Universal Prediction

Neri Merhav, *Senior Member, IEEE*, and Meir Feder, *Senior Member, IEEE*

(Invited Paper)

***Abstract*— This paper consists of an overview on universal prediction from an information-theoretic perspective. Special attention is given to the notion of probability assignment under the self-information loss function, which is directly related to the theory of universal data compression. Both the probabilistic setting and the deterministic setting of the universal prediction problem are described with emphasis on the analogy and the differences between results in the two settings.**

***Index Terms*— Bayes envelope, entropy, finite-state machine, linear prediction, loss function, probability assignment, redundancy-capacity, stochastic complexity, universal coding, universal prediction.**

I. Introduction

Can the future of a sequence be predicted based on its past? If so, how good could this prediction be? These questions are frequently encountered in many applications. Generally speaking, one may wonder why should the future be at all related to the past. Evidently, often there is such a relation, and if it is known in advance, then it might be useful for prediction. In reality, however, the knowledge of this relation or the underlying model is normally unavailable or inaccurate, and this calls for developing methods of universal prediction. Roughly speaking, a universal predictor is one that does not depend on the unknown underlying model and yet performs essentially as well as if the model were known in advance.

This is a survey that describes some of the research work on universal prediction that has been carried out throughout the years in several scientific disciplines such as information theory, statistics, machine learning, control theory, and operations research. It should be emphasized, however, that there is no attempt to cover comprehensively the entire volume of work that has been done in this problem area. Rather, the aim is to point out a few of the highlights and the principal methodologies from the authors' personal information-theoretic perspective. Also, throughout the paper there are a few new results whose derivations are given in detail.

Manuscript received December 1, 1997; revised April 8, 1998. This work was supported by the Israel Science Foundation administered by the Israeli Academy of Sciences and Humanities.

N. Merhav is with the Department of Electrical Engineering, Technion–Israel Institute of Technology, Haifa 32000, Israel (e-mail: merhav@ee.technion.ac.il).

M. Feder is with the Department of Electrical Engineering–Systems, Tel Aviv University, Tel Aviv 69978, Israel (e-mail: meir@eng.tau.ac.il).

Publisher Item Identifier S 0018-9448(98)05086-X.

Historically, the information-theoretic approach to prediction dates back to Shannon [104], who related prediction to entropy and proposed a predictive estimate of the entropy of the English language. Inspired by Haggelbarger, Shannon [105] created later a "mind-reading" machine that predicts human decisions. About that time, Kelly [59] showed the equivalence between gambling (which, in turn, is definitely a form of prediction) and information. Following Cover [17], Rissanen [89], [90], and Rissanen and Langdon [93], it is well recognized to date that universal prediction is intimately related to universal lossless source coding. In the last three decades, starting from the pioneering work of Fittingoff [42] and Davisson [27], and later Ziv [124], Lempel and Ziv [68], [125], [126], Rissanen and Langdon [93], Krichevsky and Trofimov [63], and others, the theory and practice of universal coding have been greatly advanced. The state-of-the-art knowledge in this area is sufficiently mature to shed light on the problem of universal prediction. Specifically, prediction schemes as well as fundamental performance limits (lower bounds), stemming from those of universal coding, have been derived. It is the relation between universal coding and universal prediction that is the main theme of this paper, from the point of view of both algorithms and performance bounds.

Let us now describe the prediction problem in general. An observer sequentially receives a sequence of observations $x_1, x_2, \cdots, x_t, \cdots$ over some alphabet $\mathcal{X}$. At each time instant t, after having seen $x^{t-1} = (x_1, \cdots, x_{t-1})$ but not yet x_t, the observer predicts the next outcome x_t, or more generally, makes a decision b_t based on the observed past x^{t-1}. Associated with this prediction or decision b_t, and the actual outcome x_t, there is a loss function $l(b_t, x_t)$ that measures quality. Depending on the particular setting of the prediction problem, the objective would be to minimize this instantaneous loss, or its time-average, or the expected value of either one of these quantities. Obviously, prediction in the ordinary sense is a special case of this, where $b_t = \hat{x}_t$ is an estimate of x_t based on x^{t-1} and $l(b_t, x_t) = l(\hat{x}_t, x_t)$ is some estimation performance criterion, e.g., the Hamming distance (if x_t is discrete) or the squared error $l(b_t, x_t) = (x_t - b_t)^2$ (if x_t is continuous).

Another special case, which is more general than the above examples, is based on assigning weights or probabilities to all possible values of the next outcome. For example, the weatherman may assess 70% chance of rain tomorrow, instead of making a commitment whether it will rain or not. This is clearly more informative than the ordinary prediction described above because it gives an assessment of the degree of *confidence* or *reliability* associated with the prediction. In terms of the above described prediction problem, here b_t is a conditional probability assignment of x_t given x^{t-1}, i.e., a nonnegative function $b_t(\cdot|x^{t-1})$ that integrates (or sums) to unity for every x^{t-1}. Upon observing x_t, the performance of b_t is assessed with respect to a suitable loss function l, which should decrease monotonically with the probability assigned to the actual outcome $b_t(x_t|x^{t-1})$. A very important loss function of this kind is the *self-information loss* function, which is also referred to as the *log–loss* function in the machine-learning literature. For every probability assignment $b = \{b(x), x \in \mathcal{X}\}$ over $\mathcal{X}$ and every $x \in \mathcal{X}$, this function is defined as

$$l(b, x) = -\log b(x) \tag{1}$$

where logarithms throughout this paper are taken to the base 2 unless otherwise specified. For reasons to be discussed in Section II, the self-information loss function plays a central role in the literature on prediction and hence also throughout this survey.

Let us now return to the prediction problem in its general form. Quite clearly, solutions to this problem are sought according to the particular assumptions on the data-generating mechanism and on the exact objectives. Classical statistical decision theory (see, e.g., [35]) assumes that a known probabilistic source P generates the data, and so, a reasonable objective is to minimize the expected loss. The optimum strategy b_t^* then minimizes the expected loss, given the past, i.e.,

$$E\{l(b, X_t)|X^{t-1} = x^{t-1}\} = \int_{\mathcal{X}} dP(x|x^{t-1}) l(b, x) \tag{2}$$

where random variables are denoted by capital letters. Moreover, under suitable assumptions on stationarity and ergodicity, optimum prediction $\{b_t^*\}$ in the expected loss sense, is optimum also in the sense of minimizing the almost sure asymptotic time-average of $l(b_t, X_t)$ (see, e.g., [4]). Given $X^{t-1} = x^{t-1}$, the quantity

$$U(x^{t-1}) = \inf_b \int dP(x|x^{t-1}) l(b, x)$$

is referred to as the conditional *Bayes envelope* given x^{t-1}. For example, if $\{X_t\}$ is a binary source, $b_t = \hat{x}_t$, and $l(\cdot, \cdot)$ is the Hamming distance, then

$$b_t^* = \begin{cases} 0, & \text{if } P(0|x^{t-1}) \geq P(1|x^{t-1}) \\ 1, & \text{otherwise} \end{cases} \tag{3}$$

and the conditional Bayes envelope given x^{t-1} is

$$U(x^{t-1}) = \min\{P(0|x^{t-1}), P(1|x^{t-1})\}.$$

For $l(b, x) = (b - x)^2$

$$b_t^* = E(X_t|X^{t-1} = x^{t-1})$$

and

$$U(x^{t-1}) = \mathrm{Var}\,\{X_t|X^{t-1} = x^{t-1}\}.$$

If, in addition, the underlying source P is known to be Gaussian (or, if only the class of linear predictors is allowed), then b_t^* is well-known to be a linear function of x^{t-1} given as a special case of the causal Wiener filter [119] (see also [86, Ch. 14-3]). In the self-information loss case, $b_t^*(\cdot|x^{t-1}) = P(\cdot|x^{t-1})$ minimizes $E\{-\log b(X_t|X^{t-1} = x^{t-1})\}$, namely, the best probability assignment is the true one. The conditional Bayes envelope given x^{t-1}, is the (differential) entropy of X_t given $X^{t-1} = x^{t-1}$, i.e.,

$$U(x^{t-1}) = -E \log P(X_t|X^{t-1} = x^{t-1}).$$

While classical theory (e.g., Wiener prediction theory) assumes that the source P is known, the more realistic and interesting situation occurs when P is either unknown, or nonexistent. In the second case, there is no probabilistic data-generating mechanism and the data are considered arbitrary and deterministic. Both cases fall into the category of the universal prediction problem, where the former is referred to as the *probabilistic setting* and the latter is called the *deterministic setting*. Let us now elaborate on these two settings.

A. The Probabilistic Setting

In the probabilistic setting the objective is normally to minimize the expected cumulative loss asymptotically for large n simultaneously for any source in a certain class. A universal predictor $\{b_t^u(x^{t-1})\}$ does not depend on P, and at the same time, keeps the difference between

$$E\left\{\frac{1}{n}\sum_{t=1}^{n} l(b_t^u, X_t)\right\}$$

and

$$\begin{aligned}\overline{U}_n(P) &= \frac{1}{n}\sum_{t=1}^{n} EU(X^{t-1}) \\ &= \frac{1}{n}\sum_{t=1}^{n} E\left\{\inf_b E[l(b, X_t)|X^{t-1})]\right\}\end{aligned} \tag{4}$$

vanishingly small for large n. The cumulative Bayes envelope of (4) represents the performance of the optimal predictor tuned to P. For a stationary and ergodic source, the sequence $\{\overline{U}_n(P)\}_{n\geq 1}$ has a limit $\overline{U}(P)$, referred to as the *asymptotic Bayes envelope*, that coincides (by the Cesaro theorem [23]) with $\lim_{t\to\infty} E\{U(X^t)\}$, which in turn exists by nonincreasing monotonicity. In the self-information loss case, $\overline{U}(P)$ is the entropy rate of P, which means that the goal of universal prediction is equivalent to that of universal coding.

There are essentially three levels of universality according to the degree of uncertainty regarding the source.

Universality with Respect to Indexed Classes of Sources: Suppose that the source is unknown except for being a member of a certain indexed class $\{P_\theta, \theta \in \Lambda\}$, where Λ is the index set. Most commonly, θ designates a parameter vector of a smooth parametric family, e.g., the families of finite-alphabet memoryless sources, kth-order Markov sources, M-state sources, AR (p) Gaussian sources, but other index sets (e.g., finite sets) are possible as well. There are two interesting issues here. The first is to devise universal prediction schemes that asymptotically attain $\overline{U}_n(P_\theta)$ in the above defined sense for every $\theta \in \Lambda$, and the second is performance bounds beyond $\overline{U}_n(P_\theta)$ that apply to any universal predictor. Analogously to the universal coding terminology, the extra loss beyond $\overline{U}_n(P)$ will be referred to as the *redundancy*. Redundancy bounds are useful to establish necessary conditions for the existence of universal schemes as well as limitations on the rate of convergence. Both are dictated by a certain measure of the *richness* of the class $\{P_\theta\}$. Furthermore, even if the redundancy bound does not vanish as $n \rightarrow \infty$, and hence universal schemes in the above defined sense do not exist, the question of universality can be extended to that of achieving this bound. For self-information loss prediction, we will explicitly characterize such bounds, and demonstrate achievability by certain universal schemes.

Universality with Respect to Very Large Classes of Sources: Suppose that all we know about the source is that it is Markov of an unknown finite order, or that it is stationary and ergodic, or mixing in a certain sense. For such large classes, quantitative characterizations of uniform redundancy rates do not exist [60], [106], [107]. Here, one cannot hope for more than *weak universality*, a term mentioned and defined in [27], which means that universality is attained at a nonuniform convergence rate. Sometimes even weak universality cannot be obtained, and in [60] there are necessary and sufficient conditions for the existence of universal schemes.

Hierarchical Universality: In this level, the goal is to devise universal schemes with respect to a sequence $\Lambda_1, \Lambda_2, \cdots$ of index sets of sources, which may (though not necessarily) have some structure like nesting, i.e., $\Lambda_k \subset \Lambda_{k+1}$ for every positive integer k. Perhaps the most common example is where for every k, Λ_k is the class of all kth-order Markov sources of a given alphabet. Here the only prior knowledge that one may have on the source is that its index θ belongs to $\Lambda = \bigcup_{k \geq 1} \Lambda_k$. The straightforward approach would be to consider Λ as one big class and to seek universal schemes with respect to Λ. The drawback of this approach, however, is that it is pessimistic in the sense that the convergence rate towards $\overline{U}(P_\theta)$, might be very slow, if at all existent, because Λ could be a very rich class. In the above Markov example, while each Λ_k falls within the category of the first level above, the union Λ falls in the second level. Nonetheless, it turns out that in certain situations it is possible to achieve redundancy rate that is essentially as small as if k were known *a priori*. This gives rise to an elegant compromise between the two former levels of universality. It keeps the fast convergence rates of the first level without sacrificing the generality of the class of sources of the second level.

B. The Deterministic Setting

In this setting, the observed sequence is not assumed to be randomly drawn by some probability law, but is rather an individual, deterministic sequence. There are two difficulties in defining the universal prediction problem in this context. The first is associated with setting the desired goal. Formally, for a given sequence $x_1, x_2, \cdots$, there is always the perfect prediction function defined as $b_t(x^{t-1}) = x_t$, and so, the prediction problem seemingly boils down to triviality. The second difficulty is in the other way around. For a given deterministic predictor $\{b_t(\cdot)\}_{t \geq 1}$, there is always the adversary sequence where at each time instant t, x_t is chosen to maximize $l(b_t, x_t)$.

The first difficulty is fundamental because it means that without any limitations on the class of allowed predictors, there is a severe overfitting effect, which tailors a predictor to the sequence so strongly, that it becomes, in fact, anticipating and hence completely misses the essence of prediction as a causal, sequential mechanism. Therefore, one must limit the class B of allowed predictors $\{b_t(\cdot)\}_{t \geq 1}$ in some reasonable way. For example, B could be the class of predictors that are implementable by finite-state machines (FSM's) with M states, or Markov-structured predictors of the form $b_t(x^{t-1}) = b(x_{t-k}, \cdots, x_{t-1})$, and so on. Such limitations make sense not only by virtue of avoiding these trivialities, but also because they reflect real-life situations of limited resources, like memory, computational power, and so on. Stated more formally, for a given class B of predictors, we seek a sequential predictor $\{b_t^u\}_{t \geq 1}$ that is universal in the sense of being independent of the future, and at the same time, its average loss

$$\frac{1}{n} \sum_{t=1}^{n} l(b_t^u, x_t)$$

is asymptotically the same as

$$\min_B \frac{1}{n} \sum_{t=1}^{n} l(b_t, x_t)$$

for every x^n. The universal predictor need not be necessarily in B but it must be causal, whereas the reference predictor in B, that minimizes the average loss, may (by definition) depend on the entire sequence x^n.

The second difficulty mentioned above is alleviated by allowing randomization. In other words, predictions are generated at random according to a certain probability distribution that depends on the past. Note that this is different from the above discussed case where b_t was a probability assignment, because now the assigned probability distribution is actually used for randomization.

Analogously to the probabilistic case, here we also distinguish between three levels of universality, which are now in accordance to the richness of the class B. The first level corresponds to an indexed class of predictors which is dual to the above mentioned indexed class of sources. Examples of this are parametric classes of predictors, like finite-state machines with a given number of states, fixed-order Markov predictors, predictors based on neural nets with a given number of neurons, finite sets of predictors, etc. The second level corresponds to very large classes like the class of all finite-state predictors (without specifying the number of states), operating on infinitely long sequences, etc. Finally, the third level corresponds to hierarchical universality and parallels that of the probabilistic setting. The nature of the reported results is somewhat similar to that of the probabilistic approach, but there are several important differences in algorithmic aspects as well as in existence theorems and performance bounds.

The outline of the paper is as follows. Section II is devoted to the motivation and the justification for the use of the self-information loss function as a performance criterion in prediction. In Section III, the probabilistic setting will be discussed with a great emphasis on the self-information loss case which is fairly well-understood. In Section IV, the deterministic setting will be described with special attention to the similarity and the difference from the probabilistic setting. Section V is devoted to the concept of hierarchical universality in both settings. Finally, Section VI summarizes the paper along with some open problems and directions for further research.

II. The Self-Information Loss Function

We mentioned earlier the self-information loss function and its central role in universal prediction. In this section, we discuss some motivations and justifications for using this loss function as a measure of prediction performance. As explained in Section I, predictive probability assignment for the next outcome is more general and more informative than estimating the value of the next outcome, and a reasonable loss function should be monotonically decreasing with the assigned probability of the actual outcome. The self-information loss function, defined in (1), clearly satisfies this requirement, but it also possesses many other desirable features of fundamental importance.

The first advantage of the self-information loss function is technical. It is convenient to work with because the logarithmic function converts joint probability functions, or equivalently, products of conditional probabilities into cumulative sums of loss terms. This suits the framework of the general prediction problem described above.

But beyond this technical convenience, there is a deeper significance. As is well known, the self-information manifests the degree of uncertainty, or the amount of information treasured in the occurrence of an event. The conditional self-information of the future given the past, therefore, reflects the ability to deduce information from the past into the future with minimum uncertainty.

Evidently, prediction under the self-information loss function and lossless source coding are intimately related. This relation stems from the fact that $l(b, x) = -\log b(x)$ is the *ideal codelength* of x with respect to a probability function $b(\cdot)$. This codelength can be implemented sequentially within any desired precision using arithmetic coding [88]. Conversely, any codelength function can be translated into a probability assignment rule [90], [93], [109], [117]. Another direct application of self-information loss minimization to the problem area of prediction, is that of gambling [17], [19], [38]. In this case, $b_t(\cdot|x^{t-1})$ represents the distribution of money invested in each one of the possible values of the next outcome. The self-information loss function then dictates the exponential growth rate of the amount of money with time.

