*The Annals of Applied Probability*
2004, Vol. 14, No. 1, 54–89


# UNIVERSAL PREDICTION OF RANDOM BINARY SEQUENCES IN A NOISY ENVIRONMENT

BY TSACHY WEISSMAN AND NERI MERHAV

*Stanford University and Technion*

Let $X = \{(X_t, Y_t)\}_{t\in\mathbb{Z}}$ be a stationary time series where $X_t$ is binary valued and $Y_t$, the *noisy observation* of $X_t$, is real valued. Letting $\mathbf{P}$ denote the probability measure governing the joint process $\{(X_t, Y_t)\}$, we characterize $U(l, \mathbf{P})$, the optimal asymptotic average performance of a predictor allowed to base its prediction for $X_t$ on $Y_1, \ldots, Y_{t-1}$, where performance is evaluated using the loss function $l$. It is shown that the stationarity and ergodicity of $\mathbf{P}$, combined with an additional "conditional mixing" condition, suffice to establish $U(l, \mathbf{P})$ as the fundamental limit for the almost sure asymptotic performance. $U(l, \mathbf{P})$ can thus be thought of as a generalized notion of the Shannon entropy, which can capture the sensitivity of the underlying clean sequence to noise. For the case where $\mathbf{X} = \{X_t\}$ is governed by $P$ and $Y_t$ given by $Y_t = g(X_t, N_t)$ where $g$ is any deterministic function and $\mathbf{N} = \{N_t\}$, the noise, is any i.i.d. process independent of $\mathbf{X}$ (namely, the case where the "clean" process $\mathbf{X}$ is passed through a fixed memoryless channel), it is shown that, analogously to the noiseless case, there exist *universal predictors* which do not depend on $P$ yet attain $U(l, \mathbf{P})$. Furthermore, it is shown that in some special cases of interest [e.g., the binary symmetric channel (BSC) and the absolute loss function], there exist *twofold universal predictors* which do not depend on the noise distribution either. The existence of such universal predictors is established by means of an explicit construction which builds on recent advances in the theory of prediction of individual sequences in the presence of noise.

**1. Introduction.** Let $\{(X_t, Y_t)\}_{t\in\mathbb{Z}}$ be a random process with components taking values in $\{0, 1\} \times \mathbb{R}$, where $\mathbf{X} = \{X_t\}$ is the binary-valued stationary time series of interest and $\mathbf{Y} = \{Y_t\}$ can be perceived as its real-valued noisy observation sequence. As a concrete example of practical interest, assume that $\mathbf{Y}$ is the output of a fixed and memoryless noisy channel whose input is $\mathbf{X}$. Without loss of generality, this situation can be modeled by the existence of some measurable function $g : \{0, 1\} \times \mathbb{R} \to \mathbb{R}$ such that $Y_t = g(X_t, N_t)$ where $\mathbf{N} = \{N_t\}$ is some i.i.d. process independent of $\mathbf{X}$. A predictor $F$, in this noisy context, is a sequence of measurable mappings $\{F_t\}_{t\geq 1}$, $F_t : \mathbb{R}^{t-1} \to [0, 1]$, where $F_t(Y_1, \ldots, Y_{t-1})$ is interpreted as the prediction of $F$ at the time $t$, based on the noisy past. The performance of the predictor is evaluated using a loss function

Received August 2000; revised January 2002.
*AMS 2000 subject classifications.* Primary 62M20, 60G25; secondary 94A17.
*Key words and phrases.* Prediction with noise, conditional mixing, universal prediction, Shannon entropy, filtering, asymptotic optimality, prediction with experts, sequential decisions, martingale difference, generalized ergodic theorem.

$l:\{0,1\}\times[0,1]\to[0,\infty)$, so that $l(X_t, F_t(Y_1,\ldots,Y_{t-1}))$ is the loss of the predictor $F$, at time $t$. Ultimately, the goal of the predictor is to minimize its average performance $\frac{1}{n}L_F(Y_1^n, X_1^n) =_{\text{def}} \frac{1}{n}\sum_{t=1}^n l(X_t, F_t(Y_1,\ldots,Y_{t-1}))$. As is straightforward to see (and as will be formally established in the sequel), the best predictor in the sense of minimizing the *expected* average loss $\frac{1}{n}EL_F(Y_1^n, X_1^n)$ is the Bayes optimal one, namely, that minimizing at each point in time the expected loss of the prediction, given the available noisy past. Unfortunately, there are two main types of difficulties associated with the employment of the Bayes optimal predictor. The first is an algorithmic, computational difficulty. The implementation of the Bayes predictor requires knowledge of the conditional distribution of $X_t$ given $Y_1^{t-1} = (Y_1,\ldots,Y_{t-1})$, the explicit computation of which turns out to be of prohibitively large complexity, growing exponentially with $t$ even in the simplest of situations. The second difficulty, which is deeper and more fundamental, is its dependence on $\mathbf{P}$, the distribution of the process $\mathbf{X}$. Clearly, even when the channel is known, minimizing the expected loss given the noisy past requires full knowledge of the distribution of $\mathbf{X}$ which, in the majority of practical situations, will not be available.

The main contribution of this work is in establishing, for the above-described situation, the existence of universal predictors, that is, predictors that are independent of $\mathbf{P}$ and yet asymptotically achieve optimum performance. Moreover, these predictors have a relatively simple structure that avoids an explicit estimation of the distribution of $X_t$ conditioned on its noisy past.

The existence of universal predictors for the stochastic noiseless setting is a well-known fact, extensively established by the cumulative work of many researchers (cf. [1], Section V, [15], Section II and the many references therein). The basic idea behind all universal predictors in the noiseless case in a nutshell is essentially the same: learning the unknown distribution from experience by working with the empirical distribution induced by the available past sequence. This is true even for prediction schemes that are not explicitly presented this way, for example, the predictor (cf. [7], Section V) induced by the Lempel–Ziv universal coding scheme [21].

To the best of our knowledge, however, the question of existence and the problem of construction of universal predictors for the noisy setting has not been explicitly addressed in the literature. As mentioned earlier, in the literature related to prediction, universality was investigated exclusively for the noiseless case. In the classical and well-developed theory of filtering, on the other hand, which is dedicated to a variety of situations involving noisy observations, the probabilistic model is usually assumed to be fully specified, and, accordingly, the prediction schemes obtained are distribution dependent. An exception to this is the line of work related to robust filtering, which was carried out by researchers from the signal processing and information theoretic communities in the seventies and eighties (cf. [3, 6, 8, 14, 13, 16] and the references therein). The setting considered

in robust filtering, however, is completely different from the one we consider in this paper. The typical problem considered in robust filtering is that where the power spectral densities of the signal and the noise (assumed jointly weakly stationary), rather than being fully known, are only known (or assumed) to belong to some set. One then seeks a (time-invariant) filter which is minimax optimal under mean squared error loss. The prediction (filtering) schemes obtained in robust filtering, which are guided by the minimax formulation, are not adaptive. They are essentially designed to minimize the expected loss for the worst possible signal and noise sources in the uncertainty set. This is in contrast to the approach we present here which allows one to compete with classes of sources as rich as that of all stationary and ergodic processes; whereas, under the minimax formulation of robust filtering, the uncertainty classes needed to be considerably more limited to allow for meaningful results. In addition, while in robust filtering attention is restricted to the mean square error loss as the performance criterion, in this work we consider the case of a general loss function and pay most of our attention to the actual rather than the expected loss. Accordingly, the setting considered in this work is general enough to model a wide range of noisy situations arising in data compression (e.g., predictive coding of noisy images), signal processing (filtering), learning theory, statistics and economics. One can think of $X_t$ as the label and $Y_t$ as the instance. Then our setting coincides with that of learning with unlabeled instances; see [5].

It should be emphasized that there is no unique natural and straightforward way to extend the basic idea behind the universal prediction schemes in the noiseless setting for the noisy one. This is because the approximated distribution of the observed sequence $\mathbf{Y}$, available through the accumulating observations of the noisy past, does not, in general, naturally induce an empirically approximated distribution of the next *clean* bit (conditioned on its noisy past). This conditional distribution is hard to estimate even when the channel is completely known. Furthermore, even in situations where the empirical distribution of the noisy observation sequence does naturally lead to an estimation of the distribution of the clean bit based on its noisy past (e.g., cases where it may seem natural to use the plug-in approach of deconvolving the empirical distribution of the noisy sequence with the channel), it is not clear that this estimated distribution has the desired convergence properties. Correspondingly, it is not clear that the prediction strategy that such a plug-in approach induces asymptotically attains the best achievable performance.

Our approach to the construction of the universal predictors uses recent advances in the theory of prediction of individual sequences in the presence of noise [19, 20]. The main contribution of this recent line of research was in establishing the existence of predictors, fed by a noisy version of an underlying clean individual sequence, that compete successfully with a given set of prediction schemes, uniformly for all individual sequences. The idea behind the choice

of the predictors proposed here is to use the prediction schemes of the individual sequence setting [19] and to claim that these compete with classes of prediction schemes that are on the one hand reasonably small and hence yield small redundancy rates and, on the other hand, dense in the sense of covering members whose asymptotic performance is the same as the optimum. Specifically, we construct predictors which compete with the set of all Bayesian–Markovian prediction schemes whose performance is guaranteed for each individual sequence, a fortiori, in the probabilistic setting considered here. A similar approach was recently taken in [10] and in [9] for the original noise-free case, where efficient and relatively simple predictors for ergodic time series were constructed using results from the theory of prediction of individual sequences.

As will be demonstrated in Section 4, the approach taken here to the construction of universal predictors for this noisy setting is a fruitful one. In particular, it will be established in Section 4 that when the noisy sequence is the output of an *arbitrary* memoryless channel, there always exists a *universal* predictor; that is, a predictor which almost surely asymptotically attains the best theoretically achievable performance no matter what the clean sequence is, provided that certain regularity conditions are met.

It should be emphasized that the finding that for essentially *all* memoryless channels there exists a universal predictor is far from being trivial. This is because the traditional points of view taken in the construction of universal predictors for the noise-free setting do not lend themselves easily to the case of noisy observations. In the noise-free case, universal predictors are, roughly speaking, always viewed either as prediction schemes which strive to identify and imitate predictors which have been proved efficient on the past sequence, or prediction schemes which try to learn the conditional distribution of the next outcome of the sequence from experience, given its past. Both of these points of view, however, are hard to extend to the noisy setting. This is because, due to the noise, it is unclear which predictors have been doing well on the past sequence (if one adopts the first point of view) and it is equally unclear how to "learn" the conditional distribution of the next *clean* outcome based on the past noisy sequence since the next clean outcome will never be observed. Nevertheless, as will be established in Section 4, this difficulty can be alleviated and universal predictors can be constructed for the noisy setting.

The outline of this work is as follows. In Section 2, we present our notation conventions and regularity conditions. The two main parts of this work are presented in Sections 3 and 4, respectively. Section 3 is dedicated to assessing the ultimate limits of prediction performance in the noisy setting. In Section 4, the existence of universal predictors is established.

**2. Notation conventions.** Throughout the paper, random elements will be denoted by capital letters; thus, for any integers $m \leq n$ we let $X_m^n$ denote a random binary vector $(X_m, \ldots, X_n) \in \{0, 1\}^{n-m+1}$. Infinite random binary sequences will

be denoted by boldface letters, so that $\mathbf{X} = (\ldots, X_{-1}, X_0, X_1, \ldots) \in \{0,1\}^{\infty}$. Similarly, if the components of $\{Y_i\}$ are real valued, we let $Y_m^n$ and $\mathbf{Y}$ denote the random elements of $\mathbb{R}^{n-m+1}$ and $\mathbb{R}^{\mathbb{Z}}$, respectively. Deterministic elements, or specific sample values of random elements, will be denoted by the respective lowercase letters.

Suppose now that we have a process with components $(X_t, Y_t)$ taking values in $\{0,1\} \times \mathbb{R}$. In this framework, we think of $\{X_t\}_{t \geq 1}$ as the "clean" sequence of interest to predict, which, unfortunately, we cannot directly access and thus must base our predictions for the bit $X_t$ on its "noisy past" $Y_1^{t-1}$. Thus, a predictor $F$ is a sequence of functions $F_t : \mathbb{R}^{t-1} \to [0,1]$, $t \geq 1$. For any $\mathbf{x} \in \{0,1\}^{\mathbb{Z}}$, $\mathbf{y} \in \mathbb{R}^{\mathbb{Z}}$ and a given loss function $l : [0,1] \times \{0,1\} \to [0,\infty)$, we let

$$L_F(y_{t_1}^{t_2}, x_{t_1}^{t_2}) \stackrel{\text{def}}{=} \sum_{t=t_1}^{t_2} l\big(F_t(y_1^{t-1}), x_t\big)$$

denote the average loss from time $t_1$ up to time $t_2$ of the predictor $F$ when fed with the noisy sequence $\mathbf{y}$ and judged w.r.t. the clean sequence $\mathbf{x}$. The dependence of $L_F$ on the particular loss function $l$ is suppressed in the notation as it is assumed fixed and known in any prediction problem. We will further let

$$L_F(\mathbf{y}, \mathbf{x}) \stackrel{\text{def}}{=} \limsup_{n \to \infty} \frac{1}{n} L_F(y_1^n, x_1^n)$$

denote the asymptotic performance of $F$. For a stochastic process $\{(X_t, Y_t)\}_{t \in \mathbb{Z}}$ governed by a probability measure $\mathbf{P}$, we let $E^{\mathbf{P}}$ denote expectation w.r.t. $\mathbf{P}$. We omit the superscript $\mathbf{P}$ whenever clear from the context.

For any family of random variables $\{R_i\}_{i \in I}$ (where $I$ is an arbitrary index set), we let $\sigma(\{R_i\}_{i \in I})$ denote the smallest sigma field with respect to which all the $\{R_i\}_{i \in I}$ are measurable. For any process $\mathbf{W} = \{W_t\}_{t \in \mathbb{Z}}$ we let $\mathcal{F}_W^t =_{\text{def}} \sigma(W_1^t)$ (the information known to one who observes the process from time 1 up to time $t$). Similarly, we let $\mathcal{F}_W =_{\text{def}} \sigma(\mathbf{W})$. Following the customary abuse of notation, for any random variable $R$, we shall frequently write $E(R|Y_1^t)$ instead of $E(R|\mathcal{F}_Y^t)$. Equality between random variables, when not explicitly specified, should be interpreted in the almost sure sense w.r.t. a probability measure which should be clear from the context.

Throughout this work the following will be kept intact.

ASSUMPTION 1. The function $l(\cdot,\cdot)$, together with the derivatives of $l(\cdot,0)$ and $l(\cdot,1)$ (assumed to exist), are bounded throughout by the constant $B$.

Assumption 1 is not essential for the validity of our results. The reason it is made is to facilitate a simpler exposition, emphasizing the important points and avoiding insignificant technicalities. This is also the reason for restricting attention to the case where $\mathbf{X}$ is binary valued and $\mathbf{Y}$ is real valued. The results and

approach presented in the sequel carry over straightforwardly to the case where the components of $\mathbf{X}$ belong to any finite alphabet and the components of $\mathbf{Y}$ to an arbitrary Polish space. Finally, we remark that, though in real-life prediction situations time starts at $t = 1$ and, accordingly, all quantities of interest for our setting ultimately depend only on $\{(X_t, Y_t)\}_{t\geq 1}$, Kolmogorov's extension theorem allows us to assume that the time axis is infinite in both directions.

