Minimax Pointwise Redundancy for Memoryless Models Over Large Alphabets

Wojciech Szpankowski, *Fellow, IEEE*, and Marcelo J. Weinberger, *Fellow, IEEE*

***Abstract*—We study the minimax pointwise redundancy of universal coding for memoryless models over large alphabets and present two main results. We first complete studies initiated in Orlitsky and Santhanam deriving precise asymptotics of the minimax pointwise redundancy for all ranges of the alphabet size relative to the sequence length. Second, we consider the minimax pointwise redundancy for a family of models in which some symbol probabilities are fixed. The latter problem leads to a binomial sum for functions with superpolynomial growth. Our findings can be used to approximate numerically the minimax pointwise redundancy for various ranges of the sequence length and the alphabet size. These results are obtained by analytic techniques such as tree-like generating functions and the saddle point method.**

***Index Terms*—Binomial sums, large alphabet, memoryless sources, minimax pointwise redundancy, saddle point methods, tree generating functions.**

I. Introduction

THE classical universal source coding problem [4] is typically concerned with a known source alphabet whose size is much smaller than the sequence length. In this setting, the asymptotic analysis of universal schemes assumes a regime in which the alphabet size remains fixed as the sequence length grows. More recently, the case in which the alphabet size is very large, often comparable to the length of the source sequences, has been studied from two different perspectives. In one setup (motivated by applications such as text compression over an alphabet composed of words), the alphabet is assumed unknown or even infinite (see, e.g., [2], [9], [12], [16], and [18]). In another setup (see, e.g., [15]), the alphabet is still known and finite (as in applications such as speech and image coding), but the asymptotic regime is such that both the size of the alphabet and the length of the source sequence are very large. Notice that, in this scenario, the optimality criteria and the corresponding optimal codes do not differ from the classical approach; rather, it is the asymptotic analysis that is affected.

Manuscript received November 04, 2010; revised April 05, 2012; accepted April 06, 2012. Date of publication May 9, 2012; date of current version June 12, 2012. This work was partially done while W. Szpankowski was visiting Hewlett-Packard Laboratories. W. Szpankowski was supported in part by the NSF STC Grant CCF-0939370, NSF Grants DMS-0800568 and CCF-0830140, AFOSR Grant FA8655-11-1-3076, NSA Grant H98230-08-1-0092, and MNSW Grant N206 369739. This paper was presented in part at the 2010 IEEE International Symposium on Information Theory.

W. Szpankowski is with the Department of Computer Science, Purdue University, IN 47907 USA (e-mail: spa@cs.purdue.edu).

M. J. Weinberger is with Hewlett-Packard Laboratories, CA 94304, USA (e-mail: marcelo.weinberger@hp.com).

Communicated by T. Weissman, Associate Editor for Shannon Theory.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2012.2195769

In this paper, we follow the latter scenario, targeting a classical figure of merit: the *minimax* (worst-case) *pointwise* redundancy (regret) [19]. Specifically, we derive precise asymptotic results for two memoryless model families. To recall, the pointwise redundancy of a code arises in a deterministic setting involving *individual* data sequences, where probability distributions are mere tools for describing a choice of coding strategies. In this framework, given an individual sequence, the pointwise redundancy of a code is measured with respect to a (probabilistic) model family (i.e., a collection of probability distributions that reflects limited knowledge about the data-generating mechanism). The pointwise redundancy determines by how much the code length exceeds that of the code corresponding to the best model in the family (see, e.g., [14] and [23] for an in-depth discussion of this framework). In the minimax pointwise scenario, one designs the best code for the worst-case sequence, as discussed next.

A fixed-to-variable code $C_n : \mathcal{A}^n \to \{0,1\}^*$ is an injective mapping from the set $\mathcal{A}^n$ of all sequences of length n over the finite alphabet $\mathcal{A}$ of size $m = |\mathcal{A}|$ to the set $\{0,1\}^*$ of all binary sequences. We assume that C_n satisfies the prefix condition and denote $L(C_n, x_1^n)$ the code length it assigns to a sequence $x_1^n = x_1, \ldots, x_n \in \mathcal{A}^n$. A prefix code matched to a model P (given by a probability distribution P over $\mathcal{A}^n$) encodes x_1^n with an "ideal" code length $-\log P(x_1^n)$, where $\log := \log_2$ will denote the binary logarithm throughout this paper, and we ignore the integer length constraint. Given a sequence x_1^n, the pointwise redundancy of C_n with respect to a model family $\mathcal{S}$ (such as the family of memoryless models $\mathcal{M}_0$) is, thus, given by

$$R_n(C_n, x_1^n; \mathcal{S}) = L(C_n, x_1^n) + \sup_{P \in \mathcal{S}} \log P(x_1^n).$$

Finally, the minimax pointwise redundancy $R_n^*(\mathcal{S})$ for the family $\mathcal{S}$ is given by

$$R_n^*(\mathcal{S}) = \min_{C_n} \max_{x_1^n} R_n(C_n, x_1^n; \mathcal{S}). \tag{1}$$

This quantity was studied by Shtarkov [19], who found that, ignoring the integer length constraint also for C_n (cf., [5])

$$R_n^*(\mathcal{S}) = \log \left(\sum_{x_1^n} \sup_{P \in \mathcal{S}} P(x_1^n) \right) \tag{2}$$

and is achieved with a code that assigns to each sequence a code length proportional to its maximum-likelihood probability over $\mathcal{S}$. In particular, for $\mathcal{S} = \mathcal{M}_0$, precise asymptotics of $R_n^*(\mathcal{M}_0)$ have been derived in the regime in which the alphabet size m is treated as a *constant* [20] (cf., also [23]). The minimax pointwise redundancy was also studied when both n and m are large, by

Orlitsky and Santhanam [15]. Formulating this scenario as a sequence of problems in which m varies with n, leading term asymptotics for $m = o(n)$ and $n = o(m)$, as well as bounds for $m = \Theta(n)$, are established in [15].[1] The goal of this formulation is to estimate $R_n^*(\mathcal{M}_0)$ for given values of n and m, which fall in one of the aforementioned cases.

In this paper, we first provide, in Theorem 1, precise asymptotics of $R_n^*(\mathcal{M}_0)$ for all ranges of m relative to n. Our findings are obtained by analytic methods of analysis of algorithms [8], [21]. Theorem 1 not only completes the study of [15] by covering all ranges of m (including $m = \Theta(n)$), but also strengthens it by providing more precise asymptotics. Indeed, it will be shown that the error incurred by neglecting lower order terms may actually be quite significant, to the point that, for $m = o(n)$, the first two terms of the asymptotic expansion for constant m given in [20] are a better approximation to $R_n^*(\mathcal{M}_0)$ than the leading term established in [15].

In addition, Theorem 1 also enables a precise analysis of the minimax pointwise redundancy in a more general scenario. Specifically, we consider the alphabet $\mathcal{A} \cup \mathcal{B}$, with $|\mathcal{A}| = m$ and $|\mathcal{B}| = M$, and a (memoryless) model family, denoted $\widetilde{\mathcal{M}}_0$, in which the probabilities of symbols in $\mathcal{B}$ are *fixed*, while m may be large.[2] Such *constrained* model families, which correspond to partial knowledge of the data-generating mechanism, fill the gap between two classical paradigms: one in which a code is designed for a specific distribution in $\mathcal{M}_0$ (Shannon-type coding), and universal coding in $\mathcal{M}_0$. For example, consider a situation in which data sequences from two different sources (over disjoint alphabets) are randomly interleaved (e.g., by a router), as proposed in [1], and assume that one of the sequences is (controlled) simulation data, for which the generating mechanism is known. If we further assume that the switching probabilities are also known, this situation falls under the proposed setting, where $\mathcal{B}$ corresponds to the alphabet of the simulation data. Other constrained model families have been studied in the literature as means to reduce the number of free parameters in the probability model (see [22] for an example motivated in image coding). Given our knowledge of the distribution on $\mathcal{B}$, one would expect to "pay" a smaller price for universality in terms of redundancy. In a probabilistic setting and for m treated as a constant, Rissanen's lower bound on the (average) redundancy [17] is indeed proportional to the number $m - 1$ of free parameters. Moreover, it is easy to see that the leading term asymptotics of the pointwise redundancy of a (sequential) code that uses a fixed probability assignment for symbols in $\mathcal{B}$, and one based on the Krichevskii--Trofimov scheme [13] for symbols in $\mathcal{A}$, are indeed the same as those for $R_n^*(\mathcal{M}_0)$. However, this intuition notwithstanding, notice that the minimax scheme for the combined alphabet does *not* encode the two alphabets separately. Moreover, the analysis is more complex for unbounded m, especially when we are interested in more precise asymptotics.