The paradigm of predictive probability assignment is also the basis of Dawid's *prequential principle* [31]. However, the motivation of the prequential principle was not in prediction *per se*, but rather the use of probability assignment for testing the validity of statistical models. A good probability assignment is one that behaves empirically as expected from the true probabilistic model. For example, if $\{x_t\}$ are binary, then a good sequence $\{b_t(1|x^{t-1})\}$ of probabilities assigned to $x_t = 1$ should satisfy

$$\frac{1}{n}\sum_{t=1}^{n}(x_t - b_t) \to 0$$

namely, the law of large numbers. As further discussed in [32]–[34], other requirements are based on the central limit theorem, the law of iterated logarithm, behavior of confidence intervals, and so on.

Interestingly, it turns out that predictive probability assignment under the self-information loss criterion can be useful also for the purpose of testing the validity of statistical models as described above. One reason is that when a certain source P governs the data, then it is the true conditional probability $b_t(\cdot|x^{t-1}) = P(\cdot|x^{t-1})$ that minimizes $E\{-\log b_t(X_t|X^{t-1} = x^{t-1})\}$. In simpler words, the maximum achievable assigned probability is also the true one (a property shared by very specific loss functions, see [78]). Moreover, by the Shannon–McMillan–Breiman theorem, under certain ergodicity assumptions, this is true not only in the expected value sense, but also almost surely. Thus by combining the prequential principle with the Shannon–McMillan–Breiman theorem, a good probabilistic model for the data $b_t(\cdot|x^{t-1})$ must minimize

$$\frac{1}{n}\sum_{t=1}^{n} -\log b_t(x_t|x^{t-1})$$

i.e., the average self-information loss.

From another perspective, we observe that any sequential probability assignment mechanism gives rise to a probability assignment for the entire observation vector x^n by

$$Q(x^n) = \prod_{t=1}^{n} b_t(x_t|x^{t-1}).$$

Conversely, any consistent probability assignment Q for x^n (i.e., Q that satisfies $Q(x^{t-1}) = \sum_{x_t\in\mathcal{X}} Q(x^t)$ for all t and x^{t-1}), provides a valid sequential probability assignment by

$$b_t(x_t|x^{t-1}) = \frac{Q(x^t)}{Q(x^{t-1})}. \tag{5}$$

Therefore, the choice of $\{b_t\}$ in self-information loss prediction is completely equivalent to the choice of Q that assigns maximum probability to x^n, that is, maximum-likelihood estimation.

In our discussion thus far, we focused on motivating the self-information loss function itself. Yet another motivation for studying universal prediction in the self-information loss case is that it sheds light on the universal prediction problem for other loss functions as well. Perhaps the most direct way to look at self-information loss prediction is as a mechanism that generates a probability distribution when the underlying source is unknown or nonexistent. One plausible approach to the prediction problem with a general loss function is then to generate, at each time instant, a prediction that is a functional of the self-information-loss conditional probability assignment. For example, in the squared-error loss case, a reasonable predictor would be the conditional mean associated with $b_t(\cdot|x^{t-1})$, which hopefully tends to the true conditional probability as discussed above. As will be seen in the probabilistic setting, this technique is often successful, whereas in the deterministic setting, some modification is required.

However, there is another way in which self-information loss prediction serves as a yardstick to prediction under other loss functions, and this is the notion of *exponential weighting.* In certain situations, minimization of the cumulative loss $\sum_t l(b_t, x_t)$ corresponds to maximization of the exponentiated loss $\exp(-\alpha \sum_t l(b_t, x_t)$ $(\alpha > 0)$, which in turn can be treated altogether as an auxiliary probability assignment. In certain important special cases (though not always), the solution to this probability assignment problem translates back as a solution to the original problem. We will also see the usefulness of the exponential weighting technique as a tool for deriving lower bounds that are induced from corresponding strong lower bounds of the self-information loss case.

III. The Probabilistic Setting

We begin with the problem of probability assignment for the next outcome given the past, under the self-information loss function. As explained above, this problem is completely equivalent to that of finding a probability assignment Q for the entire data sequence.

As we mentioned earlier, if the source P were known, then clearly, the optimal Q that minimizes the above expected self-information loss would be $Q = P$, i.e., the prediction induced by the true underlying source

$$b_t(\cdot|x^{t-1}) \triangleq Q(\cdot|x^{t-1}) = P(\cdot|x^{t-1}).$$

The average cumulative loss would then be the entropy $H_n(P) = -E\{\log P(X^n)\}$. If P is unknown and we wish to assign a certain probability distribution Q that does not depend upon the unknown P, then the extra loss beyond the entropy is given by

$$E\{-\log Q(X^n) - (-\log P(X^n))\} = D_n(P\|Q) \quad (6)$$

where $D_n(P\|Q)$ is the nth-order information divergence (relative entropy) between P and Q. In the corresponding lossless compression problem, $D_n(P\|Q)/n$ is the coding redundancy, i.e., the normalized per-symbol difference between the average code length and the entropy. Of course, the minimizations of $D_n(P\|Q)$ for two or more sources $\{P\}$ at the same time might be contradictory. Thus the problem of universal probability assignment is that of finding a good compromise Q that is uniformly as "close" as possible, in the information divergence sense, to every P in a given class of sources. We shall elaborate later on this notion of simultaneous divergence minimization.

As explained in Section I, the theory of universality splits into several levels according to the degree of uncertainty regarding the source. We begin with the conceptually simplest case where the source is known to belong to a given indexed class of sources $\{P_\theta, \theta \in \Lambda\}$, where θ is the index (e.g., a parameter vector) and Λ is the index set. Since we look at prediction from the viewpoint of probability assignment and we start from the self-information loss criterion, our survey in this part is largely taken from the theory of universal coding.

A. Indexed Classes of Sources

1) The Self-Information Loss Function: We first describe two common approaches to universal probability assignment for indexed classes of sources.

The Plug-in Approach versus the Mixture Approach: One natural approach to universal prediction with respect to an indexed class of sources $\{P_\theta, \theta \in \Lambda\}$ is the so-called *plug-in* approach. According to this approach, at every time instant t, the index (or the parameter) θ is estimated on-line from x^{t-1} (e.g., by using the maximum-likelihood estimator), and the estimate $\hat{\theta}_t = \hat{\theta}_t(x^{t-1})$ is then used for prediction as if it were the true parameter value, i.e., the conditional probability assigned to x_t is given by $P_{\hat{\theta}_t}(x_t|x^{t-1})$.

The plug-in approach may work quite well under certain regularity conditions. Intuitively, if the estimator $\hat{\theta}_t$ is statistically consistent and $P_\theta(x_t|x^{t-1})$ is continuous in θ for every x^{t-1} and x_t, then the estimated probability assignment may converge to the true conditional probability in the probabilistic sense. Nonetheless, this convergence property does not always hold (e.g., when θ is the center of a Cauchy density estimated by the sample mean), and even if it does, the rate of convergence might be of crucial importance. Moreover, it is not true, in general, that better estimation of the conditional probability necessarily yields better self-information loss performance. The plug-in approach is, in essence, a heuristic approach that lacks a well-substantiated, deep theoretical justification in general.

An alternative approach, henceforth referred to as the *mixture approach*, is based on generating convex combinations (mixtures) of all sources in the class $\{P_\theta, \theta \in \Lambda\}$. Specifically, given a certain nonnegative weight function $w(\theta)$ that integrates to unity (and hence can be thought of as a prior on Λ), we define the mixture probability mass (or density) function over n-tuples as

$$Q_w(x^n) = \int_\Lambda dw(\theta) P_\theta(x^n). \quad (7)$$

With an appropriate choice of the weight function w, the mixture Q_w, as we shall see later, turns out to possess certain desirable properties which motivate its definition as a *universal probability measure.* This universal measure then induces a conceptually simple sequential probability assignment mechanism defined by

$$b_t(x_t|x^{t-1}) = Q_w(x_t|x^{t-1}) = \frac{Q_w(x^t)}{Q_w(x^{t-1})}. \quad (8)$$

It is interesting to note [72, Theorem 2] that the above predictive probability function induced by the mixture of $\{P_\theta, \theta \in \Lambda\}$ can also be represented as a mixture of the conditional probability functions $\{P_\theta(x_t|x^{t-1}), \theta \in \Lambda\}$, where the weighting function is given by the *posterior* probability density function of θ given x^{t-1}, i.e.,

$$Q_w(x_t|x^{t-1}) = \int_\Lambda dw(\theta|x^{t-1}) P_\theta(x_t|x^{t-1}) \quad (9)$$

where

$$w(\theta|x^{t-1}) = \frac{w(\theta)P_\theta(x^{t-1})}{\int_\Lambda dw\,(\theta')P_{\theta'}(x^{t-1})} = \frac{w(\theta)2^{-\log 1/P_\theta(x^{t-1})}}{\int_\Lambda dw\,(\theta')P_{\theta'}(x^{t-1})} \tag{10}$$

and where the last expression manifests the interpretation of *exponential weighting* according to the probability assignment performance (given by $\log 1/P_\theta(x^{t-1})$) on data seen thus far: points in Λ that correspond to good performance in the past are rewarded exponentially higher weights in prediction of future outcomes. The exponential weighting is an important concept. We will further elaborate later on it in a broader context of lower bounds and algorithms for sequential prediction under more general loss functions in the probabilistic as well as in the deterministic setting.

For the class of binary memoryless (Bernoulli) sources with $\theta = \Pr\{x_t = 0\}$, the mixture approach, with $w(\cdot)$ being uniform over $\Lambda = [0, 1]$, leads to the well-known Laplace prediction [66], [67]. Suppose that x^{t-1} contains t_0 zeros and $t_1 = t - 1 - t_0$ ones, then

$$\begin{aligned} Q_w(x_t = 0|x^{t-1}) &= \frac{\int_0^1 \theta^{t_0+1}(1-\theta)^{t_1}\,d\theta}{\int_0^1 \theta^{t_0}(1-\theta)^{t_1}\,d\theta} \\ &= \frac{t_0+1}{(t-1)+2} = \frac{t_0+1}{t+1} \end{aligned} \tag{11}$$

which, in this case, can be thought of also as a plug-in algorithm because $(t_0+1)/(t+1)$ can be interpreted as a biased version of the maximum-likelihood estimator of θ. Such a bias is clearly desirable in a sequential regime because the naive maximum-likelihood estimator $\hat{\theta}_t = t_0/(t-1)$ would assign zero probability to the first occurrence of "1" which, in turn, would result in infinite loss. Also, this bias gives rise to the plausible symmetry consideration that in the absence of any data (i.e., $t_0 = t - 1 = 0$) one would assign equal probabilities to "0" and "1." But this would be also the case with any estimator of the form $\hat{\theta}_t = (t_0 + \beta)/(t + 2\beta)$, $\beta > 0$. Indeed, other weight functions (from the Dirichlet family) yield different bias terms and with slight differences in performance (see also [62]). This discussion carries over to general finite-alphabet memoryless sources [63] (as will be discussed later) and to Markov chains [28], [91]. However, it should be kept in mind that for a general family of sources $\{P_\theta, \theta \in \Lambda\}$, the mixture approach does not necessarily boil down to a plug-in algorithm as above, and that the choice of the weight function might have a much more dramatic impact on performance [76]. In this case, we would like to have some theoretical guidance regarding the choice of w.

This will be accomplished in the forthcoming subsection, where we establish the theoretical justification of the mixture approach in a fairly strong sense. Interestingly, in the next section, it will be motivated also in the deterministic setting, and for loss functions other than the self-information loss function.

Minimax and Maximin Universality: We have seen (6) that the excess loss associated with a given probability assignment Q while the underlying source is P_θ is given by $D_n(P_\theta\|Q)$. The first fundamental justification of the mixture approach (presented in [76]) is the following simple fact: given an arbitrary probability assignment Q, there exists another probability assignment Q_w in the convex hull of $\{P_\theta, \theta \in \Lambda\}$, (that is, a mixture) such that $D_n(P_\theta\|Q_w) \le D_n(P_\theta\|Q)$ simultaneously for every $\theta \in \Lambda$. This means that when seeking a universal probability assignment, there is no loss of optimality in any reasonable sense, if we confine attention merely to the convex hull of the class $\{P_\theta\}$. Nonetheless, this interesting fact does not tell us how to select the weight function $w(\cdot)$ of the mixture Q_w. To this end, we make a few additional observations.

As mentioned earlier, we wish to find a probability assignment Q that is independent of the unknown θ, and yet guarantees a certain level of excess loss beyond the minimum achievable loss had θ been known *a priori* (i.e., the nth-order entropy $H_n(P_\theta)$). Referring again to (6), this suggests to solve the following minimax problem:

$$\inf_Q \sup_{\theta\in\Lambda} D_n(P_\theta\|Q) = \inf_Q \sup_w \int_\Lambda dw\,(\theta) D_n(P_\theta\|Q). \tag{12}$$

The value of this quantity, after normalizing by n, is called the *minimax redundancy* and is denoted by R_n^+ in the literature of universal coding. At first glance, this approach might seem somewhat pessimistic because it is a worst case approach. Fortunately enough, in many cases of interest, $R_n^+ \to 0$ as $n \to \infty$, which means that the minimax Q asymptotically achieves the entropy rate, uniformly rapidly in Λ. Moreover, as we shall see shortly, the minimax approach, in the self-information loss case, is not at all pessimistic even if R_n^+ does not tend to zero. Again, in view of the discussion in the previous paragraph, the minimax-optimal Q is a mixture of the sources in the class.

An alternative to the minimax criterion is the maximin criterion, whose definition has a strong Bayesian flavor that gives rise to the mixture approach from a seemingly different point of view. Here is the idea: since $\theta \in \Lambda$ is unknown, let us postulate some prior probability density function $w(\theta)$ over Λ. The performance of a given probability assignment Q would be then judged with respect to the normalized weighted average redundancy $D_n(P_\theta\|Q)$, i.e.,

$$R_n(Q, w) = \frac{1}{n}\int_\Lambda dw\,(\theta)\, D_n(P_\theta\|Q). \tag{13}$$

It is easy to see that for a given w, the Q that minimizes $R_n(Q, w)$ is just the Q_w defined in (7), and that the resultant average redundancy $R_n(Q_w, w)$, is exactly the mutual information $I_w(\Theta; X^n)$ between random variables Θ and X^n whose joint probability density function is given by $\mu(\theta, x^n) = w(\theta)P_\theta(x^n)$. But w is arbitrary and the question that again arises is what would be an "appropriate" choice of w? Let us adopt again a worst case approach and use the "least favorable" prior that maximizes $\inf_Q R_n(Q, w)$, that is, solve the maximin problem

$$\sup_w \inf_Q R_n(Q, w) \tag{14}$$

whose value, when normalized by n, is referred to as the *maximin redundancy* and denoted by R_n^-. It is important to note that R_n^-, which is the supremum of $I_w(\Theta; X^n)/n$ over all allowable w's, is given the interpretation of the *capacity* of the "channel" from Θ to X^n, defined by the class of sources. In this definition, each source $P_\theta(x^n)$ is thought of as the conditional probability function of the "channel output" given the "channel input" Θ. We will refer to this channel capacity as the *capacity of the class* of sources $\{P_\theta, \theta \in \Lambda\}$ and will denote it by C_n. Thus C_n is identical to R_n^-.

These notions of minimax and maximin universality were first defined by Davisson [27] in the context of universal coding (see also [11], [28], [30], [37], [58], and others). Several years after Davisson's paper [27] it was observed (first by Gallager [45], and then independently by Davisson and Leon-Garcia [29], Ryabko [96], and others) that the minimax and the maximin solutions are equivalent, i.e., $R_n^+ = R_n^- = C_n$. Furthermore, the mixture Q_{w^*}, where w^* is the *capacity-achieving prior* (i.e., $I_{w^*}(\Theta; X^n)/n = C_n$), is both minimax and maximin optimal. This result is referred to as the *redundancy-capacity theorem* of universal coding.

The capacity C_n, therefore, measures the "richness" of the class of sources. It should be pointed out, though, that C_n is not very sensitive to "distances" among the sources in the class, but rather to the effective number of essentially distinct sources. For example, the source P_1 that generates 0's only with probability one is at infinite divergence-distance from the source P_2 that generates 1's only. Yet their mixture $\frac{1}{2}P_1 + \frac{1}{2}P_2$ (in the level of n-tuples) is within normalized divergence of $1/n$ from both, and so, the capacity of $\{P_1, P_2\}$ is very small. It is a remarkable fact that the theory of universal coding is so intimately related to that of channel capacity. Moreover, the importance and significance of the redundancy-capacity theorem are fairly deep also in the broader context of probability assignment and prediction.

On the face of it, at this point the problem of universal probability assignment, or equivalently, universal prediction under the self-information loss function with respect to an indexed class of sources, is fairly well addressed. Nonetheless, there are still several important issues to be considered.

The first concern comes from a practical aspect. Explicit evaluation of the proposed minimax/maximin probability assignment is not trivial. First of all, the capacity-achieving prior w^* is hard to evaluate in general. Furthermore, even when it can be computed explicitly, the corresponding mixture Q_{w^*} as well as the induced conditional probabilities $Q_{w^*}(x_t|x^{t-1})$ might still be hard to compute. This is in contrast to the plug-in approach, which is relatively easy to implement. Nevertheless, we shall return later to the earlier example of the mixtures of Bernoulli sources, or more generally, finite-alphabet memoryless sources, and see that fortunately enough, some satisfactory approximations are available.

