CERIAS Tech Report 2002-12

Generalized Shannon Code Minimizes The Maximal Redundancy

Michael Drmota[2], Wojciech Szpankowski [1]
Center for Education and Research in
Information Assurance and Security
&
1Department of Computer Science, Purdue University,
West Lafayette, IN 47907
2Institut Für Geometrie

Generalized Shannon Code Minimizes the Maximal Redundancy

August 2, 2001

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

Source coding, also known as data compression, is an area of information theory that deals with the design and performance evaluation of optimal codes for data compression. In 1952 Huffman constructed his optimal code that minimizes the *average* code length among all prefix codes for known sources. Actually, Huffman codes minimizes the average *redundancy* defined as the difference between the code length and the entropy of the source. Interestingly enough, no optimal code is known for other popular optimization criterion such as the *maximal redundancy* defined as the maximum of the pointwise redundancy over all source sequences. We first prove that a generalized Shannon code minimizes the maximal redundancy among all prefix codes, and present an efficient implementation of the optimal code. Then we compute precisely its redundancy for memoryless sources. Finally, we study universal codes for unknown source distributions. We adopt the minimax approach and search for the best code for the worst source. We establish that such redundancy is a sum of the likelihood estimator and the redundancy of the generalize code computed for the maximum likelihood distribution. This replaces Shtarkov's bound by an exact formula. We also compute precisely the maximal minimax for a class of memoryless sources. The main findings of this paper are established by techniques that belong to the toolkit of the "analytic analysis of algorithms" such as theory of distribution of sequences modulo 1 and Fourier series. These methods have already found applications in other problems of information theory, and they constitute the so called *analytic information theory.*

1 Introduction

The celebrated Huffman code minimizes the average code length among all prefix codes (i.e., satisfying the Kraft inequality), provided the probability distribution is known. As a matter of fact, the Huffman code minimizes the average *redundancy* that is defined as the difference between the code length and the entropy for the source. But other than the average redundancy optimization criteria were also considered in information theory. The most popular (cf. Shtarkov [10]) is the maximal redundancy defined as the maximum over all source sequences of the sum of the code length and the logarithm of the probability of source sequences. A seemingly innocent, and still open, problem is what code minimizes the maximal redundancy. To make it more precise we need to plunge a little into source coding, better known as data compression.

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

We start with a quick introduction of the *redundancy problem*. A code $C_n : \mathcal{A}^n \to \{0,1\}^*$ is defined as a mapping from the set $\mathcal{A}^n$ of all sequences of length n over the finite alphabet $\mathcal{A}$ to the set $\{0,1\}^*$ of all binary sequences. A message of length n with letters indexed from 1 to n is denoted by x_1^n, so that $x_1^n \in \mathcal{A}^n$. We write X_1^n to denote the random variable representing a message of length n. Given a probabilistic source model, we let $P(x_1^n)$ be the probability of the message x_1^n; 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 $\log := \ln$.

From Shannon's works 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. Hence $-\lg P(x_1^n)$ can be viewed as the "ideal" code length. The next natural question is to ask by how much the code length $L(C_n, x_1^n)$ differs from the ideal code length, either for individual sequences or on average. The *pointwise redundancy* $R_n(C_n, P; x_1^n)$ and the *average redundancy* $\overline{R}_n(C_n, P)$ are 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 the underlying probability measure P represents a particular source model and $\mathbf{E}$ denotes the expectation. Another natural measure of code performance is the *maximal* redundancy defined as

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

While the pointwise redundancy can be negative, maximal and average redundancies cannot, by Kraft's inequality and Shannon's source coding theorem, respectively (cf. [2]).

Source coding is an area of information theory that searches for optimal codes under various optimization criteria. It has been known from the inception of the Huffman code (cf. [2]) that its average redundancy is bounded from above by 1, but its precise characterization for memoryless sources was proposed only recently in [12]. In [3, 7, 9] conditions for optimality of the Huffman code were given for a class of weight function and cost criteria. Surprisingly enough, to the best of our knowledge, no one was looking at another natural question: What code minimizes the maximal redundancy? More precisely, we seek a prefix code C_n such that

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

We shall prove in this paper, that a generalized Shannon code[†] is the optimal code in this case, and propose an efficient algorithm to construct such a code. Our algorithm runs in $O(N \log N)$ steps if source probabilities are not sorted and in $O(N)$ steps if the probabilities are sorted, where N is the number of source sequences. We also compute precisely the maximal redundancy of the optimal generalized Shannon code. In passing we observe that Shannon codes, in one form or another, are often used in practice; e.g., in arithmetic coder.