[1]We write $f(n) = O(g(n))$ if and only if $|f(n)| \le C|g(n)|$ for some positive constant C and sufficiently large n. Also, $f(n) = \Theta(g(n))$ if and only if $f(n) = O(g(n))$ and $g(n) = O(f(n))$, $f(n) = o(g(n))$ if and only if $\lim_{n\to\infty} f(n)/g(n) = 0$, and $f(n) = \Omega(g(n))$ if and only if $g(n) = O(f(n))$.

[2]Note that the model families $\mathcal{M}_0$ and $\widetilde{\mathcal{M}}_0$ are defined over different alphabets. In addition, the family $\widetilde{\mathcal{M}}_0$ is constrained in that the probabilities of symbols in $\mathcal{B}$ take fixed values.

In this paper, we formalize this intuition by providing precise asymptotics of the minimax pointwise redundancy $R_n^*(\widetilde{\mathcal{M}}_0)$, again for all ranges of m (relative to n). We first prove that

$$R_n^*(\widetilde{\mathcal{M}}_0) = \log \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} 2^{R_k^*(\mathcal{M}_0)} \tag{3}$$

where $p = 1 - P(\mathcal{B})$. As it turns out, in order to estimate this quantity asymptotically, we need a quite precise understanding of the asymptotic behavior of $R_k^*(\mathcal{M}_0)$ for large k and m, as provided by Theorem 1.

The study of the minimax pointwise redundancy over $\mathcal{A} \cup \mathcal{B}$ expressed in (3) leads to an interesting problem for the so-called *binomial sums*, defined in general as

$$S_f(n) = \sum_{k} \binom{n}{k} p^k (1-p)^{n-k} f(k) \tag{4}$$

where $0 < p < 1$ is a fixed probability and f is a given function. In [6] and [11], asymptotics of $S_f(n)$ were derived for the *polynomially* growing function $f(x) = O(x^a)$. This result applies to our case when m is a constant, and leads to the conclusion that the asymptotics of $R_n^*(\widetilde{\mathcal{M}}_0)$ are the same as those of $R_{np}^*(\mathcal{M}_0)$, an intuitively appealing result since the length of the subsequence over $\mathcal{A}$ is np with high probability. But when m also grows, we encounter subexponential, exponential, and superexponential functions f, depending on the relation between m and n; therefore, we need more precise information about f to extract precise asymptotics of $S_f(n)$. In our second main result, Theorem 2, we use the asymptotics derived in Theorem 1 to deal with the binomial sum (3) and extract asymptotics of $R_n^*(\widetilde{\mathcal{M}}_0)$ for large n and m.

In the remainder of this paper, Section II reviews the analytic methods of analysis of algorithms that were used in [20] for estimating $R_n^*(\mathcal{M}_0)$ in the constant m case, as well as the saddle point method, whereas Section III presents our main results. These results are proved in Section IV.

II. Background

In the sequel, we will denote $d_{n,m} := R_n^*(\mathcal{M}_0)$ to emphasize the dependence of $R_n^*(\mathcal{M}_0)$ on both n and m. We will also denote $d_{n,m} := \log D_{n,m}$ which, by (2), implies

$$D_{n,m} = \sum_{x_1^n} \sup_{P \in \mathcal{M}_0} P(x_1^n). \tag{5}$$

Clearly, $D_{n,m}$ takes the form

$$D_{n,m} = \sum_{k_1+\cdots+k_m=n} \binom{n}{k_1, \ldots, k_m} \left(\frac{k_1}{n}\right)^{k_1} \cdots \left(\frac{k_m}{n}\right)^{k_m} \tag{6}$$

where k_i is the number of times symbol $i \in \mathcal{A}$ occurs in a string of length n.

The asymptotics of the sequence of numbers $\langle D_{n,m} \rangle$ (for m constant) are analyzed in [20] through its *tree-like generating function*, defined as

$$D_m(z) = \sum_{n=0}^{\infty} \frac{n^n}{n!} D_{n,m} z^n.$$

Here, we will follow the same methodology, which we review next. The first step is to use (6) to define an appropriate recurrence on $\langle D_{n,m}\rangle$ (involving both indices, n and m), and to employ the convolution formula for generating functions (cf., [21]) to relate $D_m(z)$ to the tree-like generating function of the sequence $\langle 1, 1, \ldots\rangle$, namely

$$B(z) = \sum_{k=0}^{\infty} \frac{k^k}{k!} z^k.$$

This function, in turn, can be shown to satisfy (cf., [21])

$$B(z) = \frac{1}{1 - T(z)} \tag{7}$$

for $|z| < e^{-1}$, where $T(z)$ is the well-known *tree function*, which is a solution to the implicit equation

$$T(z) = ze^{T(z)} \tag{8}$$

with $|T(z)| < 1$.[3] Specifically, the following relation is proved in [20].

Lemma 1: The tree-like generating function $D_m(z)$ of $\langle D_{n,m}\rangle$ satisfies, for $|z| < e^{-1}$,

$$D_m(z) = [B(z)]^m - 1$$

and, consequently

$$D_{n,m} = \frac{n!}{n^n}[z^n]\,[B(z)]^m \tag{9}$$

where $[z^n]f(z)$ denotes the coefficient of z^n in $f(z)$.

Defining $\beta(z) = B(z/e)$, $|z| < 1$, noticing that $[z^n]\beta(z) = e^{-n}[z^n]B(z)$, and applying Stirling's formula, (9) yields

$$D_{n,m} = \sqrt{2\pi n}\left(1 + O(n^{-1})\right)[z^n]\,[\beta(z)]^m. \tag{10}$$

Thus, it suffices to extract asymptotics of the coefficient at z^n of $[\beta(z)]^m$, for which a standard tool is Cauchy's coefficient formula [8], [21], that is

$$[z^n][\beta(z)]^m = \frac{1}{2\pi i}\oint \frac{\beta^m(z)}{z^{n+1}}dz \tag{11}$$

where the integration is around a closed path containing $z = 0$ inside which $\beta^m(z)$ is analytic.

Now, the *constant* m case is solved in [20] by use of the Flajolet and Odlyzko singularity analysis [8], [21], which applies because $[\beta(z)]^m$ has algebraic singularities. Indeed, using (7) and (8), the singular expansion of $\beta(z)$ around its singularity $z = 1$ takes the form [3]

$$\beta(z) = \frac{1}{\sqrt{2(1-z)}} + \frac{1}{3} - \frac{\sqrt{2}}{24}\sqrt{(1-z)} + O(1-z).$$

The singularity analysis then yields [20]

$$d_{n,m} = \frac{m-1}{2}\log\left(\frac{n}{2}\right) + \log\left(\frac{\sqrt{\pi}}{\Gamma(\frac{m}{2})}\right) + \frac{\Gamma(\frac{m}{2})m\log e}{3\Gamma(\frac{m}{2}-\frac{1}{2})}\cdot\frac{\sqrt{2}}{\sqrt{n}} + O\left(\frac{1}{n}\right) \tag{12}$$

for large n and constant m, where Γ is the Euler gamma function.[4]

When m also grows, which is the case of interest in this paper, the singularity analysis *does not* apply. Instead, the growth of the factor $\beta^m(z)$ determines that the saddle point method [8], [21], which we briefly review next, can be applied to (11). We will restrict our attention to a special case of the method, where the goal is to obtain an asymptotic approximation of the coefficient $a_n := [z^n]g(z)$ for some *analytic* function $g(z)$, namely

$$a_n = \frac{1}{2\pi i}\oint \frac{g(z)}{z^{n+1}}dz = \frac{1}{2\pi i}\oint e^{h(z)}dz$$

where $h(z) := \ln g(z) - (n+1)\ln z$, under the assumption that $h'(z)$ has a *real* root z_0.

