Minimax Pointwise Redundancy for Memoryless Models over Large Alphabets[*]

Wojciech Szpankowski[†]
Department of Computer Science, Purdue University
W. Lafayette, IN 47907, U.S.A., spa@cs.purdue.edu

Marcelo J. Weinberger
Hewlett-Packard Laboratories
Palo Alto, CA 94304, U.S.A., marcelo.weinberger@hp.com

***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 [15] deriving precise asymptotics of the minimax pointwise redundancy for all ranges of the alphabet size relative to the sequence length. Second, we consider the pointwise minimax redundancy for a family of models in which some symbol probabilities are fixed. The latter problem leads to a binomial sum for functions with super-polynomial 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.**

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], [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.

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 the paper, and we ignore the integer length constraint. Given a sequence

[*] The material in this paper was presented in part at the 2010 IEEE International Symposium on Information Theory, Austin, Texas, USA, June 2010.

[†] This author's contribution was partially made when visiting Hewlett–Packard Laboratories, Palo Alto, CA 94304, U.S.A. This author was also supported 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.

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 above 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] is 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.

In this paper, we formalize the above 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

[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.

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], [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 sub-sequence over $\mathcal{A}$ is np with high probability. But when m also grows, we encounter sub-exponential, exponential and super-exponential 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 precise 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 so-called *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 indexes, 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

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

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]

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

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

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 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:

(i) For $m = o(n)$

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

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

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

where

$$C_\alpha := \frac{1}{2} + \frac{1}{2}\sqrt{1 + \frac{4}{\alpha}} \tag{17}$$

$$A_\alpha := C_\alpha + \frac{2}{\alpha} \tag{18}$$

and

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

(iii) For $n = o(m)$

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

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].

[4] As mentioned, Equation (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.

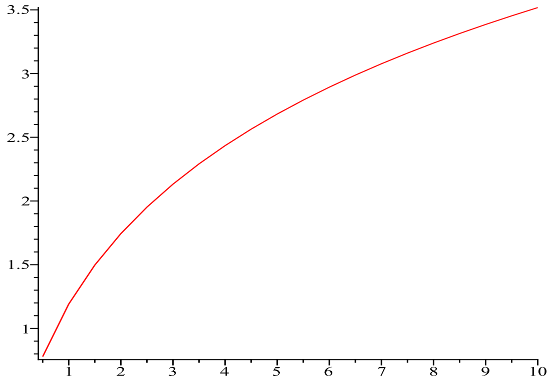

Fig. 1. Value of the constant $\log B_\alpha$ in the $\Theta(n)$ term of $d_{n,m}$ in case $m = \Theta(n)$.

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 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).

For the case $m = \Theta(n)$, the methodology of [15] allowed only to extract 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 Figure 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 $\ell(n) = 0$.

B. *Model family* $\widetilde{\mathcal{M}}_0$

In this subsection 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 below 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

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

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 sub-sequence of x consisting of i symbols from $\mathcal{B}$. Similarly, $y \in \mathcal{A}^{n-i}$ is the sub-sequence 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

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

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

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

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 $D_{n,m,M}$. As mentioned, the generic form of the sum in the lemma, given in Equation (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], [11] to conclude that $D_{n,m,M} \sim D_{np,m}$. However, when m varies with n as in our study, the above expansion does not apply. In particular, the polynomial growth of $D_{n,m}$ ceases to hold and we need to compute asymptotics anew. We state and discuss our second main result in Theorem 2 below, whose proof is presented in Section IV. In the sequel, we will use the notation m_n wherever it is desirable to explicitly show a dependence of m on n.

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 form:

(i_0) If m is fixed, then

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

(i) Let $m_n{\to}\infty$ as n grows, with $m_n{=}o(n)$. Assume:
  (a) $m(x) := m_x$ is a continuous function, as well as its derivatives $m'(x)$ and $m''(x)$.
  (b) $\Delta_n := m_{n+1} - m_n = O(m'(n))$, $m'(n) = O(m/n)$, and $m''(n){=}O(m/n^2)$, where $m'(n)$ and $m''(n)$ are derivatives of $m(x)$ at $x = n$.

