A Strong Version of the Redundancy-Capacity Theorem of Universal Coding

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

Abstract—**The capacity of the channel induced by a given class of sources is well known to be an attainable lower bound on the redundancy of universal codes with respect to this class, both in the minimax sense and in the Bayesian (maximin) sense. We show that this capacity is essentially a lower bound also in a stronger sense, that is, for "most" sources in the class. This result extends Rissanen's lower bound for parametric families. We demonstrate the applicability of this result in several examples, e.g., parametric families with growing dimensionality, piecewise-fixed sources, arbitrarily varying sources, and noisy samples of learnable functions. Finally, we discuss implications of our results to statistical inference.**

Index Terms— **Universal coding, minimax redundancy, minimum description length, channel capacity, arbitrarily varying sources, random coding.**

I. Introduction

In the traditional probabilistic setting of the problem of universal lossless source coding with respect to a given class of information sources, the objective is to design a single data compression scheme that performs well, in some sense, for every source in the class. The sources in this class are usually assumed to be indexed by some variable $\theta \in \Lambda$ (e.g., a parameter vector). The performance of a given code, with a length function $L(\cdot)$, is judged on the basis of the redundancy function $R_n(L, \theta)$, which is defined as the difference between the expected code length of $L(\cdot)$ with respect to a given source in the class P_θ and the nth-order entropy of P_θ normalized by the length n of the input vector to be encoded.

There are several notions of universality that were first defined by Davisson [6]. Two important ones are the *maximin* universality and the *minimax* universality. In the former, there is a Bayesian approach: the universal encoder treats the index θ as a random variable whose probability distribution is assumed worst in the sense of maximizing the minimum *expected* redundancy, i.e., the maximin redundancy. In the latter, the variable θ is considered deterministic and the goal is to find a code that minimizes the worst case redundancy $\sup_\theta R_n(L, \theta)$, namely, to achieve the minimax redundancy. (See also [2], [3], [7], [9], [14], and others).

Several years after Davisson's paper [6] it was established, first by Gallager [11], and then independently by Davisson and Leon-Garcia [8], Ryabko [19], and others, that the minimax and the maximin redundancies are equivalent and that they are both equal to the capacity of the "channel" whose input is θ and whose output is the random source vector $x^n = (x_1, \cdots, x_n)$ to be encoded, i.e., the channel defined by the set of conditional probability measures $P_\theta(x^n)$, of x^n given θ. This interesting result (which will be discussed in detail in Section II), has greatly contributed new insight into the theory of universal coding because it has been linked to the well-established theory of channel capacity. In particular, for parametric families of sources, that is, for the case where θ is a k-dimensional parameter vector, the minimax redundancy, and hence also the maximin redundancy, and the capacity of the corresponding channel, was shown to be given by $0.5k \log n/n$ plus higher order terms, under certain regularity conditions (see, e.g., [7]).

Rissanen [16], [17] has strengthened the notion of universality with respect to parametric families by showing that $0.5k \log n/n$ is not only an achievable lower bound in the minimax sense or the Bayesian sense, but also a lower bound for "most" sources in the class. Here by "most" sources we mean *every* point θ except for a subset of points whose volume (Lebesgue measure) vanishes as n grows. Rissanen's proof, however, relies heavily on the structure of the parametric family and essentially the main insight that can be gained from his work is that the redundancy is strongly related to the richness of the class, which in the parametric case is proportional to the dimension k of the parameter vector.

The question that arises now, in the light of these facts, is whether Rissanen's stronger notion of universality can be extended to the general case where the class of sources is not necessarily a parametric family. To make the question more concrete, what would be the analog measure of richness of a general class? Is it again the capacity of the channel corresponding to the given class of sources?

It turns out, as we show in this paper, that generally speaking, the answer to the latter question is yes. Specifically, we show in Section II that the Shannon capacity of the induced channel is a lower bound on the redundancy that holds simultaneously for all sources in the class except for a subset of points whose probability, under the capacity-achieving probability measure, is vanishing as n tends to infinity. This means that the minimax redundancy and the lower bound

Manuscript received December 15, 1993; revised July 6, 1994. This research was supported by the Wolfson Research Awards administrated by the Israel Academy of Science and Humanities. Part of this research was done while one of the authors (M.Feder) visited Sonderforschungsbereich 343, "Diskrete Strukturen in der Mathematik," Universität Bielefeld, Bielefeld, Germany. The material in this paper was presented at the IEEE/IMS Workshop on Information Theory and Statistics, Alexandria, VA, October 1994.

N. Merhav is with the Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel.

M. Feder is with the Department of Electrical Engineering–Systems, Tel Aviv University, Tel Aviv 69978, Israel.

IEEE Log Number 9410401.

essentially coincide for most choices of θ. Weighting the index θ by the capacity-achieving prior makes sense for reasons that will be explained in Section II. Moreover, if there exists an asymptotically capacity-achieving probability density that is positive almost everywhere (Lebesgue), as is normally the case in parametric families; the above result holds also for most sources in the Lebesgue measure sense and therefore Rissanen's result for parametric families can be obtained as a special case.

The proof of the bound in Section II is completely different and considerably simpler than Rissanen's proof [16]. However, it does not allow a free choice of any prior, other than the capacity-achieving prior, that might be reasonable as well for weighting the set of points that violate the bound. In Section III, we provide another variant of our result which permits any prior on the index set, but then the *random coding* capacity of the induced channel rather than its Shannon capacity is obtained as a lower bound. Here the random coding capacity refers to the normalized logarithm of the maximum number M of randomly chosen points $\theta_1, \cdots, \theta_M$ (corresponding to a random channel code), which form, with high probability, a set of distinguishable sources $P_{\theta_1}, \cdots, P_{\theta_M}$. Strictly speaking, this result is slightly weaker than the former because the random coding capacity might be smaller than the Shannon capacity in some cases (e.g., nonergodic channels). However, for most cases of practical interest in the context of universal coding (as we demonstrate in Section IV), the Shannon capacity and the random coding capacity are equivalent and hence the resulting bound is virtually as tight. We believe that another advantage of this random coding capacity result is that it may add some new insight about the relation between redundancy and capacity. Specifically, in the proof of this result the redundancy is linked directly, not only to the mathematical notion of capacity as the maximum mutual information, but also to the *operational* notion of capacity as the maximum attainable transmission rate that still enables reliable communication.

