - B_t^g|^q \to 0$ with probability one for all $1 \leq q < p$.*

4.2 Prediction from Observations with Additive Noise

Suppose now that $p = 2$. Here we consider a variant of the prediction problem in which the forecaster does not have direct access to the values of the process $\mathbf{X}$, but to noisy observations of the form

$$Y_t \;=\; X_t + N_t \quad t \geq 1, \tag{8}$$

where $\mathbf{N} = N_1, N_2, N_3, \ldots$ are zero mean random variables, defined on the same probability space as $\mathbf{X}$. In particular, we assume that

(N1) $\mathbf{N}$ and $\mathbf{X}$ are independent;

(N2) $\mathbf{N}$ is a martingale difference sequence, *i.e.*, $E(N_t \,|\, N_1^{t-1}) = 0$ wp1 for $t \geq 1$;

(N3) $\sup_{t\geq 1} E\psi(|N_t|^2) < \infty$, where $\psi(u) = u\log^2(1+u)$.

Let $\mathbf{Y} = Y_1, Y_2, \ldots$ be the available sequence of noisy observations; it follows from (A1) and (N3) that $\sup_{t\geq 1} E\psi(|Y_t|^2) < \infty$. Suppose that the performance of a scheme $F : \mathbb{R}^* \to \mathbb{R}$ over n time units is measured by its average squared loss:

$$\tilde{L}_n(F) \;=\; \tilde{L}_n(F, \mathbf{X}, \mathbf{Y}) \;=\; \frac{1}{n}\sum_{t=1}^n (X_t - F(Y^{t-1}))^2.$$

A prediction scheme G will be called Cesaro optimal for $(\mathbf{X}, \mathbf{Y})$ if for every (measurable) prediction scheme F,

$$\liminf_{n\to\infty} \left[\tilde{L}_n(F) \;-\; \tilde{L}_n(G)\right] \;\geq\; 0 \;\; \text{wp1}.$$

Under the squared loss, the Bayes prediction scheme for $\mathbf{X}$ based on $\mathbf{Y}$ is given by the conditional expectation $\tilde{B}(Y^{t-1}) := E[\,X_t \,|\, Y^{t-1}\,]$. One may establish using (N1) and (N2) that $E[\,N_t \,|\, Y^{t-1}\,] = 0$, and therefore $\tilde{B}(Y^{t-1}) = E[\,Y_t \,|\, Y^{t-1}\,]$ for each $t \geq 1$. Thus $\tilde{B}$ coincides with the Bayes prediction scheme B for $\mathbf{Y}$ under ℓ_2.

In recent work Weissman and Merhav [42, 43] studied prediction of individual and ergodic binary sequences in the presence of noise under a variety of loss functions. In [42] they exhibited Cesaro optimal schemes for several different noise models when the joint process of clean and noisy observations is ergodic and satisfies a conditional mixing condition. For additive noise satisfying (N1)-(N3), the existence and uniqueness of Cesaro optimal prediction schemes holds under very general conditions.

Proposition 1 *Suppose that (N1)-(N3) hold, and that (A1)-(A2) hold with $p = 2$. Then the Bayes scheme $\tilde{B}$ is Cesaro optimal for $(\mathbf{X}, \mathbf{Y})$. If F is any other Cesaro optimal prediction scheme for $(\mathbf{X}, \mathbf{Y})$, then $n^{-1} \sum_{t=1}^{n} |F_t - \tilde{B}_t|^2 \to 0$ with probability one.*

Proof: The proof follows that of Theorem 1. Let F be any prediction scheme and fix $c > 0$. Define $F_t = F(Y^{t-1})$, $F'_t = F_t\, I\{|F_t| \leq c\}$, and let $B_t = \tilde{B}(Y^{t-1}) = B(Y^{t-1})$. By arguments like those leading to (3) above, one finds that

$$\begin{aligned} &\liminf_{n\to\infty} [\,\tilde{L}_n(F) - \tilde{L}_n(\tilde{B})\,] \\ &\quad \geq \liminf_{n\to\infty} [\,\tilde{L}_n(F') - \tilde{L}_n(\tilde{B})\,] \;-\; \limsup_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} |X_t|^p\, I\{|X_t| > c/3\} \quad \text{wp1.} \end{aligned} \tag{9}$$

Consider the first term on the right hand side of (9). A simple calculation shows that

$$(X_t - F'_t)^2 \;-\; (X_t - B_t)^2 \;=\; (Y_t - F'_t)^2 \;-\; (Y_t - B_t)^2 \;+\; 2N_t F'_t \;-\; 2N_t B_t.$$

Thus, as $\tilde{B} = B$,

$$\tilde{L}_n(F') - \tilde{L}_n(\tilde{B}) \;=\; L_n(F') - L_n(B) \;+\; \frac{2}{n} \sum_{t=1}^{n} N_t F'_t \;-\; \frac{2}{n} \sum_{t=1}^{n} N_t B_t. \tag{10}$$

Since $\psi(|N_t B_t|) \leq \psi(2|N_t|^2) + \psi(2|B_t|^2)$, assumption (N3) and part (c) of Lemma 2 together ensure that $\sup_{t\geq 1} E\psi(|N_t B_t|)$ is finite. By Lemma A,

$$\frac{1}{n} \sum_{t=1}^{n} N_t B_t \;-\; \frac{1}{n} \sum_{t=1}^{n} E[\,N_t B_t \,|\, X_1^{t-1},\, N_1^{t-1}\,] \;\to\; 0 \quad \text{wp1.}$$

Assumptions (N1)-(N2) imply that $E[\,N_t B_t \,|\, X_1^{t-1},\, N_1^{t-1}\,] = B_t\, E[\,N_t \,|\, X_1^{t-1},\, N_1^{t-1}\,] = 0$, so $n^{-1} \sum_{t=1}^{n} N_t B_t \to 0$ with probability one. A similar argument shows that $n^{-1} \sum_{t=1}^{n} N_t F'_t \to 0$ with probability one. As B is Cesaro optimal for $\mathbf{Y}$, it follows from (10) that

$$\liminf_{n\to\infty} [\,\tilde{L}_n(F') - \tilde{L}_n(\tilde{B})\,] \;=\; \liminf_{n\to\infty} [\,L_n(F') - L_n(B)\,] \;\geq\; 0$$

In conjunction with (9), the last inequality implies that

$$\liminf_{n\to\infty} [\, \tilde{L}_n(F) - \tilde{L}_n(\tilde{B}) \,] \;\geq\; -\limsup_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} |X_t|^p \, I\{|X_t| > c/3\} \quad \text{wp1.}$$

The optimality of $\tilde{B}$ follows by letting $c \to \infty$. The proof of uniqueness is similar to that of Theorem 2 and is omitted.

5 Other Forms of Optimality

To be Cesaro optimal, the average performance of a prediction scheme must dominate or equal that of any competing scheme in the limit of increasing observations. We describe here two other forms of optimality that have also received attention in the literature. Both forms are essentially stronger than Cesaro optimality, in that they apply stronger competitive criteria. In each case, under suitable assumptions, the Bayes scheme is optimal and is unique in an appropriate sense.

5.1 Strong Optimality

Strong optimality, like the notion of calibration discussed in Section 6 below, is defined in terms of place selection schemes.

Definition: A measurable place selection scheme is a binary valued function $S : \mathbb{R}^* \to \{0, 1\}$ such that, for each $j \geq 1$, the restriction of S to j-tuples $x_1, \ldots, x_j$ is a measurable function from $\mathbb{R}^j$ to $\{0, 1\}$.

Let $\mathbf{X} = X_1, X_2, \ldots \in \mathbb{R}$ be a given stochastic process. Each place selection scheme S selects a random subsequence $X_{t_1}, X_{t_2}, \ldots$ of $\mathbf{X}$, where $t_1 = \min\{t \geq 1 : S(X^{t-1}) = 1\}$, and $t_k = \min\{t > t_{k-1} : S(X^{t-1}) = 1\}$ for $k \geq 2$. By definition, the inclusion of X_t in the subsequence depends only on the previous values X^{t-1} of the process. One way of assessing the performance of a prediction scheme F on $\mathbf{X}$ is to evaluate the difference, at each selected time t_k, between the prediction $F(X^{t_k-1})$ and the observed value X_{t_k}.

Definition: A decision scheme G is strongly optimal for a bounded process $\mathbf{X}$ under ℓ_p if for every decision scheme F, and every measurable selection scheme S,

$$\liminf_{n\to\infty} \left[\frac{\sum_{t=1}^{n} S(X^{t-1})\, \ell_p(F(X^{t-1}), X_t)}{\sum_{s=1}^{n} S(X^{s-1})} \;-\; \frac{\sum_{t=1}^{n} S(X^{t-1})\, \ell_p(G(X^{t-1}), X_t)}{\sum_{s=1}^{n} S(X^{s-1})} \right] \;\geq\; 0$$

almost surely on the event $A(\mathbf{X}, S) = \left\{\sum_{t=1}^{\infty} S(X^{t-1}) = \infty\right\}$ that S selects an infinite subsequence of $\mathbf{X}$.

Strong optimality was introduced in an equivalent form by Dawid [13] as a means of assessing the empirical validity of a prediction scheme that is applied to an individual binary sequence. To avoid pathologies arising in the individual sequence setting, he restricts attention to computable selection schemes and computable prediction rules. For a given stochastic process, such pathologies occur with probability zero, and there is no loss in considering measurable selection schemes and prediction rules, provided that one is satisfied with almost sure results. Analysis of strong optimality relies on the following analog of Lemma A. For a proof and discussion, see Dawid [11].

Lemma B *Let $X_1, X_2, \ldots$ be any process taking values in $\mathbb{R}$ and let $Z_1, Z_2, \ldots \in \mathbb{R}$ be random variables such that Z_t is a measurable function of $X_1, \ldots, X_t$. If there is a constant $L < \infty$ such that $|Z_t| \leq L$ with probability one for each $t \geq 1$, then*

$$\frac{\sum_{t=1}^{n} S(X^{t-1})\,(Z_t - E(Z_t|X^{t-1}))}{\sum_{s=1}^{n} S(X^{s-1})} \;\rightarrow\; 0$$

almost surely on $A(\mathbf{X}, S)$.

