Universal Source Coding for Monotonic and Fast Decaying Monotonic Distributions

Gil I. Shamir

***Abstract*—We study universal compression of sequences generated by monotonic distributions. We show that for a monotonic distribution over an alphabet of size k, each probability parameter costs essentially $0.5\log(n/k^3)$ bits, where n is the coded sequence length, as long as $k = o(n^{1/3})$. Otherwise, for $k = O(n)$, the total average sequence redundancy is $O(n^{1/3+\varepsilon})$ bits overall. We then show that there exists a sub-class of monotonic distributions over infinite alphabets for which redundancy of $O(n^{1/3+\varepsilon})$ bits overall is still achievable. This class contains fast decaying distributions, including many distributions over the integers such as the family of Zipf distributions and geometric distributions. For some slower decays, including other distributions over the integers, redundancy of $o(n)$ bits overall is achievable. A method to compute specific redundancy rates for such distributions is derived. The results are specifically true for finite entropy monotonic distributions. Finally, we study individual sequence redundancy behavior assuming a sequence is governed by a monotonic distribution. We show that for sequences whose empirical distributions are monotonic, individual redundancy bounds even tighter than those in the average case can be obtained. The relation of universal compression with monotonic distributions to universal compression of *patterns* of sequences is demonstrated.**

***Index Terms*—Average redundancy, individual redundancy, large alphabets, monotonic distributions, patterns, universal compression.**

I. Introduction

THE classical setting of the universal lossless compression problem [6], [9], [10] assumes that a sequence x^n of length n that was generated by a source $\boldsymbol{\theta}$ is to be compressed without knowledge of the particular $\boldsymbol{\theta}$ that generated x^n but with knowledge of the class Λ of all possible sources $\boldsymbol{\theta}$. The average performance of any given code, that assigns a length function $L(\cdot)$, is judged on the basis of the *redundancy* function $R_n\left(L, \boldsymbol{\theta}\right)$, which is defined as the difference between the expected code length of $L\left(\cdot\right)$ with respect to (w.r.t.) the given source probability mass function P_θ and the nth-order entropy of P_θ normalized by the length n of the uncoded sequence. A class of sources is said to be universally compressible in some worst sense if the redundancy function diminishes for this worst setting. Another approach to universal coding [37] considers the *individual sequence* redundancy $\hat{R}_n\left(L, x^n\right)$, defined as the normalized difference between the code length obtained by $L(\cdot)$ for x^n and the negative logarithm of the *maximum likelihood* (ML) probability of the sequence x^n, where the ML probability is within the class Λ. We thereafter refer to this negative logarithm as the *ML description length* of x^n. The individual sequence redundancy is defined for each sequence that can be generated by a source $\boldsymbol{\theta}$ in the given class Λ.

Classical literature on universal compression [6], [9], [10], [25], [37] considered compression of sequences generated by sources over finite alphabets. In fact, it was shown by Kieffer [17] (see also [15]) that there are no universal codes (in the sense of diminishing redundancy) for sources over infinite alphabets. Later work (see, e.g., [23], [30], [38]), however, bounded the achievable redundancies for i.i.d. sequences generated by sources over large and infinite alphabets. Specifically, while it was shown that the redundancy does not decay if the alphabet size is of the same order of magnitude as the sequence length n or greater, it was also shown that the redundancy does decay for alphabets of size $o(n)$.[1]

While there is no universal code for infinite alphabets, recent work [22] demonstrated that if one considers the *pattern* of a sequence instead of the sequence itself, universal codes do exist in the sense of diminishing redundancy. A pattern of a sequence, first considered, to the best of our knowledge, in [1], is a sequence of indices, where the index ψ_i at time i represents the order of first occurrence of letter x_i in the sequence x^n. Further study of universal compression of patterns [13], [22], [23], [31], [35] (and subsequently to the work in this paper in [2]) provided various lower and upper bounds to various forms of redundancy in universal compression of patterns. Another related study is that of compression of data, where the order of the occurring data symbols is not important, but their types and empirical counts are [39], [40].

This paper considers universal compression of data sequences generated by distributions that are known *a priori* to be monotonic. The order of probabilities of the source symbols is known in advance to both encoder and decoder and can be utilized as side information. Monotonic distributions, such as the *Zipf* (see, e.g., [42], [43]) and the geometric distribution over the integers, are common in applications such as language modeling, and image compression where residual signals are compressed (see, e.g., [20], [21]). One can also consider compression of the list of last or first names in a given city of a given population. Usually, there exists some monotonicity for

Manuscript received April 05, 2007; revised July 02, 2013; accepted August 01, 2013. Date of publication September 16, 2013; date of current version October 16, 2013. This work was supported by the NSF under Grant CCF-0347969. This paper was presented at the 2007 IEEE International Symposium on Information Theory.

The author is with Google, Inc., Pittsburgh, PA 15206 USA (e-mail: gshamir@ieee.org).

Communicated by W. Szpankowski, Associate Editor for Source Coding.

Digital Object Identifier 10.1109/TIT.2013.2281695

[1]For two functions $f(n)$ and $g(n)$, $f(n) = o(g(n))$ if $\forall c$, $\exists n_0$, such that $\forall n > n_0$, $f(n) < cg(n)$; $f(n) = O(g(n))$ if $\exists c$, n_0, such that $\forall n > n_0$, $0 \leq f(n) \leq cg(n)$; $f(n) = \Theta(g(n))$ if $\exists c_1$, c_2, n_0, such that $\forall n > n_0$, $c_1 g(n) \leq f(n) \leq c_2 g(n)$.

such a distribution in the given population, which both encoder and decoder may be aware of *a priori*. For example, the last name "Smith" can be expected to be much more common than the last name "Shannon." Another example is the compression of a sequence of observations of different species, where one has prior knowledge which species are more common, and which are rare. Finally, one can consider compressing data for which side information given to the decoder through a different channel gives the monotonicity order.

Monotonic distributions were studied by Elias [8], Rissanen [24], and Ryabko [26]. In [8] and [24], the study focused on *relative redundancy*, computing the ratio between average assigned code length and the source entropy. Ryabko in [26] studied codes for monotonic distributions and used the connection between redundancy and channel capacity (i.e., the *redundancy-capacity theorem*) to lower bound minimax redundancy. Much newer work by Foster *et al.* showed in [11] that (unlike the compression of patterns) there are no universal block codes in the standard sense for the complete class of monotonic distributions. The main reason is that there exist such distributions, for which much of the statistical weight lies in the long tail of the distribution in symbols that have very low probability, and most of which will not occur in a given sequence. Thus, in practice, even though one has the prior knowledge of the monotonicity of the distribution, this monotonicity is not necessarily retained in an observed sequence. Actual coding is, therefore, very similar to compressing with infinite alphabets, and the additional prior knowledge of the monotonicity is not very helpful in reducing redundancy. Despite that, Foster *et al.* demonstrated codes that obtained universal per-symbol redundancy of $o(1)$ as long as the source entropy is fixed (i.e., neither increasing with n nor infinite).

The work in [11] studied coding sequences (or blocks) generated by i.i.d. monotonic distributions, and designed codes for which the relative block redundancy could be (upper) bounded. Unlike that work, the focus in [8], [24], and [26] was on designing codes that minimize the redundancy or relative redundancy for a single symbol generated by a monotonic distribution. Specifically, in [24], *minimax* codes, which minimize the relative redundancy for the worst possible monotonic distribution over a given alphabet size, were derived. In [26], it was shown that redundancy of $O(\log \log k)$, where k is the alphabet size, can be obtained with minimax per-symbol codes. Very recent work [18] considered per-symbol codes that minimize an average redundancy over the class of monotonic distributions for a given alphabet size. Unlike [11], all these papers study per-symbol codes. Therefore, the codes designed always pay nondiminishing per-symbol redundancy.

A different line of work on monotonic distributions considered optimizing codes for a known monotonic distribution but with unknown parameters (see [20], [21] for design of codes for two-sided geometric distributions). In this line of work, the class of sources is very limited and consists of only the unknown parameters of a known distribution.

In this paper, we consider a general class of monotonic distributions that is not restricted to a specific type or a single parameter. We study standard block redundancy for coding sequences generated by i.i.d. monotonic distributions, i.e., a setting similar to the work in [11]. We do, however, restrict ourselves to smaller subsets of the complete class of monotonic distributions. First, we consider monotonic distributions over alphabets of size k, where k is either small w.r.t. n, or of $O(n)$. Then, we extend the analysis to show that under minimal restrictions of the monotonic distribution class, there exist universal codes in the standard sense, i.e., with diminishing per-symbol redundancy. In fact, not only do universal codes exist, but under mild restrictions, they achieve the same redundancy as obtained for alphabets of size $O(n)$. The restrictions on this subclass imply that some types of fast decaying monotonic distributions are included in it, and therefore, sequences generated by these distributions (without prior knowledge of either the distribution or of its parameters) can still be compressed universally in the class of monotonic distributions.

The main contributions of this paper are the development of codes and derivation of their upper bounds on the redundancies for coding i.i.d. sequences generated by monotonic distributions. Specifically, this paper gives complete characterization of the redundancy in coding with monotonic distributions over "small" alphabets ($k = o(n^{1/3})$) and "large" alphabets ($k = O(n)$). Then, it shows that these redundancy bounds carry over (in first order) to fast decaying distributions. Next, a code that achieves good redundancy rates for even slower decaying monotonic distributions is derived, and is used to study achievable redundancy rates for such distributions. Finally, even tighter upper bounds relative to the ML description length are obtained for individual sequences for which the monotonic order of the probabilities is known. The codes derived are two part codes, based on a description of any sequence using a quantized distribution describing the ML distribution of a given sequence. The redundancy consists of the distribution description cost and quantization penalty.

Lower bounds are also presented (in both average and individual sequence cases) to complete the characterization, and are shown to meet the upper bounds in the first three cases (small alphabets, large alphabets, and fast decaying distributions). The lower bounds turn out to relate to those obtained for coding patterns. The relationship to patterns is demonstrated in the proofs of the lower bounds. The main components of the average case proofs are, in fact, identical to those in [31], and the reader is referred to more details in [31]. The main steps of the proofs are still presented in appendixes here for the sake of completeness.

The universal compression problem over monotonic distributions is very related to that of patterns. For small and large alphabets, the redundancy rates attained appear to be the same. This is because in both problems the richness of the class (yielding the universal coding redundancy) is decreased by the same factor from that of the original i.i.d. class, although for different reasons. In the pattern case, sequences which are label permutations of the others are governed by the same pattern ML distribution. Here, such sequences are constrained to a distribution whose probabilities are ordered by the monotonicity constraint. However, a monotonic ML distribution requires given labels to appear in the required order, and may not equal the actual i.i.d. ML distribution. This restriction is not imposed when coding patterns, and makes this part of the analysis more difficult for monotonic distributions. Overall, in both cases, we observe a

cost of essentially $0.5\log(n/k^3)$ bits per each unknown parameter for smaller alphabets and a cost of essentially $O(n^{1/3})$ bits overall for larger alphabets. The technique that is used to prove the upper bounds of the main theorems in this paper follows the original work in [35] for upper bounding the redundancy for coding patterns. Tight upper bounds on the redundancy for coding patterns were not attained when the work presented in this paper, published originally in [36] and [34], was done. Several years subsequently to the work presented here, the general construction in [35] was followed in [2] to show an $O(n^{1/3+\varepsilon})$ upper bound for coding patterns. An upper bound for small alphabets of $(1+\varepsilon)0.5\log(n/k^3)$ bits per parameter is yet to have been derived for patterns, to the best of our knowledge. The constructions used in this paper can be applied to the pattern problem. The description costs of these constructions apply to patterns, but the computation of quantization costs is much more difficult for patterns. Specifically, the construction used in the individual sequence case for monotonic distributions can be applied to patterns.

The outline of this paper is as follows. Section II describes the notation and basic definitions. Then, in Section III, lower bounds on the redundancy for monotonic distributions are derived. Next, in Section IV, we propose codes and upper bound their redundancy for coding monotonic distributions over small and large alphabets. These upper bounds match the rates of the lower bounds. They are then extended to fast decaying monotonic distributions in Section V, which also demonstrates the use of the bounds on some standard monotonic distributions. Finally, in Section VI, we consider individual sequences.

II. Notation and Definitions

Let $x^n \triangleq (x_1, x_2, \ldots, x_n)$ denote a sequence of n symbols over the alphabet Σ of size k, where k can go to infinity. Without loss of generality, we assume that $\Sigma = \{1, 2, \ldots, k\}$, i.e., it is the set of positive integers from 1 to k. The sequence x^n is generated by an i.i.d. distribution of some source, determined by the parameter vector $\boldsymbol{\theta} \triangleq (\theta_1, \theta_2, \ldots, \theta_k)$, where θ_i is the probability of X taking value i. The components of $\boldsymbol{\theta}$ are nonnegative and sum to 1. The distributions we consider in this paper are monotonic. Therefore, $\theta_1 \geq \theta_2 \geq \ldots \geq \theta_k$. The class of all monotonic distributions will be denoted by $\mathcal{M}$. The class of monotonic distributions over an alphabet of size k is denoted by $\mathcal{M}_k$. It is assumed that prior to coding x^n both encoder and decoder know that $\boldsymbol{\theta} \in \mathcal{M}$ or $\boldsymbol{\theta} \in \mathcal{M}_k$, and also know the order of the probabilities in $\boldsymbol{\theta}$. In the more restrictive setting, k is known in advance and it is known that $\boldsymbol{\theta} \in \mathcal{M}_k$. We do not restrict ourselves to this setting. In general, boldface letters will denote vectors, whose components will be denoted by their indices in the vector. Capital letters will denote random variables. We will denote an estimator by the *hat* sign. In particular, $\hat{\boldsymbol{\theta}}$ will denote the ML estimator of $\boldsymbol{\theta}$ which is obtained from x^n.

The probability of x^n generated by $\boldsymbol{\theta}$ is given by $P_\theta(x^n) \triangleq \Pr(x^n | \boldsymbol{\Theta} = \boldsymbol{\theta})$. The average per-symbol[2] nth-order redundancy obtained by a code that assigns length function $L(\cdot)$ for $\boldsymbol{\theta}$ is

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

where E_θ denotes expectation w.r.t. $\boldsymbol{\theta}$, and $H_\theta[X]$ is the (per-symbol) entropy (rate) of the source ($H_\theta[X^n]$ is the nth-order sequence entropy of $\boldsymbol{\theta}$, and for i.i.d. sources, $H_\theta[X^n] = nH_\theta[X]$). With entropy coding techniques, assigning a universal probability $Q(x^n)$ is identical to designing a universal code for coding x^n where, up to negligible integer length constraints that will be ignored, the negative logarithm to the base of 2 of the assigned probability is considered as the code length.

The *individual* sequence redundancy (see, e.g., [37]) of a code with length function $L(\cdot)$ per sequence x^n over class Λ is

$$\hat{R}_n(L, x^n) \triangleq \frac{1}{n}\{L(x^n) + \log P_{ML}(x^n)\} \tag{2}$$

where the logarithm function is taken to the base of 2, here and elsewhere, and $P_{ML}(x^n)$ is the probability of x^n given by the ML estimator $\hat{\boldsymbol{\theta}}_\Lambda \in \Lambda$ of the governing parameter vector $\boldsymbol{\Theta}$. The negative logarithm of this probability is, up to integer length constraints, the shortest possible code length assigned to x^n in Λ. It will be referred to as the ML description length of x^n in Λ. In the general case, one considers the i.i.d. ML. However, since we only consider $\boldsymbol{\theta} \in \mathcal{M}$, i.e., restrict the sequence to one governed by a monotonic distribution, we define $\hat{\boldsymbol{\theta}}_{\mathcal{M}} \in \mathcal{M}$ as the monotonic ML estimator. Its associated shortest code length will be referred to as the monotonic ML description length. The estimator $\hat{\boldsymbol{\theta}}_{\mathcal{M}}$ may differ from the i.i.d. ML $\hat{\boldsymbol{\theta}}$, in particular, if the empirical distribution of x^n is not monotonic. The individual sequence redundancy in $\mathcal{M}$ is thus defined w.r.t. the monotonic ML description length, which is the negative logarithm of $P_{ML}(x^n) \triangleq P_{\hat{\theta}_{\mathcal{M}}}(x^n) \triangleq \Pr\left(x^n \mid \boldsymbol{\Theta} = \hat{\boldsymbol{\theta}}_{\mathcal{M}} \in \mathcal{M}\right)$.