The saddle point method is based on Taylor's expansion of $h(z)$ around z_0 which, recalling that $h'(z_0) = 0$, yields

$$h(z) = h(z_0) + \frac{1}{2}(z-z_0)^2 h''(z_0) + O(h'''(z_0)(z-z_0)^3). \tag{13}$$

After choosing a path of integration that goes through z_0, and under certain assumptions on the function $h(z)$, it can be shown (cf., e.g., [21]) that the first term of (13) gives a factor $e^{h(z_0)}$ in a_n, the second term—after integrating a Gaussian integral—leads to a factor $1/\sqrt{2\pi|h''(z_0)|}$, and finally the third term determines the error term in the expansion of a_n. The standard saddle point method described in [21, Table 8.4] then yields the following lemma.

Lemma 2: Assume that the conditions required in [21, Table 8.4] hold and let z_0 denote a real root of $h'(z)$. Then

$$a_n = \frac{e^{h(z_0)}}{\sqrt{2\pi|h''(z_0)|}} \times \left(1 + O\left(\frac{h'''(z_0)}{(h''(z_0))^{\rho}}\right)\right) \tag{14}$$

for any constant $\rho < 3/2$, provided the error term is $o(1)$.[5]

In order to control the error term, the conditions stated in [21, Table 8.4] include the requirement that, as n grows, $h''(z_0) \to \infty$. It turns out, however, that more is known for our particular $h(z)$: indeed, it will be further shown that the growth of $h''(z_0)$ is at least linear. This additional property allows us to extend Lemma 2 to the case $\rho = 3/2$. The modified lemma will be the main tool in our derivation.

III. Main Results

In this section, we present and discuss our main results, deferring their proof to Section IV.

A. Model Family $\mathcal{M}_0$

Theorem 1: For the memoryless model family $\mathcal{M}_0$ over an m-ary alphabet, where $m \to \infty$ as n grows, the minimax pointwise redundancy $d_{n,m}$ behaves asymptotically as follows.

[3] In terms of the standard *Lambert-W* function, we have $T(z) = -W(-z)$.

[4] As mentioned, (2) ignores the integer length constraint of a code, and therefore, $O(1)$ terms in (12) are arguably irrelevant. This issue is addressed in [5]; here, we focus on the probability assignment problem, which unlike coding does not entail an integer length constraint.

[5] This expression for the error term in (14) is obtained with the choice $\delta(n) = h''(z_0)^{-\rho/3}$ in [21, Table 8.4], provided certain conditions on $h(z)$ are satisfied.

i) For $m = o(n)$

$$d_{n,m} = \frac{m-1}{2}\log\frac{n}{m} + \frac{m}{2}\log e + \frac{m\log e}{3}\sqrt{\frac{m}{n}} - \frac{1}{2} - \frac{\log e}{4}\sqrt{\frac{m}{n}} + O\left(\frac{m^2}{n} + \frac{1}{\sqrt{m}}\right). \quad (15)$$

ii) For $m = \alpha n + \ell(n)$, where α is a positive constant and $\ell(n) = o(n)$

$$d_{n,m} = n\log B_\alpha + \ell(n)\log C_\alpha - \log\sqrt{A_\alpha} - \frac{\ell(n)^2\log e}{2n\alpha^2 A_\alpha} + O\left(\frac{\ell(n)^3}{n^2} + \frac{\ell(n)}{n} + \frac{1}{\sqrt{n}}\right) \quad (16)$$

where

$$C_\alpha := \frac{1}{2} + \frac{1}{2}\sqrt{1+\frac{4}{\alpha}} \quad (17)$$

$$A_\alpha := C_\alpha + \frac{2}{\alpha} \quad (18)$$

and

$$B_\alpha := \alpha C_\alpha^{\alpha+2} e^{-\frac{1}{C_\alpha}}. \quad (19)$$

iii) For $n = o(m)$

$$d_{n,m} = n\log\frac{m}{n} + \frac{3}{2}\frac{n(n-1)}{m}\log e + O\left(\frac{1}{\sqrt{n}} + \frac{n^3}{m^2}\right). \quad (20)$$

B. Discussion of Theorem 1

Significance and Related Work: The formulation of the scenario in which both n and m are large, as a sequence of problems where m varies with n, follows Orlitsky and Santhanam [15]. In a typical application of Theorem 1, for a *given* pair of values $n = n_0$ and $m = m_0$, which are deemed to fall in one of the three itemized cases, the formulas are used to approximate the minimax pointwise redundancy d_{n_0,m_0}. The leading terms of the asymptotic expansions for $m = o(n)$ and $n = o(m)$ (i.e., (15) and (20)) were derived in [15]. The asymptotic expansion in (15) reveals that the error incurred by neglecting lower order terms may be significant. Consider the example in which $n = 10^4$ and $m = 40$ (or, approximately, $m = n^{0.4}$). Then, the leading term in (15) is only 5.5 times larger than the second term, and 131 times larger than the third term. The error from neglecting these two terms is thus 15.4% (assuming that all other terms are negligible). Even for $n = 10^8$ (and $m = 1600$), the error is still over 8%. It is interesting to notice that (15) is a "direct scaling" of (12): using Stirling's approximation to replace $\Gamma(x)$ in (12) by its asymptotic value $\sqrt{2\pi/x}(x/e)^x$, and further approximating $(1+1/x)^{(x+1)/2}$ with $\sqrt{e}\,(1+1/(4x))$, indeed yields exactly (15), up to the error terms. Thus, our results reveal that the first two terms of the asymptotic expansion for fixed m given by (12) are in fact a better approximation to $d_{n,m}$ than the leading term of (15).

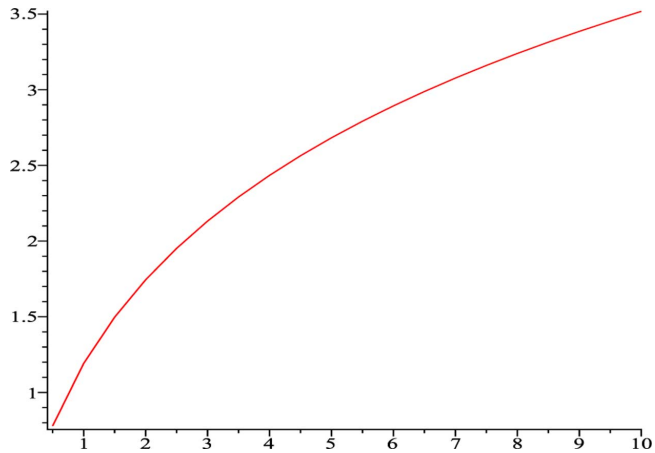

Fig. 1. Value of the constant $\log B_\alpha$ in the $\Theta(n)$ term of $d_{n,m}$ in case $m = \Theta(n)$.

For the case $m = \Theta(n)$, the methodology of [15] allowed only the extraction of the growth rate, i.e., $d_{n,m} = \Theta(n)$, but not the constant in front of n. The value of this constant, $\log B_\alpha$, where B_α is specified in (19) and (17), is plotted against α in Fig. 1. It is easy to see that when $\alpha \to 0$, $\log B_\alpha \approx (\alpha/2)\log(1/\alpha)$, in agreement with (15). Similarly, when $\alpha \to \infty$, $\log B_\alpha \approx \log\alpha$, in agreement with (20).

Finally, for the case $n = o(m)$, our results confirm that the leading term is a good approximation to $d_{n,m}$. The intuition behind this term is that, for large m, the value of the minimax game is achieved when all the symbols in x_1^n are roughly different (so that the maximum-likelihood probability of each occurring symbol tends to $1/n$) and the code assigns $\log m$ bits to each symbol, leading to a pointwise redundancy of, roughly, $n\log(m/n)$.