It must be said, however, that in practice probability distribution (i.e., source) P is unknown. So the next natural question is to find optimal codes for sources with unknown probabilities. In information theory this is handled by the so called *minimax* redundancy,

[†] Shannon's code assigns length $\lceil -\lg P(x_1^n) \rceil$ to the source sequence x_1^n for known source distribution P.

that we introduce next. In fact, for unknown probabilities, the redundancy rate can be also viewed as the penalty paid for estimating the underlying probability measure. More precisely, *universal codes* are those for which the redundancy is $o(n)$ for all $P \in \mathcal{S}$ where $\mathcal{S}$ is a class of source models (distributions). The (asymptotic) *redundancy-rate problem* consists in determining for a class $\mathcal{S}$ the rate of growth of the minimax quantities as $n \to \infty$ either on average

$$\overline{R}_n(\mathcal{S}) = \min_{C_n \in \mathcal{C}} \max_{P \in \mathcal{S}} [\overline{R}_n(C_n, P)], \tag{1}$$

or in the worst case

$$R_n^*(\mathcal{S}) = \min_{C_n \in \mathcal{C}} \max_{P \in \mathcal{S}} [R_n^*(C_n, P)], \tag{2}$$

where $\mathcal{C}$ denotes the set of all codes satisfying the Kraft inequality.

In this paper we deal with the maximal *minimax redundancy* $R_n^*(\mathcal{S})$ defined by (2). Shtarkov [10] proved that

$$\lg\left(\sum_{x_1^n} \sup_{P \in \mathcal{S}} P(x_1^n)\right) \le R_n^*(\mathcal{S}) \le \lg\left(\sum_{x_1^n} \sup_{P \in \mathcal{S}} P(x_1^n)\right) + 1. \tag{3}$$

We replace the inequalities in the above by an exact formula. Namely, we shall prove that

$$R_n^*(\mathcal{S}) = \lg\left(\sum_{x_1^n} \sup_{P \in \mathcal{S}} P(x_1^n)\right) + R^{GS}(Q^*)$$

where $R^{GS}(Q^*)$ is the maximal redundancy of the generalized Shannon code for the (known) distribution $Q^*(x_1^n) = \sup_P P(x_1^n) / \sum_{x_1^n} \sup_P P(x_1^n)$. For a class of memoryless sources we derive an asymptotic expansion for the maximal minimax redundancy $R_n^*(\mathcal{S})$.

2 Main Result

We first consider sources with known distribution P and find an optimal code that minimizes the maximal redundancy, that is, we compute

$$R_n^*(P) = \min_{C_n \in \mathcal{C}} \max_{x_1^n} [L(C_n, x_1^n) + \log_2 P(x_1^n)]. \tag{4}$$

We recall that Shannon code C_n^S assigns length $L(C_n^S, x_1^n) = \lceil -\lg P(x_1^n) \rceil$ to the source sequence x_1^n. We define a *generalized Shannon* code C_n^{GS} as

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

where $\mathcal{L} \subset \mathcal{A}^n$, and the Kraft inequality holds.

Our first main result proves that a generalized Shannon code is an optimal code with respect to the maximal redundancy.

Theorem 1 *If the probability distribution P is dyadic, i.e. $\lg P(x_1^n) \in \mathbf{Z}$ ($\mathbf{Z}$ is the set of integers) for all $x_1^n \in \mathcal{A}^n$, then $R_n^*(P) = 0$. Otherwise, let $p_1, p_2, \ldots, p_{|\mathcal{A}|^n}$ be the probabilities $P(x_1^n)$, $x_1^n \in \mathcal{A}^n$, ordered in a nondecreasing manner, that is,*

$$0 \le \langle -\lg p_1 \rangle \le \langle -\lg p_2 \rangle \le \cdots \le \langle -\lg p_{|\mathcal{A}|^n} \rangle \le 1,$$

where $\langle x \rangle = x - \lfloor x \rfloor$ is the fractional part of x. Let now j_0 be the maximal j such that

$$\sum_{i=1}^{j-1} p_i 2^{\langle -\lg p_i \rangle} + \frac{1}{2} \sum_{i=j}^{|\mathcal{A}|^n} p_i 2^{\langle -\lg p_i \rangle} \le 1, \tag{5}$$