The average *minimax* redundancy of some class Λ is defined as

$$R_n^+(\Lambda) \triangleq \min_L \sup_{\boldsymbol{\theta} \in \Lambda} R_n(L, \boldsymbol{\theta}). \tag{3}$$

Similarly, the *individual minimax* redundancy is that of the best code $L(\cdot)$ for the worst sequence x^n,

$$\hat{R}_n^+(\Lambda) \triangleq \min_L \sup_{\boldsymbol{\theta} \in \Lambda} \max_{x^n} \frac{1}{n}\{L(x^n) + \log P_\theta(x^n)\}. \tag{4}$$

The *maximin* redundancy of Λ is

$$R_n^-(\Lambda) \triangleq \sup_w \min_L \int_\Lambda w(d\boldsymbol{\theta})\, R_n(L, \boldsymbol{\theta}) \tag{5}$$

where $w(\cdot)$ is a prior on Λ. In [6], it was shown by Davisson that $R_n^+(\Lambda) \geq R_n^-(\Lambda)$. Davisson also tied the maximin redundancy to the capacity of the channel induced by the conditional probability P_θ. It was then shown independently by Gallager [12] and Ryabko [26] first, and then by Davisson and Leon-Garcia [7], that the minimax and maximin redundancies are essentially equal, hence, making the connection between the minimax redundancy and the capacity of the channel induced by P_θ. Finally, Merhav and Feder [19] tied between the capacity of this

[2] In this paper, redundancy is defined per-symbol (normalized by the sequence length n). However, when we refer to redundancy in overall bits, we address the block redundancy cost for a sequence.

channel and redundancy for almost all sources in a class proving a strong version of the theorem. The *redundancy-capacity* theorem is used to prove lower bounds in the minimax (maximin) and "almost all sources" senses for the monotonic distribution class.

III. Lower Bounds

Lower bounds on various forms of the redundancy for the class of monotonic distributions can be obtained with slight modifications of the proofs for the lower bounds on the redundancy of coding patterns in [16], [22], [23], and [31]. The bounds are presented in the following three theorems. For the sake of completeness, the main steps of the proofs of the first two theorems are presented in appendixes, and the proof of the third theorem is presented below. The reader is referred to [16], [22], [23], [30] and [31] for more details.

Theorem 1: Fix an arbitrarily small $\varepsilon > 0$, and let $n \to \infty$. Then, the nth-order average maximin and minimax universal coding redundancies for i.i.d. sequences generated by a monotonic distribution with alphabet size k are lower bounded by

$$R_n^- (\mathcal{M}_k) \geq \begin{cases} \frac{k-1}{2n} \log \frac{n^{1-\varepsilon}}{k^3} + \frac{k-1}{2n} \log \frac{\pi e^3}{2} - O\left(\frac{\log k}{n}\right), & k \leq \mathcal{T}_{\varepsilon,n}^- \\ \left(\frac{\pi}{2}\right)^{1/3} (1.5 \log e) \frac{n^{(1-\varepsilon)/3}}{n} - O\left(\frac{\log n}{n}\right), & k > \mathcal{T}_{\varepsilon,n}^- \end{cases} \tag{6}$$

where

$$\mathcal{T}_{\varepsilon,n}^- \triangleq \left(\frac{\pi n^{1-\varepsilon}}{2}\right)^{1/3}. \tag{7}$$

Theorem 2: Fix an arbitrarily small $\varepsilon > 0$, and let $n \to \infty$. Let $\mu_n(\cdot)$ be the uniform prior over points in $\mathcal{M}_k$. Then, the nth-order average universal coding redundancy for coding i.i.d. sequences generated by monotonic distributions with alphabet size k is lower bounded by

$$R_n (L, \boldsymbol{\theta}) \geq \begin{cases} \frac{k-1}{2n} \log \frac{n^{1-\varepsilon}}{k^3} - \frac{k-1}{2n} \log \frac{8\pi}{e^3} - O\left(\frac{\log k}{n}\right), & k \leq \mathcal{T}_{\varepsilon,n} \\ \frac{1.5 \log e}{2\pi^{1/3}} \cdot \frac{n^{(1-\varepsilon)/3}}{n} - O\left(\frac{\log n}{n}\right), & k > \mathcal{T}_{\varepsilon,n} \end{cases} \tag{8}$$

where

$$\mathcal{T}_{\varepsilon,n} \triangleq \frac{1}{2} \cdot \left(\frac{n^{1-\varepsilon}}{\pi}\right)^{1/3} \tag{9}$$

for every code $L(\cdot)$ and almost every i.i.d. source $\boldsymbol{\theta} \in \mathcal{M}_k$, except for a set of sources $A_\varepsilon(n)$ whose relative volume $\mu_n(A_\varepsilon(n))$ in $\mathcal{M}_k$ goes to 0 as $n \to \infty$.

Theorems 1 and 2 give lower bounds on redundancies of coding over monotonic distributions for the class $\mathcal{M}_k$. However, the bounds are more general, and the second region applies to the whole class of monotonic distributions $\mathcal{M}$. By plugging the boundary values of k into the first regions of both Theorems, the bounds of the second regions are obtained, demonstrating the threshold phenomenon of the transition between the regions. Subsequent work in [2] to the work presented in this paper slightly tightened the second region of the bound of Theorem 1 for patterns. This was done by applying a general technique that uses bounds on error correcting codes, as that described in earlier work in [27]–[29], to patterns on top of the bounding methods used in [31]. The tighter bound for that region can also be applied to monotonic distributions. As in the case of patterns [22], [31], the bounds in (6)–(8) show that each parameter costs at least $0.5 \log(n/k^3)$ bits for small alphabets, and the total universality cost is at least $\Theta(n^{1/3-\varepsilon})$ bits overall for larger alphabets. We show in Section IV that for $k = O(n)$ these bounds are asymptotically achievable for monotonic distributions. The bounds in (6)–(8) focus on large values of k that can increase with n. For small fixed k, the second-order terms of existing bounds for coding unconstrained i.i.d. sources are tighter. However, as k increases, the bounds above become tighter through their first dominant term, and second-order terms become negligible. The proofs of Theorems 1 and 2 are presented in Appendixes A and B, respectively.

Theorem 3: [3] Let $n \to \infty$. Then, the nth-order individual minimax redundancy for i.i.d. sequences with maximal letter k w.r.t. the monotonic ML description length with alphabet size k is lower bounded by

$$\hat{R}_n^+ (\mathcal{M}_k) \geq \begin{cases} \frac{k}{2n} \log \frac{ne^3}{k^3} - \frac{\log n}{2n} + O\left(\frac{k^{3/2}}{n^{3/2}}\right), & k \leq n^{1/3} \\ \frac{3}{2} (\log e) \cdot \frac{n^{1/3}}{n} - \frac{\log n}{2n} + O\left(\frac{1}{n}\right), & k > n^{1/3}. \end{cases} \tag{10}$$

Theorem 3 lower bounds the individual minimax redundancy for coding a sequence believed to have an empirical monotonic distribution. The alphabet size is determined by the maximal letter that occurs in the sequence, i.e., $k = \max\{x_1, x_2, \ldots, x_n\}$. (If k is unknown, one can use Elias' code for the integers [8] using $O(\log k)$ bits to describe k. However, this is not reflected in the lower bound.) The ML probability estimate is taken over the class of monotonic distributions. Namely, the empirical probability (standard ML) estimate $\hat{\boldsymbol{\theta}}$ is not $\hat{\boldsymbol{\theta}}_{\mathcal{M}}$ in case $\hat{\boldsymbol{\theta}}$ does not satisfy the monotonicity that defines the class $\mathcal{M}$. While the average case maximin and minimax bounds of Theorem 1 also apply to $\hat{R}_n^+ (\mathcal{M}_k)$, the bounds of Theorem 3 are tighter for the individual redundancy and are obtained using individual sequence redundancy techniques.

Proof [Theorem 3]: Using Shtarkov's normalized maximum likelihood (NML) approach [37], one can assign probability

$$Q(x^n) \triangleq \frac{P_{\hat{\theta}_{\mathcal{M}}}(x^n)}{\sum_{y^n} P_{\hat{\theta}'_{\mathcal{M}}}(y^n)} \triangleq \frac{\max_{\theta' \in \mathcal{M}} P_{\theta'}(x^n)}{\sum_{y^n} \max_{\theta'' \in \mathcal{M}} P_{\theta''}(y^n)} \tag{11}$$

to sequence x^n. This approach minimizes the individual minimax redundancy, giving individual redundancy of

$$\begin{aligned} \hat{R}_n (Q, x^n) &= \frac{1}{n} \log \frac{\max_{\theta' \in \mathcal{M}} P_{\theta'}(x^n)}{Q(x^n)} \\ &= \frac{1}{n} \log \left\{ \sum_{y^n} \max_{\theta' \in \mathcal{M}} P_{\theta'}(y^n) \right\} \end{aligned} \tag{12}$$

[3] The original submission of this paper derived a looser bound for the first region of (10). A tighter bound was obtained using results that appeared subsequently to the submission of this paper in [38].

to every x^n, specifically achieving the individual minimax redundancy.

It is now left to bound the logarithm of the sum in (12). We follow the approach used in [23, Th. 2] for bounding the redundancy for standard compression of i.i.d. sequences over large alphabets and use the results in [38] (as well as the approximation in [1]) for a precise expression of this component. We then adjust the result to monotonic distributions. Let $\mathbf{n}_x^\ell \triangleq (n_x(1), n_x(2), \ldots, n_x(\ell))$ denote the occurrence counts of the first ℓ letters of the alphabet Σ in x^n. Assuming k is the largest letter in x^n, $\sum_{i=1}^{k} n_x(i) = n$. Now, following (12),

$$\begin{aligned} & n\hat{R}_n^+(\mathcal{M}_k) \\ & \overset{(a)}{\geq} \log\left\{\sum_{y^n:\hat{\theta}(y^n)\in\mathcal{M}} P_{\hat{\theta}}(y^n)\right\} \\ & \overset{(b)}{\geq} \log\left\{\sum_{\ell=1}^{k}\sum_{\mathbf{n}_y^\ell}\frac{1}{\ell!}\binom{n}{n_y(1),\ldots,n_y(\ell)}\prod_{i=1}^{\ell}\left(\frac{n_y(i)}{n}\right)^{n_y(i)}\right\} \\ & \overset{(c)}{\geq} \log\left\{\sum_{\mathbf{n}_y^k}\frac{1}{k!}\binom{n}{n_y(1),\ldots,n_y(k)}\prod_{i=1}^{k}\left(\frac{n_y(i)}{n}\right)^{n_y(i)}\right\} \\ & \overset{(d)}{=} \frac{k-1}{2}\log\frac{n}{k} + \frac{k}{2}\log e + O\left(\frac{k^{3/2}}{n^{1/2}}\right) - \log(k!) \\ & \overset{(e)}{\geq} \frac{k}{2}\log\frac{ne^3}{k^3} - \frac{1}{2}\log n + O\left(\frac{k^{3/2}}{n^{1/2}}\right) \end{aligned} \tag{13}$$

where (a) follows from including only sequences y^n that have a monotonic empirical (i.i.d. ML) distribution in Shtarkov's sum. Inequality (b) follows from partitioning the sequences y^n into types as done in [23], first by the number of occurring symbols ℓ, and then by the empirical distribution. Unlike standard i.i.d. distributions though, monotonicity implies that only the first ℓ symbols in Σ occur, and thus the choice of ℓ out of k in the proof in [23] is replaced by 1. Like in coding patterns, we also divide by $\ell!$ because each type with ℓ occurring symbols can be ordered in at most $\ell!$ ways, where only some retain the monotonicity. (Note that this step is the reason that step (b) produces an inequality, because more than one of the orderings may be monotonic if equal occurrence counts occur.) Retaining only the term $\ell = k$ yields (c). Then, (d) follows from applying (15) in [38] (see also the approximation of (13) in [1]). Finally, (e) follows from Stirling's approximation

$$\sqrt{2\pi m}\cdot\left(\frac{m}{e}\right)^m \leq m! \leq \sqrt{2\pi m}\cdot\left(\frac{m}{e}\right)^m\cdot\exp\left\{\frac{1}{12m}\right\}. \tag{14}$$

The first region in (10) results directly from (13). The value $\ell^* = n^{1/3}$ that maximizes the summand can be retained in step (c) instead of k, for every $k \geq \ell^*$, yielding the second region of the bound. This concludes the proof of Theorem 3. ■

IV. Upper Bounds for Small and Large Alphabets

In this section, we demonstrate codes that asymptotically achieve the lower bounds for $\boldsymbol{\theta} \in \mathcal{M}_k$ and $k = O(n)$. We begin with a theorem and a corollary that show the achievable redundancies. The theorem shows a simpler bound, and the corollary (that follows the proof of the theorem) shows a tighter, more complex bound. The remainder of the section is devoted to proving both theorem and corollary, by describing codes, for which the redundancy bounds provide general bounds on the redundancy, and bounding their redundancies. The theorem is stated assuming no initial knowledge of k. The proof first considers the setting where k is known, and then shows how the same bounds are achieved even when k is unknown in advance, but as long as it satisfies the conditions.

Theorem 4: Fix an arbitrarily small $\varepsilon > 0$, and let $n \to \infty$. Then, there exists a code with length function $L^*(\cdot)$ that achieves redundancy

$$R_n(L^*, \boldsymbol{\theta}) \leq \begin{cases} (1+\varepsilon)\frac{k-1}{2n}\log\frac{n}{k^3}, & k = o\left(n^{1/3}\right), \\ (1+\varepsilon)\frac{k-1}{2n}\log\frac{n(\log n)^2}{k^3}, & k \leq n^{1/3}, \\ (1+\varepsilon)(\log n)\left(\log\frac{k}{n^{1/3-\varepsilon}}\right)\frac{n^{1/3}}{n}, & n^{1/3} < k = o(n), \\ (1+\varepsilon)\frac{2}{3}(\log n)^2\frac{n^{1/3}}{n}, & n^{1/3} < k = O(n) \end{cases} \tag{15}$$

for i.i.d. sequences generated by any source $\boldsymbol{\theta} \in \mathcal{M}_k$.

The bounds presented are asymptotic. Second-order terms are absorbed in ε. The second region contains the first, and the last contains the third. The first and third regions, however, have tighter bounds for the smaller values of k. The code designed to code a sequence x^n is a two part code [25]. First, a distribution is described, and then it is used to code x^n. The redundancy consists of the cost of describing the distribution and a quantization cost. Quantization is performed to reduce description cost, but yields the quantization cost. To achieve the lower bound, the larger the probability parameter is, the coarser its quantization. This approach was used in [30] and [31] to obtain upper bounds on the redundancy for coding over large alphabets and for coding patterns, respectively. The method in [30] and [31], however, is insufficient here, because it still results in too many quantization points due to the polynomial growth in quantization spacing. Here, we use an exponential growth as the parameters increase. This general idea was used in [35] to improve an upper bound on the redundancy of coding patterns. Since both encoder and decoder know the order of the probabilities *a priori*, this order need not be coded. It is thus sufficient to encode the quantized probabilities of the monotonic distribution, and the decoder can identify which probability is associated with which symbol using the monotonicity of the distribution. This point, in fact, complicates the proof, because the actual ML distribution $\hat{\boldsymbol{\theta}}$ of a given sequence may not be monotonic even if the sequence was generated by a monotonic distribution. Since the labels are not coded, we must quantize $\hat{\boldsymbol{\theta}}_{\mathcal{M}}$ instead. There is no such complication when coding patterns or sequences that obey distribution monotonicity side information as in Section VI.

Branching several steps from the proof of Theorem 4 below leads to the following tighter bounds on the upper regions, which are proved following the proof of Theorem 4.

Corollary 1: Fix an arbitrarily small $\varepsilon > 0$, and let $n \to \infty$. Then, for $k > n^{1/3}$, there exists a code with length function $L^*(\cdot)$ that achieves redundancy

$$R_n(L^*, \boldsymbol{\theta}) \leq \begin{cases} \left(2.3\log\frac{k}{n^{1/3-\varepsilon}} + 0.8\log n\right)\frac{n^{1/3}(\log n)^{1/3}}{n}, & k = o(n), \\ \frac{2.3n^{1/3}(\log n)^{4/3}}{n}, & k = O(n) \end{cases} \tag{16}$$

for i.i.d. sequences generated by any source $\boldsymbol{\theta} \in \mathcal{M}_k$.