Convergence: Observe that the second-order term in (15), which is $\Theta(m)$, dominates $-\log(n/m)$ whenever $m = \Omega(n^a)$ for some a, $0 < a < 1$. Hence, the leading term in the expansion is rather $(m/2)\log(n/m)$ than $(m-1)/2\log(n/m)$. In the numerical example given for this case, the choice of a growth rate $m = o(\sqrt{n})$ is due to the fact that, otherwise, the error term $O(m^2/n)$ may not even vanish, and it may dominate the constant, as well as the $\sqrt{m/n}$ terms. For any given growth rate $m = O(n^a)$, $0 < a < 1$, an expansion in which the error term vanishes can be derived; however, no expansion has this property for *every* possible value of a. The reason is that, as will become apparent in the proof of the theorem, any expansion will include an error term of the form $O(m(m/n)^{j/2})$ for some positive integer j. The same situation can be observed in (20), where one of the error terms becomes $O(n(n/m)^j)$ if a more accurate expansion is used.

A similar phenomenon is observed for the error term in (16), which is guaranteed to vanish only if $\ell(n) = o(n^{2/3})$, and it can otherwise dominate the constant term in the expansion. Again, for any given growth rate $\ell(n) = O(n^a)$, an expansion in which the error term vanishes can be derived. Notice, however, that the case $\ell(n) \neq 0$ is analyzed only for completeness since, as mentioned, a typical application of (16) would in general involve approximating d_{n_0,m_0}, for a given pair of values n_0, m_0 which are deemed to fall in Case (ii), by using (16) with $\alpha = n_0/m_0$ and $\ell(n) = 0$.

C. Model Family $\widetilde{\mathcal{M}}_0$

In this section we consider the second main topic of this paper, namely, the minimax pointwise redundancy $R_n^*(\widetilde{\mathcal{M}}_0)$ relative to the family $\widetilde{\mathcal{M}}_0$ of constrained (i.e., some parameters are fixed) memoryless models. Recall that the model family $\widetilde{\mathcal{M}}_0$ assumes an alphabet $\mathcal{A}\cup\mathcal{B}$, where $|\mathcal{A}|=m$ and $|\mathcal{B}|=M$. The probabilities of symbols in $\mathcal{A}$, denoted by $p_1,\ldots,p_m$, are allowed to vary (unknown), while the probabilities $q_1,\ldots,q_M$ of the symbols in $\mathcal{B}$ are fixed (known). Furthermore, $q=q_1+\cdots+q_M$ and $p=1-q$. We assume that $0<q<1$ is fixed (independent of the sequence length n). To simplify our notation, we also write $\mathbf{p}=(p_1,\ldots,p_m)$ and $\mathbf{q}=(q_1,\ldots,q_M)$. The output sequence is denoted $x:=x_1^n\in(\mathcal{A}\cup\mathcal{B})^n$.

Our goal is to derive asymptotics of $R_n^*(\widetilde{\mathcal{M}}_0):=d_{n,m,M}$ for large n and m, where again we introduce notation that emphasizes the dependence on m (the dependence on M will be shown to be indirect, via p, and does not affect the analysis). First, Lemma 3 relates $d_{n,m,M}$ to the minimax pointwise redundancy $d_{n,m}$ relative to $\mathcal{M}_0$, studied in Theorem 1, and to p. The lemma is stated in terms of $D_{n,m,M}:=2^{d_{n,m,M}}$ and $D_{n,m}=2^{d_{n,m}}$.

Lemma 3:

$$D_{n,m,M}=\sum_{k=0}^{n}\binom{n}{k}p^k(1-p)^{n-k}D_{k,m}\,.$$

Proof: Let $P\in\widetilde{\mathcal{M}}_0$. By (2), we have

$$D_{n,m,M}=\sum_{x\in(\mathcal{A}\cup\mathcal{B})^n}\sup_{\mathbf{p}}P(x)=\sum_{x\in(\mathcal{A}\cup\mathcal{B})^n}\widetilde{P}_n(x)\qquad(21)$$

where $\widetilde{P}_n(x)=\sup_{\mathbf{p}}P(x)$ is the maximum-likelihood (ML) estimator of $P(x)$ over $\widetilde{\mathcal{M}}_0$. To simplify (21), consider $x\in(\mathcal{A}\cup\mathcal{B})^n$ and assume that i symbols are from $\mathcal{B}$ and the remaining $n-i$ symbols are from $\mathcal{A}$. We denote by $z\in\mathcal{B}^i$ the subsequence of x consisting of i symbols from $\mathcal{B}$. Similarly, $y\in\mathcal{A}^{n-i}$ is the subsequence of x over $\mathcal{A}$. For any such pair (y,z), there are $\binom{n}{i}$ ways of interleaving the sub-sequences, all leading to the same ML probability $\widetilde{P}_n(x)$. Now, it is easy to see that $\widetilde{P}_n(x)$ takes the form

$$\widetilde{P}_n(x)=p^{n-i}\hat{P}_{n-i}(y)q^iP_i(z)$$

where $\hat{P}_{n-i}(y)$ is the ML probability of y (over the set $\mathcal{M}_0$ of memoryless sources over $\mathcal{A}$), and $P_i(z)$ is the probability of z over $\mathcal{B}$ with (given) probabilities $q_1/q,\ldots,q_M/q$. In summary, using (21), we obtain

$$\begin{aligned}D_{n,m,M}&=\sum_{i=0}^{n}\binom{n}{i}p^{n-i}q^i\sum_{y\in\mathcal{A}^{n-i}}\sum_{z\in\mathcal{B}^i}\hat{P}_{n-i}(y)P_i(z)\\&=\sum_{i=0}^{n}\binom{n}{i}p^{n-i}q^i\sum_{y\in\mathcal{A}^{n-i}}\hat{P}_{n-i}(y).\end{aligned}\qquad(22)$$

The proof is complete by noticing that the inner summation in (22) is precisely $D_{n-i,m}$. □

By Lemma 3, the robust asymptotic expression of $D_{n,m}$ derived in Theorem 1 will be our starting point for estimating $D_{n,m,M}$.[6] As mentioned, the generic form of the sum in the lemma, given in (4), is known as the binomial sum [6], [11]. If $D_{k,m}$ has a polynomial growth (i.e., $D_{k,m}=2^{d_{k,m}}=O(k^{(m-1)/2})$ when m is a constant), then we can use the asymptotic expansion derived in [6] and [11] to conclude that $D_{n,m,M}\sim D_{np,m}$. However, when m varies with n as in our study, the aforementioned expansion does not apply and we need to compute asymptotics anew. We state and discuss our second main result in Theorem 2, whose proof is presented in Section IV.

Theorem 2: Consider a family of memoryless models $\widetilde{\mathcal{M}}_0$ over the $(m+M)$-ary alphabet $\mathcal{A}\cup\mathcal{B}$, with fixed probabilities $q_1,\ldots,q_M$ of the symbols in $\mathcal{B}$, such that $q=q_1+\ldots+q_M$ is bounded away from 0 and 1. Let $p=1-q$. Then, the minimax pointwise redundancy $d_{n,m,M}$ takes the following form.

i$_0$) If m is constant, then

$$d_{n,m,M}=\frac{m-1}{2}\log\left(\frac{np}{2}\right)+\log\left(\frac{\sqrt{\pi}}{\Gamma(\frac{m}{2})}\right)+O\left(\frac{1}{\sqrt{n}}\right).\qquad(23)$$

i) Let $m\to\infty$ as n grows, with $m=o(n)$.
If $m=O(n^a)$ with $a<1/2$, then

$$\begin{aligned}d_{n,m,M}&=\frac{m-1}{2}\log\frac{np}{m}+\frac{m}{2}\log e-\frac{1}{2}\\&\quad+\frac{m}{3}\log e\sqrt{\frac{m}{np}}+O\left(\frac{1}{\sqrt{m}}+\frac{m^2}{n}\right).\end{aligned}\qquad(24)$$

Otherwise

$$\begin{aligned}d_{n,m,M}&=\frac{m}{2}\log\frac{np}{m}+\frac{m}{2}\log e+\frac{m}{3}\log e\sqrt{\frac{m}{np}}\\&\quad+O\left(\log n+\frac{m^2}{n}\right).\end{aligned}\qquad(25)$$

ii) If $m=\Theta(n)$, then

$$d_{n,m,M}=nK+o(n)\qquad(26)$$

where $\log(B_\alpha p)\le K\le\log B_\alpha$, $\alpha=\frac{m}{n}$, and B_α is defined in Theorem 1(ii).

iii) If $n=o(m)$, then

$$d_{n,m,M}=n\log\frac{m}{n}+O(n).\qquad(27)$$

D. Discussion of Theorem 2

Asymptotics: By Lemma 3, $d_{n,m,M}$ depends on $\mathcal{B}$ only through p, and it is given by the logarithm of a binomial sum, which for a generic function f takes the form (4) (in our case, $f(k)=D_{k,m}$, where m may grow with n). Intuitively, when f grows polynomially in k, the maximum under the sum occurs around $k=np$, to find asymptotics we need to sum only within the range $\pm\sqrt{n}$ around np, and $d_{n,m,M}$ behaves roughly as

[6]Notice, however, that some extra care will be needed in the application of Theorem 1 since, in the generic term $D_{k,m}$ in the sum, m grows with n, not with k.

