Hierarchical Universal Coding

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

Abstract—**In an earlier paper, we proved a strong version of the redundancy–capacity converse theorem of universal coding, stating that for "most" sources in a given class, the universal coding redundancy is essentially lower-bounded by the capacity of the channel induced by this class. Since this result holds for general classes of sources, it extends Rissanen's strong converse theorem for parametric families. While our earlier result has established strong optimality only for mixture codes weighted by the capacity-achieving prior, our first result herein extends this finding to a general prior. For some cases our technique also leads to a simplified proof of the above mentioned strong converse theorem.**

The major interest in this paper, however, is in extending the theory of universal coding to hierarchical structures of classes, where each class may have a different capacity. In this setting, one wishes to incur redundancy essentially as small as that corresponding to the active class, and not the union of classes. Our main result is that the redundancy of a code based on a two-stage mixture (first, within each class, and then over the classes), is no worse than that of any other code for "most" sources of "most" classes. If, in addition, the classes can be efficiently distinguished by a certain decision rule, then the best attainable redundancy is given explicitly by the capacity of the active class plus the normalized negative logarithm of the prior probability assigned to this class. These results suggest some interesting guidelines as for the choice of the prior. We also discuss some examples with a natural hierarchical partition into classes.

Index Terms— **Universal coding, minimax redundancy, maximin redundancy, capacity, redundancy–capacity theorem, mixtures, arbitrarily varying sources.**

I. Introduction

IN THE basic classical setting of the problem of universal coding it is assumed that, although the exact information source is unknown, it is still known to belong to a given class $\{P(\cdot|\theta), \theta \in \Lambda\}$, e.g., memoryless sources, first-order Markov sources, and so on. The performance of a universal code is measured in terms of the excess compression ratio beyond the entropy, namely, the redundancy rate $R_n(L, \theta)$, which depends on the code length function $L(\cdot)$, the source indexed by θ, and the data record length n. The *minimax* redundancy

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

defined by Davisson [9], is the minimum uniform redundancy rate that can be attained for *all* sources in the class. Gallager [13] was the first to show (see also, e.g., [11], [22]) that $R_n^+ = C_n$, where C_n is the capacity (per symbol) of the "channel" from θ to the source string $\boldsymbol{x} = (x_1, \cdots, x_n)$, i.e., the channel defined by the set of conditional probabilities $\{P(\boldsymbol{x}|\theta), \theta \in \Lambda\}$. This redundancy rate can be achieved by an encoder whose length function corresponds to a mixture of the sources in the class, where the weighting of each source θ is given by the capacity-achieving distribution. Thus the capacity $C_n = R_n^+$ actually measures the richness of class from the viewpoint of universal coding.

One may argue that the minimax redundancy is a pessimistic measure for universal coding redundancy since it serves as a lower bound to the redundancy for the *worst* source only. Nevertheless, for smooth parametric classes of sources, Rissanen [18] has shown that this (achievable) lower bound essentially applies to *most* sources in the class, namely, for all θ except for a subset B whose Lebesgue measure vanishes with n. In a recent paper [16], we have extended this result to general classes of information sources, stating that for any given L, $R_n(L, \theta)$ is essentially never smaller than C_n, simultaneously for every θ except for a "small" subset B. The subset B is small in the sense of having a vanishing measure w.r.t. the prior w^* that achieves (or nearly achieves) capacity.[1] The results in [16] strengthen the notion of Shannon capacity in characterizing the richness of a class of sources. In this context, our first contribution here is in developing a technique that both simplifies the proof and extends the result of [16] to a general prior, not only the capacity-achieving prior. In light of all these findings, this basic setting of universal coding for classes with uniform redundancy rates is now well understood.

Another category of results in universal lossless source coding corresponds to situations where the class of sources is so large and rich, that there are no uniform redundancy rates at all; for example, the class of all stationary and ergodic sources. In these situations, the goal is normally to devise data compression schemes that are universal in the weak sense only, namely, schemes that asymptotically attain the entropy of every source, but there is no characterization of the redundancy, which might decay arbitrarily slowly for some sources. In fact, this example of the class of all stationary and ergodic sources is particularly interesting because it can be thought of as a "closure" of the union of all classes Λ_i of ith-order Markov sources: every stationary and ergodic source can be approached, in the relative entropy sense, by a sequence of Markov sources of growing order. Unfortunately, existing universal encoders for stationary and ergodic sources (e.g., the Lempel–Ziv algorithm) are unable to adapt the redundancy

Manuscript received April 27, 1994; revised January 20, 1996. This work was supported in part by the S. Neaman Institute and the Wolfson Research Awards administered by the Israel Academy of Science and Humanities.

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

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

Publisher Item Identifier S 0018-9448(96)05451-X.

[1] It is explained in [16] why it is more reasonable to measure the exception set B w.r.t. w^* (or a good approximation to w^*) rather than the uniform measure.

when a source from a "small" subclass is encountered. For example, when the underlying source is Bernoulli, the redundancy of the Lempel–Ziv algorithm does not reduce to the capacity $C_n \approx 0.5 \log n/n$ of the class of Bernoulli sources.

This actually motivates the main purpose of this paper, which is to extend the scope of universal coding theory so as to deal with hierarchies of classes. Specifically, we focus on the following problem: let $\Lambda_1, \Lambda_2, \cdots,$ denote a finite or countable set of source classes with possibly different capacities $C_n(\Lambda_1), C_n(\Lambda_2), \cdots$. We know that the source belongs to some class Λ_i but we do not know i. Our challenge is to provide coding schemes with optimum "adaptation" capability in the sense that, first, the capacity of the active class $C_n(\Lambda_i)$ is always approached, and moreover, the extra redundancy due to the lack of prior knowledge of i is *minimum*.

One conceptually straightforward way to achieve this adaptation property is to apply a two-part code, where the first part is a code for the index i using some prior on the integers $\{\pi_i\}$, and the second part implements optimum universal coding within each class. By doing this, one can achieve redundancy essentially as small as $C_n(\Lambda_i) + (\log 1/\pi_i)/n$. This method, however, requires a comparison between competing codes for all $\{i\}$ or a good estimator for the true i, for example, the minimum description length (MDL) estimator [17]–[19] or some of its extensions (see, e.g., [2], [3]). Although this approach has been proved successful in certain situations, it is not clear whether it is optimal in general.

An alternative approach, proposed first by Ryabko [23] for Markov sources, is to make a further step in the Bayesian direction and to use a code that corresponds to a two-stage mixture, first within each class and then over the classes. (See also, e.g., [26] for efficient implementation of two-stage mixture codes, and [25] for other related work.) It is easy to show that the resultant redundancy is never larger than that of the above mentioned two-part code. We will see, however, that the reasoning behind the Bayesian approach to hierarchical universal coding is deeper than that. We prove that a two-stage mixture code with a given weighting is no worse than any other lossless code for "most" sources of "most" classes w.r.t. this weighting.

If, in addition, the classes $\{\Lambda_i\}$ are *distinguishable* in the sense that there exists a good estimator for i (e.g., the Markovian case where there is a consistent order estimator [24]), then the minimum attainable redundancy is essentially

$$C_n(\Lambda_i) + \frac{1}{n} \log \frac{1}{\pi_i}. \tag{1}$$

While this redundancy is well known to be achievable, here we also establish it as a lower bound. This suggests an interesting guideline with regard to the choice of the prior: It would be reasonable to choose $\{\pi_i\}$ so that the second term would be a negligible fraction of the first term, which is unavoidable. This means that the richer classes are assigned smaller weights.

In other cases, the redundancy of this two-stage mixture code, which essentially serves as a lower bound for any other code, can be decomposed into a sum of two capacity terms. The first is the *intra-class capacity* $C_n(\Lambda_i)$, representing the cost of universality within Λ_i, and the second term is the *inter-class capacity* c_n, which is attributed to the lack of prior knowledge of the index i. The goal of approaching $C_n(\Lambda_i)$ for every i is now achievable if c_n (which is independent of i) is very small compared to $C_n(\Lambda_i)$ for all i.