The second technical point has to do with the evaluation of capacity, or at least, its asymptotic behavior, which is of crucial importance. As mentioned earlier, the capacity measures the "complexity" or "richness" of the class of sources, and $C_n \to 0$ if and only if uniform redundancy rates are achievable (i.e., strong universality). This means that if the class of sources is too rich so that C_n does not vanish as n grows without bound, one can no longer hope for uniformly small redundancy rates [48], [107]. We shall see examples of this later.

Another problem that calls for attention is that the predictor, or the sequential probability assignment mechanism that we are proposing here, is not really sequential in the sense that the horizon n must be prescribed in advance. The reason is that the capacity-achieving prior w^* depends on n, in general. A possible remedy (both to this and to the problem of computability) is to seek a fixed prior w, independent of n, that achieves capacity at least asymptotically, i.e.,

$$\lim_{n\to\infty} I_w(\Theta; X^n)/(nC_n) = 1.$$

Fortunately, this is possible in some important examples.

Finally, we mentioned earlier that the minimax approach is pessimistic in essence, a fact which seems to be of special concern when $R_n^+ = C_n$ does not tend to zero as n grows. The reason is that although $D_n(P_\theta \| Q_{w^*}) \le nC_n$ for all θ, minimaxity guarantees that the lower bound

$$D_n(P_\theta \| Q) \ge nC_n \qquad \forall Q \tag{15}$$

is valid for *one* source P_θ in the class. The maximin point of view tells us further that this holds true also in the sense of the weighted average of $D_n(P_\theta \| Q)$ over θ with respect to w^*. Still, the optimality of Q_{w^*} is on seemingly somewhat weak grounds. Nonetheless, a closer inspection reveals that the right-hand side of (15) is essentially a lower bound in a much stronger sense which will now be discussed.

A Strong Converse Theorem: It turns out that in the self-information loss case, there is a remarkable "concentration" phenomenon: It is shown in [76] that

$$D_n(P_\theta \| Q) \ge (1-\epsilon)nC_n \qquad \forall Q \tag{16}$$

for every $\epsilon > 0$ and for w^*-*most* values of θ. Here, the term "w^*-most" means that the total probability mass of points with this property, with respect to w^* (or any asymptotically good approximation of w^*), tends to unity as $n \to \infty$. This means that if the right-hand side of (15) is slightly reduced, namely, multiplied by a factor $(1-\epsilon)$, it becomes a lower bound for w^*-most values of θ. Referring again to the uniform upper bound, this means that w^*-most sources in the class lie near the surface of a "sphere" (in the divergence sense) of radius nC_n, centered at Q_{w^*}. Considering the fact that we have assumed virtually nothing about the structure of the class of sources, this is quite a surprising phenomenon. The roots of this are explained and discussed in detail in [39] and [76] in relation to the competitive optimality property of the self-information function [20] (see also [61]).

There is a technical concern, however: for a class of finite-alphabet sources and any finite n, the capacity-achieving prior must be discrete with support of at most A^n points in Λ [44, p. 96, Corollary 3]. Strictly speaking, the measure w^* then ignores all points outside its support, and the term "w^*-most sources" is not very meaningful. Again, fortunately enough, in most of the important examples, one can find a smooth weight function w, which is independent of n and asymptotically achieves capacity. This solves both this difficulty and the horizon-dependency problem mentioned earlier. As an alternative remedy, there is another, more general version of this strong converse result [39], which allows for an arbitrary weight function w. It tells that Q_w is optimal for w-most points in Λ. But note that $D(P_\theta \| Q_w)$ may depend on θ for a general w, and so the uniformity property might be lost.

The above result is, in fact, a stronger version of the redundancy-capacity theorem, as detailed in [76], and it generalizes the well-known strong converse to the universal coding theorem due to Rissanen [90] for a smooth parametric family $\{P_\theta\}$ whose capacity behaves like $C_n \sim (k/2n) \log n$, where k is dimension of the parameter vector. Rissanen, in his award-winning paper [90], was the first to show such a strong converse theorem that applies to most sources at the same time. The reader is referred to [76] (see also [39]) for detailed discussion on this theorem and its significance in general, as well as in the perspective of Rissanen's work in particular. Let us now examine a few examples in light of these findings.

Examples: Perhaps the simplest example is the one where $\Lambda = \{1, 2, \cdots, N\}$, namely, there are N sources $P_1, \cdots, P_N$ in the class, and the weight function w is represented by a vector $(w_1, \cdots, w_N)$ of nonnegative numbers summing to one. In this case, the above described "concentration" phenomenon becomes even sharper [44, Theorem 4.5.1], [45] than in the general case because $D(P_i \| Q_{w^*}) = nC_n$ for every i for which $w_i^* > 0$. In other words, w^**-all* sources lie *exactly* on the surface of the divergence sphere around Q_{w^*}. If the sources $\{P_i\}$ are easily distinguishable in the sense that one can reliably identify which one of the sources generated a given vector X^n, then the redundancy-capacity of the class is nearly $\log N/n$, because the "channel input" i can be "decoded" from the "channel output" X^n with small error probability. In this case, w^* tends to be uniform over $\{1, 2, \cdots, N\}$ and the best mixture Q_{w^*} is essentially a uniform mixture. If the sources are not easily distinguishable, then the redundancy-capacity is smaller. This can be thought of as a situation where the "channel" is more "noisy," or alternatively, that the effective number of distinct sources is smaller than N. In the extreme case, where $P_1 = P_2 = \cdots = P_N$, we have $C_n = 0$ as expected, since we have, in fact, only one source in the class.

Let us now revisit the Bernoulli example, or more generally, the class of memoryless sources with a given finite alphabet of size A. This is obviously a parametric class whose natural parameterization by θ is given by the letter probabilities with $A - 1$ degrees of freedom. As mentioned earlier, w^* is discrete in the finite-alphabet case, it depends on the horizon n, and it is difficult to compute. It turns out that for smooth parametric families with a bounded parameter set Λ, like the one considered here, there is no much sensitivity to the exact shape of w (used for Q_w) as long as it is bounded away from zero across Λ. In fact, any such "nice" prior essentially achieves the leading term of the capacity, which is $\frac{A-1}{2n} \log n$. Differences in performance for different choices of w are reflected in higher order terms. Specifically, Clarke and Barron [15], [16] have derived a very accurate asymptotic formula for the redundancy associated with a mixture w

$$D_n(P_\theta \| Q_w) = \frac{A-1}{2} \ln \frac{n}{2\pi e} + \ln \frac{|I(\theta)|^{1/2}}{w(\theta)} + o(1) \quad (17)$$

where $|I(\theta)|$ is the determinant of the Fisher information matrix of $\{P_\theta\}$ (see also Takeuchi and Barron [111] for extensions to more general exponential families). In the maximin setting, the weighted average of $D_n(P_\theta \| Q_w)$ is then asymptotically maximized (neglecting the $o(1)$ term) by a prior w that maximizes the second term above, which is well known as *Jeffreys' prior* [7], [16], [57], [92]

$$w_J(\theta) = \frac{|I(\theta)|^{1/2}}{\int_\Lambda |I(\theta')|^{1/2} \, d\theta'}. \quad (18)$$

In our case, $|I(\theta)|$ is inversely proportional to the square root of the product of all letter probabilities,

$$\sqrt{\prod_{i=1}^{A} \theta_i}.$$

This, in turn, is a special case of the Dirichlet prior [63], whose general form is proportional to the product of arbitrary fixed powers of $\{\theta_i\}$. Dirichlet mixtures Q_w and conditional probabilities derived from them have easy closed-form expressions as well. Generalizing the earlier Bernoulli example to the size-A alphabet parametric family, and using Jeffreys' prior, we get the universal probability assignment

$$Q_{w_J}(x_t = j | x^{t-1}) = \frac{t_j + 1/2}{(t-1) + A/2} \quad (19)$$

where t_j is the number of occurrences of $x_\tau = j$, $1 \le \tau \le t - 1$. The uniform prior that leads to the Laplace estimator discussed earlier, is yet another special case of the Dirichlet prior. It should be noted that Jeffreys' prior asymptotically achieves capacity and so, it induces an asymptotically maximin probability assignment. Interestingly, as observed in [122], it is not asymptotically minimax, and it should be slightly modified to obtain minimax optimality. These results extend to more general parametric families under certain regularity conditions detailed in the above cited papers.

But the main point to be remembered here is that for parametric classes, the choice of w is not crucial in terms of performance. This gives rise to the freedom of selecting a prior from implementational considerations, i.e., the availability of closed-form expressions for mixtures, namely, conjugate priors [35]. We have just seen the example of the Dirichlet prior in classes of memoryless sources. As another example, consider the case where $\{P_\theta\}$ is a family of Gaussian memoryless sources with mean θ and variance 1. Clearly, Q_w with respect to a Gaussian prior w is Gaussian itself in this case. The idea of conjugate priors carries over in a natural manner to more general exponential families.

It should be pointed out that there are other recent extensions [51], [53], [54], [74], [83] of the redundancy-capacity theory to more abstract classes of sources whose capacities are proportional to k, where the number k is attributed a more general notion of dimensionality that is induced by the Hellinger distance, the Kullback–Leibler distance, the Vapnik–Chervonenkis (VC) dimension, etc. Other extensions to wider classes of sources exhibit different behavior of the redundancy-capacity [25], [123]. Still, the general underlying information-theoretic principle remains the same; the richness of the class is measured by its Shannon capacity. Other examples of classes of sources that are not necessarily parametric, are given in [39] and [76].

2) General Loss Functions: It turns out that satisfactory solutions to the universal prediction problem under the self-information loss function, may prove useful for more general loss functions. Intuitively, under suitable continuity conditions, an optimal predictor with respect to l, based on a good estimator of $P_\theta(x_t|x^{t-1})$, should be close to optimum under the true conditional probability. Generally speaking, since minimum self-information loss probability assignments are essentially maximum-likelihood estimates (cf. Section II), which are statistically consistent in most situations, this requirement is satisfied.

Specifically, in the discrete alphabet case, let P_θ denote the underlying source and consider the universal probability assignment $Q = Q_{w^*}$ for which $D_n(P_\theta \| Q) \leq nC_n$ for all $\theta \in \Lambda$. Using Pinsker's inequality (see, e.g., [24, Ch. 3, Problem 17]) and the concavity of the square root function, we have (20) shown at the bottom of this page. Now, for a general loss function l, let

$$b_t^\theta(x^{t-1}) = \arg\min_b E_\theta\{l(b, X_t)|X^{t-1} = x^{t-1}\} \tag{21}$$

where E_θ denotes expectation with respect to P_θ, and

$$b_t^u(x^{t-1}) = \arg\min_b E_Q\{l(b, X_t)|X^{t-1} = x^{t-1}\} \tag{22}$$

where E_Q denotes expectation with respect to Q. Assume that l is nonnegative and bounded by some constant $L > 0$. Then, by the inequality above, we get

$$\begin{aligned}
&E_\theta\left\{\frac{1}{n}\sum_{t=1}^n l(b_t^u, X_t)\right\} - E_\theta\left\{\frac{1}{n}\sum_{t=1}^n l(b_t^\theta, X_t)\right\} \\
&\quad = \frac{1}{n}\sum_{t=1}^n \sum_{x^{t-1}} P_\theta(x^{t-1}) \sum_{x_t} P_\theta(x_t|x^{t-1}) \\
&\qquad \cdot [l(b_t^u, x_t) - l(b_t^\theta, x_t)] \\
&\quad \leq \frac{1}{n}\sum_{t=1}^n \sum_{x^{t-1}} P_\theta(x^{t-1}) \sum_{x_t} \\
&\qquad \cdot [Q(x_t|x^{t-1}) + |P_\theta(x_t|x^{t-1}) - Q(x_t|x^{t-1})|] \\
&\qquad \cdot [l(b_t^u, x_t) - l(b_t^\theta, x_t)] \\
&\quad \leq \frac{1}{n}\sum_{t=1}^n \sum_{x^{t-1}} P_\theta(x^{t-1}) \sum_{x_t} |P_\theta(x_t|x^{t-1}) \\
&\qquad - Q(x_t|x^{t-1})|[l(b_t^u, x_t) - l(b_t^\theta, x_t)] \\
&\quad \leq \frac{L}{n}\sum_{t=1}^n \sum_{x^{t-1}} P_\theta(x^{t-1}) \sum_{x_t} |P_\theta(x_t|x^{t-1}) - Q(x_t|x^{t-1})| \\
&\quad \leq L\sqrt{2C_n \ln 2}.
\end{aligned} \tag{23}$$

In words, the optimum predictor with respect to the universal probability assignment Q_{w^*} is within $L\sqrt{2C_n \ln 2}$ close to optimum simultaneously for every $\theta \in \Lambda$. The important conclusion from this result is the following: *The existence of universal predictors with uniformly rapidly decaying redundancy rates under the self-information criterion, is a sufficient condition for the existence of such predictors for general loss functions.*

At this point, two comments are in order: first, the above assumption on boundedness of l can be weakened. For example, the leftmost side of (23), which can be thought of as a generalized divergence between P_θ and Q [75], can often be upper-bounded in terms of the variational distance between P_θ and Q. We have adopted, however, the boundedness assumption to simplify the exposition. The second comment is that the upper bound of (23) might not be tight since the true redundancy rate could be faster in certain situations. For example, minimum mean-square error, fixed-order, universal linear predictors [26], [90] have redundancy rates as small as $O(\log n/n)$, whereas the above upper bound gives $O(\sqrt{\log n/n})$. The question that arises now is whether we can provide a more precise characterization of achievable redundancy rates (tight upper and lower bounds) with respect to general loss functions.

A natural way to handle this question is to take the minimax–maximin approach similarly to the self-information loss

$$\begin{aligned}
\sqrt{C_n} &\geq \sqrt{\frac{1}{n} D_n(P_\theta\|Q)} \\
&= \sqrt{\frac{1}{n}\sum_{t=1}^n \sum_{x^{t-1}} P_\theta(x^{t-1}) \sum_{x_t} P_\theta(x_t|x^{t-1}) \log \frac{P_\theta(x_t|x^{t-1})}{Q(x_t|x^{t-1})}} \\
&\geq \sqrt{\frac{1}{2\ln 2} \cdot \frac{1}{n}\sum_{t=1}^n \sum_{x^{t-1}} P_\theta(x^{t-1}) \left[\sum_{x_t \in \mathcal{X}} |P_\theta(x_t|x^{t-1}) - Q(x_t|x^{t-1})|\right]^2} \\
&\geq \frac{1}{\sqrt{2\ln 2}} \cdot \frac{1}{n}\sum_{t=1}^n \sum_{x^{t-1}} P_\theta(x^{t-1}) \sum_{x_t} |P_\theta(x_t|x^{t-1}) - Q(x_t|x^{t-1})|.
\end{aligned} \tag{20}$$

case. The minimax predictor $\{b_t\}$ is the one that minimizes

$$\sup_{\theta\in\Lambda} E_\theta\left\{\frac{1}{n}\sum_{t=1}^{n}[l(b_t, x_t) - l(b_t^\theta, x_t)]\right\} = \sup_w \int_\Lambda dw(\theta) E_\theta\left\{\frac{1}{n}\sum_{t=1}^{n}[l(b_t, x_t) - l(b_t^\theta, x_t)]\right\}. \quad (24)$$

Unfortunately, there is no known closed-form expression for the minimax predictor for a general loss function. Nonetheless, game-theoretic arguments tell us that sometimes the minimax problem is equivalent to the maximin problem. Analogously to the self-information loss case, the maximin problem is defined as the supremum of

$$\inf_{\{b_t\}} \int_\Lambda dw(\theta) E_\theta\left\{\frac{1}{n}\sum_{t=1}^{n}[l(b_t, X_t) - l(b_t^\theta, X_t)]\right\} = \inf_{\{b_t\}} E_{Q_w}\left\{\frac{1}{n}\sum_{t=1}^{n} l(b_t, X_t)\right\} - \int_\Lambda dw(\theta)\overline{U}_n(P_\theta) \quad (25)$$

over all nonnegative weight functions $w(\cdot)$ that integrate to unity. In general, the minimax and maximin problems are well known to be equivalent for convex–concave cost functions [95]. In our case, since (25) is always affine and hence concave in w, the remaining condition is that the set of allowable predictors is convex, and that $E_\theta\{\sum_{t=1}^n l(b_t, X_t)\}$ is convex in $\{b_t\}$ for every θ. The latter condition holds, for example, if $l(b, x) = |b - x|^\alpha$, $\alpha \geq 1$.

The maximin-optimal predictor is clearly the one that minimizes $E_{Q_w}\{l(b, X_t)|X^{t-1} = x^{t-1}\}$ for the worst case choice of w, i.e., the one that maximizes

$$\overline{U}_n(Q_w) - \int_\Lambda dw(\theta)\overline{U}_n(P_\theta). \quad (26)$$

In general, the maximizing w may not agree with the capacity-achieving prior w^* that has been defined for the self-information loss case. Nonetheless, similarly as in (22), these minimax–maximin considerations again justify the approach of Bayes-optimal prediction with respect to a mixture of $\{P_\theta\}$. It should be pointed out that in certain cases (e.g., the parametric case), prediction performance is not sensitive to the exact choice of w.