$d_{np,m}$. This is indeed the case when m is a constant. While, in Case (i), the growth of f is not polynomial, it is still subexponential, and it is possible to extend the aforementioned intuition to obtain the asymptotic expansion. When $m = \Theta(n)$, however, the growth of $f(k)=D_{k,m}$ is exponential, and we need all the terms in the sum in order to extract the asymptotics. As a result, even the (bounded) factor K in front of the main asymptotic term of $d_{n,m,M}$ in (26) may differ from that in $d_{np,m}$, given by $p \log B_\nu$, where $\nu=\alpha/p$. The precise behavior of this factor remains an open question: the difficulty in its determination stems from the fact that, in this case, $D_{k,m}=O(A(k)^k)$, where $A(k)$ is *not* a constant. The dependence of $A(k)$ on k is due to the fact that this case assumes a constant ratio between m and n, and not between m and k. Finally, for $n = o(m)$, the function $f(k)$ grows superexponentially, and the asymptotics of the binomial sum are determined by the last term, that is, $k = n$. In this case, the main asymptotic terms of $d_{n,m,M}$ and $d_{np,m}$ coincide.

It is interesting to notice that the $O(\log n)$ term in (25) is the dominating error term only when $m = \Omega(n^a)$ for all $a < 1/2$ but $m = O(\sqrt{n \log n})$. It is an open question whether this term can be avoided using a different proof technique.

Alternative Model: As mentioned, a natural setup for the asymptotic analysis of $d_{n,m,M}$ is one in which m may grow with n. An alternative model (not motivated by any specific setting) is one in which, in the analysis of the binomial sum for $D_{k,m}$, the parameter m grows with k, which enables a more direct application of Theorem 1. As will be discussed in Section IV, this alternative model leads to a more precise expansion in Cases (ii) and (iii).

IV. Proofs of Main Theorems

In this section, we prove Theorem 1 using analytic tools and Theorem 2 using elementary analysis.

A. Proof of Theorem 1

The starting point is (10) which, as noted, follows from Lemma 1 and Stirling's formula, and Cauchy's coefficient formula (11), which takes the form

$$[z^n][\beta(z)]^m = \frac{1}{2\pi i} \oint e^{h(z)} dz \tag{28}$$

where

$$h(z) = m \ln \beta(z) - (n+1) \ln z. \tag{29}$$

We will apply a modification of Lemma 2 in the evaluation of (28), for which we need to check that the necessary conditions are satisfied by the function $h(z)$ of (29).

We first find an explicit real root z_0 of the saddle point equation $h'(z) = 0$, and show that it is unique in the interval $[0, 1)$. Differentiating (29), we have

$$z_0 \frac{\beta'(z_0)}{\beta(z_0)} = \frac{n+1}{m}. \tag{30}$$

Differentiating (8) and using (7), it is easy to see that

$$z \frac{\beta'(z)}{\beta(z)} = \beta(z)^2 - \beta(z). \tag{31}$$

Thus, (30) takes the form

$$\beta(z_0)^2 - \beta(z_0) = \frac{n+1}{m}. \tag{32}$$

By (7) and the definition of $T(z)$, the range of $\beta(z)$ for $0 \le z < 1$ is $[1, +\infty)$. Since the quadratic equation (32) has a unique real root in this range, we have

$$\beta(z_0) = \frac{1}{2} + \frac{1}{2}\sqrt{1 + \frac{4(n+1)}{m}} := \frac{1}{\gamma_{n,m}} \tag{33}$$

and the uniqueness of a real root z_0 in $[0, 1)$ follows from the fact that $\beta(z)$ is increasing in this interval. Moreover, by (7), (33) takes the form

$$T\left(\frac{z_0}{e}\right) = 1 - \gamma_{n,m}.$$

Hence, by (8), we finally obtain the explicit expression

$$z_0 = (1 - \gamma_{n,m}) e^{\gamma_{n,m}} \tag{34}$$

where, since

$$\gamma_{n,m} = \frac{m}{2(n+1)} \left(\sqrt{1 + \frac{4(n+1)}{m}} - 1 \right) \tag{35}$$

we have $0 < \gamma_{n,m} < 1$ and also $0 < z_0 < 1$. We then see that, by (29), (33), and (34), $h(z_0)$ takes the form

$$h(z_0) = -m \ln \gamma_{n,m} - (n+1)[\ln(1 - \gamma_{n,m}) + \gamma_{n,m}]. \tag{36}$$

In addition, differentiating (29) twice, we obtain

$$h''(z_0) = mA(z_0) + \frac{n+1}{z_0^2}$$

where

$$A(z) = \frac{d}{dz}\left[\frac{\beta'(z)}{\beta(z)}\right] = \frac{[\beta(z)^2 - \beta(z)]\,[2\beta(z)^2 - \beta(z) - 1]}{z^2} \tag{37}$$

with the second equality in (37) easily seen to follow from further differentiating (31). Thus, using (32)

$$h''(z_0) = \frac{n+1}{z_0^2}\left[\frac{2(n+1)}{m} + \beta(z_0)\right]$$

which, again by (33) and (34), can be expressed in terms of $\gamma_{n,m}$ as

$$h''(z_0) = \frac{n+1}{(1 - \gamma_{n,m})^2 e^{2\gamma_{n,m}}}\left[\frac{2(n+1)}{m} + \frac{1}{\gamma_{n,m}}\right]. \tag{38}$$

Finally, taking another derivative in (37) and further using (31) and (32), after some additional computations, we obtain

$$h'''(z_0) = \frac{n+1}{\gamma_{n,m} z_0^3}\left[\frac{n+1}{m}\left(\frac{8}{\gamma_{n,m}} - 1\right) - \frac{5}{\gamma_{n,m}} + 3\right]. \tag{39}$$

With these expressions on hand, we can now check the conditions required in Lemma 2 for the evaluation of (28). The most intricate condition to be checked is that of "tail eliminations" (denoted (SP3) in [21, Table 8.4, (8.105)]). This condition is actually shown in [7, Lemma 5] to hold in more general cases than the function $h(z)$ of (29). Also, proceeding along the lines of the proof of [21, Theorem 8.17], it can be shown that (14) of Lemma 2 holds with $\rho = 3/2$ if $h''(z_0)$ grows at least linearly and if $h'''(z_0) = o((h''(z_0))^{3/2})$. Thus, (10) and the modified Lemma 2 yield

$$d_{n,m} = h(z_0)\log e - \log\sqrt{\frac{h''(z_0)}{n}} + O\left(\frac{h'''(z_0)}{(h''(z_0))^{3/2}} + \frac{1}{n}\right) \tag{40}$$

provided that the error term is $o(1)$ and $h''(z_0)$ grows at least linearly. Consequently, to complete the proof of Theorem 1, we need to evaluate the right-hand side of (40). In view of (36) and (38), which give $h(z_0)$ and $h''(z_0)$ as functions of $\gamma_{n,m}$, the solution depends on the possible growth rates of m. We analyze next all possible cases.