**3. Fundamental limitations on prediction performance.** Let $\{(X_t, Y_t)\}_{t\in\mathbb{Z}}$ be a process taking values in the measurable space $\{\{\{0,1\}\times\mathbb{R}\}^{\mathbb{Z}}, \mathcal{F}_{XY}\}$ and distributed according to the stationary probability measure $\mathbf{P}$. We let

$$\begin{aligned} U(l,\mathbf{P}) &\stackrel{\text{def}}{=} E^{\mathbf{P}}\Big\{\min_{0\leq\alpha\leq 1}\big\{\mathbf{P}\{X_0=0|Y_{-\infty}^{-1}\}\cdot l(\alpha,0) \\ &\qquad + \mathbf{P}\{X_0=1|Y_{-\infty}^{-1}\}\cdot l(\alpha,1)\big\}\Big\} \\ &= E^{\mathbf{P}} A_l(\mathbf{P}\{X_0=1|Y_{-\infty}^{-1}\}) \end{aligned} \tag{1}$$

denote the *Bayes envelope* for our setting, where, for any loss function $l$, we define $A_l : [0,1]\to[0,\infty)$ by

$$A_l(p) \stackrel{\text{def}}{=} \inf_{0\leq\alpha\leq 1}\big\{(1-p)\cdot l(\alpha,0) + p\cdot l(\alpha,1)\big\}. \tag{2}$$

Note that our assumptions on the loss function $l$ allow replacing the infimum by a minimum in (2) and guarantee the boundedness and continuity of $A_l(\cdot)$, as the minimization is performed in a compact set over continuous functions of $p$ which are uniformly bounded. It is also easily verified that $A_l(\cdot)$ is concave. For any $0\leq p\leq 1$, we now let

$$\Phi_l(p) = \arg\min_{0\leq\alpha\leq 1}\big\{(1-p)\cdot l(\alpha,0) + p\cdot l(\alpha,1)\big\},$$

where, for concreteness, if there is more than one minimizing value, we take the lowest one. Note that $\Phi_l(p)$ can be interpreted as the best prediction to make, in the sense of minimizing expected loss, when the loss is measured with the loss function $l$ and the outcome is determined according to a flip of a coin with probability $p$ to hit one.

The following theorem establishes $U(l,\mathbf{P})$ as the acheivable limitation on prediction performance in the noisy setting.

THEOREM 1. *Let* $\mathbf{P}$ *be stationary.*

(a) *For any predictor* $F$ *we have*

$$\liminf_{n\to\infty}\frac{1}{n}E^{\mathbf{P}}L_F(Y_1^n, X_1^n) \geq U(l,\mathbf{P}). \tag{3}$$

(b) *The predictor* $F^{\mathrm{opt}}$ *given by*

$$F_t^{\mathrm{opt}}(Y_1^{t-1}) = \Phi_l(\mathbf{P}\{X_t = 1|Y_1^{t-1}\}) \tag{4}$$

*satisfies*

$$\lim_{n\to\infty} \frac{1}{n} E^{\mathbf{P}} L_{F^{\mathrm{opt}}}(Y_1^n, X_1^n) = U(l, \mathbf{P}). \tag{5}$$

(c) *If* $\mathbf{P}$ *is also ergodic then*

$$\lim_{n\to\infty} \frac{1}{n} L_{F^{\mathrm{opt}}}(Y_1^n, X_1^n) = U(l, \mathbf{P}) \qquad a.s. \tag{6}$$

PROOF. To prove (a) we note that, by the definition of $A_l(\cdot)$, we have

$$\begin{aligned} \frac{1}{n} E^{\mathbf{P}} L_F(Y_1^n, X_1^n) &\geq \frac{1}{n} \sum_{t=1}^{n} E\big[A_l(\mathbf{P}\{X_t = 1|Y_1^{t-1}\})\big] \\ &= \frac{1}{n} \sum_{t=1}^{n} E\big[A_l(\mathbf{P}\{X_0 = 1|Y_{-t-1}^{-1}\})\big]. \end{aligned} \tag{7}$$

Therefore

$$\liminf_{n\to\infty} \frac{1}{n} E^{\mathbf{P}} L_F(Y_1^n, X_1^n) \geq \lim_{t\to\infty} E\big[A_l(\mathbf{P}\{X_0 = 1|Y_{-t-1}^{-1}\})\big] \tag{8}$$

$$= E\Big[A_l\Big(\lim_{t\to\infty} \mathbf{P}\{X_0 = 1|Y_{-t-1}^{-1}\}\Big)\Big] \tag{9}$$

$$= E\big[A_l(\mathbf{P}\{X_0 = 1|Y_{-\infty}^{-1}\})\big] \tag{10}$$

$$= U(l, \mathbf{P}), \tag{11}$$

where (8) follows from Cèsaro's theorem, (9) follows by bounded convergence and the continuity of $A_l(\cdot)$ and (10) follows from martingale convergence (cf., in particular, [4], Theorem 5.21). The proof of (b) is immediate upon noting that when $F = F^{\mathrm{opt}}$, (7) holds with equality. To establish (c) note that since

$$\frac{1}{n} L_{F^{\mathrm{opt}}}(Y_1^n, X_1^n) = \frac{1}{n} \sum_{t=1}^{n} l\big(\Phi_l(\mathbf{P}\{X_t = 1|Y_1^{t-1}\}), X_t\big),$$

we have almost surely

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} L_{F^{\mathrm{opt}}}(Y_1^n, X_1^n) &= El\big(\Phi_l(\mathbf{P}\{X_0 = 1|Y_{-\infty}^{-1}\}), X_0\big) \\ &= E\big[E\big\{l\big(\Phi_l(\mathbf{P}\{X_0 = 1|Y_{-\infty}^{-1}\}), X_0\big)\big|Y_{-\infty}^{-1}\big\}\big] \\ &= E\big[A_l(\mathbf{P}\{X_0 = 1|Y_{-\infty}^{-1}\})\big] \\ &= U(l, \mathbf{P}), \end{aligned} \tag{12, 13}$$

where the first equality follows by combining the stationarity and ergodicity of $\mathbf{P}$ with [4], Theorem 5.21 and invoking Breiman's generalized ergodic theorem. The equality before last can be verified by writing out explicitly the inner conditional expectation in (12). □

As was pointed out in previous works (e.g., [1]), in which part (c) of Theorem 1 was proved for the noiseless case, the celebrated Shannon–McMillan–Breiman (SMB) theorem (the asymptotic equipartition property) is obtained from this part of the theorem by letting $l$ be the logarithmic loss function and $\mathbf{X} = \mathbf{Y}$. Thus, part (c) of Theorem 1 can be regarded as a generalization of the SMB theorem for the case where the predictor accesses a noisy version of the past and its performance is evaluated using a general loss function.

To complete the picture displayed in Theorem 1, it would be natural to determine whether the almost sure analogue of part (a) of the theorem holds (for ergodic processes), as is known to be the case in the noise-free setting (cf. [1]). Our first step in this direction is to establish the following.

LEMMA 1. *Let $\{(X_t, Y_t)\}_{t\in\mathbb{Z}}$ be conditionally mixing in the sense that*

$$\sum_{s=1}^{\infty} \sup_{t\geq 1} E\left|\Pr\{X_{t+s} = a | X_t = a, \mathcal{F}_Y^{t+s-1}\} - \Pr\{X_{t+s} = a | \mathcal{F}_Y^{t+s-1}\}\right| < \infty \tag{14}$$

*for each $a \in \{0, 1\}$. Then for any predictor $F$ we have*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \left[l\left(F_t(Y_1^{t-1}), X_t\right) - E\left\{l\left(F_t(Y_1^{t-1}), X_t\right) | \mathcal{F}_Y^{t-1}\right\}\right] = 0 \qquad \text{a.s.} \tag{15}$$

We refer to (14) as a "conditional mixing condition" as it essentially implies that for two points in time $t_1 < t_2$, $X_{t_1}$ and $X_{t_2}$ are approximately independent conditioned on $\mathcal{F}_Y^{t_2-1}$ when $t_2 - t_1$ is large. One example for a situation of interest where (14) holds is the case where $\{X_t\}$ is a first-order homogeneous noncyclic Markov chain and the noise, $\{N_t\}$, is i.i.d. In this case

$$\begin{aligned}
&\Pr\{X_{t+s} = a | \mathcal{F}_Y^{t+s-1}\} \\
&\quad = \Pr\{X_{t+s} = a | X_t = 0, \mathcal{F}_Y^{t+s-1}\} \Pr\{X_t = 0 | \mathcal{F}_Y^{t+s-1}\} \\
&\qquad + \Pr\{X_{t+s} = a | X_t = 1, \mathcal{F}_Y^{t+s-1}\} \Pr\{X_t = 1 | \mathcal{F}_Y^{t+s-1}\} \\
&\quad = \Pr\{X_{t+s} = a | X_t = 0, Y_{t+1}^{t+s-1}\} \Pr\{X_t = 0 | \mathcal{F}_Y^{t+s-1}\} \\
&\qquad + \Pr\{X_{t+s} = a | X_t = 1, Y_{t+1}^{t+s-1}\} \Pr\{X_t = 1 | \mathcal{F}_Y^{t+s-1}\}
\end{aligned}$$

and therefore

$$
\begin{aligned}
&E\big|\Pr\{X_{t+s}=a|X_t=a,\mathcal{F}_Y^{t+s-1}\}-\Pr\{X_{t+s}=a|\mathcal{F}_Y^{t+s-1}\}\big| \\
&\quad\le \max_{b,b'\in\{0,1\}} E\big|\Pr\{X_{t+s}=a|X_t=b,Y_{t+1}^{t+s-1}\} \\
&\qquad\qquad -\Pr\{X_{t+s}=a|X_t=b',Y_{t+1}^{t+s-1}\}\big| \\
&\quad\le \max_{b,b'\in\{0,1\}} E\big|\Pr\{X_s=a|X_0=b,Y_1^{s-1}\}-\Pr\{X_s=a|X_0=b',Y_1^{s-1}\}\big|.
\end{aligned}
\tag{16}
$$

It is well known from the theory of nonlinear filtering (the problem of the dependence of the filter on the initial condition of the process, cf., e.g., [2]) that (16) decays exponentially rapidly in $s$, so that (14) holds. A similar argument leads to the conclusion that (14) holds for the more general case where $\{X_t\}$ is any finite-order, ergodic noncyclic Markov process and the noise, $\{N_t\}$, is i.i.d.

Note that for the noise-free case, namely when $\mathbf{X}=\mathbf{Y}$, the assertion of Lemma 1 is that for any predictor $F$ we have

$$
\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\big[l\big(F_t(X_1^{t-1}),X_t\big)-E\big\{l\big(F_t(X_1^{t-1}),X_t\big)\big|\mathcal{F}_X^{t-1}\big\}\big]=0 \qquad \text{a.s.} \tag{17}
$$

This is because, in the noise-free case, the additional "conditional mixing" condition (14) holds trivially for any source. In the noiseless case, however, this result follows easily by making the observation that $\{l(F_t(X_1^{t-1}),X_t)-E\{l(F_t(X_1^{t-1}),X_t)|\mathcal{F}_X^{t-1}\},\mathcal{F}_X^t\}$ is a (bounded) martingale difference sequence and hence satisfies a law of large numbers (cf., e.g., Theorem 2.19 of [11] or Section III of [1]). Unfortunately, in the noisy setting, this is no longer the case. Generally, the sequence $\{l(F_t(Y_1^{t-1}),X_t)-E\{l(F_t(Y_1^{t-1}),X_t)|\mathcal{F}_Y^{t-1}\}\}$ will not be a martingale difference sequence with respect to any filtration and, therefore, will not necessarily obey the law of large numbers for martingales. Indeed, a simple example will be presented in the sequel where this sequence does not satisfy a law of large numbers even when the joint process $\{(X_t,Y_t)\}$ is stationary and ergodic. As is established in the proof that follows, condition (14) guarantees that this sequence asymptotically behaves as a martingale difference sequence in the sense of satisfying a law of large numbers. The following technical lemma will be useful in the proof of Lemma 1.

LEMMA 2. *Let $\{\delta_t\}_{t\ge1}$ be any sequence of random variables satisfying for all $t: E\delta_t=0$, $|\delta_t|\le C$ a.s., and*

$$
\sum_{s=1}^{\infty}\sup_{t\ge1}|E\delta_t\delta_{t+s}|<\infty. \tag{18}
$$

*Then*

$$
\frac{1}{n}\sum_{t=1}^{n}\delta_t\overset{a.s.}{\to}0. \tag{19}
$$

PROOF. Denoting $\overline{S}_n = \frac{1}{n}\sum_{t=1}^{n}\delta_t$, condition (18) and the bound $|\delta_t| \le C$ imply that $\mathrm{Var}(\overline{S}_n) \le C_1/n$ for some finite $C_1$. Therefore, by Chebyshev's inequality and the Borel–Cantelli lemma, it follows that $\overline{S}_{k^2} \to 0$ almost surely. With $|\delta_t|$ uniformly bounded by $C$, it is easy to check that $|\overline{S}_n| \le |\overline{S}_{k^2}| + 3C/k$ for all $n \in [k^2, (k+1)^2]$, completing the proof. □

PROOF OF LEMMA 1. Since

$$
\begin{aligned}
&l\big(F_t(Y_1^{t-1}), X_t\big) - E\big\{l\big(F_t(Y_1^{t-1}), X_t\big)\big|\mathcal{F}_Y^{t-1}\big\} \\
&\quad = \big[\mathbf{1}_{\{X_t=0\}} - \Pr\{X_t = 0|\mathcal{F}_Y^{t-1}\}\big]l\big(F_t(Y_1^{t-1}), 0\big) \qquad (20)\\
&\qquad + \big[\mathbf{1}_{\{X_t=1\}} - \Pr\{X_t = 1|\mathcal{F}_Y^{t-1}\}\big]l\big(F_t(Y_1^{t-1}), 1\big),
\end{aligned}
$$

it will suffice to establish

$$
\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\delta_t = 0 \qquad \text{a.s.} \tag{21}
$$

and

$$
\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\varepsilon_t = 0 \qquad \text{a.s.,} \tag{22}
$$

where we denote $\delta_t = [\mathbf{1}_{\{X_t=0\}} - \Pr\{X_t = 0|\mathcal{F}_Y^{t-1}\}]l(F_t(Y_1^{t-1}), 0)$ and $\varepsilon_t = [\mathbf{1}_{\{X_t=1\}} - \Pr\{X_t = 1|\mathcal{F}_Y^{t-1}\}] \times l(F_t(Y_1^{t-1}), 1)$. To this end, note that for fixed $t$ and $s$,

$$
\begin{aligned}
&E\delta_t\delta_{t+s} \qquad (23)\\
&\quad = EE\{\delta_t\delta_{t+s}|\mathcal{F}_Y^{t-1}\} \\
&\quad = E\big\{l\big(F_t(Y_1^{t-1}), 0\big)l\big(F_t(Y_1^{t-1}), 1\big) \\
&\qquad \times \big[\Pr\{X_t = 0, X_{t+s} = 0|\mathcal{F}_Y^{t+s-1}\} \qquad (24)\\
&\qquad\quad - \Pr\{X_{t+s} = 0|\mathcal{F}_Y^{t+s-1}\}\Pr\{X_t = 0|\mathcal{F}_Y^{t+s-1}\}\big]\big\}.
\end{aligned}
$$

