Precise Minimax Redundancy and Regret

January 21, 2003

Michael Drmota
Institut fur Geometrie, TU Wien,
TU Wien
A-1040 Wien,
Austria
michael.drmota@tuwien.ac.at

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

Abstract

Recent years have seen a resurgence of interest in redundancy of lossless coding. The redundancy (regret) of universal fixed–to–variable length coding for a class of sources determines by how much the actual code length exceeds the optimal (ideal over the class) code length. In a minimax scenario one finds the best code for the worst source either in the worst case (called also maximal minimax) or on average. We first study the *worst case* minimax redundancy over a class of stationary ergodic sources and replace Shtarkov's bound by an exact formula. Among others, we prove that a generalized Shannon code minimizes the worst case redundancy, derive asymptotically its redundancy, and establish some general properties. This allows us to obtain precise redundancy rates for memoryless, Markov and renewal sources. For example, we derive the exact constant of the redundancy rate for memoryless and Markov sources by showing that an integer nature of coding contributes $\log(\log m/(m-1))/\log m + o(1)$ where m is the size of the alphabet. Then we deal with the *average* minimax redundancy and regret. Our approach here is orthogonal to most recent research in this area since we aspire to show that asymptotically the average minimax redundancy is equivalent to the worst case minimax redundancy. After formulating some general bounds relating these two redundancies, we prove our assertion for memoryless and Markov sources. Nevertheless, we provide evidence that maximal redundancy of *renewal processes* does not have the same leading term as the average minimax redundancy (however, our general results show that maximal and average *regrets* are asymptotically equivalent).

Index Terms: Universal noiseless coding, universal modeling, minimax redundancy, minimax and maxmin regrets, generalized Shannon code, sequences mod 1, maximum likelihood distribution, analytic information theory.

*The work of this author was supported by NSF Grants CCR-9804760 and CCR-0208709, and contract 1419991431A from sponsors of CERIAS at Purdue.

1 Introduction

Universal coding and universal modeling are two driving forces of information theory, model selection, and statistical inference. In universal coding one is to construct a code for data sequences generated by an unknown source such that, as the length of the sequence increases, the average code length approaches the entropy of whatever processes in the family has generated the data. In seminal works of Davisson [7, 8], Rissanen [20], Krichevsky and Trofimov [17], and Shtarkov [24] it was shown how to construct such codes for finite alphabet sources. Universal codes are often characterized by the average *minimax* redundancy which is the excess over the entropy of the *best* code from a class of decodable codes for the worst process in the family.

As pointed out by Rissanen [21], over years universal coding evolved into *universal modeling* where the purpose is no longer restricted to just coding but rather to finding optimal models [2, 21]. The central question of interest in universal modeling seems to be the code length achievable for *individual* sequences. The burning question is how to measure it. The *worst case* minimax redundancy and regret became handy since they measure the worst case excess of the best code maximized over the processes in the family. In [21] Rissanen also admits that, unfortunately, the redundancy restricted to the first term cannot distinguish between codes that differ by a constant, however large. Then Rissanen points out that the constant can be large if the Fisher information of the data generating source is nearly singular. The goal of this paper is to derive precise results for redundancy and regrets, however small the additional terms are.

Let us establish some notation. A code $C_n : \mathcal{A}^n \to \{0,1\}^*$ is defined as 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 consider here only fixed-to-variable uniquely decodable coding satisfying Kraft's inequality. A source sequence of arbitrary length n is denoted by $x_1^n \in \mathcal{A}^n$. We write X_1^n to denote a random variable representing a message of length n and by $P(x_1^n)$ the probability of the message x_1^n. For a given a code C_n, we let $L(C_n, x_1^n)$ be the code length for x_1^n. Information-theoretic quantities are expressed in binary logarithms written $\lg := \log_2$. We also write $\log := \ln$.

Since Shannon we know that the entropy $H_n(P) = -\sum_{x_1^n} P(x_1^n) \lg P(x_1^n)$ is the absolute lower bound on the expected code length. The *pointwise redundancy* $R_n(C_n, P; x_1^n)$ and the *average redundancy* $\overline{R}_n(C_n, P)$ for a *given* source P are therefore defined as

$$\begin{aligned} R_n(C_n, P; x_1^n) &= L(C_n, x_1^n) + \lg P(x_1^n), \\ \overline{R}_n(C_n, P) &= \mathbf{E}_P[R_n(C_n, P; X_1^n)] = \mathbf{E}[L(C_n, X_1^n)] - H_n(P), \end{aligned}$$

where $\mathbf{E}$ denotes the expectation. As pointed out above, the excess of code length for *individual sequences* is a central issue for universal modeling, therefore we define the *maximal*

or *worst case* redundancy as

$$R_n^*(C_n, P) = \max_{x_1^n}[L(C_n, x_1^n) + \lg P(x_1^n)].$$

Observe that while the pointwise redundancy can be negative, maximal and average redundancies cannot, by Kraft's inequality and Shannon's source coding theorem, respectively.

In practice, one can only hope to have some knowledge about a family of sources $\mathcal{S}$ that generates real data. For example, we may often be able to justify restricting our attention to memoryless sources $\mathcal{S} = \mathcal{M}_0$ or Markov sources of rth order $\mathcal{S} = \mathcal{M}_r$. Sometimes, however, we must consider a larger class of non-finitely parameterized sources such as renewal sources $\mathcal{S} = \mathcal{R}_0$. Following Davisson [7] we define the average minimax redundancy $\overline{R}_n(\mathcal{S})$ and the worst case (maximal) minimax redundancy $R_n^*(\mathcal{S})$ for family $\mathcal{S}$ as follows

$$\overline{R}_n(\mathcal{S}) = \min_{C_n \in \mathcal{C}} \sup_{P \in \mathcal{S}} \left(\sum_{x_1^n} P(x_1^n) \left[L(C_n, x_1^n) + \lg P(x_1^n) \right] \right), \tag{1}$$

$$R_n^*(\mathcal{S}) = \min_{C_n \in \mathcal{C}} \sup_{P \in \mathcal{S}} \max_{x_1^n} \left[L(C_n, x_1^n) + \lg P(x_1^n) \right], \tag{2}$$

where $\mathcal{C}$ denotes the set of all fixed-to-variable codes satisfying the Kraft inequality. In words, we search for the best code for the worst source on average and for individual sequences.

We should also point out that there are other measures of optimality for coding, gambling and prediction that are used in universal modeling and coding. We refer here to minimax *regret* functions defined as follows (cf. [2, 21, 29, 30]):

$$\overline{r}_n(\mathcal{S}) = \min_{C_n \in \mathcal{C}} \sup_{P \in \mathcal{S}} \sum_{x_1^n} P(x_1^n)[L(C_n, x_1^n) + \lg \sup_{P \in \mathcal{S}} P(x_1^n)], \tag{3}$$

$$r_n^*(\mathcal{S}) = \min_{C_n \in \mathcal{C}} \max_{x_1^n} [L(C_n, x_1^n) + \lg \sup_{P \in \mathcal{S}} P(x_1^n)], \tag{4}$$

and to the maxmin regret

$$\underline{r}_n(\mathcal{S}) = \sup_{P \in \mathcal{S}} \min_{C_n \in \mathcal{C}} \sum_{x_1^n} P(x_1^n)[L(C_n, x_1^n) + \lg \sup_{P \in \mathcal{S}} P(x_1^n)]. \tag{5}$$

We call $\overline{r}_n(\mathcal{S})$ the *average* minimax regret, $r_n^*(\mathcal{S})$ the worst case (maximal) minimax regret and $\underline{r}_n(\mathcal{S})$ the maxmin regret. Clearly, $\overline{R}_n(\mathcal{S}) \leq \overline{r}_n(\mathcal{S})$, and, as easy to establish,

$$r_n^*(\mathcal{S}) = R_n^*(\mathcal{S}).$$

Thus, we will not consider $r_n^*(\mathcal{S})$ in the sequel.

Finally, we may link universal modeling with game theory and statistics by ignoring an integer nature for the coding interpretations: Suppose nature picks up a distribution P from

$\mathcal{S}$ and we try to find a distribution Q as the best guess for P. We may then reformulate the above redundancy (average and maximal) as well as regrets so that $L(C_n, x_1^n)$ is replaced by its continuous approximation, namely, $\lg 1/Q(x_1^n)$. We denote these corresponding continuous redundancy and regret by placing a tilde over R or r. For example, $\widetilde{R}_n(\mathcal{S})$ and $\widetilde{R}_n^*(\mathcal{S})$ denote continuous approximations of the average and the worst case minimax redundancies. They can be explicitly defined as

$$\widetilde{R}_n(\mathcal{S}) = \inf_Q \sup_{P\in\mathcal{S}} \left(\sum_{x_1^n} P(x_1^n) \lg \frac{P(x_1^n)}{Q(x_1^n)} \right) = \inf_Q \sup_{P\in\mathcal{S}} D_n(P||Q), \tag{6}$$

$$\widetilde{R}_n^*(\mathcal{S}) = \inf_Q \sup_{P\in\mathcal{S}} \max_{x_1^n} \left[\lg \left(P(x_1^n)/Q(x_1^n) \right) \right], \tag{7}$$

where $D_n(P||Q)$ is the Kullback divergence between Q and P. The average minimax regret is defined in a similar manner, namely,

$$\widetilde{r}_n(\mathcal{S}) = \inf_Q \sup_{P\in\mathcal{S}} \left(\sum_{x_1^n} P(x_1^n) \lg \frac{\sup_{P\in\mathcal{S}} P(x_1^n)}{Q(x_1^n)} \right).$$

Clearly, the continuous approximation of the redundancy and regrets are within one bit from the corresponding redundancy and regrets, that is,

$$\overline{R}_n(\mathcal{S}) \leq \widetilde{R}_n(\mathcal{S}) \leq \overline{R}_n(\mathcal{S}) + 1, \tag{8}$$

$$\overline{r}_n(\mathcal{S}) \leq \widetilde{r}_n(\mathcal{S}) \leq \overline{r}_n(\mathcal{S}) + 1. \tag{9}$$

Indeed, it suffices to consider Shannon code for the optimal distribution.

We now summarize our main findings in the context of recent research in the area of universal coding and modeling. We should mention that to the best of our knowledge all known results concern continuous approximation of minimax redundancy and regrets while we consider the "true" worst case minimax redundancy $R_n^*(\mathcal{S})$ and show that the integer nature of coding contributes

$$\frac{\log \frac{1}{m-1} \log m}{\log m} + o(1),$$

where m is the size of the alphabet. More precisely, we first observe that the worst case minimax redundancy can be decomposed as follows (cf. Theorem 1; this is also implicitly used by Shtarkov [24])

$$R_n^*(\mathcal{S}) = \sum_{x_1^n} \sup_{P\in\mathcal{S}} P(x_1^n) + R_n^*(Q^*), \tag{10}$$

where the maximum likelihood distribution Q^* is defined as

$$Q^*(x_1^n) := \frac{\sup_{P\in\mathcal{S}} P(x_1^n)}{\sum_{x_1^n} \sup_{P\in\mathcal{S}} P(x_1^n)}.$$

It turns out that $R_n^*(Q^*)$ is the worst case redundancy of a generalized Shannon code (designed for the maximum likelihood distribution Q^*). A generalized Shannon code C_n^{GS} is such that

$$L(x_1^n, C_n^{GS}) = \begin{cases} \lfloor \lg 1/P(x_1^n) \rfloor & \text{if} \quad x_1^n \in \mathcal{L} \\ \lceil \lg 1/P(x_1^n) \rceil & \text{if} \quad x_1^n \in \mathcal{U} \end{cases}$$

where $\mathcal{L} \cup \mathcal{U} = \mathcal{A}^n$ is a partition of the set of all source sequences of length n such that Kraft's inequality is satisfied. Observe that the first term of (10) depends only on the class of processes ("richness" of the family of distributions) while the second term is responsible for coding. We may conclude that the optimal code for the maximal minimax redundancy is the generalized Shannon code for the distribution Q^*.

In order to justify the last assertion we consider the following coding problem. For a given distribution P, we look for a prefix code C_n such that

$$R_n^*(P) = \min_{C_n} \max_{x_1^n} [L(C_n, x_1^n) + \lg P(x_1^n)].$$

Observe that when the "max" operator above is replaced by the average operator $\mathbf{E}$, then the optimal code is known to be the Huffman code [5]. But what is the optimal code in the worst case? First, notice that a code that minimizes the longest code length (i.e., the one that solves $\min_{C_n} \max_{x_1^n} [L(C_n, x_1^n)]$) is such that builds a balanced coding tree and, therefore, its optimal length is $\lceil \log_m(\text{number of codewords}) \rceil$. The situation is much more interesting when the worst case redundancy is to be optimized. We shall prove in Theorem 1 that a generalized Shannon code satisfying Kraft's inequality is the optimal one. Then using analytic tools of analysis of algorithms (in particular, sequences distributed mod 1) we show (cf. Theorem 2) that for a known binary memoryless source the optimal worst case redundancy converges to a constant equal to $\log\log(2)/\log(2) + o(1) \approx 0.5287$ when $\log(1-p)/p$ is irrational and diverges (fluctuates) otherwise, where p is the *known* probability of generating a "0". Interestingly enough, the fluctuating part disappears for a *family* of memoryless sources (cf. Theorem 3). Similar results are proved for Markovian sources (cf. Theorem 4 and Theorem 5).

Let us now review what is known about the continuous approximation

$$\widetilde{R}_n^*(\mathcal{S}) := \lg d_n(\mathcal{S}) := \lg \left(\sum_{x_1^n} \sup_{P \in \mathcal{S}} P(x_1^n) \right) \tag{11}$$

of the worst case minimax redundancy. Shtarkov [24] introduced the worst case minimax redundancy problem and gave the first asymptotics of the form $d/2 \log n + O(1)$ for memoryless sources and Markov sources where d is the number of parameters (e.g., $d = m - 1$ for m-ary alphabet memoryless source and $d = m^r(m-1)$ for Markov sources of order r). The constant of $\widetilde{R}_n^*(\mathcal{M}_0)$ was identified in [25, 30] for memoryless sources, and in Rissanen [21]

and Jacquet and Szpankowski [13] for Markov sources. Szpankowski [26], using analytic tools of analysis of algorithms such as generating functions and singularity analysis [28], derived a full asymptotic expansion for memoryless sources; the first few terms are given below

$$\begin{aligned}\widetilde{R}_n^*(\mathcal{M}_0) &= \frac{m-1}{2}\lg\left(\frac{n}{2}\right) + \lg\left(\frac{\sqrt{\pi}}{\Gamma(\frac{m}{2})}\right) + \frac{\Gamma(\frac{m}{2})m}{3\Gamma(\frac{m}{2}-\frac{1}{2})}\cdot\frac{\sqrt{2}}{\sqrt{n}} \\ &+ \left(\frac{3+m(m-2)(2m+1)}{36} - \frac{\Gamma^2(\frac{m}{2})m^2}{9\Gamma^2(\frac{m}{2}-\frac{1}{2})}\right)\cdot\frac{1}{n} + O\left(\frac{1}{n^{3/2}}\right),\end{aligned} \tag{12}$$

where Γ is the Euler gamma function. In this paper, we find the correction contributed by the integer nature of coding. In particular, we show that for memoryless sources

$$R_n^*(\mathcal{M}_0) = \frac{m-1}{2}\lg\left(\frac{n}{2}\right) - \frac{\log\frac{1}{m-1}\log m}{\log m} + \log\left(\frac{\sqrt{\pi}}{\Gamma(\frac{m}{2})}\right) + O\left(\frac{\log n}{n^{1/9}}\right).$$

Similar results are obtained for Markov sources. Finally, regarding non-Markovian sources, Csiszár and Shields [6] proved that the worst case redundancy of renewal processes is $\Theta(\sqrt{n})$, while recently Flajolet and Szpankowski [11] improved it to

$$R_n^*(\mathcal{R}_0) = \frac{2}{\log 2}\sqrt{\left(\frac{\pi^2}{6}-1\right)n} + O(\log n).$$

Interestingly enough, in a very recent development Orlitsky and his students [14] studied memoryless sources over *unbounded* alphabet and showed that the worst case redundancy behaves like $R_n^*(\mathcal{S}) = \frac{1}{2}\sqrt[3]{n} + O(\log n)$.

Let us now review known results for the *average* minimax $\overline{R}_n(\mathcal{S})$. This case seems to be harder to analyze since supremum operator and the average operator $\mathbf{E}$ do not commute as in the worst case minimax redundancy. Not surprisingly, only $\widetilde{R}_n(\mathcal{S})$ was analyzed thus far with the exception of Szpankowski [27] (cf. also [10]) who obtained the leading term in the asymptotic expansion of the Huffman code for *known* memoryless distribution. Nevertheless, an impressive body of research was built over years regarding $\widetilde{R}_n(\mathcal{S})$ that we review next and put our results in this context.

The average minimax redundancy is almost entirely considered within the framework of Bayes rule and parameterized family of distributions. Let now $\mathcal{S} = \{P^\theta\}_{\theta\in\Theta}$. The average minimax problem is then reformulated as

$$\widetilde{R}_n(\Theta) = \inf_Q \sup_{\theta\in\Theta} D_n(P^\theta||Q).$$

In the Bayesian framework, one assumes that the parameter θ is generated by the density $w(\theta)$ and the mixture density $M_n^w(x_1^n)$ is defined as

$$M_n^w(x_1^n) = \int P^\theta(x_1^n)w(d\theta).$$

Observe now

$$\begin{aligned}
\inf \mathbf{E}_w[D_n(P^\theta||Q)] &= \inf_Q \int D(P^\theta||Q)dw(\theta) \\
&= \inf_Q \left(\sum_{x_1^n} M_n^w(x_1^n) \log \frac{1}{Q(x_1^n)} - \int \sum_{x_1^n} P^\theta \log P^\theta dw(\theta) \right) \\
&\stackrel{(A)}{=} \sum_{x_1^n} M_n^w(x_1^n) \log \frac{1}{M(x_1^n)} + \int \sum_{x_1^n} P^\theta \log P^\theta dw(\theta) \\
&= \int D_n(P^\theta||M^w)dw(\theta) \\
&= I(\Theta; X_1^n),
\end{aligned}$$

where line (A) follows from the fact (cf. [5])

$$\min_Q \sum_i P_i \log \frac{1}{Q_i} = \sum_i P_i \log \frac{1}{P_i},$$

and $I(\Theta, X_1^n)$ is the mutual information between the parameter space and the source output. As pointed out by Gallager, and Davisson and Leon-Garcia[8] the minimax theorem of game theory entitles us to conclude that

$$\widetilde{R}_n(\Theta) = \inf_Q \sup_{\theta \in \Theta} D_n(P^\theta||Q) = \sup_w \int D(P^\theta||M_n^w)dw(\theta) := C(\Theta, X_1^n),$$

that is, the continuous approximation of the average minimax redundancy is equal to the channel capacity $C(\Theta, X)$ between the parameter space and source output. For a generalization of this result see [19].