We would like to point out that although we formulate our results in the context of universal coding, their interpretation need not be limited to this application alone. The broader significance is that of the universal probability assignment problem (see, e.g., [4], [18], [21]). Given a class of probability measures $\{P_\theta,\ \theta \in \Lambda\}$, we wish to find a single probability measure $Q(\cdot)$ that is simultaneously "close" to every source in the class in the sense that the information divergence $D(P_\theta||Q)$ is as small as possible for every θ. Our main result, therefore, is that the divergence must be essentially at least as large as the channel capacity for most sources. This viewpoint of universal probability assignment is more general in two respects. First, it allows other applications, such as universal gambling, learning, and prediction; and in general, it may suggest guidelines to the statistician about a reasonable choice of a statistical model in the presence of uncertainty (see, e.g., Bernardo [1]). For example, when θ is unknown, one may use the "representative" probability measure Q to make inferences on the statistical behavior of future outcomes. Secondly, the probability assignment problem has a natural extension to the continuous alphabet case, unlike the lossless data compression problem.

To summarize, in our view, the main contribution of this work is primarily the redundancy-capacity Theorem in its stronger setting for most sources. This result both strengthens the earlier minimax and maximin notions of universality, and at the same time, extends Rissanen's notion of universality to families of sources that are not necessarily parametric.

II. The Shannon Capacity Lower Bound

We first define some notation. Let $x^n = (x_1, \cdots, x_n)$ denote a given string of symbols from a finite set A. The space of n-sequences will be denoted by A^n. A random vector in A^n will be denoted by $X^n = (X_1, \cdots, X_n)$. Let $\{P_\theta, \theta \in \Lambda\}$ be a class of probability mass functions (PMF's) of n-sequences in A^n, where the index set Λ, along with an appropriate set of subsets of Λ (i.e., a sigma-field), form a measurable space. For instance, Λ can be either a finite, a countable, or an uncountable index set whose members are either numbers, vectors with fixed or variable dimensionality, functions, combinations of these entities, and so on. Let $E_\theta\{\cdot\}$ denote the mathematical expectation with respect to P_θ and the nth-order entropy of P_θ is defined as

$$H_\theta(X^n) = -E_\theta \log P_\theta(X^n) \tag{1}$$

where logarithms throughout the sequel will be taken to the base 2 unless specified otherwise. A uniquely decipherable (UD) encoder for n-sequences maps each possible source string x^n to a binary word whose length will be denoted by $L(x^n)$, where by Kraft's inequality,

$$\sum_{x^n \in A^n} 2^{-L(x^n)} \le 1. \tag{2}$$

The redundancy associated with a given encoder for n-sequences is defined as

$$R_n(L, \theta) = \frac{1}{n}[E_\theta L(X^n) - H_\theta(X^n)]. \tag{3}$$

For the sake of convenience, and essentially without any effect on the results, let us ignore the integer length constraint associated with the length function $L(\cdot)$ and allow any function from A^n to the nonnegative reals such that the Kraft inequality is satisfied. The minimax redundancy is defined as

$$R_n^+ = \min_L \sup_{\theta \in \Lambda} R_n(L, \theta). \tag{4}$$

To define the maximin redundancy, let us assign a probability measure $w(\cdot)$ on Λ and let us define the mixture source

$$P(x^n) = \int_\Lambda w(d\theta) P_\theta(x^n). \tag{5}$$

The average redundancy associated with a length function $L(\cdot)$, is defined as

$$R_n(L, w) = \int_\Lambda w(d\theta) R_n(L, \theta). \tag{6}$$

The minimum expected redundancy for a given w (which is attained by the ideal code length with respect to the mixture, $L(x^n) = -\log P(x^n)$) is defined as

$$R_n(w) = \min_L R_n(L, w). \tag{7}$$

Finally, the maximin redundancy is the worst case minimum expected redundancy among all priors w, i.e.

$$R_n^- = \sup_w R_n(w). \tag{8}$$

While the minimax redundancy has the simple interpretation of worst case redundancy, in the definition of maximin redundancy, one can think of the index variable as a random variable Θ governed by the worst prior $w(\cdot)$ in the sense of maximizing the minimum average redundancy. It is easy to see [6] that the maximin redundancy is identical to the normalized capacity of the channel defined by the conditional probability measures $P_\theta(x^n)$, i.e.

$$R_n^- = \frac{C_n}{n} = \sup_w \frac{1}{n} I_w(\Theta; X^n) \tag{9}$$

where $I_w(\Theta; X^n)$ is the mutual information induced by the joint measure $w(\theta) \cdot P_\theta(x^n)$, i.e.

$$I_w(\Theta; X^n) = \int_{\theta \in \Lambda} w(d\theta) \cdot \sum_{x^n \in A^n} P_\theta(x^n) \log \frac{P_\theta(x^n)}{\int_{\theta' \in \Lambda} w(d\theta') P_{\theta'}(x^n)}. \tag{10}$$

Gallager [11] was the first to show that if $P_\theta(x^n)$ is a measurable function of θ for every x^n then $R_n^- = R_n^+$ and hence they are both equal to C_n/n. This means that the normalized capacity is an attainable lower bound on the redundancy both in the minimax sense and in the expected redundancy sense. We shall henceforth refer to the above assumption on the measurability of $P_\theta(x^n)$ as *Assumption A*.

Suppose that the capacity C_n is attained by some prior w_n^*. Our first main result strengthens the above results by stating that C_n/n is also essentially a lower bound on $R_n(L, \theta)$ for all $\theta \in \Lambda$ except for a subset of points whose probability under w_n^* vanishes as $n \to \infty$ provided that $C_n \to \infty$ (which is normally the case). The price paid for this stronger statement is that the lower bound is reduced by a factor $1 - \epsilon$, so it becomes meaningful only asymptotically as $n \to \infty$ where an arbitrarily small ϵ is allowed.

Theorem 1: Under Assumption A, for any length function L of a UD code and every $\epsilon > 0$

$$R_n(L, \theta) > (1 - \epsilon)\frac{C_n}{n} \tag{11}$$

for all $\theta \in \Lambda$ except for points in a subset $B \subset \Lambda$ where

$$w_n^*(B) \le e \cdot 2^{-\epsilon C_n} \tag{12}$$

$w_n^*(B)$ being the probability of B under w_n^*.

Another interpretation of Theorem 1 is that for every fixed probability measure Q,

$$D(P_\theta || Q) \triangleq E_\theta \log \frac{P_\theta(X^n)}{Q(X^n)} > (1 - \epsilon)C_n \tag{13}$$

for every $\theta \in \Lambda - B$. This is true because, without loss of optimality, we may restrict ourselves to length functions that satisfy the Kraft inequality with equality, and hence $Q(x^n) = 2^{-L(x^n)}$ is a probability measure. This in turn implies that $n \cdot R_n(L, \theta) \equiv D(P_\theta || Q)$.