In the last part of the paper, we analyze the special case of finite-state (FS) arbitrarily varying sources (AVS's), where such a decomposition property takes place if the $\{\Lambda_i\}$ are defined as the type classes of all possible underlying state sequences. Here, the first term $C_n(\Lambda_i)$, which depends on the type of the state sequence, tends to a positive constant as $n \to \infty$, while the second term c_n behaves like $O(\log n/n)$. Our results indicate that the best attainable compression ratio is essentially as if the state sequence was i.i.d. with a probability distribution being the same as the *empirical* distribution of the actual underlying (deterministic) state sequence. This is different from earlier results due to Berger [4, sec. 6.1.2] and Csiszár and Körner [8, Theorem 4.3] for fixed length rate-distortion codes. According to [4] and [8], for the distortionless case, the best attainable rate is the same as if the state sequence were i.i.d. with the *worst* possible distribution in the sense of maximizing the source output entropy. Thus by applying the hierarchical approach to AVS's, we have both improved the main redundancy term and characterized the best second-order term c_n.

The outline of the paper is as follows. In Section II, some preliminaries and background of earlier work are provided. In Section III, a simplified and extended version of [16, Theorem 1] is presented. In Section IV, the main results are derived for general hierarchies of classes of sources. In Section V, the closed-form expression (1) for the best achievable redundancy is developed for the case of distinguishable classes. Finally, in Section VI, the special case of FS AVS's is studied.

II. Background

Throughout this work, we adopt the convention that a (scalar) random variable is denoted by a capital letter (e.g., X), a specific value it may take is denoted by the respective lower case letter (x), and its alphabet is denoted by the respective script letter ($\mathcal{X}$). As for vectors, a bold-type capital letter ($\boldsymbol{X}$) will denote an n-dimensional random vector $(X_1, \cdots, X_n)$, a bold-type lower case letter ($\boldsymbol{x}$) will denote a specific vector value $(x_1, \cdots, x_n)$, and the respective super-alphabet, which is the nth Cartesian power of the single-letter alphabet, will be denoted by the corresponding script letter with the superscript n ($\mathcal{X}^n$). The cardinality of a set will be denoted by $|\cdot|$, e.g., $|\mathcal{X}|$ is the size of the alphabet of X. Alphabets will be assumed finite throughout this paper. Probability mass functions (PMF's) of single letters will be denoted by lower case letters (e.g., p) and PMF's of n-vectors will be denoted by the respective capital letters (P).

A uniquely decipherable (UD) encoder for n sequences maps each possible source string $\boldsymbol{x} \in \mathcal{X}^n$ to a binary word whose length will be denoted by $L(\boldsymbol{x})$, where by Kraft's inequality

$$\sum_{\boldsymbol{X} \in \mathcal{X}^n} 2^{-L(\boldsymbol{X})} \leq 1. \tag{2}$$

For the sake of convenience, and essentially without any effect on the results, we shall ignore the integer length constraint associated with the function $L(\cdot)$ and allow any nonnegative function that satisfies Kraft's inequality.

Consider a class of information sources $\{P(\cdot|\theta)\}$ indexed by a variable $\theta \in \Lambda$. For a source $P(\cdot|\theta)$ and an encoder with length function $L(\cdot)$, the redundancy is defined as

$$R_n(L,\theta) = \frac{E[L(\boldsymbol{X})|\theta] - H(\boldsymbol{X}|\theta)}{n} \tag{3}$$

where $E[\cdot|\theta]$ denotes expectation w.r.t. $P(\cdot|\theta)$ and $H(\boldsymbol{X}|\theta)$ denotes the nth-order entropy of $P(\cdot|\theta)$, i.e.,

$$H(\boldsymbol{X}|\theta) = -\sum_{\boldsymbol{X}\in\mathcal{X}^n} P(\boldsymbol{X}|\theta)\log P(\boldsymbol{x}|\theta) \tag{4}$$

where logarithms throughout the sequel will be taken to the base 2.

Davisson [9] defined, in the context of universal coding, the *minimax* redundancy and the *maximin* redundancy in the following manner. The minimax redundancy is defined as

$$R_n^+ = \min_L \sup_{\theta\in\Lambda} R_n(L,\theta). \tag{5}$$

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

$$P_w(x^n) = \int_\Lambda w(d\theta)P_\theta(x^n). \tag{6}$$

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

$$R_n(L,w) = \int_\Lambda w(d\theta)R_n(L,\theta). \tag{7}$$

The minimum expected redundancy for a given w (which is attained by the ideal code length w.r.t the mixture, i.e., $L_w(x^n) = -\log P_w(x^n)$) is defined as

$$R_n(w) = \min_L R_n(L,w). \tag{8}$$

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

$$R_n^- = \sup_w R_n(w). \tag{9}$$

It is easy to see [9] that the maximin redundancy is identical to the capacity of the channel defined by the conditional probability measures $P(\boldsymbol{x}|\theta)$, i.e.,

$$R_n^- = C_n = \sup_w \frac{1}{n} I_w(\Theta; X^n) \tag{10}$$

where $I_w(\Theta;X^n)$ is the mutual information induced by the joint measure $w(\theta)\cdot P(\boldsymbol{x}|\theta)$. If the supremum is achieved by some prior w^* (i.e., if it is in fact a maximum), then w^* is called a *capacity-achieving prior.*[2] Gallager [13] was the first to show that if $P(\boldsymbol{x}|\theta)$ is a measurable function of θ for every $\boldsymbol{x}$ then $R_n^- = R_n^+$ and hence both are equal to C_n.

While $C_n = R_n^+$ is by definition, an attainable lower bound to $R_n(L,\theta)$ for the worst source only, it turns out to hold simultaneously for "most" points θ. Specifically, the following converse theorem to universal coding, with slight modifications in the formalism, was stated and proved in [16, Theorem 1].

Thereom 1 [16]: For every UD encoder that is independent of θ, and every positive sequence $\{\lambda_n\}$

$$R_n(L,\theta) \geq C_n - \lambda_n \tag{11}$$

for every $\theta \in \Lambda$ except for a subset $B \subseteq \Lambda$ whose probability w.r.t. w^* is less than $e \cdot 2^{-n\lambda_n}$.

The theorem is of course meaningful if $\lambda_n \ll C_n$ and, at the same time, $n\lambda_n$ tends to a large constant or even to infinity (which is possible if $nC_n \to \infty$). In this case, the lower bound on the redundancy for every $\theta \in B^c = \Lambda - B$ is essentially C_n.

In order for B^c to cover "most" sources in Λ, the capacity-achieving prior w^* must be bounded away from zero. Otherwise, the theorem, though formally correct, might be meaningless. This point is discussed extensively in [16], and it is handled in two ways. First, it is shown that a similar theorem holds for priors that nearly achieve capacity. If such a prior is also bounded away from zero (e.g., the uniform prior or Jeffreys' prior in the parametric case), then it can be used instead of w^*. Therefore, as a special case of Theorem 1, one obtains Rissanen's converse theorem to universal coding [18] for smooth parametric families with k degrees of freedom, where $C_n \approx 0.5k\log n/n$. Second, another lower bound, the random coding capacity instead of the Shannon capacity, is derived for an arbitrary prior. This bound, however, might not be tight in general. A third approach, which leads to our main results in this paper, is described in the next section.

III. Another Look at the Converse Theorem

The above discussed results not only provide performance bounds, but also indicate that an optimal universal encoder, in the sense of Theorem 1, is based on a mixture of the sources in the class w.r.t. a certain prior. It turns out, however, that the *class* of codes based on mixtures of $\{P(\cdot|\theta)\}$ is optimal in a deeper and wider sense. In [16, eq. (17)] it was shown that for every length function L that does not depend on θ, there exists a length function L' associated with some mixture over Λ, such that $R_n(L',\theta) \leq R_n(L,\theta)$ simultaneously for all $\theta \in \Lambda$. Therefore, there is no loss of optimality if universal codes are sought only among these that correspond to mixtures of $\{P(\cdot|\theta), \theta \in \Lambda\}$.

Furthermore, we next show that the redundancy of the Shannon code based on a mixture $\int_\Lambda w(d\theta)P(\boldsymbol{x}|\theta)$ with a given prior w, is optimal not only on the average w.r.t. w, but also for most θ w.r.t. w. In other words, the redundancy of any length function L is essentially lower-bounded in terms of the redundancy of L_w, which is a well-defined quantity although may not have a closed-form expression. This is more general than [16, Theorem 1] since it holds for arbitrary w, not just the capacity-achieving prior w^*. For $w = w^*$, it also leads to a considerably simpler proof of [16, Theorem 1] in some cases, e.g., when Λ is a finite set. An additional bonus is that the factor e in the upper bound on the probability of B is removed.

[2] Note that w^* may not be unique.