Therefore

$$
|E\delta_t\delta_{t+s}| \le B^2 E\big|\Pr\{X_{t+s} = 0|X_t = 0, \mathcal{F}_Y^{t+s-1}\} - \Pr\{X_{t+s} = 0|\mathcal{F}_Y^{t+s-1}\}\big|. \tag{25}
$$

Consequently, (14) implies

$$
\sum_{s=1}^{\infty}\sup_{t\ge 1}|E\delta_t\delta_{t+s}| < \infty. \tag{26}
$$

Combining (26) with the immediate facts that $E\delta_t = 0$ and that $|\delta_t| \le B$ a.s. and applying Lemma 2 gives (21). The limit in (22) is obtained similarly. □

Equipped with Lemma 1, we can now present:

THEOREM 2. *For any predictor $F$, and stationary ergodic process $\{(X_t, Y_t)\}_{t\in\mathbb{Z}}$ satisfying* (14), *we have*

$$\liminf_{n\to\infty} \frac{1}{n} L_F(Y_1^n, X_1^n) \geq U(l, \mathbf{P}) \qquad a.s. \tag{27}$$

PROOF. We have almost surely

$$\begin{aligned}
\liminf_{n\to\infty} L_F(Y_1^n, X_1^n) &= \liminf_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} l\big(F_t(Y_1^{t-1}), X_t\big) \\
&= \liminf_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} E\big\{l\big(F_t(Y_1^{t-1}), X_t\big)|\mathcal{F}_Y^{t-1}\big\} \qquad (28) \\
&= \liminf_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \mathbf{P}\{X_t = 0|Y_1^{t-1}\} l\big(F_t(Y_1^{t-1}), 0\big) \\
&\qquad + \mathbf{P}\{X_t = 1|Y_1^{t-1}\} l\big(F_t(Y_1^{t-1}), 1\big) \qquad (29) \\
&= \liminf_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} (1 - \mathbf{P}\{X_t = 1|Y_1^{t-1}\}) l\big(F_t(Y_1^{t-1}), 0\big) \\
&\qquad + \mathbf{P}\{X_t = 1|Y_1^{t-1}\} l\big(F_t(Y_1^{t-1}), 1\big) \\
&\geq \liminf_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} A_l(\mathbf{P}\{X_t = 1|Y_1^{t-1}\}) \qquad (30) \\
&= E\, A_l(\mathbf{P}\{X_0 = 1|Y_{-\infty}^{-1}\}) \qquad (31) \\
&= U(l, P, Q), \qquad (32)
\end{aligned}$$

where (28) follows from Lemma 1, (32) from the definition of $A_l(\cdot)$, and (31) from Breiman's generalized ergodic theorem. □

Theorems 1 and 2 establish $U(l, \mathbf{P})$ as the fundamental prediction performance limitation for the noisy setting. For a process $\mathbf{X}$ one may regard $U(l, \mathbf{P})$ as a generalized notion of entropy. Where the entropy rate of a process measures its compressibility (which, as mentioned above, is equivalent to its noiseless predictability w.r.t. the logarithmic loss function), $U(l, \mathbf{P})$ measures its predictability w.r.t. a general loss function $l$ when the predictor is fed with a noisy version of the past. Hence $U(l, \mathbf{P})$, in some sense, is also a measure of the sensitivity of the clean process of interest to noise. As a concrete example, for

a given process $\mathbf{X}$ let $\mathbf{P}_q$ denote the measure governing the joint process $(\mathbf{X}, \mathbf{Y})$ when $\mathbf{Y}$ is the output of a BSC with crossover probability $q$ fed with $\mathbf{X}$. Looking at $U(l, \mathbf{P}_q)$ as a function of $q$ gives an indication for the sensitivity of the process $\mathbf{X}$ to noise and, in particular, can serve as a basis to discriminate and rank the predictability of different processes which have the same noiseless predictability (cf. [20], Section 5, for a more comprehensive discussion of this point in the context of individual sequences).

The statement in Theorem 2, similarly to its analogue from the noiseless setting, seems quite intuitive and acceptable. It indeed seems natural that the best prediction strategy would be to minimize the expected loss given the available information, and that, when stationarity and ergodicity are assumed, any other prediction strategy would almost surely perform no better. One may initially be tempted to conjecture that, as in the noiseless case, (27) holds true for all stationary ergodic processes. The following simple example shows that this is generally not the case. Let the distribution of $\mathbf{X}$ be given by

$$\mathbf{X} = \begin{cases} \cdots 0\,1\,0\,1\,0 \cdots, & \text{w.p. } \frac{1}{2}, \\ \cdots 1\,0\,1\,0\,1 \cdots, & \text{w.p. } \frac{1}{2}, \end{cases}$$

and let $\mathbf{Y}$ be an i.i.d. Bernoulli$(1/2)$ source independent of $\mathbf{X}$. The process $\mathbf{X}$ is easily verified to be stationary and ergodic and, consequently, so is the joint process $\{(X_t, Y_t)\}$. Denoting the probability measure governing this process by $\mathbf{P}$, we clearly have $U(l, \mathbf{P}) = A_l(1/2)$. On the other hand, the predictor $F$ which, at time $t$, without regard to $Y_1^{t-1}$, gives zero for $t$ odd and one for $t$ even, clearly satisfies $\mathbf{P}\{\liminf_{n\to\infty} L_F(Y_1^n, X_1^n) = 0\} = 1/2$, which contradicts (27) (for any nondegenerate loss function).

The above counterexample demonstrates the insufficiency of stationarity and ergodicity alone, in the noisy setting, for a strong converse result. Theorem 2, on the other hand, assures us of the sufficiency of stationarity and ergodicity when an additional mixing-type assumption (14) is made. The question of the necessity of the latter has yet to be answered.

Note that thus far, $\mathbf{P}$ was assumed an arbitrary stationary (or stationary and ergodic when the almost sure asymptotic regime was considered) probability measure governing the joint process $\{(X_t, Y_t)\}$. A situation of particular interest is that where the measurement $Y_t$ is the output of a fixed (not necessarily memoryless) noisy channel. This can be formally modeled as follows.

DEFINITION 1. Let $\mathbf{X} = \{X_t\}_{t\in\mathbb{Z}}$ and $\mathbf{N} = \{N_t\}_{t\in\mathbb{Z}}$ be two processes taking values in the measurable spaces $\{\{0,1\}^{\mathbb{Z}}, \mathcal{F}_X\}$ and $\{\mathbb{R}^{\mathbb{Z}}, \mathcal{F}_N\}$ and distributed according to the probability measures $P$ and $Q$, respectively. Given a measurable function $g : \{0,1\} \times \mathbb{R} \to \mathbb{R}$, we let for each $t$, $Y_t = g(X_t, N_t)$. Let $\mathbf{P}(P, Q, g)$ denote the probability measure on $\{\{\{0,1\} \times \mathbb{R}\}^{\mathbb{Z}}, \mathcal{F}_{XY}\}$ according to which

$\{(X_t, Y_t)\}_{t\in\mathbb{Z}}$ is distributed when the processes $\mathbf{X}$ and $\mathbf{N}$ are independent. We define

$$U(l, P, Q, g) \stackrel{\text{def}}{=} U\big(l, \mathbf{P}(P, Q, g)\big), \tag{33}$$

where the right-hand side of (33) was defined in (1).

It is conceptually constructive to think of $g$ as the channel and of $\mathbf{N}$ as the channel noise. Admittedly, not all stationary processes $\{(X_t, Y_t)\}_{t\in\mathbb{Z}}$ can be decoupled to comply with the model of Definition 1. In other words, there exist stationary probability measures $\mathbf{P}$ on $\{\{\{0, 1\} \times \mathbb{R}\}^{\mathbb{Z}}, \mathcal{F}_{XY}\}$ for which there exists no tuple $(P, Q, g)$ (where $P$, $Q$ and $g$ are as in Definition 1) such that $\mathbf{P} = \mathbf{P}(P, Q, g)$. One simple example for such a case can be found in the Appendix. This model, however, suffices for representing all conceivable situations of practical interest in which $Y_t$ is a noisy measurement of $X_t$. In addition, the model of Definition 1 facilitates the introduction of the following two notions of universality which is motivated by Theorems 1 and 2.

DEFINITION 2. Let $P$, $Q$, $g$ and $\mathbf{P}(P, Q, g)$ be as in Definition 1. A predictor $F$ will be said to be *universal* with respect to a stationary and ergodic noise source $Q$ and a channel $g$ if

$$\lim_{n\to\infty} \frac{1}{n} L_F(Y_1^n, X_1^n) = U(P, Q, g, l), \qquad \mathbf{P}(P, Q, g)\text{-a.s.} \tag{34}$$

whenever $P$ is a stationary and ergodic probability measure and $\mathbf{P}(P, Q, g)$ is conditionally mixing in the sense of (14). It will be said to be *twofold universal* with respect to a *family* $\mathcal{Q}$ of stationary and ergodic probability measures on $\{\mathbb{R}^{\mathbb{Z}}, \mathcal{F}_N\}$ (noise sources) and a channel $g$ if (34) holds whenever $P$ is a stationary and ergodic probability measure and $Q \in \mathcal{Q}$ is such that $\mathbf{P}(P, Q, g)$ is conditionally mixing in the sense of (14).

Thus, assume, for example, that the clean sequence $\{X_t\}$ is stationary and ergodic, but its distribution is not known to the predictor. Roughly speaking, a universal predictor is one that would be guaranteed to attain the best achievable asymptotic performance no matter what the distribution of $\{X_t\}$ is. A *twofold* universal predictor is one with asymptotic performance guaranteed under channel uncertainty, in addition to the lack of information regarding the distribution of the clean sequence. Note that the probability measure governing the noise process, $Q$, is assumed stationary and ergodic in the above definition. The reason for this is that otherwise $\mathbf{P}(P, Q, g)$ will not be stationary and ergodic (except for degenerate cases). In such a case, $U(l, P, Q, g)$ would lose its significance as it would no longer necessarily be the best almost surely achievable asymptotic loss. The following section is dedicated to establishing the existence of predictors that are universal in the above sense.

**4. Universal predictors for the noisy setting.** One of the main contributions of [19] was the finding that in the noisy setting, for any given finite set of prediction schemes (the reference class), there exists one which efficiently competes with all predictors in the set, uniformly for all underlying clean individual binary sequences. An approach was presented for the construction of such a predictor for a general noise source and observation alphabet.

The general approach that we propose for the stochastic noisy setting is the following: Consider the increasing family of sets of predictors $\{\mathcal{M}_k\}_{k\in\mathbb{N}}$ such that each $\mathcal{M}_k$ contains all $k$th order Bayes predictors of the form $F_t(Y_1^{t-1}) = \Phi_l(\Pr\{X_t = 1|Y_{t-k}^{t-1}\})$, for all possible stationary ergodic distributions of the input sequence $\mathbf{X}$. Then, for every $k$, obtain a predictor $F^k$ which competes with the reference class $\mathcal{M}_k$ for all individual sequences, using the methodology introduced in [19] (cf. Section 2 therein). Finally, combine the $F^k$'s into one predictor which asymptotically competes with $\mathcal{M} = \bigcup_{k=1}^{\infty} \mathcal{M}_k$ for all individual binary sequences and hence, in particular, with the Bayes optimal predictor (which is an accumulation point of $\mathcal{M}$ even when not strictly contained in that set). Since such a predictor (if successfully constructed) attains the asymptotic performance of the Bayes optimal predictor for *each* individual sequence, it clearly competes with the Bayes predictor no matter what distribution is put on the space of binary source sequences. In particular, the source sequence can be distributed according to any stationary ergodic probability measure, in which case (up to verification of the additional conditional mixing condition), competing with the Bayes optimal predictor guarantees that the best achievable performance is attained. To wit, such a predictor is universal.

Generally, in choosing the family $\{\mathcal{M}_k\}_{k\in\mathbb{N}}$, $\mathcal{M}_k$ must be rich enough to contain approximations to all $k$th order Bayes predictors of the form $F_t(Y_1^{t-1}) = \Phi_l(\Pr\{X_t = 0|Y_{t-k}^{t-1}\})$ for all possible stationary ergodic distributions of the input sequence. On the other hand, the classes in this family must be limited enough to allow for the existence of a predictor with redundancy w.r.t. each class which becomes asymptotically negligible. As will be illustrated in the first concrete model to be considered in this section, in the case of finite alphabet observations this is not a problem, and $\mathcal{M}_k$ can simply be taken as the set of *all* $k$th-order Markov predictors, that is, the set of all time-invariant predictors which base their predictions on no more than the last $k$ noisy observations. In the case of continuous-valued observations, however, the choice of $\mathcal{M}_k$ must be made more carefully. In this case, as will be elaborated on below, the set of all $k$th-order Markov predictors is much too rich and a set must be chosen which, though containing approximations to all $k$th-order Bayes predictors (corresponding to all possible distributions of the clean process $\mathbf{X}$), is considerably more limited.

4.1. *The binary symmetric channel.* Consider the setting of Definition 1 for the case where $\mathbf{N} = \{N_t\}_{t\in\mathbb{Z}}$ is a binary noise sequence distributed according to

the probability measure $Q$, the clean binary sequence $\mathbf{X}$ is distributed according to $P$, and the channel function $g$ is given by $Y_t = g(X_t, N_t) = X_t \oplus N_t$, where $\oplus$ denotes XOR (addition modulo 2). This setting models the most general situation of a symmetric binary-input binary-output channel, namely, the situation where the channel can be arbitrarily varying and have arbitrary memory. Let $\mathbf{x} = (x_1, x_2, \ldots) \in \{0,1\}^{\mathbb{N}}$ be an individual binary sequence and $\mathbf{y} = (y_1, y_2, \ldots) \in \{0,1\}^{\mathbb{N}}$ an individual binary observation sequence. We now define the conditional $k$th-order Markov predictability of $x_1^n$ given $y_1^n$ by

$$\lambda_k(x_1^n|y_1^n) \stackrel{\text{def}}{=} \inf_{F\in\mathcal{M}_k} \frac{1}{n} L_F(y_1^n, x_1^n), \tag{35}$$

where $\mathcal{M}_k$ is the set of all $k$th-order Markov predictors. The asymptotic conditional $k$th-order Markov predictability of the infinite sequence $\mathbf{x}$ given $\mathbf{y}$ is defined as

$$\lambda_k(\mathbf{x}|\mathbf{y}) \stackrel{\text{def}}{=} \limsup_{n\to\infty} \lambda_k(x_1^n|y_1^n). \tag{36}$$

Finally, we define the conditional Markov noisy predictability of $\mathbf{x}$ given $\mathbf{y}$ by

$$\lambda(\mathbf{x}|\mathbf{y}) = \lim_{k\to\infty} \lambda_k(\mathbf{x}|\mathbf{y}), \tag{37}$$

where the limit exists for all $\mathbf{x}, \mathbf{y} \in \{0,1\}^{\mathbb{N}}$ as $\lambda_k(\mathbf{x}|\mathbf{y})$ is nonincreasing with $k$. To eliminate future concerns, we note that for any two stochastic binary sequences $\mathbf{X}$ and $\mathbf{Y}$, $\lambda_k(X_1^n|X_1^n)$, and hence also $\lambda_k(\mathbf{X}|\mathbf{Y})$ and $\lambda(\mathbf{X}|\mathbf{Y})$, are well-defined random variables. The only point that must be accounted for is the measurability of $\lambda_k(X_1^n|X_1^n)$, which is defined as an infimum over the set $\mathcal{M}_k$, which is clearly not countable. This is not a real problem, however, as the loss functions are assumed continuous. Therefore, $\mathcal{M}_k$ in the definition of $\lambda_k(X_1^n|X_1^n)$ can be replaced by (or conceived as) the set of all $k$th-order Markov predictors with predictions assuming only rational values, which is a countable set.