  If $m_n = o(\sqrt{n}/\log n)$, then

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

  Otherwise,

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

(ii) Let $m_n = \alpha n + \ell(n)$, where α is a positive constant and $\ell(n)$ is a monotonic function such that $\ell(n) = o(n)$. Then,

$$\begin{aligned} d_{n,m,M} \;\; = \;\; & n \log\left(B_\alpha p + 1 - p\right) \\ & - \log\sqrt{A_\alpha} + O\left(\ell(n) + \frac{1}{\sqrt{n}}\right) \quad (26) \end{aligned}$$

  where A_α and B_α are defined in Theorem 1(ii).

(iii) Let $n{=}o(m_n)$ and assume m_k/k is a nondecreasing sequence. Then,

$$d_{n,m,M} = n \log\left(\frac{p m_n}{n}\right) + O\left(\frac{n^2}{m_n} + \frac{1}{\sqrt{n}}\right). \quad (27)$$

Discussion of Theorem 2

Assumptions. As in Theorem 1, a natural application of our asymptotic analysis in Theorem 2 will assume some large size of the set $\mathcal{A}$, such as $m_n = n^a$ for some a, where the value of a will determine which of the three cases is relevant. In this scenario, all the

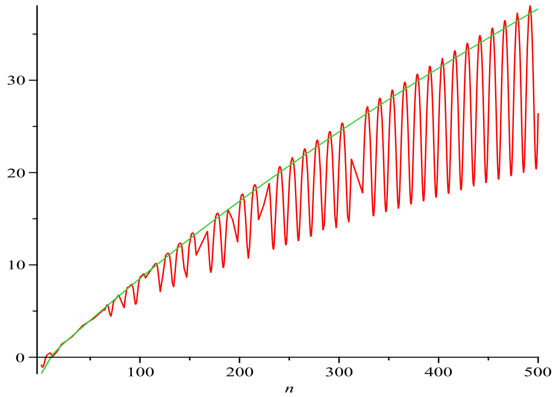

Fig. 2. Comparison of $d_{np,m} = m_{np}/2 \cdot \log(np/m_{np})$ ("zigzag curve") and $d_{n,m,M}$ when $m_n = \sqrt{n}(\sin(n+0.77)+2)$ for $p = 0.5$.

assumptions on m_n hold trivially since, in case (i) ($a < 1$), we have $\Delta_n \approx m'(n) = an^{a-1} = O(m/n)$, and $m''(n) = -a(1-a)n^{a-2} = O(m/n^2)$. We have chosen to state the theorem with more generality because the itemized assumptions actually point to the key properties that the proof will require. For the assumption $\Delta_n = O(m'(n))$ in part (i) of the theorem to hold, we need appropriate smoothness conditions (e.g., $\log m'(x)$ should be of bounded variation). In turn, for the assumption $m'(n) = O(m/n)$ to hold, it suffices to further assume that m_n/n monotonically decreases for sufficiently large n, which is natural since $m_n/n = o(1)$ in this case. Finally, $m''(n) = O(m/n^2)$ requires natural convexity assumptions.[6] If, instead, these assumptions cease to hold due to oscillations (which, as mentioned, are not natural in our context), the claim of the theorem may not hold. For example, for $m_n = \sqrt{n}(\sin(n+0.77)+2)$, we have $m'(n) = O(\sqrt{n}(\cos(n+0.77)+2))$, the assumption $\Delta_n = O(m/n)$ breaks, and, as shown in Figure 2, Theorem 2(i) is invalid.

Similarly, the assumption of a monotonic increase of m_k/k in case (iii) is also natural, since $n/m_n = o(1)$ in this case. We can replace this assumption by the weaker version $1 \leq \frac{m_k}{k} \leq C\frac{m_n}{n}$ for all $k \leq n$ and some $C > 0$, but then we can only show that

$$d_{n,m,M} = n \log\left(\frac{pm_n}{n}\right) + O(n).$$

As for case (ii), as discussed in connection with Theorem 1, the case $\ell(n) \neq 0$ is discussed for completeness only. We have assumed that $\ell(n)$ is monotonic in order to prevent certain types of fluctuations. The result holds under a weaker assumption, though, namely that there exist constants C and a such that, for every pair of positive integers i, j, if $i < j$ we have

$$e^{|\ell(i)|+(1/\sqrt{i})}\, i^a \leq Ce^{|\ell(j)|+(1/\sqrt{j})}\, j^a. \tag{28}$$