The next proposition may be established using Lemma B and arguments similar to those for Theorems 1 and 2. It should be noted that its conclusions do not imply those of Dawid [13] in the setting of computable prediction schemes, and conversely. Uniqueness of computable schemes for individual binary sequences is established in Theorem 7.1 of [13].

Proposition 2 *Let $\mathbf{X}$ be a bounded process and $p > 1$. The Bayes scheme B is strongly optimal for $\mathbf{X}$, and if F is any other strongly optimal scheme for $\mathbf{X}$, then $|F(X^{t-1}) - G(X^{t-1})| \rightarrow 0$ with probability one.*

5.2 Efficiency

Another notion of predictive optimality is that of efficiency, considered by Skouras and Dawid [38] (see also [12, 39]).

Definition: A prediction scheme F is efficient for a process $\mathbf{X}$ under ℓ_p if for every measurable decision scheme G

$$\limsup_{n\rightarrow\infty} \left[\sum_{t=1}^{n} |X_t - F(X^{t-1})|^p \;-\; \sum_{t=1}^{n} |X_t - G(X^{t-1})|^p \right] \;<\; \infty \qquad (11)$$

with probability one.

A multivariate version of the following result appears in Theorem 1 of [38] and subsequent remarks. As noted there, the case $p \neq 2$ remains open. Let $\mathrm{Var}(X_t|X^{t-1}) = E[\,(X_t - E(X_t\,|\,X^{t-1}))^2 \mid X^{t-1}]$ be the conditional variance of X_t given X^{t-1}.

Theorem A *Under the squared loss, the Bayes scheme B is efficient for $\mathbf{X}$ if $\sup_{t\geq 1} \mathrm{Var}(X_t|X^{t-1})$ is finite with probability one; in particular, (11) holds almost surely on the event where the supremum is finite. If F is any other efficient prediction scheme for $\mathbf{X}$, then $\sum_{t=1}^{\infty}(B(X^{t-1})-F(X^{t-1}))^2$ is finite with probability one.*

For bounded processes with squared error, the comparative strengths of different forms of optimality follow readily from their relation to the Bayes scheme.

Proposition 3 *If $\mathbf{X}$ is bounded then, under the squared loss, efficiency implies strong optimality, and strong optimality implies Cesaro optimality.*

6 Optimal Prediction and Calibration

6.1 Cesaro Optimality and Weak Calibration

Recall from Section 5 that a place selection scheme S selects a subsequence of a given process in a non-anticipating manner. Motivated by Dawid [13], we make the following definition.

Definition: A prediction scheme F is first order calibrated to $\mathbf{X}$ if for every measurable selection scheme S,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} S(X^{t-1})\,(F(X^{t-1}) - X_t) \;=\; 0 \tag{12}$$

with probability one. A scheme F is second order calibrated to $\mathbf{X}$ if for every measurable selection scheme S,

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} S(X^{t-1})\,(F^2(X^{t-1}) - X_t^2) \;\leq\; 0 \tag{13}$$

with probability one.

The definition above is weak in the sense that the averages in (12) and (13) are taken with respect to the time scale of the original process, rather than that of the subsequence $X_{t_1}, X_{t_2}, \ldots$. Note however that the relation (12) implies that

$$\frac{\sum_{t=1}^{n} S(X^{t-1})\,(F(X^{t-1}) - X_t)}{\sum_{s=1}^{n} S(X^{s-1})} \;\to\; 0$$

almost surely on the event $\{\liminf_n n^{-1}\sum_{t=1}^{n} S(X^{t-1}) > 0\}$, *i.e.* when the selected times $t_1, t_2, \ldots$ occupy a non-negligible fraction of the positive integers. Similar remarks apply to the relation (13). The following proposition is proved in Section 10

Proposition 4 *Under assumptions (A1) and (A2), a prediction scheme F is Cesaro optimal for* $\mathbf{X}$ *under* ℓ_2 *if and only if F is first and second order calibrated to* $\mathbf{X}$.

6.2 Strong Optimality and Strong Calibration

For bounded processes, one may strengthen in a natural way the notion of calibration studied above by requiring that the convergence in (12) hold whenever S selects any infinite subsequence of $\mathbf{X}$.

Definition: A decision scheme F is strongly calibrated to $\mathbf{X}$ if, for every measurable selection scheme S,

$$\frac{\sum_{t=1}^{n} S(X^{t-1})(F(X^{t-1}) - X_t)}{\sum_{s=1}^{n} S(X^{s-1})} \;\to\; 0$$

almost surely on the event $A(\mathbf{X}, S) = \{\sum_{t=1}^{\infty} S(X^{t-1}) = \infty\}$.

Strong calibration was introduced in [13] for individual binary sequences. The next proposition may be established using Lemma B and arguments like those for Proposition 4. An analogous result for individual binary sequences is given in Theorem 8.1 of [13].

Proposition 5 *A prediction scheme F is strongly optimal for a bounded process* $\mathbf{X}$ *under the squared loss if and only if it is strongly calibrated to* $\mathbf{X}$.

7 Threshold Prediction of Binary Processes

Here we establish a connection between the prediction of binary processes under the squared and Hamming loss functions. Let $\mathbf{X} = X_1, X_2, \ldots$ be a process with values $X_i \in \{0, 1\}$. To take a popular example, suppose that $\mathbf{X}$ is a binary record of rainfall at a specific location, with $X_i = 1$ if it rains on the i'th day, and $X_i = 0$ otherwise. Under the square loss ℓ_2, the predictions of the Bayes scheme B are the conditional probabilities

$$B(X^{t-1}) \;=\; E(X_t \,|\, X^{t-1}) \;=\; P(X_t = 1 \,|\, X^{t-1}) \;\in\; [0, 1].$$

A decision scheme $F : \mathbb{R}^* \to \mathbb{R}$ models the predictions of a weather forecaster who, on each day $t-1$, predicts the conditional probability of rain on day t by $F(X^{t-1})$ and incurs loss $(F(X^{t-1}) - X_t)^2$ when the value of X_t is revealed.

Now suppose that a forecaster employing a decision scheme F with values in $\mathbb{R}$ is restricted to make binary predictions of the form "tomorrow it will rain" or "tomorrow it will not rain", and that he incurs loss 0 or 1 depending on whether his prediction is correct or not.

This is a discrete version of the prediction problem with Hamming loss $\ell_H(u,v) = I\{u \neq v\}$. In this case the Bayes decision scheme is given by

$$\check{B}(X^{t-1}) \;=\; \arg\min_{u\in\{0,1\}} P(X_t \neq u \,|\, X^{t-1}) \;=\; I\{B(X^{t-1}) > 1/2\} \tag{14}$$

and is obtained by thresholding the Bayes scheme under ℓ_2 at $1/2$. One may readily show that $\check{B}$ is Cesaro optimal for $\mathbf{X}$. In light of (14), it is natural for the forecaster to employ the threshold scheme

$$\check{F}(X^{t-1}) \;=\; I\{F(X^{t-1}) > 1/2\}$$

in order to predict the next value of $\mathbf{X}$ based on his conditional probability estimates F. In fact, the Cesaro optimality of F implies that of $\check{F}$. A version of the following result for ergodic processes was established independently in [22].

Theorem 4 *Let $\mathbf{X} = X_1, X_2, \ldots$ be any binary process. If F is Cesaro optimal for $\mathbf{X}$ under the squared loss, then the threshold prediction scheme $\check{F}$ is Cesaro optimal for $\mathbf{X}$ under the Hamming loss.*

Proof: Let B and $\check{B}$ be as above. It can be shown (see *e.g.* the proofs of Theorems 2.1 and 2.2 in [15]) that for each $t \geq 1$,

$$P(\check{F}_t \neq X_t \,|\, X^{t-1}) \;\leq\; P(\check{B}_t \neq X_t \,|\, X^{t-1}) \;+\; 2\,|F_t - B_t|. \tag{15}$$

Fix any binary-valued prediction scheme $\check{H} : \{0,1\}^* \to \{0,1\}$ and let $\check{H}_t = \check{H}(X^{t-1})$. We wish to establish that

$$\liminf_{n\to\infty} \big[L_n(\check{H}) - L_n(\check{F})\big] \;=\; \liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} (I\{\check{H}_t \neq X_t\} - I\{\check{F}_t \neq X_t\}) \;\geq\; 0 \;\text{ wp1.}$$

By Lemma A it suffices to show that

$$\liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} [\, P(\check{H}_t \neq X_t \,|\, X^{t-1}) \;-\; P(\check{F}_t \neq X_t \,|\, X^{t-1})\,] \;\geq\; 0 \;\text{ wp1.} \tag{16}$$

Inequality (15) and equation (14) imply that

$$P(\check{H}_t \neq X_t \,|\, X^{t-1}) \;-\; P(\check{F}_t \neq X_t \,|\, X^{t-1}) \;\geq\; -2\,|F_t - B_t|,$$

and (16) follows, since $n^{-1}\sum_{i=1}^{n} |F_t - B_t| \to 0$ by Theorem 2.

A straightforward modification of the preceding proof, substituting Lemma B for Lemma A, yields the following result.

Theorem 5 *Let $\mathbf{X} = X_1, X_2, \ldots$ be any binary process. If F is strongly optimal for $\mathbf{X}$ under the squared loss, then the threshold prediction scheme $\check{F}$ is strongly optimal for $\mathbf{X}$ under the Hamming loss.*

8 Aggregating Decision Schemes

Consider again the general prediction problem described in the introduction, now with the goal of constructing a prediction scheme that is Cesaro optimal, under the p'th power loss, for a given family $\mathcal{X}$ of stochastic processes satisfying (A1) and (A2). (Recall that a prediction scheme F is Cesaro optimal for a family of processes if it is Cesaro optimal for every member of the family; analogous definitions hold for other forms of optimality.) It is clear from the definition that the Bayes scheme B for a given process $\mathbf{X} \in \mathcal{X}$ will generally not be Cesaro optimal for a *different* process $\mathbf{X}' \in \mathcal{X}$. Skouras and Dawid [38] use a Bayesian approach to combine the Bayes predictors for a given parametric family of processes $\mathcal{X} = \{\mathbf{X}_\theta : \theta \in \Theta\}$ with $\Theta \subseteq \mathbb{R}^d$. For suitable families $\mathcal{X}$, they use a positive prior π on Θ to construct prediction schemes efficient for (Lebesgue) almost every member of the family. In particular, their prediction scheme is Cesaro optimal under ℓ_2 for almost every member of a parametric family of bounded processes. Nevertheless, one may readily verify that no decision scheme is Cesaro optimal under ℓ_p for the family of all processes satisfying (A1) and (A2). The same conclusion holds if we restrict attention to bounded, or binary, processes. (Given any prediction scheme F, define recursively a sequence $\mathbf{x} = x_1, x_2, \ldots \in \{0, 1\}$ such that $|F(x^{t-1}) - x_t| \geq 1/2$ for each $t \geq 1$. If $\mathbf{X} = \mathbf{x}$ with probability one, then $L_n(F, \mathbf{X}) \geq 1/2$ for every n, whereas the cumulative loss of the Bayes scheme for $\mathbf{X}$ is equal to zero.)