THEOREM 3. *Let* $\mathbf{P}(P, Q, g)$ *be stationary, ergodic and satisfy* (14). *We then have*

$$\lambda(\mathbf{X}|\mathbf{Y}) \le U(l, P, Q, g), \qquad \mathbf{P}(P, Q, g)\text{-a.s.} \tag{38}$$

The assertion of Theorem 3 should not be surprising. Letting $F_t^{\text{opt}_K}(Y_1^{t-1}) = \Phi_l(\Pr(X_t = 1|Y_{t-k}^{t-1}))$ denote the "best" Bayesian $k$th-order Markov predictor, clearly $F^{\text{opt}_K} \in \mathcal{M}_k$. Therefore, $\lambda(\mathbf{X}|\mathbf{Y})$ should be no larger than the asymptotic average loss of any $F^{\text{opt}_K}$ for any $k$. However, when stationarity and ergodicity prevail, the losses of the $F^{\text{opt}_K}$ converge, as $k$ approaches infinity, to $U(l, P, Q, g)$. This rationale is made precise in the proof that follows.

PROOF OF THEOREM 3. For each $k$ we have almost surely

$$\lambda_k(\mathbf{X}|\mathbf{Y}) = \limsup_{n\to\infty} \inf_{F\in\mathcal{M}_k} L_F(Y_1^n, X_1^n) \tag{39}$$

$$\leq \limsup_{n\to\infty} L_{F^{\mathrm{opt}_k}}(Y_1^n, X_1^n) \tag{40}$$

$$= \limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} E\{l(F_t^{\mathrm{opt}_k}(Y_1^{t-1}), X_t)|\mathcal{F}_Y^{t-1}\} \tag{41}$$

$$\begin{aligned} &= \limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n}\big[l\big(\Phi_l\big(\Pr(X_t=1|Y_{t-k}^{t-1})\big),0\big)\Pr(X_t=0|Y_1^{t-1}) \\ &\qquad + l\big(\Phi_l\big(\Pr(X_t=1|Y_{t-k}^{t-1})\big),1\big)\Pr(X_t=1|Y_1^{t-1})\big] \end{aligned} \tag{42}$$

$$\begin{aligned} &= E\big[l\big(\Phi_l\big(\Pr(X_0=1|Y_{-k}^{-1})\big),0\big)\Pr(X_0=0|Y_{-\infty}^{-1}) \\ &\qquad + l\big(\Phi_l\big(\Pr(X_0=1|Y_{-k}^{-1})\big),1\big)\Pr(X_0=1|Y_{-\infty}^{-1})\big], \end{aligned} \tag{43}$$

where (40) follows by the fact that, as established in the preceding discussion, $F^{\mathrm{opt}_k} \in \mathcal{M}_k$. Equation (41) follows from Lemma 1. Equation (43) follows from the generalized ergodic theorem and the appropriate version of martingale convergence (cf. [4], Theorem 5.21). Consequently,

$$\lambda(\mathbf{X}|\mathbf{Y}) = \lim_{k\to\infty} \lambda_k(\mathbf{X}|\mathbf{Y})$$

$$\begin{aligned} &\leq \limsup_{k\to\infty} E\big[l\big(\Phi_l\big(\Pr(X_0=1|Y_{-k}^{-1})\big),0\big)\Pr(X_0=0|Y_{-\infty}^{-1}) \\ &\qquad + l\big(\Phi_l\big(\Pr(X_0=1|Y_{-k}^{-1})\big),1\big)\Pr(X_0=1|Y_{-\infty}^{-1})\big] \end{aligned} \tag{44}$$

$$\begin{aligned} &\leq E\Big[\limsup_{k\to\infty} l\big(\Phi_l\big(\Pr(X_0=1|Y_{-k}^{-1})\big),0\big)\Pr(X_0=0|Y_{-\infty}^{-1}) \\ &\qquad + l\big(\Phi_l\big(\Pr(X_0=1|Y_{-k}^{-1})\big),1\big)\Pr(X_0=1|Y_{-\infty}^{-1})\Big] \end{aligned} \tag{45}$$

$$= E A_l\big(\Pr(X_0=1|Y_{-\infty}^{-1})\big) \tag{46}$$

$$= U(l,P), \tag{47}$$

where (45) follows from Fatou's lemma, and (46) follows from combining martingale convergence with the fact that for all $0 \leq p \leq 1$,

$$\lim_{p'\to p}\big[l\big(\Phi_l(p'),0\big)(1-p) + l\big(\Phi_l(p'),1\big)p\big] = A_l(p). \tag{48}$$

To see why (48) holds [note that this is not completely trivial as $\Phi_l(\cdot)$ may not be continuous, e.g., for the absolute loss function], denote for convenience $\gamma(p',p) = l(\Phi_l(p'),0)(1-p) + l(\Phi_l(p'),1)p$. Clearly $|\gamma(p',p) - \gamma(p',p')| \leq$

$2B|p' - p|$, but, by the definition of $\Phi_l(\cdot)$, $\gamma(p', p') = A_l(p')$, which, by the assumed continuity of the loss function, approaches $A_l(p)$ as $p' \to p$. $\square$

The following is an immediate consequence of Theorem 3. We state it explicitly as a corollary for future reference.

COROLLARY 1. *Let $Q$ be stationary and ergodic and $F$ be any predictor such that*

$$L_F(\mathbf{Y}, \mathbf{x}) \le \lambda(\mathbf{x}|\mathbf{Y}), \qquad Q\text{-a.s. } \forall \mathbf{x} \in \{0, 1\}^\infty, \tag{49}$$

*then $F$ is universal with respect to $Q$.*

PROOF. Clearly, (49) implies that for *any* probability measure $P$ on $\{\{0, 1\}^\infty, \mathcal{F}_X\}$,

$$L_F(\mathbf{Y}, \mathbf{X}) \le \lambda(\mathbf{X}|\mathbf{Y}), \qquad \mathbf{P}(P, Q, g)\text{-a.s.} \tag{50}$$

In particular, whenever $\mathbf{P}(P, Q, g)$ is stationary and ergodic and (14) is satisfied, the right-hand side of (50) is upper bounded by $U(l, P, Q, g)$. This, by definition, implies that $F$ is universal w.r.t. $Q$. $\square$

Clearly, the existence of a predictor satisfying (49) implies, by Theorem 2, that (38) holds with equality whenever $\mathbf{P}(P, Q, g)$ is conditionally mixing. Corollary 1 justifies our approach to obtaining universal predictors for this case, which is to seek universal predictors in the sense of (49) (namely efficient prediction schemes from the individual sequence setting). Note that the treatment above did not assume that the noise components were independent (i.e., that $\mathbf{N}$ is an i.i.d. process). In the following sections, we consider universal predictors for the case of a stationary and *memoryless* channel, that is, the case where $\mathbf{N}$ is an i.i.d. Bernoulli($p$) process. This model is traditionally referred to as the binary symmetric channel (BSC). We let $Q_p$ denote the measure corresponding to the case where the channel crossover probability is $p$.

4.1.1. *The absolute loss function and twofold universality.* Consider the absolute loss function given by $l(\alpha, x) = |\alpha - x|$. Two predictors were presented in [7] for this loss function and shown to be universal for the noiseless individual sequence setting: the *increasing order Markov predictor* and the *incremental parsing predictor*. The reader is referred to [7] for a full description of and discussions on these predictors. In the next section, which will be dedicated to the case of a general loss function, we will construct a predictor which is very similar in spirit to the increasing order Markov predictor of [7]. We therefore, at this point, abstain from describing this predictor. In [20], the following was established.

THEOREM 4 (Theorem 7 of [20]). *Let $F$ be any universal predictor from the noiseless case for the $l1$ loss function (e.g., one of the two mentioned above). Then for every $p \in [0, 1/2)$,*

$$L_F(\mathbf{Y}, \mathbf{x}) \leq \lambda(\mathbf{x}|\mathbf{Y}), \qquad Q_p\text{-a.s. } \forall \mathbf{x} \in \{0,1\}^\infty. \tag{51}$$

Actually, to be more precise, the right-hand side of (51) in Theorem 7 of [20] was $\pi(\mathbf{x}|\mathbf{Y})$ rather than $\lambda(\mathbf{x}|\mathbf{Y})$, where the former denotes the *conditional finite-state noisy predictability* defined in [20] and is actually a lower bound for the latter (as is immediate from their definitions).

As extensively discussed in [20], this result is quite surprising because it tells us that not only is the Markov (respectively, finite state) predictability achievable in the noisy individual sequence setting, but that it can in fact be achieved by employing a universal predictor from the noiseless setting, no matter what the parameter governing the channel may be. The following is an immediate consequence of Corollary 1 and Theorem 4.

THEOREM 5. *The increasing order Markov predictor and the incremental parsing predictor are twofold universal with respect to $\{Q_p\}_{p\in[0,1/2)}$.*

In words, the bottom line of Theorem 5 is that by employing, for example, the incremental parsing predictor of [7] on the noisy sequence, one is guaranteed of almost surely achieving the Bayesian envelope, for all stationary, ergodic, and conditionally mixing sources and for all possible BSCs (of crossover probability less than $1/2$).

4.1.2. *General loss functions.* This section is dedicated to the construction of universal predictors for general loss functions using the methodology described at the beginning of the section. As will be evident, these predictors are not twofold universal, as they are constructed using loss estimators which depend on the channel crossover probability $p$. The question of the existence of twofold universal predictors for the BSC and a general loss function remains open.

Given a loss function $l$, we denote for convenience $l_0(\cdot) =_{\text{def}} l(0, \cdot)$ and $l_1(\cdot) =_{\text{def}} l(1, \cdot)$. Suppose now that $l_0$ and $l_1$ are twice differentiable and define

$$S(z) = l_0'(z)l_1''(z) - l_1'(z)l_0''(z)$$

and

$$R(z) = \frac{l_0'(z)l_1'(z)^2 - l_1'(z)l_0'(z)^2}{S(z)},$$

where $l_i'$ and $l_i''$ are the first and second derivatives of $l_i$, respectively, $i = 0, 1$. Now, define

$$c_l = \sup_{0<z<1} R(z).$$

If $S(z) = 0$ for some $0 < z < 1$, we write $c_L = \infty$. For a given $p \in [0, 1/2)$, we define the following auxiliary loss function, $\tilde{l}$, in terms of the original one:

$$\tilde{l}_0(z) \stackrel{\text{def}}{=} \frac{1-p}{1-2p} l_0(z) - \frac{p}{1-2p} l_1(z) + \frac{p}{1-2p} l_1(0)$$

and

$$\tilde{l}_1(z) \stackrel{\text{def}}{=} \frac{-p}{1-2p} l_0(z) + \frac{1-p}{1-2p} l_1(z) + \frac{p}{1-2p} l_0(1)$$

so that

$$\tilde{l}(y, z) = (1-y)\tilde{l}_0(z) + y\tilde{l}_1(z). \tag{52}$$

Let now $S_p(z)$, $R_p(z)$ and $c_l(p)$ be as in equations (A.2)–(A.2) for the loss function $\tilde{l}$, where if $S_p(z) = 0$ for some $0 < z < 1$, we take $c_l(p) = \infty$. We are now ready to present the main result of this section.

THEOREM 6. *Let $l$ be a loss function such that*:

(i) $l_0(0) = l_1(1) = 0$,
(ii) $l_0$ *and* $l_1$ *are three times differentiable in* $(0, 1)$,
(iii) $l_0'(z) > 0$ *and* $l_1'(z) < 0$ *for* $0 < z < 1$.

*Assume further that $c_l(p) < \infty$ and that $S(z)$ is positive for $0 < z < 1$. Then for any $0 \le p < 1/2$, there exists a universal predictor with respect to $Q_p$.*

Two examples for families of loss functions easily verified to satisfy the hypotheses of Theorem 6 (for all $0 \le p < 1/2$) are (cf. [19]):

1. *$\alpha$-powered loss function*: defined by $l(z, x) = |x - z|^\alpha$, for any $1 < \alpha \le 2$.
2. *Hellinger loss function*: defined by $l(z, x) = \frac{1}{2}((\sqrt{1-x} - \sqrt{1-z})^2 + (\sqrt{x} - \sqrt{z})^2)$. More generally, the loss function given by $l(z, 0) = 1 - (1-z)^\alpha$ and $l(z, 1) = 1 - z^\alpha$, for $1/2 \le \alpha < 1$ (note that the Hellinger loss is the case $\alpha = 1/2$).

Let $\mathcal{F}$ be a given finite set of experts and consider the following prediction algorithm due originally to Vovk ([18, 17]; cf. also [12]).

ALGORITHM 1 (Vovk [18]).

**Initialization.** Enumerate the experts arbitrarily as $\mathcal{F} = \{F^{(i)}\}_{i=1}^N$ and set the weights $w_{1,i} = 1$, $1 \le i \le N$.

**Prediction.** Let $v_{t,i} = w_{t,i} / W_t$, where $W_t = \sum_{i=1}^N w_{t,i}$. At the beginning of trial $t$, compute for $x = 0$ and $x = 1$ the value

$$\Delta(x) = -c_l \ln \sum_{i=1}^N v_{t,i} \exp\{-l(x, F_t^{(i)}(x^{t-1}))/c_l\}$$

and predict with any value $P_t$ that satisfies, for $x = 0$ and $x = 1$, the condition

$$l(x, P_t) \le \Delta(x). \tag{53}$$

If no value of $P_t$ meets (53) then the algorithm is said to fail.

**Update.** After observing $x_t$, let

$$w_{t+1,i} = w_{t,i} \exp\{-l\big(x, F_t^{(i)}(x^{t-1})\big)/c_L\}.$$

The following was established in [19], by simply applying known results for Algorithm 1 (cf., e.g., [12]) to the auxiliary loss function of (52).