Proof: For a given length function L and $\epsilon > 0$, consider the set $B = \{\theta : R_n(L, \theta) \le (1 - \epsilon)C_n/n\}$. Since $R_n(L, \theta) \le (1 - \epsilon)C_n/n$ for every $\theta \in B$, *a fortiori* the minimax redundancy within B cannot exceed $(1 - \epsilon)C_n/n$. By Assumption A, Gallager's redundancy-capacity Theorem [11] is applicable to the subset B, and hence the unnormalized minimax redundancy within B is equal to the capacity C_B of the channel from B to X^n, i.e., the highest mutual information that can be attained by a prior w that is supported by B. Thus the idea of the proof is to show that since $C_B \le (1-\epsilon)C_n$, then B cannot weigh "too much." Let Z denote a binary random variable that indicates the event $\Theta \in B$ where the random variable Θ is governed by w_n^*, namely, $\Pr\{Z = 1\} = w_n^*(B)$. Since $Z \to \Theta \to X^n$ is a Markov chain, we have

$$\begin{aligned} C_n = I_{w_n^*}(\Theta; X^n) &= I(Z; X^n) + I(\Theta; X^n | Z) \\ &= I(Z; X^n) + I(\Theta; X^n | Z = 1) \cdot \Pr\{Z = 1\} \\ &\quad + I(\Theta; X^n | Z = 0) \cdot \Pr\{Z = 0\}. \end{aligned} \tag{14}$$

However, $I(\Theta; X^n | Z = 1)$ never exceeds C_B, which in turn, is upper-bounded by $(1 - \epsilon)C_n$, and $I(\Theta; X^n | Z = 0)$ cannot be larger than C_n (by definition of C_n). Therefore, we have

$$\begin{aligned} C_n &\le I(Z; X^n) + (1 - \epsilon)C_n \cdot w_n^*(B) + C_n \cdot (1 - w_n^*(B)) \\ &\le H(Z) + (1 - \epsilon)C_n \cdot w_n^*(B) + C_n \cdot (1 - w_n^*(B)) \end{aligned} \tag{15}$$

where $H(Z) = h(w_n^*(B))$ is the entropy of Z, $h(\cdot)$ being the binary entropy function. A straightforward algebraic manipulation of (15) yields $h(w_n^*(B))/w_n^*(B) \ge \epsilon C_n$, or, equivalently

$$\log \frac{1}{w_n^*(B)} + \frac{1 - w_n^*(B)}{w_n^*(B)} \log \frac{1}{1 - w_n^*(B)} \ge \epsilon C_n. \tag{16}$$

The proof is completed by observing that the second term on the left-hand side of (16) is never larger than $\log e$. ■

Discussion: We would like to point out several aspects of the significance and the interpretation of Theorem 1.

More general alphabets: Theorem 1 in the formulation of (13) extends to classes of sources with infinite alphabet (countable and uncountable) provided that $C_n < \infty$ and that the class of sources is sufficiently regular so that redundancy-capacity Theorem [8], [11], [19] remains valid. In general, the regularity conditions should permit the interchange of summation (or integration) over the infinite alphabet with derivatives and with limit operations. For the special case of parametric families of continuous alphabet sources, Clarke and Barron [4] have studied the redundancy-capacity Theorem under suitable regularity conditions on the smoothness of the class of probability density functions. This extension to infinite alphabet sources is important because Rissanen [16] has also allowed infinite alphabets, and so Theorem 1 can now completely cover the setting studied by Rissanen.

Priors that almost achieve capacity: In certain situations it might happen that there is no prior that attains the capacity. However, there must be a sequence of priors $\{w_i\}_{i \ge 1}$ (where the subscript n is omitted for brevity) such that for a given n, $I_{w_i}(\Theta; X^n) \to C_n$ as $i \to \infty$. In this case, the assertion

of Theorem 1 holds for every θ outside a set B such that for every i

$$w_i(B) \leq \frac{1 + C_n - I_{w_i}(\Theta; X^n)}{\epsilon C_n}.$$

Of course, if i is sufficiently large the numerator of the latter expression tends to unity, while the denominator tends to infinity provided that $C_n \to \infty$. This upper bound on $w_i(B)$ is obtained from (15) by using the fact that $H(Z) \leq 1$.

In other situations, the capacity-achieving prior w_n^* might be discrete for every finite n, although Λ is a continuum (see, e.g., the parametric case described in Example A below). In this case, Theorem 1, as stated, is not very meaningful because it does not rule out "bad" points that are not in the support of w_n^*. It seems appropriate to restate Theorem 1 as follows: Suppose there exists a *fixed* prior w with the property

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

(e.g., Jeffrey's prior in the parametric case [4]). Then, similarly to the above argument, for every sufficiently large n the assertion of Theorem 1 holds for every point θ outside a set B such that

$$w(B) \leq [1 + C_n - I_w(\Theta; X^n)]/[\epsilon C_n].$$

Again, the last expression tends to zero if $C_n \to \infty$.

The uniform prior versus the capacity-achieving prior: While Rissanen's lower bound [16] holds for the parametric case for most sources in the uniform measure sense, our result holds in the sense of the capacity-achieving prior w_n^*, or the asymptotically capacity-achieving Jeffrey's prior. At first glance, the Lebesgue measure formulation seems more plausible because there is no apparent reason to consider some parameter values as more important than others and hence, when scanning all points $\theta \in \Lambda$ and judging the performance of L at each point, every such point should weigh equally.

A closer look, however, tells us that this approach is not always justified and in some cases it might yield misleading or meaningless results. Consider the following extreme example. The index set is $\Lambda = \{1, 2, \cdots, M^2\}$, and suppose that the sources $P_1, P_2, \cdots, P_M$ can be distinguished with a vanishingly small probability of error, upon observing X^n, although $M \to \infty$ as $n \to \infty$. The remaining sources $P_{M+1}, P_{M+2}, \cdots, P_{M^2}$ are all copies of P_M. In this case, C_n is nearly $\log M \to \infty$, but if we choose the length function L of the Shannon code with respect to P_M, then the redundancy is never larger than $1/n \ll C_n/n$ for $M(M-1)$ sources out of the M^2 sources, i.e., *most* sources in the uniform counting sense.

In contrast, if the indices $\{1, \cdots, M^2\}$ are weighted by the capacity-achieving prior, which tends to assign a mass $1/M$ to all sources $i = 1, 2, \cdots, M-1$ and a *total* mass of $1/M$ to the remaining sources, there is a natural adaptation to the fact that there are effectively only M different sources. The general conclusion is that this happens because the capacity-achieving prior assigns higher probability to regions of Λ where there is a better discrimination among the sources $\{P_\theta\}$ (and hence X^n is more informative about θ), and smaller probabilities to "noisier" channel inputs where the sources of the class are closer to each other. In these noisier regions there is a weaker justification for treating the sources as distinct.