Proof [Theorem 4]: The proof treats the regions $k \leq n^{1/3}$ and $k > n^{1/3}$ separately. For each region, we construct a grid of points to which a two part code can quantize the probability parameters. The main idea is that spacing between adjacent grid points is "semi"-exponentially increasing. To achieve that, the probability space is partitioned into intervals, whose length increases exponentially, and within each interval a fixed number of equally separated grid points are generated. Next, the ML probability vector of each sequence is quantized into the points of the grid. In the lower k region, a differential code is used to describe the number of points in the grid between two adjacent probability parameters, starting with the smallest one. In the upper k region, the number of probability parameters quantized to that grid point is described for every grid point. Then, the description cost, and the quantization cost are upper bounded. The sum of these two costs constitutes the description length. The redundancy is computed by subtracting the description length with the true probability parameters from the description length used. The quantized version of the true probability vector is used as an auxiliary vector to aid in upper bounding this difference.

We start with $k \leq n^{1/3}$ assuming k is known. Let $\beta = 1/(\log n)$ be a parameter (we can also choose other values). Partition the probability space into $J_1 = \lceil 1/\beta \rceil$ intervals

$$I_j = \left[\frac{n^{(j-1)\beta}}{n}, \frac{n^{j\beta}}{n} \right), \ 1 \leq j \leq J_1. \tag{17}$$

Note that $I_1 = [1/n, 2/n)$, $I_2 = [2/n, 4/n), \ldots,$ $I_j = [2^{j-1}/n, 2^j/n)$. Let $k_j = |\theta_i \in I_j|$ denote the number of probabilities in $\boldsymbol{\theta}$ that are in interval I_j. In interval j, take a grid of points with spacing

$$\Delta_j^{(1)} = \frac{\sqrt{k} n^{j\beta}}{n^{1.5}}. \tag{18}$$

Note that to complete all points in an interval, the spacing between two points at the boundary of an interval may be smaller. There are $\lceil \log n \rceil$ intervals. Ignoring negligible integer length constraints (here and elsewhere), in each interval, the number of points is bounded by

$$|I_j| \leq \frac{1}{2} \cdot \sqrt{\frac{n}{k}}, \ \forall j : j = 1, 2, \ldots, J_1, \tag{19}$$

where $|\cdot|$ denotes the cardinality of a set. Let the *grid*

$$\boldsymbol{\tau} = (\tau_1, \tau_2, \ldots) = \left(\frac{1}{n}, \frac{1}{n} + \frac{2\sqrt{k}}{n^{1.5}}, \ldots, \frac{2}{n}, \frac{2}{n} + \frac{4\sqrt{k}}{n^{1.5}}, \ldots \right) \tag{20}$$

be a vector that takes all the points from all intervals, with cardinality

$$B_1 \triangleq |\boldsymbol{\tau}| \leq \frac{1}{2} \cdot \sqrt{\frac{n}{k}} \lceil \log n \rceil. \tag{21}$$

Now, let $\boldsymbol{\varphi} = (\varphi_1, \varphi_2, \ldots, \varphi_k)$ be a monotonic probability vector, such that $\sum \varphi_i = 1, \varphi_1 \geq \varphi_2 \geq \cdots \geq \varphi_k \geq 0$, and also the smaller $k-1$ components of $\boldsymbol{\varphi}$ are either 0 or from $\boldsymbol{\tau}$, i.e., $\varphi_i \in (\boldsymbol{\tau} \cup \{0\}), i = 2, 3, \ldots, k$. One can code x^n using a two part code, assuming the distribution governing x^n is given by the parameter $\boldsymbol{\varphi}$. The code length required (up to integer length constraints) is

$$L(x^n|\boldsymbol{\varphi}) = \log k + L_R(\boldsymbol{\varphi}) - \log P_\varphi(x^n), \tag{22}$$

where $\log k$ bits are needed to describe how many letter probabilities are greater than 0 in $\boldsymbol{\varphi}$, $L_R(\boldsymbol{\varphi})$ is the number of bits required to describe the quantized points of $\boldsymbol{\varphi}$, and the last term is needed to encode x^n assuming it is governed by $\boldsymbol{\varphi}$.

The vector $\boldsymbol{\varphi}$ can be described by a code as follows. Let $\hat{k}_\varphi$ be the number of nonzero letter probabilities hypothesized by $\boldsymbol{\varphi}$. Let b_i denote the index of φ_i in $\boldsymbol{\tau}$, i.e., $\varphi_i = \tau_{b_i}$. Then, we will use the following differential code. For $\varphi_{\hat{k}_\varphi}$ we need at most $1 + \log b_{\hat{k}_\varphi} + 2\log(1 + \log b_{\hat{k}_\varphi})$ bits to code its index in $\boldsymbol{\tau}$ using Elias' coding for the integers [8]. For φ_{i-1}, we need at most $1 + \log(b_{i-1} - b_i + 1) + 2\log[1 + \log(b_{i-1} - b_i + 1)]$ bits to code the index displacement from the index of the previous parameter, where an additional 1 is added to the difference in case the two parameters share the same index. Summing up all components of $\boldsymbol{\varphi}$, and taking $b_{\hat{k}_\varphi + 1} = 0$,

$$\begin{aligned} L_R(\boldsymbol{\varphi}) &\leq \hat{k}_\varphi - 1 + \sum_{i=2}^{\hat{k}_\varphi} \log(b_i - b_{i+1} + 1) + \\ &\quad 2\sum_{i=2}^{\hat{k}_\varphi} \log[1 + \log(b_i - b_{i+1} + 1)] \\ &\overset{(a)}{\leq} (k-1) + (k-1)\log\frac{B_1 + k - 1}{k} + \\ &\quad 2(k-1)\log\log\frac{B_1 + k - 1}{k} + o(k) \\ &\overset{(b)}{=} (1+\varepsilon)\frac{k-1}{2}\log\frac{n(\log n)^2}{k^3}. \end{aligned} \tag{23}$$

Inequality (a) is obtained by applying Jensen's inequality once on the first sum, twice on the second sum utilizing the monotonicity of the logarithm function, and by bounding $\hat{k}_\varphi$ by k, and absorbing second-order terms in the resulting $o(k)$ term. Then, second-order terms are absorbed in ε, and (21) is used to obtain (b).

To code x^n, we choose $\boldsymbol{\varphi}$ which minimizes the expression in (22) over all $\boldsymbol{\varphi}$, i.e.,

$$L^*(x^n) = \min_{\boldsymbol{\varphi}: \varphi_i \in (\boldsymbol{\tau} \cup \{0\}),\ i=2,3,\ldots,k} L(x^n|\boldsymbol{\varphi}) \triangleq L(x^n|\hat{\boldsymbol{\varphi}}). \tag{24}$$

The *pointwise* redundancy for x^n is given by

$$\begin{aligned} nR_n(L^*, x^n) &= L^*(x^n) + \log P_\theta(x^n) \\ &= \log k + L_R^*(\hat{\boldsymbol{\varphi}}) + \log\frac{P_\theta(x^n)}{P_{\hat{\varphi}}(x^n)}. \end{aligned} \tag{25}$$

Note that the pointwise redundancy differs from the individual one, since it is defined w.r.t. the true probability of x^n. Thus, for a given x^n it may also be negative.

To bound the third term of (25), let $\boldsymbol{\theta}'$ be a monotonic version of $\boldsymbol{\theta}$ quantized onto $\boldsymbol{\tau}$, i.e., $\theta_i' \in (\boldsymbol{\tau} \cup \{0\})$, $i = 2, 3, \ldots, k$, where if $\theta_i > 0 \Leftrightarrow \theta_i' > 0$ as well. This implies that all positive $\theta_i < 1/n$ are quantized to $\theta_i' = 1/n$. Define the quantization error,

$$\delta_i = \theta_i - \theta_i'. \tag{26}$$

The quantization is performed from the smallest parameter θ_k to the largest, where monotonicity is maintained, as well as minimal absolute cumulative quantization error. Thus, unless there is cumulative error formed by many parameters $\theta_i < 1/n$, θ_i will be quantized to one of the two nearest grid points (one smaller and one greater than it). It also guarantees that $|\delta_1| \leq \Delta_{j_2}^{(1)} \leq \Delta_{j_1}^{(1)}$, where j_1 and j_2 are the indices of the intervals in which θ_1 and θ_2 are contained, respectively, i.e., $\theta_1 \in I_{j_1}$ and $\theta_2 \in I_{j_2}$. However, if there exists a cumulative error Δ_{offset} due to quantization of parameters $\theta_i : 0 < \theta_i < 1/n$ to $\theta_i' = 1/n$, this error is offset by decreasing every θ_i' for $\theta_i > 1/n$ by $\alpha_i \cdot \Delta_{offset} \cdot \theta_i'$, where $\alpha_i > 0$ is some constant, and quantizing the value to the nearest grid point maintaining monotonicity and minimal cumulative error. By construction, $\Delta_{offset} \leq k/n$, and thus

$$|\delta_i| \leq \frac{\sqrt{k} n^{j\beta}}{n^{1.5}} + \frac{k}{n} \alpha_i' \theta_i' \tag{27}$$

where $\alpha_i' > 0$ is a constant derived from α_i.

Now, since $\boldsymbol{\theta'}$ is included in the minimization of (24), we have, for every x^n,

$$L^*(x^n) \leq L(x^n | \boldsymbol{\theta'}), \tag{28}$$

and also

$$nR_n(L^*, x^n) \leq \log k + L_R(\boldsymbol{\theta'}) + \log \frac{P_\theta(x^n)}{P_{\theta'}(x^n)}. \tag{29}$$

Averaging over all possible x^n, the average redundancy is bounded by

$$\begin{aligned} &nR_n(L^*, \boldsymbol{\theta}) \\ &\quad = \log k + E_\theta L_R^*(\hat{\boldsymbol{\varphi}}) + E_\theta \log \frac{P_\theta(X^n)}{P_{\hat{\varphi}}(X^n)} \\ &\quad \leq \log k + E_\theta L_R(\boldsymbol{\theta'}) + E_\theta \log \frac{P_\theta(X^n)}{P_{\theta'}(X^n)}. \end{aligned} \tag{30}$$

The second term of (30) is bounded with the bound of (23), and we proceed with the third term

$$\begin{aligned} &E_\theta \log \frac{P_\theta(X^n)}{P_{\theta'}(X^n)} \\ &\quad \overset{(a)}{=} n \sum_{i=1}^{k} \theta_i \log \frac{\theta_i}{\theta_i'} \overset{(b)}{=} n \sum_{i=1}^{k} (\theta_i' + \delta_i) \log \left(1 + \frac{\delta_i}{\theta_i'}\right) \\ &\quad \overset{(c)}{\leq} n(\log e) \sum_{i=1}^{k} (\theta_i' + \delta_i) \frac{\delta_i}{\theta_i'} \overset{(d)}{=} n(\log e) \sum_{i=1}^{k} \frac{\delta_i^2}{\theta_i'} \\ &\quad \overset{(e)}{\leq} (1 + o(1)) k \log e + (1 + o(1)) \frac{2(\log e)k}{n} \sum_{j=1}^{J_1} k_j \cdot n^{j\beta} \\ &\quad \overset{(f)}{\leq} 5(1 + o(1))(\log e)k. \end{aligned} \tag{31}$$

Equality (a) is since the expectation is performed on the number of occurrences of letter i for each letter. Representing $\theta_i = \theta_i' + \delta_i$ yields equation (b). We use $\ln(1 + x) \leq x$ to obtain (c). Equality (d) is obtained since all the quantization displacements must sum to 0. The first term of inequality (e) in (31) is obtained under a worst case assumption that $\theta_i \ll 1/n$ for $i \geq 2$. Thus, it is quantized to $\theta_i' = 1/n$, and the bound $|\delta_i| \leq 1/n$ is used. In a different worst case scenario, we have from (27) and since in interval j, $\theta_i' \geq n^{(j-1)\beta}/n$,

$$\frac{\delta_i^2}{\theta_i'} \leq \frac{2kn^{j\beta}}{n^2} + \frac{2k^{1.5} n^{j\beta}}{n^{2.5}} + \frac{k^2 \alpha_i'^2 \theta_i'}{n^2} \tag{32}$$

where $n^\beta = 2$ is used to derive the equation. Since $k = o(n)$, the second term above is absorbed in the first, leading to the second term of inequality (e) of (31) after aggregating elements of the sum into intervals. The sum over i of the last term of (32) is $o(1)$ since $k = o(n)$. This sum is absorbed into the first term of inequality (e) of (31). Inequality (f) of (31) is obtained since

$$\sum_{j=1}^{J_1} k_j n^{j\beta} = \sum_{j=1}^{J_1} k_j 2^j \leq 2n. \tag{33}$$

Inequality (33) follows since $k_1 \leq n$, $k_2 \leq (n - k_1)/2$, $k_3 \leq (n - k_1)/4 - k_2/2$, and so on, until

$$k_{J_1} \leq \frac{n}{2^{J_1 - 1}} - \sum_{\ell=1}^{J_1 - 1} \frac{k_\ell}{2^{J_1 - \ell}} \Rightarrow \sum_{j=1}^{J_1} k_j 2^j \leq 2n. \tag{34}$$

The reason for these relations are the lower limits of the J_1 intervals that restrict the number of parameters inside the interval. The restriction is done in order of intervals, so that the used probabilities are subtracted, leading to the series of equations.

Plugging the bounds of (23) and (31) into (30), we obtain

$$\begin{aligned} &nR_n(L^*, \boldsymbol{\theta}) \\ &\quad \leq \log k + (1 + \varepsilon) \frac{k - 1}{2} \log \frac{n(\log n)^2}{k^3} + 5(\log e)k \\ &\quad \leq (1 + \varepsilon') \frac{k - 1}{2} \log \frac{n(\log n)^2}{k^3} \end{aligned} \tag{35}$$

where we absorb second-order terms in ε'. Replacing ε' by ε normalizing the redundancy per symbol by n, the bound of the second region of (15) is proved. Since $\log \log n$ can also be absorbed in ε, the first region is also proved. The code proposed, however, will lead to redundancy whose second order is larger than obtained for standard i.i.d. optimal codes that do not exploit the distribution monotonicity for fixed k. This is because the grid used here is too dense for the fixed k case. One can use the standard i.i.d. codes for tighter second-order bounds for fixed k.