Theorem 2: Let $L(\cdot)$ be the length function of an arbitrary UD encoder that does not depend on θ, and let $L_w(\boldsymbol{x}) = -\log P_w(\boldsymbol{x})$ where $P_w(\cdot)$ is defined as in (6). Then, for every positive sequence $\{\lambda_n\}$

$$R_n(L,\theta) \geq R_n(L_w,\theta) - \lambda_n \tag{12}$$

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

$$w(B) = \int_B w(d\theta) \leq 2^{-n\lambda_n}. \tag{13}$$

Observe that if Λ is a finite set and $w = w^*$, the capacity-achieving prior, then $R_n(L_{w^*},\theta) = C_n$ for all $\theta \in \Lambda$ with positive prior probability [12, Theorem 4.5.1], and so, we obtain [16, Theorem 1] at least for a discrete Λ as a special case. Clarke and Barron [5], [6] have shown also that for parametric classes of memoryless sources and Jeffreys' prior w_J (which nearly attains capacity), $R_n(L_{w_J},\theta)$ coincides with C_n within a term of $O(1/n)$. Therefore, Theorem 2 extends Theorem 1 in the parametric case as well.

For choices of w that are significantly different from w^*, the redundancy $R_n(L_w,\theta)$ may depend on θ. The choice of w may depend on the desired weighting that one may wish to assign to the exceptional set B according to Theorem 2. For example, for a uniform w, the quantity $w(B)$ has the meaning of a simple relative count if Λ is discrete, or the Lebesgue measure if Λ is continuous (see also [18]).

Another way to look at Theorem 2 is in terms of the relative entropy. Since one may confine attention to length functions that satisfy Kraft's inequality with equality, then $Q(\boldsymbol{x}) = 2^{-L(\boldsymbol{x})}$ can be thought of as a probability measure and so

$$\begin{aligned} R_n(L,\theta) &= \frac{1}{n} D(P(\cdot|\theta)||Q) \\ &\triangleq \frac{1}{n} \sum_{\boldsymbol{x}\in\mathcal{X}^n} P(\boldsymbol{x}|\theta) \log \frac{P(\boldsymbol{x}|\theta)}{Q(\boldsymbol{x})}. \end{aligned} \tag{14}$$

From this point of view, Theorem 1 tells us that

$$D(P(\cdot|\theta)||Q) \geq D(P(\cdot|\theta)||P_w)$$

for most θ w.r.t. w. In words, among all fixed probability measures of n-tuples, P_w is essentially the "closest" to "most" measures in the class. This inequality, which was discussed extensively in [16], continues to hold even when $\boldsymbol{x}$ takes on values in a continuous alphabet. Therefore, it is not limited merely to the context of lossless source coding.

Proof of Theorem 1: By Kraft's inequality,

$$\begin{aligned} 1 &\geq \sum_{\boldsymbol{x}\in\mathcal{X}^n} 2^{-L(\boldsymbol{x})} \\ &= \sum_{\boldsymbol{x}\in\mathcal{X}^n} P_w(\boldsymbol{X}) \cdot 2^{L_w(\boldsymbol{x})-L(\boldsymbol{x})} \\ &= \int_\Lambda w(d\theta) \sum_{\boldsymbol{X}\in\mathcal{X}^n} P(\boldsymbol{x}|\theta) 2^{L_w(\boldsymbol{X})-L(\boldsymbol{X})} \\ &\geq \int_\Lambda w(d\theta) 2^{E\{L_w(\boldsymbol{X})|\theta\}-E\{L(\boldsymbol{X})|\theta\}} \\ &= \int_\Lambda w(d\theta) 2^{n[R_n(L_w,\theta)-R_n(L,\theta)]} \end{aligned} \tag{15}$$

where the second inequality follows from the convexity of the function $f(u) = 2^u$ and Jensen's inequality. Finally, by Markov's inequality and (15), we have

$$\begin{aligned} w(B) &= w\{\theta: 2^{n[R_n(L_w,\theta)-R_n(L,\theta)]} > 2^{n\lambda_n}\} \\ &\leq \frac{\int_\Lambda w(d\theta) 2^{n[R_n(L_w,\theta)-R_n(L,\theta)]}}{2^{n\lambda_n}} \leq 2^{-n\lambda_n}. \end{aligned} \tag{16}$$

□

The proof of Theorem 2 can be viewed as an extended version of a simple technique [1] for proving the competitive optimality property [7]. Competitive optimality means that the Shannon code length is not only optimum in the expected length sense, but it also outperforms, within c bits, any other length function with probability at least $1 - 2^{-c}$. More precisely, if $L_*(\boldsymbol{x}) = -\log P(\boldsymbol{x})$ for a given source P, then for any other UD code with length function L, Kraft's inequality implies (similarly as above) that

$$1 \geq \sum_{\boldsymbol{x}} P(\boldsymbol{x}) 2^{L_*(\boldsymbol{x})-L(\boldsymbol{x})}$$

which, in turn, by Markov's inequality, leads to

$$\Pr\{L_*(\boldsymbol{x}) > L(\boldsymbol{x}) + c\} \leq 2^{-c}$$

for all c. The above proof of the universal coding result just contains a refinement that the expectation w.r.t. $\boldsymbol{x}$ is raised to the exponent, while the expectation w.r.t. θ is kept intact.

In the other direction, as will be demonstrated in the next section, the proof of Theorem 2 is easy to extend to hierarchical structures of classes of information sources.

IV. Two-Stage Mixtures are Optimal for Hierarchical Coding

Consider a sequence of classes of sources, $\Lambda_1, \Lambda_2, \cdots, \Lambda_{M_n}$. The number of classes M_n may be finite and fixed, or growing with n, or even countably infinite for all n. We know that the active source $P(\cdot|\theta)$ belongs to one of the classes Λ_i but we do not know i in advance. In view of the above findings, if one views this problem just as universal coding w.r.t. the union of classes $\Lambda = \cup_i \Lambda_i$, then the redundancy would be the capacity $C_n(\Lambda)$ associated with Λ. For example, if $\Lambda_i, 1 \leq i \leq M_n$ is the class of all finite-state sources with i states, then $C_n(\Lambda)$ is essentially the same as the redundancy associated with the maximum number of states M_n. Obviously, it is easy to do better than that as there are many ways to approach the capacity $C_n(\Lambda_i)$ of the class corresponding to the active source.

One conceptually simple approach is to apply a two-part code described as follows: For a given i, the first part (the header) encodes the index i using some prior on the integers $\{\pi_i\}$, and the second part implements $L_{w_i^*}$, which corresponds to the capacity-achieving prior w_i^* of Λ_i. The value of i is chosen so as to minimize the total length of the code. By doing this, one achieves redundancy essentially as small as $C_n(\Lambda_i) + (\log 1/\pi_i)/n$. This method, however, requires a comparison between competing codes for all $\{i\}$, or an estimator for i (e.g., the minimum description length estimator

[19]). It is not clear, however, whether this yields the best achievable redundancy in general.

In view of the optimality of the Bayesian approach for a single class, a natural alternative is to use a code that corresponds to a two-stage mixture, first over each Λ_i and then over $\{i\}$, which is obviously equivalent to a certain mixture over the entire set Λ. This idea has been first proposed by Ryabko [23] for the hierarchy of Markov sources. A simple observation is the following. Let w_i denote a prior on Λ_i and let $\pi = \{\pi_i\}$ denote a prior on the integers $1 \leq i \leq M_n$. Now, let

$$P_{w_i}(\boldsymbol{x}) = \int_{\Lambda_i} w_i(d\theta)P(\boldsymbol{x}|\theta) \tag{17}$$

$$P_\pi(\boldsymbol{x}) = \sum_{i=1}^{M_n} \pi_i P_{w_i}(\boldsymbol{x}) \tag{18}$$

and

$$L_\pi(\boldsymbol{x}) = -\log P_\pi(\boldsymbol{x}). \tag{19}$$

Since $P_\pi(\boldsymbol{x}) \geq \pi_i P_{w_i}(\boldsymbol{x})$, then by choosing $w_i = w_i^*$, the resulting redundancy would be essentially upper-bounded by that of the above described two-part code. In other words, the mixture approach is at least as good as the two-part approach.

But as discussed in the beginning of Section III, the optimality of the mixture approach follows from deeper considerations, which are relevant to the hierarchical setting as well. Indeed, by a simple extension of the proof of Theorem 2 above, we show that $L_\pi(\boldsymbol{x})$ for arbitrary weighting is essentially optimum for "most" sources of "most" classes w.r.t. this weighting.