Clearly, this condition is satisfied if $\ell(n)$ is monotonic (and therefore so is $|\ell(n)|$ for sufficiently large n). In any case, if $g(n)$ is a monotonic function such that $\ell(n) = O(g(n))$, then the theorem holds with $\ell(n)$ replaced with $g(n)$ in the error term. If $\ell(n)$ is a constant, denoted ℓ, then the constant term in (26) can be shown to be exactly $\log(C_\alpha^\ell/\sqrt{A_\alpha})$. If $\tilde{\ell}(n) := \ell(n) - (\log\sqrt{A_\alpha})/(\log C_\alpha) = \Omega(1)$, under the additional assumption that $|\tilde{\ell}(k)|/k$ is nonincreasing (which is again natural since $\ell(n) = o(n)$), the error term in (26) can be further shown to be $\Theta(\tilde{\ell}(n))$.

Asymptotics. As discussed in Section I, one would expect $d_{n,m,M}$ to behave roughly as $d_{np,m}$ (so that the redundancy depends on $\mathcal{B}$ only through p). This is indeed the case, at least for the main asymptotic terms, in cases (i) and (iii). It is interesting to notice, though, that in case (ii), even the main asymptotic term differs from that of $d_{np,m}$. In passing, let us explain intuitively the asymptotics behind Theorem 2. As shown in Lemma 3, we deal here with the binomial sum which, for a general function f, takes the form (4) (in our case, $f(k) = D_{k,m}$). Observe that, when f grows polynomially, the maximum under the sum occurs around $k = np$, and to find asymptotics we need to sum only within the range $\pm\sqrt{n}$ around np. This observation essentially explains case (i). When $m = \Theta(n)$, the growth of $f(k) = D_{k,m} = O(A^k)$ is exponential, and we need all the terms in the sum in order to extract the asymptotics. Finally, for case (iii), the function $f(k) = D_{k,m}$ grows super-exponentially, and the asymptotics of the binomial sum are determined by the last term, that is, $k = n$.

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

$$[z^n][\beta(z)]^m = \frac{1}{2\pi i}\oint e^{h(z)}dz, \tag{29}$$

where

$$h(z) = m\ln\beta(z) - (n+1)\ln z. \tag{30}$$

We will apply a modification of Lemma 2 in the evaluation of (29), for which we need to check that the

[6] For example, if m_n/n vanishes in a convex manner and m_n is concave, then it is easy to see that $m''(n) = O(m/n^2)$.

necessary conditions are satisfied by the function $h(z)$ of (30).

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 (30), we have

$$z_0 \frac{\beta'(z_0)}{\beta(z_0)} = \frac{n+1}{m}. \tag{31}$$

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

$$z \frac{\beta'(z)}{\beta(z)} = \beta(z)^2 - \beta(z). \tag{32}$$

Thus, (31) takes the form

$$\beta(z_0)^2 - \beta(z_0) = \frac{n+1}{m}. \tag{33}$$

By (7) and the definition of $T(z)$, the range of $\beta(z)$ for $0 \leq z < 1$ is $[1, +\infty)$. Since the quadratic equation (33) has a unique real root in this range, we have

$$\beta(z_0) = \frac{1}{2} + \frac{1}{2}\sqrt{1 + \frac{4(n+1)}{m}} := \frac{1}{\gamma_{n,m}} \tag{34}$$

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), (34) takes the form

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

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

$$z_0 = (1 - \gamma_{n,m}) e^{\gamma_{n,m}} \tag{35}$$

where, since

$$\gamma_{n,m} = \frac{m}{2(n+1)} \left(\sqrt{1 + \frac{4(n+1)}{m}} - 1 \right) \tag{36}$$

we have $0 < \gamma_{n,m} < 1$ and also $0 < z_0 < 1$. We then see that, by (30), (34), and (35), $h(z_0)$ takes the form

$$h(z_0) = -m \ln \gamma_{n,m} - (n+1)[\ln(1-\gamma_{n,m}) + \gamma_{n,m}]. \tag{37}$$

In addition, differentiating (30) twice, we obtain

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

where

$$A(z) = \frac{d}{dz}\left[\frac{\beta'(z)}{\beta(z)}\right] = \frac{[\beta(z)^2 - \beta(z)]\,[2\beta(z)^2 - \beta(z) - 1]}{z^2} \tag{38}$$

with the second equality in (38) easily seen to follow from further differentiating (32). Thus, using (33),