THEOREM 7 [19]. *Let $l$ be a loss function satisfying the hypotheses of Theorem* 6. *For any class of predictors $\{F^{(i)} : 1 \le i \le N\}$, let $P$ be the strongly sequential predictor given in Algorithm* 1, *with $\tilde{l}$ replacing $l$, $F_t^{(i)}(y^{t-1})$ replacing $F_t^{(i)}(x^{t-1})$ and $c_l(p)$ replacing $c_l$. Then $P$ is guaranteed never to fail and we have for all $n$,*

$$\max_{y_1^n \in \{0,1\}^n} \left\{ \hat{L}_P(y_1^n) - \min_{1 \le i \le N} \hat{L}_{F^{(i)}}(y_1^n) \right\} \le c_L(p) \ln N, \tag{54}$$

*where, for any predictor $F$, $\hat{L}_F(y^n)$ is defined by*

$$\hat{L}_F(y_1^n) \stackrel{\text{def}}{=} \sum_{t=1}^{n} \frac{(1 - y_t) - p}{1 - 2p} l_0\big(F_t(y^{t-1})\big) + \frac{y_t - p}{1 - 2p} l_1\big(F_t(y^{t-1})\big). \tag{55}$$

As was established in [19], Lemma 2, $\hat{L}_F(Y_1^n)$ is an efficient estimator for $L_F(Y_1^n, x_1^n)$ in the sense that, for any predictor $F$ and all $x \in \{0, 1\}^\infty$, we have

$$\limsup_{n \to \infty} \frac{|L_F(Y_1^n, x_1^n) - \hat{L}_F(Y_1^n)|}{\sqrt{n \log \log n}} \le C(p), \qquad Q_p\text{-a.s.}, \tag{56}$$

where $C(p)$ is a deterministic constant depending only on the noise parameter $p$.

Equipped with Theorem 7, we now proceed to construct a predictor which is universal in the sense of satisfying (49) and therefore, by Corollary 1, is universal w.r.t. $Q_p$, that is, for a given channel crossover probability. Note first that in the present setting, where the noisy observations are binary valued, the set $\mathcal{M}_k$ of all $k$th-order Markov predictors can be quite naturally parameterized by the set $[0, 1]^{2^k}$. To see the one-to-one correspondence between the elements of $\mathcal{M}_k$ and those of $[0, 1]^{2^k}$, note that for any point $\theta \in [0, 1]^{2^k}$, we can let $P^\theta$ be the $k$th-order Markov predictor satisfying $P_t^\theta(y^{t-1}) = \theta(y_{t-k}^{t-1})$, $\theta(y_{t-k}^{t-1})$ denoting the $i$th component of $\theta$ $(0 \le i \le 2^k - 1)$, where $i = \sum_{j=1}^{k} 2^{j-1} y_{t-j}$. For $\varepsilon > 0$ we let $\mathcal{M}_k^\varepsilon$ be the finite set of $k$th-order Markov predictors corresponding to the $\varepsilon$-grid in $[0, 1]^{2^k}$, namely, the set of all $\lfloor 1/\varepsilon \rfloor^{2^k}$ points in $[0, 1]^{2^k}$ with components that are integer multiples of $\varepsilon$. We can now state the following theorem.

THEOREM 8. *Let the hypotheses of Theorem* 6 *hold. Let further* $P$ *be the predictor obtained by dividing the observed noisy data* $\mathbf{y} = (y_1^{N_1} y_{N_1+1}^{N_1+N_2} \cdots y_{N_1+\cdots+N_{k-1}+1}^{N_1+\cdots+N_k} \cdots)$ *into nonoverlapping blocks and applying the predictor of Theorem* 7 *to each block where, for the kth block (of length* $N_k$*), the expert class is* $\mathcal{M}_k^{\varepsilon_k}$*. Taking* $N_k = 2^{2^k}$ *and* $\varepsilon_k = 1/N_k$*, we have for all* $k$ *and* $y \in \{0,1\}^{\mathbb{N}}$,

$$\hat{L}_P(y_1^n) - \inf_{F \in \mathcal{M}_k} \hat{L}_F(y_1^n) \le O\big(\sqrt{n}\big). \tag{57}$$

PROOF. Suppose first that $n = \sum_{j=1}^{k} N_j + t$, where $1 \le t \le N_{k+1}$. We then have the following chain of inequalities (justified below):

$$\begin{aligned}
&\hat{L}_P(y_1^n) - \inf_{F \in \mathcal{M}_k} \hat{L}_F(y_1^n) \\
&\quad \le \hat{L}_P\big(y_1^{N_1+\cdots+N_{k-1}}\big) - \inf_{F \in \mathcal{M}_k} \hat{L}_F\big(y_1^{N_1+\cdots+N_{k-1}}\big) && (58)\\
&\qquad + \hat{L}_P\big(y_{N_1+\cdots+N_{k-1}+1}^{N_1+\cdots+N_k}\big) - \inf_{F \in \mathcal{M}_k} \hat{L}_F\big(y_{N_1+\cdots+N_{k-1}+1}^{N_1+\cdots+N_k}\big) && (59)\\
&\qquad + \hat{L}_P\big(y_{N_1+\cdots+N_k+1}^{N_1+\cdots+N_k+t}\big) - \inf_{F \in \mathcal{M}_k} \hat{L}_F\big(y_{N_1+\cdots+N_k+1}^{N_1+\cdots+N_k+t}\big) && (60)\\
&\quad \le c_1 \sum_{j=1}^{k-1} N_j && (61)\\
&\qquad + c_l(p) 2^k \ln \frac{1}{\varepsilon_k} + c_2 N_k \varepsilon_k && (62)\\
&\qquad + c_l(p) 2^{k+1} \ln \frac{1}{\varepsilon_{k+1}} + c_2 t \varepsilon_{k+1} && (63)\\
&\quad \le c_1 N_{k-1}\big(1 + o(k)\big) + c_l(p) 2^k \ln N_k + c_2 \\
&\qquad + c_l(p) 2^{k+1} \ln N_{k+1} + c_2 && (64)\\
&\quad = c_1 \sqrt{N_k}\big(1 + o(k)\big) \\
&\qquad + c_l(p)\big[(\ln N_k)^2 + (\ln N_{k+1})^2\big] + 2c_2 \\
&\quad = O\big(\sqrt{N_k}\big) \\
&\quad \le O\big(\sqrt{n}\big). && (65)
\end{aligned}$$

The term in (61) bounds that in (58) for some constant $c_1$ as it clearly follows from the boundedness of the loss function and the definition of $\hat{L}_F(\cdot)$ that, for any $F, y, a, l : |\hat{L}_F(y_{a+1}^{a+l})| \le c_1 l$. To see that (62) bounds (59) and, similarly, that

(63) bounds (60), we write

$$\hat{L}_P\big(y_{N_1+\cdots+N_{k-1}+1}^{N_1+\cdots+N_k}\big) - \inf_{F\in\mathcal{M}_k} \hat{L}_F\big(y_{N_1+\cdots+N_{k-1}+1}^{N_1+\cdots+N_k}\big)$$

$$\le \hat{L}_P\big(y_{N_1+\cdots+N_{k-1}+1}^{N_1+\cdots+N_k}\big) - \inf_{F\in\mathcal{M}_k^{\varepsilon_k}} \hat{L}_F\big(y_{N_1+\cdots+N_{k-1}+1}^{N_1+\cdots+N_k}\big) + c_2 N_k \varepsilon_k \tag{66}$$

$$\le c_l(p) \ln |\mathcal{M}_k^{\varepsilon_k}| + c_2 N_k \varepsilon_k \tag{67}$$

$$= c_l(p) \ln \lfloor 1/\varepsilon_k \rfloor^{2^k} + c_2 N_k \varepsilon_k$$

$$\le c_l(p) 2^k \ln \frac{1}{\varepsilon_k} + c_2 N_k \varepsilon_k, \tag{68}$$

where inequality (67) follows from Theorem 7. Inequality (66) follows from the fact that for some constant $c_2$ and all $\varepsilon > 0$ we have

$$\inf_{F\in\mathcal{M}_k^{\varepsilon}} \hat{L}_F\big(y_{N_1+\cdots+N_{k-1}+1}^{N_1+\cdots+N_k}\big) \le \inf_{F\in\mathcal{M}_k} \hat{L}_F\big(y_{N_1+\cdots+N_{k-1}+1}^{N_1+\cdots+N_k}\big) + c_2 N_k \varepsilon. \tag{69}$$

This follows from the facts that for any predictor in $\mathcal{M}_k$ there exists one in $\mathcal{M}_k^{\varepsilon}$ which $\varepsilon$-approximates all its $N_k$ predictions on the $k$th block and that $L_0(\cdot)$ and $L_1(\cdot)$ are Lipschitz [as $L_0'(\cdot)$ and $L_1'(\cdot)$ are bounded]. To conclude, note that for any $n' > n$, there exists $k' \ge k$ such that $n' = \sum_{j=1}^{k'} N_j + t$, where $1 \le t \le N_{k'+1}$. Since $\mathcal{M}_k \subset \mathcal{M}_{k'}$, (65) gives us

$$\hat{L}_P\big(y_1^{n'}\big) - \inf_{F\in\mathcal{M}_k} \hat{L}_F\big(y_1^{n'}\big) \le \hat{L}_P\big(y_1^{n'}\big) - \inf_{F\in\mathcal{M}_{k'}} \hat{L}_F\big(y_1^{n'}\big) \tag{70}$$

$$\le O\big(\sqrt{n'}\big). \tag{71}$$

□

We can now prove Theorem 6 by establishing the universality of the predictor of Theorem 8.

PROOF OF THEOREM 6. By Corollary 1 it will suffice to establish

$$L_P(\mathbf{Y}, \mathbf{x}) \le \lambda(\mathbf{x}|\mathbf{Y}), \qquad Q_p\text{-a.s. } \forall\, \mathbf{x} \in \{0,1\}^\infty, \tag{72}$$

where $P$ is the predictor of Theorem 8. To this end, fix $\varepsilon > 0$, $k \in \mathbb{N}$, $x \in \{0,1\}^{\mathbb{N}}$. We have $Q_p$-almost surely

$$L_P(\mathbf{Y}, \mathbf{x}) - \lambda_k(\mathbf{x}|\mathbf{Y}) \tag{73}$$

$$= \limsup_{n\to\infty} \frac{1}{n} L_P(Y_1^n, x_1^n) - \limsup_{n\to\infty} \lambda_k(x_1^n|Y_1^n) \tag{74}$$

$$\le \limsup_{n\to\infty} \left[\frac{1}{n} L_P(Y_1^n, x_1^n) - \lambda_k(x_1^n|Y_1^n)\right] \tag{75}$$

$$= \limsup_{n\to\infty} \left[\frac{1}{n} L_P(Y_1^n, x_1^n) - \frac{1}{n}\hat{L}_P(Y_1^n) + \frac{1}{n}\hat{L}_P(Y_1^n)\right. \tag{76}$$

$$- \inf_{F\in\mathcal{M}_k^\varepsilon} \frac{1}{n}\hat{L}_F(Y_1^n) + \inf_{F\in\mathcal{M}_k^\varepsilon} \frac{1}{n}\hat{L}_F(Y_1^n) \tag{77}$$
$$- \inf_{F\in\mathcal{M}_k} \frac{1}{n}L_F(Y_1^n, x_1^n)\Big]$$

$$\le \limsup_{n\to\infty}\Big[\frac{1}{n}L_P(Y_1^n, x_1^n) - \frac{1}{n}\hat{L}_P(Y_1^n) + \frac{1}{n}\hat{L}_P(Y_1^n) \tag{78}$$
$$- \inf_{F\in\mathcal{M}_k} \frac{1}{n}\hat{L}_F(Y_1^n) + \min_{F\in\mathcal{M}_k^\varepsilon} \frac{1}{n}\hat{L}_F(Y_1^n) \tag{79}$$
$$- \min_{F\in\mathcal{M}_k^\varepsilon} \frac{1}{n}L_F(Y_1^n, x_1^n) + c\varepsilon\Big]$$

$$\le \limsup_{n\to\infty}\Big[\frac{1}{n}\big|L_P(Y_1^n, x_1^n) - \hat{L}_P(Y_1^n)\big| + \frac{1}{n}\Big(\hat{L}_P(Y_1^n) - \inf_{F\in\mathcal{M}_k} \hat{L}_F(Y_1^n)\Big) \tag{80}$$
$$+ \max_{F\in\mathcal{M}_k^\varepsilon} \frac{1}{n}\big|\hat{L}_F(Y_1^n) - L_F(Y_1^n, x_1^n)\big| + c\varepsilon\Big] \tag{81}$$

$$\le \limsup_{n\to\infty} \frac{1}{n}\big|L_P(Y_1^n, x_1^n) - \hat{L}_P(Y_1^n)\big| \tag{82}$$

$$+ \limsup_{n\to\infty} \frac{1}{n}\Big(\hat{L}_P(Y_1^n) - \inf_{F\in\mathcal{M}_k} \hat{L}_F(Y_1^n)\Big) \tag{83}$$

$$+ \limsup_{n\to\infty} \max_{F\in\mathcal{M}_k^\varepsilon} \frac{1}{n}\big|\hat{L}_F(Y_1^n) - L_F(Y_1^n, x_1^n)\big| + c\varepsilon \tag{84}$$

$$\le 0 + 0 + 0 + c\varepsilon, \tag{85}$$

where (79) bounds (77) as it follows from the definition of $\mathcal{M}_k^\varepsilon$ that $\inf_{F\in\mathcal{M}_k^\varepsilon} \frac{1}{n} \times \hat{L}_F(Y_1^n, x_1^n) \ge \inf_{F\in\mathcal{M}_k} \frac{1}{n}\hat{L}_F(Y_1^n, x_1^n)$ and that $\inf_{F\in\mathcal{M}_k} \frac{1}{n}L_F(Y_1^n, x_1^n) \ge \inf_{F\in\mathcal{M}_k^\varepsilon} \frac{1}{n} \times L_F(Y_1^n, x_1^n) - c\varepsilon$ for some constant $c$ which depends only on the loss function (as the Lipschitz condition is satisfied by the loss function which has bounded derivatives) and is independent of $\varepsilon$. The limit suprema in (82) and (84) are zero from (56) (note that $\mathcal{M}_k^\varepsilon$ is a finite set). The limit supremum in (83) is upper bounded by zero by Theorem 8. The remaining transitions are self evident. So we have established

$$L_P(\mathbf{Y}, \mathbf{x}) \le \lambda_k(\mathbf{x}|\mathbf{Y}) + c\varepsilon, \tag{86}$$

which, by the arbitrariness of $k$, $\varepsilon$ and $\mathbf{x}$, implies (72). □

Note that given the strength of (54) and (56), it is easy to see that in the above proof [cf., in particular, (85)] we have actually shown that for all $\delta > 0$,

$$\limsup_{n\to\infty} \frac{n(\frac{1}{n}L_P(Y_1^n, x_1^n) - \lambda_k(\mathbf{x}|\mathbf{Y}) - \delta)}{\sqrt{\log\log n}} \le C(p), \qquad Q_p\text{-a.s. } \forall\, \mathbf{x} \in \{0,1\}^\infty,$$

which implies

$$\limsup_{n\to\infty} \frac{n(\frac{1}{n}L_P(Y_1^n, x_1^n) - \lambda(\mathbf{x}|\mathbf{Y}) - \delta)}{\sqrt{\log\log n}} \le C(p), \qquad Q_p\text{-a.s. } \forall\, \mathbf{x} \in \{0,1\}^\infty$$

and consequently, by Theorem 3, for all $0 \le p < 1/2$ and $P$ such that $\mathbf{P}(P, Q_p)$ satisfies (14),

$$\limsup_{n\to\infty} \frac{n(\frac{1}{n}L_P(Y_1^n, X_1^n) - U(l, P, Q_p) - \delta)}{\sqrt{\log\log n}} \le C(p), \qquad \mathbf{P}(P, Q_p)\text{-a.s.}$$