By definition, vanishingly small minimax redundancy rates guarantee uniform convergence to the Bayes envelope. However, unlike the self-information loss case, for a general loss function, there is not necessarily a "concentration phenomenon" where w-most points of Λ lie at nearly the same redundancy level. For example, in the Bernoulli case with $l(b, x)$ being the Hamming distance between b and x [77] there are only two optimal predictors: one predicts always "1" and the other predicts always "0," according to whether $\Pr\{x_t = 1\}$ is smaller or larger than $1/2$. Thus it is easy to find a zero-redundancy predictor for one half of the sources in the class, and hence there cannot be a nontrivial lower bound on the redundancy that applies to most sources. Nevertheless, by using the concept of exponential weighting, in some cases it is possible to derive strong lower bounds that hold for w-most points in Λ at the same time.

Specifically, let us assume that b is an estimate of x, the subtraction operation $x - b$ is well-defined, and that the loss function is of the form $l(b, x) = \rho(x - b)$, where the function $\rho(z)$ is monotonically increasing for $z > 0$, monotonically decreasing for $z < 0$, and $\rho(0) = 0$. We next derive a lower bound on

$$E_\theta\left\{\frac{1}{n}\sum_{t=1}^{n}\rho(X_t - b_t(X^{t-1}))\right\}$$

which holds for w^*-most points in Λ, and for any predictor $\{b_t\}$ that does not depend on θ. This will extend the lower bound on universal minimum mean-square error prediction of Gaussian autoregressive moving average (ARMA) processes given by Rissanen [90].

We assume that $\rho(\cdot)$ is sufficiently "steep" in the sense that $\int e^{-s\rho(z)}\,dz < \infty$ for every $s > 0$, and define the log-moment generating function

$$\psi(s) = -\log\left[\int e^{-s\rho(z)}\,dz\right], \qquad s > 0 \quad (27)$$

and

$$\phi(d) = \inf_{s>0}[sd - \psi(s)], \qquad d > 0. \quad (28)$$

The function $\phi(d)$ can be interpreted as the (differential) entropy associated with the probability function

$$q_s(z) = e^{-s\rho(z)+\psi(s)}$$

where s is tuned so that $E_s\rho(Z) = d$, E_s being the expectation operation with respect to q_s. For a given predictor $\{b_t\}$, consider the following probability assignment:

$$Q(x^n) = \int_0^\infty ds\,\nu(s)\prod_{t=1}^{n} q_s(x_t - b_t(x^{t-1})) \quad (29)$$

where $\nu(\cdot)$ is a locally bounded away from zero "prior" on s. According to [103], $-\log Q(x^n)$ can be approximated as follows:

$$\begin{aligned}-\log Q(x^n) &= n\cdot\inf_{s>0}\left[s\cdot\frac{1}{n}\sum_{t=1}^{n}\rho(x_t - b_t(x^{t-1})) - \psi(s)\right] + \frac{1}{2}\log n + R(x^n)\\ &= n\cdot\phi\left(\frac{1}{n}\sum_{t=1}^{n}\rho(x_t - b_t(x^{t-1}))\right) + \frac{1}{2}\log n + R(x^n) \quad (30)\end{aligned}$$

where $R(x^n)$ is a small remainder term. If $E_\theta R(X^n) = O(1)$ for all θ, then following the strong converse of the self-information loss case (16), we have that for w^*-most points of Λ

$$\begin{aligned}\frac{1}{n}E_\theta\{-\log Q(X^n)\} &= E_\theta\phi\left(\frac{1}{n}\sum_{t=1}^{n}\rho(X_t - b_t(X^{t-1}))\right) + \frac{\log n}{2n} + O\left(\frac{1}{n}\right)\\ &\geq \frac{H_n(P_\theta)}{n} + (1-\epsilon)C_n \quad (31)\end{aligned}$$

for every $\epsilon > 0$ and all sufficiently large n. Since $\phi(\cdot)$ is concave ($\bigcap$), interchanging the order between the expectation operator and the function ϕ would not decrease the expression on the right-hand side of the first line of (31), and so

$$\phi\left(E_\theta\left\{\frac{1}{n}\sum_{t=1}^{n}\rho(X_t - b_t(X^{t-1}))\right\}\right) \geq \frac{H_n(P_\theta)}{n} + (1-\epsilon)C_n - \frac{\log n}{2n} - O\left(\frac{1}{n}\right) \quad (32)$$

for every $\epsilon > 0$, n sufficiently large, and w^*-most $\theta \in \Lambda$. Since ϕ is monotonically nondecreasing, this gives a lower bound on

$$E_\theta\left\{\frac{1}{n}\sum_{t=1}^{n}\rho(X_t - b_t(X^{t-1}))\right\}.$$

The above lower bound is not always tight. Evidently, tightness depends on whether the above defined Q also satisfies the reverse inequality in (31) for some predictor. This, in turn, is the case whenever the self-information lower bound is achievable by universal *predictive coding*, which models the prediction error $e_t = x_t - b_t(x^{t-1})$ as a memoryless process with q_s being the marginal for some $s > 0$. Referring to the case where $C_n \to 0$, the above bound is nontrivial if $\phi(\overline{U}(P_\theta)) = \overline{H}(P_\theta)$, the entropy rate of P_θ. When this is the case, our lower bound suggests a converse to the previous statement on conditions for uniform redundancy rates: *The existence of universal predictors with uniformly rapidly decaying redundancy rates under the self-information criterion (i.e., $C_n \to 0$), is a necessary condition for the existence of such predictors for general loss functions.* In summary, under suitable regularity conditions, there is a uniform redundancy rate for a general l, *if and only if* there is one for the self-information loss function. Furthermore, even if $\phi(\overline{U}(P_\theta)) = \overline{H}(P_\theta)$, there is another requirement for the bound to be nontrivial, which is $C_n > \frac{\log n}{2n}$. Indeed, in the Bernoulli case, where it is possible to achieve zero redundancy for half of the sources (as mentioned earlier), $C_n \sim \frac{\log n}{2n}$ and the bound becomes meaningless.

Let us consider an important example where the above bound is useful. For $\rho(z) = z^2$, q_s is the zero-mean Gaussian density function with variance $1/(2s)$. Therefore, the log-moment generating function is given by $\psi(s) = \frac{1}{2}\ln\frac{s}{\pi}$, and the differential entropy is $\phi(d) = \frac{1}{2}\ln(2\pi e d)$. Thus we have

$$E_\theta\left\{\frac{1}{n}\sum_{t=1}^{n}(X_t - b_t(X^{t-1}))^2\right\} \geq \frac{1}{2\pi e}\exp\left\{\frac{H_n(P_\theta)}{n} + (1-\epsilon)C_n - \frac{\ln n}{2n} - O\left(\frac{1}{n}\right)\right\}. \quad (33)$$

If $\{P_\theta\}$ is the class of Gaussian ARMA(p, q) sources with driving noise of variance σ^2, then $H_n(P_\theta) = \frac{1}{2}\ln(2\pi e\sigma^2)$ and $C_n \sim (p+q+1)\ln n/(2n)$, and we further obtain

$$E_\theta\left\{\frac{1}{n}\sum_{t=1}^{n}(X_t - b_t(X^{t-1}))^2\right\} \geq \sigma^2\exp\left\{(1-\epsilon)(p+q)\frac{\ln n}{n}\right\} \geq \sigma^2\left[1 + (1-\epsilon)(p+q)\frac{\ln n}{n}\right]. \quad (34)$$

This bound has been obtained by Rissanen [90], and it is known to be tight at least in the autoregressive case [26]. Another example of a class of Gaussian sources is the one where $x_t = \theta_t + v_t$, $\{v_t\}$ being zero-mean independent and identically distributed (i.i.d.) Gaussian noise with power σ^2, and $\theta = \{\theta_t\}_{t\geq 1}$ is a deterministic signal with power,

$$\limsup_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\theta_t^2$$

limited to S and relative bandwidth (normalized by 2π) limited to $0 \leq W \leq 1$. Here again,

$$H_n(P_\theta) = \frac{1}{2}\ln(2\pi e\sigma^2)$$

for every θ, but now

$$C_n = \frac{W}{2}\ln\left(1 + \frac{S}{\sigma^2 W}\right) + o(1)$$

the capacity of the band-limited Gaussian channel, which gives

$$\liminf_{n\to\infty} E_\theta\left\{\frac{1}{n}\sum_{t=1}^{n}(X_t - b_t(X^{t-1}))^2\right\} \geq \sigma^2\exp\left\{(1-\epsilon)W\ln\left(1 + \frac{S}{\sigma^2 W}\right)\right\} = \sigma^2\left(1 + \frac{S}{\sigma^2 W}\right)^{(1-\epsilon)W}. \quad (35)$$

As for achievability of the above bound, recall that the corresponding universal probability assignment problem is solved by the mixture Q_w with respect to the capacity-achieving input which is Gaussian, and therefore Q_w itself is Gaussian. When Q_w is in turn factored to a product of $Q_w(x_t|x^{t-1})$, each one of these conditional densities is again a Gaussian density, whose exponent depends only on $(x_t - b_t(x^{t-1}))^2$, where $b_t(\cdot)$ is a linear predictor, and the asymptotic variance is given by

$$\exp\left\{\frac{1}{2\pi}\int_0^{2\pi}\ln(F(\omega) + \sigma^2)\,d\omega\right\}$$

$F(\omega)$ being the power spectral density of the capacity-achieving input process. It can be shown (using techniques similarly as in [41]) that this Bayesian linear predictor asymptotically attains the above bound.

Another approach to derivation of lower bounds on performance of universal schemes has been proposed in the broader context of the multi-armed bandit problem [1], [2], [64], [108]. In this line of work, tight upper and lower bounds on redundancy rates have been given for a class of *uniformly good* schemes in the sense of adapting to the underlying source. However, these results are confined to the case where Λ is a finite set.

B. Very Large Classes of Sources

So far we have discussed classes of sources where there exists a uniform redundancy rate, which is given in terms of the capacity C_n, at least in the self-information loss case. The capacity may or may not tend to zero as $n \to \infty$, but even if it does not, the predictive self-information performance, or the compression ratio of the corresponding universal code, $H_n(\theta)/n + C_n$, might still be less than $\log A$ (where A is the alphabet size) for all $\theta \in \Lambda$, provided that n is sufficiently large. This means that *some* degree of compression (or nonuniform probability assignment) is still achievable for all sources at the same time, although there is no longer hope to approach the entropy for every θ.

In this section, we focus on much wider classes of sources where even this property does no longer exist. These classes are so rich that, in the self-information loss case, for every finite n and every predictive probability assignment Q, there exists a source in the class such that

$$E\{-\log Q(X^n)\} \geq n \log A - o(n).$$

In other words, there is a total "breakdown" in terms of self-information loss performance, and similar behavior with other loss functions. This happens, for instance, with the class of all stationary and ergodic sources [56], [106], [107] the class of all finite-order Markov sources (without limiting the order), and many other classes that can be represented as infinite unions of nested index sets $\Lambda_1 \subset \Lambda_2 \subset \cdots$. Nonetheless, universal schemes that approach the entropy rate, or more generally, the asymptotic Bayes envelope, may still exist if we do not insist on uniform redundancy rates. In other words, *weakly universal* schemes [27] are sometimes available. For example, the Lempel–Ziv algorithm (and hence also the predictive probability assignment that it induces [65]) is weakly universal over the class of all stationary and ergodic sources with a given finite alphabet [126]. Necessary and sufficient conditions for the existence of weak universality can be found in [60].

One straightforward observation that we can now make from an analysis similar to that of (23), is that a sufficient condition for the existence of a weakly universal predictor for a general (bounded) loss function is the existence of such predictor for probability assignment in the self-information case. Thus the predictive probability assignment with respect to the self-information loss function is again of crucial importance. In view of this fact, the fundamental problem, in this context, is that of estimating conditional probabilities.

Cover [18] has raised the question whether it is possible to produce consistent estimates of conditional probabilities with $|b_t(X_t = x|X^{t-1}) - P(X_t = x|X^{t-1})| \to 0$ almost surely as $t \to \infty$. Bailey [6] gave a negative answer to this question (see also Ryabko [98, Proposition 3]), but pointed out a positive result (Orenstein [85]) to a similar question. It states that for a two-sided stationary binary process, it is possible to estimate the value of $P(X_0 = x|X_{-1}, \cdots, X_{-t})$ strongly consistently as $t \to \infty$. The proposed estimates are based on finite-order Markov approximations where the order depends on the data itself. A similar estimator for $P(X_t = x|X^{t-1})$ turns out to converge to the true value in the $L^1(P)$ sense, which is weaker than the almost sure sense. This estimator has been shown by Bailey [6] to give

$$\frac{1}{n}\sum_t \log[P(X_t|X^{t-1})/b_t(X_t|X^{t-1})] \to 0$$

almost surely as $n \to \infty$. Algoet [3] gave an extension of Orenstein's results to more general alphabets, which was later simplified by Morvai *et al.* [80]. In a more recent paper, Morvai *et al.* [81] have simplified the estimator (which is based on empirical averages) for the finite-alphabet case, at the expense of losing the strong consistency property. Their estimator is consistent in the self-information sense, i.e., for every stationary P

$$\lim_{t\to\infty} E\left\{\log \frac{P(X_0|X_{-1}, X_{-2}, \cdots)}{b_t(X_0|X_{-1}, \cdots, X_{-t})}\right\} = 0 \qquad (36)$$

which implies consistency in the $L^1(P)$ sense.

Another line of research work concentrates on the square-error loss function. Since the minimum mean-square-error predictor for a known source is the conditional mean

$$b_t(x^{t-1}) = E\{X_t|X^{t-1} = x^{t-1}\}$$

most of the work in this direction focuses on consistent estimation of the conditional mean. For Gaussian processes with unknown covariance function, Davisson [26] has shown that a kth-order linear predictor, based on empirical covariances gives asymptotic cumulative mean-square error that behaves like $\sigma^2(k)(1+k \ln n/n)$, where $\sigma^2(k)$ is the residual error of optimal kth-order linear prediction with known covariances. Thus by letting k grow sufficiently slowly with time, the conditional mean, given the infinite past, can be eventually attained. For general stationary processes, Scarpellini [102] used sample averages with certain spacing between time instants in order to estimate $E\{X_k|X_0, X_{-1}, \cdots\}$ where $k > 0$ is a fixed time instant. Modha and Masry [79] considered mixing processes and proposed an estimator based on slow increase of the prediction memory, using complexity regularization methods. The limitation of their method is that it depends on knowledge of the mixing rate. Meir [73] proposed a complexity regularization method in the same spirit, where for a given complexity, the class of allowable predictors is limited by a finite Vapnik–Chervonenkis (VC) dimension.

Finally, for a general loss function l, Algoet [4] (see also [5] for the special case of log-optimum investment) has proved strong ergodic theorems on the cumulative loss. First, for a known stationary and ergodic source, it is shown that the strategy that minimizes the conditional mean of $l(b, X_t)$ given the past, is also optimal in the almost sure (and L^1) limit of the time-average loss. When P is unknown, empirical estimates of the conditional probability are provided. By plugging in these estimates instead of the true P, universal schemes are obtained with the same ergodic property as above.

IV. The Deterministic Setting

In the traditional, probabilistic setting of prediction, that was described in the previous section, one assumes that the data are generated by a mechanism that can be characterized in statistical terms, such as a memoryless source, Markov source, or more generally, an arbitrary stationary and ergodic source. As we have seen, the observer estimates on-line either explicitly (plug-in approach) or implicitly (mixture approach) the conditional probability of the next outcome given the past, and then uses this estimate for prediction of future outcomes.

But when it comes to the deterministic setting of individual data sequences, the underlying philosophy must be substantially different. There is no longer an assumption of an ensemble of sequences generated by an underlying probabilistic mechanism, but rather only one arbitrary, deterministic, individual sequence. What is the best prediction strategy that one can possibly use for this fixed sequence?

We realize that, as stated, this question is completely trivial and meaningless. As explained in Section I, formally, for any sequence, there is a perfect predictor that suffers zero loss along this particular sequence. But at the same time, this particular predictor might be extremely bad for many other sequences. Evidently, we are over-tailoring a predictor to one particular sequence, and there is no hope to track the strategy of this predictor in the sequential regime that is inherent to the task of prediction. The root of this "overfitting" effect lies in the fact that we allowed, in the above discussion, too much freedom in the choice of the predictor. Loosely speaking, so much freedom that the amount of information treasured in the *choice* of this predictor is as large as the amount of information conveyed by the sequence itself! Roughly speaking, in these situations the algorithm "learns the data by heart" instead of performing the task we expect. The unavoidable conclusion is that we must limit the freedom of the choice of predictors to a certain class. This limited class of allowable predictors will be henceforth referred to as the *comparison class* (or target class) and will be denoted by B.

We would like to have a *single* universal predictor b_t^* that competes with the best predictor in B, simultaneously for every x^n, in the sense that

$$\frac{1}{n}\sum_{t=1}^{n} l(b_t^*, x_t)$$

is asymptotically the same as

$$\min_B \frac{1}{n}\sum_{t=1}^{n} l(b_t, x_t).$$

The universal predictor need not be necessarily in B but it must be the same predictor for every x^n, whereas the choice of the reference predictor in B, that minimizes the average loss, may depend (by definition) on the entire sequence x^n. The difference between the performance of the sequential universal predictor and the best predictor in B for x^n actually manifests our *regret*, because the choice of this optimal predictor is the best we could have done in retrospect within B had we known the entire sequence in advance.