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

which, again by (34) and (35), can be expressed in terms of $\gamma_{n,m}$ as

$$h''(z_0) = \frac{n+1}{(1-\gamma_{n,m})^2 e^{2\gamma_{n,m}}}\left[\frac{2(n+1)}{m} + \frac{1}{\gamma_{n,m}}\right]. \tag{39}$$

Finally, taking another derivative in (38) and further using (32) and (33), after some additional computations, we obtain

$$h'''(z_0) = \frac{n+1}{\gamma_{n,m} z_0^3}\left[\frac{n+1}{m}\left(\frac{8}{\gamma_{n,m}} - 1\right) - \frac{5}{\gamma_{n,m}} + 3\right]. \tag{40}$$

With these expressions on hand, we can now check the conditions required in Lemma 2 for the evaluation of (29). 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 (30). Also, proceeding along the lines of the proof of [21, Theorem 8.17]), it can be shown that Equation (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

$$d_{n,m} = h(z_0)\log e - \log\sqrt{\frac{h''(z_0)}{n}} + O\left(\frac{h'''(z_0)}{(h''(z_0))^{3/2}} + \frac{1}{n}\right) \tag{41}$$

provided 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 (41). In view of (37) and (39), 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 Equation (36), 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 (37) and (39), 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

$$\ln\frac{h''(z_0)}{n} = \ln\frac{n}{m} + \ln 2 + \frac{1}{2}\sqrt{\frac{m}{n}} + O\left(\frac{m}{n}\right). \tag{42}$$

From (40), and noticing that, in this case, Equation (35) yields $z_0 \to 1$, we further obtain

$$h'''(z_0) = \Theta\left(\frac{n^3}{m^2}\right). \tag{43}$$

Theorem 1(i) follows from substituting these equations into (41), observing that (42) and (43) 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 (35) 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 (41) completes the proof of Theorem 1(ii), after observing, again, that the necessary conditions for the modified Lemma 2 hold.

CASE: $n = o(m)$.

Letting $n/m \to 0$ in Equation (36), 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 (37) and (39), 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 (40), and noticing that, in this case, Equation (35) 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 (41), 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

$$S_f(n) = \sum_k \binom{n}{k} p^k (1-p)^{n-k} f(k) \qquad (44)$$

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

CASE: $m_n = o(n)$.

We first 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. Our basic evaluation technique will rely on the concentration of X around its mean np. The following lemma is a straightforward consequence of this concentration.

Lemma 4: Let $g(k)$ be a function satisfying the following condition: There exist constants C and a such that, for every pair of positive integers i, j, with $i < j$, we have $|g(i)|i^a \leq C|g(j)|j^a$. Then, $S_g(n) = O(g(n))$ and $S_{1/|g|}(n) = \Omega(1/g(n))$.

Proof: By Hoeffding's inequality [10], for any $\epsilon > 0$ we have

$$\Pr\{X < n(p - \epsilon)\} \leq e^{-\frac{1}{2}n\epsilon^2}.$$

Therefore,

$$S_g(n) \leq e^{-\frac{1}{2}n\epsilon^2} \max_{1 \leq k \leq n} |g(k)| + \max_{n(p-\epsilon) \leq k \leq n} |g(k)|. \quad (45)$$

By the assumed condition on g, $C|g(n)|n^a$ is an upper bound on $r^a|g(k)|$ for all k in the range $[r, n]$. Letting r take the values $r = 1$ and $r = n(p - \epsilon)$, (45) implies

$$S_g(n) \leq C|g(n)|[e^{-\frac{1}{2}n\epsilon^2} n^a + (p - \epsilon)^{-a}] \leq C'|g(n)|$$

for some constant C'. Similarly,

$$\begin{aligned} S_{1/|g|}(n) &\geq \Pr\{X > n(p - \epsilon)\} \min_{n(p-\epsilon) \leq k \leq n} (1/|g(k)|) \\ &> \frac{(p - \epsilon)^a}{2C} \frac{1}{|g(n)|}. \end{aligned}$$

□

Lemma 4 applies, e.g., to functions that vanish polynomially fast without excessive fluctuations. It holds trivially for nondecreasing functions.

[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 $m = O(n^a)$, 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 $h(z_0)$.