Another advantage of the capacity-achieving prior is its invariance under one-to-one mappings of θ. The assertion of Theorem 1 should not change under different representations of the index set. If $\eta = f(\theta)$ then the capacity-achieving prior with respect to θ induces on η a probability measure that attains the capacity of the channel from η to X^n. This requirement is not satisfied, in general, by the uniform prior: If Θ is a uniform random variable on $[0, 1]$, then Θ^3 is, of course, not uniform.

In spite of these facts, we shall demonstrate in Section IV that Rissanen's result with uniform weighting can be obtained as a special case of Theorem 1. This happens because in the parametric case, the density of w_n^* is normally strictly positive almost everywhere (Lebesgue), so a small probability of B under w_n^* implies also a small probability under the uniform probability measure.

Achievability: In many examples it is known that a good choice of a universal code is a *mixture code*, i.e.

$$L(x^n) = -\log \int_\Lambda w(d\theta) P_\theta(x^n)$$

for some weight function $w(\cdot)$ that integrates to unity. This turns out to be generally true in a fairly strong sense: For any universal code (i.e., independent of θ) $L(x^n)$ there exists a mixture code $L'(x^n)$ such that for *every* $\theta \in \Lambda$, $R_n(L', \theta) \leq R_n(L, \theta)$. This statement follows straightforwardly from certain properties of projections of probability measures on convex sets [5]. Specifically, assume that $K(x^n) = 2^{-L(x^n)}$ is a probability measure of n-vectors. Now construct $L'(x^n) = -\log Q(x^n)$ where Q is the I-projection of K on the set E of all possible mixtures

$$\int_\Lambda w(d\theta) P_\theta(x^n)$$

i.e.

$$Q = \text{argmin}_{P\in E} D(P||K)$$

and the minimizer is unique (if existent) because of the convexity of E. Now, by [5, Theorem 2.2],

$$D(P||K) \geq D(P||Q) + D(Q||K) \geq D(P||Q) \qquad (17)$$

for every $P \in E$ and hence, in particular, for every $P = P_\theta$, $\theta \in \Lambda$. But $D(P_\theta||K)$ and $D(P_\theta||Q)$ are exactly the unnormalized redundancies $n \cdot R_n(L, \theta)$ and $n \cdot R_n(L', \theta)$, respectively, and thus the claim follows. This means that when seeking universal codes, one can restrict himself to mixture codes without loss of optimality, virtually in any reasonable sense. Therefore, the minimax redundancy must take the form

$$\inf_L \sup_{\theta\in\Lambda} R_n(L, \theta) = \frac{1}{n} \inf_w \sup_{\theta\in\Lambda} \sum_{x^n \in A^n} P_\theta(x^n) \cdot \log \frac{P_\theta(x^n)}{\int_{\theta'\in\Lambda} w(d\theta') P_{\theta'}(x^n)} \qquad (18)$$

which is known under Assumption A to coincide with C_n. Consequently, the mixture code L^* corresponding to the best

choice of w guarantees that $R_n(L^*,\theta) \le C_n/n$ for every $\theta \in \Lambda$, namely, the upper bound and lower bound meet for almost every θ.

Partitioning: There are situations where it is natural to partition Λ into disjoint regions and to consider the performance in each region separately. In such cases, the lower bound on the redundancy might depend on θ via the region to which θ belongs. This is different from the usual lower bound for parametric families which is independent of θ. We shall see that in the example of the arbitrarily varying source (AVS) there is a natural partitioning of the underlying state sequences (according to their types) and the bound depends explicitly on the underlying state sequence.

III. The Random Coding Capacity Lower Bound

In this section, we develop a variant of Theorem 1, in which the lower bound on the redundancy is given in terms of the random coding capacity; namely, the logarithm of the maximum number of randomly chosen points $\{\Theta_i\}$ that form a set of distinguishable sources. This capacity depends on the probability distribution $w(\cdot)$ under which the points are chosen, and the bound holds for all $\theta \in \Lambda$ except for a set whose measure vanishes with respect to $w(\cdot)$. As mentioned in the Introduction, here, unlike in Theorem 1, $w(\cdot)$ can be any probability distribution, not necessarily the capacity-achieving distribution.

The derivation of the bound follows from an analysis of a special case where $\Lambda = \{1, 2, \cdots, M\}$ and M is a finite integer being either fixed or growing with n. In other words, our class contains M known information sources $P_1, \cdots, P_M$. Let Ω_i denote the decision region associated with the ML rule, i.e.

$$\Omega_i = \{x^n \in A^n : P_i(x^n) \ge \max_{j \ne i} P_j(x^n)\} \tag{19}$$

where ties are broken arbitrarily, and let the probability of error be defined as

$$P_e = \frac{1}{M}\sum_{i=1}^{M} P_i(\Omega_i^c) \tag{20}$$

where the superscript c denotes the complementary set. Similarly to the notation in Section II, $R_n(L,i)$ will denote the redundancy of L with respect to P_i.

Theorem 2: For every $\epsilon > 0$, and every UD code

$$R_n(L,i) \ge (1-\epsilon)\frac{\log M}{n} \tag{21}$$

for every $1 \le i \le M$ except for a subset of indices B where

$$\frac{|B|}{M} \le \frac{P_e \log M + 2}{\epsilon \log M}. \tag{22}$$

The Theorem tells us that if $M \to \infty$ (or just if M is very large) but still $P_e \to 0$ as $n \to \infty$, then the redundancy is essentially lower-bounded by $n^{-1}\log M$ for all M sources except for a negligibly small fraction of them. Generally speaking, this is a special case of Theorem 1 above because the capacity of the induced channel is about $\log M$ when the sources are well distinguishable (small P_e) and the uniform prior is nearly optimal, so the probability of B under the uniform prior (which is the same as $|B|/M$) is small.

Proof of Theorem 2: For a uniform prior $w = \mu$ on $\Lambda = \{1, \cdots, M\}$ we have by Fano's inequality

$$\begin{aligned} I_\mu(\Theta; X^n) &= H_\mu(\Theta) - H_\mu(\Theta|X^n) \\ &\ge (1-P_e)\log M - h(P_e) \end{aligned} \tag{23}$$

where $H_\mu(\Theta) \equiv \log M$, and $H_\mu(\Theta|X^n)$ are the unconditional and conditional entropies of Θ, respectively. Now apply (15) with w^* replaced by μ, where C_n is lower bounded by $(1-P_e)\log M - h(P_e)$ on the left-most side, and upper bounded by $\log M$ on the right-most side. This results in

$$\begin{aligned} (1-P_e)\log M - h(P_e) \le h(\mu(B)) &+ (1-\epsilon)\mu(B)\log M \\ &+ (1-\mu(B))\log M. \end{aligned} \tag{24}$$

The proof is now completed by using the facts that $h(\mu(B)) \le 1$ and $h(P_e) \le 1$.