Theorem 3: Let $L(\cdot)$ be the length function of an arbitrary UD encoder that does not depend on θ or i, and let $\{\lambda_n\}$ be a positive sequence. Then for every i, except for a subset of $\{1, 2, \cdots, M_n\}$ whose total weight w.r.t. π is less than $2^{-n\lambda_n}$, Λ_i has the following property:

$$R_n(L, \theta) \geq R_n(L_\pi, \theta) - 2\lambda_n \tag{20}$$

for every $\theta \in \Lambda_i$ except for points a subset $B_i \subseteq \Lambda_i$ where

$$w_i(B_i) \leq 2^{-n\lambda_n}. \tag{21}$$

Proof: Similarly as in the proof of Theorem 2, we obtain

$$\sum_{i=1}^{M_n} \pi_i \int_{\Lambda_i} w_i(d\theta)P(\boldsymbol{x}|\theta)2^{n[R_n(L_\pi,\theta)-R_n(L,\theta)]} \leq 1. \tag{22}$$

Thus by Markov's inequality

$$\int_{\Lambda_i} w_i(d\theta)P(\boldsymbol{x}|\theta)2^{n[R_n(L_\pi,\theta)-R_n(L,\theta)]} \leq 2^{n\lambda_n} \tag{23}$$

for all i except for a subset of integers in $\{1, 2, \cdots, M_n\}$ whose total weight w.r.t. π is less than $2^{-n\lambda_n}$. Now, for every nonexceptional i, we have by another application of Markov's inequality

$$w_i\{\theta \in \Lambda_i: R_n(L_\pi, \theta) \geq R_n(L, \theta) + 2\lambda_n\} \leq 2^{-n\lambda_n}. \tag{24}$$

□

Let us take a closer look at the redundancy of the two-stage mixture code $R_n(L_\pi, \theta)$.

$$\begin{aligned} nR_n(L_\pi, \theta) &= E[-\log P_\pi(\boldsymbol{X})|\theta] - E[-\log P(\boldsymbol{X}|\theta)|\theta] \\ &= (E[-\log P_{w_i}(\boldsymbol{X})|\theta] - E[-\log P(\boldsymbol{X}|\theta)|\theta]) \\ &\quad + (E[-\log P_\pi(\boldsymbol{X})|\theta] - E[-\log P_{w_i}(\boldsymbol{X})|\theta]) \\ &= nR_n(L_{w_i}, \theta) + E\left[\log \frac{P_{w_i}(\boldsymbol{X})}{P_\pi(\boldsymbol{X})}\middle|\theta\right] \end{aligned} \tag{25}$$

where L_{w_i} is the length function of the Shannon code w.r.t. P_{w_i}. Thus the redundancy of L_π is decomposed into two terms. The first is $R_n(L_{w_i}, \theta)$, the redundancy within Λ_i, and the second is

$$r_n(\theta) \triangleq \frac{1}{n}E\left[\log \frac{P_{w_i}(\boldsymbol{X})}{P_\pi(\boldsymbol{X})}\middle|\theta\right]. \tag{26}$$

As mentioned earlier, since $P_\pi(\boldsymbol{x})$ is never smaller than $\pi_i P_{w_i}(\boldsymbol{x})$, it is readily seen that

$$\sup_{\theta \in \Lambda_i} r_n(\theta) \leq n^{-1}\log(1/\pi_i).$$

In the next section, we show that if the classes are efficiently *distinguishable* upon observing $\boldsymbol{x}$ by a good estimator of i, then not only is this bound tight, but moreover, $r_n(\theta) \approx n^{-1}\log(1/\pi_i)$ for "most" θ w.r.t w_i.

Returning to the general case, a natural question that arises at this point is how to choose the priors $\{w_i\}$ and π. There are two reasonable guidelines that we may suggest. The first is to put more mass on sources and classes which are considered "more important" in the sense of Theorem 3. If all classes and all sources in each class are equally important, use uniform distributions. A second reasonable choice (for the same reasons as explained in [16]) is $w_i = w_i^*$ for all i, and $\pi = \pi^*$, where π^* achieves the capacity c_n of the "channel" from i to $\boldsymbol{x}$, as defined by $P_{w_i^*}(\boldsymbol{x})$. Note that in this case, since the expectation of $r_n(\theta)$ w.r.t. w_i^* is c_n [12, Theorem 4.5.1], we have

$$\sup_{\theta \in \Lambda_i} R_n(L, \theta) \geq C_n(\Lambda_i) + c_n \tag{27}$$

for all i with $\pi_i^* > 0$. Namely, the maximum redundancy is lower-bounded by the sum of two capacity terms: the *intra-class capacity* $C_n(\Lambda_i)$ associated with universality within each class, and the *inter-class capacity* c_n, which is the cost attributed to the lack of knowledge of i.

In Section VI, we provide the example of finite-state (FS) arbitrarily varying sources (AVS's), where inequality (27) becomes an equality for every source θ in the class. This happens because in the special case of the AVS, $r_n(\theta)$ turns out to be independent of θ and so

$$r_n(\theta) = \sum_{\theta' \in \Lambda_i} w_i(\theta')r_n(\theta') = c_n$$

for all θ.

V. Distinguishable Classes of Sources

It was mentioned earlier that

$$\sup_{\theta\in\Lambda_i} r_n(\theta) \le n^{-1}\log(1/\pi_i).$$

An interesting question is: under what conditions exactly is this bound tight?

To answer this question, we pause for a moment from our original problem and consider the problem of universal coding for a class with a countable number of sources defined by arbitrary PMF's on $\mathcal{X}^n$, denoted $Q(\cdot|i), i=1,2,\cdots,M_n$. In the next lemma, we provide bounds on the redundancy of the mixture

$$Q_\pi(\boldsymbol{x}) = \sum_i \pi_i Q(\boldsymbol{x}|i)$$

w.r.t. every $Q(\cdot|i)$. Let $g:\mathcal{X}^n \to \{1,2,\cdots,M_n\}$ denote an arbitrary estimator of the index i of $Q(\cdot|i)$, and let $Q(e|i) = Q\{\boldsymbol{x}: g(\boldsymbol{x}) \ne i|i\}$ denote the error probability given i. Similarly, let $Q(c|i) = 1 - Q(e|i)$, and

$$Q(e) = \sum_i \pi_i Q(e|i)$$

for the given prior π. Then, we have the following result:

Lemma 1: For every estimator g and every $1 \le i \le M_n$

$$\log\left(\frac{1}{\pi_i}\right) \ge nD(Q(\cdot|i)||Q_\pi) \ge Q(c|i)\log\left[\frac{Q(c|i)}{\pi_i + Q(e)}\right] + Q(e|i)\log Q(e|i). \tag{28}$$

The proof appears in Appendix I.

The lemma tells us that if there exists a consistent estimator g, i.e., $Q(e|i)$ for every i, and so $Q(e)$, tend to zero as $n\to\infty$, then the rightmost side tends to $\log(1/\pi_i)$ and hence so does $nD(Q(\cdot|i)||Q_\pi)$. In other words, for a discrete set of sources $\{Q(\cdot|i)\}$ that are *distinguishable* upon observing $\boldsymbol{x}$ by some decision rule g, the redundancy of the mixture Q_π w.r.t. $Q(\cdot|i)$ behaves like $n^{-1}\log(1/\pi_i)$ for large n.

The relevance of this lemma to our problem becomes apparent by letting $Q(\boldsymbol{x}|i) = P_{w_i}(\boldsymbol{x})$, and then $Q(e|i)$ is interpreted as the average error probability given Λ_i w.r.t. w_i. Specifically, for a given $\theta\in\Lambda_i$, let us denote

$$P(e|\theta) = \sum_{\boldsymbol{x}:g(\boldsymbol{x})\ne i} P(\boldsymbol{x}|\theta) \tag{29}$$

$$\overline{P}(e|i) = Q(e|i) = \int_{\Lambda_i} w_i(d\theta)P(e|\theta) \tag{30}$$

$$\overline{P}(c|i) = Q(c|i) = 1 - \overline{P}(e|i) \tag{31}$$

and

$$\overline{P}(e) = Q(e) = \sum_{i=1}^{M_n} \pi_i \overline{P}(e|i). \tag{32}$$

We also note that this substitution gives

$$\int_{\Lambda_i} w_i(d\theta)r_n(\theta) = D(Q(\cdot|i)||Q_\pi) \tag{33}$$

and so it immediately leads to the following corollary to Lemma 1.