that is, the Kraft inequality holds for a generalized Shannon code. Then

$$R_n^*(P) = 1 - \langle -\lg p_{j_0} \rangle. \tag{6}$$

Proof. First we want to recall that we are only considering codes satisfying Kraft's inequality

$$\sum_{x_1^n} 2^{-L(C_n, x_1^n)} \le 1.$$

Especially we will use the fact that for any choice of positive integers $l_1, l_2, \ldots, l_{|\mathcal{A}|^n}$ with

$$\sum_{i=1}^{|\mathcal{A}|^n} 2^{-l_i} \le 1$$

there exists a (prefix) code C_n with code lengths l_i, $1 \le i \le |\mathcal{A}|^n$.

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

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

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

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

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

The idea of the proof is to find some properties of the optimal code. Especially we will show that there exists an optimal code $C_n^* \in \mathcal{C}^*$ with

(i)
$$\lfloor -\lg P(x_1^n) \rfloor \le L(C_n^*, x_1^n) \le \lceil -\lg P(x_1^n) \rceil \tag{7}$$

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

$$L(C_n^*, x_1^n) = \lfloor \lg 1/P(x_1^n) \rfloor \quad \text{if} \quad \langle \lg 1/P(x_1^n) \rangle < s_0 \tag{8}$$

and

$$L(C_n^*, x_1^n) = \lceil \lg 1/P(x_1^n) \rceil \quad \text{if} \quad \langle \lg 1/P(x_1^n) \rangle \ge s_0, \tag{9}$$

that is, C_n^* is the generalized Shannon code. 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 (8) and (9). It is clear that (5) is just Kraft's inequality for codes of that kind. The optimal choice is $j = j_0$ and consequently $R_n^*(P) = 1 - \langle -\lg p_{j_0} \rangle$.

It remains to prove the above properties (i) and (ii). Assume that C_n^* is an optimal code. First of all, the upper bound in (7) is obviously satisfied for C_n^*. Otherwise we would have

$$\max_{x_1^n}[L(C_n^*, x_1^n) + \log_2 P(x_1^n)] > 1$$

which contradicts Shtarkov's bound (3). 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, $\widetilde{C}_n^*$ is optimal, too. This proves (i).

Now consider an optimal code C_n^* satisfying (7) and let x_1^{n*} be a sequence with $R_n^*(P) = 1 - \langle -\lg P(x_1^{n*}) \rangle$. Thus, $L(C_n^*, x_1^n) = \lfloor \lg 1/P(x_1^n) \rfloor$ for all x_1^n with $\langle -\lg P(x_1^n) \rangle < \langle -\lg P(x_1^{n*}) \rangle$. This proves (8) with $s_0 = \langle -\lg P(x_1^{n*}) \rangle$. Finally, if (9) 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). ■

Thus, we proved that the following generalized Shannon code code is the desired optimal code and it satisfies

$$L(C_n^{GS}, x_1^n) = \begin{cases} \lfloor \lg 1/P(x_1^n) \rfloor & \text{if} \quad x_1^n \in \mathcal{L}_{s_0} \\ \lceil \lg 1/P(x_1^n) \rceil & \text{if} \quad x_1^n \in \mathcal{A}^n \setminus \mathcal{L}_{s_0}, \end{cases}$$

where

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

and $s_0 = \langle -\lg p_{j_0} \rangle$ is defined in (5).

The next question is how to construct efficiently the optimal generalized Shannon code? This turns out to be quite simple due to property (ii) (cf. (8) and (9)). The algorithm is presented below.

Algorithm GS–CODE

Input: Probabilities $P(x_1^n)$.
Output: Optimal generalized Shannon code.
1. Let $s_i = \langle -\lg P(x_1^n) \rangle$ for $i = 1, 2, \ldots, N$, where $N \le |\mathcal{A}^n|$.
2. Sort $s_1, \ldots, s_N$.
3. Use *binary search* to find the largest j_0 such that (5) holds, and set $s_0 = 1 - s_{j_0} = 1 - \langle -\lg p_{j_0} \rangle$.

5. Set code length $l_i = \lfloor -\lg p_i \rfloor$ for $i \le j_0$, otherwise $l_i = \lceil -\lg p_i \rceil$.
end

Observe that property (ii) above was crucial to justify the application of the binary search in Step 3 of the algorithm. Obviously, Step 2 requires $O(N \log N)$ operations which determines the complexity of the algorithm. If probabilities are sorted, then the complexity is determined by Step 5 and it is equal to $O(N)$, as for the Huffman code construction (cf. [8]).