We now consider the larger values of k, i.e., $n^{1/3} < k = O(n)$. The idea of the proof is the same. However, we need to partition the probability space to different intervals, the spacing within an interval must be optimized, and the parameters' description cost must be bounded differently, because now there are more parameters quantized than points in the quantization grid. Define the jth interval as

$$I_j = \left[\frac{n^{(j-1)\beta}}{n^2}, \frac{n^{j\beta}}{n^2} \right), \ 1 \leq j \leq J_2, \tag{36}$$

where $J_2 = \lceil 2/\beta \rceil = \lceil 2 \log n \rceil$. Again, let $k_j = |\theta_i \in I_j|$ denote the number of probabilities in $\boldsymbol{\theta}$ that are in interval I_j. It could be possible to use the intervals as defined in (17), but this

would not guarantee bounded redundancy in the rate we require if there are very small probabilities $\theta_i \ll 1/n$. Note that the smallest nonzero component of $\hat{\boldsymbol{\theta}}$ is $1/n$. However, this is not necessarily the case for $\hat{\boldsymbol{\theta}}_{\mathcal{M}}$. The latter may consist of smaller nonzero probabilities for sequences that do not obey the monotonicity of the distribution. Therefore, the interval definition in (17) can be used for larger alphabets only if the probabilities of the symbols are known to be bounded. Define the spacing in interval j as

$$\Delta_j^{(2)} = \frac{n^{j\beta}}{n^{2+\alpha}}, \tag{37}$$

where $\alpha > 0$ is a parameter to be optimized. Similarly to (19), the interval cardinality here is

$$|I_j| \le 0.5 \cdot n^{\alpha} \ \forall j : j = 1, 2, \ldots, J_2. \tag{38}$$

In a similar manner to the definition of $\boldsymbol{\tau}$ in (20), we define

$$\begin{aligned}\boldsymbol{\eta} &= (\eta_1, \eta_2, \ldots) \\ &= \left(\frac{1}{n^2}, \frac{1}{n^2} + \frac{2}{n^{2+\alpha}}, \ldots, \frac{2}{n^2}, \frac{2}{n^2} + \frac{4}{n^{2+\alpha}}, \ldots\right). \end{aligned} \tag{39}$$

The cardinality of $\boldsymbol{\eta}$ is

$$B_2 \triangleq |\boldsymbol{\eta}| \le 0.5 \cdot n^{\alpha} \lceil 2 \log n \rceil \le n^{\alpha} \lceil \log n \rceil. \tag{40}$$

We now perform the encoding similarly to the small k case, where we allow quantization to nonzero values to the components of $\boldsymbol{\varphi}$ up to $i = n^2$. (This is more than needed but is possible since $\eta_1 = 1/n^2$.) Encoding is performed similarly to the small k case. Thus, similarly to (30), we have

$$nR_n\left(L^*, \boldsymbol{\theta}\right) \le 2 \log n + E_{\theta} L_R\left(\boldsymbol{\theta}'\right) + E_{\theta} \log \frac{P_{\theta}\left(X^n\right)}{P_{\theta'}\left(X^n\right)}, \tag{41}$$

where the first term is due to allowing up to $\hat{k} = n^2$. Since usually in this region $k \ge B_2$ (except the low end), the description of vectors $\boldsymbol{\varphi}$ and $\boldsymbol{\theta}'$ is done by coding the cardinality of $|\varphi_i = \eta_j|$ and $|\theta_i' = \eta_j|$, respectively, i.e., for each grid point the code describes how many letters have probability quantized to this point. This idea resembles coding profiles of patterns, as done in [22]. However, unlike the method in [22], here, many probability parameters of symbols with different occurrences are mapped to the same grid point by quantization. The number of parameters mapped to a grid point of $\boldsymbol{\eta}$ is coded using Elias' representation of the integers. Hence, in a similar manner to (23),

$$\begin{aligned} & L_R(\boldsymbol{\theta}') \qquad (42) \\ & \overset{(a)}{\le} \sum_{j=1}^{B_2} \{1 + \log(|\theta_i' = \eta_j| + 1) + \\ & \quad 2 \log[1 + \log(|\theta_i' = \eta_j| + 1)]\} \\ & \overset{(b)}{\le} B_2 + B_2 \log \frac{k + B_2}{B_2} + 2B_2 \log\log \frac{k + B_2}{B_2} + o(B_2) \\ & \overset{(c)}{\le} \begin{cases} (1+\varepsilon)(\log n)(\log \frac{k}{n^{\alpha-\varepsilon}})n^{\alpha}, & n^{\alpha} < k = o(n), \\ (1+\varepsilon)(1-\alpha)(\log n)^2 n^{\alpha}, & n^{\alpha} < k = O(n). \end{cases} \end{aligned}$$

The additional 1 term in the logarithm in (a) is for 0 occurrences, (b) is obtained similarly to step (a) of (23), absorbing all second-order terms in the last term. To obtain (c), we first assume, for the first region, that $kn^{\varepsilon} \gg B_2$ (an assumption that must be later validated with the choice of α). Then, second-order terms are absorbed in ε. The extra n^{ε} factor is unnecessary if $k \gg B_2$. The second region is obtained by upper bounding k without this factor. It is possible to separate the first region into two regions, eliminate this factor in the lower region, and obtain a more complicated, yet tighter, expression in the upper region, where $k \sim \Theta(n^{1/3})$.

Now, similarly to (31), we obtain

$$\begin{aligned} E_{\theta} \log \frac{P_{\theta}(X^n)}{P_{\theta'}(X^n)} &\le n(\log e) \sum_{i=1}^{k} \frac{\delta_i^2}{\theta_i'} \\ &\overset{(a)}{\le} O(1) + \frac{2 \log e}{n^{1+2\alpha}} \sum_{j=1}^{J_2} k_j n^{j\beta} \\ &\overset{(b)}{\le} 4(\log e) n^{1-2\alpha} + O(1). \end{aligned} \tag{43}$$

The first term of inequality (a) is obtained under the assumption that $k = O(n)$, $\theta_i' \ge 1/n^2$, and $|\delta_i| \le 1/n^2$. Similarly to the last two terms of (32), we obtain an additional $O(1)$ term for extra offset costs of the larger probability symbols due to many small probability symbols if they exist. For the second term $|\delta_i| \le n^{j\beta}/n^{2+\alpha}$, and $\theta_i' \ge n^{(j-1)\beta}/n^2$. Inequality (b) is obtained in a similar manner to inequality (f) of (31), where the sum is shown similarly to be $2n^2$.

Summing up the contributions of (42) and (43) in (41), $\alpha = 1/3$ is shown to minimize the total cost (to first order). This choice of α also satisfies the assumption of step (c) in (42). Using $\alpha = 1/3$, absorbing all second-order terms in ε and normalizing by n, we obtain the remaining two regions of the bound in (15). It should be noted that the proof here would give a bound of $O(n^{1/3+\varepsilon})$ up to $k = O(n^{4/3})$. If the intervals in (17) were used for bounded distributions, the coefficients of the last two regions will be reduced by a factor of 2.

The proof up to this point assumes that k is known in advance. This is important for the code resulting in the bounds for the first two regions because the quantization grid depends on k. Specifically, if in building the grid, k is underestimated, the description cost of $\boldsymbol{\varphi}$ increases. If k is overestimated, the quantization cost will increase. Also, if the code of larger k's is used for a smaller k, a larger bound than necessary results. To solve this, the optimization that chooses $L^*(x^n)$ is done over all possible values of k (greater than or equal to the maximal symbol occurring in x^n), i.e., every greater k in the first construction, and the construction of the code for the top regions. For fixed k, a standard optimal code for nonmonotonic distributions can also be constructed. For every small k, a different construction is done, using the appropriate k to determine the spacing in each interval. The value of k yielding the shortest code word is then used. Elias' coding for the integers can be used to designate k with $O(\log k)$ prefix bits. The analysis continues as before. This does not change the redundancy to first order, giving all four regions of the bound in (15), even if k is unknown in advance. This concludes the proof of Theorem 4. ■

Proof [Corollary 1]: The proof branches off the proof of Theorem 4 by improving on several steps, and mainly on the choice of α. First, like the partitioning of the probability space into three intervals in [35], we can partition the probability space into two intervals here, $(0, 1/n^{\alpha}]$ and $(1/n^{\alpha}, 1)$. (Since we can

have probabilities smaller than $1/n$, we cannot use a bottom interval of $(0, n^\alpha/n]$ here.) In the top interval, we need at most $(1+\varepsilon)n^\alpha \log n$ bits to describe the monotonic ML probabilities of at most n^α symbols whose probabilities are in this interval. Quantizing all these probabilities with $1/n$ resolution yields $o(1)$ additional quantization cost. (This can be shown following similar steps as (43) with a choice of $\alpha = (1+o(1))/3$.) Using $1/n^\alpha$ as the upper limit on the total number of intervals in (36) instead of 1 now yields $J_2' = (2-\alpha)\log n$. It then follows, similarly to (40), that $B_2' \leq 0.5(2-\alpha)n^\alpha \log n$. Next, the description costs in (42) reduce by the factor $0.5(2-\alpha)$. Combining the costs in (41), using the new description cost, the quantization cost of (43), and absorbing the cost of the probability parameter top interval and other second-order terms in ε yields

$$\begin{aligned} &nR_n\left(L^*, \boldsymbol{\theta}\right) \\ &\quad \leq (1+\varepsilon)0.5(2-\alpha)(1-\alpha)n^\alpha(\log n)^2 + 4(\log e)n^{1-2\alpha} \end{aligned} \tag{44}$$

for $k = O(n)$. (Similarly, a more complex expression can be written for $k = o(n)$.) A choice of

$$\alpha = \frac{1}{3} - \frac{2}{3}\frac{\log\log n}{\log n} + \frac{\log \nu}{\log n} \tag{45}$$

for some parameter ν minimizes (44) yielding

$$nR_n\left(L^*, \boldsymbol{\theta}\right) \leq (1+\varepsilon)\left(\frac{5\nu}{9} + \frac{4\log e}{\nu^2}\right) n^{1/3}(\log n)^{4/3}. \tag{46}$$

Taking $\nu = (72(\log e)/5)^{1/3} \approx 2.75$ minimizes (46), yielding a coefficient of less than 2.3 in (46). Letting $n \to \infty$, absorbing all second-order terms in the gap of the coefficient to 2.3 proves the second region of (16). Using the same value of ν for the resulting terms for the first region in a similar manner, proves the first region. A slightly tighter bound can be obtained for the first region if the value of ν is optimized for the specific value of k. ■

V. Upper Bounds for Fast Decaying Distributions

This section shows that with some mild conditions on the source distribution, the same redundancy upper bounds achieved for finite monotonic distributions can be achieved even if the monotonic distribution is over an infinite alphabet. The key observation that leads to this result is that a distribution that decays fast enough will result in only a small number of occurrences of letters from its tail in x^n. Occurrences of these letters will likely not retain the monotonicity. Since there are few such occurrences, they can be handled without increasing the asymptotic behavior of the coding cost. More precisely, fast decaying monotonic distributions can be viewed as if they have some effective bounded alphabet size. Occurrences of symbols outside this limited alphabet are rare. We present two theorems and a corollary that upper bound the redundancy when coding with such unknown monotonic distributions. The first theorem also provides a slightly stronger bound (with smaller coefficient) for $k = O(n)$. For slower decays with more occurring symbols from the distribution tail, the redundancy order does increase due to the penalty of identifying these symbols in a sequence. However, we show, consistently with the results in [11], that as long as the entropy of the source is finite, a universal code, in the sense of diminishing redundancy per symbol, still exists. We begin with stating the two theorems and the corollary, then the proofs are presented. The section is concluded with three examples of typical monotonic distributions over the integers, that demonstrate both cases of fast and slow decays.

A. Upper Bounds

We begin with some notation. Fix an arbitrary small $\varepsilon > 0$, and let $n \to \infty$. Define $m \triangleq m_\rho \triangleq n^\rho$ as the effective alphabet size, where $\rho > \varepsilon$. (Note that $\rho = (\log m)/(\log n)$.) Let

$$\begin{aligned} &\mathcal{R}_n(m) \\ &\triangleq \begin{cases} \frac{m-1}{2}\log\frac{n}{m^3}, & m = o(n^{1/3}), \\ \frac{1}{2}\left(\rho+\frac{1}{3}\right)\left(\rho+\varepsilon-\frac{1}{3}\right)(\log n)^2 n^{1/3}, & \text{o.w.} \end{cases} \end{aligned} \tag{47}$$

Theorem 5:

I. Fix an arbitrarily small $\varepsilon > 0$, and let $n \to \infty$. Let x^n be generated by an i.i.d. monotonic distribution $\boldsymbol{\theta} \in \mathcal{M}$. If there exists m^*, such that

$$\sum_{i>m^*} n\theta_i \log i = o\left[\mathcal{R}_n\left(m^*\right)\right], \tag{48}$$

then, there exists a code with length function $L^*(\cdot)$, such that

$$R_n\left(L^*, \boldsymbol{\theta}\right) \leq \frac{(1+\varepsilon)}{n}\mathcal{R}_n\left(m^*\right) \tag{49}$$

for the monotonic distribution $\boldsymbol{\theta}$.

II. If there exists m^* for which $\rho^* = o\left(n^{1/3}/(\log n)\right)$, such that

$$\sum_{i>m^*} \theta_i \log i = o(1), \tag{50}$$

then, there exists a universal code with length function $L^*(\cdot)$, such that

$$R_n\left(L^*, \boldsymbol{\theta}\right) = o(1). \tag{51}$$

Theorem 5 shows that redundancy bounds of the same order as those obtained for finite alphabets are achievable for monotonic distributions that decay fast enough (with effective alphabet that does not exceed $O(n^\rho)$ symbols for a fixed ρ). Specifically, very fast decaying distributions, although over infinite alphabets, may even behave like monotonic distributions with $o\left(n^{1/3}\right)$ symbols. The condition in (48) merely means that the cost that a code would incur in order to code very rare symbols, that are larger than the effective alphabet size, is negligible w.r.t. the total cost obtained from other, more likely, symbols. Note that for $m = n$, the bound is tighter than that of the last region of Theorem 4, and a constant of 4/9 replaces 2/3. The second part of the theorem states that if the decay is slow, but the cost of coding rare symbols is still diminishing per symbol, a universal code still exists for such distributions. However, in this case the redundancy will be dominated by coding the rare (out of order) symbols.

Applying the additional steps used to prove Corollary 1 to the proof of the first part of Theorem 5 yields a tighter expression for the second region of $\mathcal{R}_n(m)$ in (47), which for fixed ρ is

$\Theta\left(n^{1/3}(\log n)^{4/3}\right)$. While Theorem 5 bounds the redundancy decay rate for two extremes, a more general theorem can provide the redundancy rates for coding an unknown monotonic distribution whose decay rate is between these extremes. As the examples at the end of this section show, the next theorem is very useful for slower decaying distributions. It also encapsulate the derivation of a tighter bound as that in Corollary 1 for the more general case.

Theorem 6: Fix an arbitrarily small $\varepsilon > 0$, and let $n \to \infty$. Let x^n be generated by an i.i.d. monotonic distribution $\boldsymbol{\theta} \in \mathcal{M}$. Then, there exists a code with length function $L^*(\cdot)$, that achieves redundancy

$$nR_n(L^*, \boldsymbol{\theta}) \leq (1+\varepsilon)\cdot \min_{\alpha>0, \rho: \rho \geq \alpha+\varepsilon} \left\{ \frac{(\rho+\alpha)(\rho-\alpha)(\log n)^2 n^\alpha}{2} + \frac{5n^{1-2\alpha}}{\ln 2} + \left(1+\frac{1}{\rho}\right) n \sum_{i>n^\rho} \theta_i \log i \right\} \tag{52}$$

for coding sequences generated by the source $\boldsymbol{\theta}$.

The theorems above lead to the following corollary.

Corollary 2: As $n \to \infty$, sequences generated by monotonic distributions with $H_\theta(X) = O(1)$ are universally compressible in the average sense.

Corollary 2 shows that sequences generated by finite entropy monotonic distributions can be compressed in the average with diminishing per symbol redundancy. This result is consistent with the results shown in [11].

We continue with proving the two theorems and the corollary.

Proof: The proof of both theorems is constructive in a similar manner to the proof of Theorem 4. This time, however, the main idea is first separating the more likely symbols from the unlikely ones. The code first determines the point of this separation $m = n^\rho$. (Note that ρ can be greater than 1.) All symbols $i \leq m$ are considered likely and are quantized and described in a similar manner as in the codes for smaller alphabets. Unlike bounded alphabets, though, a more robust grid is used here to allow larger values of m. The unlikely symbols are coded hierarchically. They are first merged into a single innovation symbol. Then, they are encoded within this symbol by coding their actual values. As long as the decay is fast enough, the average cost of conveying these symbols becomes negligible w.r.t. the cost of coding the likely symbols. If the decay is slower, but still fast enough, as the case described in condition (50), the coding cost of the rare symbols dominates the redundancy, which is still diminishing. The description length of likely symbols is bounded as in the proof of Theorem 4, consisting of description of the probability grid points and the quantization cost. In order to determine the best value of m for a given sequence, all values are tried and the one yielding the shortest description is used for coding a specific x^n. The steps described prove both Theorems 5 and 6.

Let $m \geq 2$ determine the number of likely symbols in the alphabet. For a given m, define

$$S_m \triangleq \sum_{i>m} \theta_i, \tag{53}$$

as the total probability of the remaining symbols. Given $\boldsymbol{\theta}$, m and S_m, a probability

$$P\left(x^n | m, S_m, \boldsymbol{\theta}\right) \triangleq \left[\prod_{i=1}^{m} \theta_i^{n_x(i)}\right] \cdot S_m^{n_x(x>m)} \cdot \prod_{i>m} \left(\frac{n_x(i)}{n_x(x>m)}\right)^{n_x(i)} \tag{54}$$

can be computed for x^n, where $n_x(i)$ counts occurrences of symbol i in x^n, and $n_x(x > m)$ is the count of all symbols greater than m in x^n. This probability mass function clusters all symbols greater than m into one innovation symbol. Then, it uses the ML estimate of each to distinguish among them in the clustered symbol.