Loosely speaking, there is a fairly strong duality between the probabilistic and the deterministic setting. While in the former, we make certain assumptions and limitations on the data sequences that we are likely to encounter, but no prior limitations on the class of prediction algorithms, in the latter, it is the other way around. Yet, the deterministic setting is frequently considered stronger and more appealing, because the underlying model seems to be better connected to practical situations: There is no (known) probabilistic mechanism that generates the data, but on the other hand, our algorithmic resources are, after all, limited.

Perhaps one of the facts that shed even more light on this duality between the probabilistic and the deterministic setting, is that quite frequently, the comparison class B is defined as a collection of predictors that are obtained as optimal solutions for a certain class of sources in the parallel probabilistic setting. For example, fixed predictors, where $b_t(x^{t-1})$ is a constant independently of x^{t-1}, are optimal for memoryless stationary sources, linear predictors are sufficient for the Gaussian case, Markov predictors are adequate for Markov processes, and so on. In these cases, there is a remarkable degree of duality and analogy between results obtained in the deterministic setting and those of the corresponding probabilistic setting, notwithstanding the considerable difference between the two concepts. Specifically, many of the results of the individual-sequence setting are completely analogous to their probabilistic counterparts, where the probabilistic source is replaced by the empirical measure extracted from the individual sequence with respect to certain sufficient statistics that are induced by B. Indeed, the structure of this section is similar to that of the previous section, so as to manifest this analogy. Nonetheless, there are still certain aspects in which the two scenarios diverge from each other, as we shall see later.

Similarly as in the previous section, our emphasis here is on the information-theoretic point of view, and as such, it again largely focuses on the self-information loss function.

A. Indexed Comparison Classes

In analogy to the indexed class of sources, that was extensively discussed in the previous section on the probabilistic setting, there has been considerable attention in the literature to the dual comparison classes in the deterministic setting. An indexed comparison class of predictors is a class B that can be represented as $\{b^\theta, \theta \in \Lambda\}$, where θ designates the index and Λ is the index set. Similarly as in Section III-A, the index set Λ could be a finite set $\{1, \cdots, N\}$ (N—positive integer), where N may or may not grow with n, a countably infinite set, a continuum, e.g., a compact subset of the real-line or a higher dimensional Euclidean space (when θ is a parameter of a smooth parametric class), or some combination of these. As was already noted above, in many cases, b^θ could be defined as the optimum predictor for a certain member P_θ of an indexed class of sources (cf. Section III-A).

1) Self-Information Loss: In analogy to Section III, let us consider first the self-information loss function, or equivalently, the probability assignment problem for individual sequences. In other words, our goal is to sequentially assign a universal probability mass function

$$Q(x^n) = \prod_{t=1}^{n} b_t(x_t|x^{t-1}) \tag{37}$$

to the observed sequence x^n, so that $-\frac{1}{n}\log Q(x^n)$ would be essentially as small as

$$-\frac{1}{n}\log \max_\theta \prod_{t=1}^{n} b^\theta(x_t|x^{t-1})$$

for every sequence x^n, uniformly if possible.

Shtarkov [109] has demonstrated that this is indeed possible by minimizing over Q the quantity

$$\max_{x^n} \frac{1}{n}\left[-\log Q(x^n) - \left(-\log \max_\theta \prod_{t=1}^{n} b^\theta(x_t|x^{t-1})\right)\right]. \tag{38}$$

Specifically, the minimax-optimal probability assignment is attained by the normalized maximum-likelihood function

$$Q_n^*(x^n) = \frac{1}{K_n}\max_\theta \prod_{t=1}^{n} b^\theta(x_t|x^{t-1}) \tag{39}$$

where K_n is a normalization factor, i.e.,

$$K_n = \sum_{x^n}\max_\theta \prod_{t=1}^{n} b^\theta(x_t|x^{t-1}). \tag{40}$$

Indeed, it is readily seen that, by definition of Q_n^*

$$-\frac{1}{n}\log Q_n^*(x^n) = -\frac{1}{n}\log \max_\theta \prod_{t=1}^{n} b^\theta(x_t|x^{t-1}) + \frac{1}{n}\log K_n \tag{41}$$

and so, the universal probability function Q_n^* essentially assigns uniformly as high probabilities as those assigned by the best member in the comparison class, provided that K_n does not grow exponentially rapidly with n.

If, for example, $\{b^\theta\}$ is the class of finite-alphabet memoryless probability assignments (i.e., $b^\theta(x_t|x^{t-1}) = b^\theta(x_t)$) with θ designating the vector of $k = A-1$ free letter probabilities, then it is easy to show (e.g., by using the method of types [24]) that K_n grows asymptotically in proportion to $n^{k/2}$ and thus (38) behaves like $\frac{k}{2n}\log n$. This in turn is the same behavior that was obtained for smooth parametric families in the probabilistic setting.

The number $\Gamma_n = n^{-1}\log K_n$ is therefore given the interpretation of the deterministic analog to the minimax redundancy-capacity C_n, where the maximization of redundancy over θ in the probabilistic setting is now replaced by maximization over all possible sequences x^n. Intuitively, Γ_n is another measure for the richness of the comparison class of predictors, in addition to the capacity C_n of the probabilistic setting. Moreover, it turns out that there are relations between these two quantities. To demonstrate this relation between Γ_n and the operational notion of capacity as the maximum reliable transmission rate, we note that when $\Lambda = \{1, \cdots, N\}$, the quantity K_n can be interpreted as $N \cdot P_c$ where P_c is the probability of correct decision of an N-hypotheses testing problem involving the sources

$$P_i(x^n) = \prod_{t=1}^{n} b^i(x_t|x^{t-1}), \qquad 1 \le i \le N$$

that are induced by the predictors, with a uniform prior on i. This is true because

$$P_c = \frac{1}{N}\sum_{x^n}\max_i P_i(x^n) = \frac{K_n}{N}. \tag{42}$$

This means that if the sources $\{P_i\}$ are "far apart" and distinguishable with high probability, then the minimax redundancy is essentially $\log N$ (compare with the first example in Section III). If Λ is countably infinite or a continuum, then any finite subset $\{\theta_i, i = 1, \cdots, N\}$ of Λ gives a lower bound on K_n in the above manner. As N grows, P_c normally decreases, but the product NP_c can be kept large at least as long as N is smaller than 2^{nC_n} so as to "transmit" at a rate below capacity, which allows for keeping P_c close to unity. But the maximum achievable product NP_c might be achieved at rates beyond capacity.

It is easy to show directly that Γ_n is never smaller than C_n for the same class of sources or probability assignments indexed by Λ. This implies that a necessary condition for the existence of minimax universality in the deterministic setting is the existence of the parallel property in the dual probabilistic setting. In the smooth parametric case both C_n and Γ_n behave like $\frac{k}{2n}\log n$. More precisely (see, e.g., Rissanen [92])

$$C_n = \frac{k}{2n}\log\frac{n}{2\pi e} + \frac{1}{n}\log\int_\Lambda |I(\theta)|^{1/2}\,d\theta + o\left(\frac{1}{n}\right) \tag{43}$$

whereas

$$\Gamma_n = \frac{k}{2n}\log\frac{n}{2\pi} + \frac{1}{n}\log\int_\Lambda |I(\theta)|^{1/2}\,d\theta + o\left(\frac{1}{n}\right). \tag{44}$$

It turns out, however, that richer indexed classes may exhibit a considerably larger gap between these two quantities (see, e.g., the example of arbitrarily varying sources in [76]).

The main drawback of the maximum-likelihood (ML) probability assignment Q_n^* is obviously on the practical side: not only Q_n^* is hard to compute in general, but more importantly, it is again horizon-dependent, i.e., the sequence length n must be prescribed. To alleviate this difficulty, the maximum-likelihood $\max_\theta \prod_t b_t(x_t|x^{t-1})$ can be exponentially approximated by a mixture using Laplace integration [67]. Specifically, for the case of stationary memoryless probability assignments, Shtarkov [109] proposed, following Krichevsky and Trofimov [63], the Dirichlet-$\left(\frac{1}{2}, \cdots, \frac{1}{2}\right)$ (Jeffreys' prior) mixture, which leads to the purely sequential probability assignment

$$b_t(x_t = a|x^{t-1}) = \frac{t(a) + \frac{1}{2}}{(t-1) + \frac{A}{2}} \tag{45}$$

where $t(a)$ is the number of occurrences of the letter a in x^{t-1}. We have mentioned earlier, in Section III, the family of sequential probability assignments that arise from Dirichlet weighting in general. But the interesting property of the Dirichlet-$\left(\frac{1}{2}, \cdots, \frac{1}{2}\right)$ (in addition to being Jeffreys' prior for this family), is that it is asymptotically as good as the ML probability assignment. Specifically, with $b_t(\cdot|x^{t-1})$ defined as above

$$\max_{x^n}\left[-\log\prod_{t=1}^{n} b_t(x_t|x^{t-1}) - \left(-\log\max_{\theta}\prod_{t=1}^{n} b^{\theta}(x_t|x_{t-1})\right)\right] \le \frac{k}{2}\log n + \text{Const} + o(1) \tag{46}$$

where only the constant here is larger than the one obtained by Q_n^*.

Further refinements and extensions of this result have been recently carried out, e.g., in [92] and [121]. Specifically, Xie and Barron [121] introduce also the dual notion of the maximin redundancy (or regret) whose value coincides with Γ_n as well, and show that Jeffreys' mixture is asymptotically maximin with asymptotically constant regret for sequences whose empirical pmf's are internal to the simplex. Similarly as in the probabilistic setting, it is not asymptotically minimax though because of problematic sequences on the boundary of the simplex. Nevertheless, a slight modification of Jeffreys' mixture (which again, depends on n and hence makes it again horizon-dependent), is both asymptotically minimax and maximin.

Finally, Weinberger, Merhav, and Feder [117] have studied the problem of universal probability assignment for individual sequences under the self-information loss function with respect to the comparison class of all probability assignments that are implementable by finite-state machines with a fixed number of states. There are no such accurate formulas therein regarding the higher order redundancy terms. However, it is shown that the $\frac{k}{2}\log n$ behavior is not only minimax over all sequences, but moreover, it is a tight lower bound for *most* sequences of most *types* defined with respect to those finite-state probability assignments. This result parallels the w-almost everywhere optimality of universal probability assignments in the probabilistic setting (cf. Section III). In this context, it is interesting to note, as shown in [117], that in contrast to the probabilistic setting, the plug-in approach fails, in general, when it comes to individual sequences. We will elaborate on these results further in Section V in the context of hierarchical comparison classes.

2) General Loss Functions: The problem of universal sequential prediction or decision-making for individual sequences under general loss functions, is definitely a much wider problem area than that of the special case of probability assignment under the self-information loss function that we discussed thus far in this section. In fact, most of the classical work in this problem area, in various scientific disciplines, has concentrated primarily on the case of *constant* predictors, i.e., predictors for which each b^{θ} yields a certain fixed prediction, regardless of the observed past. For example, b^{θ}, for a certain value of θ, may suggest to predict *always* "0" as the next outcome of a binary sequence, or, it may always assign a probability of 0.8 for the next outcome being "1." This is seemingly not a very interesting comparison class because past information is entirely ignored.

Nonetheless, the motivation for carefully studying this simple comparison class is that it is fundamental for examining comparison classes of more sophisticated predictors. For example, a first-order Markov predictor, characterized by $b^{\theta}(x_t|x^{t-1}) = b^{\theta}(x_t|x_{t-1})$, can be thought of (in the binary case) as a combination of two fixed predictors operating, respectively, on two subsequences of $\boldsymbol{x}^n$: the one corresponding to all time instants $\{t\}$ that follow $x_{t-1} = 0$, and the other—where $x_{t-1} = 1$. Having made this observation, the problem then boils down back to that of constant predictors.

One example, which is still closely related to the self-information, is that of portfolio selection for optimal investment in the stock market [3]–[5], [21]. In this model, the goal is to maximize the asymptotic exponential growth rate of the capital, where the current investment strategy depends on the past. The corresponding loss function, in our framework, is then $l(b, x) = -\log(b^T x)$, with both b and x being m-dimensional vectors of nonnegative components, where in the former these components sum to unity. The vector x represents the return per monetary unit in several investment opportunities (stocks), whereas the vector b characterizes the fraction of the current capital allocated to each stock. Cover [21] and Cover and Ordentlich [22] have used techniques similar to those of the self-information loss described above, to develop a sequential investment algorithm and related it again to universal coding with results of a similar flavor. Again, their universal sequential strategy competes with the best constant investment strategy. These results can be viewed as an extension of the self-information loss because the latter is actually a special case where the vector x is always all-zero except for one component (corresponding to the current alphabet letter), which is 1.

3) The Sequential-Compound Decision Problem: Other examples of loss functions are not so closely related to that of the self-information loss, and consequently, the techniques and the results are considerably different. The comparison class of constant strategies for more general loss functions has been studied in a somewhat more general setting, referred to as the *sequential-compound decision problem*, which was first presented by Robbins [94] and has been thoroughly investigated later by many researchers from disciplines of mathematical statistics, game theory, and control theory (see, e.g., [8], [9], [49], [50], and [112]). Perhaps the most fundamental findings of the compound sequential decision problem are summarized in the theory of Bayes decision rules, that includes the notion of *Bayes envelope* (that is, the best achievable target performance as a functional of the empirical pmf of the sequence) and an analysis of its basic properties. This in turn has been combined with approachability–excludability theory, that provides simple necessary and sufficient conditions under which one player (in our case, the predictor) of a repeated zero-sum game can reach a certain performance level (in our case, the Bayes envelope) for every strategy of the opponent player (in our case, Nature that chooses an adversary sequence x^n).

The sequential compound decision problem is more general than our setting in the sense that the observer is assumed to access only noisy versions of the sequence x^n, yet the loss function to be minimized is still associated with the clean sequence (e.g., the expected cumulative loss, or its probabilistic limit with respect to the ensemble of noise processes). Hannan [49] has taken a game-theoretic approach to develop upper bounds on the decay rate on the regret, showing a convergence rate of $O(n^{-1/2})$ in the finite-alphabet, finite-strategy space case, and a rate of $O\left(n^{-1}\sum_{t=1}^{n} t^{-\alpha}\right)$ in the continuous case, provided that the loss-minimizing strategy b^* as a functional of the underlying empirical pmf of x^n, that is, the *Bayes response*, satisfies a Lipschitz condition of order $\alpha > 0$. Thus for $\alpha = 1$, which is normally the case, this means a convergence rate of $\log n/n$, similarly to the self-information loss case that we have seen above.

One of the essential ideas underlying the analysis techniques, is the following simple "sandwich" argument (see, e.g., [75]): It is easy to show that

$$\min_B \frac{1}{n}\sum_{t=1}^{n} l(b_t, x_t)$$

i.e., the Bayesian envelope, is upper- and lower-bounded by the average loss associated with two strategies. The current strategy for the upper bound is optimal within B for the data seen thus far x^{t-1}, and for the lower bound, it is an (imagined) strategy that is allowed to access x^t for this optimization within B. Thus the strategy of the lower bound sees merely one more outcome than that of the upper bound. When the comparison class is that of constant strategies, the Bayes envelope depends on the sequence only through its empirical pmf, and this additional observation perturbs the current empirical pmf by a term proportional to $1/t$. Therefore, under the appropriate smoothness conditions ($\alpha = 1$ above), the instantaneous losses of the upper and lower bound differ also by a quantity that scales proportionally to $1/t$, which when averaged over the integers $1, \cdots, n$, gives $O(\log n/n)$. *A fortiori*, the difference between the upper bound and the Bayes envelope, i.e., the regret, cannot exceed $O(\log n/n)$.

In some important special cases, however, the loss function and the Bayes response are discontinuous. This happens, for example, in prediction of binary sequences under the criterion of relative frequency of mispredicted outcomes, where the Bayes response with respect to the class of constant predictors is binary itself and it depends on whether the relative frequency of zeros is below or above $1/2$. In this case, randomization of the sequential prediction strategy around the discontinuity point (see, e.g., [40], [99], and [100]) is necessary in order to achieve the target performance for problematic sequences whose empirical pmf's visit infinitely often (as $n \to \infty$) these discontinuity points. The cost of this randomization, however, is a considerable slowdown in the rate of convergence towards the Bayes envelope. In the above binary case, for example, the rate of convergence is $O(1/\sqrt{n})$, whereas in the parallel probabilistic setting, where such a randomization is not needed (cf. Section III), it is as fast as $O(1/n)$.

Van Ryzin [112] has shown that even in the former case of smooth loss functions, the convergence rate can be more tightly upper-bounded by $O(\log n/n)$ under certain regularity conditions on the channel through which the observer receives the noisy measurements. Gilliland [46] further investigated convergence rates for the special case of the square loss function $l(b, x) = (x - b)^2$ under various sets of assumptions. Several later papers [82], [113] deal with the more general case where the comparison class consists of Markov strategies, whose importance will be emphasized later.

On-Line Prediction Using Expert Advice: A completely different point of view has been taken more recently, primarily by learning theorists in their studies of a paradigm referred to as *on-line prediction using expert advice* (see, e.g., [12], [13], [36], [43], [69], [84], and [115]). In the previously defined terminology, the basic assumption is that the comparison class consists of finitely many predictors $b^1, \cdots, b^N$, referred to as *experts*. There are absolutely no assumptions on any structure or relationships among these experts. The goal is to devise a sequential universal prediction algorithm that performs essentially as well as the best of these experts along every individual sequence.