In view of the above, the average minimax redundancy problem is reduced to finding the optimal prior distribution $\overline{w}^*(\theta)$ for the Bayes rule. This was accomplished by Bernardo [3] who proved that asymptotically

$$\overline{w}^*(\theta) = \frac{\sqrt{\det \mathbf{I}(\theta)}}{\int \sqrt{\det \mathbf{I}(x)}dx},$$

where $\mathbf{I}(\theta)$ is the Fisher information matrix

$$\mathbf{I}(\theta) = \left\{ -\mathbf{E} \left[\frac{\partial^2 w(\theta)}{\partial \theta_i \theta_j} \right] \right\}_{\theta_i, \theta_j \in \Theta}.$$

For example, for a binary memoryless source with one parameter θ

$$I(\theta) = \frac{1}{\theta(1-\theta)}, \qquad \overline{w}^*(\theta) = \frac{1}{\pi\sqrt{\theta(1-\theta)}},$$

while for a memoryless source over m-ary alphabet the optimal prior is Dirichlet$(1/2, \ldots, 1/2)$ density (cf. [4, 29]). Finally, we observe that the optimal *coding* distribution $\overline{Q}^* = M_n(\overline{w}^*)$, that is, it is a mixture with $\overline{w}^*$ prior.

There is a long list of contributing authors to our current understanding of the average minimax redundancy (formally, its continuous approximation). Krichevsky and Trofimov [17] show that $\overline{R}_n(\mathcal{M}_0) = (m-1)/2 \cdot \log n + O(1)$ for memoryless sources while Rissanen proves in [20] that for any code and almost all θ the average maxmin redundancy is bounded from below by $d/2 \log n - o(\log n)$ where d is the dimension of Θ. Clark and Barron [4], and Xie and Barron [29] find explicit constant in $\widetilde{R}_n(\mathcal{M}_0) = d/2 \log n + c_\theta + o(1)$ for codes based on mixtures. Atteson [1] extended the result of Clark and Barron to Markov sources. We should point out that the computation involved in these analyses are appreciately more complicated (mostly based on a subtle application of the multidimensional saddle point method). Finally, Csiszár and Shields [6] proved that the average minimax for renewal processes is $\Theta(\sqrt{n})$, while Shields [23] shown that there are no universal redundancy rates for general general stationary ergodic processes.

In view of these difficulties and our better understanding of the worst case minimax redundancy, we propose in this paper an orthogonal approach to the average minimax redundancy and regrets. Based on previous results (cf. [2, 4, 29, 30]) we put forward the following conjecture: *for a large class of sources $\mathcal{S}$ the maximal and the average redundancy are asymptotically equivalent for large n, that is*

$$\overline{R}_n(\mathcal{S}) \sim R_n^*(\mathcal{S}), \tag{13}$$

where $a_n \sim b_n$ means $\lim_{n\to\infty} a_n/b_n = 1$.

We now summarize our main findings in this area, In Theorem 6 we basically show that if the maximum likelihood distribution Q^* belongs to the convex hull of $\mathcal{S}$, then

$$\widetilde{R}_n(\mathcal{S}) - \widetilde{R}_n^*(\mathcal{S}) = O(c_n(\mathcal{S}))$$

where

$$c_n(\mathcal{S}) = \sup_{P\in\mathcal{S}} \sum_{x_1^n} P(x_1^n) \lg \frac{\sup_{P\in\mathcal{S}} P(x_1^n)}{P(x_1^n)}.$$

In particular, if we can show that $c_n(\mathcal{S}) = o(\lg d_n)$ (cf. (11) above for the definition of d_n), then our conjecture (13) is true. We shall prove it for memoryless sources $\mathcal{M}_0$ (cf. 9) and Markov sources $\mathcal{M}_r$ (cf. Theorem 5). But, to our surprise we indicate in Lemma 3 that the conjecture seems to fail for renewal processes (i.e., we show that $c_n(\mathcal{R}_0) = \Theta(\sqrt{n})$).

Finally, we wrestle with the worst case redundancy $R_n^*(\mathcal{S})$ and the average regret $\overline{r}_n(\mathcal{S})$. Interestingly enough, we prove in Theorem 7 that $\widetilde{R}_n^*(\mathcal{S}) = \widetilde{r}_n(\mathcal{S})$ provided Q^* belongs to the convex hull of $\mathcal{S}$. More precisely, we show that

$$R_n^*(\mathcal{S}) = \overline{r}_n(\mathcal{S}) + O(1).$$

We finally approaching the end of this long introduction. We now only mention that the paper is organized in the following manner. In the next section we summarize formally

our main results. In Section 3 we derive most of our results concerning the worst case redundancy for memoryless and Markov sources, while in Section 4 we prove our main findings about the relation between the average and the worst case redundancy and regrets together with specific results concerning memoryless, Markov, and renewal processes.

2 Summary of Main Results

In this section we formulate our main results concerning the worst case (maximal) minimax redundancy (cf. Section 2.1) and the average minimax redundancy and regrets (cf. Section 2.2). In the latter section we relate the average minimax redundancy to the maximal redundancy. Most of the proofs will be delayed till Sections 3 and 4.

2.1 The Worst Case Minimax Redundancy

In 1987 Shtarkov [24] proved that

$$\lg\left(\sum_{x_1^n}\sup_{P\in\mathcal{S}}P(x_1^n)\right)\le R_n^*(\mathcal{S})\le\lg\left(\sum_{x_1^n}\sup_{P\in\mathcal{S}}P(x_1^n)\right)+1. \tag{14}$$

We need a more precise result for the maximal minimax redundancy $R_n^*(\mathcal{S})$ that will replace the inequalities above by an equality. For convenience we set

$$d_n=d_n(\mathcal{S}):=\sum_{x_1^n}\sup_{P\in\mathcal{S}}P(x_1^n)$$

and

$$Q^*(x_1^n):=\frac{\sup_{P\in\mathcal{S}}P(x_1^n)}{d_n}. \tag{15}$$

The distribution Q^* is called the *maximum likelihood distribution*. We also write, as already mentioned in the introduction, $\widetilde{R}_n^*(\mathcal{S}):=\lg d_n(\mathcal{S})$ for the continuous approximation of $R_n^*(\mathcal{S})$.

We start with a simple result that decomposes the maximal minimax redundancy into two terms: the first one depends only on the underlying class of processes while the second one involves coding (which is implicitly also used in [24]).

Lemma 1 *Let $\mathcal{S}$ be a system of probability distributions P on $\mathcal{A}^n$ and let Q^* be the maximum likelihood distribution defined by (15). Then*

$$R_n^*(\mathcal{S})=R_n^*(Q^*)+\widetilde{R}_n^*(\mathcal{S})=R_n^*(Q^*)+\lg d_n(\mathcal{S}) \tag{16}$$

where

$$R_n^*(Q^*)=\min_{C_n\in\mathcal{C}}\max_{x_1^n}\left(L(C_n,x_1^n)+\lg Q^*(x_1^n)\right)$$

is the worst case redundancy of a single source $\mathcal{S}=\{Q^*\}$.

Proof. By definition and noting that max and sup commute, we have

$$\begin{aligned} R_n^*(\mathcal{S}) &= \min_{C_n\in\mathcal{C}} \sup_{P\in\mathcal{S}} \max_{x_1^n} (L(C_n, x_1^n) + \lg P(x_1^n)) \\ &= \min_{C_n\in\mathcal{C}} \max_{x_1^n} \left(L(C_n, x_1^n) + \sup_{P\in\mathcal{S}} \lg P(x_1^n) \right) \\ &= \min_{C_n\in\mathcal{C}} \max_{x_1^n} (L(C_n, x_1^n) + \lg Q^*(x_1^n) + \lg d_n) \\ &= R_n^*(Q^*) + \lg d_n. \end{aligned}$$

which proves the lemma. ∎

In passing we observe an interesting property of the continuous part $\widetilde{R}_n^*(\mathcal{S})$ of the maximal minimax redundancy, namely that it is a *non-decreasing* function of n. Indeed, we have

$$\begin{aligned} \widetilde{R}_n^*(\mathcal{S}) &= \lg \left(\sum_{x_1^n} \sup_{P\in\mathcal{S}_n} P(x_1^n) \right) \\ &= \lg \left(\sum_{x_1^n} \sup_{P\in\mathcal{S}_n} \sum_{z\in\mathcal{A}} \widetilde{P}(x_1^n z) \right) \\ &\le \lg \left(\sum_{x_1^{n+1}} \sup_{\widetilde{P}\in\mathcal{S}_{n+1}} \widetilde{P}(x_1^{n+1}) \right) = \widetilde{R}_{n+1}^*(\mathcal{S}), \end{aligned}$$

that is, $\widetilde{R}_n^*(\mathcal{S}) \le \widetilde{R}_{n+1}^*(\mathcal{S})$. The second part $R_n^*(Q^*)$ of $R_n^*(\mathcal{S})$ is the maximal redundancy of an optimal code for the distribution Q^* and its behavior might be quite erratic, as discussed below. We next compute a closed form formula for this maximal redundancy and find the optimal code.

Let us introduce a natural generalization of the Shannon code that we call a *generalized Shannon code* denoted as C_n^{GS}. For a *given* distribution $P \in \mathcal{S}$ we define its code length as

$$L(x_1^n, C_n^{GS}) = \begin{cases} \lfloor \lg 1/P(x_1^n) \rfloor & \text{if} \quad x_1^n \in \mathcal{L} \\ \lceil \lg 1/P(x_1^n) \rceil & \text{if} \quad x_1^n \in \mathcal{U}, \end{cases}$$

where $\mathcal{L} \cup \mathcal{U} = \mathcal{A}^n$ is a partition of the set of all source sequences of length n. In addition, we shall postulate Kraft's inequality is to hold, that is, for a binary alphabet[†]

$$\sum_{x_1^n\in\mathcal{L}} P(x_1^n) 2^{\langle -\lg P(x_1^n)\rangle} + \frac{1}{2} \sum_{x_1^n\in\mathcal{U}} P(x_1^n) 2^{\langle -\lg P(x_1^n)\rangle} \le 1,$$

where $\langle x \rangle = x - \lfloor x \rfloor$ is the fractional part of x.

[†]From now on we mostly work with a binary alphabet $\mathcal{A} = \{0, 1\}$.

Our first main result proves that there exists a generalized Shannon code which is optimal with respect to the maximal redundancy. Let us define the following *partitions* of $\mathcal{A}^n$

$$\mathcal{L}_t := \{x_1^n \in \mathcal{A}^n : \langle -\lg P(x_1^n)\rangle < t\}, \tag{17}$$

$$\mathcal{U}_t := \{x_1^n \in \mathcal{A}^n : \langle -\lg P(x_1^n)\rangle \geq t\} \tag{18}$$

where $0 < t < 1$. It is should be clear that that for such a code

$$R_n^*(C_n^{GS}, P) = 1 - t$$

for a given distribution P.

As already mentioned above, all our results are formulated for a binary alphabet to simplify exposition. Also, throughout we write $\mathbb{Z}$ for the set of integers, $\mathbb{Q}$ for rational numbers, $\mathbb{R}$ for the set of reals, and $\langle x\rangle = x - \lfloor x \rfloor$ for the fractional part of $x \in \mathbb{R}$. In Section 3 we prove the following result.

Theorem 1 *Suppose that $\mathcal{S}$ is a set of probability distributions P on $\mathcal{A}^n$ and let Q^* be the maximum likelihood distribution defined by (15). If the probability distribution Q^* is dyadic, i.e. $\lg Q^*(x_1^n) \in \mathbb{Z}$ for all $x_1^n \in \mathcal{A}^n$, then*

$$R_n^*(\mathcal{S}) = \lg d_n(\mathcal{S}). \tag{19}$$

Otherwise, set $T = T(Q^*) := \{\langle -\lg Q^*(x_1^n)\rangle : x_1^n \in \mathcal{A}^n\}$ and let $t_0 \in T$ be the largest t such that

$$\sum_{x_1^n \in \mathcal{L}_t} Q^*(x_1^n) 2^{\langle -\lg Q^*(x_1^n)\rangle} + \frac{1}{2} \sum_{x_1^n \in \mathcal{U}_t} Q^*(x_1^n) 2^{\langle -\lg Q^*(x_1^n)\rangle} \leq 1, \tag{20}$$

where $\mathcal{L}_t$ and $\mathcal{U}_t$ are defined in (17) and (18), respectively. Then

$$R_n^*(\mathcal{S}) = \lg d_n(\mathcal{S}) + 1 - t_0 \tag{21}$$

and the optimum is obtained for a generalized Shannon code with $\mathcal{L} = \mathcal{L}_{t_0}$ and $\mathcal{U} = \mathcal{U}_{t_0}$, that is, $R_n^*(Q^*) = 1 - t_0$.

Observe that for $\mathcal{S} = \{P\}$ (a single known source) $\lg d_n(P) = 0$ and $R_n^*(\mathcal{S}) = R_n^*(P) = 1 - t_0$ where t_0 is the largest t for which the Kraft inequality (20) holds. In summary, we just found that a prefix code solving the following optimization problem

$$R_n^*(P) = \min_{C_n \in \mathcal{C}} \max_{x_1^n} [L(C_n, x_1^n) + \lg P(x_1^n)] \tag{22}$$

is the generalized Shannon code with parameter t_0. (This should be compared to the Huffman code that is optimal for the *average* redundancy.) We will prove it more formally in Lemma 4 of Section 3 that will automatically imply Theorem 1.

But how to compute $t_0 = 1 - R_n^*(P)$ and more generally $R_n^*(Q^*)$ for a set of sources? We next find precise formulas for the maximal redundancy for a given source $R_n^*(P)$ and the maximal minimax redundancy for memoryless sources $\mathcal{M}_0$ and Markovian sources $\mathcal{M}_r$.

We start with a binary **memoryless source** over a binary alphabet. We consider the distribution $P_p(x_1^n) = p^k(1-p)^{n-k}$ where k is the number of "0" in x_1^n and p is the probability of generating a "0". We first assume that p is known, thus $\mathcal{S} = \{P_p\}$ is a single distribution. We prove in Section 3 the following surprising result.

Theorem 2 *Suppose that* $\lg \frac{1-p}{p} \notin \mathbb{Q}$ *is irrational. Then as* $n \to \infty$

$$R_n^*(P_p) = -\frac{\log\log 2}{\log 2} + o(1) = 0.5287\ldots + o(1),$$

where the term $o(1)$ *depends on* p *and cannot be generally improved. If* $\lg \frac{1-p}{p} = \frac{N}{M} \in \mathbb{Q}$ *(for some coprime integers* $M, N \in \mathbb{Z}$*) is rational and non-zero, then as* $n \to \infty$

$$R_n^*(P_p) = -\frac{\lfloor M \lg(M(2^{1/M}-1)) - \langle Mn \lg 1/(1-p)\rangle \rfloor + \langle Mn \lg 1/(1-p)\rangle}{M} + o(1).$$

Finally, if $\lg \frac{1-p}{p} = 0$ *then* $p = \frac{1}{2}$ *and* $R_n^*(P_{1/2}) = 0$.

Next we consider a *class* of (unknown) memoryless sources P_p such that $p \in [a,b]$ for some $0 \le a < b \le 1$. Interesting enough, in this case the rational case of $R_n^*(Q^*)$ disappears and we obtain a precise result for the maximal minimax redundancy of memoryless sources.

Theorem 3 *Let* $0 \le a < b \le 1$ *be given and let* $\mathcal{M}_0^{a,b} = \{P_p : a \le p \le b\}$. *Then as* $n \to \infty$

$$R_n^*(\mathcal{M}_0^{a,b}) = \frac{1}{2}\lg n + \lg\, C_{a,b} - \frac{\log\log 2}{\log 2} + O_{a,b}\left(n^{-1/9}\log n\right),$$

where

$$C_{a,b} = \frac{1}{\sqrt{2\pi}}\int_a^b \frac{dx}{\sqrt{x(1-x)}} = \sqrt{\frac{2}{\pi}}(\arcsin\sqrt{b} - \arcsin\sqrt{a}).$$

Remark. If $a = 0$ and $b = 1$, then we find $C_{0,1} = \sqrt{\pi/2}$ which agrees with known results (cf. [25, 26, 30]), however, the term $\log\log 2/\log 2$ was not derived before. If $a > 0$ and $b < 1$ the error term can be improved to $O_{a,b}(n^{-1/3}\log n)$. As pointed out in the introduction, for m-ary alphabet the second constant term becomes $\log\log m^{\frac{1}{m-1}}/\log m$. Also, in (12) we quoted a result from [26] that provides a full asymptotic expansion of the term $\widetilde{R}_n^*(\mathcal{M}_0)$ for an m-ary alphabet.

Similar results can be proved for a class of **Markov sources** $\mathcal{M}_r$ of order r. To simplify our exposition, we only present results for binary Markov sources of order one, that is, we set now $\mathcal{S} = \mathcal{M}_1$.

From now on we assume that the transition matrix of the Markov source is

$$\mathbf{P} = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix},$$

where $p_{ij} = \Pr\{X_{k+1} = j \,|X_k = i\}$. The stationary distribution is then

$$p_0 = \frac{p_{10}}{p_{10} + p_{01}} \quad \text{and} \quad p_1 = \frac{p_{01}}{p_{10} + p_{01}}.$$

The probability of a message x_1^n becomes

$$P(x_1^n) = \hat{p}\, p_{00}^{k_{00}} p_{01}^{k_{01}} p_{10}^{k_{10}} p_{11}^{k_{11}},$$

where $\hat{p} = p_0$ if $x_0 = 0$ and $\hat{p} = p_1$ if $x_0 = 1$ and k_{ij} is the number of $k \in \{1, 2, \ldots, n-1\}$ such that $(x_k, x_{k+1}) = (i, j)$. Note that $k_{00} + k_{01} + k_{10} + k_{11} = n - 1$ and that $k_{01} = k_{10}$ if $x_1 = x_n$ and $k_{01} = k_{10} \pm 1$ if $x_1 \neq x_n$ (cf. [13, 31]).

The following result is an analogue of Theorem 2 that we also prove in Section 3.

Theorem 4 *Consider a stationary binary Markov source $\mathcal{S} = \{P\}$ with known transition matrix $\mathbf{P}$. If*

$$\lg \frac{p_{00}}{\sqrt{p_{10}p_{01}}} \quad \text{or} \quad \lg \frac{p_{11}}{\sqrt{p_{10}p_{01}}}$$

is irrational, then as $n \to \infty$,

$$R_n^*(P) = -\frac{\log\log 2}{\log 2} + o(1) = 0.5287\ldots + o(1)$$

where the term $o(1)$ in general cannot be improved

A generalization of Theorem 3 to Markov sources $\mathcal{M}_1$ (for which the transition matrix $\mathbf{P}$ is unknown) is presented below.

Theorem 5 *Let $\mathcal{M}_1$ be the set of Markov sources over a binary alphabet. Then*

$$R_n^*(\mathcal{M}_1) = \lg n + \lg\left(\frac{8}{\pi}\sum_{j\geq 0}\frac{(-1)^j}{(2j+1)^2}\right) - \frac{\log\log 2}{\log 2} + o(1) \tag{23}$$

as $n \to \infty$.