One approach for taking advantage of the concentration of X consists of applying Taylor's theorem to $f(x)$ (the extension of $f(n)$ to the real line) around the mean $x = np$, and estimating $f''(n)$. However, notice that Theorem 1 does not provide enough information about $f(n)$ to obtain such an estimate, since the behavior of $f''(n)$ could be dominated by the error term of $f(n)$. We circumvent this problem by appropriately defining functions f_1 and f_2 such that

$$f(n) = f_1(n)[1 + O(f_2(n))]$$

where $f_2(n)$ is a vanishing function that satisfies the condition of Lemma 4, and

$$\max_{0\le x\le n} |f_1(x)| = O(f_1(n)). \tag{46}$$

It then follows from Lemma 4 and (46) that

$$S_f(n) - S_{f_1}(n) = O(f_1(n)f_2(n)). \tag{47}$$

Next, we estimate $S_{f_1}(n)$ by applying Taylor's theorem to $f_1(x)$ around $x = np$, which yields

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

for some x' that lies between x and np. Letting

$$\xi(n) := \max_{0\le x\le n} |f_1''(x)|$$

we obtain

$$f_1(x) - f_1(np) - (x - np)f_1'(np) = \frac{(x-np)^2}{2} O(\xi(n)). \tag{48}$$

Taking expectations with respect to X in (48), and noting that $\mathbf{E}_X[X] = np$ and $\mathbf{Var}\ [X] = npq$, yields

$$S_{f_1}(n) - f_1(np) = O(n\xi(n))$$

which, together with (47), implies

$$S_f(n) - f_1(np) = O(n\xi(n) + f_1(n)f_2(n)).$$

By (46) we then have

$$S_f(n) = f_1(np)\left[1 + O\left(\frac{n\xi(n)}{f_1(n)} + f_2(n)\right)\right]. \tag{49}$$

As we will show, this bound leads to a precise asymptotic estimate of $S_f(n)$ provided that $n\xi(n) = o(f_1(n))$. In this case, (49) implies

$$d_{n,m,M} = \log S_f(n) = \log f_1(np) + O\left(\frac{n\xi(n)}{f_1(n)} + f_2(n)\right). \tag{50}$$

In the fixed m case we have, by (12),

$$f(n) = Kn^{(m-1)/2}\left[1 + O\left(1/\sqrt{n}\right)\right]$$

where K is a constant. Thus, we can choose $f_1(n) = Kn^{(m-1)/2}$, $f_2(n) = 1/\sqrt{n}$, and all the necessary conditions are obviously satisfied. Hence, Theorem 2(i$_0$) holds. A more precise asymptotic expansion can be found using tools from [6], [11].

Let us now consider part (i) of Theorem 2, that is, we assume that $m \to \infty$ and $m = o(n)$. If we further assume, first, that $m = o(\sqrt{n})$, the error term in (15) dominates the $O(\sqrt{m/n})$ term, and we can then choose

$$f_1(n) = \left(\frac{ne}{m}\right)^{\frac{m}{2}} \sqrt{\frac{m}{2n}}\, e^{\frac{m}{3}\sqrt{\frac{m}{n}}} \tag{51}$$

which clearly satisfies (46), and

$$f_2(n) = O\left(\frac{m^2}{n} + \frac{1}{\sqrt{m}}\right) \tag{52}$$

which vanishes polynomially fast. In order to check the applicability of (50), we need to estimate $\xi(n)/f_1(n)$, for which we will use two of the additional assumptions in this part of the lemma, namely that $O(m'(n)) = O(m/n)$ and $O(m''(n)) = O(m/n^2)$. Now, since for any function g we have $g''/g = [(\ln g)']^2 + (\ln g)''$, it is relatively simple to compute that

$$\frac{nf_1''(n)}{f_1(n)} = O\left(\frac{m^2\log^2 n}{n}\right). \tag{53}$$

Moreover, due to the continuity of m, m', and m'' (which implies the continuity of f_1''), and to the fact that $[f_1(n)m^2\log^2 n]/n^2$ is increasing for sufficiently large n, it is easy to see that (53) holds also when $\xi(n)$ replaces $f_1''(n)$ in the right-hand side. When $m = o(\sqrt{n}/\log n)$, we have $n\xi(n)/f_1(n) = o(1)$ and (24) follows from (50), (51), and (52).