We have actually examined earlier this scenario in the context of the self-information loss function and a finite index set $\Lambda = \{1, \cdots, N\}$, where our conclusion was that the necessary minimax price of universality need not exceed $\log N/n$ in the worst case, namely, when the probability assignments b^i correspond to distinguishable sources. Interestingly, this behavior essentially continues to take place for general (but sufficiently regular) loss functions. Vovk [115] and Littlestone and Warmuth [70] proposed independently a sequential prediction algorithm, whose regret with respect to the best expert never exceeds $c_l \log N$, where c_l is a constant that depends solely on the loss function l. At the heart of this algorithm, there is a remarkable similarity to the mixture approach, or, more concretely, the notion of exponential weighting that was discussed in Section III in the special case of the self-information loss.

Here is the idea: let $\eta > 0$ be a given constant (to be chosen later) and consider the weighted average of $e^{-\eta l(b_t^i, x_t)}$, i.e.,

$$\sum_{i=1}^{N} w_t(i)\, e^{-\eta l(b_t^i, x_t)} \tag{47}$$

where b_t^i is the prediction of the ith expert at time t, and $w_t(i)$ is the *weight* assigned to this expert at this time. The weights, at each time instant, are nonnegative numbers summing to unity. Intuitively, we would like to assign higher weights to experts who were proven better in the past. Therefore, a reasonable thing to do, following (10), is to assign to each expert a weight $w_t(i)$ that is proportional to

$$\exp\left(-\eta \sum_{\tau=0}^{t-1} l(b_\tau^i, x_\tau)\right)$$

where for $t = 0$ the summation will be defined as zero (i.e., uniform initial weighting). Now, if we are fortunate enough that there exists a strategy b such that for every x

$$e^{-\eta l(b, x)} \geq \sum_{i=1}^{N} w_t(i)\, e^{-\eta l(b_t^i, x)} \tag{48}$$

then it is easy to see that this strategy will serve our purpose. This is true because the above condition suggests the following conceptually simple algorithm:

0) **Initialization:** Set $w_0(i) = 1/N$ for $1 \le i \le N$ and then $t = 1$.
1) **Prediction:** Choose a prediction b_t^* at time t that satisfies (48).
2) **Update:** Upon receiving x_t, update the weight function according to

$$w_{t+1}(i) = \frac{w_t(i)\, e^{-\eta l(b_t^i, x_t)}}{\sum_{j=1}^{N} w_t(j)\, e^{-\eta l(b_t^j, x_t)}}. \tag{49}$$

3) **Iteration:** Increment t and go to 1).

It follows immediately from the definition of the algorithm that the exponent of the cumulative loss associated with $\{b_t^*\}$ satisfies

$$\begin{aligned}\exp\left(-\eta \sum_{t=1}^{n} l(b_t^*, x_t)\right) &\ge \frac{1}{N} \sum_{i=1}^{N} \exp\left(-\eta \sum_{t=1}^{n} l(b_t^i, x_t)\right) \\ &\ge \frac{1}{N} \max_i \exp\left(-\eta \sum_{t=1}^{N} l(b_t^i, x_t)\right)\end{aligned} \tag{50}$$

and so

$$\sum_{t=1}^{n} l(b_t^*, x_t) \le \min_i \sum_{t=1}^{n} l(b_t^i, x_t) + \frac{1}{\eta} \ln N. \tag{51}$$

Thus the crucial question that remains to be addressed is regarding the conditions under which (48) is satisfied. To put this question in perspective, first, observe that for the self-information loss function and $\eta = 1$, the functions

$$\exp\left(-\eta \sum_{t=1}^{n} l(b_t^i, x_t)\right)$$

are probability measures of n-tuples. Therefore, their weighted average (mixture) is itself a probability measure and as such, can be represented by

$$\exp\left(-\eta \sum_{t=1}^{n} l(b_t^*, x_t)\right)$$

for a certain $\{b_t^*\}$, which is the probability assignment corresponding to the finite mixture. However, in general, the function $e^{-\eta l(\cdot,\cdot)}$ may not be closed to convex combinations. Fortunately, it is shown that under fairly mild regularity conditions (see [52], [115], and [116] for details), it is guaranteed that (48) always holds provided that η is chosen to be at most $1/c_l$ and that $c_l < \infty$, in which case the regret can be made as small as $c_l \ln N/n$. Many of the important loss functions, like the self-information loss and the square-error loss, satisfy these conditions. For example, if the function $e^{-\eta l(b,x)}$ is concave ($\cap$) in b for every x (which is the case in linear prediction and squared-error loss under some conditions [110]), namely,

$$\exp\left[-\eta l\left(\sum_{i=1}^{N} w_t(i) b_t^i, x\right)\right] \ge \sum_{i=1}^{N} w_t(i)\, e^{-\eta l(b_t^i, x)} \tag{52}$$

then it is clear that the weighted average of the experts' predictions will be a suitable solution. Unfortunately, there are also other important loss functions (like the L_1 loss function, $l(b,x) = |x-b|$) for which $c_l = \infty$. This means that for these loss functions, the regret does not behave like $O(\log N/n)$, but rather decays at a slower rate with n, e.g., like $1/\sqrt{n}$. These cases should be handled separately.

What makes this algorithm even more interesting is the fact that it turns out to be minimax-optimal in the sense that $c_l \ln N/n$ is also an asymptotic lower bound on the maximum regret. Unfortunately, the weak point of this lower bound is that this maximum is taken not only over all sequences $\{x^n\}$, but also over all possible sets of N experts! The algorithm is, therefore, asymptotically optimal in an extremely pessimistic sense, which is of special concern when N is large. What is left to be desired then is a stronger bound that depends on the relationships among the experts. As an extreme example, if all experts are identical then there is in fact only one expert, not N, and we would expect to obtain zero regret. Intuitively, we would like the formal number of experts N to be replaced by some notion of an "effective" number of distinct experts, in analogy and as an extension of the role played by capacity C_n or by Γ_n in the self-information loss case. To the best of our knowledge, to date, there are no reported results of this kind in the literature except for Cesa-Bianchi and Lugosi [14] who characterized the minimax regret along with upper and lower bounds for binary sequences and the Hamming loss function, but without any constructive algorithm yet.

Another drawback is associated with the algorithm itself. To use this algorithm in practice, one should actually implement in parallel the prediction algorithms proposed by all N experts, which might be computationally demanding for large N. This is in contrast to the situation in certain special cases, e.g., when the experts correspond to all finite-state machines with a given number of states [38], [40], [75], [126]. In these cases, there is no explicit implementation of all finite-state machines in parallel.

In spite of these shortcomings, the problem of on-line prediction with expert advice has attracted fairly much attention over the last few years and there are quite a few reported extensions, modifications, and other variations on the theme (see, e.g., [10] for a summary of recent work in on-line learning). One extension that would be especially interesting is to tie it with the setting of the compound sequential decision problem in the sense that the predictor accesses only noisy observations, whereas the loss function remains in terms of the clean outcomes. Clearly, the above weighting algorithm, in its present form, is not directly implementable since there is no perfect feedback on the loss associated with past expert advice.

B. Very Large Comparison Classes

We end this section with a natural analog to the case of very large classes of sources in the probabilistic setting, namely, very large comparison classes of predictors for which there are normally no uniform redundancy rates.

In the general level, consider a nested infinite sequence of index sets $\Lambda_1 \subset \Lambda_2 \subset \cdots$, and their union $\Lambda = \cup_{k \ge 1} \Lambda_k$.

Strictly speaking, Λ is itself an index set, whose members are of the form (k, θ), where k is the smallest integer such that $\theta \in \Lambda_k$. However, the basic property that makes Λ herein different than the index sets of Section IV-A is that it is so rich, that for every finite sequence x^n, the minimum cumulative loss over all predictors indexed by Λ is zero. In other words, there is too much freedom within Λ, and we are confronting again the undesirable overfitting effect discussed earlier. This happens in many important examples, e.g., when Λ consists of the class of all finite-state predictors with an undetermined (but finite) number of states, or the class of all Markov predictors, or even more specifically, all linear predictors with an unspecified finite order, etc. Quite clearly, in all these situations, there are enough degrees of freedom to tailor a perfect predictor for any finite sequence x^n, and thus our earlier definition (cf. Section IV-A) of the target performance $\min_\Lambda \sum_t l(b_t, x_t)$ becomes meaningless.

We are lead then to the conclusion that we must modify the definition of the target performance. The key principle for doing this is to keep an asymptotic regime of $n \gg k$. To fix ideas, consider an infinite sequence $\boldsymbol{x} = (x_1, x_2, \cdots)$, where x^n always designates the first n outcomes of $\boldsymbol{x}$. First, similarly as in Section IV-A, let us define

$$u_k(x^n) = \min_{\Lambda_k} \frac{1}{n} \sum_{t=1}^{n} l(b_t, x_t) \tag{53}$$

where it is assumed that each Λ_k is an index set of the type discussed in Section IV-A. As for asymptotics, we let first n grow without bound, and define

$$u_k(\boldsymbol{x}) = \limsup_{n\to\infty} u_k(x^n) \tag{54}$$

where the $\limsup$ operation manifests a worst case approach: since the sequence $\boldsymbol{x}$ is not necessarily ergodic, i.e., the limit may not exist, one must worry about the worst performance level obtained infinitely often along $\boldsymbol{x}$. Finally, we define our target performance as

$$u(\boldsymbol{x}) = \lim_{k\to\infty} u_k(\boldsymbol{x}) \tag{55}$$

where now the limit clearly exists since $\{u_k(\boldsymbol{x})\}_{k\geq 1}$ is a monotonically nonincreasing sequence whose elements are obtained from minimizations over increasing sets of predictors. Since the limit $n \to \infty$ is taken first, the asymptotic regime here indeed meets the above mentioned requirement that $n \gg k$. The problem is now to devise a universal prediction algorithm $\{b_t^*\}_{t\geq 1}$ that asymptotically achieves $u(\boldsymbol{x})$.

One of the most popular applications of this general scenario is the one where Λ consists of all strategies that are implementable by finite-state machines, which means that each Λ_S, $S = 1, 2, \cdots$, corresponds to the class of finite-state machines with no more than S states. Specifically, each member of Λ_S is defined by two functions f and g. The function g, referred to as the *next-state function*, describes the evolution of the state of the machine, $s_t \in \{1, \cdots, k\}$, according to the recursion

$$s_t = g(x_{t-1}, s_{t-1}), \qquad t = 1, 2, \cdots \tag{56}$$

where the initial state s_0 is fixed. The function f describes the strategy b_t at time t, which depends only on s_t by

$$b_t = f(s_t). \tag{57}$$

The idea behind this model is that the state variable s_t represents the limited information that the machine can "memorize" from the past x^{t-1} for the purpose of choosing the current strategy. An important special case of a finite-state machine with $S = A^k$ states is that of a kth-order Markov machine (also called finite-memory machine), where $s_t = (x_{t-k}, \cdots, x_{t-1})$.

Ziv and Lempel described, in their famous paper [126], a target performance in this spirit in the context of data compression of individual sequences using finite-state machines. The best $\limsup$ compression ratio obtained by finite-state encoders over infinitely long individual sequences (in the above defined sense) has been referred to as the *finite-state compressibility* of $\boldsymbol{x}$, and the well-known Lempel–Ziv algorithm (LZ'78) has been shown to achieve the finite-state compressibility for every sequence. In a later paper [127], Ziv and Lempel extended this definition to compression of two-dimensional arrays (images), where the additional ingredient is in defining also a scanning strategy.

In [38], results in the same spirit have been obtained for sequential gambling over individual sequences, where again the comparison class is that of gambling strategies that are implementable by finite-state machines. Since the gambling problem is completely analogous to that of data compression, or more precisely, probability assignment under the self-information loss function (see also [117] discussed in Section IV-A), the results therein are largely similar to those of Ziv and Lempel [126]. The formal setting of [38], however, is somewhat more compliant than [126] to our general definition of cumulative loss minimization, where each loss term depends on one outcome x_t only.

The results of [38] in turn provided the trigger to a later work [40], where the comparison class of finite-state predictors for binary sequences was studied under the Hamming loss function, defined as $l(b, x) = 0$ if $x = b$, and $l(b, x) = 1$ otherwise. In other words, in this case, $b_t = f(s_t)$ is simply an estimate of the value of the next outcome x_t, and the performance measure is the relative frequency of prediction errors. Analogously to [126], the quantity $u(\boldsymbol{x})$, in this special case, is called the *finite-state predictability* of $\boldsymbol{x}$. Similarly, when Λ_k is further confined to the class of kth-order Markov predictors, then the correspondingly defined $u(\boldsymbol{x})$ is called the *Markov predictability* of $\boldsymbol{x}$. There are two main conclusions pointed out in [40].

The first is that the finite-state predictability and the Markov predictability are always equivalent, which means that it is sufficient to confine attention to Markov predictors in order to achieve the finite-state predictability. It is worthwhile to note that in the probabilistic setting, such a result would have been expected under certain mixing conditions because the effect of the remote past fades away as time evolves, and only the immediate past (that is stored as the state of a Markov predictor) should be essential. Yet, when it comes to individual sequences this finding is not at all trivial since the sequence is arbitrary and there is no parallel assumption on mixing or fading memory. The proof of this result stems from pure information-theoretic considerations.

The second conclusion, which is largely based on the first one, is on the algorithmic side. It turns out that a prediction strategy that corresponds to probability assignments based on the incremental parsing procedure of the LZ algorithm (see also [65] and [114]) asymptotically achieves the finite-state predictability. The incremental parsing procedure sequentially parses a sequence into distinct phrases, where each new phrase is the shortest string that is not identical to any previously parsed phrase. The reason is that the incremental parsing procedure works like a Markov predictor of time-varying order $k(t)$, where in the long run, $k(t)$ is very large most of the time because the phrases become longer and longer. Consequently, the Markov predictability, and hence also the finite-state predictability, are eventually attained. But the deep point here lies in the simple fact that the incremental parsing algorithm, which was originally developed as a building block of a compression algorithm, serves also as the engine of a probability-assignment mechanism, which is useful for prediction.

This gives rise to the idea that this probability assignment induces a universal probability measure in the context of individual sequences. Loosely speaking, it means that the universal probability measure is proportional to $2^{-\mathrm{LZ}(x^n)}$, where $\mathrm{LZ}(x^n)$ is the LZ codeword length for x^n [38], [65]. This in turn can be thought of as an extension of Shtarkov's ML probability assignment because $2^{-\mathrm{LZ}(x^n)}$ is well known [87] to be an upper bound (within vanishingly small terms) of $\max_P P(x^n)$, where the maximum is taken over all finite-state sources with a fixed number of states.

The problem of [40] was later extended [75] in several directions simultaneously: the alphabet of x^n and the loss function were assumed to be more general. Also, classes of predictors other than that of deterministic finite-state predictors were considered, e.g., randomized finite-state predictors (where the next-state function is randomized), families of linear predictors, etc. Many of the results of [40] turn out to carry over to this more general case.

Finally, one additional result of [75, Theorem 3] (see, also [126]) relates the individual-sequence setting back to the probabilistic setting. It tells us that under suitable regularity conditions, for a stationary and ergodic process $\cdots, X_{-1}, X_0, X_1, \cdots$, the quantity $u(X_1, X_2, \cdots)$, defined with respect to finite-state or Markov predictors, agrees almost surely with the probabilistic performance measure $\inf_b E\{l(b, X_0)|X_{-1}, X_{-2}, \cdots\}$. One special case of this result [126] is that the finite-state compressibility is almost surely equal to the entropy rate of a stationary and ergodic source. Another important example corresponds to the case where Λ_k is the class of all linear predictors of order k, and hence $u(\boldsymbol{x})$ is the *linear predictability.* In the stationary and ergodic case, the above cited result suggests that with probability one, $u(X_1, X_2, \cdots)$ coincides with the variance of the innovation process (that is, the residual linear prediction error) given by

$$\epsilon^2 = \exp\left[\frac{1}{2\pi}\int_0^{2\pi} \ln S(e^{j\omega})\, d\omega\right]$$

where $S(e^{j\omega})$ is the power spectral density of the process.

While the duality between certain classes of sources and the corresponding classes of predictors was quite straightforward in relatively small indexed (parametric) classes, the above result establishes a parallel duality between the very large class of stationary and ergodic sources and the very large class of finite-state predictors or Markov predictors.

V. Hierarchical Universality

So far we have focused on two substantially different situations of universal prediction, both of which take place in the probabilistic setting as well as in the deterministic setting: Universality with respect to an indexed class,[1] which is relatively "small," as opposed to universality with respect to a very large class, where no uniform redundancy rates exist. These two extreme situations reflect the interplay between two conflicting goals, namely, fast decay of redundancy rates on the one hand, and universality with respect to classes as wide and general as possible, on the other. For example, the Lempel–Ziv algorithm for data compression (or for predictive probability assignment) is universal for all stationary and ergodic sources, but when a memoryless source is encountered, this algorithm gives a redundancy rate that might be much slower than that of a universal scheme which is tailored to the class of memoryless sources; see [71], [87], and [101].

Our basic assumption throughout this section is that the large class Λ of sources (in the probabilistic setting) or predictors (in the deterministic setting) can be represented as a countable union of a sequence of index sets $\{\Lambda_k\}_{k\geq 1}$, which may, but not necessarily, have a certain structure, such as nestedness $\Lambda_1 \subset \Lambda_2 \subset \cdots$. In the probabilistic setting, perhaps the first example that naturally comes into one's mind is where each Λ_k is the class of discrete kth-order Markov sources, and hence the union Λ is the large class of all finite-order Markov sources. Furthermore, in the finite-alphabet case, if we slightly extend this class and take its "closure" with respect to the information divergence "distance" measure, it would include the class of all stationary sources. This is because every stationary source can be approximated, in the divergence sense, by a sequence of Markov sources of growing order [44, Theorem 3.5.1, p. 57], [47, Theorem 2.6.2, p. 52]. A few other examples of hierarchical probabilistic models are the following: i) finite-state sources with deterministic/randomized next-state functions, ii) tree sources (FSMX), iii) noisy versions of signals that are representable by countable families of basis functions, iv) arbitrarily varying sources [76], v) sources with countable alphabets (referred to as sequences of classes of growing alphabets), and vi) piecewise-stationary memoryless sources. Most of these examples have dual comparison classes in the deterministic setting.