In the absence of "universal" schemes, one useful way to assess the quality of a given prediction scheme is to compare, for each process $\mathbf{X} \in \mathcal{X}$, the asymptotic performance of that scheme with the best asymptotic performance among a finite or countable family $\mathcal{F}$ of competing schemes. In this way attention shifts from absolute to comparative measures of performance. A central problem in the comparative framework is how to construct a single scheme that competes favorably with every member of a given family $\mathcal{F}$ on a wide variety of processes. In many cases, this may be accomplished by suitably combining, or aggregating, the decisions of the individual schemes in $\mathcal{F}$. Aggregating methods, and corresponding bounds on the difference between the loss of an aggregate scheme and that of the best scheme in the family, have been established in a variety of settings. Representative work and further references can be found in [41, 17, 27, 10, 9, 25]. Foster and Vohra [19] give an account of the aggregating problem and its history, and Merhav and Feder [28] give an overview of prediction from individual sequences. In recent work, Weissman and Merhav [43] established finite sample aggregation bounds for the prediction of individual binary

sequences observed in additive, independent noise, under p'th power loss, with $1 \le p \le 2$.

Here we describe an aggregating method for prediction schemes that is based on weighted majority techniques [41, 27, 8] for predicting individual binary sequences. Fix a countable family $\mathcal{F} = \{F^{(1)}, F^{(2)}, \ldots\}$ of prediction schemes and let $p > 1$. Assume that each scheme $F^{(r)}$ in $\mathcal{F}$ is bounded, in the sense that

$$|F^{(r)}| \;:=\; \sup_{t \ge 1} \sup_{x^{t-1}} |F^{(r)}(x^{t-1})| \;<\; \infty.$$

Let $\mathcal{F}_j = \{F^{(r)} : 1 \le r \le 2^j\}$ contain the first 2^j prediction schemes in $\mathcal{F}$, and let $x_1, x_2, \ldots \in \mathbb{R}$. For $j \ge 0$ and $2^j \le t < 2^{j+1}$ define

$$\tilde{F}(x^{t-1}) \;=\; \sum_{F \in \mathcal{F}_j} w_t(F)\, F(x^{t-1}) \tag{17}$$

to be a weighted sum of the predictions made by schemes $F \in \mathcal{F}_j$ at time t, with weights given by

$$w_t(F) \;=\; \frac{\exp\{-c_j \sum_{s=2^j}^{t-1} \ell_p(F(x^{s-1}), x_s)\}}{\sum_{F' \in \mathcal{F}_j} \exp\{-c_j \sum_{s=2^j}^{t-1} \ell_p(F'(x^{s-1}), x_s)\}} \qquad c_j = 2^{-j+j^{1/2}}. \tag{18}$$

Note that for $2^j \le t < 2^{j+1}$ the weight assigned to a scheme $F \in \mathcal{F}_j$ at time t depends on the cumulative loss of its predictions from time 2^j to time $t-1$. When $t = 2^j$, each $F \in \mathcal{F}_j$ has equal weight $w_t(F) = |\mathcal{F}_j|^{-1} = 2^{-j}$. A related weight assignment was used in [24] and more recently in [43] to combine binary predictors.

The next proposition shows that, for suitable processes $\mathbf{X}$, the long run average average loss of $\tilde{F}$ is less than or equal to the long run average loss of every scheme in $\mathcal{F}$. The proof is given in Section 10.3. Foster [18] established an analogous result for bounded processes under the squared loss using a recursive construction.

Proposition 6 *Suppose that $|F^{(r)}| = O(r^{(1-\delta)/2p})$ for some $\delta > 0$. Let $\mathbf{X} = X_1, X_2, \ldots$ be any stochastic process such that (i) $\sup_{t \ge 1} E|X_t|^q$ is finite for some $q > p$, and (ii) $\limsup_n n^{-1} \sum_{t=1}^n |X_t|^p$ is finite with probability one. Then with probability one,*

$$\limsup_{n \to \infty} L_n(\tilde{F}, \mathbf{X}) \;\le\; \limsup_{n \to \infty} L_n(F, \mathbf{X}) \quad \forall F \in \mathcal{F} \tag{19}$$

and

$$\liminf_{n \to \infty} \Big[L_n(F, \mathbf{X}) - L_n(\tilde{F}, \mathbf{X}) \Big] \ge 0 \quad \forall F \in \mathcal{F} \tag{20}$$

Remark: Given any countable family $\mathcal{F} = \{F^{(r)}\}$ of bounded prediction schemes, one may ensure, by replicating schemes $F^{(r)}$ as necessary, that the growth condition $|F^{(r)}| =$

$O(r^{(1-\delta)/2p})$ is satisfied. Thus the conclusions of Proposition 6 will hold for any such family. For bounded processes under the squared loss one can exhibit aggregate prediction schemes $\tilde{H}$ such that

$$L_n(\tilde{H}, \mathbf{X}) \;\leq\; \inf_F \left[L_n(F, \mathbf{X}) \;+\; \frac{c \ln \pi(F)}{n} \right] \quad n = 1, 2, \ldots \quad \text{wp1} \tag{21}$$

where c is a universal constant and $\pi(\cdot)$ is a prior distribution on the elements of $\mathcal{F}$ (*c.f.* [22]). The proof of Proposition 6 relies on a weaker, but more general, inequality of this sort, that is based on arguments of Cesa-Bianchi [8] (see Lemma 3 below).

9 Sequential Prediction of Ergodic Processes

Let X be a random variable, defined on a probability space $(\Omega, \mathcal{S}, P)$, with $E|X|^p < \infty$. For each sub-sigma field $\mathcal{S}' \subseteq \mathcal{S}$ define

$$\pi_p(X|\mathcal{S}') \;=\; \arg\min_{a \in \mathbb{R}} E[\,|X - a\,|^p \;|\, \mathcal{S}'\,]. \tag{22}$$

The definition ensures that $\pi_p(X|\mathcal{S}')$ is an $\mathcal{S}'$-measurable random variable, and that $E|X - \pi_p(X|\mathcal{S}')|^p \;\leq\; E|X - Y|^p$ for any $\mathcal{S}'$-measurable random variable Y. Thus $\pi_p(X|\mathcal{S}')$ is the natural L_p-projection of X onto the family of $\mathcal{S}'$-measurable random variables. The properties of such projections were studied by Ando and Amemiya [1], who established the following result.

Theorem B *If $\mathcal{S}_1 \subseteq \mathcal{S}_2 \subseteq \cdots$ are increasing sub-sigma fields of $\mathcal{S}$ with limit $\mathcal{S}_\infty = \sigma(\cup_{k \geq 1} \mathcal{S}_k)$, then the following relations hold:*

(a) $\pi_p(X|\mathcal{S}_k) \;\to\; \pi_p(X|\mathcal{S}_\infty)$ with probability one;

(b) $E[\,\sup_{k \geq 1} |\pi_p(X|\mathcal{S}_k)|^p\,] \;\leq\; \gamma\, E|X|^p$ for some constant $\gamma = \gamma(p) < \infty$.

Remark: As an alternative to the approach in [1], one may establish (a) using the definition (22) and the fact that, with probability one, $f_k(a) = E[\,|X - a|^p \;|\, \mathcal{S}_k\,]$, $k \geq 1$, are convex functions converging pointwise in a (and hence uniformly on bounded intervals) to the convex function $f_\infty(a) = E[\,|X - a|^p \;|\, \mathcal{S}_\infty\,]$. Part (b) of the theorem shows that the expected value of the supremum is finite. Under the stronger moment assumptions made here, this may be established by more direct arguments. Indeed, by an obvious extension of Lemma 2,

$$\sup_{k \geq 1} |\pi(X|\mathcal{S}_k)|^p \;\leq\; 2^p \sup_{k \geq 1} E[\,|X|^p \,|\, \mathcal{S}_k\,].$$

Moreover $\{E[\,|X|^p\,|\,\mathcal{S}_k\,]:k\geq 1\}$ is a uniformly integrable martingale that converges with probability one, and in expectation, to the integrable random variable $E[\,|X|^p\,|\,\mathcal{S}_\infty\,]$. It follows from the maximal inequality for submartingales and standard bounds (see Theorems 3.2 and 3.4' of Doob [16]) that $E(\sup_{k\geq 1}E[\,|X|^p\,|\,\mathcal{S}_k\,])$ is finite if $E\,|X|^p\log(1+|X|^p)$ is finite.

Using Theorem B and Breiman's ergodic theorem, one may give a simple characterization of Cesaro optimal prediction schemes for ergodic processes in terms of their limiting average loss, without reference to competing prediction schemes. This characterization is given in the next proposition, which can be deduced from the results of Algoet [2]. We sketch a more direct, simpler proof for completeness. By standard arguments (c.f. Breiman [7], Chapter 6) we may assume without loss of generality that any ergodic process $\mathbf{X}$ under consideration has a doubly infinite time index, and is defined on a probability space $(\Omega,\mathcal{S},P)$, where Ω consists of all doubly infinite sequences of real numbers, $\mathcal{S}$ is generated by finite dimensional cylinder sets, and $X_i(\omega)=w_i$ for each $-\infty<i<\infty$ and each $\omega\in\Omega$.