We next show how Theorem 2 can be used to derive the more general desired result. Let us return to the general class of information sources $\{P_\theta, \theta \in \Lambda\}$ satisfying Assumption A. Let $w(\cdot)$ be an arbitrary probability measure on Λ and let $\Theta_1, \cdots, \Theta_M$ denote M independent random points selected from Λ under the probability measure w. Suppose, without loss of generality, that Θ_1 has generated X^n. Let $\bar{P}_e(M,n,w)$ denote the average error probability; namely, the probability that $\Theta_1, \cdots, \Theta_M$ and X^n are such that for some $2 \le i \le M$, $P_{\Theta_i}(X^n) \ge P_{\Theta_1}(X^n)$. In the mathematical language,

$$\begin{aligned} \bar{P}_e(M,n,w) = 1 - \int_\Lambda w(d\theta) \sum_{x^n \in A^n} P_\theta(x^n)\big[1 - w\{\theta' : P_{\theta'}(x^n) \\ \ge P_\theta(x^n)\}\big]^{M-1}. \end{aligned} \tag{25}$$

Now let $M(n,\delta,w)$ be the largest integer M such that

$$\bar{P}_e(M,n,w) \le \delta \tag{26}$$

and finally, define the *random coding δ-capacity with respect to w* as

$$C_R(n,\delta,w) \triangleq \log M(n,\delta,w). \tag{27}$$

Theorem 3: Let $\{P_\theta, \theta \in \Lambda\}$ satisfy Assumption A and let w be any probability measure on Λ. Then for every $\epsilon > 0$, $0 < \delta < 1$, and every UD code

$$R_n(L,\theta) \ge (1-\epsilon)\frac{C_R(n,\delta,w)}{n} \tag{28}$$

for every $\theta \in \Lambda$ except for a subset of points $B' \in \Lambda$ such that

$$w(B') \le \frac{\delta C_R(n,\delta,w) + 2}{\epsilon C_R(n,\delta,w)}. \tag{29}$$

Equation (29) is of course meaningful if δ is kept very small compared to ϵ and if $C_R(n,\delta,w) \to \infty$. The tightest lower bound results from optimization of w in the sense of maximizing $C_R(n,\delta,w)$. The resulting maximum denoted $C_R(n,\delta)$, and henceforth referred to as the *random coding δ-capacity*, is essentially the lower bound for every θ except for a set whose probability under the *optimal* prior is small. This result is in the same spirit as in Theorem 1.

Proof of Theorem 3: The idea of the proof is the following. Consider a two-step procedure of randomly selecting a point in Λ, where first, we select M independent random points $\Theta_1, \cdots, \Theta_M$ under $w(\cdot)$, and then, we randomly select an index $1 \le i \le M$ of θ_i with equal probabilities. On one hand, this is obviously equivalent to a direct random selection of Θ_i under $w(\cdot)$. On the other hand, under the above described two-step random selection, it will be convenient to show (using the result of Theorem 2), that the lower bound holds for most choices of Θ_i.

For a given set of $M = M(n, \delta, w)$ randomly chosen points $\Theta_1, \cdots, \Theta_M$, let $P_e(\Theta_1^M)$ denote the error probability associated with optimal hypothesis testing (channel decoding) among the M sources $P_{\Theta_1}, \cdots, P_{\Theta_M}$ similarly as in (20). By Theorem 2, the redundancy associated with each of the resulting sources $\{P_{\Theta_i}\}$ is lower-bounded by $(1-\epsilon)C_R(n, \delta, w)/n$ except for a fraction of sources less than

$$\frac{P_e(\Theta_1^M)C_R(n, \delta, w) + 2}{\epsilon C_R(n, \delta, w)}.$$

Since $w(B')$ is identical to the average probability of the exceptional indices, we have

$$\begin{aligned} w(B') &\le \frac{E\{P_e(\Theta_1^M)\} \cdot C_R(n, \delta, w) + 2}{\epsilon C_R(n, \delta, w)} \\ &\le \frac{\delta C_R(n, \delta, w) + 2}{\epsilon C_R(n, \delta, w)} \end{aligned} \tag{30}$$

where the second inequality follows since

$$E\{P_e(\Theta_1^M)\} \equiv \bar{P}_e(M, n, w) \le \delta$$

by definition of $M(n, \delta, w)$. This completes the proof of Theorem 3. ■

Several comments about Theorem 3 are in order.

Bounds for most sources in the uniform measure sense: Since the measure w in Theorem 3 is arbitrary, one can obtain a lower bound in the uniform measure sense and thereby extend more directly Rissanen's result [16]. This can be done by redefining $C_R(n, \delta)$ as the supremum of $C_R(n, \delta, w)$ over all priors w whose densities are bounded away from zero, i.e., all w such that $dw/d\mu \ge \Delta > 0$, where $dw/d\mu$ is the Radon–Nykodim derivative of w with respect to the uniform probability measure μ. In this case, Theorem 3 holds, with $C_R(n, \delta, w)$ as a lower bound, outside B' with

$$\mu(B') \le \frac{\delta C_R(n. \delta, w) + 2}{\epsilon \Delta C_R(n. \delta, w)}.$$

However, the restriction to w's that are uniformly bounded away from zero might result in a strictly smaller random coding capacity.

More general random coding distributions: The proof of Theorem 3 would continue to hold if rather than using M independent copies of a random variable governed by w, one would use any joint probability measure of $(\Theta_1, \cdots, \Theta_M)$ with the property that the mixture of its marginals, $M^{-1}\sum_{i=1}^{M} w_i(\cdot)$, agrees with $w(\cdot)$. In this case, one should modify accordingly the expression for $\bar{P}_e(M, n, w)$ instead of using (25), which has been derived under the assumption that $\{\Theta_i\}$ are drawn independently. For instance, let $\Lambda = [0, 1]$ and consider the grid $\theta_i = i/M$, $i = 0, \cdots, M-1$. Now, let $\Theta_i = \theta_i + \alpha$, where α is uniformly distributed in $[0, 1/M]$. Then, the mixture of marginals of $\{\Theta_i\}$ is uniform on $[0, 1]$. This might be useful when the evaluation of the capacity is easier with this constellation than with that of Theorem 3 (see Examples A and B in the next section). In spite of this fact, we preferred to formulate Theorem 3 with independent selections of $\{\Theta_i\}$ because this is in the spirit of conventional random coding techniques.