In view of the discussion in the above two paragraphs, a natural question that arises, at this point, is the following: can one devise a universal predictor that enjoys both the benefits of a small indexed class and a large class? In other words, we would like to have, if possible, a universal predictor with respect to the large class, but with the additional property that it also performs essentially as well as the best universal predictor

[1] Since this refers to both the probabilistic and the deterministic setting, the term "class" here corresponds both to a class of sources in the probabilistic setting, and a comparison class of predictors in the deterministic setting.

within every given indexed subclass Λ_k of Λ. In the probabilistic setting, this means that if we are so fortunate that the source happens to be a member of a relatively small indexed class (e.g., a memoryless source), then the redundancy, or the regret, would be essentially the same as that of the best universal predictor for this smaller class. In the analog deterministic setting, we would like the universal predictor of this large class to behave similarly as the best universal predictor within a certain indexed comparison subclass. Note that the above question is meaningful even if Λ is merely a finite (rather than a countably infinite) union of $\{\Lambda_k\}_{k\geq 1}$. The reason is that the uniform redundancy rate of Λ, that is, the redundancy-capacity, denoted by $C_n(\Lambda)$ in the self-information loss case, might still be larger than that of any subset $C_n(\Lambda_k)$. Therefore, even in this case, treating Λ just as one big class might not be the best thing to do.

In the probabilistic setting, Ryabko [97] was the first to address this interesting question for the above described nested sequence of classes of Markov sources, and for the self-information loss (universal coding). Generally speaking, Ryabko's idea is to apply the following conceptually simple two-part code, referred to as a *twice-universal* code. The first part of the code is a codeword for an integer i whose length is $L(i) = \log i + O(\log\log i)$, and the second part is a universal code with respect to Λ_i, where i is chosen so as to minimize the total codeword length. Clearly, this code attains redundancy of

$$\min_i [C_n(\Lambda_i) + L(i)/n] \tag{58}$$

which obviously never exceeds $C_n(\Lambda_k) + L(k)/n$ for the true value of k. Since $C_n(\Lambda_k)$ behaves like $O(\log n/n)$ in the Markov case, the additional $O(1/n)$ term does not affect the rate of convergence within each Λ_k. Thus although there cannot be uniform redundancy rates simultaneously over the entire class of Markov sources Λ there is still asymptotically optimal behavior within every Λ_k.

An alternative to this two-part code, which cannot be transformed easily into a prediction scheme, is the mixture approach. Specifically, for the problem of prediction with self-information loss, the suggested solution is based on a probability assignment formed by two-stage mixture, first within each Λ_k, and then over the integers $k = 1, 2, \cdots$ [98]. The first observation is that the mixture approach, with appropriately chosen weight functions, is no worse than the above two-part scheme. To see this, let us assume that $\{L(i)\}_{i\geq 1}$ satisfy Kraft's inequality with equality (otherwise, they can be improved), and consider the two-stage mixture

$$\begin{aligned} Q(x^n) &= \sum_{i\geq 1} 2^{-L(i)} \int_\Lambda dw_i^*(\theta) P_\theta(x^n) \\ &= \sum_{i\geq 1} 2^{-L(i)} Q_{w_i^*}(x^n) \end{aligned} \tag{59}$$

where w_i^* is the capacity-achieving prior of Λ_i. Then

$$\begin{aligned} -\log Q(x^n) &\leq -\log \left[\max_{i\geq 1} (2^{-L(i)} Q_{w_i^*}(x^n))\right] \\ &= \min_{i\geq 1} [-\log Q_{w_i^*}(x^n) + L(i)] \end{aligned} \tag{60}$$

where the left-most side corresponds to the performance of the mixture approach and the right-most side corresponds to the performance of the two-part scheme with an optimum mixture within each class. The message here is that for every *individual sequence*, the mixture approach is no worse than the two-part approach. In [117] this point is further explored and developed for several examples of hierarchical classes (finite-state machines and others) in view of the fact that the first term of the right-most side above is also a lower bound for "most" sequences in a fairly strong sense (cf. Section III). Of course, the last chain of inequalities continues to hold after taking expectations in the probabilistic setting.

It turns out though, that in the probabilistic setting the mixture approach is not only no worse than the two-part approach, but moreover, it is an optimal approach in a much sharper and deeper sense. As an extension to the result of w-almost everywhere optimality of Q_w (cf. Section II), the following holds for hierarchies of classes [39, Theorem 3]: the two-stage mixture with arbitrary weight functions $\{w_i(\cdot)\}_{i\geq 1}$ within the classes, and $\pi = \{\pi_i\}_{i\geq 1}$, $\pi_i = 2^{-L(i)}$, over the positive integers, simultaneously minimizes in essence redundancy for w_i-most points in Λ_i of π-most classes $\{\Lambda_i\}$. If, in addition, $w_i = w_i^*$ is the capacity-achieving prior for all i, then this minimum redundancy can be decomposed into a sum of two terms, the first of which is $C_n(\Lambda_k)$, the capacity within the underlying class Λ_k, and the second is an extra redundancy term that reflects the additional cost of universality with respect to the unknown k. The latter term is always upper-bounded by $\frac{1}{n}\log 1/\pi_k = L(k)/n$. However, if we further assume that the classes are "easily distinguishable" in the sense that there exists a good (model order) estimator for k with small average error probability [39, Theorem 4], then $L(k)/n$ is an asymptotically tight bound. This means that in the case of distinguishable classes, $C_n(\Lambda_k) + L(k)/n$ is the optimal performance even at the level of the higher order term $L(k)/n$, which might be considerably larger for large k. However, if the classes are not easily distinguishable, the mixture approach yields a smaller second-order redundancy term whereas the two-part coding approach continues to give $L(k)/n$. Some guidelines regarding the choice of π (or, equivalently, $\{L(i)\}$) are given in [39]. It should be noted that for any monotone nonincreasing sequence of probabilities, $\pi_i \leq 1/i$ for all i, namely, $L(i) \geq \log i$, and so $C_n(\Lambda_k) + (\log k)/n$ is optimum redundancy in the distinguishable case, as it can be asymptotically attained by a universal code for the integers.

From the viewpoint of sequential predictive probability assignment, however, both the two-part method and the method of mixtures are not directly implementable because in the former, the minimizing i depends on the entire x^n, and in the latter, $\{w_i^*\}$ may depend on n. A possible alternative to the nonsequential minimization over i could be on-line estimation of i and plug-in. An algorithm in this spirit has been proposed by Weinberger, Rissanen, and Feder [118] for hierarchies of tree sources in the probabilistic setting, where the estimator of i (which is associated the context, in this case) was based on algorithm Context. Fortunately, the probability of error in estimating i decays sufficiently rapidly, so as to leave the leading redundancy term unaffected. In the deterministic setting, however, it can be shown [117] that the method based on the plug-in estimate of i does not work, i.e., there are sequences for which the resulting "redundancy" is higher than achieved when the class Λ_i is known in advance.

The mixture approach, however, is useful in both the probabilistic setting and the deterministic setting, giving us yet another reason to prefer it. To overcome the problem mentioned above, namely, the fact that the weights of the mixture over the index i depend on the horizon, we use fixed-weight functions. Fortunately, as mentioned in Section II, in many cases w_i^* are replaceable by mixture weights that do not depend on n and yet asymptotically achieve capacity.

At this point it is necessary to address a major practical concern: is it computationally feasible to implement the two-stage mixture probability assignment? More specifically, we have seen (Sections III and IV) that in some important examples the mixture within a single indexed class is easily implementable, but is it still reasonably easy to implement the second-stage mixture among (possibly infinitely) many classes. Unfortunately, there is no positive answer to this question in the general level. Nonetheless, Willems, Shtarkov, and Tjalkens, in their award-winning paper [120] provided a positive answer to this question for finite hierarchies of classes of tree sources, using an efficient recursive method, referred to as *context-tree weighting.* Their method is optimal for every individual sequence in the sense of (60). For hierarchies of countably infinitely many classes, however, the implementation issue is still unresolved. In [117] several examples are demonstrated where the countably infinite mixture over i actually collapses to a finite one. This happens because the contributions of mixtures corresponding to all i beyond a certain threshold i_0 turn out to be identical and then can be merged with the combined weight $\sum_{i \geq i_0} \pi_i$. The problem is, though, that i_0 normally grows with n, and so, the computational burden of computing i_0 mixtures at every time instant becomes explosively large as time elapses.

So far, we have discussed hierarchical universal prediction solely under the self-information loss function. What can be said about other loss functions? Apparently, we can deduce from the self-information loss function to other loss functions in the same way that this has been done in Sections III and IV. Beyond that, we are not aware of much reported work on this topic. We will mention only two directions that have been pursued explicitly. The first one is by Helmbold and Schapire [55], who have combined the exponential weighting mechanism of on-line prediction using expert advice [115] (with respect to the absolute error loss function) together with the context-tree weighting algorithm of Willems, Shtarkov, and Tjalkens [120] for competing with the best pruning of a decision tree.

Other recent work is in hierarchical linear prediction for individual sequences under the square error loss function [41], [110]. In these papers, the linear prediction problem is transformed into a Gaussian sequential probability assignment problem. The universal assignment is obtained by a two-stage mixture, over the linear prediction coefficients and over the model order. For the mixture over the parameters, a Gaussian prior is used, and the mixture can be evaluated analytically. The probability assignment attained by the mixture does not correspond directly to a universal predictor, but fortunately, such correspondence can be made for a certain range of values of the predicted sequence. Thus by a proper choice of prior, the predictor can be scaled to any finite range of the sequence values. In addition, the mixture over the model order is performed in a computationally efficient way, since using lattice filters, all possible linear predictors with model order up to some largest order M can be weighted in an efficient recursive procedure whose complexity is not larger than that for a conventional linear predictor of the model order M. It was also noted, following [75], that a plug-in estimator of the parameter (resulting from the recursive least squares (RLS) algorithm) leads to universal prediction albeit at a slower rate than the mixture approach. The resulting universal linear predictor has been implemented and tested experimentally in several practical communication and signal processing problems [110].

VI. Conclusion and Future Directions

In this paper, an attempt has been made to provide an overview on the current state-of-the-art in the problem area of universal prediction. As explained in Section I, it is definitely not, and not meant to be, a full encyclopedic survey of all scientific work that has ever been done on this topic. The aim was to mention several important concepts from the authors' point of view. Let us summarize some of these concepts very briefly.

We have seen that the problem of universal prediction has been studied extensively both in the probabilistic and the deterministic setting. There are many common features shared by these two settings. First of all, in both of them the self-information loss case plays a central role, which stems from several facts. i) It is an important loss function on its own right for reasons that were explained in Section II. One of the main reasons is that we view the prediction problem as one of probability assignment, and as such, the self-information loss function arises in a very natural manner. ii) In the self-information loss case the theory is fairly mature and well understood. iii) Results (both lower bounds and algorithms) for other loss functions can be obtained from the self-information loss function. The second common feature of the probabilistic and the deterministic settings is in the large degree of parallelism between the theories of universal prediction: universality with respect to small indexed classes, universality with respect to very large classes, and hierarchical universality, which actually bridges them. There is also a remarkable degree of analogy between the quantitative results obtained in both settings in some cases. One of the fundamental connections is that for stationary and ergodic sequences, the best attainable performance level of the deterministic definition agrees almost surely with its probabilistic counterpart.

However, there are a few differences as well: sometimes minimax redundancy rates of the deterministic setting are different from those of the probabilistic setting. The plug-in approach for predictive probability assignment works well in many instances of the probabilistic setting, but it is normally not a good approach in the deterministic setting. The minimax redundancy of the deterministic setting is different from that of the probabilistic setting. Randomization is sometimes necessary in the deterministic setting, but not in the probabilistic setting.

Perhaps one of the interesting messages is that although the term "probability assignment" originally comes from the probabilistic world, it is still meaningful in the pure deterministic setting as well. This fact is far from being trivial. Moreover, there are very efficient algorithmic tools for obtaining good probability assignments, and one of them is the incremental parsing procedure of the Lempel–Ziv algorithm.

We also see a few more theoretical problems which might be interesting to consider for future research. Some of them have been mentioned in the body of the paper.

- Develop a more solid and general theory of universal prediction for general loss functions, in parallel and extension of the theory of the self-information loss function. Derive tighter and stronger lower bounds for general loss functions both in the probabilistic setting and in the deterministic setting. For example, in the framework of prediction using expert advice, take into account relations among the experts rather than assuming the worst set of experts.
- Extend results on universal prediction with respect to the comparison class of finite-state machines to the case noisy observations.
- Impose limitations on the resources of the universal sequential predictor. For example, if the comparison class is that of finite-state predictors, how many states should the universal predictor have to guarantee redundancy below a certain level?

Some of these challenges have defied the best efforts of many researchers so far. Others are yet to be explored.

REFERENCES

[1] R. Agrawal, D. Teneketzis, and V. Anantharam, "Asymptotic adaptive allocation schemes for controlled i.i.d. processes: Finite parameter case," *IEEE Trans. Automat. Contr.*, vol. 34, pp. 258–267, Mar. 1989.

[2] ——, "Asymptotic adaptive allocation schemes for controlled Markov chains: Finite parameter case," *IEEE Trans. Automat. Contr.*, vol. 34, pp. 1249–1259, Mar. 1989.

[3] P. H. Algoet, "Universal schemes for prediction, gambling and portfolio selection," *Ann. Probab.*, vol. 20, pp. 901–941, Apr. 1992.

[4] ——, "The strong law of large numbers for sequential decision under uncertainty," *IEEE Trans. Inform. Theory*, vol. 40, pp. 609–633, May 1994.

[5] P. H. Algoet and T. M. Cover, "Asymptotic optimality and asymptotic equipartition properties of log-optimal investment," *Ann. Probab.*, vol. 16, no. 2, pp. 876–898, 1988.

[6] D. H. Bailey, "Sequential schemes for classifying and predicting ergodic processes," Ph.D. dissertation, Stanford Univ., Stanford, CA, 1976.

[7] J. M. Bernardo, "Reference posterior distributions for Bayesian inference," *J. Roy. Statist. Soc. B*, vol. 41, no. 2, pp. 113–147, 1979.

[8] D. Blackwell, "An analog to the minimax theorem for vector payoffs," *Pac. J. Math.*, vol. 6, pp. 1–8, 1956.

[9] ——, "Controlled random walks," in *Proc. Int. Congress Math.*, vol. 3. Amsterdam, The Netherlands: North Holland, 1956, pp. 336–338.

[10] A. Blum, "On-line algorithms in machine learning." [Online.] Available WWW: http://www.cs.cmu.edu/afs/cs.cmu.edu/user/ avrim/www/ Papers/pubs.html.

[11] A. C. Blumer, "Minimax universal noiseless coding for unifilar and Markov sources," *IEEE Trans. Inform. Theory*, vol. IT-33, pp. 925–930, Nov. 1987.

[12] N. Cesa-Bianchi, Y. Freund, D. P. Helmbold, D. Haussler, R. E. Schapire, and M. K. Warmuth, "How to use expert advice," in *Annu. ACM Symp. Theory of Computing*, 1993, pp. 382–391.

[13] N. Cesa-Bianchi, Y. Freund, D. P. Helmbold, and M. K. Warmuth, "On-line prediction and conversion strategies," in *Proc. EUROCOLT'93*, (Oxford, U.K., 1993), pp. 205–216.

[14] N. Cesa-Bianchi and G. Lugosi, "On sequential prediction of individual sequences relative to a set of experts," 1998, preprint.

[15] B. S. Clarke and A. R. Barron, "Information-theoretic asymptotics of Bayesian methods," *IEEE Trans. Inform. Theory*, vol. 36, pp. 453–471, May 1990.

[16] B. S. Clarke and A. R. Barron, "Jeffreys' prior is asymptotically least favorable under entropy risk," *J. Statist. Plan. Inform.*, vol. 41, pp. 37–60, Aug. 1994.

[17] T. M. Cover, "Universal gambling schemes and the complexity measures of Kolmogorov and Chaitin," Tech. Rep. 12, Dept. Statist., Stanford Univ., Stanford, CA, Oct. 1974.

[18] ——, "Open problems in information theory," in *Proc. Moscow Information Theory Workshop.* New York: IEEE Press, 1975, pp. 35–36.

[19] T. M. Cover and R. King, "A convergent gambling estimate of the entropy of English," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 413–421, July 1978.

[20] T. M. Cover, "On the competitive optimality of Huffman code," *IEEE Trans. Inform. Theory*, vol. 37, pp. 172–174, Jan. 1991.

[21] ——, "Universal portfolios," *Math. Finance*, vol. 1, no. 1, pp. 1–29, Jan. 1991.

[22] T. M. Cover and E. Ordentlich, "Universal portfolios with side information," *IEEE Trans. Inform. Theory*, vol. 42, pp. 348–363, Mar. 1996.

[23] T. M. Cover and J. A. Thomas, *Elements of Information Theory.* New York: Wiley, 1991.

[24] I. Csiszár and J. Körner, *Information Theory–Coding Theorems for Discrete Memoryless Systems.* New York: Academic, 1981.