Remark. It was known since Rissanen [20] that for Markov sources $\mathcal{M}_r$ of order r over an m-ary alphabet

$$\widetilde{R}_n^*(\mathcal{M}_r) \sim \frac{m^r(m-1)}{2}\lg n$$

as $n \to \infty$. However, the computation of the next term is only completed now in the above theorem. Rissanen [21], and Jacquet and Szpankowski [13] computed the constant term of

$\widetilde{R}_n^*(\mathcal{M}_r)$ using probabilistic and analytic tolls, respectively. For example, we know from [13] that for an m-ary alphabet

$$\widetilde{R}_n^*(\mathcal{M}_1) = \frac{m(m-1)}{2} \lg\left(\frac{n}{2\pi}\right) + \lg A_m + O\left(\frac{1}{n}\right)$$

where the constant A_m has the following integral representation

$$A_m = \int_{\mathcal{K}(1)} mF_m(y_{ij}) \prod_i \frac{\sqrt{\sum_j y_{ij}}}{\prod_j \sqrt{y_{ij}}} \prod_{ij} dy_{ij}.$$

In the above, $\mathcal{K}(1) = \{y_{ij} : y_{ij} \geq 0,\ \sum_{ij} y_{ij} = 1,\ \forall i : \sum_j y_{ij} = \sum_j y_{ji}\}$ and $F_m(\cdot)$ is a polynomial expression of degree $m-1$. (More precisely, $F_m(\mathbf{y}) = \sum_k \det_{kk}(1-\mathbf{y}^*)$, and $\mathbf{y}^*$ is the matrix whose ij-th coefficient is $y_{ij}/\sum_{j'} y_{ij'}$, while $\det_{kk}(\mathbf{I}-\mathbf{y})$ is the determinant of the kk-th cofactor of the matrix $\mathbf{I}-\mathbf{y}$.) In particular, for $m=2$ the authors of [13] computed $A_2 = 16 \cdot G$ where $G = \sum_i \frac{(-1)^i}{(2i+1)^2} \approx 0.915965594$ is the Catalan constant.

2.2 The Average Minimax Redundancy and Regrets

In the introduction we concluded that the evaluation of the average minimax redundancy (and regrets) is more challenging due to the difficulties encountered in solving the Bayes problem. Since $\overline{R}(\mathcal{S}) \leq R_n^*(\mathcal{S})$ it is natural to ask whether $\overline{R}(\mathcal{S})$ is asymptotically well approximated by the maximal minimax redundancy $R_n^*(\mathcal{S})$ that we can evaluate quite precisely, as demonstrated in the previous section.

We start with a word of caution. We cannot expect that $\overline{R}_n(\mathcal{S}) \sim R_n^*(\mathcal{S})$ for all classes of sources $\mathcal{S}$ since even for the simplest class $\mathcal{S} = \{P\}$ consisting of one distribution we know that this asymptotic equivalence is violated. Indeed, the result from Theorem 2 should be compared with the average redundancy of the Huffman code. In particular, in [27] it is proved that for a binary memoryless source with known probability p the average redundancy $\overline{R}_n^H$ of the Huffman code is asymptotically equal to

$$\overline{R}_n^H = \begin{cases} \frac{3}{2} - \frac{1}{\ln 2} + o(1) \approx 0.057304 & \text{if} \quad \lg\frac{1-p}{p} \quad \text{irrational} \\ \\ \frac{3}{2} - \frac{1}{M}\left(\langle \beta M n\rangle - \frac{1}{2}\right) - \frac{1}{M(1-2^{-1/M})} 2^{-\langle n\beta M\rangle/M} + O(\rho^n) & \text{if} \quad \lg\frac{1-p}{p} = \frac{N}{M} \end{cases}$$

where $\rho < 1$, $\beta = -\lg(1-p)$ and N, M are integers such that $\gcd(N,M) = 1$.

Nevertheless, we put forward the following conjecture that we shall verify and slightly modify at the end of this section. Since, as observed above, the average and the maximal redundancies differ at least on order $O(1)$, we shall work from now on with the continuous average minimax redundancy $\widetilde{R}_n(\mathcal{S})$ knowing that $\overline{R}_n(\mathcal{S}) \leq \widetilde{R}_n(\mathcal{S}) \leq \overline{R}_n(\mathcal{S}) + 1$.

Conjecture *For a class of sources $\mathcal{S}$ with at least one unknown parameter the average minimax redundancy $\overline{R}_n(\mathcal{S})$ and the average minimax $\overline{r}_n(\mathcal{S})$ are asymptotically equivalent to the maximal minimax redundancy $R_n^*(\mathcal{S})$, that is,*

$$R_n^*(\mathcal{S}) \sim \overline{R}_n(\mathcal{S}) \sim \overline{r}_n(\mathcal{S}) \tag{24}$$

for large n.

When trying to establish this conjecture, we realize that it is prudent to restrict the class of sources to those for which the maximum likelihood distribution Q^* belongs to a convex hull of $\mathcal{S}$, where the convex hull of $\mathcal{S}$ is just the set of all finite convex combinations of elements of $\mathcal{S}$. (We assume no topology on the set of probability measures.) We formulate it as the following postulate.

(H) The maximum likelihood distribution Q^* can be represented as a linear combination of distributions from $\mathcal{S}$, that is, for $P_1, P_2, \ldots, P_N \in \mathcal{S}$

$$Q^* = \sum_{i=1}^{N} \alpha_i P_i \tag{25}$$

where $\alpha_i \geq 0$ and $\sum_{i=1}^{N} \alpha_i = 1$.

Under this assumption we prove in Section 4 the following crucial lemma that will allow us to establish some relationships between the maximal redundancy and the average redundancy and regrets.

Lemma 2 *Let $\mathcal{S}$ be a subset of probability distributions P on $\mathcal{A}^n$. Then for all probability distributions $\widetilde{Q}$ contained in the convex hull of $\mathcal{S}$ we have*

$$\inf_{Q} \sup_{P \in \mathcal{S}} \left(\sum_{x \in X} P(x) \lg \frac{\widetilde{Q}(x)}{Q(x)} \right) = 0.$$

Now we are equipped with all the necessary tools to state and prove our first main result of this section. We recall that $d_n(\mathcal{S}) = \sum_{x_1^n} \sup_{P \in \mathcal{S}} P(x_1^n)$ and $\widetilde{R}_n^*(\mathcal{S}) = \lg d_n(\mathcal{S})$.

Theorem 6 (i) [Upper Bound] *For any set of probability distributions $\mathcal{S}$ on $\mathcal{A}^n$, we have*

$$\widetilde{R}_n(\mathcal{S}) \leq \lg d_n(\mathcal{S}) - \inf_{P \in \mathcal{S}} \left(\sum_{x_1^n} P(x_1^n) \lg \frac{\sup_{P \in \mathcal{S}} P(x_1^n)}{P(x_1^n)} \right). \tag{26}$$

(ii) [Lower Bound] *If the hypothesis (H) holds, that is, if Q^* is contained in the convex hull of $\mathcal{S}$, then*

$$\widetilde{R}_n(\mathcal{S}) \geq \lg d_n(\mathcal{S}) - \sup_{P \in \mathcal{S}} \left(\sum_{x_1^n} P(x_1^n) \lg \frac{\sup_{P \in \mathcal{S}} P(x_1^n)}{P(x_1^n)} \right). \tag{27}$$

Remark If one does not know that Q^* is contained in the convex hull of $\mathcal{S}$ but it is known that there exists a probability distribution $\widetilde{Q}$ in the convex hull of $\mathcal{S}$ such that

$$\max_{x_1^n} \left| \lg \frac{Q^*(x_1^n)}{\widetilde{Q}(x_1^n)} \right| \leq C$$

then one gets the weaker lower bound

$$\widetilde{R}_n(\mathcal{S}) \geq \lg d_n(\mathcal{S}) - C - \sup_{P \in \mathcal{S}} \left(\sum_{x_1^n} P(x_1^n) \lg \frac{\sup_{P \in \mathcal{S}} P(x_1^n)}{P(x_1^n)} \right).$$

Proof. For part (i) we just observed trivially

$$\begin{aligned} \widetilde{R}_n(\mathcal{S}) &= \inf_Q \sup_{P \in \mathcal{S}} D(P||Q) \leq \sup_{P \in \mathcal{S}} D(P||Q^*) \\ &= \lg d_n(\mathcal{S}) - \inf_{P \in \mathcal{S}} \left(\sum_{x_1^n} P(x_1^n) \lg \frac{\sup_{P \in \mathcal{S}} P(x_1^n)}{P(x_1^n)} \right). \end{aligned}$$

For the lower bound (ii) we need to use Lemma 2, hence we need to assume (H). Then we have

$$\begin{aligned} \widetilde{R}_n(\mathcal{S}) &= \inf_Q \sup_{P \in \mathcal{S}} D(P||Q) \\ &= \lg d_n(\mathcal{S}) + \inf_Q \sup_{P \in \mathcal{S}} \left(\sum_{x_1^n} P(x_1^n) \lg \frac{Q^*(x_1^n)}{Q(x_1^n)} - \sum_{x_1^n} P(x_1^n) \lg \frac{\sup_{P \in \mathcal{S}} P(x_1^n)}{P(x_1^n)} \right) \\ &\overset{\text{Lemma 2}}{\geq} \lg d_n(\mathcal{S}) - \sup_{P \in \mathcal{S}} \left(\sum_{x_1^n} P(x_1^n) \lg \frac{\sup_{P \in \mathcal{S}} P(x_1^n)}{P(x_1^n)} \right), \end{aligned}$$

as needed. ∎

In view of the above our conjecture holds if

$$c_n(\mathcal{S}) := \sup_{P \in \mathcal{S}} \left(\sum_{x_1^n} P(x_1^n) \lg \frac{\sup_{P \in \mathcal{S}} P(x_1^n)}{P(x_1^n)} \right) = o(\lg d_n(\mathcal{S})). \tag{28}$$

We will verify this condition below for memoryless and Markovian sources, but indicate that it may not be satisfied for renewal sources.

Interestingly we have a stronger result for the minimax regret $\overline{r}(\mathcal{S})$ that basically shows that the conjecture is true under postulate (H).

Theorem 7 *Let hypothesis (H) hold. Then*

$$\widetilde{r}_n(\mathcal{S}) = \lg d_n(\mathcal{S}) = \widetilde{R}_n^*(\mathcal{S}). \tag{29}$$

Proof. We start with an upper bound that actually holds without assumption (H). We have for any $\mathcal{S}$

$$\begin{aligned}
\widetilde{r}_n(S) &= \lg d_n(\mathcal{S}) + \inf_Q \sup_{P\in\mathcal{S}} \left(\sum_{x_1^n} P(x_1^n) \lg \frac{Q^*(x_1^n)}{Q(x_1^n)} \right) \\
&\leq \lg d_n(\mathcal{S}) + \sup_{P\in\mathcal{S}} \left(\sum_{x_1^n} P(x_1^n) \lg \frac{Q^*(x_1^n)}{Q^*(x_1^n)} \right) \\
&= \lg d_n(\mathcal{S}).
\end{aligned}$$

For the lower bound we need to use Lemma 2 that requires (H). We proceed as follows

$$\begin{aligned}
\widetilde{r}_n(\mathcal{S}) &= \lg d_n(\mathcal{S}) + \inf_Q \sup_{P\in\mathcal{S}} \left(\sum_{x_1^n} P(x_1^n) \lg \frac{Q^*(x_1^n)}{Q(x_1^n)} \right) \\
&= \lg d_n(\mathcal{S})
\end{aligned}$$

which proves the theorem. ∎

Finally, for the maxmin regret $\underline{r}_n(\mathcal{S})$ we also can establish a precise result.

Theorem 8 *Let Q^* be defined as (15) and let $\overline{R}_n^H(P)$ be the average minimax redundancy for the Huffman code for the distribution P. Then*

$$\begin{aligned}
\underline{r}_n(\mathcal{S}) &= \lg d_n(\mathcal{S}) + \sup_{P\in\mathcal{S}} (\overline{R}_n^H(P) - D(P||Q^*)) \\
&= \lg d_n(\mathcal{S}) - \inf_{P\in\mathcal{S}} D(P||Q^*) + O(1) \\
&= \sup_{P\in\mathcal{S}} \left(\sum_{x_1^n} P(x_1^n) \lg \frac{\sup_{P\in\mathcal{S}} P(x_1^n)}{P(x_1^n)} \right) + O(1).
\end{aligned}$$

If $Q^* \in \mathcal{S}$, then $\underline{r}_n(\mathcal{S}) = \lg d_n(\mathcal{S}) + O(1) = R_n^*(\mathcal{S}) + O(1)$.

Proof. The calculations are straightforward:

$$\begin{aligned}
\underline{r}_n(\mathcal{S}) &= \sup_{P\in\mathcal{S}} \min_{C_n\in\mathcal{C}} \sum_{x_1^n} P(x_1^n)[L(C_n, x_1^n) + \lg \sup_{P\in\mathcal{S}} P(x_1^n)] \\
&= \lg d_n(\mathcal{S}) + \sup_{P\in\mathcal{S}} \min_{C_n} \sum_{x_1^n} P(x_1^n)[L(C_n, x_1^n) + \lg Q^*(x_1^n)] \\
&= \lg d_n(\mathcal{S}) + \sup_{P\in\mathcal{S}} \min_{C_n} \sum_{x_1^n} P(x_1^n)[L(C_n, x_1^n) + \lg Q^*(x_1^n) + \lg P(x_1^n) - \lg P(x_1^n)] \\
&= \lg \left(\sum_{x_1^n\in\mathcal{A}^n} \sup_{P\in\mathcal{S}} P(x_1^n) \right) + \sup_{P\in\mathcal{S}} \min_{C_n} \sum_{x_1^n} P(x_1^n)[L(C_n, x_1^n) + \lg P(x_1^n)] - \sum_{x_1^n} P(x_1^n) \lg \left(\frac{P(x_1^n)}{Q^*(x_1^n)} \right)
\end{aligned}$$

$$
\begin{aligned}
&= \lg d_n(\mathcal{S}) + \sup_{P\in\mathcal{S}}(\overline{R}_n^H(P) - D(P||Q^*)) \\
&= \lg d_n(\mathcal{S}) - \inf_{P\in\mathcal{S}} D(P||Q^*) + O(1) \\
&= \sup_{P\in\mathcal{S}} \left(\sum_{x_1^n} P(x_1^n) \lg \frac{\sup_{P\in\mathcal{S}} P(x_1^n)}{P(x_1^n)} \right) + O(1).
\end{aligned}
$$

which proves the desired result. ∎

In view of the above results, the conjecture is true for the minimax regret provided the postulate (H) holds. In fact, we just proved that under (H)

$$\overline{r}_n(\mathcal{S}) = R_n^*(\mathcal{S}) + O(1).$$

Furthermore, we have

$$\underline{r}_n(\mathcal{S}) = c_n(\mathcal{S}) \geq R_n^*(\mathcal{S}) - \overline{R}_n(\mathcal{S}) + O(1).$$

But, if $Q^* \in \mathcal{S}$, then $\underline{r}_n(\mathcal{S}) = R_n^*(\mathcal{S}) + O(1)$.

On the other hand, in order to verify the conjecture for the average minimax redundancy we need to evaluate $c_n(\mathcal{S})$ defined in (28). In Section 4 we do it for memoryless sources $\mathcal{M}_0$ and Markov sources $\mathcal{M}_r$. In particular, for binary alphabet we prove the following two results.

Theorem 9 *For a class of binary memoryless sources $\mathcal{M}_0$ the following holds*

$$\overline{R}_n(\mathcal{M}_0) = R_n^*(\mathcal{M}_0) + O(1) = \frac{1}{2}\lg n + O(1). \tag{30}$$

as $n \to \infty$.

Theorem 10 *Let $\mathcal{M}_1$ denote the set of all Markov sources over a binary alphabet. Then*

$$\overline{R}_n(\mathcal{M}_1) = R_n^*(\mathcal{M}_1) + O(1) = \lg n + O(1). \tag{31}$$

as $n \to \infty$.

As already discussed above, generalizations of the above to m-ary alphabet and r-th order Markov is quite straightforward although may be technically involved.

Finally, we deal with the renewal process $\mathcal{R}_0$ introduced by Csiszár and Shields [6]: Let $T_1, T_2 \ldots$ be a sequence of i.i.d. positive–valued random variables with common distribution $R(j) = \Pr\{T_1 = j\}$. Throughout we assume that $\mathbf{E}[T_1] < \infty$. With such a renewal process there is associated a *binary renewal sequence* that is a $0,1$-sequence in which the 1's occur exactly at the renewal epochs $T_1,\ T_1 + T_2$, etc. Accordingly , we start the renewal sequence

with a 1 put at the zeroth position followed by a run of zeros. In passing we observe that since $P(x_1^n)$ and $R(j)$ determine one another, we identify the underlying probability measure P defined on $\{0,1\}^\infty$ with the distribution R that it induces. For such a renewal source $\mathcal{R}_0$, Csiszár and Shields [6] proved that the average minimax redundancy $\overline{R}_n(\mathcal{R}_0)$ is of order $\sqrt{n}$. More precisely, there exist two constants $C_1, C_2 > 0$ such that

$$C_1\sqrt{n} \le \overline{R}_n(\mathcal{R}_0) \le C_2\sqrt{n}.$$

Flajolet and Szpankowski [11] recently showed that the worst case minimax redundancy $R_n^*(\mathcal{R}_0)$ of this process is

$$R_n^*(\mathcal{R}_0) = \left(\frac{2}{\log 2}\sqrt{\frac{\pi^2}{6}-1}\right)\cdot\sqrt{n} + O(\log n) \approx 2.317\cdot\sqrt{n} + O(\log n),$$

thus $C_2 \le 2.317\ldots$. The question is whether $C_1 = C_2$, and hence $\overline{R}_n(\mathcal{R}_0) \sim R_n^*(\mathcal{R}_0)$. Unfortunately, this seems to be not true as we prove in Section 4 the following lemma.

Lemma 3 *For all probability distributions $P \in \mathcal{R}_0$ we have*

$$c_n(\mathcal{R}_0) = \sup_{P\in\mathcal{R}_0}\left(\sum_{x_1^n} P(x_1^n)\lg\frac{\sup_{P\in\mathcal{R}_0} P(x_1^n)}{P(x_1^n)}\right) \le \frac{1+\frac{2}{e}}{\log 2\sqrt{\frac{\sqrt{2}}{3}+\frac{2}{e}}}\cdot\sqrt{n} + O(1) \approx 2.278\cdot\sqrt{n}.$$

Theorem 6 and Lemma 3 suggest (if we verify hypothesis (H)) that

$$0.039\ldots\cdot\sqrt{n} \le \overline{R}_n(\mathcal{R}_0) \le 2.317\ldots\cdot\sqrt{n}.$$

A direct implication of this result is that our conjecture seems to be not true for $\overline{R}_n(\mathcal{S})$ for general class of processes $\mathcal{S}$. Therefore, we modify it slightly, and expect that for general class of sources $\mathcal{S}$ the following holds

$$R_n^*(\mathcal{S}) \asymp \overline{R}_n(\mathcal{S}),$$

where $a_n \asymp b_n$ if there are constants $c_1, c_2 > 0$ such that $c_1 a_n \le b_n \le c_2 a_n$ for large n.