Corollary 1: For every estimator g and every $1 \le i \le M_n$

$$\log\left(\frac{1}{\pi_i}\right) \ge n\int_{\theta\in\Lambda_i} w_i(d\theta)r_n(\theta) \ge \overline{P}(c|i)\log\left[\frac{\overline{P}(c|i)}{\pi_i + \overline{P}(e)}\right] + \overline{P}(e|i)\log\overline{P}(e|i). \tag{34}$$

The corollary tells us that if there exists an index estimator g that is *consistent for "most"* $\theta\in\Lambda_i, i=1,2,\cdots,M_n$, in the sense that for every $\epsilon>0, w_i\{\theta: P(e|\theta)\ge\epsilon\}\to 0$ as $n\to\infty$, then the lower bound will be essentially $\log(1/\pi_i)$.

A common example is where Λ_i is the class of all unifilar finite-state sources with i states. A unifilar finite-state source is characterized by

$$P(\boldsymbol{x}|\theta) = \prod_{t=1}^{n} p(x_i|s_i)$$

where $\theta = \{p(x|s)\}$, $\boldsymbol{s} = (s_1,\cdots,s_n)$ is a state sequence whose elements are taking values in $\{1,\cdots,i\}$, and $s_t, t=2,3,\cdots$, is given by a deterministic function of x_{t-1} and s_{t-1}, while the initial state s_1 is assumed fixed. In this example, there is a consistent estimator [24] for i provided that M_n is fixed or grows sufficiently slowly with n. (See also Hannan and Quinn [14], Kieffer [15], and Rudich [21] for earlier related work on model order selection.) It should be pointed out that in [24] it has not been established explicitly that $\overline{P}(e)\to 0$ for the model estimator proposed therein. Nevertheless, this can be easily deduced from the following consideration: For every $\epsilon>0$, the set $\{\theta\in\Lambda_i: P(e|\theta)\ge\epsilon\}$ has a vanishingly small probability w.r.t. w_i as $n\to\infty$, provided that w_i does not put too much mass near the boundaries between Λ_i and Λ_{i-1}.

Let us denote the lower bound of Corollary 1 by $[\log(1/\pi_i) - \epsilon_n(i)]$, i.e.,

$$\epsilon_n(i) = \log\frac{1}{\pi_i} - \overline{P}(c|i)\log\left[\frac{\overline{P}(c|i)}{\pi_i+\overline{P}(e)}\right] - \overline{P}(e|i)\log\overline{P}(e|i) \tag{35}$$

keeping in mind that if the classes $\{\Lambda_i\}$ are distinguishable in the sense that such an estimator g exists, then $\epsilon_n(i)\to 0$ for every fixed i. There are two immediate conclusions from Corollary 1. First, it implies that

$$n\sup_{\theta\in\Lambda_i} r_n(\theta) \ge \log(1/\pi_i) - \epsilon_n(i)$$

and since we have already seen that

$$n\sup_{\theta\in\Lambda_i} \le \log(1/\pi_i)$$

we conclude that

$$n\sup_{\theta\in\Lambda_i} r_n(\theta) \approx \log(1/\pi_i).$$

Second, since the supremum is *upper*-bounded by $\log(1/\pi_i)$, while the expectation is *lower*-bounded by $\log(1/\pi_i) - \epsilon_n(i)$, then obviously, "most" points in Λ_i must have $nr_n(\theta) \approx \log(1/\pi_i)$. More precisely, for $\Delta > 0$ let

$$S = \left\{\theta: nr_n(\theta) < \log\frac{1}{\pi_i} - \Delta\right\}. \tag{36}$$

Then, we have

$$\begin{aligned}\log\frac{1}{\pi_i} - \epsilon_n(i) &\le \int_S w_i(d\theta)nr_n(\theta) + \int_{S^c} w_i(d\theta)nr_n(\theta)\\ &\le w_i(S)\left(\log\frac{1}{\pi_i} - \Delta\right) + [1 - w_i(S)]\log\frac{1}{\pi_i}\end{aligned} \tag{37}$$

which implies that

$$w_i(S) \le \frac{\epsilon_n(i)}{\Delta}. \tag{38}$$

By combining Theorem 3, where $\lambda_n = \lambda/n$ ($\lambda > 0$), and (38), both with $w_i = w_i^*$ for all i, we obtain a lower bound on the redundancy of an arbitrary UD encoder with length function L. Specifically

$$\begin{aligned}R_n(L,\theta) &\ge R_n(L_{w_i^*},\theta) + r_n(\theta) - \frac{2\lambda}{n}\\ &\ge C_n(\Lambda_i) + r_n(\theta) - \frac{3\lambda}{n}\\ &\ge C_n(\Lambda_i) + \frac{1}{n}\left(\log\frac{1}{\pi_i} - \Delta - 3\lambda\right).\end{aligned} \tag{39}$$

The first inequality, which is a restatement of Theorem 3, applies to "most" sources w.r.t. w_i^* of "most" classes w.r.t. π. The second inequality, which follows from Theorem 1, and the third inequality, which we have now established, both hold for "most" $\theta \in \Lambda_i$ w.r.t. w_i^*.

Thus we have just proved the following Theorem, which provides a lower bound for hierarchical universal coding, for the case of distinguishable classes of sources.

Theorem 4: Let g be an estimator of the index of the class such that

$$\overline{P}^*(e|i) = \int_{\Lambda_i} w_i^*(d\theta)P(e|\theta) \to 0$$

as $n \to \infty$ uniformly for all $1 \le i \le M_n$. Let L be the length function of an arbitrary UD encoder that does not depend on θ or i, and let $\lambda > 0$ and $\Delta > 0$ be arbitrary constants. Then, for every i, except for a subset of $\{1, 2, \cdots, M_n\}$ whose total weight w.r.t. π is less than $2^{-\lambda}$, every class Λ_i has the following property:

$$R_n(L,\theta) \ge C_n(\Lambda_i) + \frac{1}{n}\left(\log\frac{1}{\pi_i} - \Delta - 3\lambda\right) \tag{40}$$

for every $\theta \in \Lambda_i$ except for points in a subset $B_i \subset \Lambda_i$ such that

$$w_i^*(B_i) \le 2^{-(\lambda-1)} + \frac{\epsilon_n^*(i)}{\Delta} \tag{41}$$

where $\epsilon_n^*(i)$ is defined as in (35), with average error probabilities being defined w.r.t. $\{w_i^*\}$.

Again, as mentioned after Theorem 1, it should be kept in mind that if necessary, each w_i^* can be essentially replaced by a prior that is bounded away from zero, and at the same time, nearly achieves $C_n(\Lambda_i)$ (see also [16]).

The second term of the lower bound might not be meaningful if $\log(1/\pi_i)$ is of the same order of magnitude as $\Delta + 3\lambda$, which in turn should be reasonably large so as to keep the mass of B_i small. However, if we fix λ and Δ so that $w_i^*(B_i)$ would be fairly small, say 0.01, and if M_n is very large (M_n may tend to infinity), then for most classes (in the uniform counting sense), π_i must be very small, and so $\log(1/\pi_i)$ would be large compared to $\Delta + 3\lambda$. Thus the assertion of the theorem is meaningful if π is chosen such that for "most" values of i w.r.t. π, $\log(1/\pi_i)$ is large. This can happen only if π has a large entropy, i.e., it is close to the uniform distribution in some sense. Of course, if π is exactly uniform then $\log(1/\pi_i) = \log M_n$ for all i. This interpretation of Theorem 4, however, should be taken carefully, because if i is allowed to grow with n, and hence π_i decays with n, then $\epsilon_n^*(i)$ is small only if

$$\overline{P}^*(e) = \sum_i \pi_i \overline{P}^*(e|i)$$

is small compared to π_i (see Corollary 1). In other words, Theorem 4 is meaningful only for i that is sufficiently small compared to n. This is guaranteed for all i when M_n grows sufficiently slowly.

Roughly speaking, the theorem tells us that if the classes $\{\Lambda_i\}$ are distinguishable in the sense that there exists a good estimator g, then the minimum achievable redundancy is approximately

$$C_n(\Lambda_i) + \frac{1}{n}\log\frac{1}{\pi_i}. \tag{42}$$

Note that if, in addition, $\{\pi_i\}$ is a monotonically nonincreasing sequence, then $\pi_i \le 1/i$, and so $\log(1/\pi_i)$ is further lower-bounded by $\log i$. This is still nearly achievable by assigning the universal prior on the integers or $\pi_i \propto 1/i^{1+\delta}$ where $\delta > 0$ if $M_n = \infty$. This means that

$$C_n(\Lambda_i) + \frac{\log i}{n} \tag{43}$$

is the minimum attainable redundancy w.r.t. any monotone weighting of the indices $\{i\}$.