Proposition 7 *Let $\mathbf{X}=\{X_i:-\infty<i<\infty\}$ be a stationary ergodic process such that $E\,|X_0|^p\log^2(1+|X_0|^p)<\infty$. A prediction scheme F is Cesaro optimal for $\mathbf{X}$ under ℓ_p if and only if*

$$L_n(F,\mathbf{X})\;\rightarrow\;L^*(\mathbf{X})\;=\;E|X_0-\pi_p(X_0|X_{-\infty}^{-1})\,|^p\quad wp1, \tag{23}$$

where $\pi_p(X_0|X_{-\infty}^{-1})=\pi_p(X_0|\sigma(X_{-1},X_{-2},\cdots))$. The optimal limiting average loss can also be written as

$$L^*(\mathbf{X})\;=\;\inf_{k\geq 1}\inf_{f_k}E|X_{k+1}-f_k(X_1^k)|^p, \tag{24}$$

where the second infimum is over all bounded, uniformly continuous functions $f_k:\mathbb{R}^k\rightarrow\mathbb{R}$.

Proof: Theorem 1 and relation (4) imply that F is Cesaro optimal for $\mathbf{X}$ if and only $L_n(F)-L_n(B)\rightarrow 0$ with probability one, where B is the Bayes scheme for $\mathbf{X}$ under ℓ_p. Thus to establish (23) it suffices to consider the case $F=B$. Note that

$$L_n(B,\mathbf{X})\;=\;\frac{1}{n}\sum_{t=1}^{n}|\pi_p(X_t|X_1^{t-1})-X_t|^p\;=\;\frac{1}{n}\sum_{t=1}^{n}|\pi_p(X_0|X_{-t+1}^{-1})-X_0|^p\circ T^t$$

where $T:\Omega\rightarrow\Omega$ is the left shift operator. By assumption, T is P-preserving and ergodic. Thus (23) will follow from the last expression and Breiman's generalized ergodic theorem (see [2] for a proof) if (a) $|X_0-\pi_p(X_0|X_{-t}^{-1})|^p\rightarrow|X_0-\pi_p(X_0|X_{-\infty}^{-1})|^p$ with probability one as $t\rightarrow\infty$, and (b) $E(\sup_{t\geq 1}|X_0-\pi_p(X_0|X_{-t}^{-1})|^p)<\infty$. Both these relations follow

immediately from Theorem B. To establish (24), note that (a)-(b) and the dominated convergence theorem imply that

$$L^*(\mathbf{X}) \;=\; E|\, X_0 \,-\, \pi_p(X_0|X_{-\infty}^{-1})\,|^p \;=\; \lim_{k\to\infty} E|\, X_0 \,-\, \pi_p(X_0|X_{-k}^{-1})\,|^p.$$

By definition of $\pi_p(\cdot|\cdot)$, the k'th term in the limit is equal to

$$E|\, X_0 \,-\, \pi_p(X_0|X_{-k}^{-1})\,|^p \;=\; \inf_{f_k} E|X_0 - f_k(X_{-1},\ldots,X_{-k})|^p,$$

where the infimum is over all measurable functions $f_k : \mathbb{R}^k \to \mathbb{R}$. These expectations are decreasing in k, and equation (24) follows as bounded, uniformly continuous functions are dense in $L_p(X_{-1},\ldots,X_{-k})$.

9.1 Universal Prediction Schemes for Ergodic Processes

Recall that a decision scheme F is Cesaro optimal for a family $\mathcal{X}$ of processes if it is Cesaro optimal for every process $\mathbf{X} \in \mathcal{X}$. Algoet [2] established the existence of Cesaro optimal schemes for families of ergodic processes in the general setting of sequential decision problems. His schemes are derived from estimates $\hat{P}(X^{t-1})$ of the conditional probabilities $P(X_t|X^{t-1})$ with the property that $\hat{P}(X_{-1},\ldots,X_{-t})$ converges weakly to $P(X_0|X_{-1},X_{-2},\ldots)$ with probability one for every ergodic process. (For more on such estimates, see [3, 30].) Specialized to the setting of this paper, the results of [2] establish that, for every $p > 1$ and every $M < \infty$, there exist Cesaro optimal schemes under ℓ_p for the family of all ergodic processes with values in $[-M, M]$. Below we describe prediction schemes $\tilde{H}$ that are Cesaro consistent for unbounded ergodic processes, under relatively weak moment conditions. The schemes here are based on the elementary aggregating method described in the previous section, and avoid the use of conditional probability estimates. Modha and Masry [29] exhibited in-probability consistent estimates of $E(X_0|X_{-\infty}^{-1})$ for bounded, alpha-mixing processes for which the mixing coefficients decay at a known exponential rate. Under additional conditions, they established rates of convergence for estimates of $E(X_0|X_{-k}^{-1})$ when $\mathbf{X}$ has finite memory k.

Let $\pi_1 \geq \pi_2 \geq \cdots$ be a nested sequence of finite partitions of $\mathbb{R}$ whose constituent cells shrink, in the sense that for each $x \in \mathbb{R}$,

$$\lim_{r\to\infty} \mathrm{diam}(\pi_r[x]) \;=\; 0. \tag{25}$$

Here $\pi_r[x]$ is the unique cell of π_r containing x, and $\mathrm{diam}(A) = \sup_{u,v\in A} |u-v|$ denotes the maximum distance between any two points in A. As π_r is finite, it must necessarily have

unbounded cells. However, the condition (25) ensures that the sequence of cells containing a fixed point $x \in \mathbb{R}$ will eventually shrink down to x. The partition π_r may be obtained, for example, by dividing $[-r, r)$ into intervals of length 2^{-r}, and letting the complement of $[-r, r)$ comprise a single cell.

Fix $p > 1$ and $0 < \delta < 1$. For each $k \geq 1$ define a k'th order Markov prediction scheme $H^{(k)}$ as follows. For $t \leq k+1$, set $H^{(k)}(x^{t-1}) = 0$; for $t \geq k+2$ and each $x_1, \ldots, x_{t-1} \in \mathbb{R}$ let

$$H^{(k)}(x^{t-1}) \;=\; \underset{-a_k \leq u \leq a_k}{\arg\min} \sum_{s=k+1}^{t-1} \ell(u, x_s)\, I\{x_{s-1} \in \pi_k[x_{t-1}], \ldots, x_{s-k} \in \pi_k[x_{t-k}]\},$$

where $a_k = k^{(1-\delta)/2p}$. To understand the definition, let us say that a k-match occurs at position s if the k vectors preceding x_s lie in the same cells of π_k as the k vectors preceding x_t. Then $H^{(k)}(x^{t-1})$ is the element $u \in \mathbb{R}$ that minimizes the sum of the losses $\ell(u, x_s)$ occurring at the k-match positions $s \leq t-1$. Note that as k increases the predictions of $H^{(k)}$ are based on longer and more precise matches. Prediction schemes analogous to $H^{(k)}$ are briefly discussed by Algoet [2]; similar, randomized, schemes were proposed in [24] for the prediction of binary processes. Note that no randomization is required in the present setting. The proof of the following theorem is given in Section 10.4.

Theorem 6 *Let $\tilde{H}$ be the aggregate prediction scheme derived from $\mathcal{H} = \{H^{(k)} : k \geq 1\}$ via (17)-(18). Then $\tilde{H}$ is Cesaro optimal under the p'th power loss for any ergodic process $\mathbf{X}$ such that $E|X_1|^q < \infty$ for some $q > p$.*

The existence of Cesaro optimal schemes for general, bounded loss functions was established by Algoet [2]. In Theorem 8 of [4], a Cesaro optimal schemes for bounded processes under the squared error is described, and the possible consistency of the scheme for unbounded processes is briefly discussed. Using aggregation bounds of the form (21), Györfi and Lugosi [22] have independently established the Cesaro optimality of a prediction scheme similar to $\tilde{H}$ for bounded processes under the squared loss. They also consider the Cesaro optimality of prediction schemes based on generalized linear estimates, and obtain rates of convergence for predicting Gaussian processes. The results of [2, 4, 22] on Cesaro optimal prediction assume boundedness of the loss function, or that the X_i take values in a prespecified bounded interval. No such assumptions are made in Theorem 6.

9.2 Strongly Optimal Schemes for Ergodic Processes

For $A \subseteq \mathbb{R}$ let $\mathcal{E}(A)$ be the family of ergodic processes $\mathbf{X}$ taking values in A. If A is bounded then the aggregate scheme $\tilde{H}$ defined in the previous section is Cesaro optimal for $\mathcal{E}(A)$ under ℓ_p. By contrast, no prediction scheme is *strongly* optimal for $\mathcal{E}(A)$ under the squared loss if A has more than one elements. To illustrate this, let $\mathcal{E}(\{0,1\})$ be the family of all binary ergodic processes, and let $\ell_2(u,v) = (u-v)^2$ be the squared loss. For each $\mathbf{X} \in \mathcal{E}(\{0,1\})$ the corresponding Bayes scheme is of the form $B(x^{t-1}) = P(X_t = 1|X^{t-1} = x^{t-1})$. If F is strongly optimal for $\mathcal{E}(\{0,1\})$, then it follows from Proposition 2 that for every binary ergodic process $\mathbf{X}$,

$$|F(X^{t-1}) - P(X_t = 1|X^{t-1})| \to 0 \quad \text{wp1.}$$

However, it is known [6, 37, 23] that no such "on-line" estimation scheme exists. Therefore no prediction scheme is strongly optimal for $\mathcal{E}(\{0,1\})$, and by the same reasoning, no prediction scheme is strongly optimal under the squared loss for $\mathcal{E}(A)$ if A has cardinality greater than one. A similar negative conclusion holds for efficient prediction schemes.