Now, we turn our attention to universal codes for which the probability distribution P is unknown. We assume that P belongs to a set $\mathcal{S}$ (e.g., class of memoryless sources with unknown parameters). The following result summarizes our next finding. It transforms the Shtarkov bound (3) into an equality.

Theorem 2 *Suppose that $\mathcal{S}$ is a system of probability distributions P on $\mathcal{A}^n$ and set*

$$Q^*(x_1^n) := \frac{\sup_{P\in\mathcal{S}} P(x_1^n)}{\sum_{y_1^n\in\mathcal{A}^n} \sup_{P\in\mathcal{S}} P(y_1^n)}.$$

If the probability distribution Q^* is dyadic, i.e. $\lg Q^*(x_1^n) \in \mathbf{Z}$ for all $x_1^n \in \mathcal{A}^n$, then

$$R_n^*(\mathcal{S}) = \lg\left(\sum_{x_1^n\in\mathcal{A}^n} \sup_{P\in\mathcal{S}} P(x_1^n)\right). \tag{10}$$

Otherwise, let $q_1, q_2, \ldots, q_{|\mathcal{A}|^n}$ be the probabilities $Q^*(x_1^n)$, $x_1^n \in \mathcal{A}^n$, ordered in such a way that

$$0 \le \langle -\lg q_1\rangle \le \langle -\lg q_2\rangle \le \cdots \le \langle -\lg q_{|\mathcal{A}|^n}\rangle \le 1,$$

and let j_0 be the maximal j such that

$$\sum_{i=1}^{j-1} q_i 2^{\langle -\lg q_i\rangle} + \frac{1}{2}\sum_{i=j}^{|\mathcal{A}|^n} q_i 2^{\langle -\lg q_i\rangle} \le 1. \tag{11}$$

Then

$$R_n^*(\mathcal{S}) = \lg\left(\sum_{x_1^n\in\mathcal{A}^n} \sup_{P\in\mathcal{S}} P(x_1^n)\right) + R_n^*(Q^*), \tag{12}$$

where $R_n^*(Q^*) = 1 - \langle -\lg q_{j_0}\rangle$ is the maximal redundancy of the optimal generalized Shannon code designed for the distribution Q^*.

Proof. By definition we have

$$\begin{aligned}
R_n^*(\mathcal{S}) &= \min_{C_n\in\mathcal{C}} \sup_{P\in\mathcal{S}} \max_{x_1^n} (L(C_n, x_1^n) + \lg P(x_1^n)) \\
&= \min_{C_n\in\mathcal{C}} \max_{x_1^n} \left(L(C_n, x_1^n) + \sup_{P\in\mathcal{S}} \lg P(x_1^n) \right) \\
&= \min_{C_n\in\mathcal{C}} \max_{x_1^n} \left(L(C_n, x_1^n) + \lg Q^*(x_1^n) + \lg\left(\sum_{y_1^n\in\mathcal{A}^n} \sup_{P\in\mathcal{S}} P(y_1^n)\right)\right) \\
&= R_n^*(Q^*) + \lg\left(\sum_{y_1^n\in\mathcal{A}^n} \sup_{P\in\mathcal{S}} P(y_1^n)\right),
\end{aligned}$$

where $R_n^*(Q^*) = 1 - \langle -\lg q_{j_0} \rangle$, and by Theorem 1 it can be interpreted as the maximal redundancy of the optimal generalized Shannon code designed for the distribution Q^*. Theorem 2 is proved. ∎

3 Memoryless Sources

Let us consider a binary memoryless source with $P_p(x_1^n) = p^k(1-p)^{n-k}$ where k is the number of "0" in x_1^n and p is the probability of generating a "0". In the next theorem we compute the maximal redundancy $R_n^*(P_p)$ of the optimal generalized Shannon code assuming p is *known*.

Theorem 3 *Suppose that* $\lg \frac{1-p}{p}$ *is irrational. Then as* $n \to \infty$

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

If $\lg \frac{1-p}{p} = \frac{N}{M}$ *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* $R_n^*(P_{1/2}) = 0$.