3 Analysis of the Worst Case Minimax Redundancy

In this section we prove Theorems 1 – 5. In most of the proofs we apply analytic techniques of analysis of algorithms that can be reviewed from [28].

3.1 Proof of Theorem 1

We recall that we are to prove the following decomposition of the worst case minimax redundancy $R_n^*(\mathcal{S})$:

$$R_n^*(\mathcal{S}) = \lg d_n(\mathcal{S}) + 1 - t_0$$

where t_0 is the largest $t \in T$ for which the Kraft inequality (20) holds.

It turns out that Theorem 1 follows directly from Lemma 1 and the next result in which we consider the following optimization problem: For a given source P find a prefix code that minimizes the maximal redundancy $R_n^*(P)$, that is,

$$R_n^*(P) = \min_{C_n \in \mathcal{C}} \max_{x_1^n} [L(C_n, x_1^n) + \lg P(x_1^n)]. \tag{32}$$

The next lemma proves that a generalized Shannon codes solves the above optimization problem.

Lemma 4 *If the probability distribution P is dyadic, i.e. $\lg P(x_1^n) \in \mathbb{Z}$ for all $x_1^n \in \mathcal{A}^n$, then $R_n^*(P) = 0$. Otherwise, let $T = T(P) := \{\langle -\lg P(x_1^n)\rangle : x_1^n \in \mathcal{A}^n\}$ and $t_0 \in T$ be the largest t such that*

$$\sum_{x_1^n \in \mathcal{L}_t} P(x_1^n) 2^{\langle -\lg P(x_1^n)\rangle} + \frac{1}{2} \sum_{x_1^n \in \mathcal{U}_t} P(x_1^n) 2^{\langle -\lg P(x_1^n)\rangle} \le 1, \tag{33}$$

where

$$\mathcal{L}_t := \{x_1^n \in \mathcal{A}^n : \langle -\lg P(x_1^n)\rangle < t\}$$

and

$$\mathcal{U}_t := \{x_1^n \in \mathcal{A}^n : \langle -\lg P(x_1^n)\rangle \ge t\}.$$

Then

$$R_n^*(P) = 1 - t_0 \tag{34}$$

and the optimum is obtained for a generalized Shannon code with $\mathcal{L} = \mathcal{L}_{t_0}$ and $\mathcal{U} = \mathcal{U}_{t_0}$.

Proof. If P is dyadic then the numbers $l(x_1^n) := -\lg P(x_1^n)$ are positive integers satisfying

$$\sum_{x_1^n} 2^{-l(x_1^n)} = 1.$$

Kraft's inequality holds and consequently there exists a (prefix) code C_n with $L(C_n, x_1^n) = l(x_1^n) = -\lg P(x_1^n)$, and this implies C_n with $L(C_n, x_1^n) = l(x_1^n) = -\lg P(x_1^n)$ and $R_n^*(P) = 0$.

Now assume that P is not dyadic and let C_n^* denote the set of optimal codes, i.e.

$$\mathcal{C}_n^* = \{C_n \in \mathcal{C} : R_n^*(C_n, P) = R_n^*(P)\}.$$

The idea of the proof is to establish several properties of an optimal code. In particular, we will show that there exists an optimal code $C_n^* \in \mathcal{C}_n^*$ with the following two properties:

(i)

$$\lfloor -\lg P(x_1^n) \rfloor \le L(C_n^*, x_1^n) \le \lceil -\lg P(x_1^n) \rceil \tag{35}$$

(ii) There exists $s_0 \in [0,1]$ such that

$$L(C_n^*, x_1^n) = \lfloor \lg 1/P(x_1^n) \rfloor \quad \text{if} \quad \langle \lg 1/P(x_1^n) \rangle < s_0 \tag{36}$$

and

$$L(C_n^*, x_1^n) = \lceil \lg 1/P(x_1^n) \rceil \quad \text{if} \quad \langle \lg 1/P(x_1^n) \rangle \ge s_0. \tag{37}$$

Observe that w.l.o.g. we may assume that $s_0 = 1 - R_n^*(P)$. Thus, in order to compute $R_n^*(P)$ we just have to consider codes satisfying (36) and (37). As already mentioned, (33) is just Kraft's inequality for codes of that kind. The optimal choice is $t = t_0$ which also equals s_0. Consequently $R_n^*(P) = 1 - t_0$.

In view of the above, it suffices to prove properties (i) and (ii). Assume that C_n^* is an optimal code. First of all, the upper bound in (35) is obviously satisfied for C_n^*. Otherwise we would have

$$\max_{x_1^n} [L(C_n^*, x_1^n) + \lg P(x_1^n)] > 1$$

which contradicts Shtarkov's bound (14). Second, if there exists x_1^n such that $L(C_n^*, x_1^n) < \lfloor \lg 1/P(x_1^n) \rfloor$, then (in view of Kraft's inequality) we can modify this code to a code $\widetilde{C}_n^*$ with

$$\begin{aligned} L(\widetilde{C}_n^*, x_1^n) &= \lceil \lg 1/P(x_1^n) \rceil \quad \text{if } L(C_n^*, x_1^n) = \lceil \lg 1/P(x_1^n) \rceil, \\ L(\widetilde{C}_n^*, x_1^n) &= \lfloor \lg 1/P(x_1^n) \rfloor \quad \text{if } L(C_n^*, x_1^n) \le \lfloor \lg 1/P(x_1^n) \rfloor. \end{aligned}$$

By construction $R_n^*(\widetilde{C}_n^*, P) = R_n^*(C_n^*, P)$. Thus, $\widetilde{C}_n^*$ i s optimal, too. This proves (i).

Now consider an optimal code C_n^* satisfying (35) and let $\widetilde{x}_1^n$ be a word with $R_n^*(P) = 1 - \langle -\lg P(\widetilde{x}_1^n) \rangle$. Thus, $L(C_n^*, x_1^n) = \lfloor \lg 1/P(x_1^n) \rfloor$ for all x_1^n with $\langle -\lg P(x_1^n) \rangle < \langle -\lg P(\widetilde{x}_1^n) \rangle$. This proves (36) with $s_0 = \langle -\lg P(\widetilde{x}_1^n) \rangle$. Finally, if (37) is not satisfied, then (in view of Kraft's inequality) we can modify this code to a code $\widetilde{C}_n^*$ with

$$\begin{aligned} L(\widetilde{C}_n^*, x_1^n) &= \lceil \lg 1/P(x_1^n) \rceil \quad \text{if } \langle \lg 1/P(x_1^n) \rangle \ge s_0, \\ L(\widetilde{C}_n^*, x_1^n) &= \lfloor \lg 1/P(x_1^n) \rfloor \quad \text{if } \langle \lg 1/P(x_1^n) \rangle < s_0. \end{aligned}$$

By construction $R_n^*(\widetilde{C}_n^*, P) = R_n^*(C_n^*, P)$. Thus, $\widetilde{C}_n^*$ is optimal, too. This proves (ii) and the lemma. ∎

3.2 Proof of Theorem 2

We now consider a binary memoryless source $P_p(x_1^n) = p^k(1-p)^{n-k}$ where p is a *given* probability of generating a "0". Our goal is to estimate precisely the maximal redundancy of the optimal generalized code just constructed in Lemma 4 of the previous section.

The proof of Theorem 2, as well as some others in this section, rely heavily on properties of sequences modulo 1 that we review first (the reader is referred to [9, 27] for more detailed exposition). We start with a definition of P-uniformly distributed sequences modulo 1.

Definition 1 (P-u.d. mod 1) *A sequence $x_n \in \mathbb{R}$ is said to be P-uniformly distributed modulo 1 (P-u.d. mod 1) with respect to the set of probability distributions $P = \{(p_{n,k})_{k\geq 0} : n \geq 0\}$ if*

$$\lim_{n\to\infty} \sum_{k\geq 0} p_{n,k} I_A(\langle x_k \rangle) = \lambda(A) \tag{38}$$

holds uniformly for every interval $A \subset [0,1]$, where $I_A(x_n)$ is the characteristic function of A (i.e., it equals 1 if $x_n \in A$ and 0 otherwise) and $\lambda(A)$ is the Lebesgue measure of A.

In particular, we will use the probability distributions $p_{n,k} = \binom{n}{k} p^k (1-p)^{n-k}$ and call a sequence x_n Bernoulli distributed mod 1 if x_n is P-u.d. mod 1 for this particular P.

The following result summarizes the main property of P-u.d. modulo 1 sequences. It provides the leading term of asymptotics for sums like $\sum_k p_{n,k} f(\langle x_k + y \rangle)$, where x_k is P-u.d. mod 1 and y is a shift and f is a Riemann integrable function.

Theorem 11 *Suppose that the sequence x_n is P-uniformly distributed modulo 1. Then for every Riemann integrable function $f : [0,1] \to \mathbb{R}$*

$$\lim_{n\to\infty} \sum_{k\geq 0} p_{n,k} f(\langle x_k + y \rangle) = \int_0^1 f(t)\, dt, \tag{39}$$

where the convergence is uniform for all shifts $y \in \mathbb{R}$.

To apply Theorem 11, one needs easy criteria to verify whether a sequence x_k is P-u.d. mod 1. Fortunately, such a result exists and is due to H. Weyl.

Theorem 12 (Weyl, 1916) *A sequence x_n is P-u.d. mod 1 if and only if*

$$\lim_{n\to\infty} \sum_{k\geq 0} p_{n,k} e^{2\pi i m x_k} = 0 \tag{40}$$

holds for all $m \in \mathbb{Z} \setminus \{0\}$.

Now we are in position to prove the first part of Theorem 2, that is, *if* $\lg \frac{1-p}{p}$ *is irrational, then as* $n \to \infty$

$$R_n^*(P_p) = -\frac{\log\log 2}{\log 2} + o(1) = 0.5287\ldots + o(1). \tag{41}$$

Set

$$\alpha_p = \lg \frac{1-p}{p}, \qquad \beta_p = \lg \frac{1}{1-p}.$$

Then

$$-\lg(p^k(1-p)^{n-k}) = \alpha_p k + \beta_p n.$$

Since α_p is irrational we know from [9, 27] that $\langle \alpha_p n \rangle$ is a Bernoulli-u.d modulo 1 sequence (it also easy to prove directly by noting that Weyl's criterion holds), and therefore by Theorem 11 we have

$$\lim_{n\to\infty} \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} f(\langle \alpha_p k + \beta_p n \rangle) = \int_0^1 f(x)\, dx. \tag{42}$$

Now set $f_{s_0}(x) = 2^x$ for $0 \le x < s_0$ and $f_{s_0}(x) = 2^{x-1}$ for $s_0 \le x \le 1$. We find

$$\lim_{n\to\infty} \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} f_{s_0}(\langle \alpha k + \beta n \rangle) = \frac{2^{s_0-1}}{\log 2}.$$

In particular, for

$$s_0 = 1 + \frac{\log\log 2}{\log 2} = 0.4712\ldots$$

we obtain $\int_0^1 f(x)\, dx = 1$. This implies that

$$\lim_{n\to\infty} R_n^*(P_p) = 1 - s_0 = 0.5287\ldots$$

which proves (41) and the first part of Theorem 2.

Now we establish the second part of Theorem 2, that is, *if* $\lg \frac{1-p}{p} = \frac{N}{M}$ *is rational and non-zero (with coprime integers* N, M*) then, as* $n \to \infty$

$$R_n^*(P_p) = -\frac{\lfloor M \lg(M(2^{1/M}-1)) - \langle Mn \lg 1/(1-p) \rangle \rfloor + \langle Mn \lg 1/(1-p) \rangle}{M} + o(1). \tag{43}$$

As in [27] we first observe that

$$\begin{aligned} \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} f(\langle \alpha_p k + \beta_p n \rangle) &= \frac{1}{M} \sum_{m=0}^{M-1} f\left(\left\langle \frac{mN}{M} + \beta_p n \right\rangle \right) + o(1) \\ &= \frac{1}{M} \sum_{m=0}^{M-1} f\left(\frac{m + \langle M\beta_p n \rangle}{M} \right) + o(1). \end{aligned}$$

As before, we use $f_{s_0}(x)$, where s_0 is of the form

$$s_0 = \frac{m_0 + \langle M\beta_p n \rangle}{M}$$

and choose m_0 maximal such that

$$\begin{aligned}\frac{1}{M}\sum_{m=0}^{M-1} f_{s_0}\left(\frac{m+\langle M\beta_p n\rangle}{M}\right) &= \frac{2^{\langle M\beta_p n\rangle}/M}{M}\left(\sum_{m=0}^{m_0-1} 2^{m/M}+\sum_{m=m_0}^{M-1} 2^{m/M-1}\right)\\ &= \frac{2^{(\langle M\beta_p n\rangle+m_0)/M-1}}{M(2^{1/M}-1)}\\ &\le 1.\end{aligned}$$

Thus

$$m_0 = M + \lfloor M\lg(M(2^{1/M}-1)) - \langle Mn\lg 1/(1-p)\rangle\rfloor,$$

and consequently

$$\begin{aligned}R_n^*(P_p) &= 1-s_0+o(1)\\ &= 1-\frac{m_0+\langle M\beta_p n\rangle}{M}+o(1)\\ &= -\frac{\lfloor M\lg(M(2^{1/M}-1))-\langle Mn\lg 1/(1-p)\rangle\rfloor+\langle Mn\beta_p\rangle}{M}+o(1).\end{aligned}$$

This completes the proof of Theorem 2.

3.3 Proof of Theorem 3

Now we consider a class of binary memoryless sources $P_p(x_1^n) = p^k(1-p)^{n-k}$ such that $p\in[a,b]$ for some $0\le a<b\le 1$ and prove Theorem 3 which states that for $\mathcal{M}_0^{a,b} = \{P_p : a\le p\le b\}$

$$R_n^*(\mathcal{M}_0^{a,b}) = \frac{1}{2}\lg n + \lg\, C_{a,b} - \frac{\log\log 2}{\log 2} + O\left(\frac{\log n}{n^{1/9}}\right), \tag{44}$$

where $C_{a,b} = \sqrt{\frac{2}{\pi}}(\arcsin\sqrt{b}-\arcsin\sqrt{a})$.

After observing that

$$Q^*(x_1^n) = \sup_{p\in[a,b]} p^k(1-p)^{n-k} = \begin{cases} a^k(1-a)^{n-k} & \text{for } 0\le k<na,\\ \left(\frac{k}{n}\right)^k\left(1-\frac{k}{n}\right)^{n-k} & \text{for } na\le k\le nb,\\ b^k(1-b)^{n-k} & \text{for } nb<k\le n.\end{cases}$$

we express $d_n(\mathcal{M}_0^{a,b}) = \sum_{x_1^n}\sup_P P(x_1^n)$ as follows

$$\begin{aligned}d_n := d_n(\mathcal{M}_0^{a,b}) &= \sum_{k<na}\binom{n}{k}a^k(1-a)^{n-k}+\sum_{na\le k\le nb}\binom{n}{k}\left(\frac{k}{n}\right)^k\left(1-\frac{k}{n}\right)^{n-k}\\ &+ \sum_{k>nb}\binom{n}{k}b^k(1-b)^{n-k}.\end{aligned}$$

It is easy to show that

$$\sum_{k<na}\binom{n}{k}a^k(1-a)^{n-k}=\frac{1}{2}+O(n^{-1/2}),$$

and

$$\sum_{k>nb}\binom{n}{k}b^k(1-b)^{n-k}=\frac{1}{2}+O(n^{-1/2}).$$

Furthermore, we have (uniformly for $an\le k\le bn$) by Stirling's formula

$$\binom{n}{k}\left(\frac{k}{n}\right)^k\left(1-\frac{k}{n}\right)^{n-k}=\frac{1}{\sqrt{2\pi}}\sqrt{\frac{n}{k(n-k)}}+O(n^{-3/2}).$$

Consequently

$$\begin{aligned}\sum_{na\le k\le nb}\binom{n}{k}\left(\frac{k}{n}\right)^k\left(1-\frac{k}{n}\right)^{n-k} &= \sqrt{\frac{n}{2\pi}}\int_a^b\frac{dx}{\sqrt{x(1-x)}}+O(n^{-1/2})\\ &= 2\sqrt{\frac{n}{2\pi}}(\arcsin\sqrt{b}-\arcsin\sqrt{a})+O(n^{-1/2}),\end{aligned}$$

which gives

$$d_n=C_{a,b}\sqrt{n}+1+O(n^{-1/2})$$

and

$$\widetilde{R}_n^*(\mathcal{M}_0^{a,b})=\lg d_n=\frac{1}{2}\lg n+\lg C_{a,b}+O(n^{-1/2}).$$

Next, we need to compute $R_n^*(Q^*)$, as we did in the previous section for a *given* distribution. We recall that by Theorem 1 evaluation of the redundancy $R_n^*(Q^*)$ reduces to a verification of the Kraft's inequality, which in our case becomes

$$\sum_{x_1^n}Q^*(x_1^n)f_{s_0}(\langle-\lg Q^*(x_1^n)\rangle)$$

where $f_{s_0}(x)=2^{-\langle s_0-x\rangle+s_0}$ for some $0\le s_0<1$. Thus, the problem is to evaluate the following sum

$$\begin{aligned}&\sum_{k=0}^n\binom{n}{k}\frac{\sup_{p\in[a,b]}p^k(1-p)^{n-k}}{d_n}f_{s_0}\left(-\lg\left(\sup_{p\in[a,b]}p^k(1-p)^{n-k}\right)+\lg d_n\right)\\ &\quad=\frac{1}{d_n}\sum_{k<an}\binom{n}{k}a^k(1-a)^{n-k}f_{s_0}(-\lg(a^k(1-a)^{n-k})+\lg d_n)\\ &\quad+\frac{1}{d_n}\sum_{an\le k\le bn}\binom{n}{k}\left(\frac{k}{n}\right)^k\left(1-\frac{k}{n}\right)^{n-k}f_{s_0}\left(-\lg\left(\left(\frac{k}{n}\right)^k\left(1-\frac{k}{n}\right)^{n-k}\right)+\lg d_n\right)\\ &\quad+\frac{1}{d_n}\sum_{k>bn}\binom{n}{k}b^k(1-b)^{n-k}f_{s_0}(-\lg(b^k(1-b)^{n-k})+\lg d_n)\\ &\quad=S_1+S_2+S_3.\end{aligned}$$

Obviously, the first and third sum can be estimated by

$$S_1 = O(n^{-1/2}) \quad \text{and} \quad S_3 = O(n^{-1/2}).$$

Thus, it remains to study S_2.

We prove the following lemma.

Lemma 5 *For every (Riemann integrable) function $f : [0,1] \to \mathbf{R}$ of bounded variation $V_0^1(f)$ and for every sequence $x_{n,k}$, $an \le k \le bn$, such that*

$$x_{n,k} = k \lg k + (n-k) \lg(n-k) + c_n,$$

where c_n is an arbitrary sequence, we have

$$\frac{1}{d_n} \sum_{an \le k \le bn} \binom{n}{k} \left(\frac{k}{n}\right)^k \left(1 - \frac{k}{n}\right)^{n-k} f(\langle x_{n,k} \rangle) = \int_0^1 f(x)\, dx + O\left(V_0^1(f) \frac{\log n}{n^{1/9}}\right) \tag{45}$$

Remark If $a > 0$ and $b < 1$ then we even have a better error term of the form $O\left(V_0^1(f)(\log n) n^{-1/3}\right)$.