Lower bounds on $M(n, \delta, w)$: In some cases it will be easier to use Theorem 3 by computing a lower bound on $M(n, \delta, w)$ rather than trying to compute it precisely. This can be done by using upper bounds on $\bar{P}_e(M, n, w)$. One simple upper bound is the union bound associated with the pairwise error probabilities $P_{\Theta_i}\{P_{\Theta_j}(X^n) \ge P_{\Theta_i}(X^n)\}$. Specifically, by the union bound the average probability of error defined in (25) is upper bounded by

$$\begin{aligned} \bar{P}_e(M, n, w) \le (M-1) \cdot \int_{\Lambda} w(d\theta) \\ \cdot \sum_{x^n \in A^n} P_\theta(x^n) w\{\theta' : P_{\theta'}(x^n) \ge P_\theta(x^n)\} \end{aligned} \tag{31}$$

and so

$$\begin{aligned} M(n, \delta, w) \\ \ge 1 + \frac{\delta}{\int_{\Lambda} w(d\theta) \sum_{x^n \in A^n} P_\theta(x^n) w\{\theta' : P_{\theta'}(x^n) \ge P_\theta(x^n)\}}. \end{aligned} \tag{32}$$

Thus $C_R(n, \delta, w)$ is essentially lower-bounded by the negative logarithm of the average pairwise error probability. Another bound (which can be combined with the first one) results from replacing the ML decision rule by a suboptimal decision rule that is easier to analyze.

IV. Examples

In this section we demonstrate the applicability of our results in several specific examples of classes of sources. We start from the case where $\{P_\theta, \theta \in \Lambda\}$ is a parametric family and obtain Rissanen's lower bound [16] as a special case. Later on, we study two important cases where the model class attempts to deal with nonstationarity. One is the case of an abrupt change in the statistics of the source (i.e., piecewise-fixed sources), and the other is the AVS. Finally, we discuss the problem of universal coding of noisy versions of outputs of binary functions given the inputs as side information, and relate the necessary and sufficient conditions for the existence of universal codes to the learnability of these binary functions.

A. Parametric Families

Let $\{P_\theta, \theta \in \Lambda\}$ be a parametric family, where Λ is a compact subset of the k-dimensional Euclidean space. For parametric families of memoryless sources Clarke and Barron [4] have derived the following refined expression for the unnormalized redundancy: $n \cdot R_n(w, \theta)$ of the mixture code

with respect to w at θ.

$$n \cdot R_n(w,\theta) = \frac{k}{2}\log\frac{n}{2\pi e} + \log\frac{\sqrt{\det I(\theta)}}{dw(\theta)/d\theta} + o(1) \quad (33)$$

where the $o(1)$ term is uniformly small on compact subsets interior to Λ [4], and $I(\theta)$ is the one-sample Fisher information matrix given by

$$I(\theta) = E_\theta[\partial \log P_\theta(x)/\partial\theta]^2. \quad (34)$$

Averaging with respect to w yields

$$n \cdot R_n(w) = I_w(\Theta; X^n) = \frac{k}{2}\log\frac{n}{2\pi e} + \int_\Lambda w(d\theta)\log\frac{\sqrt{\det I(\theta)}}{dw(\theta)/d\theta} + o(1). \quad (35)$$

There is no closed-form expression for the exact maximizer of (35) for any finite n. However, by Jensen's inequality the middle term (which depends strongly on w) is maximized for

$$\frac{dw^*(\theta)}{d\theta} = \frac{\sqrt{\det I(\theta)}}{\int\limits_{\theta'\in\Lambda}\sqrt{\det I(\theta')}d\theta'} \quad (36)$$

which is known as *Jeffrey's prior*. It is well known (see, e.g., [4]) that Jeffrey's prior asymptotically attains capacity in the sense of the second comment after Theorem 1 (second paragraph). A point to observe is that $dw^*/d\theta$ is a monotonic function of det $I(\theta)$ which is in agreement with the intuition that w^* puts more mass on values of θ that are easier to estimate (see, Discussion of Section II). However, if det $I(\theta)$ is positive almost everywhere (Lebesgue), then a small probability of B under w^* (according to Theorem 1) implies a small probability under the uniform measure as stated in [16].

The limitation of the above derivation of Clarke and Barron, however, is that the required smoothness conditions about P_θ are considerably demanding. In particular, these conditions are more restrictive than Rissanen's condition, which requires the existence of an estimator $\hat\theta = \hat\theta(x^n)$ for θ, such that for every $\theta \in \Lambda$, every $c > 0$ and every $n \geq n_0(c)$

$$P_\theta\left\{x^n : ||\hat\theta(x^n) - \theta|| > \frac{c}{\sqrt{n}}\right\} \leq \epsilon(c) \quad (37)$$

where $||\cdot||$ denotes the Euclidean norm (or the maximal component norm) and $\epsilon(c) \to 0$ as $c \to \infty$.

We next show that the existence of such an estimator *implies* that for the given parametric family, the maximum number of distinguishable randomly chosen points in Λ is proportional to $n^{k/2}$ and hence this condition is *weaker* than Rissanen's condition for the validity of the lower bound $0.5k\log n/n$.

In this case it is easier to apply Theorem 3 by using uniform grids with random offsets $\Theta_1, \cdots, \Theta_M$, rather than by independent random selection of the M points (see second comment after Theorem 3). Specifically, let $G_n = \{\theta_1, \cdots, \theta_M\}$ be the grid of k-dimensional vectors $\theta_i \in \Lambda$ whose components are all integer multiples of $2c/\sqrt{n}$. In our random set $\{\Theta_1, \cdots, \Theta_M\}$ let every Θ_i be given by $\theta_i + \eta \in \Lambda$ where η is a random vector uniformly distributed within the k-dimensional cube all sides of which are $2c/\sqrt{n}$. By an argument similar to that of the proof of Theorem 3, if we show that Theorem 2 is applicable to $\{\Theta_1, \cdots, \Theta_M\}$ for *every* η, then this will imply that the lower bound holds for most points in Λ in the Lebesgue measure sense. This, in turn, is immediate if we consider the suboptimal decision rule which picks the ith source if $\hat\theta$ falls in the cube whose center is Θ_i. Since this rule is suboptimal its average error probability is greater than the average error probability associated with the ML rule $\bar P_e(M, n, w)$, but smaller than $\epsilon(c)$ for *every* η, as observed from (37). However, $\epsilon(c)$ in turn, can be made smaller than δ if c is sufficiently large. Since the number of grid points is proportional to $n^{k/2}$ the desired result follows.