We need a different approach for the remaining $m = o(n)$ cases, since in those cases the error term $O(m^2(\log^2 n)/n)$ does not vanish. Observe that we always have

$$\begin{aligned}\frac{e^{-1/(12np(1-p))}}{\sqrt{2\pi np(1-p)}} f(np) &\le S_f(n)\\ &\le n\max_k\left(\binom{n}{k}p^k(1-p)^{n-k}f(k)\right)\end{aligned} \tag{54}$$

where we have used Stirling's inequality to lower-bound the term corresponding to $k = \lceil np\rceil$ in the sum (44).[9] We need to find $k = k^*$ that maximizes the right-hand side of (54). Let

$$F(k) = \binom{n}{k}p^k(1-p)^{n-k}f(k).$$

[9] If, instead of bounding $S_f(n)$, we use (47) and bound $S_{f_1}(n)$, the fact that $f_1''(x) > 0$ for sufficiently large x immediately implies that $S_{f_1}(n) \ge f_1(np)$ (up to an exponentially decaying term that accounts for the range of values of x for which $f_1''(x) < 0$, if any), which is stronger than the claimed lower bound. However, the $O(\log n)$ term resulting from the use of Stirling's inequality is asymptotically inconsequential.

Then, k^* satisfies

$$\frac{F(k^*+1)}{F(k^*)} \approx 1. \tag{55}$$

We first observe that for our $f(k) = D_{k,m_k}$, using (15) and our assumption that $\Delta_k = O(m_k/k)$, we obtain, after some computations,

$$\frac{f(k+1)}{f(k)} = 1 + O\left(\frac{m_k}{k}\log\frac{k}{m_k}\right).$$

Thus, (55) takes the form

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

which yields

$$k^* = np + O(m_n\log(n/m_n)).$$

Applying Stirling's formula it can then be shown that

$$\log F(k^*) = \log f(k^*) + O(\log n) + O\left(\frac{m_n^2}{n}\log^2\frac{n}{m_n}\right) \tag{56}$$

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,

$$\begin{aligned}\log f(k^*) \;&=\; \frac{m_{np}-1}{2}\log\left(\frac{np}{m_{np}}\right) + \frac{m_{np}}{2}\log e \\ &\quad + \frac{m_{np}}{3}\log e\sqrt{\frac{m_{np}}{n}} + O\left(\frac{m_n^2}{n}\log^2\frac{n}{m_n}\right)\end{aligned} \tag{57}$$

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 (54), (56), and (57) imply (25) of Theorem 2(i), where the growth rate of m_n further determines the dominating error terms.

Case: $m_n = \Theta(n)$.

By (16), since $\ell(n)/n = o(1)$,

$$f(k) = D_{k,m_k} \le A_\alpha^{-\frac{1}{2}} B_\alpha^k 2^{|f_1(k)|}$$

where $f_1(k) = O(\ell(k) + 1/\sqrt{k})$, and the inequality is needed because $\ell(k)$ could be negative. Thus,

$$\begin{aligned} S_f(n) \;&\le\; A_\alpha^{-\frac{1}{2}}(B_\alpha p+q)^n \\ &\quad\cdot \sum_{k=0}^{n}\binom{n}{k}\left(\frac{B_\alpha p}{B_\alpha p+q}\right)^k\left(\frac{q}{B_\alpha p+q}\right)^{n-k} 2^{|f_1(k)|}.\end{aligned}$$

The above sum is upper-bounded by the binomial sum (with parameter $B_\alpha p/(B_\alpha p+q)$ rather than p) for the function $2^{C'(|\ell(k)|+1/\sqrt{k})}$ for some constant C'. Since $\ell(n)$ is assumed monotonic, Condition (28) is satisfied (see discussion on Theorem 2), and therefore we can apply Lemma 4 to this new binomial sum, to obtain

$$S_f(n) \le A_\alpha^{-\frac{1}{2}}(B_\alpha p+q)^n O(2^{f_2(n)}) \tag{58}$$

where $f_2(n) = C'(|\ell(n)| + 1/\sqrt{n})$. Since $2^{f_2(n)} \ge 1$, we conclude that

$$\log S_f(n) \le n\log(B_\alpha p+q) - \log\sqrt{A_\alpha} + O(\ell(n) + 1/\sqrt{n}) \tag{59}$$

where we notice that (59) is in fact an equality whenever $\ell(n) \ge 0$.