For every m, we can define a quantization grid $\boldsymbol{\xi}_m$, in a similar manner to the proof of Theorem 4, for the first m components of $\boldsymbol{\theta}$. If $m = o(n^{1/3})$, we use $\boldsymbol{\xi}_m = \boldsymbol{\tau}_m$, where $\boldsymbol{\tau}_m$ is the grid defined in (20) with m replacing k. Otherwise, we can use the definition of $\boldsymbol{\eta}$ in (39). However, to obtain tighter bounds for large m, we define a different grid for the larger values of m following similar steps to those in (36)–(40). First, define the jth interval as

$$I_j = \left[\frac{n^{(j-1)\beta}}{n^{\rho+2\alpha}}, \frac{n^{j\beta}}{n^{\rho+2\alpha}}\right), \; 1 \leq j \leq J_\rho, \tag{55}$$

where $\rho = (\log m)/(\log n)$ as defined above, $\alpha > 0$ is a parameter, and $\beta = 1/(\log n)$ as before. Within the jth interval, we define the spacing in the grid by

$$\Delta_j^{(\rho)} = \frac{n^{j\beta}}{n^{\rho+3\alpha}}. \tag{56}$$

As in (38),

$$|I_j| \leq 0.5 \cdot n^\alpha \; \forall j : j = 1, 2, \ldots, J_\rho, \tag{57}$$

and the total number of intervals to describe probabilities less up to $1/n^\alpha$ is

$$J_\rho = \lceil (\rho+\alpha) \log n \rceil. \tag{58}$$

As in the proof of Corollary 1, we use $O\left(n^\alpha \log n\right)$ bits to describe and quantize probabilities greater than $1/n^\alpha$. Similarly to (39), $\boldsymbol{\xi}_m$ is defined as

$$\begin{aligned} \boldsymbol{\xi}_m &= (\xi_1, \xi_2, \ldots) \\ &= \left(\frac{1}{n^{\rho+2\alpha}}, \frac{1}{n^{\rho+2\alpha}} + \frac{2}{n^{\rho+3\alpha}}, \ldots, \frac{2}{n^{\rho+2\alpha}}, \right. \\ &\qquad \left. \frac{2}{n^{\rho+2\alpha}} + \frac{4}{n^{\rho+3\alpha}}, \ldots\right). \end{aligned} \tag{59}$$

The cardinality of $\boldsymbol{\xi}_m$ is thus

$$B_\rho \triangleq |\boldsymbol{\xi}_m| \leq 0.5 \cdot n^\alpha \lceil (\rho+\alpha) \log n \rceil. \tag{60}$$

An mth order quantized version $\boldsymbol{\theta}'_m$ of $\boldsymbol{\theta}$ is obtained by quantizing $\theta_i \le 1/n^\alpha$, $i = 2, 3, \ldots, m$ onto $\boldsymbol{\xi}_m$, such that $\theta'_i \in \boldsymbol{\xi}_m$ for these values of i. Then, the remaining cluster probability S_m is quantized into $S'_m \in [1/n, 2/n, \ldots, 1]$. The parameter θ'_1 is constrained by the quantization of the other parameters. Quantization is performed again in a manner that minimizes the cumulative error but retains monotonicity, and probabilities smaller than ξ_1 are offset by larger symbols as before.

Now, for any $m \ge 2$, let $\boldsymbol{\varphi}_m$ be any monotonic probability vector of cardinality m whose last $m-1$ components are quantized into $\boldsymbol{\xi}_m$ (or coded separately in the upper interval $(1/n^\alpha, 1)$ if such values exist), and let $\sigma_m \in [1/n, 2/n, \ldots, 1]$ be a quantized value of the innovation symbol, such that $\sum_{i=1}^{m} \varphi_{i,m} + \sigma_m = 1$, where $\varphi_{i,m}$ is the ith component of $\boldsymbol{\varphi}_m$. If m, σ_m, and $\boldsymbol{\varphi}_m$ are known, a given x^n can be coded using $P(x^n|m, \sigma_m, \boldsymbol{\varphi}_m)$ as defined in (54), with σ_m replacing S_m, and the m components of $\boldsymbol{\varphi}_m$ replacing the first m components of $\boldsymbol{\theta}$. However, in the universal setting, none of these parameters are known in advance. Furthermore, neither the symbols greater than m nor their conditional ML probabilities are known in advance. Therefore, the total cost of coding x^n using these parameters requires universality costs for describing them. The additional universality cost of coding x^n with probability $P(x^n|m, \sigma_m, \boldsymbol{\varphi}_m)$ thus consists of the following five components: 1) m should be described using Elias' representation with at most $1 + \rho \log n + 2\log(1 + \rho \log n)$ bits. 2) The value of σ_m in its quantization grid should be coded using $\log n$ bits. 3) The m components of $\boldsymbol{\varphi}_m$ require $L_R(\boldsymbol{\varphi}_m)$ bits. 4) The number $c_x(x > m)$ of distinct letters in x^n greater than m is coded using $\log n$ bits. 5) Each letter $i > m$ in x^n is coded. Elias' coding for the integers using $1 + \log i + 2\log(1 + \log i)$ bits can be used, but to simplify the derivation we can also use the code, also presented in [8], that uses no more than $1 + 2\log i$ bits to describe i. In addition, at most $\log n$ bits are required for describing $n_x(i)$ in x^n. For $n \to \infty$, $m \gg 1$, and $\varepsilon > 0$ arbitrarily small, this yields a total cost of

$$\begin{aligned} &L(x^n|m, \sigma_m, \boldsymbol{\varphi}_m) \\ &\quad \le -\log P(x^n|m, \sigma_m, \boldsymbol{\varphi}_m) + L_R(\boldsymbol{\varphi}_m) + \\ &\qquad [(1+\varepsilon)\rho + c_x(x > m) + 2]\log n \\ &\qquad + c_x(x > m) + 2 \sum_{i > m, i \in x^n} \log i \end{aligned} \tag{61}$$

where we assume m is large enough to bound the cost of describing m by $(1 + \varepsilon)\rho \log n$.

The description cost of $\boldsymbol{\varphi}_m$ for $m = o(n^{1/3})$ is bounded by

$$L_R(\boldsymbol{\varphi}_m) \le (1+\varepsilon)\frac{m-1}{2}\log\frac{n}{m^3} \tag{62}$$

using (23) with m replacing k. The $(\log n)^2$ factor in (23) can be absorbed in ε since we limit m to $o(n^{1/3})$, unlike the derivation in (23). For larger values of m, we describe symbol probabilities of $\boldsymbol{\varphi}_m$ in the grid $\boldsymbol{\xi}_m$ in a similar manner to the description of $O(n)$ symbol probabilities in the grid $\boldsymbol{\eta}$ in the proof of Corollary 1. Similarly to (42), we have

$$\begin{aligned} &L_R(\boldsymbol{\varphi}_m) \\ &\quad \le B_\rho + B_\rho \log\frac{n^\rho + B_\rho}{B_\rho} + 2B_\rho \log\log\frac{n^\rho + B_\rho}{B_\rho} + O(B_\rho) \\ &\quad \overset{(a)}{\le} \frac{(1+\varepsilon)}{2}(\rho + \alpha)(\rho + \varepsilon - \alpha)(\log n)^2 n^\alpha \end{aligned} \tag{63}$$

where the term $O(B_\rho)$ absorbs the cost of probabilities larger than $1/n^\alpha$. To obtain inequality (a), we first multiply n^ρ by n^ε in the numerator of the argument of the logarithm. This is only necessary for $\rho \to \alpha$ to guarantee that $n^{\rho+\varepsilon} \gg B_\rho$. Substituting the bound on B_ρ from (60), absorbing second-order terms in the leading ε yields the bound.

A sequence x^n can now be coded using the universal parameters that minimize the sequence description length, i.e.,

$$\begin{aligned} &L^*(x^n) \\ &\quad \triangleq \min_{m' \ge 2} \min_{\sigma_{m'} \in \left[\frac{1}{n}, \frac{2}{n}, \ldots, 1\right]} \min_{\boldsymbol{\varphi}_{m'}: \varphi_i \in \boldsymbol{\xi}_{m'}, i \ge 2} L(x^n|m', \sigma_{m'}, \boldsymbol{\varphi}_{m'}) \\ &\quad \le L(x^n|m, S'_m, \boldsymbol{\theta}'_m) \end{aligned} \tag{64}$$

where the minimization over φ_i also includes values larger than $1/n^\alpha$, using their designated description. The values $\boldsymbol{\theta}'_m$ and S'_m are the true source parameters quantized in the manner described above, and the inequality holds for every m. The minimization on m' should be performed only up to the maximal symbol that occurs in x^n.

Following (61)–(64), up to negligible integer length constraints, the average redundancy using $L^*(\cdot)$ is bounded, for every $m \ge 2$, by

$$\begin{aligned} &nR_n(L^*, \boldsymbol{\theta}) \\ &\quad = E_\theta[L^*(X^n) + \log P_\theta(X^n)] \\ &\quad \overset{(a)}{\le} E_\theta[L(X^n \mid m, S'_m, \boldsymbol{\theta}'_m) + \log P_\theta(X^n)] \\ &\quad \overset{(b)}{\le} E_\theta \log\frac{P_\theta(X^n)}{P(X^n \mid m, S'_m, \boldsymbol{\theta}'_m)} + L_R(\boldsymbol{\theta}'_m) + \\ &\qquad 2\sum_{i > m} P_\theta(i \in X^n)\log i + \\ &\qquad (1+\varepsilon)[E_\theta C_x(X > m) + \rho + 2]\log n \end{aligned} \tag{65}$$

where (a) follows from (64), and (b) follows from averaging on (61) with $\sigma_m = S'_m$, and $\boldsymbol{\varphi}_m = \boldsymbol{\theta}'_m$ with the average on $c_x(x > m)$ absorbed in the leading ε.

Expressing $P_\theta(x^n)$ as

$$P_\theta(x^n) = \left[\prod_{i \le m} \theta_i^{n_x(i)}\right] \cdot S_m^{n_x(x>m)} \cdot \prod_{i > m}\left(\frac{\theta_i}{S_m}\right)^{n_x(i)} \tag{66}$$

and defining $\delta_S \triangleq S_m - S'_m$, the first term of (65) is bounded, for the upper region of m, by

$$\begin{aligned}
&E_\theta \log \frac{P_\theta(X^n)}{P(X^n \mid m, S'_m, \boldsymbol{\theta}'_m)} \\
&\quad \le E_\theta \left[\sum_{i=1}^{m} N_x(i) \log \frac{\theta_i}{\theta'_{i,m}} + N_x(X > m) \log \frac{S_m}{S'_m} + \right. \\
&\qquad \left. \sum_{i>m} N_x(i) \log \frac{\theta_i / S_m}{N_x(i)/N_x(X>m)} \right] \\
&\quad \overset{(a)}{\le} n \cdot \sum_{i=1}^{m} \theta_i \log \frac{\theta_i}{\theta'_{i,m}} + n S_m \log \frac{S_m}{S'_m} \\
&\quad \overset{(b)}{\le} n(\log e) \left[\left(\sum_{i=1}^{m} \frac{\delta_i^2}{\theta'_{i,m}} \right) + \frac{\delta_S^2}{S'_m} \right] \\
&\quad \overset{(c)}{\le} (\log e) \cdot \frac{n \cdot n^\rho}{n^{\rho+2\alpha}} + 2(\log e) n^{1-\rho-4\alpha} \cdot \sum_{j=1}^{J_\rho} k_j n^{j\beta} + \log e \\
&\quad \overset{(d)}{\le} 5(\log e) n^{1-2\alpha} + \log e
\end{aligned} \tag{67}$$

where (a) is since for the third term, the conditional ML probability used for coding is greater than the actual conditional probability assigned to all letters greater than m for every x^n. Hence, the third term is bounded by 0. Expectation is performed for the other terms. Inequality (b) is obtained similarly to (31) where quantization includes the first m components of $\boldsymbol{\theta}$ and the parameter S_m. Then, inequality (c) follows the same reasoning as step (a) of (43). The first term bounds the worst case in which all n^ρ symbols are quantized to $1/n^{\rho+2\alpha}$ with $|\delta_i| \le 1/n^{\rho+2\alpha}$. The second term is obtained where $\theta'_{i,m} \ge n^{(j-1)\beta}/n^{\rho+2\alpha}$ and $|\delta_i| \le n^{j\beta}/n^{\rho+3\alpha}$ for $\theta_i \in I_j$, and $k_j = |\theta_i \in I_j|$ as before. Offsetting of probabilities smaller than ξ_1, if required, results, similarly to (27), in $|\delta_i| \le n^{j\beta}/n^{\rho+3\alpha} + \gamma'_i \theta'_i / n^{2\alpha}$ where $\gamma'_i > 0$ is some constant, and adds negligibly to both terms. The last term of (c) is since $S'_m \ge 1/n$ and $|\delta_S| \le 1/n$. Finally, (d) is obtained similarly to step (b) of (43), where as in (33), $\sum k_j n^{j\beta} \le 2n^{\rho+2\alpha}$. For $m = o(n^{1/3})$, the same initial steps up to step (b) in (67) are applied. The remaining steps in (31) are then applied with m replacing k, yielding a total quantization cost of $5(1+o(1))(\log e)m + \log e$.

To bound the third and fourth terms of (65),

$$P_\theta(i \in X^n) = 1 - (1-\theta_i)^n \le n\theta_i. \tag{68}$$

Similarly,

$$E_\theta C_x(X > m) = \sum_{i>m} P_\theta(i \in X^n) \le n S_m. \tag{69}$$

Combining the dominant terms of the third and fourth terms of (65), we have

$$\begin{aligned}
&2 \sum_{i>m} P_\theta(i \in X^n) \log i + (1+\varepsilon) E_\theta C_x(X > m) \log n \\
&\quad \overset{(a)}{=} \sum_{i>m} P_\theta(i \in X^n) \left[2 \log i + (1+\varepsilon) \log n \right] \\
&\quad \overset{(b)}{\le} \left(2 + \frac{1+\varepsilon}{\rho} \right) \sum_{i>m} P_\theta(i \in X^n) \log i \\
&\quad \overset{(c)}{\le} \left(2 + \frac{1+\varepsilon}{\rho} \right) n \sum_{i>m} \theta_i \log i
\end{aligned} \tag{70}$$

where (a) is because $E_\theta C_x(X > m) = \sum_{i>m} P_\theta(i \in X^n)$, (b) is because for $i > m = n^\rho$, $\log i > \rho \log n$, and (c) follows from (68). Given $\rho > \varepsilon$ for an arbitrary *fixed* $\varepsilon > 0$, the resulting coefficient above is upper bounded by some constant κ.

Summing up the contributions of the terms of (65) from (31), (62), and (70), absorbing second-order terms in a leading ε', we obtain that for $m = o(n^{1/3})$,

$$nR_n(L^*, \boldsymbol{\theta}) \le (1+\varepsilon') \frac{m-1}{2} \log \frac{n}{m^3} + \kappa n \sum_{i>m} \theta_i \log i. \tag{71}$$

For the second region, substituting $\alpha = 1/3$, and summing up the contributions of (67), (63), and (70) to (65), absorbing second-order terms in ε', we obtain

$$\begin{aligned}
&nR_n(L^*, \boldsymbol{\theta}) \\
&\quad \le (1+\varepsilon') \frac{1}{2} \left(\rho + \frac{1}{3} \right) \left(\rho + \varepsilon' - \frac{1}{3} \right) (\log n)^2 n^{1/3} + \\
&\qquad \kappa n \sum_{i>m} \theta_i \log i.
\end{aligned} \tag{72}$$

Using the value of α in (45) instead would yield a tighter expression of $\Theta\left(n^{1/3}(\log n)^{4/3}\right)$ for the first term, and then the value of ν can be optimized to minimize the leading coefficient. Since (71) and (72) hold for every $m > n^\varepsilon$, there exists m^* for which the minimal bound is obtained. To bound the redundancy, we choose this m^*. Now, if the condition in (48) holds, then the second term in (71) and (72) is negligible w.r.t. the first term. Absorbing it in a leading ε, normalizing by n, yields the upper bound of (49), and concludes the proof of the Part I of Theorem 5.