Proof. We first show that $x_{n,k}$ is Q^*-u.d. sequence modulo 1. If view of Theorem 11 this proves (45) except the the error term that we analyze precisely below.

First, we consider the the following exponential sum

$$S := \sum_{an \le k \le cn} e(h(k \lg k + (n-k) \lg(n-k)),$$

where $e(x) = e^{2\pi i x}$, $c \in [a,b]$, and h is an arbitrary non-zero integer. By Van-der-Corput's method (see [16, p. 31]) we know that

$$|S| \ll \frac{|F'(cn) - F'(an)| + 1}{\sqrt{\lambda}},$$

where $\lambda = \min\limits_{an \le y \le cn} |F''(y)| > 0$ and

$$F(y) = h(y \lg y + (n-y) \lg(n-y)).$$

Since $|F'(y)| \ll h \log n$, and $|F''(y)| \gg h/n$ (uniformly for $an \le y \le cn$) we immediately find

$$|S| \ll \log n \sqrt{hn}$$

and consequently

$$\left| \sum_{an \le k \le cn} e(h x_{nk}) \right| \ll \log n \sqrt{hn}.$$

Note that all these estimates are uniform for $c \in [a,b]$. Next we consider the following exponential sum which is in fact the one appearing in the Weyl criterion

$$\widetilde{S} := \sum_{an \le k \le bn} a_{n,k} e(hx_{nk}),$$

where

$$a_{n,k} = \binom{n}{k} \left(\frac{k}{n}\right)^k \left(1 - \frac{k}{n}\right)^{n-k}.$$

By elementary calculations we obtain (uniformly for $an \le k \le bn$) $a_{n,k} \ll \min(k, n-k)^{-1/2}$ and

$$|a_{n,k+1} - a_{n,k}| \ll \min(k, n-k)^{-3/2}.$$

Thus, if $a > 0$ and $b < 1$ we have $a_{n,k} \ll n^{-1/2}$ and $|a_{n,k+1} - a_{n,k}| \ll n^{-3/2}$. Consequently by partial summation

$$\begin{aligned} |\widetilde{S}| &\le a_{n,bn} \left| \sum_{an \le k \le bn} e(hx_{n,k}) \right| \\ &+ \sum_{an \le k < bn} |a_{n,k+1} - a_{n,k}| \left| \sum_{an \le \ell \le k} e(hx_{n,\ell}) \right| \\ &\ll n^{-1/2} \log n \sqrt{hn} + n n^{-3/2} \log n \sqrt{hn} \\ &\ll \sqrt{h}\, \log n. \end{aligned}$$

This means that for every non-zero integer h we have

$$\lim_{n\to\infty} \frac{1}{d_n} \sum_{an \le k \le bn} a_{n,k} e(hx_{n,k}) = O\left(\sqrt{h} \frac{\log n}{\sqrt{n}}\right) \to 0. \tag{46}$$

Therefore, Weyl's criterion holds in that case. (Note that the *shifting* sequence c_n does not chance the absolute value of the exponential sums $\widetilde{S}$.)

Now suppose that $a = 0$ and $b < 1$. Set $\varepsilon = (h/n)^{1/4}$. Then

$$\begin{aligned} |\widetilde{S}| &\le \sum_{k \le \varepsilon n} a_{n,k} + \left| \sum_{\varepsilon n \le k \le bn} a_{n,k} e(hx_{nk}) \right| \\ &\ll \sqrt{\varepsilon n} + n^{-1/2} \log n \sqrt{hn} + n(\varepsilon n)^{-3/2} \log n \sqrt{hn} \\ &\ll (\log n) h^{1/8} n^{3/8} \end{aligned}$$

which proves

$$\lim_{n\to\infty} \frac{1}{d_n} \sum_{an \le k \le bn} a_{n,k} e(hx_{n,k}) = O\left(h^{1/8} \frac{\log n}{n^{1/8}}\right) \to 0 \tag{47}$$

in this case. The final case $a = 1$, $b = 1$ is completely similar.

Now, with help of Erdős-Turán's inequality and Koksma-Hlawka's inequality (see [9]) one gets

$$\left|\frac{1}{d_n}\sum_{an\le k\le bn}\binom{n}{k}\left(\frac{k}{n}\right)^k\left(1-\frac{k}{n}\right)^{n-k} f(\langle x_{n,k}\rangle)-\int_0^1 f(x)\,dx\right|$$
$$\ll V_0^1(f)\cdot\left(\frac{1}{H}+\sum_{h=1}^{H}\frac{1}{h}\left|\frac{1}{d_n}\sum_{an\le k\le bn} a_{n,k}e(hx_{n,k})\right|\right).$$

Thus, if $0<a<b<1$ we set $H=n^{1/3}$ and get from (46)

$$\frac{1}{H}+\sum_{h=1}^{H}\frac{1}{h}\sqrt{h}\frac{\log n}{\sqrt{n}}\ll\frac{\log n}{n^{1/3}}.$$

If $a=1$ or $b=1$ we choose $H=n^{1/9}$ and one obtains

$$\frac{1}{H}+\sum_{h=1}^{H}\frac{1}{h}h^{1/8}\frac{\log n}{n^{1/8}}\ll\frac{\log n}{n^{1/9}}.$$

This proves the lemma. ∎

To complete the proof of Theorem 3 we note that we are now in a similar situation as in the proof of Theorem 2. We apply (45) with $f_{s_0}(x)$ for $s_0=-\log\log 2/\log 2$, and (44) follows. (Note that $f_s(x)$ has variation $V_0^1(f_s)=2<\infty$.)

3.4 Proof of Theorem 4

Our goal in this section is to extend Theorem 2 to (binary) Markov sources. We first assume that the Markov transition matrix $\mathbf{P}=\{p_{ij}\}_{i,j=0}^{1}$ is given. We recall that the stationary distribution is given by

$$p_0=\frac{p_{10}}{p_{10}+p_{01}}\quad\text{and}\quad p_1=\frac{p_{01}}{p_{10}+p_{01}}$$

and

$$P(x_1^n)=\hat{p}\,p_{00}^{k_{00}}p_{01}^{k_{01}}p_{10}^{k_{10}}p_{11}^{k_{11}},\tag{48}$$

where $\hat{p}=p_0$ if $x_0=0$ and $\hat{p}=p_1$ if $x_0=1$ and k_{ij} is the number of $k\in\{1,2,\ldots,n-1\}$ such that $(x_k,x_{k+1})=(i,j)$. Note that $k_{00}+k_{01}+k_{10}+k_{11}=n-1$ and that $k_{01}=k_{10}$ if $x_1=x_n$ and $k_{01}=k_{10}\pm 1$ if $x_1\neq x_n$ (cf. [13, 31]). The latter condition is called the *conservation flow property* and is crucial to determine Markov redundancy and Markov types as discussed in [13].

Here we aim at proving that if

$$\lg\frac{p_{00}}{\sqrt{p_{10}p_{01}}}\quad\text{or}\quad\lg\frac{p_{11}}{\sqrt{p_{10}p_{01}}}\tag{49}$$

is irrational, then

$$R_n^*(P) = -\frac{\log\log 2}{\log 2} + o(1) = 0.5287\ldots + o(1) \tag{50}$$

for a given source P.

Before we proceed with a rigorous proof, let us first propose an intuitive explanation of the above result. Observe that $k_{00} + k_{11} + 2k_{01} = n - 1$, hence

$$\lg P(x_1^n) = k_{00} \lg \frac{p_{00}}{\sqrt{p_{10}p_{01}}} + k_{11} \lg \frac{p_{11}}{\sqrt{p_{10}p_{01}}} + \frac{n-1}{2} \lg(p_{01}p_{10}) + O(1).$$

Therefore, as already seen in the proof of Theorem 2, one expects that the sequence $\lg P(x_1^n)$ is P-u.d. if conditions (49) hold. We prove it below by showing that the Weyl criterion is satisfied.

For $\mathbf{k} = (k_{00}, k_{01}, k_{10}, k_{11})$ let $N_{\mathbf{k}}$ be the number of binary sequences x_1^n of length $n = k_{00} + k_{01} + k_{10} + k_{11} + 1$ with k_{ij} pairs $(x_k, x_{k+1}) = (i, j)$. In fact, $N_{\mathbf{k}}$ is the type of the underlying Markov source, and it was studied extensively before (cf. [13, 31]). In particular, using results from [13, 31] one can compute

$$\begin{aligned} G(z) = \sum_{n\geq 1} \sum_{x_1^n} \widetilde{P}(x_1^n)\, z^n &= \sum_{n\geq 1} \sum_{\mathbf{k}} N_{\mathbf{k}} \hat{p}\, p_{00}^{k_{00}} p_{01}^{k_{01}} p_{10}^{k_{10}} p_{11}^{k_{11}} \\ &= \frac{A(z)}{\det(\mathbf{I} - z\mathbf{P})} \\ &= z\frac{p_0(1 - p_{11}z + p_{01}z) + p_1(1 - p_{00}z + p_{10}z)}{1 - z(p_{00} + p_{11}) + z^2(p_{00}p_{11} - p_{01}p_{10})}, \end{aligned} \tag{51}$$

where $\widetilde{P}(x_1^n)$ is the same as (48) except that p_{ij} are considered to be *complex variables*. Observe that if $p_{00} + p_{01} = p_{10} + p_{11} = 1$, then the right hand side of the above is equal to $z/(1-z)$ as it should.

Let us now evaluate the Weyl sum (40) for our case. Therefore, we replace p_{kl} in (51) by $p_{kl}e(h \lg p_{kl}) = p_{kl}e^{2\pi i h \lg p_{kl}}$ for any integer $h \neq 0$, and must prove that such a sum converges to zero for any $h \neq 0$. We obtain

$$\sum_{n\geq 1} \sum_{x_1^n} P(x_1^n) e(h \lg P(x_1^n))\, z^n = \tag{52}$$

$$= \frac{z p_0 e(h\lg p_0)(1 - p_{11}e(h\lg p_{11})z + p_{01}e(h\lg p_{01})z)}{1 - z(p_{00}e(h\lg p_{00}) + p_{11}e(h\lg p_{11})) + z^2(p_{00}p_{11}e(h\lg(p_{00}p_{11}) - p_{01}p_{10}e(h\lg(p_{01}p_{10})))}$$

$$+ \frac{z p_1 e(h\lg p_1)(1 - p_{00}e(h\lg p_{00})z + p_{10}e(h\lg p_{10})z)}{1 - z(p_{00}e(h\lg p_{00}) + p_{11}e(h\lg p_{11})) + z^2(p_{00}p_{11}e(h\lg(p_{00}p_{11}) - p_{01}p_{10}e(h\lg(p_{01}p_{10})))}.$$

After tedious algebra, one can show that both zeros of the denominator are of absolute value greater than 1 if $\lg(p_{00}/\sqrt{p_{10}p_{01}})$ or $\lg(p_{11}/\sqrt{p_{10}p_{01}})$ is irrational. Therefore, for all integers $h \neq 0$

$$\lim_{n\to\infty} \sum_{x_1^n} P(x_1^n) e(h \lg P(x_1^n)) = 0.$$

Consequently, the Weyl condition holds and by Theorem 11 we conclude that

$$\lim_{n\to\infty} \sum_{x_1^n} P(x_1^n) f(\langle -\lg P(x_1^n)\rangle) = \int_0^1 f(x)\,dx$$

for all Riemann integrable functions $f : [0,1] \to \mathbf{R}$. Applying this relation to functions $f_{s_0}(x)$ defined by $f_{s_0}(x) = 2^x$ for $0 \le x < s_0$ and $f_{s_0}(x) = 2^{x-1}$ for $s_0 \le x \le 1$ we immediately derive (50), as in the proof of Theorem 2.

3.5 Proof of Theorem 5

Finally, we consider a class of binary Markov sources with unknown transition matrix $\mathbf{P}$ and prove Theorem 5, that is, we derive

$$R_n^*(\mathcal{M}_1) = \lg n + \lg\left(\frac{8}{\pi}\sum_{j\ge 0}\frac{(-1)^j}{(2j+1)^2}\right) - \frac{\log\log 2}{\log 2} + o(1).$$

As we already observed in Section 2, the leading terms $\lg n + \lg\left(\frac{8}{\pi}\sum_{j\ge 0}\frac{(-1)^j}{(2j+1)^2}\right) + o(1)$ that correspond to $\lg d_n(\mathcal{M}_1)$ have been recently determined in [13]. Thus we must concentrate on the redundancy $R_n(Q^*)$ of the maximal likelihood distribution Q^*. However, we cannot work directly with Q^*. We have to approximate it appropriately.

Lemma 6 *Suppose that x_1^n is a binary word of length n. Let k_{ij} denote the number of $k \in \{1,2,\dots,n-1\}$ such that $(x_k, x_{k+1}) = (i,j)$ and k_i the number of $k \in \{1,2,\dots,n\}$ such that $x_k = i$ Then there exist two constants $C_1, C_2 > 0$ such that the ratio of*

$$S^{**}(x_1^n) := \frac{\frac{\hat{k}}{\widetilde{k}}}{\frac{k_{01}}{k_{00}+k_{01}} + \frac{k_{10}}{k_{10}+k_{11}}} \left(\frac{k_{00}}{k_{00}+k_{01}}\right)^{k_{00}} \left(\frac{k_{01}}{k_{00}+k_{01}}\right)^{k_{01}} \left(\frac{k_{10}}{k_{10}+k_{11}}\right)^{k_{10}} \left(\frac{k_{11}}{k_{10}+k_{11}}\right)^{k_{11}}$$

(where $\hat{k} = k_{10}$ and $\widetilde{k} = k_{10}+k_{11}$ if $\hat{p} = p_0$ and $\hat{k} = k_{01}$ and $\widetilde{k} = k_{00}+k_{01}$ if $\hat{p} = p_1$) and

$$S^*(x_1^n) = \sup_{p_{ij}} \left(\hat{p}\, p_{00}^{k_{00}} p_{01}^{k_{01}} p_{10}^{k_{10}} p_{11}^{k_{11}}\right)$$

is always contained in the interval $[C_1, C_2]$. Furthermore, there are constant $C_3 > 0, C_4 > 0$ such that for every $\varepsilon > 0$ there exists n_0 such that

$$1 - C_3\varepsilon \le \frac{S^*(x_1^n)}{S^{**}(x_1^n)} \le 1 + C_3\varepsilon \tag{53}$$

for all words x_1^n of length $n \ge n_0$ and k_{ij} with $\varepsilon n \le k_{ij} \le (1-\varepsilon)n$ and that

$$\sum_{\varepsilon n \le k_{ij} \le (1-\varepsilon)n} S^*(x_1^n) \ge (1 - C_4\sqrt{\varepsilon}) \sum_{x_1^n} S^*(x_1^n). \tag{54}$$

Proof (Sketch). Suppose that $x_1 = x_n = 0$. Then $\hat{p} = p_0 = p_{10}/(p_{01} + p_{10})$, $k_0 = k_{00} + k_{01} + 1$ and $k_1 = k_{10} + k_{11}$. The supremum

$$\sup_{0 \le p_{01}, p_{10} \le 1} \left(\frac{p_{10}}{p_{01} + p_{10}} (1 - p_{01})^{k_{00}} p_{01}^{k_{01}} p_{10}^{k_{10}} (1 - p_{10})^{k_{11}} \right)$$

is obtained for those p_{01}, p_{10} which satisfy the system of equations

$$\begin{aligned} -\frac{1}{p_{01} + p_{10}} - \frac{k_{00}}{1 - p_{01}} + \frac{k_{01}}{p_{01}} &= 0 \\ \frac{1}{p_{10}} - \frac{1}{p_{01} + p_{10}} + \frac{k_{10}}{p_{10}} - \frac{k_{11}}{1 - p_{10}} &= 0. \end{aligned}$$

Assume for a moment that $k_{ij} \ge 2$. By using the "Ansatz"

$$p_{01} = \frac{k_{01} - r}{k_{00} + k_{10}}, \quad p_{10} = \frac{k_{10} - s}{k_{10} + k_{11}}$$

one easily sees that there exists a solution with $0 \le r \le 1$ and $0 \le s \le 1$. Thus, the ratio $S^*(x_1^n)/S^{**}(x_1^n)$ is bounded from below and above. (The cases $k_{ij} = 0$ and $k_{ij} = 1$ can treated similarly.)

The proof of (53) and (54) is also not very difficult. From $k_{ij} \ge \varepsilon n$ one gets $r = O(1/n)$ and $s = O(1/n)$ where the O-constants depend on ε. Consequently

$$p_{01} = \frac{k_{01}}{k_{00} + k_{10}} \left(1 + O\left(\frac{1}{n}\right) \right), \quad p_{10} = \frac{k_{10}}{k_{10} + k_{11}} \left(1 + O\left(\frac{1}{n}\right) \right)$$

and (53) and (54) follow immediately. ∎

We will further need the following asymptotic expansions which can be found in [13, Theorem 5] and Whittle [31].

Lemma 7 *For* $\mathbf{k} = (k_{00}, k_{01}, k_{10}, k_{11})$ *and* $a, b \in \{0, 1\}$, *let* $N_{\mathbf{k}}^{a,b}$ *denote the number of* 0-1*-sequences of length* $n = k_{00} + k_{01} + k_{10} + k_{11} + 1$, *where* $x_0 = a$, $x_n = b$, *and* k_{ij} *is the number of* $k \in \{1, 2, \ldots, n-1\}$ *such that* $(x_k, x_{k+1}) = (i, j)$. *Then*

$$\begin{aligned} N_{\mathbf{k}}^{0,0} &\sim \frac{k_{10}}{k_{10} + k_{11}} \binom{k_{00} + k_{01}}{k_{00}} \binom{k_{10} + k_{11}}{k_{10}}, \\ N_{\mathbf{k}}^{0,1} &\sim \frac{k_{01}}{k_{00} + k_{01}} \binom{k_{00} + k_{01}}{k_{00}} \binom{k_{10} + k_{11}}{k_{10}}, \\ N_{\mathbf{k}}^{1,0} &\sim \frac{k_{10}}{k_{10} + k_{11}} \binom{k_{00} + k_{01}}{k_{00}} \binom{k_{10} + k_{11}}{k_{10}}, \\ N_{\mathbf{k}}^{0,0} &\sim \frac{k_{01}}{k_{00} + k_{01}} \binom{k_{00} + k_{01}}{k_{00}} \binom{k_{10} + k_{11}}{k_{10}} \end{aligned}$$

for those $\mathbf{k}$ *which are admissible (i.e. if* $a = b$ *then* $k_{01} = k_{10}$ *and if* $a \ne b$ *then* $k_{01} = k_{10} \pm 1$*).*

We will use these expansions in the range $\varepsilon n \le k_{ij} \le (1-\varepsilon)n$ for which it is easy to prove that they are uniform (compare with [13]). In what follows we will also use the notation $N_{\mathbf{k}} = N_{\mathbf{k}}^{0,0} + N_{\mathbf{k}}^{0,1} + N_{\mathbf{k}}^{1,0} + N_{\mathbf{k}}^{1,1}$.