[25] I. Csiszár and P. C. Shields, "Redundancy rates for renewal and other processes," *IEEE Trans. Inform. Theory*, vol. 42, pp. 2065–2072, Nov. 1996.

[26] L. D. Davisson, "The prediction error of stationary Gaussian time series of unknown covariance," *IEEE Trans. Inform. Theory*, vol. IT-11, pp. 527–532, Oct. 1965.

[27] ——, "Universal noiseless coding," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 783–795, Nov. 1973.

[28] ——, "Minimax noiseless universal coding for Markov sources," *IEEE Trans. Inform. Theory*, vol. IT-29, pp. 211–215, Mar. 1983.

[29] L. D. Davisson and A. L. Garcia, "A source matching approach to finding minimax codes," *IEEE Trans. Inform. Theory*, vol. IT-26, pp. 166–174, Mar. 1980.

[30] L. D. Davisson, R. J. McEliece, M. B. Pursley, and M. S. Wallace, "Efficient universal noiseless source codes," *IEEE Trans. Inform. Theory*, vol. 27, pp. 269–278, May 1981.

[31] A. P. Dawid, "Present position and potential developments: Some personal views on statistical theory the prequential approach (with discussion)," *J. Roy. Statist. Soc. A*, vol. 147, pt. 2, pp. 278–292, 1984.

[32] ——, "Fisherian inference in likelihood and prequential frames of reference (with discussion)," *J. Roy. Statist. Soc. B*, vol. 53, pp. 79–109, 1991.

[33] ——, "Prequential data analysis," in *Current Issues in Statistical Inference* (IMS Lecture Notes—Monograph Series 17), M. Ghosh and P. K. Pathak, Eds., 1992, pp. 113–126.

[34] A. P. Dawid and V. G. Vovk, "Prequential probability: Principles and properties." [Online.] Available WWW: http://www-stat.wharton.upenn.edu/Seq96/members/vovk/index. html.

[35] M. H. DeGroot, *Optimal Statistical Decisions.* New York: McGraw-Hill, 1970.

[36] A. DeSantis, G. Markowsky, and M. Wegman, "Learning probabilistic prediction functions," in *Proc. 29th IEEE Symp. Foundations of Computer Science*, 1988, pp. 110–119.

[37] P. Elias, "Minimax optimal universal codeword sets," *IEEE Trans. Inform. Theory*, vol. IT-29, pp. 491–502, July 1983.

[38] M. Feder, "Gambling using a finite state machine," *IEEE Trans. Inform. Theory*, vol. 37, pp. 1459–1465, Sept. 1991.

[39] M. Feder and N. Merhav, "Hierarchical universal coding," *IEEE Trans. Inform. Theory*, vol. 42, pp. 1354–1364, Sept. 1996.

[40] M. Feder, N. Merhav, and M. Gutman, "Universal prediction of individual sequences," *IEEE Trans. Inform. Theory*, vol. 38, pp. 1258–1270, July 1992.

[41] M. Feder and A. Singer, "Universal data compression and linear prediction," in *Proc. DCC'98*, 1998, pp. 511–520.

[42] B. Fittingoff, "Universal methods of coding for the case of unknown statistics," in *Proc. 5th Symp. Information Theory* (Moscow/Gorky, USSR, 1972), pp. 129–135.

[43] Y. Freund and R. Schapire, "Game theory, on-line prediction and boosting," in *Proc. 9th Ann. Workshop Computational Learning Theory*, 1996, pp. 89–98.

[44] R. G. Gallager, *Information Theory and Reliable Communications.* New York: Wiley, 1968.

[45] ——, "Source coding with side information and universal coding," unpublished manuscript; also presented at the *Int. Symp. Information Theory*, Oct. 1974.

[46] D. C. Gilliland, "Sequential compound estimation," *Ann. Math. Statist.*, vol. 39, no. 6, pp. 1890–1904, 1968.

[47] R. M. Gray, *Entropy and Information Theory.* New York: Springer-Verlag, 1990.

[48] L. Györfi, I. Pali, and E. C. van der Meulen, "There is no universal source code for an infinite source alphabet," *IEEE Trans. Inform. Theory*, vol. 40, pp. 267–271, Jan. 1994.

[49] J. F. Hannan, "Approximation to Bayes risk in repeated plays," in *Contributions to the Theory of Games, Ann. Math. Studies* (Princeton Univ., Princeton, NJ), vol. 3, no. 39, pp. 97–139, 1957.

[50] J. F. Hannan and H. Robbins, "Asymptotic solutions of the compound decision problem for two completely specified distributions," *Ann. Math. Statist.*, vol. 26, pp. 37–51, 1955.

[51] D. Haussler, "A general minimax result for relative entropy," *IEEE Trans. Inform. Theory*, vol. 43, pp. 1276–1280, July 1997.

[52] D. Haussler, J. Kivinen, and M. K. Warmuth, "Sequential prediction of individual sequences under general loss functions," *IEEE Trans. Inform. Theory*, to be published.

[53] D. Haussler and M. Opper, "Mutual information, metric entropy and cumulative relative entropy risk," *Ann. Statist.*, vol. 25, no. 6, Dec. 1997, to be published.

[54] ——, "General bounds on the mutual information between a parameter and n conditionally independent observations," in *Proc. 8th Annu. Workshop Computational Learning Theory (COLT'95)*, 1995, pp. 402–411.

[55] D. P. Helmbold and R. E. Schapire, "Predicting nearly as well as the best pruning of a decision tree," *Mach. Learn.*, vol. 27, pp. 51–68, 1997.

[56] Y. Hershkovitz and J. Ziv, "On fixed-database universal data compression with limited memory," *IEEE Trans. Inform. Theory*, vol. 43, pp. 1966–1976, Nov. 1997.

[57] H. Jeffreys, "An invariant form of for the prior probability in estimation problems," *Proc. Roy. Soc. London*, vol. 186, pt. A, pp. 453–461, 1946.

[58] D. Kazakos, "Robust noiseless source coding through a game theoretic approach," *IEEE Trans. Inform. Theory*, vol. IT-29, pp. 576–583, July 1983.

[59] J. L. Kelly, Jr., "A new interpretation of information rate," *Bell Syst. Tech. J.*, vol. 35, pp. 917–926, 1956.

[60] J. C. Kieffer, "A unified approach to weak universal source coding," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 674–682, Nov. 1978.

[61] ——, "An ergodic theorem for constrained sequences of functions," *Bull. Amer. Math. Soc.*, vol. 21, pp. 249–253, 1989.

[62] R. E. Krichevskiy, "Laplace's law of succession and universal encoding," *IEEE Trans. Inform. Theory*, vol. 44, pp. 296–303, Jan. 1998.

[63] R. E. Krichevski and V. E. Trofimov, "The performance of universal encoding," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 199–207, Mar. 1981.

[64] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Adv. Appl. Math.*, vol. 6, pp. 4–22, 1985.

[65] G. G. Langdon, "A note on the Lempel–Ziv model for compressing individual sequences," *IEEE Trans. Inform. Theory*, vol. IT-29, pp. 284–287, 1983.

[66] P. S. Laplace, "Memoire sur la probabilite des causes par les evenemens," *Memoires de l'Academie Royale del Sciences*, no. 6, pp. 612–656, 1774, reprinted in *Laplace Complete Work*, vol. 8. Paris, France: Gauthier-Villars, pp. 27–65. English translation by S. M. Stigler, 1986.

[67] ——, "Memoire sur les approximations des formulas qui sont functions de tres grands nombres et sur leur application aux probabilies," *Memoires de l'Academie des Sciences de Paris*, pp. 353–415, 559–565, 1810; reprinted in *Laplace Complete Work*, vol. 12. Paris, France: Gauthier-Villars,, pp. 301–353; English translation by S. M. Stigler, 1986.

[68] A. Lempel and J. Ziv, "On the complexity of finite sequences," *IEEE Trans. Inform. Theory*, vol. IT-22, pp. 75–81, Jan. 1976.

[69] N. Littlestone, P. Long, and M. K. Warmuth, "On-line learning of linear functions," in *Proc. 23rd Annu. ACM Symp. Theory of Computing*, 1991, pp. 382–391.

[70] N. Littlestone and M. K. Warmuth, "The weighted majority algorithm," *Inform. Comput.*, vol. 108, no. 2, pp. 212–261, 1994.

[71] G. Louchard and W. Szpankowski, "On the average redundancy rate of the Lempel–Ziv code," *IEEE Trans. Inform. Theory*, vol. 43, pp. 2–8, Jan. 1997.

[72] T. Matsushima, H. Inazumi, and S. Hirawasa, "A class of distortionless codes designed by Bayes decision theory," *IEEE Trans. Inform. Theory*, vol. 37, pp. 1288–1293, Sept. 1991.

[73] R. Meir, "Performance bounds for nonlinear time-series prediction," 1997, preprint.

[74] R. Meir and N. Merhav, "On the stochastic complexity of learning realizable and unrealizable rules," *Mach. Learn.*, vol. 19, no. 3, pp. 241–261, 1995.

[75] N. Merhav and M. Feder, "Universal schemes for sequential decision from individual sequences," *IEEE Trans. Inform. Theory*, vol. 39, pp. 1280–1292, July 1993.

[76] ——, "A strong version of the redundancy-capacity theorem of universal coding," *IEEE Trans. Inform. Theory*, vol. 41, pp. 714–722, May 1995.

[77] N. Merhav, M. Feder, and M. Gutman, "Some properties of sequential predictors of binary Markov sources," *IEEE Trans. Inform. Theory*, vol. 39, pp. 887–892, May 1993.

[78] J. W. Miller, R. Goodman, and P. Smyth, "On loss functions which minimize to conditional expected values and posterior probabilities," *IEEE Trans. Inform. Theory*, vol. 39, pp. 1404–1408, July 1993.

[79] D. S. Modha and E. Masry, "Universal prediction of stationary random processes," 1996, preprint.

[80] G. Morvai, S. J. Yakowitz, and L. Györfi, "Nonparametric inference for ergodic, stationary time series," *Ann. Statist.*, vol. 24, pp. 370–379, 1996.

[81] G. Morvai, S. J. Yakowitz, and P. H. Algoet, "Weakly convergent nonparametric forecasting of stationary time series," *IEEE Trans. Inform. Theory*, vol. 43, pp. 483–497, Mar. 1997.

[82] Y. Nogami, "The k-extended set-compound estimation problem in nonregular family of distributions," *Ann. Inst. Stat. Math.*, vol. 31A, pp. 169–176, 1979.

[83] M. Opper and D. Haussler, "Bounds for predictive errors in the statistical mechanics of supervised learning," *Phys. Rev. Lett.*, vol. 75, pp. 3772–3775, 1995.

[84] ——, "Worst case prediction over sequences under log loss," in *The Mathematics of Information Coding, Extraction and Distribution*, G. Cybenko, D. O'Leary, and J. Rissanen, Eds. New York: Springer-Verlag, 1997.

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

[86] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 3rd ed. (McGraw-Hill Series in Electrical Engineering). New York: McGraw-Hill, 1991.

[87] E. Plotnik, M. J. Weinberger, and J. Ziv, "Upper bounds on the probability of sequences emitted by finite-state sources and on the redundancy of the Lempel–Ziv algorithm," *IEEE Trans. Inform. Theory*, vol. 38, pp. 66–72, Jan. 1992.

[88] J. Rissanen, "Generalized Kraft's inequality and arithmetic coding," *IBM J. Res. Develop.*, vol. 20, no. 3, pp. 198–203, 1976.

[89] ——, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[90] ——, "Universal coding, information, prediction, and estimation," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 629–636, July 1984.

[91] ——, "Complexity of strings in the class of Markov sources," *IEEE Trans. Inform. Theory*, vol. IT-32, pp. 526–532, 1986.

[92] ——, "Fisher information and stochastic complexity," *IEEE Trans. Inform. Theory*, vol. 42, pp. 40–47, Jan. 1996.

[93] J. Rissanen and G. G. Langdon, "Universal modeling and coding," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 12–23, Jan. 1984.

[94] H. Robbins, "Asymptotically subminimax solutions of compound statistical decision problems," in *Proc. 2nd Berkeley Symp. Math. Stat. Prob.*, 1951, pp. 131–148.

[95] R. T. Rockafeller, *Convex Analysis.* Princeton, NJ: Princeton Univ. Press, 1970.

[96] B. Ya. Ryabko, "Encoding a source with unknown but ordered probabilities," *Probl. Inform. Transm.*, pp. 134–138, Oct. 1979.

[97] ——, "Twice-universal coding," *Probl. Inform. Transm.*, vol. 20, no. 3, pp. 173–177, July/Sept. 1984.

[98] ——, "Prediction of random sequences and universal coding," *Probl. Inform. Transm.*, vol. 24, no. 2, pp. 87–96, Apr./June 1988.

[99] E. Samuel, "Asymptotic solution of the sequential compound decision problem," *Ann. Math. Statist.*, pp. 1079–1095, 1963.

[100] ——, "Convergence of the losses of certain decision rules for the sequential compound decision problem," *Ann. Math. Statist.*, pp. 1606–1621, 1964.

[101] S. A. Savari, "Redundancy of the Lempel–Ziv incremental parsing rule," *IEEE Trans. Inform. Theory*, vol. 43, pp. 9–21, Jan. 1997.

[102] B. Scarpellini, “Conditional expectations of stationary processes,” *Z. Wahrscheinlichkeitstheorie Verw. Gebiete*, vol. 56, pp. 427–441, 1981.
[103] G. Schwarz, “Estimating the dimension of a model,” *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, 1978.
[104] C. E. Shannon, “Prediction and entropy of printed English,” *Bell Syst. Tech. J.*, vol. 30, pp. 5–64, 1951.
[105] ——, “The mind reading machine,” in *Shannon’s Collected Papers*, A. D. Wyner and N. J. A. Sloane, Eds. New York: IEEE Press, 1993, pp. 688–689.
[106] P. C. Shields, “Uniform redundancy rates do not exist,” *IEEE Trans. Inform. Theory*, vol. 39, pp. 520–524, Mar. 1993.
[107] P. C. Shields and B. Weiss, “Universal redundancy rates for the class of B-processes do not exist,” *IEEE Trans. Inform. Theory*, vol. 41, pp. 508–512, Mar. 1995.
[108] N. Shimkin, “Dynamic decision problems in multi-user systems,” Ph.D. dissertation, Technion–Israel Inst. Technol., Haifa, Isreal, Nov. 1991.
[109] Y. M. Shtar’kov, “Universal sequential coding of single messages,” *Problems of Inform. Trans.*, vol. 23, no. 3, pp. 175–186, July/Sept. 1987.
[110] A. Singer and M. Feder, “Universal linear prediction over parameters and model orders,” *IEEE Trans. Signal Processing*, to be published.
[111] J.-I. Takeuchi and A. R. Barron, “Asymptotically minimax regret for exponential and curved exponential families,” 1998, preprint.
[112] J. van Ryzin, “The sequential compound decision problem with $m \times n$ finite loss matrix,” *Ann. Math. Statist.*, vol. 37, pp. 954–975, 1966.
[113] S. B. Vardeman, “Admissible solutions of k-extended finite state set and the sequence compound decision problems,” *J. Multiv. Anal.*, vol. 10, pp. 426–441, 1980.
[114] J. S. Vitter, “Optimal prefetching via data cxompression,” in *Proc. Conf. Foundations of Computer Science*, 1991, pp. 121–130.
[115] V. G. Vovk, “Aggregating strategies,” in *Proc. 3rd Annu. Workshop Computational Learning Theory* (San Mateo, CA, 1990), pp. 371–383.
[116] ——, “A game of prediction with expert advice,” in *Proc. 3rd Annu. Workshop Computational Learning Theory* (New York, NY, 1995), pp. 51–60.
[117] M. J. Weinberger, N. Merhav, and M. Feder, “Optimal sequential probability assignment for individual sequences,” *IEEE Trans. Inform. Theory*, vol. 40, pp. 384–396, Mar. 1994.
[118] M. J. Weinberger, J. Rissanen, and M. Feder, “A universal finite memory source,” *IEEE Trans. Inform. Theory*, vol. 41, pp. 643–652, May 1995.
[119] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series.* Cambridge, MA: MIT Press, 1949.
[120] F. M. J. Willems, Y. M. Shtarkov, and T. Tjalkens, “The context-tree weighting method: Basic properties,” *IEEE Trans. Inform. Theory*, vol. 41, pp. 653–664, May 1995.
[121] Q. Xie and A. R. Barron, “Asymptotic minimax regret for data compression, gambling, and prediction,” *IEEE Trans. Inform. Theory*, 1996, submitted for publication.
[122] ——, “Minimax redundancy for the class of memoryless sources,” *IEEE Trans. Inform. Theory*, vol. 43, pp. 646–657, Mar. 1997.
[123] B. Yu, “Lower bounds on expected redundancy for nonparametric classes,” *IEEE Trans. Inform. Theory*, vol. 42, pp. 272–275, Jan. 1996.
[124] J. Ziv, “Coding of sources with unknown statistics—Part I: Probability of encoding error,” *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 384–394, May 1972.
[125] J. Ziv and A. Lempel, “A universal algorithm for sequential data compression,” *IEEE Trans. Inform. Theory*, vol. IT-23, pp. 337–343, July 1977.
[126] ——, “Compression of individual sequences via variable-rate coding,” *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 530–536, Sept. 1978.
[127] ——, “Universal coding of two-dimensional data,” *IEEE Trans. Inform. Theory*, vol. IT-32, pp. 2–8, Jan. 1986.