The minimum redundancy (42) is attained by a two-stage mixture where $w_i = w_i^*$. The choice of π, in this case, can be either based on the guidelines provided in the previous section or on the following consideration: We would like the extra redundancy term $\log(1/\pi_i)$ to be a small fraction of the first redundancy term $C_n(\Lambda_i)$ that we must incur anyhow. Specifically, if possible, we would like to choose $n^{-1}\log(1/\pi_i) \approx \epsilon C_n(\Lambda_i)$, which leads to

$$\pi_i = \frac{2^{-\epsilon n C_n(\Lambda_i)}}{K_n(\epsilon)} \tag{44}$$

where $K_n(\epsilon)$ is a normalizing factor. This means that the rich and complex classes are assigned a smaller prior probability. The redundancy would then be $(1+\epsilon)C_n(\Lambda_i) + n^{-1}\log K_n(\epsilon)$, where now the second term does not depend on i. For example, if Λ_i is the class of ith-order Markov sources, then

$$C_n(\Lambda_i) \approx 0.5A^i(A-1)\log n/n$$

(see, e.g., [10], [20]), and so

$$\pi_i = \frac{\exp_2\left[-\frac{1}{2}\epsilon A^i(A-1)\log n\right]}{K_n(\epsilon)} = \frac{n^{-0.5\epsilon A^i(A-1)}}{K_n(\epsilon)}. \quad (45)$$

As for the normalization factor

$$\begin{aligned} K_n(\epsilon) &\le \sum_{i=0}^{\infty} n^{-0.5\epsilon A^i(A-1)} \\ &\le \sum_{i=1}^{\infty} n^{-0.5\epsilon i} \\ &= \frac{1}{n^{0.5\epsilon}-1} \to 0 \end{aligned} \quad (46)$$

and therefore the term $n^{-1}\log K_n(\epsilon)$ has a negative contribution. Note that if $M_n < \infty$ and ϵ is chosen very small (so that the coefficient in front of $C_n(\Lambda_i)$ would be close to unity), then π is close to uniform. This agrees with the conclusion of our earlier discussion that π should be uniform or nearly uniform.

We have mentioned before the hierarchy of classes of unifilar finite-state sources as an example where the classes are distinguishable. In the next section, we examine another example—FS AVS's—where the natural hierarchical partition does not yield distinguishable classes, yet the universal coding redundancy can be characterized quite explicitly.

VI. Arbitrarily Varying Sources

An FS AVS is a nonstationary memoryless source characterized by the PMF

$$P(\boldsymbol{X}|\boldsymbol{s}) = \prod_{i=1}^{n} p(x_i|s_i) \quad (47)$$

where $\boldsymbol{x} = (x_1, \cdots, x_n)$ is again the source sequence to be encoded, and $\boldsymbol{s} = (s_1, \cdots, s_n)$ is an unknown *arbitrary* sequence of states corresponding to $\boldsymbol{x}$, where each s_i takes on values in a finite set $\mathcal{S}$. We shall assume, for the sake of simplicity, that the parameters of the AVS $\{p(x|s)\}_{x\in\mathcal{X},s\in\mathcal{S}}$ are known, and then only universality w.r.t. the unknown state sequence will be studied. This is clearly a special case of our problem with $\theta = \boldsymbol{s}$ and $\Lambda = \mathcal{S}^n$.

Obviously, since C_n for all n is given by the capacity C of the memoryless channel $p(x|s)$, it does not vanish with n and so, universal coding in the sense of approaching the entropy, is not feasible for this large class of sources. Yet, universal coding in the sense of attaining the lower bound remains a desirable goal. The capacity-achieving prior on $\mathcal{S}^n$ is the i.i.d. measure w^* that achieves the capacity of the memoryless channel (47). Therefore, most of the mass is assigned by w^* to state sequences whose empirical distributions are close to w^*. Consequently, if $\Lambda = \mathcal{S}^n$ is treated as one big class of sources, [16, Theorem 1] and Theorem 2 herein tell us very little about the redundancy incurred at all other state sequences. We are then led to treat separately each type class of state sequences with the same empirical distribution, in other words, to use the hierarchical approach.

We, therefore, pause to provide a few definitions associated with type classes. For a given state sequence $\boldsymbol{s} \in \mathcal{S}^n$, the empirical PMF is the vector $w_{\boldsymbol{s}} = \{w_{\boldsymbol{s}}(s), s \in \mathcal{S}\}$ where $w_{\boldsymbol{s}}(s) = n_{\boldsymbol{s}}(s)/n$, $n_{\boldsymbol{s}}(s)$ being the number of occurrences of the state $s \in \mathcal{S}$ in the sequence $\boldsymbol{s}$. The set of all empirical PMF's of sequences $\boldsymbol{s}$ in $\mathcal{S}^n$, i.e., rational PMF's with denominator n, will be denoted by $\mathcal{P}_n$. The type class $T_{\boldsymbol{s}}$ of a state sequence $\boldsymbol{s}$ is the set of all state sequences $\boldsymbol{s}' \in \mathcal{S}^n$ such that $w_{\boldsymbol{s}'} = w_{\boldsymbol{s}}$. We shall also denote the type classes of state sequences by $\{T_i\}$ where the index i is w.r.t. some arbitrary but fixed ordering in $\mathcal{P}_n$.

We will now consider $\Lambda = \mathcal{S}^n$ as the union of all type classes $\Lambda_i = T_i, i = 1, 2, \cdots, M_n = |\mathcal{P}_n|$. Note that since the empirical PMF of $\boldsymbol{s}$ can be estimated with precision no better than $O(1/\sqrt{n})$, it is clear that in this case, the assumption on a good estimator of the exact class $\Lambda_i = T_i$ is not met. Therefore, we are led to use one of the guidelines described in Section IV regarding the choice of the priors. (We will elaborate on this point at the end of this section.)

Let us focus on the two-stage mixture code L_π, where $w_i = w_i^*$ attains the capacity within each type class. Following [12, Theorem 4.5.2], it is readily seen that the intra-class capacity $C_n(T_i)$ is attained by a uniform distribution on T_i, i.e.,

$$w_i^*(\boldsymbol{s}) = u_i(\boldsymbol{s}) \triangleq \begin{cases} \frac{1}{|T_i|}, & \boldsymbol{s} \in T_i \\ 0, & \text{elsewhere.} \end{cases} \quad (48)$$

It is shown in Appendix II that if T_i corresponds to an empirical PMF on $\mathcal{S}$ that tends to a fixed PMF $w = \{w(s), s \in \mathcal{S}\}$, then $C_n(T_i)$ tends to

$$I_w(S;X) \triangleq \sum_{s\in\mathcal{S}} w(s) \sum_{x\in\mathcal{X}} p(x|s) \log \frac{p(x|s)}{\sum_{s'\in\mathcal{S}} w(s')p(x|s')}. \quad (49)$$

The second redundancy term $r_n(\boldsymbol{s}) = r_n(\theta)$ associated with L_π is given by

$$r_n(\boldsymbol{s}) = \frac{1}{n} E\left[\log \frac{P_{u_i}(\boldsymbol{X})}{P_\pi(\boldsymbol{X})}\middle| \boldsymbol{s}\right] \quad (50)$$

where

$$P_{u_i}(\boldsymbol{x}) = \frac{1}{|T_i|} \sum_{\boldsymbol{s}\in T_i} P(\boldsymbol{x}|\boldsymbol{s}). \quad (51)$$

Observe that $P_{u_i}(\boldsymbol{x})$, and hence also $P_\pi(\boldsymbol{x})$ (which is a mixture of $\{P_{u_i}(\boldsymbol{x})\}$), are invariant to permutations of $\boldsymbol{x}$. Consequently, the expectation on the right-hand side of (50) is the same for all $\boldsymbol{s} \in T_i$, and so, the second-order redundancy term $r_n(\boldsymbol{s})$ is exactly the normalized divergence between P_{u_i} and P_π. If, in addition, $\pi = \pi^*$, the capacity-achieving prior of the channel from i to $\boldsymbol{x}$ defined by $\{P_{u_i}(\boldsymbol{x})\}$, then this divergence coincides with the capacity c_n of this channel for every i with $\pi_i^* > 0$. Clearly

$$c_n \le n^{-1}\log|\mathcal{P}_n| = O(\text{log}n/n).$$

In summary, for AVS's it is natural to apply Theorem 3 with uniform weighting within each type. The best attainable compression ratio (in the sense of Theorem 3) is given by $H(\boldsymbol{X}|\boldsymbol{s})/n + C_n(T_{\boldsymbol{s}}) + c_n$, where

$$H(\boldsymbol{X}|\boldsymbol{s}) = -n\sum_{s\in\mathcal{S}} w_{\boldsymbol{s}}(s)\sum_{x\in\mathcal{X}} p(x|s)\log p(x|s). \tag{52}$$