9.3 Properties of Universal Schemes under Squared Loss

Throughout this section let H^* be the aggregate prediction scheme derived via (17)-(18) from the Markov schemes $\mathcal{H} = \{H^{(k)} : k \geq 1\}$ under the squared loss. Theorem 6 ensures that H^* is Cesaro optimal for every ergodic process $\mathbf{X}$ such that $E|X_0|^q < \infty$ for some $q > 2$. Let $\mathbf{X}$ be any such process. By Theorem 2 one has

$$\frac{1}{n}\sum_{t=1}^{n}(H^*(X^{t-1}) - E(X_t|X^{t-1}))^2 \ \to \ 0 \ \ \text{wp1.} \tag{26}$$

(This same property is derived for bounded processes in [22] by different arguments.) Moreover, Proposition 7 ensures that

$$\frac{1}{n}\sum_{t=1}^{n}(H^*(X^{t-1}) - X_t)^2 \ \to \ E(X_0 - E(X_0|X_{-\infty}^{-1}))^2 \ \ \text{wp1.}$$

Suppose now that $\mathbf{N} = \{N_i : -\infty < i < \infty\}$ are i.i.d., zero-mean random variables that are independent of $\mathbf{X}$, and satisfy $E|N_0|^q < \infty$. Let $Y_i = X_i + N_i$, $-\infty < i < \infty$, be observations of $\mathbf{X}$ corrupted by additive noise. If $H^*(Y^{t-1})$ is used to predict the "clean" value X_t, then the limiting average cumulative loss of H^* has a natural form. Related results for binary processes $\mathbf{X}$ under more general loss functions can be found in [42].

Proposition 8 *If* $\mathbf{X}$ *and* $\mathbf{N}$ *are as above then*

$$\frac{1}{n}\sum_{t=1}^{n}(H^*(Y^{t-1}) - X_t)^2 \;\to\; E(X_0 - E(X_0 \,|\, Y_{-\infty}^{-1}))^2 \quad \textit{wp1.}$$

Proof: Let $H_t^* = H^*(Y^{t-1})$ and let $B_t = E(Y_t \,|\, Y^{t-1}) = E(X_t \,|\, Y^{t-1})$. By Theorem 2, $n^{-1}\sum_{t=1}^{n}(H_t^* - B_t)^2 \to 0$. Using this fact and the Cauchy-Schwartz inequality, one may readily show that

$$\left|\frac{1}{n}\sum_{t=1}^{n}(H_t^* - X_t)^2 \;-\; \frac{1}{n}\sum_{t=1}^{n}(B_t - X_t)^2\right| \;\to\; 0 \quad \text{wp1.}$$

It follows from Breiman's ergodic theorem, or alternatively by arguments similar to those in the proof of Proposition 1, that $n^{-1}\sum_{t=1}^{n}(B_t - X_t)^2 \to E(X_0 - E(X_0 \,|\, Y_{-\infty}^{-1}))^2$ with probability one.

An interesting question, which has received some attention in the literature, is how to estimate the conditional expectation $E(X_0 \,|\, X_{-\infty}^{-1})$ from observations $X_{-1}, X_{-2}, \ldots$ of an ergodic process $\mathbf{X} = \{X_i : -\infty < i < \infty\}$. Ornstein [33] described almost surely consistent estimates for binary processes; the case of bounded, real valued processes was studied by Algoet [3] (see also [30, 31]), the final word being [4]. The prediction scheme H^* yields estimates of $E(X_0 \,|\, X_{-\infty}^{-1})$ that are consistent in the weaker, expectation sense.

Proposition 9 *If* $\mathbf{X} = \{X_i : -\infty < i < \infty\}$ *is bounded and ergodic, then the estimate*

$$\phi(X_{-n}^{-1}) = \frac{1}{n}\sum_{t=0}^{n-1} H^*(X_{-t}^{-1})$$

converges in probability to $E(X_0 \,|\, X_{-\infty}^{-1})$ *as* $n \to \infty$.

Proof: Let $B(X^{t-1}) = E(X_t|X^{t-1})$ be the Bayes prediction scheme for $\mathbf{X}$. The identity $B(X_{-t}^{-1}) = E(X_0|X_{-t}^{-1})$ and the martingale convergence theorem imply that $\tilde{\phi}(X_{-n}^{-1}) = n^{-1}\sum_{t=0}^{n-1} B(X_{-t}^{-1})$ converges in expectation to $E(X_0 \,|\, X_{-\infty}^{-1})$. Thus it suffices to show that $E|\phi(X_{-n}^{-1}) - \tilde{\phi}(X_{-n}^{-1})| \to 0$. However, this expectation is at most

$$\begin{aligned}\frac{1}{n}\sum_{t=1}^{n} E|H^*(X_t^{-1}) - B(X_{-t}^{-1})| &= \frac{1}{n}\sum_{t=1}^{n} E|H^*(X_1^{t-1}) - B(X_1^{t-1})| \\ &= E\left[\frac{1}{n}\sum_{t=1}^{n}|H^*(X_1^{t-1}) - B(X_1^{t-1})|\right]\end{aligned}$$

where the first equality follows from the stationarity of $\mathbf{X}$. The final expectation above tends to zero by Theorem 2 and the bounded convergence theorem.

Now let $\mathcal{E}(\{0,1\})$ be the family of binary ergodic processes. The calibration of H^* and Breiman's ergodic theorem have the following elementary corollary. Let $\mathbf{X} \in \mathcal{E}(\{0,1\})$, and suppose that $p \in (0,1)$ and $\epsilon > 0$ are such that $P\{|E(X_0|X_{-\infty}^{-1}) - p| \leq \epsilon\} > 0$. Then the ratio

$$\gamma_n(p,\epsilon) \;=\; \frac{\sum_{t=1}^n I\{|H^*(X^{t-1}) - p\,| \leq \epsilon\}\, X_t}{\sum_{s=1}^n I\{|H^*(X^{s-1}) - p\,| \leq \epsilon\}}.$$

is such that

$$p - \epsilon \;\leq\; \liminf_{n\to\infty} \gamma_n(p,\epsilon) \;\leq\; \limsup_{n\to\infty} \gamma_n(p,\epsilon) \;\leq\; p + \epsilon.$$

Suppose again that $\mathbf{X}$ is the binary record of rainfall at some location. The inequalities above show that if H^* is used to predict the probability of rain on the next day then, among those days for which H^*'s predicted probability of rain is near p, the fraction of days on which it actually rained is also near p. In other words, H^* is calibrated in the classical sense (*c.f.* [32]). Note also that if $\check{H}^*(X^{t-1}) = I\{H^*(X^{t-1}) > 1/2\}$ is the threshold prediction scheme associated with H^*, then for each $X \in \mathcal{E}(\{0,1\})$,

$$\frac{1}{n}\sum_{t=1}^n I\{\check{H}^*(X^{t-1}) \neq X_t\} \;\to\; E\min\{P(X_0 = 0|X_{-\infty}^{-1}), P(X_0 = 1|X_{-\infty}^{-1})\} \quad \text{wp1}$$

as $\check{H}^*$ is Cesaro optimal for $\mathbf{X}$.

10 Additional Derivations

10.1 Proof of Theorem 2, Case $1 < p < 2$.

Suppose now that $1 \leq q < p < 2$, and that F is a Cesaro optimal decision scheme for $\mathbf{X}$ under ℓ_p. Let $B_t = B(X^{t-1})$. We claim that

$$(i)\ \limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^n |B_t|^p \;<\; \infty \quad (ii)\ \limsup_{n\to\infty} L_n(B,\mathbf{X}) \;<\; \infty \quad (iii)\ \limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^n |F_t|^p \;<\; \infty$$

Indeed, (i) follows from Lemma 2 and assumption (A2). Relation (ii) is a consequence of (i) and (A2), and relation (iii) follows from (ii), (4), and the elementary inequality $|F_t|^p \leq 2^{p-1}(|F_t - X_t|^p + |X_t|^p)$. Suppose now that for some $\alpha > 0$,

$$P\left\{\limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^n |F_t - B_t|^q \;\geq\; \alpha.\right\} \;>\; \alpha. \tag{27}$$

Let $\alpha_1 = 3^{-(q+1)}\alpha$. As $q < p$, relations (i) and (iii) imply that there exists $c < \infty$ such that

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^n |F_t|^q I\{|F_t| > c\} \;<\; \alpha_1 \;\text{ and }\; \limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^n |B_t|^q I\{|B_t| > c\} \;<\; \alpha_1. \tag{28}$$

Let $F'_t = |F_t| I\{|F_t| \leq c\}$ and $B'_t = |B_t| I\{|B_t| \leq c\}$ be truncated versions of B_t and F_t, respectively. Then (27) and (28) imply that

$$P\left\{\limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} |F'_t - B'_t|^q \;\geq\; \alpha_1\right\} \;>\; \alpha. \tag{29}$$

As F'_t and B'_t are bounded, it follows from (29) that for some $\alpha_2 > 0$,

$$P\left\{\limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} I\{|F'_t - B'_t| \geq \alpha_2\} \;>\; \alpha_2\right\} \;>\; \alpha.$$

Finally, (A2) and the last expression imply that for some $c' < \infty$ and $\alpha_3 > 0$,

$$P\left\{\limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} I\{|F'_t - B'_t| \geq \alpha_3\} I\{|X_t| \leq c'\} \;>\; \alpha_3\right\} \;>\; \frac{\alpha}{2}. \tag{30}$$

The continuity and strict convexity of $f(u) = |u|^p$ ensure that $\Gamma(u-x, v-x)$ (see equation (6) above) is positive for all values of u, v, x, and that $\Gamma(u-x, v-x) \geq \beta$ for some $\beta > 0$ on the compact set $\{(u,v,x) : |u-v| \geq \alpha_3,\ |u|, |v| \leq c,\ |x| \leq c'\}$. It then follows from (30) and inequality (5) that

$$\begin{aligned}
\liminf_{n\to\infty} [L_n(H) - L_n(B)] \;&\leq\; \liminf_{n\to\infty} \frac{-\beta}{n}\sum_{t=1}^{n} I\{|F'_t - B'_t| \geq \alpha_3\} I\{|X_t| \leq c'\} \\
&=\; -\beta \limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} I\{|F'_t - B'_t| \geq \alpha_3\} I\{|X_t| \leq c'\} \\
&\leq\; -\beta\alpha_3
\end{aligned}$$

with positive probability. This contradicts the Cesaro optimality of B. Thus (27) fails to hold for any $\alpha > 0$, and the result follows.