4.2. *Continuous-valued observations.* Throughout this section, we assume a fixed loss function $l : [0,1] \times \{0,1\} \to [0,\infty]$. The output alphabet is now the real line. We assume a memoryless noisy channel with a fixed noise distribution characterized by the density function $f(\cdot|\cdot)$, which is really two density functions (w.r.t. Lebesgue measure): $f(y_t|x_t = 0)$ and $f(y_t|x_t = 1)$. The only assumption on the channel which we will need for the main result of this section is the trivially mildest possible one, namely, that $f(\cdot|x_t = 0)$ and $f(\cdot|x_t = 1)$ or, more precisely, the corresponding cumulative distribution functions, are not identical. Any channel satisfying this requirement will henceforth be referred to as *distinguishable*. Note that this setting complies with the model of Definition 1. To see this concretely, let, for example, $Q$ be the probability measure making the process $\mathbf{N}$ of Definition 1 an i.i.d. process with $N_1$ uniformly distributed on $[0,1]$. Let the function $g$ be given by $g(i, N_t) = F_i^{-1}(N_t)$ for $i \in \{0,1\}$, where $F_i^{-1} : [0,1] \to \mathbb{R}$, the (possibly pseudo-) inverse of the distribution function $F_i(\cdot) = \int_{-\infty}^{\cdot} f(y|x_t = i)\, dy$, is given by $F_i^{-1}(\beta) =_{\text{def}} \inf\{\alpha : \int_{-\infty}^{\alpha} f(y|x_t = i)\, dy \ge \beta\}$. It is then easy to see that for the above described choices of $Q$ and $g$, $\mathbf{P}(P, Q, g)$ is the measure according to which the joint process $\{(X_t, Y_t)\}_{t\in\mathbb{Z}}$ is distributed. Since in this setting, the measure according to which $\{(X_t, Y_t)\}_{t\in\mathbb{Z}}$ is distributed is fully determined by the $\mathbf{X}$-marginal $P$ and by the conditional density $f(\cdot|\cdot)$, we henceforth slightly alter the notation introduced in Definition 1 and let $\mathbf{P}(P, f)$ replace $\mathbf{P}(P, Q, g)$ and $U(l, P, f)$ replace $U(l, P, Q, g)$, where, for a given chanel $f$, $Q$ and $g$ are constructed as specified above. The terminology of Definition 2 will be altered analogously, so that we will say that a predictor $F$ is universal w.r.t. the channel $f$ rather than w.r.t. the coresponding $Q$ and $g$.

To present our main result for this section, we first define, for any $B > 0$, the following auxiliary "loss function" $l_B : [0,1] \times [0,1] \to [0,\infty]$:

$$\begin{aligned} l_B(u, z) = {} & \{1 - [(2B+1)u - B]\} \cdot l_0(z) \\ & + [(2B+1)u - B] \cdot l_1(z) + B l_1(0). \end{aligned} \tag{87}$$

Further, let $S_B(z)$, $R_B(z)$ and $c_l(B)$ be defined as in equations (A.2), (A.2) and (A.2), respectively, with $l_{B0}$ and $l_{B1}$ replacing $l_0$ and $l_1$ where $l_{B0}(\cdot) = l_B(0, \cdot)$ and $l_{B1}(\cdot) = l_B(1, \cdot)$. We will also need the following notion of regularity.

DEFINITION 3. Let $\mathbf{X} = \{X_t\}_{t\in\mathbb{Z}}$ be a stationary binary sequence and define

$$\alpha_{\mathbf{X}}(k) \stackrel{\text{def}}{=} \min_{\{v\in\{0,1\}^k:\, \Pr(X_1^k=v)>0\}} \Pr(X_1^k = v).$$

The process $\mathbf{X}$ (or the probability measure governing it) will be said to be *exponentially regular* if $\{\alpha_{\mathbf{X}}(k)\}_{k\geq 1}$ decays exponentially fast at the most, that is,

$$\liminf_{k\to\infty} \frac{1}{k} \log \alpha(k) > -\infty. \tag{88}$$

Note that Markov processes of all finite orders, processes which are mixing in any reasonable sense and almost any stationary processes that one can think of are exponentially regular. One may even be tempted to conjecture that all stationary binary sequences are exponentially regular. To see that this is not the case, let $\mathbf{x}(k) \in \{0,1\}^{\mathbb{Z}}$ be a periodic deterministic binary sequence of period $k$ obtained from the infinite concatenation of the block of length $k$ consisting of $k-1$ 1's followed by a 0. Let now $\mathbf{X}(k)$ be the stationary sequence obtained by randomly shifting $\mathbf{x}(k)$ anywhere between 1 and $k$ steps (equiprobably, say, to the right). Denote the measure corresponding to $\mathbf{X}(k)$ by $\mathbf{P}_k$. For a given sequence of nonnegative weights summing to unity, $\{w_k\}_{k\geq 1}$, we now let $\mathbf{P} = \sum_{k=1}^{\infty} w_k \mathbf{P}_k$. $\mathbf{P}$ is clearly stationary and we have

$$\begin{aligned} \alpha_{\mathbf{P}}(k) &\stackrel{\text{def}}{=} \min_{\{v\in\{0,1\}^k:\, \mathbf{P}(X_1^k=v)>0\}} \mathbf{P}(X_1^k = v) \\ &\leq \mathbf{P}(X_1^k = 1\,1\cdots 1) \qquad (89) \\ &\leq \sum_{j=k+1}^{\infty} w_k. \end{aligned}$$

Clearly, the weights $\{w_k\}_{k\geq 1}$ can be chosen such that the sum in (89) decays superexponentially with $k$ (e.g., take $w_k = C2^{-2^k}$ where $C$ is the normalizing constant), thus yielding a stationary process which is not exponentially regular. We merely remark here that, while the $\mathbf{P}$ constructed in this example is clearly not ergodic, one can similarly (although the details are somewhat more involved) construct a stationary *and ergodic* $\mathbf{P}$ for which $\alpha_{\mathbf{P}}(k)$ decays superexponentially as well. This notion of regularity will play a role in the proof of the main result of this section which we are now ready to present.

THEOREM 9. *Let $l$ be such that $l_0$ and $l_1$ are three times differentiable in $(0,1)$. Assume further that for all $B > 0$ $c_l(B)$ is finite and $S_B(z)$ is positive for $0 < z < 1$. Finally, assume that for all $u, a, b \in [0,1]$, the function $g_B$, defined by $g_B(u,a,b) = (l_B(u,a) - l_B(u,b))/c_l(B)$, satisfies*

$$\frac{\partial^2 g_B(u,a,b)}{\partial u^2} + \left(\frac{\partial g_B(u,a,b)}{\partial u}\right)^2 \geq 0. \tag{90}$$

*Then for any distinguishable channel $f$ there exists a predictor $P$ which is universal w.r.t. $f$ in the sense that*

$$\lim_{n\to\infty}\frac{1}{n}L_P(Y_1^n, X_1^n) = U(l, P_{\mathrm{er}}, f), \qquad \mathbf{P}(P_{\mathrm{er}}, f)\text{-a.s.} \tag{91}$$

*for all $P_{\mathrm{er}}$'s that are exponentially regular and such that $\mathbf{P}(P_{\mathrm{er}}, f)$ satisfies* (14).

Two examples of families of loss functions shown in [19] to comply with the hypotheses of Theorem 9 are presented after Theorem 6.

The proof of Theorem 9, to which the remainder of this section is dedicated, establishes the existence of a universal predictor for each $f$ by means of an explicit construction. The significance of the exponential regularity assumption will be made clear and elaborated on in the sequel, as it will naturally arise at one of the stages of the proof.

Let us define $\lambda_k^r(x_1^n|y_1^n)$, $\lambda_k^r(\mathbf{x}|\mathbf{y})$ and $\lambda^r(\mathbf{x}|\mathbf{y})$ similarly to the definitions for $\lambda_k(x_1^n|y_1^n)$, $\lambda_k(\mathbf{x}|\mathbf{y})$ and $\lambda(\mathbf{x}|\mathbf{y})$ of (35)–(37), respectively, with the crucial difference that now $\mathcal{M}_k$, rather than denoting the set of *all* $k$th order Markov experts, denotes the set of all $k$th order Markov experts of the form

$$F_t(Y_1^{t-1}) = \Phi_l\left(\frac{\sum_{x_{t-k}^{t-1}}\alpha_{x_{t-k}^{t-1}}\beta_{x_{t-k}^{t-1}}\prod_{i=t-k}^{t-1} f(Y_i|x_i)}{\sum_{x_{t-k}^{t-1}}\beta_{x_{t-k}^{t-1}}\prod_{i=t-k}^{t-1} f(Y_i|x_i)}\right), \tag{92}$$

where $\alpha_{x_{t-k}^{t-1}} \in [0,1]$ for all $x_{t-k}^{t-1} \in \{0,1\}^k$ and $\{\beta_{x_{t-k}^{t-1}}\}_{x_{t-k}^{t-1}\in\{0,1\}^k}$ belongs to the simplex $\beta_{x_{t-k}^{t-1}} \geq 0$ for all $x_{t-k}^{t-1} \in \{0,1\}^k$ and $\sum_{x_{t-k}^{t-1}\in\{0,1\}^k}\beta_{x_{t-k}^{t-1}} = 1$. Note, in particular, the strong dependence of the sets $\mathcal{M}_k$ on the channel $f$. To motivate this choice of $\mathcal{M}_k$, note that the "best" Bayesian $k$th order predictor for this setting would be

$$F_t^{\mathrm{opt}_k}(Y_1^{t-1}) = \Phi_l\big(\Pr(X_t = 1|Y_{t-k}^{t-1})\big) \tag{93}$$

$$= \Phi_l\left(\frac{\sum_{x_{t-k}^{t-1}}\Pr(X_t = 1|x_{t-k}^{t-1})\Pr(x_{t-k}^{t-1})\prod_{i=t-k}^{t-1} f(Y_i|x_i)}{\sum_{x_{t-k}^{t-1}}\Pr(x_{t-k}^{t-1})\prod_{i=t-k}^{t-1} f(Y_i|x_i)}\right), \tag{94}$$

so that clearly $F^{\mathrm{opt}_k} \in \mathcal{M}_k$.

THEOREM 10. *Let $P$ be stationary and ergodic and $f$ be such that $\mathbf{P}(P, f)$ satisfies* (14). *Then*

$$\lambda^r(\mathbf{X}|\mathbf{Y}) \leq U(l, P, f), \qquad \mathbf{P}(P, f)\text{-a.s.} \tag{95}$$

Since, as will be established below, $\lambda^r(\mathbf{X}|\mathbf{Y})$ is achievable by a (universal) predictor, the inequality in (95) actually holds with equality (by Theorem 2). The proof of Theorem 10 is similar to that of Theorem 3 and is therefore omitted. Its implication is that, similarly as was done in the previous section where the BSC

was considered, universality can be attained by achieving $\lambda^r(\mathbf{X}|\mathbf{Y})$, which, in turn, can be done if one finds a way for efficiently competing with the sets $\mathcal{M}_k$ for all $k$'s.

The main reason why in this setting, where the observations take values in a continuum, it is imperative to modify the definition of $\mathcal{M}_k$ from the set of all $k$th order Markov predictors to the set of $k$th order Markov predictors of the form (92) is the following. The richness of the set of all $k$th order Markov predictors for this case renders the originally defined $\lambda_k(x_1^n|y_1^n)$, $\lambda_k(\mathbf{x}|\mathbf{y})$ and $\lambda_k(\mathbf{x}|\mathbf{y})$ meaningless for this case. In fact, when, for example, the conditional distributions characterizing the channel $f(\cdot|\cdot)$ are absolutely continuous with respect to, say, Lebesgue measure (i.e., when the noisy observations are purely continuous-valued random variables), the originally defined $\lambda_k(X_1^n|Y_1^n)$, $\lambda_k(\mathbf{X}|\mathbf{Y})$ and $\lambda_k(\mathbf{X}|\mathbf{Y})$ are almost surely all zero. To see this, note that in this case, for any $n$, the probability that $Y_1^n$ will have two identical components is zero. Therefore, there almost surely exists a (continuous) function $h:\mathbb{R}\to[0,1]$ such that $h(Y_i)=X_i$ for all $1\le i\le n$. Consequently, this implies that the originally defined $\lambda_1(X_1^n|Y_1^n)$ equals zero almost surely for all $n$ which, in turn [by the monotonicity of $\lambda_k(X_1^n|Y_1^n)$ in $k$], implies the same for $\lambda_k(X_1^n|Y_1^n)$, $\lambda_k(\mathbf{X}|\mathbf{Y})$ and $\lambda(\mathbf{X}|\mathbf{Y})$. Hence, the ambition to compete with all Markov predictors of all orders is clearly excessively optimistic (in fact, successful competition is impossible as it would contradict Theorem 2). The considerably more modest ambition to compete with the modified classes $\mathcal{M}_k$, however, as we plan to show next, is indeed realistic, and will suffice for universality by Theorem 10.

As discussed in [19], the hypotheses of Theorem 11 are satisfied by most loss functions of interest, with the exception of the absolute loss function. It will be clear from the proof that follows, however, that for the absolute loss function, a universal predictor can be similarly constructed by replacing the basic predictors used as building blocks for the universal predictor of Theorem 9 with standard exponential weighting predictors (cf., e.g., [19]).

To set the scene for the proof of Theorem 9, we define, analogously to the cumulative loss estimator introduced in Section 4.1, for any predictor $F$,

$$\hat{L}_F(y_1^n) \stackrel{\text{def}}{=} \sum_{t=1}^{n}(1-y_t)l_0\big(F_t(y_1^{t-1})\big)+y_t l_1\big(F_t(y_1^{t-1})\big). \tag{96}$$

As will be argued in the proof below, $\hat{L}_F(Y_1^n)$, under certain assumptions, is an efficient estimator for $L_F(Y_1^n,x_1^n)$ (in a sense to be made precise). The predictors which will be used as building blocks in the construction of our universal predictor are those that were proven effective in [19] for the individual sequence setting. For convenience, prior to the proof of Theorem 9, we recall the relevant results from [19]. For a fixed $B>0$, consider the predictor given in the following algorithm (which is guaranteed never to fail as long as $y_t\in[-B,B+1]$ for all $1\le t\le n$), for the case where $\mathcal{F}$ is an arbitrary finite class of predictors.

ALGORITHM 2.

**Initialization.** Enumerate the $N$ predictors in $\mathcal{F}$ and set their weights $w_{1,i} = 1$, $1 \le i \le N$.

**Prediction.** Let $v_{t,i} = w_{t,i}/W_t$, where $W_t = \sum_{i=1}^{N} w_{t,i}$. At the beginning of trial $t$, compute for $u = 0$ and $u = 1$ the value

$$\Delta_B(u) = -c_l(B) \ln \sum_{i=1}^{N} v_{t,i} \exp\{-l_B(u, F_t^{(i)}(y^{t-1}))/c_l(B)\}$$

and predict with any value $P_t$ that satisfies for $u = 0$ and $u = 1$ the condition

$$l_B(u, P_t) \le \Delta_B(u).$$

If no such $P_t$ exists, the algorithm is said to fail.