Another advantage of the above technique is that it is easy to generalize to some parametric models where the dimension of the parameter vector k_n grows with n. Consider, for example, a source whose output results from very fine quantization of the process

$$x_i = \sum_{j=1}^{k_n} a_j\phi_j(i) + w_i, i = 1, \cdots, n \quad (38)$$

where $k_n \leq n$, $\{\phi_j(\cdot)\}_{j=1}^n$ is a complete set of orthonormal sequences

$$\Lambda = \{(a_1, \cdots, a_{k_n}) : \sum_{j=1}^{k_n} a_j^2 \leq P\}$$

and $\{w_i\}$ are i.i.d zero-mean Gaussian random variables with variance σ^2. Here the ML decision rule picks the source (parameter vector) that minimizes

$$\sum_{i=1}^{n}(x_i - \sum_{j=1}^{k_n} a_j\phi_j(i))^2$$

which is equivalent to the selection of the source that minimizes

$$\sum_{j=1}^{k_n}(a_j - \hat a_j)^2$$

where

$$\hat a_j = n^{-1}\sum_{i=1}^{n} x_i\phi_i(j).$$

By a derivation similar to the above, it is not difficult to show that for every $\delta > 0$ the leading term of $C_R(n, \delta, \mu)$ is given by

$$\frac{k_n}{2}\log\left(1 + \frac{P}{\sigma^2}\cdot\frac{n}{k_n}\right). \quad (39)$$

This agrees with the well-known results in two extreme cases: If $k_n = k$ is fixed, this expression is dominated by $0.5k\log n$ as before, and if $k_n = \alpha n$ for some fixed $0 < \alpha \leq 1$, we get (as expected) the capacity of a discrete-time band-limited Gaussian channel whose bandwidth is a fraction α of the entire Nyquist bandwidth. Thus (39) bridges these two extremes for any sublinear growth rate of k_n.

B. Piecewise-Fixed Sources

Let Q_1 and Q_2 be two given distinct probability assignments on data sequences, and for $i = 0, 1, \cdots, n$ let us define

$$P_i(x^n) = Q_1(x_1^i) \cdot Q_2(x_{i+1}^n) \tag{40}$$

where x_i^j denotes the substring $(x_i, \cdots, x_j)$ for $i \le j$ and the empty string for $i > j$. The probability of the empty string (under both Q_1 and Q_2) is defined as unity.

This is a very simple model of a single abrupt change from one probability law to another, where the two (stationary) segments x_1^i and x_{i+1}^n are assumed independent. The boundary point i is assumed unknown to both encoder and decoder. We shall show that if Q_1 and Q_2 are sufficiently "distant" the capacity associated with the class $\{P_i, i = 0, 1, \cdots, n\}$ behaves like $\log n$.

More precisely, assume that Q_1 and Q_2 are such that for every $\epsilon > 0$, $|i - j| > n^\epsilon$ implies that

$$P_i\{x^n : P_j(x^n) > P_i(x^n)\} = o(1/n).$$

(For instance, if Q_1 and Q_2 are memoryless sources then $P_i\{x^n : P_j(x^n) > P_i(x^n)\}$ decays exponentially with $|i - j|$ and the assumption definitely holds). Again, in this example, it is easier to estimate the capacity by using its operational significance rather than to calculate the mutual information. We shall use again the idea of a random grid as in Example A.

Fix $\epsilon > 0$ and select $M = n^{1-\epsilon}$ integers $i_j = l + (j-1) \cdot n^\epsilon$, $j = 1, \cdots, M$, where l is uniformly distributed on the integers $\{1, \cdots, n^\epsilon\}$. Since $\{i_1, \cdots, i_M\}$ are such that $|i_j - i_k| \ge n^\epsilon$ for every $j \ne k$, then by the union bound, the probability of error is less than

$$\begin{aligned}(M-1) &\max_{\{k:|i_k - i_j| > n^\epsilon\}} P_{i_j}\{P_{i_k}(x^n) > P_{i_j}(x^n)\} \\ &\le n \max_{\{k:|i_k - i_j| > n^\epsilon\}} P_{i_j}\{P_{i_k}(x^n) > P_{i_j}(x^n)\} \end{aligned} \tag{41}$$

which vanishes with n (independently of l) under the above assumption. Therefore, $C_n > (1 - \epsilon) \log n$. A similar result has been derived in [15] by using a more direct extension of Rissanen's bound and by parametrizing the transition point i as $i = \lfloor n \cdot \alpha \rfloor$ where $\alpha \in (0, 1)$. The bound was then shown to hold for almost every α. The bound is achievable by an encoder that seeks the best partitioning of the data in the sense of minimizing the total Shannon code length $[-\log Q_1(x_1^i)] + [-\log Q_2(x_{i+1}^n)]$ and adds $\log n$ bits to encode the best choice of i.

C. Arbitrarily Varying Sources

The memoryless arbitrarily varying source (AVS) is defined by

$$P(x^n|s^n) = \prod_{t=1}^{n} p(x_t|s_t) \tag{42}$$

where $s^n = (s_1, \cdots, s_n)$. The variable s_t, which is referred to as the *state* of the source at time t, is assumed to take on values in some finite set S. If we scan all possible state sequences s^n, we can view the AVS as a class of sources indexed by $\theta = s^n$ and the index set Λ is S^n. Thus the problem of universal coding for the AVS is that of efficient data compression in the absence of knowledge about the underlying state sequence.

The AVS can be viewed also as a memoryless channel with transition probabilities $\{p(x|s)\}_{x \in A, s \in S}$, where the input is s^n and the output is x^n. Therefore, the notion of capacity is more natural in this example than in the other examples. In order to use Theorem 1 in a meaningful manner in this example let us partition the state sequence space S^n into equivalence classes defined by the compositions (types) of the different state sequences. The type $T(s^n)$ of a state sequence s^n is the set of all state sequences $\sigma^n \in S^n$ for which the number of occurrences of every element $s \in S$ in σ^n is equal to the number of occurrences of the same element in s^n. The relative frequency of s in s^n will be denoted by $\hat{q}(s)$.

Now, if we treat every given type $T(s^n)$ as a separate index set, the best prior w^* is the uniform distribution μ on $T(s^n)$ which asymptotically achieves the per-symbol mutual information

$$I(s^n) = \sum_{x \in A} \sum_{s \in S} \hat{q}(s) p(x|s) \log \frac{p(x|s)}{\sum_{s' \in S} \hat{q}(s') p(x|s')}. \tag{43}$$

where the dependence on s^n expresses the fact that $\hat{q}$ depends on s^n. Thus asymptotically C_n/n is never less than $I(s^n)$ and hence, by Theorem 1, the normalized redundancy is lower-bounded by $(1 - \epsilon) I(s^n)$ for almost every state sequence in $T(s^n)$. This bound can be shown to be achievable for every s^n by a mixture code (see [10]).