10.2 Proof of Proposition 4

Let $\mathbf{X}$ satisfy (A1) and (A2). Under ℓ_2 the Bayes prediction B_t is equal to $E[\, X_t \,|\, X^{t-1}\,]$. Two applications of Lemma A, with $Z_t = X_t$ and $Z_t = X_t^2$, show that B is first and second order weakly calibrated to $\mathbf{X}$. Suppose now that F is Cesaro optimal for $\mathbf{X}$. Then

$U_n = n^{-1}\sum_{t=1}^n (F_t - B_t)^2 \to 0$ by Theorem 2. Note that for each selection scheme S,

$$\begin{aligned}
&\left| \frac{1}{n}\sum_{t=1}^n S_t(F_t^2 - X_t^2) \;-\; \frac{1}{n}\sum_{t=1}^n S_t(B_t^2 - X_t^2) \right| \\
&\quad\leq \quad \frac{1}{n}\sum_{t=1}^n |F_t^2 - B_t^2| \\
&\quad\leq \quad \sqrt{\frac{1}{n}\sum_{t=1}^n (F_t - B_t)^2} \cdot \sqrt{\frac{1}{n}\sum_{t=1}^n (F_t + B_t)^2} \\
&\quad\leq \quad U_n^{1/2} \sqrt{\frac{2}{n}\sum_{t=1}^n F_t^2 \;+\; \frac{2}{n}\sum_{t=1}^n B_t^2}.
\end{aligned}$$

By arguments like those in the proof of Theorem 2 for the case $1 < p < 2$, the time averages of B_t^2 and F_t^2 are bounded. Thus the final term above tends to zero with increasing n, and the second order calibration of F follows from that of B. A similar argument shows that F is first order calibrated to $\mathbf{X}$.

Suppose now that F is first and second order calibrated to $\mathbf{X}$. Note that B is Cesaro optimal for $\mathbf{X}$, and that

$$\left| \sqrt{L_n(F)} - \sqrt{L_n(B)} \right| \;\leq\; \sqrt{\frac{1}{n}\sum_{t=1}^n (F_t - B_t)^2}\,,$$

As F, B are second order calibrated to $\mathbf{X}$, the sequences $L_n(F)$ and $L_n(B)$ are bounded; thus to establish the optimality of F it suffices to show that $n^{-1}\sum_{t=1}^n (F_t - B_t)^2 \to 0$ with probability one. Application of inequality (13) to the selection schemes $S_t' \equiv 1$ and $S_t'' = I\{|F_t| > c\}$ shows that

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^n F_t^2 < \infty\,, \quad \limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^n F_t^2 I\{|F_t| \geq c\} \;\leq\; \limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^n X_t^2 I\{|F_t| \geq c\}.$$

Since $X_t^2\, I\{|F_t| \geq c\} \leq X_t^2\, I\{|X_t| \geq c\} + c^2 I\{|F_t| \geq c\}$, it follows from the last display, assumption (A2) and Lemma 1 that

$$\lim_{c\to\infty} \left[\limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^n F_t^2 I\{|F_t| \geq c\} \right] \;=\; 0 \quad \text{wp1.}$$

This same relation holds for the Bayes scheme B, as B is second order calibrated to $\mathbf{X}$. Thus, for any given $\delta > 0$, there exists $c = c(\delta) < \infty$ such that

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^n (F_t - B_t)^2 \;\leq\; \limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^n (F_t - B_t)^2 S_t \;+\; \delta,$$

where $S_t \stackrel{\triangle}{=} I\{\max|F_t|, |B_t| \le c\}$. To establish the proposition, it is therefore enough to show that for fixed $c < \infty$, and S_t defined as above,

$$\frac{1}{n}\sum_{t=1}^{n}(F_t - B_t)^2 S_t \;\to\; 0 \quad \text{wp1.} \tag{31}$$

As each term in the sum is uniformly bounded by $4c^2$, the relation (31) holds if only if for every $\epsilon > 0$ each of the events

$$A \;=\; \left\{\limsup_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n} S_t\, I\{F_t - B_t \ge \epsilon\} > 0\right\} \quad B \;=\; \left\{\limsup_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n} S_t\, I\{B_t - F_t \ge \epsilon\} > 0\right\}$$

has probability zero. Fix $\epsilon > 0$, and define a new selection scheme $S'_t = S_t \cdot I\{F_t - B_t \ge \epsilon\}$. Then clearly

$$\frac{\epsilon}{n}\sum_{t=1}^{n} S_t\, I\{F_t - B_t \ge \epsilon\} \;=\; \frac{\epsilon}{n}\sum_{t=1}^{n} S'_t \;\le\; \frac{1}{n}\sum_{t=1}^{n} S_t(F_t - B_t).$$

Moreover, Lemma A and the first order calibration of F together imply that

$$\limsup_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n} S_t(F_t - B_t) \;=\; \limsup_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n} S_t(F_t - X_t) \;=\; 0 \quad \text{wp1.}$$

Therefore $P(A) = 0$. A similar argument shows that $P(B) = 0$, and (31) holds as desired.

10.3 Proof of Proposition 6

Let $\mathcal{F}$ be a finite family of prediction schemes and let $p > 1$ be fixed. Given numbers $x_1, x_2, \ldots$ and integers $1 \le u < v$, define for $t = u, \ldots, v-1$ the composite prediction scheme

$$C(x^{t-1}) \;=\; \sum_{F\in\mathcal{F}} w_t(F)\cdot F(x^{t-1}),$$

where $w_u(F) = 1/|\mathcal{F}|$, and for $u < t < v$,

$$w_t(F) \;=\; \frac{\exp\{-c\sum_{s=u}^{t-1}\ell_p(F(x^{s-1}), x_s)\}}{\sum_{F'\in\mathcal{F}}\exp\{-c\sum_{s=u}^{t-1}\ell_P(F'(x^{s-1}), x_s)\}}$$

with c a fixed positive constant. Thus $C(x^{t-1})$ is, for each $t = u, \ldots, v-1$, a convex combination of the predictions $F(x^{t-1})$ made by the individual schemes in $\mathcal{F}$, where the weight assigned to F at time t depends on its success in predicting the values of $x_u, \ldots, x^{t-1}$. The proof of the following lemma follows closely an argument of Cesa-Bianchi [8] for individual binary sequences.

Lemma 3 *Fix $p > 1$. Over the time interval $u \le t < v$, the cumulative loss of the composite decision scheme C satisfies the following inequality:*

$$\sum_{t=u}^{v-1} \ell_p(C(x^{t-1}), x_t) \;\le\; \min_{F\in\mathcal{F}} \sum_{t=u}^{v-1} \ell_p(F(x^{t-1}), x_t) \;+\; \frac{1}{c}\ln|\mathcal{F}| \;+\; \frac{c}{2}\sum_{t=u}^{v-1}\Lambda^2(x^t),$$

where $\Lambda(x^t) = \max_{F\in\mathcal{F}} |F(x^{t-1}) - x_t|^p$.

Proof: The proof is based on a telescoping argument. To this end, set $W_u = |\mathcal{F}|$, and for $j = u+1, \ldots, v$ define

$$W_j \;=\; \sum_{F\in\mathcal{F}} \exp\{-c\sum_{s=u}^{j-1} \ell_p(F(x^{s-1}), x_s)\}.$$

Then it is clear that for $t = u, \ldots, v-1$,

$$\frac{W_{t+1}}{W_t} \;=\; \sum_{F\in\mathcal{F}} w_t(F)\exp\{-c\,\ell_p(F(x^{t-1}), x_t))\}.$$

The right hand side of the last equation is the moment generating function of the random variable $Y_t = \ell_p(F(x^{t-1}), x_t))$ when F is chosen according to the distribution $w_t(\cdot)$. Clearly, Y_t takes values in the interval $[0, \Lambda(x^t)]$. Thus, centering Y_t at its expectation, and applying Hoeffding's inequality [26] for the moment generating function of a bounded random variable, we find that

$$\begin{aligned}
\ln\frac{W_{t+1}}{W_t} &\le -c\sum_{F\in\mathcal{F}} w_t(F)\cdot\ell_p(F(x^{t-1}), x_t) \;+\; \frac{c^2\Lambda^2(x^t)}{2} \\
&\le -c\cdot\ell_p\left(\sum_{F\in\mathcal{F}} w_t(F)\cdot F(x^{t-1}), x_t\right) \;+\; \frac{c^2\Lambda^2(x^t)}{2} \\
&= -c\cdot\ell_p(C(x^{t-1}), x_t) \;+\; \frac{c^2\Lambda^2(x^t)}{2},
\end{aligned}$$

where the second inequality above is a consequence of the convexity of $\ell_p(\cdot, x_t)$. It follows by summing over t that

$$\ln\frac{W_v}{W_u} \;\le\; -c\sum_{t=u}^{v-1}\ell_p(C(x^{t-1}), x_t) \;+\; \frac{c^2}{2}\sum_{t=u}^{v-1}\Lambda^2(x^t). \tag{32}$$

On the other hand, it is clear that $W_v \ge \max_{F\in\mathcal{F}} \exp\{-c\sum_{s=u}^{v-1}\ell_p(F(x^{s-1}), x_s)\}$, and therefore

$$\ln\frac{W_v}{W_u} \;\ge\; -c\min_{F\in\mathcal{F}}\sum_{t=u}^{v-1}\ell_p(F(x^{t-1}), x_t) \;-\; \ln|\mathcal{F}|. \tag{33}$$

Combining inequalities (32) and (33) completes the proof of the lemma.

Proof of Proposition 6: To simplify notation, define $b_j = 2^j$ for $j \geq 0$. Let $F_t = F(X^{t-1})$ for $F \in \mathcal{F}$, and let $\tilde{F}_t = \tilde{F}(X^{t-1})$. Fix $\delta > 0$ such that $|F^{(r)}| = O(r^{(1-\delta)/2p})$ and let $M(1) \leq M(2) \leq \cdots$ be increasing constants such that $|F^{(r)}| \leq M(r)$ for each r and $M(r) = O(r^{(1-\delta)/2p})$. The definition of $\tilde{F}$ ensures that $|\tilde{F}(X^{t-1})| \leq M(b_j)$ for $1 \leq t < b_{j+1}$.