**Update.** After observing the $t$th outcome $y_t$, let

$$w_{t+1,i} = w_{t,i} \exp\left\{-l_B\left(\frac{y_t + B}{2B+1}, F_t^{(i)}(y^{t-1})\right)\Big/ c_l(B)\right\}.$$

THEOREM 11 ([19], Theorem 28). *Let $l : \{0,1\} \times [0,1] \to [0,\infty]$ be a loss function such that $l_1(0) = l_0(1) < \infty$, $l_0$ and $l_1$ are three times differentiable in $(0,1)$, and $l_0'(z) > 0$, $l_1'(z) < 0$ for $0 < z < 1$. Assume that $c_l(B)$ is finite and $S_B(z)$ is positive for $0 < z < 1$. Assume further that for all $u, a, b \in [0,1]$, the function $g_B$, defined by $g_B(u,a,b) = (l_B(u,a) - l_B(u,b))/c_l(B)$, satisfies*

$$\frac{\partial^2 g_B(u,a,b)}{\partial u^2} + \left(\frac{\partial g_B(u,a,b)}{\partial u}\right)^2 \ge 0. \tag{97}$$

*Then the predictor $P$ of Algorithm 2 is guaranteed never to fail for all $n$ and $y_1^n \in [-B,B]^n$ and we have*

$$\forall n \ge 1 : \max_{y^n \in [-B,B]^n} \left\{\hat{L}_P(y^n) - \min_{F \in \mathcal{F}} \hat{L}_F(y^n)\right\} \le c_l(B) \ln |\mathcal{F}|. \tag{98}$$

To construct our predictor, we redefine now $\mathcal{M}_k^\varepsilon$, to be the notion of an $\varepsilon$-grid of $\mathcal{M}_k$ suitable for the setting of the present section. Namely, we let $\mathcal{M}_k^\varepsilon$ be the subset of $\mathcal{M}_k$ such that any $F \in \mathcal{M}_k^\varepsilon$, which clearly has the form

$$F_t(Y_1^{t-1}) = \Phi_l\left(\frac{\sum_{x_{t-k}^{t-1}} \alpha_{x_{t-k}^{t-1}} \beta_{x_{t-k}^{t-1}} \prod_{i=t-k}^{t-1} f(Y_i|x_i)}{\sum_{x_{t-k}^{t-1}} \beta_{x_{t-k}^{t-1}} \prod_{i=t-k}^{t-1} f(Y_i|x_i)}\right), \tag{99}$$

is such that $\alpha_{x_{t-k}^{t-1}}$ and $\beta_{x_{t-k}^{t-1}}$ are integer multiples of $\varepsilon$ for all $x_{t-k}^{t-1} \in \{0,1\}^k$. We clearly have $|\mathcal{M}_k^\varepsilon| < \lceil 1/\varepsilon \rceil^{2^{k+1}}$. The set $\mathcal{M}_k^\varepsilon$ is shown in the Appendix to be an $\varepsilon$-cover of $\mathcal{M}_k$ in the following sense.

LEMMA 3. *For any predictor $F \in \mathcal{M}_k$ of the form $F_t(y_1^{t-1}) = \Phi_l(e(y_{t-k}^{t-1}))$, where*

$$e(y_{t-k}^{t-1}) = \frac{\sum_{x_{t-k}^{t-1}} \alpha_{x_{t-k}^{t-1}} \beta_{x_{t-k}^{t-1}} \prod_{i=t-k}^{t-1} f(Y_i|x_i)}{\sum_{x_{t-k}^{t-1}} \beta_{x_{t-k}^{t-1}} \prod_{i=t-k}^{t-1} f(Y_i|x_i)}, \tag{100}$$

*there exists a predictor $G \in \mathcal{M}_k^{\varepsilon}$ of the form $G_t(y_1^{t-1}) = \Phi_l(g(y_{t-k}^{t-1}))$ such that*

$$\left|e(y_{t-k}^{t-1}) - g(y_{t-k}^{t-1})\right| \le \varepsilon\left(\frac{2}{\beta - \varepsilon} + 1\right), \tag{101}$$

*for all $\varepsilon < \beta$, where $\beta = \min\{\beta_{x_{t-k}^{t-1}} : \beta_{x_{t-k}^{t-1}} > 0\}$.*

Observe, in particular, that Lemma 3 and the boundedness of $l_0'$ and $l_1'$ (which, in turn, implies that $l_0$ and $l_1$ are Lipschitz) assure us of the fact that for any $k$, $\varepsilon < \beta_k =_{\text{def}} \min_{\{v\in\{0,1\}^k : \Pr(X_1^k = v) > 0\}} \Pr(X_1^k = v)$, and binary-valued process $\mathbf{X}$ we have

$$\min_{F\in\mathcal{M}_k^{\varepsilon}} \hat{L}_F(y_1^t) \le \hat{L}_{F^{\text{opt}_k}}(y_1^t) + Ct\varepsilon\left(\frac{2}{\beta_k - \varepsilon} + 1\right), \tag{102}$$

where $F^{\text{opt}_k}$ is as defined in (93) for the process $\mathbf{X}$, and the constant $C$ (which depends on the loss function) is independent of $\varepsilon$, $t$, $k$ and the process $\mathbf{X}$. Equipped with Theorem 11 and Lemma 3, we now procede to construct a predictor in a manner similar to that by which the predictor in Theorem 8 was constructed for the case of the BSC.

THEOREM 12. *Let the hypotheses of Theorem* 9 *hold. Let further $P$ be the predictor obtained by dividing the data $\mathbf{y} = (y_1^{N_1} y_{N_1+1}^{N_1+N_2} \cdots y_{N_1+\cdots+N_{k-1}+1}^{N_1+\cdots+N_k} \cdots)$ into nonoverlapping blocks and applying the predictor of Theorem* 11 *to each block where, for the kth block (of length $N_k$), the expert class is $\mathcal{M}_k^{\varepsilon_k}$. Taking $N_k = 2^{2^k}$ and $\varepsilon_k = 1/N_k$, we have for all sufficiently large $k$, $y \in [-B, B]^{\infty}$ and exponentially regular stationary binary process $\mathbf{X}$,*

$$\hat{L}_P(y_1^n) - \hat{L}_{F^{\text{opt}_k}}(y_1^n) \le O\left(\sqrt{n}\right), \tag{103}$$

*where the predictors $\{F^{\text{opt}_k}\}_{k\ge 0}$ are as defined in* (93) *for the process $\mathbf{X}$.*

To see why exponential regularity is needed, note that by taking for the $k$th block $\varepsilon_k$ which decays super exponentially (as in the above theorem) guarantees the validity of inequality (102) (with $\varepsilon = \varepsilon_k$) for all sufficiently large $k$. This fact will be exploited in the proof that follows.

PROOF OF THEOREM 12. Suppose first that $n = \sum_{j=1}^{k} N_j + t$, where $1 \le t \le N_{k+1}$. We then have for sufficiently large $k$,

$$\hat{L}_P(y_1^n) - \hat{L}_{F^{\mathrm{opt}_k}}(y_1^n)$$

(104) $$= \hat{L}_P\big(y_1^{N_1+\cdots+N_{k-1}}\big) - \hat{L}_{F^{\mathrm{opt}_k}}\big(y_1^{N_1+\cdots+N_{k-1}}\big)$$

(105) $$+ \hat{L}_P\big(y_{N_1+\cdots+N_{k-1}+1}^{N_1+\cdots+N_k}\big) - \hat{L}_{F^{\mathrm{opt}_k}}\big(y_{N_1+\cdots+N_{k-1}+1}^{N_1+\cdots+N_k}\big)$$

(106) $$+ \hat{L}_P\big(y_{N_1+\cdots+N_k+1}^{N_1+\cdots+N_k+t}\big) - \hat{L}_{F^{\mathrm{opt}_k}}\big(y_{N_1+\cdots+N_k+1}^{N_1+\cdots+N_k+t}\big)$$

(107) $$\le c_1 \sum_{j=1}^{k-1} N_j$$

(108) $$+ c_l(B)2^{k+1}\ln N_k + 2C/\beta_k$$

(109) $$+ c_l(B)2^{k+2}\ln N_{k+1} + 2C/\beta_{k+1}$$

(110) $$\begin{aligned} &\le c_1 N_{k-1}\big(1+o(k)\big) + c_l(B)2^{k+1}\ln N_k + 2C/\beta_k \\ &\quad + c_l(B)2^{k+2}\ln N_{k+1} + 2C/\beta_{k+1} \\ &= c_1\sqrt{N_k}\big(1+o(k)\big) + 2c_l(B)\big[(\ln N_k)^2 + (\ln N_{k+1})^2\big] \\ &\quad + 4C/\beta_{k+1} \end{aligned}$$

(111) $$= O\big(\sqrt{N_k}\big)$$

(112) $$\le O\big(\sqrt{n}\big),$$

where $\beta_k =_{\mathrm{def}} \min_{\{v\in\{0,1\}^k:\,\Pr(X_1^k=v)>0\}} \Pr(X_1^k = v)$. The term in (107) bounds that in (104) for some constant $c_1$ as it clearly follows from the boundedness of the loss function and the definition of $\hat{L}_F(\cdot)$ that, for any $F$, $y \in [-B, B]$, $a, l : |\hat{L}_F(y_{a+1}^{a+l})| \le c_1 l$. To see that (108) bounds (105) and, similarly, that (109) bounds (106), we write

(113) $$\hat{L}_P\big(y_{N_1+\cdots+N_{k-1}+1}^{N_1+\cdots+N_k}\big) - \hat{L}_{F^{\mathrm{opt}_k}}\big(y_{N_1+\cdots+N_{k-1}+1}^{N_1+\cdots+N_k}\big)$$

$$\le \hat{L}_P\big(y_{N_1+\cdots+N_{k-1}+1}^{N_1+\cdots+N_k}\big) - \inf_{F\in\mathcal{M}_k^{\varepsilon_k}} \hat{L}_F\big(y_{N_1+\cdots+N_{k-1}+1}^{N_1+\cdots+N_k}\big)$$

(114) $$+ CN_k\varepsilon_k\frac{1}{\beta_k-\varepsilon_k}$$

(115) $$\le c_l(B)\ln|\mathcal{M}_k^{\varepsilon_k}| + C\frac{1}{\beta_k-\varepsilon_k}$$

(116) $$\le c_l(B)\ln\lceil 1/\varepsilon_k\rceil^{2^{k+1}} + C\frac{1}{\beta_k-\varepsilon_k}$$

(117) $$\le c_l(B)2^{k+1}\ln N_k + 2C\frac{1}{\beta_k},$$

where inequality (115) follows from Theorem 11. Inequality (114) follows from Lemma 3 [the constant $C$ is that appearing in inequality (102)]. Finally, equality (111) follows since, by definition, $N_k$ grows superexponentially with $k$ while our exponential regularity assumption on the process $\mathbf{X}$ assures us that $1/\beta_k$ grows, at most, exponentially rapidly. To conclude, the only thing left to verify is that for any $n' > n$ of the form $n' = \sum_{j=1}^{k'} N_j + t$, where $k' > k$ and $1 \le t \le N_{k'+1}$, we still have

$$\hat{L}_P\big(y_1^{n'}\big) - \hat{L}_{F^{\mathrm{opt}_k}}\big(y_1^{n'}\big) \le O\big(\sqrt{n'}\big). \tag{118}$$

This fact is established easily as at the end of the proof of Theorem 8.

PROOF OF THEOREM 9. Suppose first that for some $B > 0$ we have

$$\int_{-B}^{B} f(y_t|x_t)\,dy_t = 1 \tag{119}$$

and that $Y_t$ is an unbiased estimator for $x_t$, that is,

$$\int_{-B}^{B} y_t f(y_t|x_t)\,dy_t = x_t. \tag{120}$$

For this case it was established in [19], Lemma 2, that $\hat{L}_F(Y_1^n)$ is an efficient estimator for $L_F(Y_1^n, x_1^n)$ in the sense that, for any predictor $F$ and all $\mathbf{x} \in \{0,1\}^\infty$ we have

$$\limsup_{n\to\infty} \frac{|L_F(Y_1^n, x_1^n) - \hat{L}_F(Y_1^n)|}{\sqrt{n \log\log n}} \le C(B), \qquad Q_f\text{-a.s.}, \tag{121}$$

where $C(B)$ is a deterministic constant depending only on $B$ and we let $Q_f$ denote the probability measure corresponding to the case where the channel is given by $f(\cdot|\cdot)$. In particular, this gives

$$\lim_{n\to\infty} \frac{1}{n}\big|L_F(Y_1^n, x_1^n) - \hat{L}_F(Y_1^n)\big| = 0, \qquad Q_f\text{-a.s.} \tag{122}$$

Consequently, by combining Theorem 12 with (122), we have for all $k$ and $\mathbf{x} \in \{0,1\}^{\mathbb{N}}$,

$$L_P(\mathbf{Y}, \mathbf{x}) \le L_{F^{\mathrm{opt}_k}}(\mathbf{Y}, \mathbf{x}), \qquad Q_f\text{-a.s.}, \tag{123}$$

where $P$ is the predictor of Theorem 12. This implies that if $\mathbf{X}$ is any stationary and ergodic sequence then almost surely

$$L_P(\mathbf{Y}, \mathbf{X}) \le L_{F^{\mathrm{opt}_k}}(\mathbf{Y}, \mathbf{X}) \tag{124}$$

$$= \lambda_k^r(\mathbf{X}|\mathbf{Y}), \tag{125}$$

where the equality follows from the obvious fact that $F^{\mathrm{opt}_k} \in \mathcal{M}_k$ almost surely has the best asymptotic performance among all members of $\mathcal{M}_k$. Since $k$ in (125) is arbitrary, we have almost surely

$$L_P(\mathbf{Y}, \mathbf{X}) \le \lambda^r(\mathbf{X}|\mathbf{Y}). \tag{126}$$

Applying Theorem 10 completes the proof for the case where the channel satisfies (119) and (120). To establish the proof for the general case, we now argue that any other situation can be reduced to a channel which satisfies (119) and (120). To see this, note that for any process $\{(X_t, Y_t)\}$ with Bayesian envelope $U(l, \mathbf{P})$, and for any measurable injection $h$, the process $\{(X_t, Z_t)\}$ where $Z_t = h(Y_t)$ has the same Bayesian envelope (as $Y_t$ and $Z_t$ clearly carry the same information). Therefore, even when the channel does not satisfy (119) and (120), the existence of some $h(\cdot)$ for which $Z_t = h(Y_t)$ and $f(z_t|x_t)$ does satisfy these equations will suffice as, clearly, a universal predictor can be tailored for the joint process $\{(X_t, Z_t)\}$, and this predictor can be employed when the process is $\{(X_t, Y_t)\}$ by simply feeding the predictor with $h(Y_t)$. Clearly, as $\{(X_t, Y_t)\}$ and $\{(X_t, Z_t)\}$ have the same Bayesian envelope, such a predictor would be universal for $\{(X_t, Y_t)\}$. Obtaining such a function, namely an $h(\cdot)$ under which $f(z_t|x_t)$ satisfies (119) and (120), is a straightforward task. To see this concretely, let, for example, $\tilde{h}: \mathbb{R} \to (-1, 1)$ be the (increasing and continuous) bijection given by

$$\tilde{h}(t) = 2\,\mathrm{sign}(t)\left[\frac{1}{1+e^{-t^2}} - 1/2\right].$$

Let further $a(x_t) =_{\mathrm{def}} \int_{-\infty}^{\infty} \tilde{h}(y_t) f(y_t|x_t)\, dy_t$ denote the expected value of $\tilde{h}(Y_t)$ conditioned on $X_t = x_t$. It is straightforward to check that, when $a(0) \neq a(1)$, letting $Z_t = h(Y_t)$, where $h(\cdot) = \frac{1}{a(1)-a(0)}\tilde{h}(\cdot) - \frac{a(0)}{a(1)-a(0)}$, gives $f(z_t|x_t)$ which satisfies (119) and (120) (with $B = \frac{1+|a(0)|}{|a(1)-a(0)|}$). Clearly, in the ($a$-typical) case where $a(0) = a(1)$, the above suggested $\tilde{h}(\cdot)$ can be replaced by any other injection (with a bounded range) under which $a(0) \neq a(1)$ (such an injection will always exist whenever the channel is distinguishable). □

We remark that, though the components of the noise process considered in this section were assumed independent (i.e., a memoryless channel), Theorem 9 can actually be shown to hold for much more general noise processes (cf. discussion in [19] regarding the generality of the noise process). Also, the assumption that the noise components have a density which is absolutely continuous w.r.t. Lebesgue measure can be significantly relaxed.