For Part II of Theorem 5, we consider the bound of the second region in (72). If there exists $\rho^* = o\left(n^{1/3}/(\log n)\right)$ for which the condition in (50) holds, then both terms of (72) are of $o(n)$, yielding a total redundancy per symbol of $o(1)$. The proof of Theorem 5 is concluded. □

Now, consider the upper region in (65) with parameters α and ρ taking any valid value. (The code leading to the bound of the upper region can be applied even if the actual effective alphabet size is in the lower region.) We can sum up the contributions of (67), (63), and (70) to (65), absorbing second-order terms in ε. Equation (63) is valid without the middle ε term as long as $\rho \ge \alpha + \varepsilon$. Since, in the upper region of m, $i \ge m$ is large enough, Elias' code for the integers can be used costing $(1+\varepsilon)\log i$ to code i, with $\varepsilon > 0$ which can be made arbitrarily small. Hence, the leading coefficient of the bound in (70) can be replaced by $(1+\varepsilon)(1+1/\rho)$. This yields the expression bounding the redundancy in (52). This expression applies to every valid choice of α and ρ, including the choice that minimizes the expression. Thus, the proof of Theorem 6 is concluded. □

To prove Corollary 2, we use Wyner's inequality [41], which implies that for a finite entropy monotonic distribution

$$\sum_{i \ge 1} \theta_i \log i = E_\theta[\log X] \le H_\theta[X]. \tag{73}$$

Fix an arbitrarily small $\varepsilon > 0$. Since the sum on the left-hand side of (73) is finite if $H_\theta[X]$ is finite, there must exist some n_0 such that $\sum_{i>n_0} \theta_i \log i < \varepsilon$. Let $n > n_0$, then for $m^* = n$

and $\rho^* = 1$, using Theorem 6 with any $\alpha \in (0,1)$, we obtain $R_n(L^*, \boldsymbol{\theta}) < \kappa\varepsilon$ for some fixed constant $\kappa > 0$. The proof of Corollary 2 is concluded. ■

B. Examples

We demonstrate the use of the bounds of Theorems 5 and 6 with three typical distributions over the integers. We specifically show that the redundancy rate of $O\left(n^{1/3+\varepsilon}\right)$ bits overall is achievable when coding sequences generated by many of the typical monotonic distributions, and, in fact, for many distributions faster convergence rates are achievable with the codes proposed. The examples render the assumption reflected in conditions (48) and (50), that very few large symbols appear in x^n, very practical. Specifically, in the phone book example, there may be many rare names, but only very few of them may occur in a certain city. The more common names can constitute most of a possible phone book sequence.

1) Zipf Distribution: Consider the monotonic distributions over the integers [42], [43] of the form

$$\theta_i = \frac{a}{i^{1+\gamma}},\ i = 1, 2, \ldots, \tag{74}$$

where $\gamma > 0$, and a is a normalization coefficient that guarantees that the probabilities over all integers sum to 1. Approximating summation by integration, we can show that

$$S_m \le \frac{a}{\gamma m^{\gamma}} \tag{75}$$

$$\sum_{i>m} \theta_i \log i \le \frac{a}{\ln 2}\left[\frac{\ln m}{\gamma m^{\gamma}} + \frac{1}{\gamma^2 m^{\gamma}}\right] = (1+\varepsilon)\,\frac{a \log m}{\gamma m^{\gamma}} \tag{76}$$

where the last equality holds for $m \to \infty$ with some fixed $\varepsilon > 0$. For $m = n^{\rho}$ and fixed ρ, the sum in (48) is thus $O\left(n^{1-\rho\gamma}\log n\right)$, which is $o\left(n^{1/3}(\log n)^2\right)$ (and even $o\left(n^{1/3}(\log n)^{4/3}\right)$ if the tighter form of the bound is considered) for every $\rho \ge 2/(3\gamma)$, thus satisfying the negligibility condition (48) at least relative to the second region of (47). As long as $\gamma \le 2$ (slow decay), the minimal value of ρ required to guarantee negligibility of the sum in (48) is greater than 1/3. Using Theorem 5, this implies that for $\gamma \le 2$, the second (upper) region of the upper bound in (49) holds with the minimal choice of $\rho^* = 2/(3\gamma)$. Plugging in this value in the second region of (47) [i.e., in (49)] yields the upper bound shown below for this region. For $\gamma > 2$, $2/(3\gamma) < 1/3$. Hence, (48) holds for $m^* = o\left(n^{1/3}\right)$. This means that for the distribution in (74) with $\gamma > 2$, the effective alphabet size is $o\left(n^{1/3}\right)$, and thus the achievable redundancy is in the first region of the bound of (49). Thus, even though the distribution is over an infinite alphabet, its compressibility behavior is similar to a distribution over a relatively small alphabet. To find the exact redundancy rate, we balance between the contributions of (62) and (70) in (65). As long as $1 - \rho\gamma < \rho$, condition (48) holds, and the contribution of rare letters in (70) is negligible w.r.t. the other terms of the redundancy. Equality, implying $\rho^* = 1/(1+\gamma)$, achieves the minimal redundancy rate. Thus, for $\gamma > 2$,

$$\begin{aligned} nR_n(L^*, \boldsymbol{\theta}) &\overset{(a)}{\le} (1+\varepsilon)\left[\frac{a(2\rho^*+1)}{\gamma} n^{1-\rho^*\gamma}\log n + \frac{n^{\rho^*}}{2}(1-3\rho^*)\log n\right] \\ &\overset{(b)}{=} (1+\varepsilon)\left(\frac{a^{\frac{3+\gamma}{1+\gamma}}}{\gamma} + \frac{1-\frac{3}{1+\gamma}}{2}\right) n^{\frac{1}{1+\gamma}}\log n \end{aligned} \tag{77}$$

where the first term in (a) follows from the bounds in (70) and (76), with $m = n^{\rho^*}$, and the second term from that in (62), and (b) follows from $\rho^* = 1/(1+\gamma)$. Note that for a fixed ρ^*, the factor 3 in the first term can be reduced to 2 with Elias' coding for the integers. The results described are summarized in the following corollary.

Corollary 3: Let $\boldsymbol{\theta} \in \mathcal{M}$ be defined in (74). Then, there exists a universal code with length function $L^*(\cdot)$ that has only prior knowledge that $\boldsymbol{\theta} \in \mathcal{M}$, that can achieve universal coding redundancy

$$R_n(L^*, \boldsymbol{\theta}) \le \begin{cases} (1+\varepsilon)\,\frac{1}{18}\left(1+\frac{2}{\gamma}\right)\left(\frac{2}{\gamma}+\varepsilon-1\right)\frac{n^{1/3}(\log n)^2}{n}, & \gamma \le 2, \\ (1+\varepsilon)\left(\frac{a^{\frac{3+\gamma}{1+\gamma}}}{\gamma} + \frac{1-\frac{3}{1+\gamma}}{2}\right)\frac{n^{\frac{1}{1+\gamma}}\log n}{n}, & \gamma > 2. \end{cases} \tag{78}$$

Corollary 3 gives the redundancy rates for all distributions defined in (74). With a tighter form of the bound (choosing α as in (45) and applying to Theorem 6), a tighter bound of $\Theta\left(n^{1/3}(\log n)^{4/3}/n\right)$ can be obtained for the first region. Using the looser bound of Corollary 3, if, for example, $\gamma = 1$, the redundancy is $O\left(n^{1/3}(\log n)^2\right)$ bits overall with coefficient 1/6. For $\gamma = 3$, $O(n^{1/4}\log n)$ bits are required. For faster decays (greater γ) even smaller redundancy rates are achievable.

2) Geometric Distributions: Geometric distributions given by

$$\theta_i = p(1-p)^{i-1};\ i = 1, 2, \ldots, \tag{79}$$

where $0 < p < 1$, decay even faster than the Zipf distribution in (74). Thus, their effective alphabet sizes are even smaller. This implies that a universal code can have even smaller redundancy than that presented in Corollary 3, when coding sequences generated by a geometric distribution (even if this is unknown in advance, and the only prior knowledge is that $\boldsymbol{\theta} \in \mathcal{M}$). Choosing $m = \ell \cdot \log n$, the contribution of low probability symbols in (70) to (65) can be upper bounded by

$$\begin{aligned} &2n\sum_{i>m}\theta_i(\log i + \log n) \\ &\overset{(a)}{\le} 2n(1-p)^m \log n + O\left(n(1-p)^m \log m\right) \\ &\overset{(b)}{=} 2n^{1+\ell\log(1-p)}(\log n) + O\left(n^{1+\ell\log(1-p)}\log\log n\right) \end{aligned} \tag{80}$$

where (a) follows from computing S_m using geometric series, and bounding the second term, and (b) follows from substituting $m = \ell \log n$ and representing $(1-p)^{\ell\log n}$ as $n^{\ell\log(1-p)}$. As long as $\ell \ge 1/(-\log(1-p))$, the expression in (80) is $O(\log n)$,

thus negligible w.r.t. the redundancy upper bound of (49) with $m^* = \ell^* \log n = (\log n)/(-\log(1-p))$. Substituting this m^* in (49), we obtain the following corollary.

Corollary 4: Let $\boldsymbol{\theta} \in \mathcal{M}$ be a geometric distribution defined in (79). Then, there exists a universal code with length function $L^*(\cdot)$ that has only prior knowledge that $\boldsymbol{\theta} \in \mathcal{M}$, that can achieve universal coding redundancy

$$R_n\left(L^*, \boldsymbol{\theta}\right) \leq \frac{1+\varepsilon}{-2\log(1-p)} \cdot \frac{(\log n)^2}{n}. \tag{81}$$

Corollary 4 shows that if $\boldsymbol{\theta}$ parameterizes a geometric distribution, sequences governed by $\boldsymbol{\theta}$ can be coded with average universal coding redundancy of $O\left((\log n)^2\right)$ bits. Their effective alphabet size is $O(\log n)$, implying that larger symbols are very unlikely to occur. For example, for $p = 0.5$, the effective alphabet size is $\log n$, and $0.5(\log n)^2$ bits are required for a universal code. For $p = 0.75$, the effective alphabet size is $(\log n)/2$, and $(\log n)^2/4$ bits are required by a universal code.

3) Slow Decaying Distributions Over the Integers: Up to now, we considered fast decaying distributions, which all achieved the $O(n^{1/3+\varepsilon}/n)$ redundancy rate. We now consider a slowly decaying monotonic distribution over the integers, given by

$$\theta_i = \frac{a}{i\left(\log i\right)^{2+\gamma}},\ i = 2, 3, \ldots \tag{82}$$

where $\gamma > 0$ and a is a normalizing factor (see, e.g., [14], [32], [33]). This distribution has finite entropy only if $\gamma > 0$ (but is a valid infinite entropy distribution for $\gamma > -1$). Unlike the previous distributions, we need to use Theorem 6 to bound the redundancy for coding sequences generated by this distribution. Approximating the sum with an integral, the order of the third term of (52) is

$$n \sum_{i>m} \theta_i \log i = O\left(\frac{n}{(\log m)^{\gamma}}\right). \tag{83}$$

In order to minimize the redundancy bound of (52), we define $\rho = n^{\ell}$. For the minimum rate, all terms of (52) must be balanced. To achieve that, we must have

$$\alpha + 2\ell = 1 - 2\alpha = 1 - \gamma\ell. \tag{84}$$

The solution is $\alpha = \gamma/(4+3\gamma)$ and $\ell = 2/(4+3\gamma)$. Substituting these values in the expression of (52), with $\rho = n^{\ell}$, results in the first term in (52) dominating, and yields the following corollary.

Corollary 5: Let $\boldsymbol{\theta} \in \mathcal{M}$ be defined in (82) with $\gamma > 0$. Then, there exists a universal code with length function $L^*(\cdot)$ that has only prior knowledge that $\boldsymbol{\theta} \in \mathcal{M}$, that can achieve universal coding redundancy

$$R_n\left(L^*, \boldsymbol{\theta}\right) \leq (1+\varepsilon)\, \frac{n^{\frac{\gamma+4}{3\gamma+4}} (\log n)^2}{2n}. \tag{85}$$

In a similar manner to the Zipf distribution, the tighter form of the general upper bound can be used, reducing the $(\log n)^2$ term to $\Theta\left((\log n)^{4/3}\right)$ (with a different leading coefficient). Due to the slow decay rate of the distribution in (82), the effective alphabet size is much greater here. For $\gamma = 1$, for example, it is $n^{n^{2/7}}$. This implies that very large symbols are likely to appear in x^n. As γ increases though, the effective alphabet size decreases, and as $\gamma \to \infty$, $m \to n$. The redundancy rate increases due to the slow decay. For $\gamma \geq 1$, it is $O\left(n^{5/7}(\log n)^2/n\right)$. As $\gamma \to \infty$, since the distribution tends to decay faster, the redundancy rate tends to the finite alphabet rate of $O\left(n^{1/3}(\log n)^2/n\right)$. However, as the decay rate is slower $\gamma \to 0$, a nondiminishing redundancy rate is approached. Note that the proof of Theorem 6 does not limit the distribution to a finite entropy one. Therefore, the bound of (85) applies, in fact, also to $-1 < \gamma \leq 0$. However, for $\gamma \leq 0$, the per-symbol redundancy is no longer diminishing.

VI. Individual Sequences

In this section, we show that if we have side information of the monotonicity of the distribution governing an individual sequence (i.e., its ML distribution), we can universally compress the individual sequence as well as (and even better than) the average case. We next show that in this case the lower bound of Theorem 3 is asymptotically achieved. Moreover, the upper bound derived here for the upper region is tighter than the bounds obtained in Theorem 4 and Corollary 1 for the average case. The reason is that, with the additional side information that $\hat{\boldsymbol{\theta}} \in \mathcal{M}$, we restrict the smallest nonzero symbol probability to $1/n$. This is not the case in the average case, where symbols from a long tail of the distribution can have unordered occurrences in a given sequence. For a specific sequence, we can have $\hat{\boldsymbol{\theta}} \notin \mathcal{M}$, but we still need to describe $\hat{\boldsymbol{\theta}}_{\mathcal{M}} \in \mathcal{M}$ for that sequence. The distributions $\hat{\boldsymbol{\theta}}_{\mathcal{M}}$ may have probability parameters smaller than $1/n$ for symbols $i \notin x^n$ and $j \in x^n$, where $j > i$ (recalling the assumption that for $\boldsymbol{\theta} \in \mathcal{M}$, we must have $\theta_i \geq \theta_j$).

The side information assumed restricts the set of allowable sequences to those which obey the monotonicity, omitting all sequences for which $\hat{\boldsymbol{\theta}} \neq \hat{\boldsymbol{\theta}}_{\mathcal{M}}$ from the set considered. This means that the class considered is smaller than the class considered for the lower bound in Theorem 3. However, in proving Theorem 3, all sequences that do not obey the monotonicity are excluded from the Shtarkov sum [step (b) of (13)], essentially rendering the bound also as a bound on the class containing only sequences that obey the monotonicity requirement.

If one assumes some monotonicity on the symbol probabilities, but the observed sequence diverges from this assumption, the code proposed can still be used to describe the probabilities of the symbols that obey the monotonicity. An additional description is added as a prefix to the code to describe the number of symbols that do not obey the monotonicity, and then $O(\log n)$ bits are used for each such symbol to describe its occurrence count. If the maximal symbol in x^n is $\hat{k}$, as long as $o(\hat{k})$ symbols are out of order for $\hat{k} \leq n^{1/3}$ or $o(n^{1/3})$ symbols are out of order for greater $\hat{k}$, the additional cost of coding the symbols violating the monotonicity is negligible. The method described below can thus still be used. Moreover, it can be shown (see, e.g., [34]) that as long as the largest symbol is polynomial

in n and there are not too many symbols larger than n, diminishing redundancy w.r.t. the monotonic ML probability $\hat{\boldsymbol{\theta}}_{\mathcal{M}}$ can be achieved coding any such x^n. However, this result does not imply cheaper total description length than the one using the true ML $\hat{\boldsymbol{\theta}}$ of x^n, as the loss in using $\hat{\boldsymbol{\theta}}_{\mathcal{M}}$ instead of $\hat{\boldsymbol{\theta}}$ may dominate over the redundancy savings.

Finally, the class of the distributions of all sequences with $\hat{k}$ symbols that obey the monotonicity is identical to the class of the distributions of all patterns with $\hat{k}$ indices for a given $\hat{k}$. The Shtarkov sum on the ML sequence probabilities is not equal in these cases because the pattern ML sequence probability is the sum over the probabilities of all permutations of these sequences. However, the method used for describing the ML i.i.d. distribution within this class can be used to derive tight bounds for coding patterns. (Bounding the quantization cost for patterns, on the other hand, is more complicated.) This was not the case when addressing the average case, as the description cost in the average case for monotonic distributions is more complicated due to the use of the monotonic ML over all sequences, including those not obeying the monotonicity. This section is concluded with the theorem that upper bounds the individual sequence redundancy and its proof.