Fix $j_0 \geq 1$, let $n \geq b_{j_0+1}$ and define $k = k_n = \lfloor \log_2 n \rfloor$. The cumulative loss of $\tilde{F}$ on $X_1, \ldots, X_n$ may be written as follows:

$$\sum_{t=1}^{n} \ell_p(\tilde{F}_t, X_t) \;=\; \sum_{1 \leq t < b_{j_0}} \ell_p(\tilde{F}_t, X_t) \;+\; \sum_{j=j_0}^{k-1} \sum_{b_j \leq t < b_{j+1}} \ell_p(\tilde{F}_t, X_t) \;+\; \sum_{t=b_k}^{n} \ell_p(\tilde{F}_t, X_t).$$

Define $\Lambda_j(x^t) = \max_{F \in \mathcal{F}_j} |F(x^{t-1}) - x_t|^p$. Repeated application of Lemma 3 shows that the sum of the last two terms above is at most

$$\begin{aligned}
&\sum_{j=j_0}^{k-1} \min_{F \in \mathcal{F}_j} \sum_{b_j \leq t < b_{j+1}} \ell_p(F_t, X_t) \;+\; \min_{F \in \mathcal{F}_k} \sum_{t=b_k}^{n} \ell_p(F_t, X_t) \;+\; \sum_{j=j_0}^{k} \frac{\log |\mathcal{F}_j|}{c_j} \\
&\qquad\qquad + \sum_{j=j_0}^{k} \sum_{b_j \leq t < b_{j+1}} \frac{c_j \, \Lambda_j^2(X^t)}{2} \\
&\leq \min_{F \in \mathcal{F}_{j_0}} \sum_{t=1}^{n} \ell_p(F_t, X_t) \;+\; \sum_{j=0}^{k} \frac{\log |\mathcal{F}_j|}{c_j} \;+\; \sum_{j=0}^{k} \sum_{b_j \leq t < b_{j+1}} \frac{c_j \, \Lambda_j^2(X^t)}{2}
\end{aligned}$$

The two previous displays show that for $n \geq b_{j_0+1}$,

$$\begin{aligned}
L_n(\tilde{F}, \mathbf{X}) \;&\leq\; \min_{F \in \mathcal{F}_{j_0}} L_n(F, \mathbf{X}) \;+\; \frac{1}{n} \sum_{1 \leq t < b_{j_0}} \ell_p(\tilde{F}_t, X_t) \;+\; \sum_{j=0}^{k} \frac{\log |\mathcal{F}_j|}{n c_j} \\
&\quad + \sum_{j=0}^{k} \sum_{b_j \leq t < b_{j+1}} \frac{c_j \, \Lambda_j^2(X^t)}{2n}
\end{aligned} \tag{34}$$

We wish to show that the last three terms in (34) tend to zero as n tends to infinity. As $\tilde{F}_t$ and X_t are finite with probability one, it is clear that

$$\frac{1}{n} \sum_{1 \leq t < b_{j_0}} \ell_p(\tilde{F}_t, X_t) \;\to\; 0 \quad \text{wp1.} \tag{35}$$

Moreover, as $\log |\mathcal{F}_j| = j$ and $n \geq b_k$,

$$\sum_{j=0}^{k} \frac{\log |\mathcal{F}_j|}{n c_j} \;\leq\; \sum_{j=0}^{k} \frac{j}{b_k \, c_j} \;\leq\; \frac{k(k+1)}{b_k \, c_k} \;=\; \frac{k(k+1)}{2^{\sqrt{k}}} \tag{36}$$

which tends to zero, since $k = k_n$ tends to infinity with n. Now note that for $b_j \leq t < b_{j+1}$,

$$\Lambda_j^2(X^t) \;\leq\; 2^p \max_{1 \leq r \leq b_j} |F^{(r)}|^{2p} \;+\; 2^p \, |X_t|^{2p} \;\leq\; 2^p \, M(t)^{2p} \;+\; 2^p \, |X_t|^{2p}.$$

By assumption, there is some $\gamma > 0$ such that $\sup_{t\geq 1} E|X_t|^{p(1+\gamma)}$ is finite. Fix $0 < \eta < \delta$ such that $(1+\gamma)(1-\eta) > 1$. Then for $j \geq j_1 := \lceil \eta^{-2} \rceil$ the constants c_j satisfy $c_j \leq 2^{-j(1-\eta)}$. In particular, $c_j \leq 2t^{-(1-\eta)}$ for $j \geq j_1$ and $b_j \leq t < b_{j+1}$. Therefore

$$\sum_{j=0}^{k} \sum_{b_j \leq t < b_{j+1}} \frac{c_j \, \Lambda_j^2(X^t)}{2n} \;\leq\; \frac{2^{(2p+j_1)}}{n} \sum_{t=1}^{2n} \frac{M(t)^{2p}}{t^{1-\eta}} \;+\; \frac{2^{(2p+j_1)}}{n} \sum_{t=1}^{2n} \frac{|X_t|^{2p}}{t^{1-\eta}}. \tag{37}$$

Since $M(t)^{2p} = O(t^{(1-\delta)})$ and $\eta < \delta$, the first term on the right side of (37) converges to zero as $n \to \infty$. As for the second term, fix $\alpha > 0$ and note by Markov's inequality,

$$P\left\{ \frac{|X_t|^p}{t^{1-\eta}} \geq \alpha \right\} \;=\; P\{|X_t|^{p\,(1+\gamma)} \;\geq\; (\alpha\, t^{1-\eta})^{1+\gamma}\} \;\leq\; \frac{\sup_{s\geq 1} E|X_s|^{p\,(1+\gamma)}}{\alpha^{1+\gamma}\, t^{(1-\eta)(1+\gamma)}}.$$

By assumption, the supremum above is finite, and since $(1-\delta)(1+\gamma) > 1$, the sum over t of the rightmost probabilities is finite. It follows from the Borel Cantelli lemma that, with probability one, $|X_t|^p/t^{1-\eta} \geq \alpha$ for only finitely many value of t. Therefore

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{t=1}^{2n} \frac{|X_t|^{2p}}{t^{1-\eta}} \;\leq\; 2\alpha \cdot \limsup_{m\to\infty} \frac{1}{m} \sum_{t=1}^{m} |X_t|^p \quad \text{wp1}$$

As $\alpha > 0$ was arbitrary and the limit supremum is finite by assumption (ii), we conclude that

$$\frac{2^{(2p+j_1)}}{n} \sum_{t=1}^{2n} \frac{|X_t|^{2p}}{t^{1-\eta}} \;\to 0 \quad \text{wp1.}$$

Combining these relations with inequality (34), it follows that

$$\limsup_{n\to\infty} L_n(\tilde{F}, \mathbf{X}) \;\leq\; \min_{F\in\mathcal{F}_{j_0}} \limsup_{n\to\infty} L_n(F, \mathbf{X}) \quad \text{wp1}$$

and

$$\limsup_{n\to\infty} \left[L_n(\tilde{F}, \mathbf{X}) - \min_{F\in\mathcal{F}_{j_0}} L_n(F, \mathbf{X}) \right] \;\leq\; 0 \quad \text{wp1.}$$

As $j_0 \geq 1$ was arbitrary, inequalities (19) and (20) are immediate.

10.4 Proof of Theorem 6

Definition: Let $\mathcal{C}(k)$ be the family of functions $g : \mathbb{R}^k \to [-a_k, a_k]$ that are measurable with respect to the sigma field generated by sets of the form $C_1 \times \cdots \times C_k$ with $C_i \in \pi_k$. Thus each $g \in \mathcal{C}(k)$ is of the form

$$g(x_1, \ldots, x_k) \;=\; \sum_{C_1,\ldots,C_k \in \pi_k} u(C_1, \ldots, C_k)\, I\{x_1 \in C_1, \ldots, x_k \in C_k\},$$

where, for every choice of $C_1, \ldots, C_k \in \pi_k$, $u(C_1, \ldots, C_k)$ is a fixed number in $[-a_k, a_k]$.

Proof of Theorem 6: Let $\mathbf{X}$ be a stationary ergodic process such that $E|X_1|^q < \infty$ for some $q > p$. Let $k \geq 1$ and consider for the moment a sequence of cells $C_1^k = C_1, \ldots, C_k \in \pi_k$ such that $P\{X_k \in C_1, \ldots, X_1 \in C_k\} > 0$. For $u \in [-a_k, a_k]$ define

$$\phi(u, C_1^k) \;:=\; E[\,\ell_p(u, X_{k+1}) I\{X_k \in C_1, \ldots, X_1 \in C_k\}\,],$$

and for each $t \geq k+1$ define

$$\hat{\phi}_t(u, C_1^k) \;:=\; \frac{1}{t-k-1} \sum_{s=k+1}^{t-1} \ell_p(u, X_s) I\{X_{s-1} \in C_1, \ldots, X_{s-k} \in C_k\}.$$

Note that $\phi(u, C_1^k)$ is a bounded, strictly convex function of $u \in [-a_k, a_k]$, and that the same is true, with probability one, of the functions $\hat{\phi}_t(u, C_1^k)$ for each t so large that $\sum_{s=k+1}^{t} I\{X_{s-1} \in C_1, \ldots, X_{s-k} \in C_k\} \geq 1$. The ergodic theorem implies that, with probability one, $\hat{\phi}_t(u, C_1^k) \to \phi(u, C_1^k)$ for each $u \in [-a_k, a_k]$. It then follows from standard results in convex analysis (see [35]) that this convergence is uniform, in the sense that

$$\sup_{u \in [-a_k, a_k]} |\hat{\phi}_t(u, C_1^k) - \phi(u, C_1^k)| \;\to\; 0 \;\; \text{wp1} \;\; \text{as} \;\; t \to \infty. \tag{38}$$

Define minima

$$\hat{u}_t(C_1^k) \;:=\; \operatorname*{arg\,min}_{u \in [-a_k, a_k]} \hat{\phi}_t(u, C_1^k) \quad \text{and} \quad u^*(C_1^k) \;:=\; \operatorname*{arg\,min}_{u \in [-a_k, a_k]} \phi(u, C_1^k).$$

For t sufficiently large, the strict convexity of $\hat{\phi}_t$ and ϕ imply that both minima are achieved, and are unique. Moreover, (38) guarantees that $\hat{u}_t(C_1^k) \to u^*(C_1^k)$ with probability one as $t \to \infty$.