Now we are in position to sketch the last part of the **proof of Theorem 5**, that is, to find $R_n^*(Q^*)$. For this we only need to verify the Weyl criterion which will imply that for every Riemann integrable function $f : [0,1] \to \mathbf{R}$

$$\lim_{n\to\infty} \sum_{x_1^n} Q^*(x_1^n) f(\langle \lg Q_n^*(x_1^n)\rangle) = \int_0^1 f(x)\,dx. \tag{55}$$

Then, as in the previous section, we directly prove that $R_n^*(Q_n^*) = -(\log\log 2)/(\log 2) + o(1)$.

In view of Lemma 6 and Weyl's criterion, it suffices to show that

$$\lim_{n\to\infty} \frac{1}{T_n} \sum_{\varepsilon n \le k_{ij} \le (1-\varepsilon)n} N_{\mathbf{k}} S^{**}(\mathbf{k}) e\left(h \lg S^{**}(\mathbf{k})\right) = 0 \tag{56}$$

for all integers $h \neq 0$, where

$$T_n = \sum_{\mathbf{k}} N_{\mathbf{k}} S^{**}(\mathbf{k}).$$

We shall follow the footsteps of the the proof of Theorem 3, that is, we first consider the exponential sums

$$S = \sum_{\mathbf{k}} e\left(h \lg S^{**}(\mathbf{k})\right).$$

By assuming that $k_0 = k_n = 0$ (the other cases are similar) and by using the relations $k_{01} = k_{10} = (n - 1 - k_{00} - k_{11})/2$ we find

$$S = \sum_{k_{00},k_{11}} e\left(h \cdot F(k_{00}, k_{11})\right),$$

where

$$\begin{aligned} F(k_{00}, k_{11}) &= \lg\left(\frac{k_{00} - k_{11} + n + 1}{2}\right) - \log n \\ &+ k_{00} \lg k_{00} + k_{11} \lg k_{11} \\ &+ (n - 1 - k_{00} - k_{11}) \lg \frac{n - 1 - k_{00} - k_{11}}{2} \\ &- \frac{k_{00} - k_{11} + n + 1}{2} \lg\left(\frac{k_{00} - k_{11} + n + 1}{2}\right) \\ &- \frac{k_{11} - k_{00} + n - 1}{2} \lg\left(\frac{k_{11} - k_{00} + n - 1}{2}\right). \end{aligned}$$

As in the proof of Theorem 3, we can estimate S by Van-der-Corput's method. In the final step, we use partial summation to obtain estimates for

$$\widetilde{S} = \sum_{k_{00},k_{11}} \frac{k_{10}}{k_{10} + k_{11}} \binom{k_{00} + k_{01}}{k_{00}} \binom{k_{10} + k_{11}}{k_{10}} e\left(h \cdot F(k_{00}, k_{11})\right)$$

and Lemma 7 to derive (56). This proves Theorem 5.

4 Analysis of the Average Minimax Redundancy

In this section we first establish Lemma 2 that directly implies Theorems 6, 7 and 8, as already proved in Section 2. Then we deal with a class of memoryless sources, Markov sources, and finally renewal sources.

4.1 Proof of Lemma 2

We start with establishing the crucial Lemma 2 that we repeat below for the reader's convenience.

Lemma 2 *Suppose that $\mathcal{S}$ is a subset of probability distributions P on a finite set X. Then for all probability distributions $\widetilde{Q}$ contained in the convex hull of $\mathcal{S}$ we have*

$$\inf_{Q} \sup_{P \in \mathcal{S}} \left(\sum_{x \in X} P(x) \lg \frac{\widetilde{Q}(x)}{Q(x)} \right) = 0. \tag{57}$$

Proof. Suppose that $P_1, \ldots, P_N$ are probability distributions on a finite set X and $\widetilde{Q}$ is a convex convex combination of $P_1, \ldots, P_N$, i.e.

$$\widetilde{Q} = \sum_{i=1}^{N} \alpha_i P_i$$

with $\alpha_i \geq 0$ and $\sum_{i=1}^{N} \alpha_i = 1$. We first show that for every $Q \neq \widetilde{Q}$

$$\sum_{i=1}^{N} \alpha_i D(P_i||\widetilde{Q}) \leq \sum_{i=1}^{N} \alpha_i D(P_i||Q). \tag{58}$$

By using the definition of $D(P||Q)$ it immediately follows that

$$\sum_{i=1}^{N} D(P_i||Q) = \sum_{i=1}^{N} D(P_i||\widetilde{Q}) + N\, D(\widetilde{Q}||Q) > \sum_{i=1}^{N} D(P_i||\widetilde{Q}).$$

Hence, (58) holds for α_i of the form $\alpha_i = 1/N$.

It is clear that the case of rational numbers $\alpha_i = M_i/M$ (with a common denominator M) can be reduced to this case. We just have to set $\beta_j = 1/M$ for $1 \leq M$ and $Q_j = P_i$ for $M_1 + \cdots + M_{i-1} + 1 \leq j \leq M_1 + \cdots + M_i$. Then

$$\sum_{i=1}^{N} \alpha_i D(P_i||Q) = \sum_{j=1}^{M} \beta_j D(Q_j||Q),$$

and the lemma is proved since α_i real can be viewed as a limit of the rational case.

We now prove (57). Obviously we have

$$\inf_{Q} \sup_{P \in \mathcal{S}} \left(\sum_{x \in X} P(x) \lg \frac{\widetilde{Q}(x)}{Q(x)} \right) \leq 0.$$

(We only have to choose $Q = \widetilde{Q}$.)

The converse inequality can be proved indirectly. Let $\widetilde{Q}$ be contained in the convex hull of $\mathcal{S}$, i.e. there are finitely many $P_1, \ldots, P_N \in \mathcal{S}$ such that $\widetilde{Q}$ is convex combination of the form

$$\widetilde{Q} = \sum_{i=1}^{N} \alpha_i P_i$$

Suppose that there exists Q such that for all P

$$\sum_{x \in X} P(x) \lg \frac{\widetilde{Q}(x)}{Q(x)} = D(P||Q) - D(P||\widetilde{Q}) < 0.$$

Then we also have

$$\sum_{i=1}^{N} \alpha_i D(P_i||Q) < \sum_{i=1}^{N} \alpha_i D(P_i||\widetilde{Q})$$

which is of course a contradiction to (58). Thus

$$\inf_{Q} \sup_{P \in \mathcal{S}} \left(\sum_{x \in X} P(x) \lg \frac{\widetilde{Q}(x)}{Q(x)} \right) \geq 0,$$

and this proves the lemma. ∎

As said above, Lemma 2 directly implies Theorem 6, however, verification that Q^* belongs a convex hall of $\mathcal{S}$ might be quite troublesome. Therefore, we relax this condition such that Theorem 6 still holds.

Corollary 1 *Suppose that there exists a probability distribution $\widetilde{Q}$ in the convex hull of $\mathcal{S}$ such that*

$$\max_{x_1^n} \left| \lg \frac{Q^*(x_1^n)}{\widetilde{Q}(x_1^n)} \right| \leq C, \tag{59}$$

then still

$$\overline{R}_n(\mathcal{S}) \geq \lg d_n - C - \sup_{P \in \mathcal{S}} \left(\sum_{x_1^n} P(x_1^n) \lg \frac{\sup_{P \in \mathcal{S}} P(x_1^n)}{P(x_1^n)} \right) + O(1). \tag{60}$$

Proof. The proof of (60) is a trivial extension of that of (27). In fact, the maximal deviation can be bounded by

$$\sup_{P \in \mathcal{S}} \left| \sum_{x_1^n} P(x_1^n) \lg \frac{Q^*(x_1^n)}{\widetilde{Q}(x_1^n)} \right| \leq C$$

for some constant C. ∎

4.2 Proof of Theorem 9

We now study a class of binary memoryless sources $\mathcal{M}_0$ and show that

$$\overline{R}_n(\mathcal{M}_0) = R_n^*(\mathcal{M}_0) + O(1) = \frac{1}{2}\lg n + O(1). \tag{61}$$

By Theorem 6 we have to prove that $c_n(\mathcal{M}_0)$ defined in (28) is $O(1)$ and that Q^* is contained in the convex hull of $\mathcal{M}_0$. Observe that $c_n(\mathcal{M}_0)$

$$c_n(\mathcal{M}_0) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \lg \frac{\left(\frac{k}{n}\right)^k \left(1-\frac{k}{n}\right)^{n-k}}{p^k(1-p)^k}.$$

Lemma 8 *For every $p \in [0,1]$ we have*

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \lg \frac{\left(\frac{k}{n}\right)^k \left(1-\frac{k}{n}\right)^{n-k}}{p^k(1-p)^k} \leq \frac{1}{\log 2}.$$

Proof. By using the inequality $\log x \leq x - 1$ we get

$$\begin{aligned}
\log \frac{\left(\frac{k}{n}\right)^k \left(1-\frac{k}{n}\right)^{n-k}}{p^k(1-p)^{n-k}} &= k \log \frac{k/n}{p} + (n-k) \log \frac{1-k/n}{1-p} \\
&\leq k\left(\frac{k/n}{p} - 1\right) + (n-k)\left(\frac{1-k/n}{1-p} - 1)\right) \\
&= n\left(\frac{(k/n)^2}{p} + \frac{(1-k/n)^2}{1-p} - 1\right).
\end{aligned}$$

Since

$$n \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \frac{(k/n)^2}{p} = np + (1-p),$$

we find

$$\begin{aligned}
n \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \left(\frac{(k/n)^2}{p} + \frac{(1-k/n)^2}{1-p} - 1\right) \\
= np + 1 - p + n(1-p) + p - n = 1,
\end{aligned}$$

which completes the proof of the lemma. ∎

In view of this we conclude that that for memoryless sources

$$c_n(\mathcal{M}_0) = \sup_{P \in \mathcal{S}} \left(\sum_{x_1^n} P(x_1^n) \lg \frac{\sup_{P \in CM_0} P(x_1^n)}{P(x_1^n)} \right) = O(1).$$

On passing we observe that we can be much more precise if we consider just p with $a \leq p \leq b$, where $0 < a < b < 1$. Here we have uniformly for $a \leq p \leq b$ as $n \to \infty$ (cf. [4])

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \lg \frac{\left(\frac{k}{n}\right)^k \left(1-\frac{k}{n}\right)^{n-k}}{p^k(1-p)^k} = \frac{1}{2} + O\left(n^{-1/2}\right).$$

Note that this relation is not true if $a = 0$ or $b = 1$ (cf. [30]).

In order to complete the proof of Theorem 9 we must establish that Q^* belongs to the convex hull of $\mathcal{M}_0$. By Corollary 1 it suffices to prove the following lemma.

Lemma 9 *Suppose that $0 \leq a < b \leq 1$ are given and set $\mathcal{M}_0^{a,b} = \{P_p : a \leq p \leq b\}$. Furthermore, let Q^* be the maximum likelihood distribution corresponding to $\mathcal{M}_0^{a,b}$.*

There exists a convex combination $\widetilde{Q}$ of the probability distributions $P_{k/n}$ $(an \leq k \leq bn)$ such that

$$\max_{x_1^n} \left| \lg \frac{Q^*(x_1^n)}{\widetilde{Q}(x_1^n)} \right| = O(1).$$

as $n \to \infty$.

Proof. We start by considering the case $a = 0$ and $b = 1$. Recall that

$$\sup_{0 \leq p \leq 1} p^k (1-p)^{n-k} = \left(\frac{k}{n}\right)^k \left(1-\frac{k}{n}\right)^{n-k}.$$

Our goal is to show that there exist positive numbers β_l such that the sums

$$s_k := \sum_{l=0}^{n} \beta_l l^k (n-l)^{n-k}$$

satisfy

$$\max_{0 \leq k \leq n} \left| \lg \frac{s_k}{k^k (n-k)^{n-k}} \right| = O(1). \tag{62}$$

Then we just have to define $\widetilde{Q}$ by normalizing s_k.

We now show that we can use

$$\beta_k := \left(\sum_{l=0}^{n} \left(\frac{l}{k}\right)^k \left(\frac{n-l}{n-k}\right)^{n-k} \right)^{-1}.$$

It is an easy exercise to show that β_k can be written as

$$\beta_k = \sqrt{\frac{n}{2\pi k(n-k)}} \left(1 + O\left(\frac{n}{k(n-k)}\right)\right).$$

Hence, after some algebra we arrive at

$$\begin{aligned}\frac{s_k}{k^k(n-k)^{n-k}} &= \sum_{l=0}^{n}\beta_l\left(\frac{l}{k}\right)^k\left(\frac{n-l}{n-k}\right)^{n-k}\\ &= 1+\sum_{l=0}^{n}(\beta_l-\beta_k)\left(\frac{l}{k}\right)^k\left(\frac{n-l}{n-k}\right)^{n-k}\\ &= 1+O\left(\sqrt{\frac{n}{k(n-k)}}\right)\end{aligned}$$

and this proves (62).

Now suppose that $0<a<b<1$. We have

$$\sup_{p\in[a,b]}p^k(1-p)^{n-k}=\begin{cases}a^k(1-a)^{n-k} & \text{for } 0\le k<na,\\ \left(\frac{k}{n}\right)^k\left(1-\frac{k}{n}\right)^{n-k} & \text{for } na\le k\le nb,\\ b^k(1-b)^{n-k} & \text{for } nb<k\le n.\end{cases}$$

Now our goal is to show that there exist positive numbers β_l $(an\le l\le bn)$ such that the sum

$$s_k:=\sum_{an\le l\le bn}\beta_l l^k(n-l)^{n-k}$$

satisfies

$$\max_{an\le k\le bn}\left|\lg\frac{s_k}{k^k(n-k)^{n-k}}\right|=O(1), \tag{63}$$

$$\max_{0\le k<an}\left|\lg\frac{s_k}{(an)^k(n-an)^{n-k}}\right|=O(1), \tag{64}$$

and

$$\max_{bn<k\le n}\left|\lg\frac{s_k}{(bn)^k(n-bn)^{n-k}}\right|=O(1). \tag{65}$$

We define β_k by

$$\beta_k:=\begin{cases}1/\sqrt{n} & \text{for } \lceil an\rceil<k<\lfloor bn\rfloor,\\ 1 & \text{for } k=\lceil an\rceil \text{ and } k=\lfloor bn\rfloor.\end{cases}$$

First, (64) follows from

$$\begin{aligned}(an)^k(n-an)^{n-k} &\le (an)^k(n-an)^{n-k}+\frac{1}{\sqrt{n}}\sum_{l>an}l^k(n-l)^{n-k}\\ &= O((an)^k(n-an)^{n-k})\end{aligned}$$

if $k\le an$. Observe that (65) is just the symmetric case. Thus it remains to show (63).

It is clear that the mapping $l\mapsto l^k(n-l)^{n-k}$ attains its maximum for $l=k$. A local expansion (around this optimal value) shows that for every fixed $0<a<b<1$ there exist two positive constants c_1,c_2 such that

$$c_1\sqrt{n}k^k(n-k)^{n-k}\le\sum_{an\le k\le bn}l^k(n-l)^{n-k}\le c_2\sqrt{n}k^k(n-k)^{n-k}$$

for all k with $an \le k \le bn$. Thus, (63) follows.

Finally, if $a = 0$ and $b < 1$ or $a > 1$ and $b = 1$, then we have to combine the two preceding cases. This completes the proof. ∎

4.3 Proof of Theorem 10

Let us now consider a class of binary Markov sources $\mathcal{M}_1$ and establish Theorem 10, that is,

$$\overline{R}_n(\mathcal{M}_0) = R_n^*(\mathcal{M}_0) + O(1) = \lg n + O(1). \tag{66}$$

As in the case of memoryless sources we refer to Theorem 6 and Corollary 1. We start with the following lemma.

Lemma 10 *There exists a constant $C > 0$ such that for any transition matrix* $\mathbf{P}$

$$\sum_{x_1^n} P(x_1^n) \lg \frac{S^*(x_1^n)}{P(x_1^n)} \le C$$

where

$$S^*(x_1^n) = \sup_{p_{ij}} \left(\hat{p}\, p_{00}^{k_{00}} p_{01}^{k_{01}} p_{10}^{k_{10}} p_{11}^{k_{11}} \right).$$

Proof. By Lemma 6 it suffices to prove that

$$\mathbf{E}\left[\lg \frac{S^{**}(x_1^n)}{P(x_1^n)} \right] = \sum_{x_1^n} P(x_1^n) \lg \frac{S^{**}(x_1^n)}{P(x_1^n)} \le C', \tag{67}$$

where

$$S^{**}(x_1^n) := \frac{\frac{\hat{k}}{\hat{\hat{k}}}}{\frac{k_{01}}{k_{00}+k_{01}} + \frac{k_{10}}{k_{10}+k_{11}}} \left(\frac{k_{00}}{k_{00}+k_{01}} \right)^{k_{00}} \left(\frac{k_{01}}{k_{00}+k_{01}} \right)^{k_{01}} \left(\frac{k_{10}}{k_{10}+k_{11}} \right)^{k_{10}} \left(\frac{k_{11}}{k_{10}+k_{11}} \right)^{k_{11}}.$$

We split the sum (67) into five natural terms $S_1, \ldots, S_5$ with the first term being

$$S_1 = \sum_{x_1^n} P(x_1^n) \lg \frac{\frac{\hat{k}}{\hat{\hat{k}}}}{\hat{p}\left(\frac{k_{01}}{k_{00}+k_{01}} + \frac{k_{10}}{k_{10}+k_{11}} \right)} \le \frac{1}{\log 2} \sum_{x_1^n} P(x_1^n) \frac{1}{\hat{p}}.$$

From the generating function $G(z)$ computed in (51) we conclude that

$$\begin{aligned} \sum_{n \ge 1} \sum_{x_1^n} P(x_1^n) \frac{1}{\hat{p}}\, z^n &= z \frac{1 - p_{11}z + p_{10}z + 1 - p_{00}z + p_{01}z}{1 - (p_{00}+p_{11})z + (p_{00}p_{11} - p_{01}p_{10})z^2} \\ &= \frac{2z}{1-z} \end{aligned}$$

and hence $S_1 \le 2/\log 2$.