CASE: $m = o(n)$

Letting $m/n \to 0$ in (35), it is easy to see that

$$\gamma_{n,m} = \sqrt{\frac{m}{n}}\left(1 - \frac{1}{2}\sqrt{\frac{m}{n}}\right) + O\left(\frac{m^{3/2}}{n^{3/2}}\right).$$

Substituting into (36) and (38), we obtain

$$h(z_0) = \frac{m}{2}\ln\frac{n}{m} + \frac{m}{2} + \frac{m}{3}\sqrt{\frac{m}{n}} + O\left(\frac{m^2}{n}\right)$$

and

$$\ln\frac{h''(z_0)}{n} = \ln\frac{n}{m} + \ln 2 + \frac{1}{2}\sqrt{\frac{m}{n}} + O\left(\frac{m}{n}\right). \tag{41}$$

From (39), and noticing that, in this case, (34) yields $z_0 \to 1$, we further obtain

$$h'''(z_0) = \Theta\left(\frac{n^3}{m^2}\right). \tag{42}$$

Theorem 1(i) follows from substituting these equations into (40), observing that (41) and (42) guarantee that the necessary conditions for the modified Lemma 2 to hold for $h(z)$ are satisfied.[7]

CASE: $m = \Theta(n)$

Since z_0 is given by (34) where, in this case, $m = \alpha n + \ell(n)$ and $\ell(n) = o(n)$, we can view z_0 as a function of $m/(n+1)$, which we expand around α. The value of this function at α is

$$z_\alpha = (1 - C_\alpha^{-1})e^{1/C_\alpha} = \alpha^{-1}C_\alpha^{-2}e^{1/C_\alpha}$$

where C_α is given by (17). It is is then easy to see that

$$z_0 = z_\alpha - z_\alpha\alpha^{-1}A_\alpha^{-1}\delta(n) + O(\delta(n)^2)$$

where $\delta(n) := (\ell(n) - \alpha)/(n+1) = o(1)$ and A_α is given by (18). With this value of z_0, we can then compute, with a Taylor expansion around z_α

$$\begin{aligned} h(z_0) &= n\ln(C_\alpha^\alpha z_\alpha^{-1}) + \ell(n)\ln C_\alpha \\ &\quad - \ln z_\alpha - n\delta(n)^2\frac{1}{2\alpha^2 A_\alpha} + O(n\delta(n)^3) \\ \ln\frac{h''(z_0)}{n} &= \ln(A_\alpha z_\alpha^{-2}) + O(\delta(n)) \\ h'''(z_0) &= O(n). \end{aligned}$$

Substitution into (40) completes the proof of Theorem 1(ii), after observing, again, that the necessary conditions for the modified Lemma 2 hold.

CASE: $m = o(m)$

Letting $n/m \to 0$ in (35), it is easy to see that

$$\gamma_{n,m} = 1 - \frac{n+1}{m} + \frac{2(n+1)^2}{m^2} + O\left(\frac{n^3}{m^3}\right).$$

Substituting into (36) and (38), we obtain

$$h(z_0) = (n+1)\ln\frac{m}{n+1} + \frac{3}{2}\frac{(n+1)^2}{m} + O\left(\frac{n^3}{m^2}\right)$$

and

$$\ln\frac{h''(z_0)}{n+1} = 2\ln\frac{m}{(n+1)e} + 9\frac{n+1}{m} + O\left(\frac{n^2}{m^2}\right).$$

From (39), and noticing that, in this case, (34) yields $z_0 = \Theta(1 - \gamma_{n,m}) = \Theta(n/m)$, we further obtain

$$h'''(z_0) = \Theta\left(\frac{m^3}{n^2}\right).$$

Putting everything together, substituting into (40), and observing that the necessary conditions for the modified Lemma 2 hold, we prove Theorem 1(iii).[8] □

B. Proof of Theorem 2

By Lemma 3, in order to prove Theorem 2 we need to evaluate the binomial sum

$$S_f(n) = \sum_k \binom{n}{k} p^k (1-p)^{n-k} f(k) \tag{43}$$

for $f(k) = D_{k,m}$ that, for $m \to \infty$, grows faster than any polynomial. We observe that

$$S_f(n) = \mathbf{E}_X[f(X)]$$

where $\mathbf{E}_X$ denotes expectation with respect to a binomially distributed random variable X. Since $D_{k,m}$ is nondecreasing in k

[7]Taking more terms in the expansion of $\gamma_{n,m}$, an $O(m(m/n)^{j/2})$ error term for $h(z_0)$ can be obtained, where j is as large as desired. Thus, while no value of j guarantees a vanishing error for every m, for each given $m = O(n^\alpha)$, a choice of j exists that guarantees $o(1)$ error.

[8]We can take more terms in the expansion of $\gamma_{n,m}$ also in this case, leading to an $O(n(n/m)^j)$ error term for $h(z_0)$.

(notice that m depends on n, and not on k), the function is maximum at $k = n$. Therefore

$$p^n D_{n,m} \le S_f(n) \le D_{n,m}$$

where the lower bound follows from taking only the last term in the summation. Thus, Cases (ii) and (iii) follow from taking logarithms and applying Theorem 1, Cases (ii) and (iii), respectively.

For Cases (i_0) and (i), we need a more accurate evaluation technique, which will rely on the concentration of X around its mean np. To this end, we break the summation (43) into three parts. Let $r > 0$ denote an arbitrary constant such that $r < p$, and consider a function $g_0(n)$, to be specified later, such that $g_0(n) > np$. We consider a first partial sum restricted to the first nr terms, a second partial sum from $k = nr$ to $g_0(n)$, and a third partial sum given by the remaining terms, that is,

$$S_f^{(1)}(n) := \sum_{k<nr} \tilde{f}(k), \quad S_f^{(2)}(n) := \sum_{k=nr}^{g_0(n)} \tilde{f}(k)$$
$$S_f^{(3)}(n) := \sum_{k>g_0(n)} \tilde{f}(k)$$

where

$$\tilde{f}(k) := \binom{n}{k} p^k (1-p)^{n-k} f(k) = \Pr\{X{=}k\} f(k) \quad (44)$$

so that

$$S_f(n) = S_f^{(1)}(n) + S_f^{(2)}(n) + S_f^{(3)}(n). \quad (45)$$

Lemma 4:

$$S_f^{(1)}(n) = O\left(e^{-\frac{1}{2}n(p-r)^2} f(nr)\right).$$

Proof: Since $D_{k,m}$ ia a nondecreasing functions of k, we have

$$S_f^{(1)}(n) \le \Pr\{X < rn\} f(rn). \quad (46)$$

The lemma then follows from Hoeffding's inequality [10], which states that

$$\Pr\{X < nr\} \le e^{-\frac{1}{2}n(p-r)^2}. \quad \square$$

To estimate $S_f^{(2)}(n)$, the key idea is to apply Taylor's theorem to $f(x)$ (the extension of $f(n)$ to the real line) around the mean $x = np$, and estimate $f''(x)$ at a point close to np. First, we notice that, in the relevant region, $m = o(k)$ and, therefore, in Case (i), $f(k)$ is well approximated using the asymptotic expansion (15) (this would not necessarily be the case for $k = o(n)$). Second, we notice that the behavior of the derivatives of $f(x)$ could, in principle, be dominated by the error terms in (12) (Case (i_0)) and (15) (Case (i)). To deal with this situation, we define a new function, $f_1(k)$, which differs from $f(k)$ in that it does not include error terms, namely, in Case (i_0)

$$f_1(k) = C(m) k^{\frac{m-1}{2}}$$

where $C(m)$ is a constant that depends on m (see (12)), whereas in Case (i), and further assuming $m = o(\sqrt{n})$

$$f_1(k) = \left(\frac{ke}{m}\right)^{\frac{m}{2}} \sqrt{\frac{m}{2k}}\, e^{\frac{m}{3}\sqrt{\frac{m}{k}}} \quad (47)$$

where we note that, in this subcase, the error term in (15) dominates the $O(\sqrt{m/n})$ term. Next, we approximate $S_f^{(2)}(n)$ with $S_{f_1}^{(2)}(n)$. To this end, we let $f_2(k)$ denote the (vanishing) error terms given by (12) in Case (i_0), and (15) in Case (i) (specifically, $f_2(k) = 1/\sqrt{k}$ in Case (i_0) and $f_2(k) = (m^2/k) + 1/\sqrt{m}$ in Case (i)).