Theorem 7: Fix an arbitrarily small $\varepsilon > 0$, and let $n \to \infty$. Let x^n be a sequence for which $\hat{\boldsymbol{\theta}} \in \mathcal{M}$, i.e., $\hat{\theta}_1 \geq \hat{\theta}_2 \geq \ldots$. Let $k = \hat{k}$ be the number of letters occurring in x^n. Then, there exists a code $L^*(\cdot)$ that achieves individual sequence redundancy w.r.t. $\hat{\boldsymbol{\theta}}_{\mathcal{M}} = \hat{\boldsymbol{\theta}}$ for x^n which is upper bounded by

$$\hat{R}_n\left(L^*, x^n\right) \leq \begin{cases} (1+\varepsilon)\,\frac{k-1}{2n}\log\frac{n}{k^3}, & k = o\left(n^{1/3}\right) \\ (1+\varepsilon)\,\frac{k-1}{2n}\log\frac{n(\log n)^2}{k^3}, & k \leq n^{1/3} \\ \frac{\left(0.79\log\frac{k}{n^{1/3-\varepsilon}}+0.14\log n\right)(n\log n)^{1/3}}{n}, & n^{1/3} < k = o(n) \\ \frac{0.4(\log n)^{4/3}n^{1/3}}{n}, & k = O(n). \end{cases} \tag{86}$$

Note that by the monotonicity constraint, the number of symbols $\hat{k}$ occurring in x^n also equals to the maximal symbol in x^n. Since, in the individual sequence case, this maximal symbol defines the class considered and also to be consistent with Theorem 3, we use k to characterize the alphabet size of a given sequence. Since $\hat{\boldsymbol{\theta}}$ is monotonic, $\hat{\boldsymbol{\theta}}_{\mathcal{M}} = \hat{\boldsymbol{\theta}}$.

Proof [Theorem 7]: The proof enhances on that of Theorem 4 and Corollary 1. Both regions of the proof apply here, where instead of quantizing $\boldsymbol{\theta}$ to $\boldsymbol{\theta}'$, we quantize $\hat{\boldsymbol{\theta}}$ to $\hat{\boldsymbol{\theta}}'$ in a similar manner, and do not need to average over all sequences. Instead of using any general $\hat{\boldsymbol{\varphi}}$ to code x^n, we can use $\hat{\boldsymbol{\theta}}'$ without any additional optimizations, where $\log n$ bits describe k. The first two regions of (86) are then proved similarly to these regions in Theorem 4.

To prove the bounds of the upper regions, which are tighter than those of Corollary 1, we make several modifications based on now using three major intervals (as in [35]) instead of two. To describe $\hat{\boldsymbol{\theta}}'$, using parameter α, describe the components of $\hat{\boldsymbol{\theta}}'$ separately for three intervals $(1/n, n^{\alpha}/n]$, $(n^{\alpha}/n, 1/n^{\alpha}]$, and $(1/n^{\alpha}, 1]$. For the bottom interval, use $n^{\alpha}\log n$ bits to describe all probability parameters in this interval. For each of the n^{α} points in this interval use at most $\log n$ bits to describe the multiplicity of these values in $\hat{\boldsymbol{\theta}}$. The top interval consists of at most n^{α} probability parameters. Use at most $\log n$ bits to describe the value of each. For both intervals, no quantization is necessary, and the components of $\hat{\boldsymbol{\theta}}'$ are identical to those of $\hat{\boldsymbol{\theta}}$.

As in [35], the middle interval is the one in which the parameters need to be quantized. Partition this interval into $J_2' \triangleq J_2^+ - J_2^-$ smaller intervals, in a similar manner to (17)

$$I_j = \left[\frac{n^{(j-1)\beta}}{n}, \frac{n^{j\beta}}{n}\right),\ J_2^- \leq j \leq J_2^+ \tag{87}$$

where J_2^- and J_2^+ coincide with the end points of the large middle interval. This results in $J_2' \leq (1-2\alpha)\log n$. Partition each interval into grid points with the spacing

$$\Delta_j^{\prime(2)} = \frac{n^{j\beta}}{n^{1+\alpha}}. \tag{88}$$

Similarly to (40), this yields

$$B_2' \leq 0.5(1-2\alpha)n^{\alpha}\log n \tag{89}$$

grid points. Following a similar derivation to that in (42), the description cost of $\hat{\boldsymbol{\theta}}'$ is bounded by

$$L_R(\hat{\boldsymbol{\theta}}') \leq \begin{cases} \frac{1+\varepsilon}{2}(1-2\alpha)(\log n)\left(\log\frac{k}{n^{\alpha-\varepsilon}}\right)n^{\alpha}, & n^{\alpha} < k \leq n^{1-\alpha} \\ \frac{1+\varepsilon}{2}(1-2\alpha)^2(\log n)^2 n^{\alpha}, & k > n^{1-\alpha} \end{cases} \tag{90}$$

where the description cost of the upper and lower large intervals is absorbed in second-order terms, and the second $1-2\alpha$ factor in the upper region results from $k_{mid} \leq n^{1-\alpha}$ due to the lower limit $n^{\alpha-1}$ of the middle large interval.

The number of symbols with parameters in small interval J_2^- is upper bounded by $k_{J_2^-} \leq n^{1-\alpha}$, then, $k_{J_2^-+1} \leq (n^{1-\alpha} - k_{J_2^-})/2$, and so on. Similarly to (33), we have

$$\sum_{j=J_2^-}^{J_2^+} k_j 2^j \leq n^{1-\alpha}\cdot 2^{J_2^- - 1} = n. \tag{91}$$

Thus, following (43) and using (87) and (88), the quantization cost can be upper bounded by

$$n(\log e)\sum_{i=1}^{k}\frac{\delta_i^2}{\hat{\theta}_i'} \leq \frac{2\log e}{n^{2\alpha}}\sum_{j=J_2^-}^{J_2^+} k_j 2^j = 2(\log e)n^{1-2\alpha}. \tag{92}$$

There is thus a factor of 2 reduction over (43) because of the increased lower limit of the first point of quantized parameters.

Combining (90) and (92) for the second region of (90)

$$n\hat{R}_n(L^*, x^n) \leq \frac{1+\varepsilon}{2}(1-2\alpha)^2 n^{\alpha}(\log n)^2 + 2(\log e)n^{1-2\alpha}. \tag{93}$$

Choosing α from (45) yields

$$n\hat{R}_n\left(L^*, x^n\right) \leq (1+\varepsilon)\left(\frac{\nu}{18} + \frac{2\log e}{\nu^2}\right)n^{1/3}(\log n)^{4/3} \tag{94}$$

for this region. Taking $\nu = (72\log e)^{1/3} \approx 4.7$ minimizes (94), resulting in coefficient of less than 0.4 in (94). Letting $n \to \infty$, absorbing all second-order terms in the gap of the coefficient to

0.4 proves the last region of (86). Using the same value of ν for the resulting terms for the third region in a similar manner, proves the third region. A slightly tighter bound can be obtained for the third region if the value of ν is optimized for the specific value of k. ■

VII. Summary and Conclusion

Universal compression of sequences generated by monotonic distributions was studied. We showed that for finite alphabets, if one has the prior knowledge of the monotonicity of a distribution, one can reduce the cost of universality. For alphabets of $o(n^{1/3})$ letters, this cost reduces from $0.5\log(n/k)$ bits per each unknown probability parameter to $0.5\log(n/k^3)$ bits per each unknown probability parameter. Otherwise, for alphabets of $O(n)$ letters, one can compress such sources with overall redundancy of $O(n^{1/3+\varepsilon})$ bits. This is a significant decrease in redundancy from $O(k\log n)$ or $O(n)$ bits overall that can be achieved if no side information is available about the source distribution. Redundancy of $O(n^{1/3+\varepsilon})$ bits overall can also be achieved for much larger alphabets including infinite alphabets for fast decaying monotonic distributions. Sequences generated by slower decaying distributions can also be compressed with diminishing per-symbol redundancy costs under some mild conditions and specifically if they have finite entropy rates. Examples for well-known monotonic distributions demonstrated how the diminishing redundancy decay rates can be computed by applying the bounds that were derived. The general results were shown to also apply to individual sequences whose empirical distributions obey the monotonicity. The techniques used for individual sequences can also be applied to bounding redundancy coding patterns.

Appendix

A. Proof of Theorem 1

The proof follows the same steps used in [30] and [31] to lower bound the maximin redundancies for large alphabets and patterns, respectively, using the weak version of the redundancy-capacity theorem [6]. This version ties between the maximin universal coding redundancy and the capacity of a channel defined by the conditional probability $P_{\theta}(x^n)$. We define a set $\mathbf{\Omega}_{\mathcal{M}_k}$ of points $\boldsymbol{\theta} \in \mathcal{M}_k$. Then, show that these points are *distinguishable* by observing X^n, i.e., the probability that X^n generated by $\boldsymbol{\theta} \in \mathbf{\Omega}_{\mathcal{M}_k}$ appears to have been generated by another point $\boldsymbol{\theta}' \in \mathbf{\Omega}_{\mathcal{M}_k}$ diminishes with n. Then, using Fano's inequality [4], the number of such distinguishable points is a lower bound on $R_n^-(\mathcal{M}_k)$. Since $R_n^+(\mathcal{M}_k) \geq R_n^-(\mathcal{M}_k)$, it is also a lower bound on the average minimax redundancy. The two regions in (6) result from a threshold phenomenon, where there exists a value k_m of k that maximizes the lower bound, and can be applied to all $\mathcal{M}_k$ for $k \geq k_m$.

We begin with defining $\mathbf{\Omega}_{\mathcal{M}_k}$. Let $\boldsymbol{\omega}$ be a vector of grid components, such that the last $k-1$ components θ_i, $i = 2, \ldots, k$, of $\boldsymbol{\theta} \in \mathbf{\Omega}_{\mathcal{M}_k}$ must satisfy $\theta_i \in \boldsymbol{\omega}$. Let ω_b be the bth point in $\boldsymbol{\omega}$, and define $\omega_0 = 0$ and

$$\omega_b \triangleq \sum_{j=1}^{b} \frac{2(j-\frac{1}{2})}{n^{1-\varepsilon}} = \frac{b^2}{n^{1-\varepsilon}},\ b = 1, 2, \ldots. \tag{A.1}$$

Then, for the bth point in $\boldsymbol{\omega}$,

$$b = \sqrt{\omega_b} \cdot \sqrt{n}^{1-\varepsilon}. \tag{A.2}$$

To count the number of points in $\mathbf{\Omega}_{\mathcal{M}_k}$, let us first consider the standard i.i.d. case, where there is no monotonicity requirement, and count the number of points in $\mathbf{\Omega}$, which is defined similarly, but without the monotonicity requirement (i.e., $\mathbf{\Omega}_{\mathcal{M}_k} \subseteq \mathbf{\Omega}$). Let b_i be the index of θ_i in $\boldsymbol{\omega}$, i.e., $\theta_i = \omega_{b_i}$. Then, from (A.1) and (A.2) and since the components of $\boldsymbol{\theta}$ are probabilities

$$\sum_{i=2}^{k} \frac{b_i^2}{n^{1-\varepsilon}} = \sum_{i=2}^{k} \omega_{b_i} = \sum_{i=2}^{k} \theta_i \leq 1. \tag{A.3}$$

It follows that for $\boldsymbol{\theta} \in \mathbf{\Omega}$,

$$\sum_{i=2}^{k} b_i^2 \leq n^{1-\varepsilon}. \tag{A.4}$$

Hence, since the components b_i are nonnegative integers

$$\begin{aligned} M &\triangleq |\mathbf{\Omega}| \\ &\geq \sum_{b_2=0}^{\lfloor\sqrt{n^{1-\varepsilon}}\rfloor} \sum_{b_3=0}^{\lfloor\sqrt{n^{1-\varepsilon}-b_2^2}\rfloor} \cdots \sum_{b_k=0}^{\left\lfloor\sqrt{n^{1-\varepsilon}-\sum_{i=2}^{k-1} b_i^2}\right\rfloor} 1 \\ &\overset{(a)}{\geq} \int_0^{\sqrt{n^{1-\varepsilon}}} \int_0^{\sqrt{n^{1-\varepsilon}-x_2^2}} \cdots \int_0^{\sqrt{n^{1-\varepsilon}-\sum_{i=2}^{k-1} x_i^2}} dx_k \cdots dx_2 \\ &\overset{(b)}{\triangleq} \frac{V_{k-1}\left(\sqrt{n}^{1-\varepsilon}\right)}{2^{k-1}} \end{aligned} \tag{A.5}$$

where $V_{k-1}\left(\sqrt{n}^{1-\varepsilon}\right)$ is the volume of a $k-1$ dimensional sphere with radius $\sqrt{n}^{1-\varepsilon}$, (a) follows from monotonic decrease of the function in the integrand for all integration arguments, and (b) follows since its left-hand side computes the volume of the positive quadrant of this sphere. Note that this is a different proof from that used in [30] and [31] for this step. Applying the monotonicity constraint, all permutations of $\boldsymbol{\theta}$ that are not monotonic must be taken out of the grid. Hence,

$$M_{\mathcal{M}_k} \triangleq |\mathbf{\Omega}_{\mathcal{M}_k}| \geq \frac{V_{k-1}\left(\sqrt{n}^{1-\varepsilon}\right)}{k! \cdot 2^{k-1}}, \tag{A.6}$$

where dividing by $k!$ is a worst-case assumption, yielding a lower bound and not an equality. This leads to a lower bound equal to that obtained for patterns in [31] on the number of points in $\mathbf{\Omega}_{\mathcal{M}_k}$. Specifically, the bound achieves a maximal value for $k_m = \left(\pi n^{1-\varepsilon}/2\right)^{1/3}$ and then decreases to eventually

become smaller than 1. However, for $k > k_m$, one can consider a monotonic distribution for which all components $\theta_i; i > k_m$, of $\boldsymbol{\theta}$ are zero, and use the bound for k_m.

Distinguishability of $\boldsymbol{\theta} \in \boldsymbol{\Omega}_{\mathcal{M}_k}$ is a direct result of distinguishability of $\boldsymbol{\theta} \in \boldsymbol{\Omega}$, which is shown in Lemma 3.1 in [30]. The lemma states the following: there exits an estimator $\hat{\boldsymbol{\Theta}}_g(X^n) \in \boldsymbol{\Omega}$ for which the estimate $\hat{\boldsymbol{\theta}}_g$ satisfies $\lim_{n\to\infty} P_\theta\left(\hat{\boldsymbol{\theta}}_g \neq \boldsymbol{\theta}\right) = 0$ for all $\boldsymbol{\theta} \in \boldsymbol{\Omega}$. Since this is true for all points in $\boldsymbol{\Omega}$, it is also true for all points in $\boldsymbol{\Omega}_{\mathcal{M}_k} \subseteq \boldsymbol{\Omega}$, where now, $\hat{\boldsymbol{\theta}}_g \in \boldsymbol{\Omega}_{\mathcal{M}_k}$. Assuming all points in $\boldsymbol{\Omega}_{\mathcal{M}_k}$ are equally probable to generate X^n, we can define an average error probability $P_e \triangleq \Pr\left[\hat{\boldsymbol{\Theta}}_g(X^n) \neq \boldsymbol{\Theta}\right] = \sum_{\boldsymbol{\theta}\in\boldsymbol{\Omega}_{\mathcal{M}_k}} P_\theta\left(\hat{\boldsymbol{\theta}}_g \neq \boldsymbol{\theta}\right)/M_{\mathcal{M}_k}$. Using the redundancy-capacity theorem,

$$\begin{aligned} nR_n^-\left[\mathcal{M}_k\right] &\geq C\left[\mathcal{M}_k \to X^n\right] \\ &\overset{(a)}{\geq} I[\boldsymbol{\Theta}; X^n] = H\left[\boldsymbol{\Theta}\right] - H\left[\boldsymbol{\Theta}|X^n\right] \\ &\overset{(b)}{=} \log M_{\mathcal{M}_k} - H\left[\boldsymbol{\Theta}|X^n\right] \\ &\overset{(c)}{\geq} (1-P_e)\left(\log M_{\mathcal{M}_k}\right) - 1 \\ &\overset{(d)}{\geq} (1-o(1))\log M_{\mathcal{M}_k} \end{aligned} \tag{A.7}$$

where $C\left[\mathcal{M}_k \to X^n\right]$ denotes the capacity of the channel between $\mathcal{M}_k$ and the observation X^n, and $I[\boldsymbol{\Theta}; X^n]$ is the mutual information induced by the joint distribution $\Pr\left(\Theta = \theta\right) \cdot P_\theta\left(X^n\right)$. Inequality (a) follows from the definition of capacity, equality (b) from the uniform distribution of $\boldsymbol{\Theta}$ in $\boldsymbol{\Omega}_{\mathcal{M}_k}$, inequality (c) from Fano's inequality, and (d) follows since $P_e \to 0$. Lower bounding the expression in (A.6) for the two regions (obtaining the same bounds as in [31]), then using (A.7), normalizing by n, and absorbing second-order terms in ε, yields the two regions of the bound in (6). The proof of Theorem 1 is concluded. □

B. Proof of Theorem 2

To prove Theorem 2, we use the *random-coding* strong version of the redundancy-capacity theorem [19]. The idea is similar to the weak version used in Appendix A. We assume that grids $\boldsymbol{\Omega}_{\mathcal{M}_k}$ of points are uniformly distributed over $\mathcal{M}_k$, and one grid is selected randomly. Then, a point in the selected grid is randomly selected under a uniform prior to generate X^n. The random choice of a grid and then of a source in the grid must uniformly cover the whole space $\mathcal{M}_k$. Showing distinguishability within a selected grid, for every possible random choice of $\boldsymbol{\Omega}_{\mathcal{M}_k}$, implies that a lower bound on the cardinality of $\boldsymbol{\Omega}_{\mathcal{M}_k}$ for every possible choice is essentially a lower bound on the overall sequence redundancy for most sources in $\mathcal{M}_k$.