To analyze the other terms of $\lg S^{**}(x_1^n)/P(x_1^n)$, we introduce the following notations:

$$\mathbf{E}_n[f(\mathbf{k})] := \sum_{x_1^n} P(x_1^n) f(\mathbf{k}) = \sum_{\mathbf{k}} N_{\mathbf{k}}\, \hat{p} \prod_{1\le i,j\le 1} p_{ij}^{k_{ij}} \cdot f(\mathbf{k}),$$

and (for $a, b \in \{0,1\}$)

$$\mathbf{E}_n^{a,b}[f(\mathbf{k})] := \sum_{x_1^n, x_1=a, x_n=b} P(x_1^n) f(\mathbf{k}) = \sum_{\mathbf{k}} N_{\mathbf{k}}^{a,b}\, \hat{p} \prod_{1\le i,j\le 1} p_{ij}^{k_{ij}} \cdot f(\mathbf{k}),$$

where $\mathbf{k}$ is an abbreviation for $\mathbf{k} = (k_{00}, k_{01}, k_{10}, k_{11})$. (In the sequel we will also make use of the short hand notations $\mathbf{E}_n^{\cdot,0} = \mathbf{E}_n^{0,0} + \mathbf{E}_n^{1,0}$ and $\mathbf{E}_n^{\cdot,1} = \mathbf{E}_n^{0,1} + \mathbf{E}_n^{1,1}$.) From

$$\frac{k_{00}}{k_{00}+k_{01}} N_{\mathbf{k}}^{0,0} = N_{\mathbf{k}'}^{0,0}, \quad \frac{k_{00}}{k_{00}+k_{01}} N_{\mathbf{k}}^{1,0} = N_{\mathbf{k}'}^{1,0}$$

and

$$\frac{k_{00}}{k_{00}+k_{01}-1} N_{\mathbf{k}}^{0,1} = N_{\mathbf{k}'}^{0,1}, \quad \frac{k_{00}}{k_{00}+k_{01}-1} N_{\mathbf{k}}^{1,1} = N_{\mathbf{k}'}^{1,1},$$

where $\mathbf{k}' = (k_{00}-1, k_{01}, k_{10}, k_{11})$ (and where we assume that $k_{00} > 0$), we derive

$$\mathbf{E}_n^{0,0}\left[\frac{k_{00}}{k_{00}+k_{01}} f(\mathbf{k})\right] = p_{00}\, \mathbf{E}_{n-1}^{0,0}[f(k_{00}+1, k_{01}, k_{10}, k_{11})],$$

a corresponding identity for $\mathbf{E}_n^{0,1}$, and slightly modified identities for $\mathbf{E}_n^{1,0}$ and $\mathbf{E}_n^{1,1}$. Since $1/(k_{00}+k_{01}-1) \le 1/(k_{00}+k_{01})$ we also have

$$\mathbf{E}_n\left[\frac{k_{00}}{k_{00}+k_{01}} f(\mathbf{k})\right] \le p_{00}\, \mathbf{E}_{n-1}[f(k_{00}+1, k_{01}, k_{10}, k_{11})]$$

if $f \ge 0$.

In particular, by using the inequality $\log x \le x-1$ we obtain the following for the second sum S_2

$$\begin{aligned} S_2 &= \mathbf{E}_n\left[k_{00} \lg \frac{k_{00}}{p_{00}(k_{01}+k_{10})}\right] \\ &\le \frac{1}{\log 2} \mathbf{E}_n\left[\frac{k_{00}^2}{p_{00}(k_{00}+k_{01})} - k_{00}\right] \\ &\le \frac{1}{\log 2}\left(\mathbf{E}_{n-1}[k_{00}+1] - \mathbf{E}_n[k_{00}]\right). \end{aligned}$$

In view of (51) these terms can be handled with the help of generating functions. By differentiating the generating function $G(z)$ derived in (51) with respect to p_{00}, multiplying by p_{00} and setting $p_{00}+p_{01}=1$ and $p_{10}+p_{11}=1$ we find

$$\sum_{n\ge 0} \left(\mathbf{E}_n[k_{00}]\right) z^n = p_0 p_{00} \frac{z^2}{(1-z)^2}$$

and consequently $\mathbf{E}_n[k_{00}] = p_0 p_{00}(n-1)$. (Similarly, we find $\mathbf{E}_n[k_{01}] = p_0 p_{01}(n-1)$, $\mathbf{E}_n[k_{10}] = p_1 p_{10}(n-1)$, and $\mathbf{E}_n[k_{11}] = p_1 p_{11}(n-1)$). This immediately implies

$$\mathbf{E}_{n-1}[k_{00}+1] - \mathbf{E}_n[k_{00}] = p_0 p_{00}(n-2) + 1 - p_0 p_{00}(n-1) = 1 - p_0 p_{00} \le 1$$

and, hence, $S_2 \le 1/\log 2$.

The next sum

$$S_3 = \mathbf{E}_n\left[k_{01} \lg \frac{k_{01}}{p_{01}(k_{00}+k_{01})}\right]$$

is much more delicate. We have to split it into three terms:

$$\begin{aligned}
S_{3,1} &= \mathbf{E}_n^{\cdot,0}\left[k_{01} \lg \frac{k_{01}}{p_{01}(k_{01}+k_{10})}\right], \\
S_{3,2} &= \mathbf{E}_n^{\cdot,1}\left[\mathbf{1}_{[k_{01}\le 1]} k_{01} \lg \frac{k_{01}}{p_{01}(k_{01}+k_{10})}\right], \\
S_{3,3} &= \mathbf{E}_n^{\cdot,1}\left[\mathbf{1}_{[k_{01}> 1]} k_{01} \lg \frac{k_{01}}{p_{01}(k_{01}+k_{10})}\right].
\end{aligned}$$

By using (again) the inequality $\log x \le x - 1$ and the identity

$$\frac{k_{01}^2}{k_{00}+k_{01}} = \frac{k_{00}^2}{k_{00}+k_{01}} - k_{00} + k_{01}$$

we obtain

$$\begin{aligned}
S_{3,1} &\le \frac{1}{\log 2}\mathbf{E}_n^{\cdot,0}\left[\frac{k_{01}^2}{p_{01}(k_{00}+k_{01})} - k_{01}\right] \\
&= \frac{1}{p_{01}\log 2}\left(\mathbf{E}_n^{\cdot,0}\left[\frac{k_{00}^2}{(k_{00}+k_{01})}\right] - \mathbf{E}_n^{\cdot,0}[k_{00}] + \mathbf{E}_n^{\cdot,0}[k_{01}] - p_{01}\mathbf{E}_n^{\cdot,0}[k_{01}]\right) \\
&= \frac{1}{p_{01}\log 2}\left(p_{00}\mathbf{E}_{n-1}^{\cdot,0}[k_{00}+1] - \mathbf{E}_n^{\cdot,0}[k_{00}] + p_{00}\mathbf{E}_n^{\cdot,0}[k_{01}]\right).
\end{aligned}$$

The second sum $S_{3,2}$ can be handled explicitly.[‡] (We also assume that $n > 2$. The cases $n = 1$ and $n = 2$ are easy to check separately.)

$$\begin{aligned}
S_{3,2} &= \mathbf{E}_n^{0,1}\left[\mathbf{1}_{[k_{01}=1]} \lg \frac{1}{p_{01}(k_{00}+1)}\right] + \mathbf{E}_n^{1,1}\left[\mathbf{1}_{[k_{01}=1]} \lg \frac{1}{p_{01}(k_{00}+1)}\right] \\
&= S'_{3,2} + S''_{3,2}.
\end{aligned}$$

For the sake of brevity we just consider the first term $S'_{3,2}$. Note that any source sequence x_1^n with $x_1 = 0$, $x_n = 1$ and $k_{01} = 1$ is of the form $x_1^n = 0^\ell 1^{n-\ell}$ with $1 \le \ell \le n-1$. Thus

$$S'_{3,2} = \frac{p_{01}p_{10}}{(p_{01}+p_{10})\log 2}\sum_{\ell=1}^{n-1} p_{00}^{\ell-1} p_{11}^{n-\ell-1} \log \frac{1}{\ell p_{01}}.$$

[‡]This is in fact crucial because it would not be bounded (but of order $\log n$) if we use the inequality $\log x \le x - 1$ here.

If $p_{11} \leq p_{00}$ we get (by using the inequality $p_{00} = 1 - p_{01} \leq e^{-p_{01}}$)

$$\begin{aligned} S'_{3,2} &\leq \frac{p_{01}p_{10}}{(p_{01}+p_{10})\log 2} p_{00}^{n-2} \left((n-2)\log \frac{1}{(n-2)p_{01}} + O(n-2) \right) \\ &\leq \frac{p_{10}}{(p_{01}+p_{10})\log 2} (n-2)p_{01} e^{-(n-2)p_{01}} \left(\log \frac{1}{(n-2)p_{01}} + O(1) \right) \\ &= O(1). \end{aligned}$$

If $p_{00} \leq p_{11}$ then $\frac{1}{p_{01}} \leq \frac{1}{p_{10}}$ and consequently

$$\begin{aligned} S'_{3,2} &\leq \frac{p_{01}p_{10}}{(p_{01}+p_{10})\log 2} p_{11}^{n-2} \left((n-2)\log \frac{1}{(n-2)p_{10}} + O(n-2) \right) \\ &\leq \frac{p_{01}}{(p_{01}+p_{10})\log 2} ((n-1)p_{10}) e^{-(n-2)p_{10}} \left(\log \frac{1}{(n-2)p_{10}} + O(1) \right) \\ &= O(1). \end{aligned}$$

Thus, the second sum $S_{3,2}$ is bounded.

The third sum $S_{3,3}$ can be manipulated in a similar manner as the first sum, namely

$$\begin{aligned} S_{3,3} &\leq \frac{1}{\log 2} \mathbf{E}_n^{\cdot,1} \left[\mathbf{1}_{[k_{01}>1]} \frac{k_{01}^2}{p_{01}(k_{00}+k_{01})} - k_{01} \right] \\ &= \frac{1}{p_{01}\log 2} \Bigg(\mathbf{E}_n^{\cdot,1} \left[\mathbf{1}_{[k_{01}>1]} \frac{k_{00}^2}{(k_{00}+k_{01})} \right] \\ &\qquad - \mathbf{E}_n^{\cdot,1}[\mathbf{1}_{[k_{01}>1]} k_{00}] + p_{00}\mathbf{E}_n^{\cdot,1}[\mathbf{1}_{[k_{01}>1]} k_{01}] \Bigg). \end{aligned}$$

However, we must be more careful. We start by considering the following term:

$$\begin{aligned} \mathbf{E}_n^{\cdot,1} \left[\mathbf{1}_{[k_{01}>1]} \frac{k_{00}^2}{(k_{00}+k_{01})} \right] &= \mathbf{E}_n^{\cdot,1} \left[\mathbf{1}_{[k_{01}>1,k_{00}>0]} \frac{k_{00}^2}{(k_{00}+k_{01}-1)} \right] \\ &\quad - \mathbf{E}_n^{\cdot,1} \left[\mathbf{1}_{[k_{01}>1,k_{00}>0]} \frac{k_{00}^2}{(k_{00}+k_{01})(k_{00}+k_{01}-1)} \right] \\ &= p_{00}\mathbf{E}_{n-1}^{\cdot,1} \left[\mathbf{1}_{[k_{01}>1]} (k_{00}+1) \right] - p_{00}\mathbf{E}_{n-1}^{\cdot,1} \left[\mathbf{1}_{[k_{01}>1]} \frac{k_{00}+1}{(k_{00}+k_{01}+1)} \right] \\ &= p_{00}\mathbf{E}_{n-1}^{\cdot,1} \left[\mathbf{1}_{[k_{01}>1]} (k_{00}+1) \right] \\ &\quad - p_{00}\mathbf{E}_{n-1}^{\cdot,1} \left[\mathbf{1}_{[k_{01}>1]} \right] + p_{00}\mathbf{E}_{n-1}^{\cdot,1} \left[\mathbf{1}_{[k_{01}>1]} \frac{k_{01}}{(k_{00}+k_{01}+1)} \right]. \end{aligned}$$

Since $k_{01} > 1$ we have $k_{01} \leq 2(k_{01}-1)$ and consequently

$$\begin{aligned} \mathbf{E}_{n-1}^{\cdot,1} \left[\mathbf{1}_{[k_{01}>1]} \frac{k_{01}}{(k_{00}+k_{01}+1)} \right] &\leq 2\mathbf{E}_{n-1}^{\cdot,1} \left[\mathbf{1}_{[k_{01}>1]} \frac{k_{01}-1}{(k_{00}+k_{01}-1)} \right] \\ &= 2\mathbf{E}_{n-1}^{\cdot,1} \left[\mathbf{1}_{[k_{01}>1]} \right] - 2\mathbf{E}_{n-1}^{\cdot,1} \left[\mathbf{1}_{[k_{01}>1]} \frac{k_{00}}{(k_{00}+k_{01}-1)} \right] \end{aligned}$$

$$
\begin{aligned}
&= 2\mathbf{E}^{\cdot,1}_{n-1}\left[\mathbf{1}_{[k_{01}>1]}\right] - 2p_{00}\mathbf{E}^{\cdot,1}_{n-2}\left[\mathbf{1}_{[k_{01}>1]}\right] \\
&= 2P_{n-1}(x_{n-1}=1, k_{01}>1) - 2p_{00}P_{n-2}(x_{n-2}=1, k_{01>1}) \\
&= 2P_{n-2}(x_{n-2}=1, k_{01}>1)p_{11} + 2P_{n-2}(x_{n-2}=0, k_{01}>0)p_{01} \\
&-2p_{00}P_{n-2}(x_{n-2}=1, k_{01}>1) \\
&\leq 2\left(P_{n-2}(x_{n-2}=1, k_{01}>1) + P_{n-2}(x_{n-2}=0, k_{01}>0)\right)p_{01} \\
&\leq 2p_{01}.
\end{aligned}
$$

Next, if $k_{01}=1$ and $x_n=1$ then there is only one 0-sequence in x_1^n. Hence, there is a natural bijection between words of length n and $n-1$ of that kind (by deleting resp. inserting a zero if one assumes that there is a least one zero in the longer word x_1^n). Consequently

$$
\begin{aligned}
&\mathbf{E}^{\cdot,1}_{n}\left[\mathbf{1}_{[k_{01}\leq 1]}k_{00}\right] - p_{00}\mathbf{E}^{\cdot,1}_{n-1}\left[\mathbf{1}_{[k_{01}\leq 1]}k_{00}\right] \\
&\quad = \mathbf{E}^{\cdot,1}_{n}\left[\mathbf{1}_{[k_{01}=1,k_{00}>0]}k_{00}\right] - p_{00}\mathbf{E}^{\cdot,1}_{n-1}\left[\mathbf{1}_{[k_{01}=1]}k_{00}\right] \\
&\quad = p_{00}\mathbf{E}^{\cdot,1}_{n-1}\left[\mathbf{1}_{[k_{01}=1]}(k_{00}+1)\right] - p_{00}\mathbf{E}^{\cdot,1}_{n-1}\left[\mathbf{1}_{[k_{01}=1]}k_{00}\right] \\
&\quad = p_{00}\mathbf{E}^{\cdot,1}_{n-1}\left[\mathbf{1}_{[k_{01}=1]}\right] = p_{00}\mathbf{E}^{\cdot,1}_{n-1}\left[\mathbf{1}_{[k_{01}\leq 1]}k_{01}\right].
\end{aligned}
$$

Furthermore we have

$$
\begin{aligned}
&\mathbf{E}^{\cdot,1}_{n-1}\left[\mathbf{1}_{[k_{01}=1]}\right] - \mathbf{E}^{\cdot,1}_{n}\left[\mathbf{1}_{[k_{01}=1]}\right] \\
&\quad = P_{n-1}(x_{n-1}=1, k_{01}=1) - P_n(x_n=1, k_{01}=1) \\
&\quad = P_{n-1}(x_{n-1}=1, k_{01}=1) \\
&\quad -P_{n-1}(x_{n-1}=1, k_{01}=1)p_{11} - P_{n-1}(x_{n-1}=0, k_{01}=0)p_{01} \\
&\quad = p_{10}P_{n-1}(x_{n-1}=1, k_{01}=1) + O(p_{01}) \\
&\quad = p_{10}\left(p_0\sum_{\ell=1}^{n-2}p_{00}^{\ell-1}p_{01}p_{11}^{n-l-2} + p_1\sum_{\ell=1}^{n-2}\sum_{j=1}^{n-1-\ell}p_{11}^{\ell+j-2}p_{10}p_{01}p_{00}^{n-\ell-j-2}\right) + O(p_{01}) \\
&\quad \leq p_{10}\left(p_0p_{01}\frac{1}{1-p_{11}} + p_1p_{10}p_{01}\frac{1}{(1-p_{11})^2}\right) + O(p_{01}) \\
&\quad = p_{01} + O(p_{01}) = O(p_{01})
\end{aligned}
$$

This implies

$$
\begin{aligned}
&p_{00}\mathbf{E}^{\cdot,1}_{n-1}\left[\mathbf{1}_{[k_{01}>1]}k_{00}\right] - \mathbf{E}^{\cdot,1}_{n}\left[\mathbf{1}_{[k_{01}>1]}k_{00}\right] + p_{00}\mathbf{E}^{\cdot,1}_{n}\left[\mathbf{1}_{[k_{01}>1]}k_{01}\right] \\
&\quad = p_{00}\mathbf{E}^{\cdot,1}_{n-1}\left[k_{00}\right] - \mathbf{E}^{\cdot,1}_{n}\left[k_{00}\right] + p_{00}\mathbf{E}^{\cdot,1}_{n}\left[k_{01}\right] + O(p_{01}).
\end{aligned}
$$

Thus,

$$
S_{3,3} \;\leq\; \frac{1}{p_{01}\log 2}\Big(p_{00}\mathbf{E}^{\cdot,1}_{n-1}\left[\mathbf{1}_{[k_{01}>1]}(k_{00}+1)\right] - p_{00}\mathbf{E}^{\cdot,1}_{n-1}[\mathbf{1}_{[k_{01}>1]}] + O(p_{01})
$$

$$
\begin{aligned}
& -\mathbf{E}_n^{\cdot,1}[\mathbf{1}_{[k_{01}>1]}k_{00}] + p_{00}\mathbf{E}_n^{\cdot,1}[\mathbf{1}_{[k_{01}>1]}k_{01}]\Big) \\
= \; & \frac{1}{p_{01}\log 2}\left(p_{00}\mathbf{E}_{n-1}^{\cdot,1}\left[k_{00}\right] - \mathbf{E}_n^{\cdot,1}\left[k_{00}\right] + p_{00}\mathbf{E}_n^{\cdot,1}\left[k_{01}\right] + O(p_{01})\right).
\end{aligned}
$$

Now, we add up the two upper bounds for $S_{3,1}$ and $S_{3,3}$, note that $\mathbf{E}_{n-1}^{\cdot,0}[1] = \Pr\{x_{n-1} = 0\} = p_0$, and obtain

$$
\begin{aligned}
S_{3,1} + S_{3,3} \; \leq \; & \frac{1}{p_{01}\log 2}\Big(p_{00}\mathbf{E}_{n-1}\left[k_{00}\right] + p_{00}\mathbf{E}_{n-1}^{\cdot,0}[1] \\
& \qquad -\mathbf{E}_n\left[k_{00}\right] + p_{00}\mathbf{E}_n\left[k_{01}\right] + O(p_{01})\Big) \\
= \; & \frac{1}{p_{01}\log 2}\Big(p_0 p_{00}^2(n-2) + p_0 p_{00} - p_0 p_{00}(n-1) \\
& \qquad +p_0 p_{00} p_{01}(n-1) + O(p_{01})\Big) \\
= \; & \frac{1}{p_{01}\log 2}\left(p_0 p_{00} p_{01} + O(p_{01})\right) \\
= \; & O(1)
\end{aligned}
$$

Thus, $S_3 = O(1)$, too. The remaining two terms are completely symmetric to S_2 and S_3 and are (also) bounded. This completes the proof of the lemma. ∎

To complete the proof of Theorem 10 we need to verify condition (59) of Corollary 1.