Proof. Set

$$\begin{aligned} \alpha_p &= \lg \frac{1-p}{p}, \\ \beta_p &= \lg \frac{1}{1-p}. \end{aligned}$$

Then

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

First we assume that α_p is irrational. We know from [12] that for every Riemann integrable function $f : [0,1] \to \mathbf{R}$ we have

$$\lim_{n\to\infty} \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} f(\langle \alpha_p k + \beta_p n \rangle) = \int_0^1 f(x)\, dx. \tag{13}$$

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 obtain

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

In particular, for

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

we get $\int_0^1 f(x)\, dx = 1$ so that (5) holds. This implies that

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

If $\alpha_p = \frac{N}{M}$ is rational and non-zero then we have (cf. [12] or [13] Chap. 8)

$$\lim_{n\to\infty} \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} f(\langle \alpha_p k + \beta_p n \rangle) = \frac{1}{M} \sum_{m=0}^{M-1} f\left(\left\langle \frac{mN}{M} + \beta_p n \right\rangle\right) \tag{14}$$

$$= \frac{1}{M} \sum_{m=0}^{M-1} f\left(\frac{m + \langle M\beta_p n \rangle}{M}\right). \tag{15}$$

Of course, we have to 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 maximal m_0 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 the theorem. ∎

The next step is to consider memoryless sources P_p such that p is *unknown* and say contained in an interval $[a, b]$, i.e. we restrict $\mathcal{S}_{ab}$ to the class of memoryless sources with $p \in [a, b]$. Here, the result reads as follows.

Theorem 4 *Let* $0 \le a < b \le 1$ *be given and let* $\mathcal{S}_{a,b} = \{P_p : a \le p \le b\}$. *Then as* $n \to \infty$

$$R_n^*(\mathcal{S}_{a,b}) = \frac{1}{2} \lg n + \lg\, C_{a,b} - \frac{\log\log 2}{\log 2} + o(1), \tag{16}$$

where

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

Proof. First observe that

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

By Theorem 2 we must evaluate $T_n = \sum_{x_1^n} \sup_{P\in\mathcal{S}_{ab}} P(x_1^n)$, which becomes

$$T_n := \sum_{k<na} \binom{n}{k} a^k(1-a)^{n-k} + \sum_{na\le k\le nb} \binom{n}{k} \left(\frac{k}{n}\right)^k \left(1-\frac{k}{n}\right)^{n-k} + \sum_{k>nb} \binom{n}{k} b^k(1-b)^{n-k}.$$

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 $an \le k \le bn$)

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

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

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

and

$$\lg T_n = \frac{1}{2}\lg n + \lg C_{a,b} + O(n^{-1/2}).$$

To complete the proof we must evaluate the redundancy $R_n^*(Q^*)$ of the optimal generalized Shannon code designed for the maximum likelihood distribution Q^*. We proceed as in the proof of Theorem 3, and define a function $f_{s_0} = 2^x$ for $x \le s_0$ and otherwise $f_{s_0} = 2^{x-1}$. In short, $f_{s_0}(x) = 2^{-\langle s_0 - x\rangle + s_0}$ (now considered as a periodic function with period 1). The problem is to evaluate the sum (cf. (11))

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

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, is remains to consider S_2.

We will use the property that for every (Riemann integrable) function $f : [0,1] \to \mathbf{C}$ and for every sequence $x_{n,k}$, $an \le k \le bn$, of the kind

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

where c_n is an arbitrary sequence, we have

$$\lim_{n\to\infty} \frac{1}{T_n} \sum_{an \le k \le bn} \binom{n}{k} \left(\frac{k}{n}\right)^k \left(1-\frac{k}{n}\right)^{n-k} f(\langle x_{n,k}\rangle) = \int_0^1 f(x)\,dx. \tag{17}$$

Note that we are now in a similar situation as in the proof of Theorem 3. We apply (17) with $f_{s_0}(x)$ for $s_0 = -\log\log 2/\log 2$, and (16) follows.

For the proof of (17), we verify the Weyl criteria (cf. [4, 13]), that is, we first consider the following exponential sums

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

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 [6, 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 conclude

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

and consequently

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

Note that all these estimates are uniform for $c \in [a,b]$. Next we consider exponential sums

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

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 get (uniformly for $an \le k \le bn$) $a_{n,k} \ll n^{-1/2}$ and

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

Thus, by Abel's partial summation (cf. [13])

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

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

$$\lim_{n\to\infty} \frac{1}{T_n} \sum_{an \le k \le bn} a_{n,k} e(hx_{n,k}) = 0. \tag{18}$$