Lemma 5:

$$S_f^{(2)}(n) = S_{f_1}^{(2)}(n)(1 + O(f_2)).$$

Proof: Writing $f = f_1 \times (f/f_1)$, we obtain

$$S_f^{(2)}(n) \le S_{f_1}^{(2)}(n) \max_{rn \le k \le n} [f(k)/f_1(k)]. \quad (48)$$

By the definition of f_1, for $k = \Theta(n)$, we have $\log f(k) - \log f_1(k) = O(f_2(k))$ which, since $f_2(k) = o(1)$, implies that $f(k) = f_1(k)(1 + O(f_2(k))$. Thus, since $f_2(k)$ is decreasing for $k > rn$ and sufficiently large n

$$\max_{rn \le k \le n} [f(k) - f_1(k)]/f_1(k) = O(f_2(rn)).$$

The lemma then follows from (48), observing that $f_2(rn) = O(f_2(n))$. $\square$

Next, we estimate $S_{f_1}^{(2)}(n)$ for $m = O(n^a)$, with $a < 1/2$, by applying Taylor's theorem to $f_1(x)$ around $x = np$, which yields

$$f_1(x) = f_1(np) + (x - np) f_1'(np) + \frac{(x-np)^2}{2} f_1''(x') \quad (49)$$

for some x' that lies between x and np. Noting that, for the binomially distributed random variable X, $\mathbf{E}_X[X] = np$ and the variance is $\mathrm{Var}_X[X] = npq$, and that, for $k > rn$ and n sufficiently large, $f_1''(k)$ is positive and increasing, (49) implies

$$S_{f_1}^{(2)}(n) = f_1(np) + f_1'(np)\left(S_{np-x}^{(1)}(n) - S_{x-np}^{(3)}(n)\right) + A \quad (50)$$

where

$$0 \le A \le \frac{npq}{2} f_1''(g_0(n)).$$

Proceeding as in the proof of Lemma 4, we obtain

$$0 \le S_{np-x}^{(1)}(n) \le npe^{-\frac{1}{2}n(p-r)^2}$$

and it is easy to see that

$$\frac{n f_1'(np)}{f_1(np)} = O(m).$$

To estimate A, it is also easy to see that

$$\frac{nf_1''(g_0(n))}{f_1(np)} = O\left(\frac{m^2 f_1(g_0(n))}{nf_1(np)}\right). \tag{51}$$

In Case (i_0), we can simply choose $g_0(n)=n$ so that the left-hand side of (51) is $O(1/n)$. Since the third region collapses, dividing (45) by $f_1(np)$, Lemma 4 (where we notice that $f(nr) = O(f_1(np))$, Lemma 5, and (50) yield, after taking logarithms

$$\log S_f(n) = \log f_1(np) + O(f_2 + 1/n).$$

Theorem 2 (i_0) then follows from (12) and the definitions of f_1 and f_2. A more precise asymptotic expansion can be found using tools from [6] and [11].

The analysis is less straightforward in Case (i) (where we recall that, so far, we are assuming $a < 1/2$) because, since $f_1(n)/f_1(np) = \Theta((1/p)^{\frac{m}{2}})$ and $m \to \infty$, $m^2/(np^{m/2})$ does not vanish unless $m = o(\log n)$. Here, denote $g_0(n) := np(1+\epsilon_0(n))$, where $\epsilon_0(n) > 0$. Thus

$$\frac{nf_1''(g_0(n))}{f_1(np)} = O\left(\frac{m^2(1+\epsilon_0(n))^{m/2}}{n}\right) \tag{52}$$

and, choosing $\epsilon_0(n)=n^{-a}=O(1/m)$, the left-hand side of (52) is $O(m^2/n)$. In addition, with this choice, the term $S^{(3)}_{x-np}(n)$ in (50) can be bounded again as in the proof of Lemma 4 and is therefore $O(\exp\{-n^{(1-2a)}p^2/2\})$, which is dominated by the $O(m^2/n)$ term. Consequently, (50) takes the form

$$\frac{S^{(2)}_{f_1}(n)}{f_1(np)} = 1 + O\left(\frac{m^2}{n}\right). \tag{53}$$

Finally, we need to consider the third partial sum on the right-hand side of (45) for $a < 1/2$. To this end, in addition to the function $g_0(n)$, we choose a sequence of functions (to be specified later) such that $g_0(n)<g_1(n)<\cdots<g_j(n)=n$, $j\geq 1$. Since f is nondecreasing, we can upper-bound $S^{(3)}_f(n)$ by summing over segments of the form $(g_i(n), g_{i+1}(n)]$ to obtain

$$\frac{S^{(3)}_f(n)}{f_1(np)} < \sum_{i=0}^{j-1} \Pr\{X > g_i(n)\}\frac{f(g_{i+1}(n))}{f_1(np)}. \tag{54}$$

Letting $g_i(n) := np(1+\epsilon_i(n))$, $i = 1, \ldots, j-1$, we use again Hoeffding's inequality to obtain

$$\Pr\{X > g_i(n)\} \leq e^{-\frac{1}{2}np^2\epsilon_i^2}.$$

In addition,

$$\frac{f(g_{i+1}(n))}{f_1(np)} = O\left((1+\epsilon_{i+1}(n))^{m/2}\right).$$

Thus, choosing $\epsilon_i(n)=n^{-a_i}=o(1)$, $i=1,\ldots,j-1$, where a_i are constants to be specified later, the first $j-1$ terms in the summation on the right-hand side of (54) are $O(\exp\{-\gamma n\epsilon_i^2+\delta m\epsilon_{i+1}\})$, $0 \leq i < j-1$, for some positive constants γ and δ, whereas the last term is $O(\exp\{-\gamma n\epsilon_{j-1}^2+\delta' m\})$, where again δ' is a constant. It can be readily verified that, choosing $j=1$ for $a<1/3$ and

$$j = \left\lceil \log\frac{1}{1-2a} \right\rceil > 1$$

otherwise, together with

$$a_i = \frac{1}{2} - 2^i\left(\frac{1}{2} - a\right),\ i = 1, \ldots, j-1$$

the following relations hold:

$$\frac{1}{2} > a := a_0 > a_1 > \cdots > a_{j-1} > a_j := 0$$
$$1 - 2a_i > a - a_{i+1},\ i = 0, \ldots, j-1.$$

Since, in addition, $m=O(n^a)$, all the exponents are of the form $-n^b$ for some positive constant b, and we conclude that $S^{(3)}_f(n)/f_1(np)$ is dominated by the $O(m^2/n)$ error term.

To put all the pieces together, we divide (45) by $f_1(np)$, and use Lemma 4 (where $f(nr) = O(f_1(np))$, Lemma 5, and (53), to conclude, after taking logarithms, that

$$\log S_f(n) = \log f_1(np) + O(f_2 + m^2/n).$$

Theorem 2(i) for $a < 1/2$ then follows from (15) and the definitions of f_1 and f_2.

We need a different approach for the second and third partial sums for the remaining $m = o(n)$ cases, in which $m = \Omega(n^a)$ for all $a < 1/2$. Letting $g_0(n) = n$ (thus collapsing the third region), by Lemmas 4 and 5, we need to estimate $S^{(2)}_{f_1}(n)$. Since f_1' and f_1'' are positive for sufficiently large n, (50) implies that $S^{(2)}_{f_1}(n) \geq f_1(np)$. Therefore

$$f_1(np) \leq S^{(2)}_{f_1}(n) \leq n\max_k \tilde{f}_1(k) \tag{55}$$

where we recall the definition of $\tilde{f}(k)$ from (44). We need to find $k = k^*$ that maximizes the right-hand side of (55), which satisfies

$$\frac{\tilde{f}_1(k^*+1)}{\tilde{f}_1(k^*)} \approx 1. \tag{56}$$

By (15)

$$\frac{f_1(k+1)}{f_1(k)} = O\left((1+1/k)^{\frac{m-1}{2}}\right) = 1 + O(m/k). \tag{57}$$

Thus, (56) takes the form

$$\frac{n-k}{k+1} = \frac{1-p}{p} - O\left(\frac{m}{k}\right)$$

which yields

$$k^* = np + O(m).$$

Applying Stirling's formula, it can then be shown that

$$\log \tilde{f}_1(k^*) = \log f_1(k^*) + O(\log n) + O(m^2/n) \qquad (58)$$

where the first error term is due to the $1/\sqrt{n}$ factor in the formula, and the second error term is due to the discrepancy between k^* and np. In addition

$$\log f_1(k^*) = \frac{m-1}{2}\log\left(\frac{np}{m}\right) + \frac{m}{2}\log e - \frac{1}{2} + \frac{m}{3}\log e\sqrt{\frac{m}{np}} + O\left(\frac{m^2}{n}\right) \qquad (59)$$

where again the error term is due to the discrepancy between k^* and np and is easily seen to dominate other terms in (15). Equations (55), (58), and (59), together with Lemmas 4 and 5, imply (25) of Theorem 2 (i), where the growth rate of m further determines the dominating error terms. □

Remark 1: Notice that one of the error terms generated by the "sandwich argument" of (55), used in the proof of (25), is $O(\log n)$, independent of the value of m. Therefore, this method is not suitable for the $m = O(\log n)$ cases (addressed via a Taylor expansion in the proof of (24)) as this error term would dominate one of the other terms. Moreover, for fixed m, the method cannot even provide the main asymptotic term, which is also $O(\log n)$.