The construction of $\boldsymbol{\Omega}_{\mathcal{M}_k}$ is identical to that used in [31] to construct a grid of sources that generate patterns. We pack spheres of radius $n^{-0.5(1-\varepsilon)}$ in the parameter space defining $\mathcal{M}_k$. The set $\boldsymbol{\Omega}_{\mathcal{M}_k}$ consists of the center points of the spheres. To cover the space $\mathcal{M}_k$, we randomly select a random shift of the whole lattice under a uniform distribution. The cardinality of $\boldsymbol{\Omega}_{\mathcal{M}_k}$ is lower bounded by the relation between the volume of $\mathcal{M}_k$, which equals (as shown in [31]) $1/[(k-1)!k!]$, and the volume of a single sphere, with factoring also of a packing density (see, e.g., [3]). This yields (55) in [31]

$$M_{\mathcal{M}_k} \geq \frac{1}{(k-1)!\cdot k!\cdot V_{k-1}\left(n^{-0.5(1-\varepsilon)}\right)\cdot 2^{k-1}} \tag{B.1}$$

where $V_{k-1}\left(n^{-0.5(1-\varepsilon)}\right)$ is the volume of a $k-1$ dimensional sphere with radius $n^{-0.5(1-\varepsilon)}$ (see, e.g., [3] for computation of this volume).

For distinguishability, it is sufficient to show that there exists an estimator $\hat{\boldsymbol{\Theta}}_g(X^n) \in \boldsymbol{\Omega}_{\mathcal{M}_k}$ such that $\lim_{n\to\infty} P_\Theta\left[\hat{\boldsymbol{\Theta}}_g(X^n) \neq \boldsymbol{\Theta}\right] = 0$ for every choice of $\boldsymbol{\Omega}_{\mathcal{M}_k}$ and for every choice of $\boldsymbol{\Theta} \in \boldsymbol{\Omega}_{\mathcal{M}_k}$. This is already shown in [30, Lemma 4.1] for a larger grid $\boldsymbol{\Omega}$ of i.i.d. sources, which is constructed identically to $\boldsymbol{\Omega}_{\mathcal{M}_k}$ over the complete $k-1$ dimensional probability simplex. The lemma states the following: let $\boldsymbol{\Theta} \in \boldsymbol{\Omega}$ be a randomly selected point in a grid $\boldsymbol{\Omega}$. Let a random sequence X^n be governed by $P_\Theta\left(X^n\right)$. Then, there exists a decision rule that chooses a point $\hat{\Theta}_g\left(X^n\right) \in \boldsymbol{\Omega}$, such that $\lim_{n\to\infty} P_\Theta\left[\hat{\boldsymbol{\Theta}}_g(X^n) \neq \boldsymbol{\Theta}\right] = 0$. By the monotonicity requirement, for every $\boldsymbol{\Omega}_{\mathcal{M}_k}$, there exists an i.i.d. $\boldsymbol{\Omega}$, such that $\boldsymbol{\Omega}_{\mathcal{M}_k} \subseteq \boldsymbol{\Omega}$. Since [30, Lemma 4.1] holds for $\boldsymbol{\Omega}$, it then must also hold for the smaller grid $\boldsymbol{\Omega}_{\mathcal{M}_k}$. Now, since all the conditions of the strong random-coding version of the redundancy-capacity theorem hold, taking the logarithm of the bound in (B.1), absorbing second-order terms in ε, and normalizing by n, leads to the first region of the bound in (8). By [19, Th. 3], since for any fixed arbitrarily small $\varepsilon > 0$ we have $P_e \triangleq P_\Theta\left[\hat{\boldsymbol{\Theta}}_g(X^n) \neq \boldsymbol{\Theta}\right] \to 0$, then, $\mu_n\left(A_\varepsilon(n)\right) \to 0$, thus completing the proof for the first region of the bound.

The second region of the bound is handled in a manner related to the second region of the bound of Theorem 1. However, here, we cannot simply set the probability of all symbols $i > k_m$ to zero, because all possible valid sources must be included in one of the grids $\boldsymbol{\Omega}_{\mathcal{M}_k}$ to achieve a complete covering of $\mathcal{M}_k$. As was done in [31], we include sources with $\theta_i > 0$ for $i > k_m$ in the grids $\boldsymbol{\Omega}_{\mathcal{M}_k}$, but do not include them in the lower bound on the number of grid points. Instead, for $k > k_m$, we bound the number of points in a k_m-dimensional cut of $\mathcal{M}_k$ for which the remaining $k - k_m$ components of $\boldsymbol{\theta}$ are very small (and insignificant). This analysis is valid also for $k > n$. In proving distinguishability, however, we must take into account the effect of the additional sources in the grid, and make sure that the existence of these sources in $\boldsymbol{\Omega}_{\mathcal{M}_k}$ does not lead to nondiminishing error probability. Lemma 6.1 in [31] shows that $\lim_{n\to\infty} P_\Theta\left[\hat{\boldsymbol{\Theta}}_g(X^n) \neq \boldsymbol{\Theta}\right] = 0$ for $k > k_m$ for i.i.d. nonmonotonically restricted grid of sources $\boldsymbol{\Omega}$. The proof is given in [31, Appendix D]. As before, it carries over to monotonic distributions, since as before, for each $\boldsymbol{\Omega}_{\mathcal{M}_k}$, there exists an unrestricted corresponding $\boldsymbol{\Omega}$, such that $\boldsymbol{\Omega}_{\mathcal{M}_k} \subseteq \boldsymbol{\Omega}$. The choice of $k_m = 0.5(n^{1-\varepsilon}/\pi)^{1/3}$ gives the maximal bound w.r.t. k. Since, again, all conditions of the strong version of the redundancy-capacity theorem are satisfied, the second region of the bound is obtained. This concludes the proof of Theorem 2. □

ACKNOWLEDGMENT

The author would like to express gratitude to the associate editor, W. Szpankowski, for handling this paper and for very valuable comments that helped improving this paper, and also to an anonymous reviewer for providing valuable feedback.

REFERENCES

[1] J. Åberg, Y. M. Shtarkov, and B. J. M. Smeets, "Multialphabet coding with separate alphabet description," in *Proc. Compress. Complexity Seq.*, Jun. 1997, pp. 56–65.

[2] J. Acharya, H. Das, and A. Orlitsky, "Tight bounds on profile redundancy and distinguishability," in *Proc. Natural Inf. Process. Syst.*, Dec. 2012, pp. 3266–3274.

[3] J. H. Conway and N. J. A. Sloane, *Sphere Packings, Lattices and Groups*, 3rd ed. New York, NY, USA: Springer-Verlag, 1998.

[4] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York, NY, USA: Wiley, 2006.

[5] I. Csiszar and J. Korner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York, NY, USA: Academic, 1981.

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

[7] L. D. Davisson and A. Leon-Garcia, "A source matching approach to finding minimax codes," *IEEE Trans. Inf. Theory*, vol. IT-26, no. 2, pp. 166–174, Mar. 1980.

[8] P. Elias, "Universal codeword sets and representation of the integers," *IEEE Trans. Inf. Theory*, vol. IT-21, no. 2, pp. 194–203, Mar. 1975.

[9] B. M. Fitingof, "Optimal coding in the case of unknown and changing message statistics," *Probl. Inf. Transmiss.*, vol. 2, no. 2, pp. 1–7, 1966.

[10] B. M. Fitingof, "The compression of discrete information," *Probl. Inf. Transmiss.*, vol. 3, no. 3, pp. 22–29, 1967.

[11] D. P. Foster, R. A. Stine, and A. J. Wyner, "Universal codes for finite sequences of integers drawn from a monotone distribution," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1713–1720, Jun. 2002.

[12] R. G. Gallager, "Source coding with side information and universal coding," Sep. 1976, unpublished.

[13] A. Garivier, "A lower-bound for the maximin redundancy in pattern coding," *Entropy*, vol. 11, pp. 634–642, 2009.

[14] G. M. Gemelos and T. Weissman, "On the entropy rate of pattern processes," *IEEE Trans. Inf. Theory*, vol. 52, no. 9, pp. 3994–4007, Sep. 2006.

[15] L. Györfi, I. Páli, and E. C. van der Meulen, "There is no universal code for an infinite source alphabet," *IEEE Trans. Inf. Theory*, vol. 40, no. 1, pp. 267–271, Jan. 1994.

[16] N. Jevtić, A. Orlitsky, and N. P. Santhanam, "A lower bound on compression of unknown alphabets," *Theoretical Comput. Sci.*, vol. 332, no. 1–3, pp. 293–311, 2005.

[17] J. C. Kieffer, "A unified approach to weak universal source coding," *IEEE Trans. Inf. Theory*, vol. IT-24, no. 6, pp. 674–682, Nov. 1978.

[18] M. Khosravifard, H. Saidi, M. Esmaeili, and T. A. Gulliver, "The minimum average code for finite memoryless monotone sources," *IEEE Trans. Inf. Theory*, vol. 52, no. 3, pp. 955–975, Mar. 2007.

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

[20] N. Merhav, G. Seroussi, and M. J. Weinberger, "Optimal prefix codes for sources with two-sided geometric distributions," *IEEE Trans. Inf. Theory*, vol. 46, no. 1, pp. 121–135, Jan. 2000.

[21] N. Merhav, G. Seroussi, and M. J. Weinberger, "Coding of sources with two-sided geometric distributions and unknown parameters," *IEEE Trans. Inf. Theory*, vol. 46, no. 1, pp. 229–236, Jan. 2000.

[22] A. Orlitsky, N. P. Santhanam, and J. Zhang, "Universal compression of memoryless sources over unknown alphabets," *IEEE Trans. Inf. Theory*, vol. 50, no. 7, pp. 1469–1481, Jul. 2004.

[23] A. Orlitsky and N. P. Santhanam, "Speaking of infinity," *IEEE Trans. Inf. Theory*, vol. 50, no. 10, pp. 2215–2230, Oct. 2004.

[24] J. Rissanen, "Minimax codes for finite alphabets," *IEEE Trans. Inf. Theory*, vol. IT-24, no. 3, pp. 389–392, May 1978.

[25] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Trans. Inf. Theory*, vol. IT-30, no. 4, pp. 629–636, Jul. 1984.

[26] B. Y. Ryabko, "Coding of a source with unknown but ordered probabilities," *Probl. Inf. Transmiss.*, vol. 15, no. 2, pp. 134–138, Oct. 1979.

[27] G. I. Shamir and D. J. Costello, Jr., "On the redundancy of universal lossless coding for general piecewise stationary sources," *Commun. Inf. Syst.*, vol. 1, no. 3, pp. 305–322, Sep. 2001.

[28] G. I. Shamir and D. J. Costello, Jr., "Universal lossless coding for sources with repeating statistics," *IEEE Trans. Inf. Theory*, vol. 50, no. 8, pp. 1620–1635, Aug. 2004.

[29] G. I. Shamir, "Applications of coding theory to universal lossless source coding performance bounds," in *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, A. Ashikhmin and A. Barg, Eds. Providence, RI, USA: American Mathematical Society, 2005, vol. 68, pp. 21–55.

[30] G. I. Shamir, "On the MDL principle for i.i.d. sources with large alphabets," *IEEE Trans. Inf. Theory*, vol. 52, no. 5, pp. 1939–1955, May 2006.

[31] G. I. Shamir, "Universal lossless compression with unknown alphabets—The average case," *IEEE Trans. Inf. Theory*, vol. 52, no. 11, pp. 4915–4944, Nov. 2006.

[32] G. I. Shamir, "Entropy of patterns of i.i.d. sequences—Part I: General bounds," *IEEE Trans. Inf. Theory*, vol. 54, no. 5, pp. 2263–2277, May 2008.

[33] G. I. Shamir, "Entropy of patterns of i.i.d. sequences—Part II: Bounds for some distributions," *IEEE Trans. Inf. Theory*, vol. 54, no. 5, pp. 2263–2277, May 2008 [Online]. Available: http://arxiv.org/abs/0711.2102, submitted for publication

[34] G. I. Shamir, "Universal source coding for monotonic and fast decaying monotonic distributions," Apr. 2007, Arxiv:cs.IT/0704.0838.

[35] G. I. Shamir, "A new redundancy bound for universal lossless compression of unknown alphabets," in *Proc. 38th Annu. Conf. Inf. Sci. Syst.*, Princeton, NJ, USA, Mar. 17–19, 2004, pp. 1175–1179.

[36] G. I. Shamir, "Universal source coding for monotonic and fast decaying monotonic distributions," in *Proc. IEEE Int. Symp. Inf. Theory*, Nice, France, Jun. 24–29, 2007, pp. 1956–1960.

[37] Y. M. Shtarkov, "Universal sequential coding of single messages," *Probl. Inf. Transmiss.*, vol. 23, no. 3, pp. 3–17, Jul.–Sep. 1987.

[38] Szpankowski and M. J. Weinberger, "Minimax pointwise redundancy for memoryless models over large alphabets," *IEEE Trans. Inf. Theory*, vol. 58, no. 7, pp. 4094–4104, Jul. 2012.

[39] L. R. Varshney and V. K. Goyal, "Ordered and disordered source coding," presented at the Inf. Theory Appl. Workshop, San Diego, CA, USA, Feb. 6–10, 2006.

[40] L. R. Varshney and V. K. Goyal, "On universal coding of unordered data," presented at the Inf. Theory Appl. Workshop, San Diego, CA, USA, Jan. 29, 2007.

[41] A. D. Wyner, "An upper bound on the entropy series," *Inf. Control*, vol. 20, pp. 176–181, 1972.

[42] G. K. Zipf, *The Psychobiology of Language*. New York, NY, USA: Houghton-Mifflin, 1935.

[43] G. K. Zipf, *Human Behaviour and the Principle of Least-Effort*. Cambridge, MA, USA: Addison-Wesley, 1949.

Gil I. Shamir received the B.Sc. (Cum Laude), and M.Sc. degrees from the Technion, Israel Institute of Technology, Haifa, Israel in 1990 and 1997, respectively, and the Ph.D. degree from the University of Notre Dame, Notre Dame, IN, USA, in 2000, all in electrical engineering.

From 1990 to 1995 he participated in research and development of signal processing and communication systems. From 1995 to 1997 he was with the Electrical Engineering Department at the Technion—Israel Institute of Technology, as a graduate student and teaching assistant. From September 1997 to May 2000 he was a Ph.D. student and a research assistant in the Electrical Engineering Department at the University of Notre Dame, and then a post-doctoral fellow until July 2001. During his tenure at Notre Dame he was a fellow of the Center for Applied Mathematics of the university. Between 2001 and 2008 he was with the Electrical and Computer Engineering Department at the University of Utah, and between 2008 and 2009 he was with Seagate Research. He is currently with Google Inc. His main research interests include information theory, machine learning, coding, and communication theory. Dr. Shamir received an NSF CAREER award in 2003.