By definition, the prediction $H^k(X^{t-1})$ is equal to $\hat{u}_t(\pi_k(X_{t-1}), \ldots, \pi_k(X_{t-k}))$. Define a new prediction scheme $G^k(X^{t-1}) = u^*(\pi_k(X_{t-1}), \ldots, \pi_k(X_{t-k}))$. Then the difference $|L_n(H^k) - L_n(G^k)|$ is at most

$$\sum_{C_1^k} \left[\frac{1}{n} \sum_{t=1}^{n} |\, \ell_p(\hat{u}_t(C_1^k), X_t) - \ell_p(u^*(C_1^k), X_t)\,| \; I\{X_{t-1} \in C_1, \ldots, X_{t-k} \in C_k\} \right].$$

We claim that each term in the sum over C_1^k tends to zero as n tends to infinity. If $P\{X_k \in C_1, \ldots, X_1 \in C_k\} = 0$, then the corresponding average is zero for each n with probability one. Suppose then that $P\{X_k \in C_1, \ldots, X_1 \in C_k\} > 0$. By an application of Lemma 1, it suffices to include in the second sum only those t for which $X_t \leq K$, where $K < \infty$ is fixed, but arbitrary. Under this restriction, the average tends to zero with

increasing n as $\hat{u}_t(C_1^k) \to u^*(C_1^k)$. Therefore

$$\begin{aligned}
\limsup_{n\to\infty} L_n(H^k) &= \limsup_{n\to\infty} L_n(G^k) \\
&= \lim_{n\to\infty} \sum_{C_1^k} \left[\frac{1}{n} \sum_{t=k+1}^{n} \ell_p(u^*(C_1^k), X_t) I\{X_{t-1} \in C_1, \ldots, X_{t-k} \in C_k\} \right] \\
&= \sum_{C_1^k} E[\, \ell_p(u^*(C_1^k), X_{k+1})\, I\{X_k \in C_1, \ldots, X_1 \in C_k\}] \\
&= \sum_{C_1^k} \min_{u\in[-a_k,a_k]} E[\, \ell_p(u, X_{k+1})\, I\{X_k \in C_1, \ldots, X_1 \in C_k\}\,] \\
&= \min_{g\in\mathcal{C}(k)} \sum_{C_1^k} E[\, \ell_p(g(X^k), X_{k+1})\, I\{X_k \in C_1, \ldots, X_1 \in C_k\}] \\
&= \min_{g\in\mathcal{C}(k)} E\ell_p(g(X^k), X_{k+1}). \qquad (39)
\end{aligned}$$

Now let $f : \mathbb{R}^r \to \mathbb{R}$ be bounded and uniformly continuous, and fix $\epsilon > 0$. For sufficiently large k, there exists $g \in \mathcal{C}(k)$ such that $E|X_{k+1} - g(X_1^k)|^p \le E|X_{r+1} - f(X_1^r)|^p + \epsilon$. It follows from Proposition 6 and the relation (39) that

$$\begin{aligned}
\limsup_{n\to\infty} L_n(\tilde{H}, \mathbf{X}) &\le \min_{k\ge 1} \limsup_{n\to\infty} L_n(H^k, \mathbf{X}) \\
&\le \min_{k\ge 1} \min_{g\in\mathcal{C}(k)} E[\ell(g(X^k), X_{k+1})] \\
&\le E|X_{r+1} - f(X_1^r)|^p + \epsilon.
\end{aligned}$$

As r, f, and $\epsilon > 0$ were arbitrary, $\tilde{H}$ is Cesaro optimal by Lemma 7.

References

[1] T. Ando and I. Amemiya, "Almost everywhere convergence of prediction sequences in L_p," *Z. Wahrscheinlichkeitstheorie verw. Geb.*, vol. 4, pp. 113-120, 1965.

[2] P.H. Algoet, "The strong law of large numbers for sequential decisions under uncertainty," *IEEE Trans. Info. Theory*, vol. 40, pp. 609-633, 1994.

[3] P.H. Algoet, "Universal schemes for prediction, gambling and portfolio selection," *Ann. Probab.*, vol. 20, pp. 901-941, 1992. Correction: *ibid*, vol. 23, pp. 474-478, 1995.

[4] P.H. Algoet, "Universal schemes for learning the best nonlinear predictor given the infinite past and side information," *IEEE Trans. Info. Theory*, vol. 45, pp. 1165-1185, 1999.

[5] K. Azuma, "Weighted sums of certain dependent random variables," *Tohoku Math. Journal*, vol. 68, pp. 357-367, 1967.

[6] D.H. Bailey, "Sequential schemes for classifying and predicting ergodic processes," PhD. thesis, Department of Mathematics, Stanford University, Stanford, CA, 1976.

[7] L. Breiman, *Probability*. SIAM, Philadelphia, PA, 1992.

[8] N. Cesa-Bianchi, "Analysis of two gradient-based algorithms for on-line regression," in *Proc. 12'th Annual Conference on Computational Learning Theory*, pp. 163-170, ACM Press, 1999.

[9] N. Cesa-Bianchi, Y. Freund, D.P. Helmbold, and M.K. Warmuth, "On-line prediction and conversion strategies," *Machine Learning*, vol. 25, pp. 71-110, 1996.

[10] N. Cesa-Bianchi, Y. Freund, D. Haussler, D.P. Helmbold, R.E. Schapire, and M.K. Warmuth, "How to use expert advice," *J. Assoc. Comp. Mach.*, vol. 44, pp. 427-485, 1997.

[11] A.P. Dawid, "The well-calibrated Bayesian (with discussion)," *J. Amer. Stat. Assoc.*, vol. 77, pp. 605-613, 1982.

[12] A.P. Dawid, "Statistical theory. The prequential approach (with discussion)," *J. Roy. Statist. Soc. A*, vol. 147, pp. 278-292, 1984.

[13] A.P. Dawid, "Calibration-based empirical probability (with discussion)," *Ann. Statist.*, vol. 13, pp. 1251-1285, 1985.

[14] A.P. Dawid and V.G. Vovk, "Prequential probability: Principles and properties," *Bernoulli*, vol. 5, pp. 125-162, 1999.

[15] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. Springer, New York, 1996.

[16] J.L. Doob, *Stochastic Processes*. Wiley, New York, 1953.

[17] M. Feder, N. Merhav, and M. Gutman, "Universal prediction of individual sequences," *IEEE Trans. Info. Theory*, vol. 38, pp. 1258-1270, 1992.

[18] D.P. Foster, "Prediction in the worst case," *Ann. Statist.*, vol. 19, pp. 1084-1090, 1991.

[19] D.P. Foster and R. Vohra, "Regret in the on-line decision problem," *Games and Economic Behavior*, vol. 29, pp. 1084-1090, 1999.

[20] T.S. Ferguson, *Mathematical Statistics*. Academic Press, San Diego, 1967.

[21] R.M. Gray, *Probability, Random Processes, and Ergodic Properties*. Springer, New York, 1988.

[22] L. Györfi and G. Lugosi, "Strategies for sequential prediction of stationary time series," in *Modeling Uncertainty: an Examination of its Theory, Methods, and Applications*, M. Dror, P. L'Ecuyer and F. Szidarovszky Eds, Kluwer, 2001.

[23] L. Györfi, G. Morvai, and S.J. Yakowitz, "Limits to consistent on-line forecasting for ergodic time series," *IEEE Trans. Info. Theory*, vol. 44, pp. 886-892, 1998.

[24] L. Györfi, G. Lugosi, and G. Morvai, "A simple randomized algorithm for consistent sequential prediction of ergodic time series," *IEEE Trans. Info. Theory*, vol. 45, pp. 2642-2650, 1999.

[25] D.H. Haussler, J. Kivenen, and M.K. Warmuth, "Sequential prediction of individual sequences under general loss functions," *IEEE Trans. Info. Theory*, vol. 44, pp. 1906-1925, 1998.

[26] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, pp. 13-30, 1963.

[27] N. Littlestone and M.K. Warmuth, "The weighted majority algorithm," *Info. and Comput.*, vol. 108, pp. 212-261, 1994.

[28] N. Merhav and M. Feder, "Universal Prediction," *IEEE Trans. Info. Theory*, vol. 44, pp. 2124-2147, 1998.

[29] D.S. Modha and E. Masry, "Memory-universal prediction of stationary random processes," *IEEE Trans. Info. Theory*, vol. 44, pp. 117-133, 1998.

[30] G. Morvai, S. Yakowitz, and L. Györfi, "Nonparametric inference for ergodic, stationary time series," *Ann. Stat.*, vol. 24, pp. 370-379, 1996.

[31] G. Morvai, S. Yakowitz, and P. Algoet, "Weakly convergent nonparametric forecasting of stationary time series," *IEEE Trans. Info. Theory*, vol. 43, pp. 483-498, 1997.

[32] A.H. Murphy and R.L. Winkler, "Reliability of subjective probability forecasts of precipitation and temperature," *JRSS Series C*, vol. 26, pp. 41-47, 1977.

[33] D.S. Ornstein, "Guessing the next output of a stationary process," *Israel J. Math*, vol. 30, pp. 292-296, 1978.

[34] K.R. Parthasarathy, *Probability Measures on Metric Spaces*. Academic Press, New York, 1967.

[35] R.T. Rockafellar, *Convex Analysis*. Princeton Univ. Press, Princeton, NJ, 1970.

[36] H.L. Royden, *Real Analysis*, 3'rd edition. Prentice Hall, NJ, 1988.

[37] B.Y. Ryabko, "Prediction of random sequences and universal coding," *IEEE Trans. Info. Theory*, vol. 44, pp. 2124-2147, 1988.

[38] K. Skouras and A.P. Dawid, "On efficient point prediction systems," *J. Roy. Statist. Soc. B*, vol. 60, pp. 765-780, 1998.

[39] K. Skouras and A.P. Dawid, "On efficient probability forecasting systems," *Biometrika*, vol. 86, pp. 765-784, 1999.

[40] W.F. Stout, *Almost Sure Convergence*. Academic Press, New York, 1974.

[41] V. Vovk, "Aggregating strategies," in *Proc. 3'rd Annual Workshop on Computational Learning Theory*, pp. 371-383, Morgan Kaufman, San Mateo, 1990

[42] T. Weissman and N. Merhav, "Universal prediction of random binary sequences in a noisy environment," Preprint, 2000.

[43] T. Weissman and N. Merhav, "Universal prediction of individual binary sequences in the presence of noise," *IEEE Trans. Info. Theory*, vol. 47, pp. 2151-2173, 2001.