D. Noisy Samples of Learnable Functions

An interesting special case of the AVS occurs when the the state variable s_i is given by some binary function f of another variable $z_i \in Z$, i.e., $s_i = f(z_i)$. For simplicity, let us assume that the channel corresponding to $p(x|s)$ is additive, namely, $x_i = f(z_i) + w_i$, where w_i is a binary random variable with $\Pr(w_i = 1) = p$ and the summation is modulo 2. Here $z_1, \cdots, z_n$ is some fixed sequence and the index set Λ is a certain class F of allowable binary functions f.

This model has been widely studied in the context of computational learning theory where the problem is to learn a function $f \in F$ given a sequence of pairs $\{(z_i, x_i)\}_{i=1}^n$ that serve as examples of inputs and noisy outputs of this function, and learnability means the ability to predict reasonably well (based on past examples) the next output x_{n+1} upon seeing the next input z_{n+1}. One way to assess the predictability of future outcomes is to measure their compressibility (see, e.g., [12], [13]).

It turns out that there is a relation between the learnability of functions in F and the existence of universal codes for the noisy labels $\{x_i\}$ given $\{z_i\}$ as side information. This happens because both these properties are connected to the *capacity* (or the Vapnik–Chervonenkis (VC) dimension) of F which is defined as the length n of the longest existing input sequence $z_1, \cdots, z_n$ for which all 2^n possible binary sequences $(f(z_1), \cdots, f(z_n))$ are obtained when f scans F.

Following [20] the maximum number of binary sequences $\{f(z_1), \cdots, f(z_n)\}$ that can be obtained by a class F whose

VC dimension is d, is 2^n for $n \leq d$ and upper bounded by

$$\sum_{k=0}^{d} n!/[k!(n-k)!] = O(n^d)$$

for $n > d$. Thus following Example C, the best attainable normalized redundancy C_n/n is given by $1 - h(p)$ for $n < d$, and upper-bounded by $(1 + o(1))d \log n/n$ for $n > d$. In other words, universal coding (in the sense of approaching the entropy) is possible if and only if the VC dimension of F is finite, and if this is the case the least attainable redundancy is essentially never larger than $d \log n/n$. For many parametric families the VC dimension d agrees with the dimension k of the parameter vector. Nevertheless, the upper bound on the redundancy is in general given by the factor d and not $d/2$ (as opposed to Example A) because here there are no assumptions on the smoothness of the parametric family. For instance, Example B corresponds to a class of step functions $\{f\}$ parametrized by the location α of the transition, so the VC dimension is $d = 1$, and the upper bound is tight in this case.

ACKNOWLEDGMENT

The authors wish to thank Dr. R. Meir and Dr. A. Orlitsky for interesting discussions.

REFERENCES

[1] J. M. Bernardo, "Reference posterior distributions for Bayesian inference," *J. Roy. Statist. Soc. B*, vol. 41, no. 2, pp. 113–147, 1979.
[2] A. C. Blumer, "Minimax universal noiseless coding for unifilar and Markov sources," *IEEE Trans. Inform. Theory*, vol. IT-33, no. 6, pp. 925–930, Nov. 1987.
[3] W.-C. Chen and R. J. Fontana, "On total boundedness for existence of weakly minimax universal codes," *IEEE Trans. Inform. Theory*, vol. IT-27, no. 6, pp. 781–784, Nov. 1981.
[4] B. S. Clarke and A. R. Barron, "Jeffrey's prior is asymptotically least favorable under entropy risk," *J. Statist. Plan. Inform.*, vol. 41, pp. 37–60, 1994.
[5] I. Csiszár, "I-divergence geometry of probability distributions and minimization problems," *Ann. Probab.*, vol. 3, no. 1, pp. 146–158, 1975.
[6] L. D. Davisson, "Universal noiseless coding," *IEEE Trans. Inform. Theory*, vol. IT-19, no. 6, pp. 783–795, Nov. 1973.
[7] ——, "Minimax noiseless universal coding for Markov sources," *IEEE Trans. Inform. Theory*, vol. IT-29, no. 2, pp. 211–215, Mar. 1983.
[8] L. D. Davisson and A. Leon-Garcia, "A source matching approach to finding minimax codes," *IEEE Trans. Inform. Theory*, vol. IT-26, no. 2, pp. 166–174, Mar. 1980.
[9] L. D. Davisson, R. J. McEliece, M. B. Pursely, and M. S. Wallace, "Efficient universal noiseless source codes," *IEEE Trans. Inform. Theory*, vol. IT-27, no. 3, pp. 269–279, May 1981.
[10] M. Feder and N. Merhav, "Universal coding for arbitrarily varying sources," submitted for publication.
[11] R. G. Gallager, "Source coding with side information and universal coding," unpublished manuscript, Sept. 1976.
[12] D. Haussler and A. R. Barron, "How well do Bayes methods work for on-line prediction of $\{\pm 1\}$ values?" Tech. Rep. UCSC-CRL-92-37, Comput. Inform. Sci., Univ. of Calif. at Santa Cruz, July 1992.
[13] D. Haussler, M. Kearns, and R. Schapire, "Bounds on the sample complexity of Bayesian learning using information theory and the VC dimension," Tech. Rep. UCSC-CRL-91-44, Comput. Inform. Sci., Univ. of Calif. at Santa Cruz, Mar. 1992.
[14] D. Kazakos, "Robust noiseless source coding through a game theoretic approach," *IEEE Trans. Inform. Theory*, vol. IT-29, no. 4, pp. 576–583, July 1983.
[15] N. Merhav, "On the minimum description length principle for sources with piecewise constant parameters," *IEEE Trans. Inform. Theory*, vol. 39, no. 6, pp. 1962–1967, Nov. 1993.
[16] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Trans. Inform. Theory*, vol. IT-30, no. 4, pp. 629–636, July 1984.
[17] ——, "Complexity of strings in the class of Markov sources," *IEEE Trans. Inform. Theory*, vol. IT-32, no. 4, pp. 526-532, July 1986.
[18] ——, "Fisher information and stochastic complexity," IBM Res. Rep., 1993.
[19] B. Ya. Ryabko, "Encoding a source with unknown but ordered probabilities," *Probl. Inform. Transmission*, pp. 134–138, Oct. 1979.
[20] V. N. Vapnik and A. Ya. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory of Probab. and its Appl.*, vol. XVI, pp. 264–280, 1971.
[21] 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.