To obtain a matching lower bound, we have

$$f(k) \ge A_\alpha^{-\frac{1}{2}} B_\alpha^k 2^{-|f_1(k)|}$$

so that, proceeding as in the upper bound,

$$\begin{aligned} S_f(n) \;&\ge\; A_\alpha^{-\frac{1}{2}}(B_\alpha p+q)^n \\ &\quad\cdot \sum_{k=0}^{n}\binom{n}{k}\left(\frac{B_\alpha p}{B_\alpha p+q}\right)^k\left(\frac{q}{B_\alpha p+q}\right)^{n-k} 2^{-f_2(k)}.\end{aligned}$$

We can now apply the second statement in Lemma 4, to obtain

$$S_f(n) \ge A_\alpha^{-\frac{1}{2}}(B_\alpha p+q)^n \Omega(2^{-f_2(n)})$$

which, after taking logarithms, yields the desired lower bound and, hence, Equation (26) of Theorem 2(ii). A more precise estimate is discussed in Remark 2.

When $\ell(n)$ is a constant, denoted ℓ, the constant term in (16) includes an additional $\ell\log C_\alpha$, which is added also in $d_{n,m_n,M}$, and the error term becomes $O(1/\sqrt{n})$.

Case: $n = o(m)$.

By (20),

$$f(k) = D_{k,m_k} = g_1(k)(1+g_2(k))$$

where $g_2(k) = O(1/\sqrt{k} + k/m_k)$, and

$$\begin{aligned} g_1(k) \;&=\; \left(\frac{m_k}{k}\right)^k e^{3k(k-1)/(2m_k)} \\ &=\; \left(\frac{m_k}{k}\right)^k \left(1 + \frac{3(k-1)}{2m_k} + O\left(\frac{k^2}{m_k^2}\right)\right)^k \\ &=\; \left(\frac{m_k}{k} + \frac{3}{2} + O\left(\frac{k}{m_k}\right) + O\left(\frac{1}{k}\right)\right)^k. \end{aligned}$$

We first use our assumption that $1 \le (m_k/k) \le (m_n/n)$ for all $k \le n$ to obtain the upper bound

$$\begin{aligned} S_{g_1}(n) \;&=\; \sum_{k=1}^{n}\binom{n}{k}p^k q^{n-k} g_1(k) \\ &\le\; \sum_{k=1}^{n}\binom{n}{k}\left(p\left(\frac{m_n}{n} + \frac{3}{2} + O\left(\frac{k}{m_k} + \frac{1}{k}\right)\right)\right)^k q^{n-k} \\ &\le\; \left(\frac{pm_n}{n} + K\right)^n \end{aligned}$$

for some constant K, where we have upper-bounded the $O(k/m_k)$ and $O(1/k)$ terms with a constant, since $k/m_k = o(1)$. In addition, proceeding as in the derivation of (58),

$$\begin{aligned}&\sum_{k=1}^{n}\binom{n}{k}p^k q^{n-k} g_1(k)g_2(k)\\ &\le \left(\frac{pm_n}{n}+K\right)^n O\left(\frac{1}{\sqrt{n}}+\frac{n}{m_n}\right)\end{aligned}$$

where we have used again Lemma 4.[10] Thus,

$$S_f(n) \le \left(\frac{pm_n}{n}+K\right)^n \left(1+O\left(\frac{1}{\sqrt{n}}+\frac{n}{m_n}\right)\right)$$

or

$$\begin{aligned}\log S_f(n) &\le n\log\left(\frac{pm_n}{n}+K\right)+O\left(\frac{1}{\sqrt{n}}+\frac{n}{m_n}\right)\\ &= n\log\frac{pm_n}{n}+O\left(\frac{1}{\sqrt{n}}+\frac{n^2}{m_n}\right). \qquad (60)\end{aligned}$$

On the other hand, we can lower-bound the binomial sum (44) with the term corresponding to $k = n$, namely $p^n D_{n,m_n}$, to obtain

$$\log S_f(n) \ge n\log p + d_{n,m_n}. \qquad (61)$$

Theorem 2(iii) then follows from (60), (61), and (20). If $(m_k/k) \le C(m_n/n)$, we obtain an additional term $n\log C$, thus the error term is $O(n)$. □