4.3. *Remark.* In this section, we defined the Markov noisy predictability $\lambda(\mathbf{x}|\mathbf{y})$ for the case of binary observations and $\lambda^r(\mathbf{x}|\mathbf{y})$ for the case of continuous-valued observations. Our construction actually gave rise to a predictor which was universal in the "individual sequence" noisy setting, that is, for which

$$L_F(\mathbf{Y}, \mathbf{x}) \leq \lambda(\mathbf{x}|\mathbf{Y}), \qquad Q\text{-a.s. } \forall\, \mathbf{x} \in \{0,1\}^{\infty}, \tag{127}$$

where $Q$ is the product measure governing the noise components and in the continuous case the right-hand side is replaced by $\lambda^r(\mathbf{x}|\mathbf{Y})$. Note that (127) implies that $L_F(\mathbf{Y}, \mathbf{X}) \leq \lambda(\mathbf{X}|\mathbf{Y})$ [resp., $L_F(\mathbf{Y}, \mathbf{X}) \leq \lambda^r(\mathbf{X}|\mathbf{Y})$], $\mathbf{P}(P, Q, g)$-a.s.

for *any* $P$. In particular, when $\mathbf{P}(P, Q, g)$ satisfies (14), then we have shown in the previous sections that $\lambda(\mathbf{X}|\mathbf{Y}) \le U(l, P, Q, g)$ [resp., $\lambda^r(\mathbf{X}|\mathbf{Y}) \le U(l, P, Q, g)$], $\mathbf{P}(P, Q, g)$-a.s. and hence the predictor satisfying (127) is universal in the sense of Definition 2. Thus, an alternative definition of universality may be directly through $\lambda(\mathbf{X}|\mathbf{Y})$ [resp., $\lambda^r(\mathbf{X}|\mathbf{Y})$] rather than through $U(l, P, Q, g)$, and the predictors constructed here would be universal under such a definition as well.

## APPENDIX

PROOF OF LEMMA 3. Recall that $\beta = \min\{\beta_{x_{t-k}^{t-1}} : \beta_{x_{t-k}^{t-1}} > 0\}$ and fix $\varepsilon < \beta$. Throughout this proof we let, for any $a \in [0, 1]$, $[a]$ denote the integer multiple of $\varepsilon$ which is closest to $a$. Denoting, for convenience, $\gamma_{x_{t-k}^{t-1}} = \prod_{i=t-k}^{t-1} f(Y_i|x_i)$, and $I = \{0, 1\}^k$, it will clearly be more than enough to establish, for all $\{\alpha_i\}_{i\in I} \in [0, 1]^I$, $\{\beta_i\}_{i\in I}$ in the simplex ($\beta_i \ge 0$, $\sum_{i\in I} \beta_i = 1$), and $\{\gamma_i\}_{i\in I} \in [0, \infty)^I$, that

$$\left| \frac{\sum_{i\in I} \alpha_i\beta_i\gamma_i}{\sum_{i\in I} \beta_i\gamma_i} - \frac{\sum_{i\in I} [\alpha_i][\beta_i]\gamma_i}{\sum_{i\in I} [\beta_i]\gamma_i} \right| \le \varepsilon\left(\frac{2}{\beta - \varepsilon} + 1\right). \tag{A.1}$$

To this end, note first that, since we can write

$$\frac{\sum_{i\in I} \alpha_i\beta_i\gamma_i}{\sum_{i\in I} \beta_i\gamma_i} = \frac{\sum_{i\in I} \alpha_i\beta_i(\gamma_i / \max_{i'\in I} \gamma_{i'})}{\sum_{i\in I} \beta_i(\gamma_i / \max_{i'\in I} \gamma_{i'})}, \tag{A.2}$$

there is no loss of generality in henceforth assuming that $\{\gamma_i\}_{i\in I} \in [0, 1]^I$. Now

$$\left| \frac{\sum_{i\in I} \alpha_i\beta_i\gamma_i}{\sum_{i\in I} \beta_i\gamma_i} - \frac{\sum_{i\in I} [\alpha_i]\beta_i\gamma_i}{\sum_{i\in I} \beta_i\gamma_i} \right| \le \frac{\sum_{i\in I} |\alpha_i - [\alpha_i]|\beta_i\gamma_i}{\sum_{i\in I} \beta_i\gamma_i} \tag{A.3}$$

$$\le \varepsilon. \tag{A.4}$$

In addition, letting $\varepsilon_i = [\beta_i] - \beta_i$ (so that $|\varepsilon_i| \le \varepsilon$), we have

$$\begin{aligned}
&\left| \frac{\sum_{i\in I} \alpha_i\beta_i\gamma_i}{\sum_{i\in I} \beta_i\gamma_i} - \frac{\sum_{i\in I} \alpha_i[\beta_i]\gamma_i}{\sum_{i\in I} [\beta_i]\gamma_i} \right| \\
&\quad = \left| \frac{\sum_{i\in I} \alpha_i\beta_i\gamma_i}{\sum_{i\in I} \beta_i\gamma_i} - \frac{\sum_{i\in I} \alpha_i\beta_i\gamma_i + \sum_{i\in I} \varepsilon_i\alpha_i\gamma_i}{\sum_{i\in I} \beta_i\gamma_i + \sum_{i\in I} \varepsilon_i\gamma_i} \right| \\
&\quad = \left| \frac{\sum_{i\in I} \alpha_i\beta_i\gamma_i}{\sum_{i\in I} \beta_i\gamma_i} \left[ 1 - \left( 1 + \frac{\sum_{i\in I} \varepsilon_i\gamma_i}{\sum_{i\in I} \beta_i\gamma_i} \right)^{-1} \right] - \frac{\sum_{i\in I} \varepsilon_i\alpha_i\gamma_i}{\sum_{i\in I} (\beta_i + \varepsilon_i)\gamma_i} \right| \\
&\quad \le \left| 1 - \left( 1 + \frac{\sum_{i\in I} \varepsilon_i\gamma_i}{\sum_{i\in I} \beta_i\gamma_i} \right)^{-1} \right| + \left| \frac{\sum_{i\in I} \varepsilon_i\alpha_i\gamma_i}{\sum_{i\in I} (\beta_i + \varepsilon_i)\gamma_i} \right| \\
&\quad \le \frac{\sum_{i\in I} (|\varepsilon_i|/\beta_i)\beta_i\gamma_i}{\sum_{i\in I} \beta_i\gamma_i} + \frac{\varepsilon \sum_{i\in I} \gamma_i}{\beta \sum_{i\in I} (1 - \varepsilon/\beta)\gamma_i}
\end{aligned} \tag{A.5}$$

$$\leq \varepsilon\big(1/\beta + 1/(\beta-\varepsilon)\big)$$

$$\leq \varepsilon\frac{2}{\beta-\varepsilon}. \tag{A.6}$$

By the arbitrariness of $\{\alpha_i\}$, (A.6) implies also

$$\left|\frac{\sum_{i\in I}[\alpha_i]\beta_i\gamma_i}{\sum_{i\in I}\beta_i\gamma_i} - \frac{\sum_{i\in I}[\alpha_i][\beta_i]\gamma_i}{\sum_{i\in I}[\beta_i]\gamma_i}\right| \leq \varepsilon\frac{2}{\beta-\varepsilon}. \tag{A.7}$$

Finally, (A.3) and (A.7), combined with the triangle inequality, give (A.1). □

*Example of a stationary process* $\{(X_t, Y_t)\}$ *which does not comply with the model of Definition* 1: Let $\mathbf{P}$ be the probability measure under which $\{(X_t, Y_t)\}_{t\in\mathbb{Z}}$ is distributed as follows:

$$\cdots \begin{matrix} X_{-2} & X_{-1} & X_0 & X_1 & X_2 \\ Y_{-2} & Y_{-1} & Y_0 & Y_1 & Y_2 \end{matrix} \cdots = \begin{cases} \cdots \begin{matrix} 0&0&0&0&0 \\ 1&1&1&1&1 \end{matrix} \cdots, & \text{w.p. } \frac{1}{4}, \\ \cdots \begin{matrix} 1&0&1&0&1 \\ 1&0&1&0&1 \end{matrix} \cdots, & \text{w.p. } \frac{1}{4}, \\ \cdots \begin{matrix} 0&1&0&1&0 \\ 0&1&0&1&0 \end{matrix} \cdots, & \text{w.p. } \frac{1}{4}, \\ \cdots \begin{matrix} 1&1&1&1&1 \\ 0&0&0&0&0 \end{matrix} \cdots, & \text{w.p. } \frac{1}{4}. \end{cases} \tag{A.8}$$

Clearly, $\mathbf{P}$ is stationary. In addition, we observe that

$$\begin{aligned} \mathbf{P}(Y_1 = 1 | X_1 = 1) &= 1/2 \\ &< 1 \\ &= \mathbf{P}(Y_1 = 1, Y_2 = 0 | X_1 = 1, X_2 = 0). \end{aligned} \tag{A.9}$$

Assume, by contradiction, that there exists a tuple $(P, Q, g)$ such that $\mathbf{P} = \mathbf{P}(P, Q, g)$. This would imply

$$\begin{aligned} &\mathbf{P}(Y_1 = 1, Y_2 = 0 | X_1 = 1, X_2 = 0) \\ &\quad = \mathbf{P}\big(g(X_1, N_1) = 1, g(X_2, N_2) = 0 | X_1 = 1, X_2 = 0\big) \\ &\quad = \mathbf{P}\big(g(1, N_1) = 1, g(0, N_2) = 0 | X_1 = 1, X_2 = 0\big) \\ &\quad = \mathbf{P}\big(g(1, N_1) = 1, g(0, N_2) = 0\big) && \text{(A.10)} \\ &\quad \leq \mathbf{P}\big(g(1, N_1) = 1\big) && \text{(A.11)} \\ &\quad = \mathbf{P}\big(g(X_1 = 1, N_1) = 1 | X_1 = 1\big) && \text{(A.12)} \\ &\quad = \mathbf{P}(Y_1 = 1 | X_1 = 1), && \text{(A.13)} \end{aligned}$$

where (A.10) and (A.12) follow from the independence of $\mathbf{X}$ and $\mathbf{N}$ under $\mathbf{P}(P, Q, g)$. Clearly, the inequality established between the two ends of the above chain contradicts (A.9).

**Acknowledgment.** The authors are grateful to R. Atar for providing the example for a stationary sequence which is not exponentially regular.

## REFERENCES

[1] ALGOET, P. H. (1994). The strong law of large numbers for sequential decisions under uncertainty. *IEEE Trans. Inform. Theory* **40** 609–634.

[2] ATAR, R. and ZEITOUNI, O. (1997). Lyapunov exponents for finite state nonlinear filtering. *SIAM J. Control Optim.* **35** 36–55.

[3] BREIMAN, L. (1973). A note on minimax filtering. *Ann. Probab.* **1** 175–179.

[4] BREIMAN, L. (1992). *Probability*. SIAM, Philadelphia, PA.

[5] CASTELLI, V. and COVER, T. M. (1995). On the exponential value of labeled samples. *Pattern Recognition Letters* **16** 105–111.

[6] CHEN, C. and KASSAM, S. A. (1984). Robust Wiener filtering for multiple inputs with channel distortion. *IEEE Trans. Inform. Theory* **30** 674–677.

[7] FEDER, M., MERHAV, N. and GUTMAN, M. (1992). Universal prediction of individual sequences. *IEEE Trans. Inform. Theory* **38** 1258–1270.

[8] FRANKE, J. (1985). Minimax: Robust prediction of discrete time series. *Z. Wahrsch. Verw. Gebiete* **68** 337–364.

[9] GYÖRFI, L. and LUGOSI, G. (2001). 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, Dordrecht.

[10] GYÖRFI, L., LUGOSI, G. and MORVAI, G. (1999). A simple randomized algorithm for sequential prediction of ergodic time series. *IEEE Trans. Inform. Theory* **45** 2642–2650.

[11] HALL, P. and HEYDE, C. C. (1980). *Martingale Limit Theory and Its Application*. Academic Press, New York.

[12] HAUSSLER, D., KIVINEN, J. and WARMUTH, M. K. (1998). Sequential prediction of individual sequences under general loss functions. *IEEE Trans. Inform. Theory* **44** 1906–1925.

[13] KASSAM, S. A. and LIM, T. L. (1977). Robust Wiener filters. *J. Franklin Inst.* **304** 171–185.

[14] KASSAM, S. A. and POOR, H. V. (1985). *Robust Techniques for Signal Processing*: *A Survey*. *Proceedings of the IEEE* **73**.

[15] MERHAV, N. and FEDER, M. (1998). Universal prediction. *IEEE Trans. Inform. Theory* **44** 2124–2147.

[16] VASTOLA, K. S. and POOR, H. V. (1984). Robust Wiener–Kolmogorov theory. *IEEE Trans. Inform. Theory* **30** 316–327.

[17] VOVK, V. (1990). Aggregating strategies. In *Proc. Third Annual Workshop on Computational Learning Theory* 371–383. Kaufmann, San Mateo, CA.

[18] VOVK, V. (1995). A game of prediction with expert advice. In *Proc. Eighth Annual Workshop on Computational Learning Theory* 51–60. ACM Press, NY.

[19] WEISSMAN, T. and MERHAV, N. (2001). Universal prediction of individual binary sequences in the presence of noise. *IEEE Trans. Inform. Theory* **47** 2151–2173.

[20] WEISSMAN, T., MERHAV, N. and BARUCH, A. (2001). Twofold universal prediction schemes for achieving the finite-state predictability of a noisy individual binary sequence. *IEEE Trans. Inform. Theory* **47** 1849–1866.

[21] ZIV, J. and LEMPEL, A. (1978). Compression of individual sequences via variable rate coding. *IEEE Trans. Inform. Theory* **24** 530–536.

DEPARTMENT OF ELECTRICAL ENGINEERING
STANFORD UNIVERSITY
PACKARD 256
STANFORD, CALIFORNIA 94305-9510
USA
E-MAIL: tsachy@stanford.edu

DEPARTMENT OF ELECTRICAL ENGINEERING
TECHNION—ISRAEL INSTITUTE OF TECHNOLOGY
TECHNION CITY, HAIFA 32000
ISRAEL
E-MAIL: merhav@ee.technion.ac.il