Consequently, by standard tools in Fourier analysis (18) implies (for every Riemann integrable function $f : [0,1] \to \mathbf{C}$)

$$\lim_{n\to\infty} \frac{1}{T_n} \sum_{an \le k \le bn} a_{n,k} f(\langle x_{n,k} \rangle) = A_0(f),$$

where A_0 is the zero-th Fourier coefficient

$$A_0 = \int_0^1 f(x)\, dx.$$

This means that we have proved (17). ■

Remark 1. We can derive a full asymptotic expansion for the maximal minimax redundancy $R_n^*(\mathcal{S})$ for memoryless sources. Indeed, for a change consider an m–ary alphabet $\mathcal{A}$ ($m \ge 2$). Following the footsteps of the above derivation, and using the approach from [11] for $p \in (0,1)$, we arrive at

$$\begin{aligned} R_n^*(\mathcal{S}) &= \frac{m-1}{2} \log\left(\frac{n}{2}\right) - \frac{\ln \frac{1}{m-1} \ln m}{\ln m} + \log\left(\frac{\sqrt{\pi}}{\Gamma(\frac{m}{2})}\right) + \frac{\Gamma(\frac{m}{2})m}{3\Gamma(\frac{m}{2} - \frac{1}{2})} \cdot \frac{\sqrt{2}}{\sqrt{n}} \\ &+ \left(\frac{3 + m(m-2)(2m+1)}{36} - \frac{\Gamma^2(\frac{m}{2})m^2}{9\Gamma^2(\frac{m}{2} - \frac{1}{2})}\right) \cdot \frac{1}{n} + O\left(\frac{1}{n^{3/2}}\right) \end{aligned}$$

for large n. To the best of our knowledge, the above formula is the first asymptotic expansion with the correct constant term (i.e., containing the term $R_n^*(Q^*)$). This is of some importance since some authors (cf. [14]) propose to design optimal codes that optimize the constant term.

Remark 2. Parker [9] (cf. also [3, 7]) investigated other than average cost functions but such for which the Huffman construction still produces optimal code. For example, Campbell [3] shown that Huffman's code is optimal if the average code length is replaced by

$$W(r) = \frac{1}{r}\log_m\left(\sum_{x_1^n} P(x_1^n) m^{rL(x_1^n)}\right)$$

where $m = |\mathcal{A}|$, $r > 0$ is any positive number, and $L(x_1^n)$ is the code length. Observe that $\lim_{r\to 0} W(r) = \mathbf{E}[L(C_n, X_1^n)]$, while $\lim_{r\to\infty} W(r) = \max_{x_1^n} L(C_n, x_1^n)$ $(= \lceil \log_m N \rceil)$. In this paper, we proved that when the $\max_{x_1^n} L(C_n, x_1^n)$ is replaced by the maximal redundancy $R_n^* = \max_{x_1^n}[L(C_n, x_1^n) + \log P(x_1^n)]$, then the Huffman code is not any more optimal. In general, let us define the r-th redundancy R_n^r $(r > 0)$ as

$$R_n^r = \left(\sum_{x_1^n} P(x_1^n)\left[L(C_n, x_1^n) + \log P(x_1^n)\right]^r\right)^{1/r}.$$

Observe that the average redundancy is $\overline{R}_n = R_n^1$, while the maximal redundancy is $R_n^* = R_n^\infty$. The open question is what code minimizes the r-th redundancy R_n^r?

References

[1] J. Abrahams, "Code and Parse Trees for Lossless Source Encoding, *Proc. of Compression and Complexity of SEQUENCE'97*, Positano, IEEE Press, 145–171, 1998.

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

[3] L. Campbell, A Coding Theorem and Rényi's Entropy, *Information and Control*, 8, 423–429, 1965.

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

[5] D. E. Knuth, Dynamic Huffman Coding, *J. Algorithms*, 6, 163-180, 1985.

[6] E. Krätzel, Lattice Points, Kluwer, Dordrecht, 1988.

[7] On a Coding Theorem Connected with Rényi's Entropy, *Information and Control*, 29, 234–242, 1975.

[8] J. van Leeuwen, On the Construction of the Huffman Trees, *Proc. ICALP'76*, 382–410, 1976.

[9] D. S. Parker, Conditions for Optimiality of the Huffman Algorithm, *SIAM J. Compt.*, 9, 470–489, 1980.

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

[11] W. Szpankowski, On Asymptotics of Certain Recurrences Arising in Universal Coding, *Problems of Information Transmission*, 34, No.2, 55-61, 1998.

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

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

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