Remark 1. Notice that one of the error terms generated by the "sandwich argument" of (54), used in the proof of (25), is $O(\log n)$, independently 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. In part (ii), under the additional assumptions that $\tilde{\ell}(n) := \ell(n) - (\log\sqrt{A_\alpha})/(\log C_\alpha) = \Omega(1)$ and $|\tilde{\ell}(k)|/k$ is nonincreasing, we can further prove that the error term is $\Theta(\tilde{\ell}(n))$. Clearly, our assumptions imply that $\tilde{\ell}(k)$ has constant sign. Assume $\tilde{\ell}(k) > 0$; a similar argument can be used for $\tilde{\ell}(k) < 0$. Then,

$$f(k) = B_\alpha^k 2^{\Theta(\tilde{\ell}(k))} = [B_\alpha 2^{\Theta(\tilde{\ell}(k)/k)}]^k = [B_\alpha 2^{\Omega(\tilde{\ell}(n)/n)}]^k.$$

Therefore, using a bounding technique similar to part (iii), we obtain

$$S_f(n) = [B_\alpha p + q + \Omega(\tilde{\ell}(n)/n)]^n$$

and, after taking the logarithm,

$$d_{n,m,M} = n\log(B_\alpha p + q) + \Omega(\tilde{\ell}(n)).$$

Together with (26), we conclude that the error term is $\Theta(\tilde{\ell}(n))$.

[10] Notice that if m_n/n grows faster than any polynomial, Lemma 4 can still be applied to the $O(1/\sqrt{n})$ term, which will dominate the $O(n/m_n)$ term.

REFERENCES

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

[2] S. Boucheron, A. Garivier and E. Gassiat, "Coding on countably infinite alphabets," *IEEE Trans. Information Theory*, 55, pp. 358–373, 2009.

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

[4] L. D. Davisson, "Universal noiseless coding," *IEEE Trans. Information Theory*, 19, pp. 783–795, 1973.

[5] M. Drmota and W. Szpankowski, "Precise minimax redundancy and regret," *IEEE Trans. Information Theory*, 50, pp. 2686–2707, 2004.

[6] P. Flajolet, "Singularity analysis and asymptotics of Bernoulli sums," *Theoretical Computer Science*, 215, pp. 371–381, 1999.

[7] P. Flajolet and W. Szpankowski, "Analytic variations on redundancy rates of renewal processes," *IEEE Trans. Information Theory*, 48, pp. 2911–2921, 2002.

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

[9] L. Györfi, I. Pali and E. der Meulen, "There is no universal source code for an infinite source alphabet," *IEEE Trans. Information Theory*, 40, pp. 267–271, 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. Information Theory*, 45, pp. 1072–1081, 1999.

[12] J. Kieffer, "A unified approach to weak universal source coding," *IEEE Trans. Information Theory*, 24, pp. 674–682, 1978.

[13] R. E. Krichevskii and V. K. Trofimov, "The performance of universal encoding," *IEEE Trans. Information Theory*, 27, pp. 199–207, 1981.

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

[15] A. Orlitsky and N. Santhanam, "Speaking of infinity," *IEEE Trans. Information Theory*, 50, pp. 2215–2230, 2004.

[16] A. Orlitsky, N. Santhanam, and J. Zhang, "Universal compression of memoryless sources over unknown alphabets," *IEEE Trans. Information Theory*, 50, pp. 1469–1481, 2004.

[17] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Trans. Information Theory*, 30, pp. 629–636, 1984.

[18] G. Shamir, "Universal lossless compression with unknown alphabets: The average case," *IEEE Trans. Information Theory*, 52, pp. 4915–4944, 2006.

[19] Y. Shtarkov, "Universal sequential coding of single messages," *Problems of Information Transmission*, 23, pp. 175–186, 1987.

[20] W. Szpankowski, "On asymptotics of certain recurrences arising in universal coding," *Problems of Information Transmission*, 34, pp. 55–61, 1998.

[21] W. Szpankowski, *Average case analysis of algorithms on sequences*, Wiley, New York, 2001.

[22] M.J. Weinberger and G. Seroussi, "Sequential prediction and ranking in universal context modeling and data compression," *IEEE Trans. Information Theory,* 43, pp. 1697–1706, 1997.

[23] Q. Xie, A. Barron, “Asymptotic minimax regret for data compression, gambling, and prediction,” *IEEE Trans. Information Theory*, 46, pp. 431–445, 2000.