Remark 2: Consider the alternative model mentioned in Section III, where the value of m in the binomial sum grows with k (rather than with n). To analyze this scenario, further assumptions on the growth of $m = m(k)$ with k are needed in Case (i) since, in the computation of the derivatives in (51), as well as of the ratio in (57), we can no longer assume m to be a constant. Assuming that $m(k)$ and its derivatives, $m'(k)$ and $m''(k)$, are continuous functions of k, and that $m(k+1) - m(k) = O(m'(k))$, $m'(k) = O(m/k)$, and $m''(k)=O(m/k^2)$,[9] the same proof can be used, and (24) and (25) remain valid with m replaced with $m(np)$ and the $O(m^2/n)$ error terms replaced with error terms which are $O((m^2/n)\log^2(n/m))$, where the additional factor in the error terms is due to the effect of the variability of m in (51) and (57). In Case (ii), it is easy to see that (26) holds with $K = \log(B_\alpha p + 1 - p)$, a constant (in fact, more terms in the asymptotic expansion can be obtained). Indeed, in this case, the main term under the binomial sum is

$$\tilde{f}(k) = \binom{n}{k} p^k (1-p)^{n-k} B_\alpha^k$$

which leads to a closed-form expression for the summation, namely $(B_\alpha p + 1 - p)^n$ (thus, we avoid the difficulty mentioned in the discussion in Section III regarding the variability of the ratio m/k when m is assumed to grow with n). Finally, if $n=o(m)$, we can also obtain a more precise estimate, under the assumption that $m(k)/k$ is a nondecreasing sequence (which is also natural, since $k/m(k) = o(1)$ in this case): indeed, it is easy to see that the main redundancy term is $n\log(pm/n)$.

[9]These assumptions hold if, e.g., $m(k)/k$ monotonically decreases for sufficiently large k (which is natural since $m(k)/k = o(1)$ in this case) and under natural convexity assumptions.

REFERENCES

[1] T. Batu, S. Guha, and S. Kannan, "Inferring mixtures of Markov chains," *Computational Learning Theory-—COLT*, pp. 186–199, 2004.

[2] S. Boucheron, A. Garivier, and E. Gassiat, "Coding on countably infinite alphabets," *IEEE Trans. Inf. Theory*, vol. 55, no. 1, pp. 358–373, Jan. 2009.

[3] R. Corless, G. Gonnet, D. Hare, D. Jeffrey, and D. Knuth, "On the Lambert W function," *Adv. Comput. Math.*, vol. 5, pp. 329–359, 1996.

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

[5] M. Drmota and W. Szpankowski, "Precise minimax redundancy and regret," *IEEE Trans. Inf. Theory*, vol. 50, no. 11, pp. 2686–2707, Nov. 2004.

[6] P. Flajolet, "Singularity analysis and asymptotics of Bernoulli sums," *Theor. Comput. Sci.*, vol. 215, pp. 371–381, 1999.

[7] P. Flajolet and W. Szpankowski, "Analytic variations on redundancy rates of renewal processes," *IEEE Trans. Inf. Theory*, vol. 48, no. 11, pp. 2911–2921, Nov. 2002.

[8] P. Flajolet and R. Sedgewick, *Analytic Combinatorics*. Cambridge: Cambridge Univ. Press, 2008.

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

[10] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Amer. Stat. Assoc. J.*, pp. 13–30, 1963.

[11] P. Jacquet and W. Szpankowski, "Entropy computations via analytic depoissonization," *IEEE Trans. Inf. Theory*, vol. 45, no. 4, pp. 1072–1081, May 1999.

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

[13] R. E. Krichevskii and V. K. Trofimov, "The performance of universal encoding," *IEEE Trans. Inf. Theory*, vol. 27, no. 2, pp. 199–207, Mar. 1981.

[14] N. Merhav and M. Feder, "Universal prediction," *IEEE Trans. Inf. Theory*, vol. 44, pp. 2124–2147, 1998.

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

[16] A. Orlitsky, N. 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.

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

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

[19] Y. Shtarkov, "Universal sequential coding of single messages," *Probl. Inf. Transmiss.*, vol. 23, pp. 175–186, 1987.

[20] W. Szpankowski, "On asymptotics of certain recurrences arising in universal coding," *Probl. Inf. Transmiss.*, vol. 34, pp. 55–61, 1998.

[21] W. Szpankowski, *Average Case Analysis of Algorithms on Sequences*. New York: Wiley, 2001.

[22] M. J. Weinberger and G. Seroussi, "Sequential prediction and ranking in universal context modeling and data compression," *IEEE Trans. Inf. Theory*, vol. 43, no. 5, pp. 1697–1706, Sep. 1997.

[23] Q. Xie and A. Barron, "Asymptotic minimax regret for data compression, gambling, and prediction," *IEEE Trans. Inf. Theory*, vol. 46, no. 2, pp. 431–445, Mar. 2000.

Wojciech Szpankowski (F'04) is a Saul Rosen Professor of Computer Science and (by courtesy) Electrical and Computer Engineering at Purdue University where he teaches and conducts research in analysis of algorithms, information theory, bioinformatics, analytic combinatorics, random structures, and stability problems of distributed systems. He received the M.S. and Ph.D. degrees in electrical and computer Engineering from Gdansk University of Technology.

He held several Visiting Professor/Scholar positions, including McGill University, INRIA, France, Stanford, Hewlett-Packard Labs, Universite de Versailles, University of Canterbury, New Zealand, Ecole Polytechnique, France, and the Newton Institute, Cambridge, U.K. He is the Erskine Fellow.

In 2010 he received the Humboldt Research Award. In 2001 he published the book *Average Case Analysis of Algorithms on Sequences* (Wiley, 2001). He has been a Guest Editor and Editor of technical journals, including the *Theoretical Computer Science*, the *ACM Transaction on Algorithms*, the IEEE TRANSACTIONS ON INFORMATION THEORY, *Foundation and Trends in Communications and Information Theory*, *Combinatorics*, *Probability*, *and Computing*, and *Algorithmica*. In 2008 he launched the interdisciplinary Institute for Science of Information, and in 2010 he became the Director of the newly established NSF Science and Technology Center for Science of Information.

Marcelo J. Weinberger (M'90–SM'98–F'07) received the Electrical Engineer degree from the Universidad de la República, Montevideo, Uruguay, in 1983, and the M.Sc. and D.Sc. degrees from Technion—Israel Institute of Technology, Haifa, Israel, in 1987 and 1991, respectively, both in electrical engineering.

From 1985 to 1992, he was with the Department of Electrical Engineering at Technion, joining the faculty for the 1991–1992 academic year. During 1992–1993, he was a Visiting Scientist at IBM Almaden Research Center, San Jose, CA. Since 1993, he has been with Hewlett-Packard Laboratories, Palo Alto, CA, where he is a Distinguished Scientist and leads the Information Theory Research group. His research interests include source coding, sequential decision problems, statistical modeling, and image compression. He is a coauthor of the algorithm at the core of the JPEG-LS lossless image compression standard, and was an editor of the standard specification. He also contributed to the coding algorithm of the JPEG2000 image compression standard.

Dr. Weinberger served as an Associate Editor for Source Coding of the IEEE TRANSACTIONS ON INFORMATION THEORY from 1999 to 2002. He is a corecipient of the 2006 IEEE Communications/Information Theory Societies Joint Paper Award.