Lemma 11 *There exists a convex combination $\widetilde{Q}$ of the probability distributions induces by Markov sources such that as $n \to \infty$*

$$
\max_{x_1^n}\left|\lg\frac{Q^*(x_1^n)}{\widetilde{Q}(x_1^n)}\right| = O(1).
$$

Proof (Sketch). In view of Lemma 6 we can replace the *real* distribution Q^* by Q^{**} which is defined as

$$
Q^{**}(x_1^n) = \frac{1}{d_n}\frac{\hat{k}}{n}\left(\frac{k_{00}}{\widetilde{k}_0}\right)^{k_{00}}\left(\frac{k_{01}}{\widetilde{k}_0}\right)^{k_{01}}\left(\frac{k_{10}}{\widetilde{k}_1}\right)^{k_{10}}\left(\frac{k_{11}}{\widetilde{k}_1}\right)^{k_{11}} = \frac{S^{**}(x_1^n)}{d_n},
$$

where $\widetilde{k}_0 = k_{00} + k_{01}$, $\widetilde{k}_1 = k_{10} + k_{11}$, and

$$
d_n = \sum_{\mathbf{k}} N_{\mathbf{k}}\frac{\hat{k}}{n}\left(\frac{k_{00}}{\widetilde{k}_0}\right)^{k_{00}}\left(\frac{k_{01}}{\widetilde{k}_0}\right)^{k_{01}}\left(\frac{k_{10}}{\widetilde{k}_1}\right)^{k_{10}}\left(\frac{k_{11}}{\widetilde{k}_1}\right)^{k_{11}}.
$$

Now it suffices to prove that there are $\beta_{\mathbf{k}} > 0$ such that the numbers

$$
s_{\mathbf{k}} := \sum_{\mathbf{l}} \beta_{\mathbf{l}}\frac{\hat{l}}{n}\left(\frac{l_{00}}{\widetilde{l}_0}\right)^{k_{00}}\left(\frac{l_{01}}{\widetilde{l}_0}\right)^{k_{01}}\left(\frac{l_{10}}{\widetilde{l}_1}\right)^{k_{10}}\left(\frac{l_{11}}{\widetilde{l}_1}\right)^{k_{11}}
$$

satisfy

$$C_1 \leq s_{\mathbf{k}} \leq C_2$$

for some absolute constants $C_1, C_2 > 0$. It turns out that one can use

$$\beta_{\mathbf{k}} := \left(\sum_{\mathbf{l}} \frac{\hat{l}}{n} \left(\frac{l_{00}}{\widetilde{l}_0} \right)^{k_{00}} \left(\frac{l_{01}}{\widetilde{l}_0} \right)^{k_{01}} \left(\frac{l_{10}}{\widetilde{l}_1} \right)^{k_{10}} \left(\frac{l_{11}}{\widetilde{l}_1} \right)^{k_{11}} \right)^{-1}.$$

The calculations are quite similar to those of Lemma 9 but much more involved. We leave the details to the reader. ∎

4.4 Proof of Lemma 3

We start with a recollection of some definitions. Let $q = (q_0, q_1, q_2, \ldots)$ be a probability distribution on the non-negative integers. A renewal process (with law q) generates a random 0-1-sequence in the following way. It starts with a "1" followed by a series of 0s of length α_0 (with probability q_{α_0}) followed by a 1. Then there is again a series of 0s of length α_1(with probability q_{α_1} and independent of α_0) followed by a 1 and so on.

Certainly, such a renewal process induces (marginal) distributions on binary sequences of length n. Suppose that x_1^n is of the form

$$x_1^n = 0^{\alpha_0} 10^{\alpha_1} 1 \cdots 10^{\alpha_r} 10^{k^*}, \tag{68}$$

and let k_m denote the number of α_i with $\alpha_i = m$. Then

$$P(x_1^n) = q_0^{k_0} q_1^{k_1} \cdots q_{n-1}^{k_{n-1}} (1 - q_0 - q_1 - \cdots - q_{k^*-1}).$$

Note that

$$k_0 + 2k_1 + \cdots + nk_{n-1} + k^* = n$$

which is the integer partition of n.

To proceed we need a good approximation for the maximum likelihood distribution. We state it in the next lemma that is easy to prove (cf. [11]), so we omit it here.

Lemma 12 *Let $\mathcal{R}_0$ denote the set of all probability distributions on binary sequences of length n induced by renewal processes. Suppose that x_1^n is of the form (68) and let $k_0, k_1, \ldots, k_{n-1}$ and k^* as above and set $k' = k_0 + \cdots + k_{k^*-1}$. Then*

$$\sup_{P \in \mathcal{R}_0} P(x_1^n) = \prod_{i=0}^{n-1} \left(\frac{k_i}{k+1} \right)^{k_i} \left(1 + \frac{1}{k-k'} \right)^{k-k'} \left(1 - \frac{k'}{k+1} \right).$$

Consequently, there exist two constants $C_1, C_2 > 0$ such that the ratio between $\sup_{P \in \mathcal{R}_0} P(x_1^n)$ and

$$\prod_{i=0}^{n-1} \left(\frac{k_i}{k} \right)^{k_i} \left(1 - \frac{k'}{k} \right)$$

is contained in the interval $[C_1, C_2]$.

We want to prove Lemma 3 that we repeat below for the reader's convenience.

Lemma 3 *For all probability distributions $P \in \mathcal{R}_0$ we have*

$$c_n(\mathcal{R}_0) = \sum_{x_1^n} P(x_1^n) \lg \frac{\sup_{P\in\mathcal{R}_0} P(x_1^n)}{P(x_1^n)} \leq \frac{1+\frac{2}{e}}{\log 2\sqrt{\frac{\sqrt{2}}{3}+\frac{2}{e}}} \cdot \sqrt{n} + O(1) \approx 2.278 \cdot \sqrt{n}. \quad (69)$$

Proof. For reader's convenience, let K_n denote the set of all n-tuples $\mathbf{k} = (k_0, k_1, \ldots, k_{n-1})$ of non-negative integers such that $k_0 + 2k_1 + \cdots nk_{n-1} \leq n$. Furthermore, for fixed $q = (q_0, \ldots, q_{n-1})$ we will use the notation

$$\mathbf{E}_n[f(\mathbf{k})] := \sum_{\mathbf{k}\in K_n} \frac{(k_0+\cdots+k_{n-1})!}{k_0!\cdots k_{n-1}!} \prod_{i=0}^{n-1} q_i^{k_i} \cdot (1-q_0-\cdots-q_{k^*-1})\, f(\mathbf{k}).$$

In what follows we will use the relation

$$\mathbf{E}_n\left[\frac{k_i}{k_0+\cdots+k_{n-1}} f(k_0,\ldots,k_i,\ldots)\right] = q_i \mathbf{E}_{n-i-1}[f(k_0,\ldots,k_i+1,\ldots)] \quad (70)$$

which is a direct consequence of the above definition. We will also use the inequality

$$\mathbf{E}_{n-1}[k_i] \leq \mathbf{E}_n[k_i] \quad (71)$$

that can be proved in the following way: Set $q(x) = q_0 x + q_1 x^2 + \cdots + q_{n-1}x^n$. Then

$$G(x,u) := \sum_{n\geq 0} \sum_{x_1^n} P(x_1^n) u^{k_i}\, x^n = \frac{1-q(x)}{(1-x)(1-q(x)-(u-1)q_i x^{i+1})}.$$

Consequently

$$\frac{dG(x,u)}{du}\,|_{u=1} = \sum_{n\geq 0} (\mathbf{E}_n k_i)\, x^n = \frac{q_i x^{i+1}}{(1-x)(1-q(x))}$$

and

$$\sum_{n\geq 0} (\mathbf{E}_n[k_i] - \mathbf{E}_{n-1}[k_i])\, x^n = \frac{q_i x^{i+1}}{1-q(x)}.$$

Since the function $q_i x^{i+1}/1 - q(x)$ has only non-negative Taylor coefficients, (71) follows immediately.

Now we start with the proof of (69). By Lemma 12 it suffices to use

$$\prod_{i=0}^{n-1} \left(\frac{k_i}{k}\right)^{k_i} \left(1-\frac{k'}{k}\right)$$

instead of $\sup_{P\in\mathcal{R}_0} P(x_1^n)$. Thus, we will have to deal with the following sum (here and in what follows k is always an abbreviation for $k_0+\cdots+k_{n-1}$):

$$c_n(\mathcal{R}_0) = \sum_{x_1^n} P(x_1^n) \left(\lg \frac{1-\frac{k_0+\cdots+k_{k^*-1}}{k}}{1-q_0-\cdots-q_{k^*-1}} + \sum_{i=0}^{n-1} k_i \lg \frac{k_i}{q_i k} \right).$$

By using the inequality $\log x \le x$ we first find

$$\sum_{x_1^n} P(x_1^n) \lg \frac{1 - \frac{k_0+\cdots+k_{k^*-1}}{k}}{1 - q_0 - \cdots - q_{k^*-1}} \le \sum_{\mathbf{k}\in K_n} \frac{k!}{k_0! \cdots k_{n-1}!} \prod_{i=0}^{n-1} q_i^{k_i} \cdot \left(1 - \frac{k_0 + \cdots + k_{k^*-1}}{k}\right) := A_n$$

Now the generating function of the last sum is given by

$$A(z) := \sum_{n\ge 0} A_n x^n = \frac{1 - q(x^2)}{(1-x)(1-q(x))}.$$

Since $q(1) \le 1$ and the degree of $q(x)$ is $\le n$ it follows that the Taylor coefficients of $(1 - q(x^2))/((1-x)(1-q(x)))$ are bounded by 2.

Next consider the sum

$$S_2 = \sum_{x_1^n} P(x_1^n) k_i \lg \frac{k_i}{q_i k}.$$

By using the *weights* $P(x_1^n)k_i/\mathbf{E}_n[k_i]$ and the concavity of the logarithm it follows from Jensen's inequality that

$$S_2 = \sum_{x_1^n} P(x_1^n) k_i \lg \frac{k_i}{q_i k} \le \mathbf{E}_n[k_i] \cdot \lg \left(\frac{\mathbf{E}_n\left[\frac{k_i^2}{q_i k}\right]}{\mathbf{E}_n[k_i]} \right).$$

By (70) and (71) we have

$$\mathbf{E}_n\left[\frac{k_i^2}{q_i k}\right] = \mathbf{E}_{n-i}[k_i] + 1 \le \mathbf{E}_n[k_i] + 1,$$

and consequently

$$\sum_{x_1^n} P(x_1^n) k_i \lg \frac{k_i}{q_i k} \le \mathbf{E}_n[k_i] \cdot \lg \left(1 + \frac{1}{\mathbf{E}_n[k_i]}\right).$$

Thus, we we are led to estimate the sum

$$\sum_{i=0}^{n-1} \mathbf{E}_n[k_i] \cdot \lg \left(1 + \frac{1}{\mathbf{E}_n[k_i]}\right).$$

Since $\sum_{i=0}^{n-1}(i+1)\mathbf{E}_n k_i \le n$ we obtain an upper bound by determining the maximum of the sum

$$\sum_{i=1}^{n} y_i \lg \left(1 + \frac{1}{y_i}\right) = \frac{1}{\log 2} \sum_{i=1}^{n} y_i \log \left(1 + \frac{1}{y_i}\right)$$

provided $y_i \ge 0$ and $\sum_{i=1}^n i y_i \le n$. By Lagrange's method we have to solve the system of equations

$$\log \left(1 + \frac{1}{y_i}\right) - \frac{1}{1+y_i} - \lambda i = 0,$$

where $\sum_{i=1}^{n} iy_i = n' \leq n$. For this purpose consider the function $y = y(x)$ defined by

$$\log\left(1 + \frac{1}{y}\right) - \frac{1}{1+y} = x.$$

It is easy to show that $y(x)$ is asymptotically given by

$$y(x) \sim \frac{1}{\sqrt{2x}} \qquad \text{for } x \to 0$$

and by

$$y(x) \sim e^{-x} \qquad \text{for } x \to \infty.$$

Thus, for $i < 1/\lambda$ we have $y_i \sim 1/\sqrt{2\lambda i}$ and for $i > 1/\lambda$ we get $y_i \sim e^{-\lambda i}$. Consequently

$$\begin{aligned}
\sum_{i=1}^{n} iy_i &\sim \sum_{i=1}^{1/\lambda} \sqrt{i}/\sqrt{2\lambda} + \sum_{i>1/\lambda} ie^{-\lambda i} \\
&\sim \frac{1}{\sqrt{2\lambda}} \int_0^{1/\lambda} \sqrt{x}dx + \int_{1/\lambda}^{\infty} xe^{-\lambda x}dx \\
&\sim \frac{1}{\sqrt{2\lambda}} \frac{2}{3} \left(\frac{1}{\lambda}\right)^{3/2} + \frac{2/e}{\lambda^2} \\
&= \left(\frac{\sqrt{2}}{3} + \frac{2}{e}\right) \frac{1}{\lambda^2} = n'
\end{aligned}$$

and

$$\sum_{i=1}^{n} y_i \log\left(1 + \frac{1}{y_i}\right) \sim \left(1 + \frac{2}{e}\right) \frac{1}{\lambda}.$$

Hence we have to choose

$$\lambda = \frac{\sqrt{\frac{\sqrt{2}}{3} + \frac{2}{e}}}{\sqrt{n'}}$$

and finally get the proposed bound. This completes the proof of Lemma 3. ∎

References

[1] K. Atteson, The Asymptotic Redundancy of Bayes Rules for Markov Chains, *IEEE Trans. on Information Theory*, 45, 2104–2109, 1999.

[2] A. Barron, J. Rissanen, and B. Yu, The Minimum Description Length Principle in Coding and Modeling, *IEEE Trans. Information Theory*, 44, 2743-2760, 1998.

[3] J. Bernardo, Reference Posterior Distributions for Bayesian Inference, *J. Roy. Stat. Soc. B.*, 41, 113–147, 1979.

[4] B. Clarke and A. Barron, Information-theoretic Asymptotics of Bayes Methods, *IEEE Trans. Information Theory*, 36, 453–471, 1990.

[5] T. Cover and J.A. Thomas, *Elements of Information Theory*, John Wiley & Sons, New York 1991.

[6] I. Csiszár and P. Shields, Redundancy Rates for Renewal and Other Processes, *IEEE Trans. Information Theory*, 42, 2065–2072, 1996.

[7] L. Davisson, Universal Noiseless Coding, *IEEE Trans. Inform. Theory*, 19, 783–795, 1973.

[8] L. Davisson and A. Leon-Garcia, A Source Matching Approach to Finding Minimax Codes, *IEEE Trans. Inform. Theory*, 26, 166–174, 1980.

[9] M. Drmota and R. Tichy, *Sequences, Discrepancies, and Applications*, Springer Verlag, Berlin Heidelberg, 1997.

[10] M. Drmota, H-K. Hwang, and W. Szpankowski, Precise Average Redundancy of an Idealized Arithmetic Coding, *Data Compression Conference*, 222-231, Snowbirds, 2002.

[11] P. Flajolet and W. Szpankowski, Analytic Variations on Redundancy Rates of Renewal Processes, *IEEE Trans. Information Theory*, 48, 2911 -2921, 2002.

[12] P. Jacquet and W. Szpankowski, Asymptotic Behavior of the Lempel-Ziv Parsing Scheme and Digital Search Trees, *Theoretical Computer Science*, 144, 161-197, 1995.

[13] P. Jacquet and W. Szpankowski, A Combinatorial Problem Arising in Information Theory: Precise Minimax Redundancy for Markov Sources *Proc. Colloquium on Mathematics and Computer Science II: Algorithms, Trees, Combinatorics and Probabilities*, 311-328, Birkhäuser, 2002.

[14] N. Jevtic, A. Orlitsky, and Prasad Santhanam, Universal compression of unknown alphabets, preprint (see also *IEEE International Symposium on Information Theory*, June 2002).

[15] I. Kontoyiannis, Pointwise Redundancy in Lossy Data Compression and Universal Lossy Data Compression, *IEEE Trans. Inform. Theory*, 46, 136-152, 2000.

[16] E. Krätzel, *Lattice Points*, Kluwer, Dordrecht, 1988.

[17] R. Krichevsky and V. Trofimov, The Performance of Universal Coding, *IEEE Trans. Information Theory*, 27, 199-207, 1983.

[18] G. Louchard and W. Szpankowski, On the Average Redundancy Rate of the Lempel-Ziv Code, *IEEE Trans. Information Theory*, 43, 2–8, 1997.

[19] N. Merhav and M. Feder, A Strong Version of the Redundancy-Capacity Theory of Universal Coding, *IEEE Trans. Information Theory*, 41, 714–722, 1995.

[20] J. Rissanen, Universal Coding, Information, Prediction, and Estimation, *IEEE Trans. Information Theory*, 30, 629–636, 1984.

[21] J. Rissanen, Fisher Information and Stochastic Complexity, *IEEE Trans. Information Theory*, 42, 40–47, 1996.

[22] S. Savari, Redundancy of the Lempel-Ziv Incremental Parsing Rule, *IEEE Trans. Information Theory*, 43, 9–21, 1997.

[23] P. Shields, Universal Redundancy Rates Do Not Exist, *IEEE Trans. Information Theory*, 39, 520-524, 1993.

[24] Y. Shtarkov, Universal Sequential Coding of Single Messages, *Problems of Information Transmission*, 23, 175–186, 1987.

[25] Y. Shtarkov, T. Tjalkens and F.M. Willems, Multi-alphabet Universal Coding of Memoryless Sources, *Problems of Information Transmission*, 31, 114-127, 1995.

[26] W. Szpankowski, On Asymptotics of Certain Recurrences Arising in Universal Coding, *Problems of Information Transmission*, 34, 55-61, 1998.

[27] W. Szpankowski, Asymptotic Redundancy of Huffman (and Other) Block Codes, *IEEE Trans. Information Theory*, 46, 2434-2443, 2000.

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

[29] Q. Xie, A. Barron, Minimax Redundancy for the Class of Memoryless Sources, *IEEE Trans. Information Theory*, 43, 647-657, 1997.

[30] Q. Xie, A. Barron, Asymptotic Minimax Regret for Data Compression, Gambling, and Prediction, *IEEE Trans. Information Theory*, 46, 431-445, 2000.

[31] P. Whittle, Some Distribution and Moment Formulæ for Markov Chain, *J. Roy. Stat. Soc.,* Ser. B., 17, 235–242, 1955.

[32] A. J. Wyner, The Redundancy and Distribution of the Phrase Lengths of the Fixed-Database Lempel-Ziv Algorithm, *IEEE Trans. Information Theory*, 43, 1439–1465, 1997.