While the third term c_n decays at the rate of $\log n/n$, the first two terms tend to constants if $w_{\boldsymbol{s}}$ tends to a fixed w. The sum of these constants is $H_w(X)$, the entropy of a memoryless source with letter probabilities given by

$$p_w(x) = \sum_{s\in\mathcal{S}} w(s)p(x|s). \tag{53}$$

This is different from earlier results on source coding for the AVS due to Berger [4] and Csiszár and Körner [8], who considered fixed-length rate-distortion codes that satisfy an average distortion constraint for every state sequence. In their setting, for the distortionless case, the best achievable rate is $\max_w H_w(X)$. Thus our results coincide with the earlier result only if the underlying state sequence happens to belong to the type that corresponds to the worst empirical PMF that maximizes $H_w(X)$. In other words, by using the hierarchical approach and allowing variable-length codes, we enable "adaptation" to the unknown underlying state sequence rather than using the worst case strategy.

We have then, both improved the main redundancy term and characterized the best attainable second-order performance in the sense of Theorem 3.

An interesting special case is where $\mathcal{S} = \mathcal{X}$ and $p(x|s) = 1$ if $x = s$ and zero otherwise, in other words, $\boldsymbol{x}$ is always identical to $\boldsymbol{s}$. In this case, $H(\boldsymbol{X}|\boldsymbol{s}) = 0$. If, in addition, $\boldsymbol{x}$ is such that the relative frequencies of all letters are bounded away from zero, then

$$C_n(T_{\boldsymbol{s}}) = \frac{\log|T_{\boldsymbol{s}}|}{n} \approx H_{\boldsymbol{X}}(X) - \frac{(|\mathcal{X}|-1)}{2n}\log n \tag{54}$$

where $H_{\boldsymbol{x}}(X)$ is the entropy associated with the empirical PMF of $\boldsymbol{x}$ and

$$c_n = \frac{\log|\mathcal{P}_n|}{n} \approx (|\mathcal{X}|-1)\frac{\log n}{n}. \tag{55}$$

Therefore, we conclude that the total minimum description length (MDL) is approximately

$$nH_{\boldsymbol{X}}(X) + \frac{(|\mathcal{X}|-1)}{2}\log n \tag{56}$$

in the deterministic sense. This coincides with a special case of one of the main results in [25], where optimum length functions assigned by sequential finite state machines for individual sequences were investigated, and the above minimum length corresponds to a single-state machine.

Finally, the following comment is in order. We mentioned earlier that the exact index i of T_i cannot be estimated by observing $\boldsymbol{x}$ and hence Theorem 4 is inapplicable. Nevertheless, if $|\mathcal{S}| \le |\mathcal{X}|$ and the rank of transition probability matrix $\{p(x|s)\}$ is $|\mathcal{S}|$, then the empirical PMF of $\boldsymbol{s}$ can be estimated with precision $O(1/\sqrt{n})$. This can be done by solving the linear equations

$$\sum_{s\in\mathcal{S}} w_{\boldsymbol{s}}(s)p(a|s) = q_{\boldsymbol{x}}(a), \quad a\in\mathcal{X}$$

where $q_{\boldsymbol{x}}(a)$ is the relative frequency of a in $\boldsymbol{x}$. This means that if we define Λ_i as unions of all neighboring type classes whose corresponding empirical PMF's differ by $O(1/\sqrt{n})$, then the assumption about the existence of a good estimator becomes valid. In this case, it is difficult, however, to determine w_i^* and to assess the redundancy term $C_n(\Lambda_i)$.

Appendix I
Proof of Lemma 1

The first inequality is obvious since $Q_\pi(\boldsymbol{x}) \ge \pi_i Q(\boldsymbol{x}|i)$ for every $\boldsymbol{x}$ and every i. As for the second inequality, let us denote by Ω_j the set of all $\boldsymbol{x} \in \mathcal{X}^n$ for which $g(\boldsymbol{x}) = j$, and let Ω_j^c denote the complementary set. Since data processing cannot increase the relative entropy, $D(Q(\cdot|i)||Q_\pi)$ is lower-bounded by

$$D(Q(\cdot|i)||Q_\pi) \ge Q(\Omega_i|i)\log\frac{Q(\Omega_i|i)}{Q_\pi(\Omega_i)} + Q(\Omega_i^c|i)\log\frac{Q(\Omega_i^c|i)}{Q_\pi(\Omega_i^c)}. \tag{A1}$$

The proof is now completed by observing that

$$Q(c|i) = Q(\Omega_i|i), Q(e|i) = Q(\Omega_i^c|i), Q_\pi(\Omega_i^c) \le 1$$

and

$$\begin{aligned} Q_\pi(\Omega_i) &= \sum_j \pi_j Q(\Omega_i|j) \\ &\le \pi_i + \sum_{j\ne i} \pi_j Q(\Omega_i|j) \\ &\le \pi_i + \sum_i\sum_{j\ne i} \pi_j Q(\Omega_i|j) \\ &= \pi_i + Q(e). \end{aligned} \tag{A2}$$

Appendix II
Asymptotic Behavior of $C_n(T_{\boldsymbol{s}})$

In this Appendix, we prove that if $w_{\boldsymbol{s}}$ tends to a fixed PMF w on $\mathcal{S}$, then $C_n(T_i)$ of the corresponding type $T_i = T_{\boldsymbol{s}}$ tends to $I_w(X;S)$. The quantity $C_n(T_i)$ is given by

$$C_n(T_i) = \frac{1}{n}\left[H_i(\boldsymbol{X}) - \frac{1}{|T_i|}\sum_{\boldsymbol{s}\in T_i} H(\boldsymbol{X}|\boldsymbol{s})\right] \tag{A3}$$

where $H_i(\boldsymbol{X})$ is the entropy associated with n-vectors governed by

$$P_{u_i}(\boldsymbol{x}) = \frac{1}{|T_i|}\sum_{\boldsymbol{s}\in T_i} P(\boldsymbol{x}|\boldsymbol{s}) \tag{A4}$$

and

$$\begin{aligned} H(\boldsymbol{X}|\boldsymbol{s}) &= -\sum_{\boldsymbol{x}\in\mathcal{X}^n} P(\boldsymbol{x}|\boldsymbol{s})\log P(\boldsymbol{x}|\boldsymbol{s}) \\ &= -n\sum_{s\in\mathcal{S}} w_{\boldsymbol{s}}(s)\sum_{x\in\mathcal{X}} p(x|s)\log p(x|s). \end{aligned} \tag{A5}$$

Since $H(\boldsymbol{X}|\boldsymbol{s})/n$ is the same for all $\boldsymbol{s} \in T_i$, and since it tends to

$$H_w(X|S) = -\sum_{s\in\mathcal{S}} w(s) \sum_{x\in\mathcal{X}} p(x|s) \log p(x|s) \tag{A6}$$

so does the average of $H(\boldsymbol{x}|\boldsymbol{s})$ over $\boldsymbol{s} \in T_i$. Therefore, it will be sufficient to show that $H_i(\boldsymbol{X})/n$ tends to the entropy of a memoryless source with letter probabilities given by

$$p_w(x) = \sum_{s\in\mathcal{S}} w(s)p(x|s).$$

To this end, we shall introduce the following notation. Similarly as in the definition of type classes of state sequences, the empirical PMF of the sequence $\boldsymbol{x}$ will be denoted by $\{q_{\boldsymbol{x}}(x), x \in \mathcal{X}\}$, where $q_{\boldsymbol{x}}(x)$ is the relative frequency of x in $\boldsymbol{x}$. The respective type will be denoted by $T_{\boldsymbol{x}}$, and the associated empirical entropy will be denoted by $H_{\boldsymbol{x}}(X)$. For a sequence pair $(\boldsymbol{x}, \boldsymbol{s}) \in \mathcal{X}^n \times \mathcal{S}^n$ the joint empirical PMF is defined by the joint empirical PMF of x and s, and the joint type $T_{\boldsymbol{xs}}$ of $(\boldsymbol{x}, \boldsymbol{s})$ is the set of all pair sequences $(\boldsymbol{x}', \boldsymbol{s}') \in \mathcal{X}^n \times \mathcal{S}^n$ with the same empirical joint PMF as $(\boldsymbol{x}, \boldsymbol{s})$. The empirical joint entropy is denoted by $H_{\boldsymbol{xs}}(X, S)$.

A conditional type $T_{\boldsymbol{s}|\boldsymbol{x}}$ for a given $\boldsymbol{x}$ is the set of all sequences $\boldsymbol{s}'$ in $\mathcal{S}^n$ for which $(\boldsymbol{x}, \boldsymbol{s}') \in T_{\boldsymbol{xs}}$. The corresponding empirical conditional entropy is given by

$$H_{\boldsymbol{s}|\boldsymbol{x}}(S|X) = H_{\boldsymbol{xs}}(X, S) - H_{\boldsymbol{x}}(X). \tag{A7}$$

Similar definitions and notations apply when the roles of $\{x, X, \boldsymbol{x}, \mathcal{X}\}$ and $\{s, S, \boldsymbol{s}, \mathcal{S}\}$ are interchanged.

For two sequences $\{a_n\}$ and $\{b_n\}$, the notation $a_n = b_n$ means that

$$\lim_{n\to\infty} n^{-1} \log (a_n/b_n) = 0.$$

It is well known [8] that $|T_{\boldsymbol{s}}| = 2^{nH_{\boldsymbol{s}}(S)}$ and $|T_{\boldsymbol{s}|\boldsymbol{x}}| = 2^{nH_{\boldsymbol{s}|\boldsymbol{x}}(S|X)}$. Using these facts together with the fact that $P(\boldsymbol{x}|\boldsymbol{s}) \leq 2^{-nH_{\boldsymbol{x}|\boldsymbol{s}}(X|S)}$ we have

$$\begin{aligned} P_{u_i}(\boldsymbol{x}) &= \frac{1}{|T_{\boldsymbol{s}}|} \sum_{\boldsymbol{s}'\in T_{\boldsymbol{s}}} P(\boldsymbol{x}|\boldsymbol{s}') \\ &\leq \frac{1}{|T_{\boldsymbol{s}}|} \sum_{T_{\boldsymbol{s}|\boldsymbol{x}}\subset T_{\boldsymbol{s}}} |T_{\boldsymbol{s}|\boldsymbol{x}}| 2^{-nH_{\boldsymbol{x}|\boldsymbol{s}}(X|S)} \\ &= 2^{-nH_{\boldsymbol{s}}(S)} \sum_{T_{\boldsymbol{s}|\boldsymbol{x}}\subset T_{\boldsymbol{s}}} 2^{nH_{\boldsymbol{s}|\boldsymbol{x}}(S|X)} \cdot 2^{-nH_{\boldsymbol{x}|\boldsymbol{s}}(X|S)} \\ &= 2^{-nH_{\boldsymbol{x}}(X)} \end{aligned} \tag{A8}$$

where in the last step we have used the fact that the number of conditional types classes is polynomial in n. Therefore

$$-\log P_{u_i}(\boldsymbol{x}) \geq nH_{\boldsymbol{x}}(X) + o(n). \tag{A9}$$

If the empirical PMF of $\boldsymbol{s}$ tends to w, then by the strong law of large numbers, for every $\boldsymbol{s}' \in T_{\boldsymbol{s}}, q_{\boldsymbol{x}}(x) \to p_w(x)$ with probability one, and so the expected value of $H_{\boldsymbol{x}}(X)$ given every $\boldsymbol{s}' \in T_{\boldsymbol{S}}$ tends to the entropy of $\{p_w(x), x \in \mathcal{X}\}$. *A fortiori*, the overall expectation after averaging over $T_{\boldsymbol{s}}$ tends to the same entropy. Thus

$$\liminf_n H_i(\boldsymbol{X})/n \geq H_w(X).$$

For the converse inequality, note that the entropy $H_i(X)/n$ of a vector $X = (X_1, \cdots, X_n)$ governed by P_{u_i} is never larger than the average of the marginal entropies

$$n^{-1} \sum_{t=1}^{n} H(X_t).$$

Since X_t is governed by $p(\cdot|s_t)$, then by the concavity of the entropy function, the latter expression in turn, is upper-bounded by the entropy of the i.i.d. measure

$$n^{-1} \sum_{t=1}^{n} p(x|s_t) = \sum_{s\in\mathcal{S}} w_{\boldsymbol{s}}(s)p(x|s)$$

which again tends to $p_w(x)$. Thus

$$\limsup_n H_i(\boldsymbol{X})/n \leq H_w(X)$$

completing the proof of the claim.

Acknowledgment

The authors wish to thank the anonymous reviewers for their useful comments.

References

[1] A. R. Barron, "Logically smooth density estimation," Ph.D dissertation, Stanford University, Stanford, CA, 1985.

[2] A. R. Barron and T. M. Cover, "Minimum complexity density estimation," *IEEE Trans. Inform. Theory*, vol. 37, no. 4, 1034–1054, July 1991.

[3] A. R. Barron and C. H. Sheu, "Approximation of density functions by sequences of exponential families," *Ann. Statist.*, vol. 1, no. 3, pp. 1347–1369, 1991.

[4] T. Berger, *Rate Distortion Theory.* Englewood Cliffs, NJ: Prentice-Hall, 1971.

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

[6] ——, "Jeffrey's prior is asymptotically least favorable under entropy risk," *J. Statist. Plan. Inform.*, vol. 41, pp. 37–60, 1994.

[7] T. M. Cover, "On the competitive optimality of Huffman codes," *IEEE Trans. Inform. Theory*, vol. 37, no. 1, pp. 172–174, Jan. 1991.

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

[9] L. D. Davisson, "Universal noiseless coding," *IEEE Trans. Inform. Theory*, vol. IT-19, no. 6, pp. 783–795, Nov. 1973.

[10] ——, "Minimax noiseless universal coding for Markov sources," *IEEE Trans. Inform. Theory*, vol. IT-29, no. 2, pp. 211–215, Mar. 1983.

[11] 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.

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

[13] ——, "Source coding with side information and universal coding," unpublished manuscript, Sept. 1976. (Also, presented at the International Symposium on Inform. Theory, Oct. 1974.)

[14] E. J. Hannan and B. G. Quinn, "The determination of the order of an autoregression," *J. Roy Statist. Soc. Ser. B*, vol. 41, pp. 190–195, 1979.

[15] J. C. Kieffer, "Strongly consistent MDL-based of a model class for a finite alphabet source," preprint, 1993.

[16] N. Merhav and M. Feder, "A strong version of the redundancy-capacity theorem of universal coding," *IEEE Trans. Inform. Theory*, vol. 41, no. 3, pp. 714–722, May 1995.

[17] J. Rissanen, "A universal data compression system, *IEEE Trans. Inform. Theory*, vol. IT-29, no. 5, pp. 656–664, Sept. 1983.

[18] ——, "Universal coding, information, prediction, and estimation," *IEEE Trans. Inform. Theory*, vol. IT-30, no. 4, pp. 629–636, July 1984.

[19] ——, "Stochastic complexity and modeling," *Ann. Statist.*, vol. 14, pp. 1080–1100, 1986.

[20] ——, "Complexity of strings in the class of Markov sources," *IEEE Trans. Inform. Theory*, vol. IT-32, no. 4, pp. 526–532, July 1986.

[21] S. Rudich, "Inferring the structure of a Markov chain from its output," in *Proc. 26th IEEE Symp. on Foundations of Computer Science*, 1985, pp. 321–326.

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

[23] ——, "Twice-universal coding," *Probl. Inform. Transm.*, pp. 173–177, July–Sept., 1984.

[24] M. J. Weinberger and M. Feder, "Predictive stochastic complexity and model estimation for finite-state processes," *J. Statist. Plan. Inf.*, Vol. 39, pp. 353–372, 1994.

[25] M. J. Weinberger, N. Merhav, and M. Feder, "Optimal sequential probability assignment for individual sequences," *IEEE Trans. Inform. Theory*, vol. 40, no. 2, pp. 384–396, Mar. 1994.

[26] F. M. J. Willems, Y. M. Shtarkov, and T. J. Tjalkens, "The context-tree weighting methods: basic properties," *IEEE Trans. Inform. Theory*, vol. 41, no. 3, pp. 653–664, May 1995.