The Precise Minimax Redundancy

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

1 Introduction

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 $x_1^n = (x_1, \ldots, x_n)$ of length n over the finite alphabet $\mathcal{A}$ to the set $\{0,1\}^*$ of all binary sequences. 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.

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, where $\lg := \log_2$ denotes the binary logarithm. 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 length $L(C_n, x_1^n)$ of a code differs from the ideal code length, either for individual sequences or on average. The *pointwise redundancy* $R_n(C_n, P; x_1^n) = L(C_n, x_1^n) + \lg P(x_1^n)$, while the *average redundancy* $\overline{R}_n(C_n, P)$ and the *maximal* redundancy $R_n^*(C_n, P)$ are defined, respectively, as

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

where the underlying probability measure P represents a particular source model and $\mathbf{E}$ denotes the expectation. 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.

It has been known from the inception of the Huffman code (cf. [3]) that its average redundancy is bounded from above by 1, but its precise characterization for memoryless sources was proposed only recently in [11]. In [2] 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 here (cf. Theorem 1), that a generalized Shannon code is the optimal code in this case. We also compute precisely the maximal redundancy of the optimal generalized Shannon code for memoryless sources (cf. Theorem 4).

It must be said, however, that in practice the probability distribution (i.e., source) P is unknown. So the next question is to find optimal codes for sources with unknown probabilities. In fact, for unknown probabilities, the redundancy rate can be also viewed as the penalty paid

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

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.

First, we deal with the maximal *minimax redundancy* $R_n^*(\mathcal{S})$ defined by (2). Shtarkov [9] proved that

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

We replace the inequalities in the above by an exact formula. Namely, we shall prove in Theorem 2 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 a properly chosen 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 $R_n^*(\mathcal{S})$ (cf. Theorem 5).

Finally, we deal with the *most challenging* problem. We have just argued that we can derive precise formula for the *maximal* minimax redundancy $R_n^*(\mathcal{S})$. Can we infer from this an asymptotic expansion for the *average* minimax redundancy $\overline{R}_n(\mathcal{S})$ which is *much harder* to evaluate? We prove in Theorem 3 that under certain additional conditions (i.e, $\sum_{x_1^n} P(x_1^n) \lg[\sup P(x_1^n)/P(x_1^n)]$ is of order magnitude smaller than the leading term of $R_n^*(\mathcal{S})$, that is, $\lg\left(\sum_{x_1^n} \sup_{P \in \mathcal{S}} P(x_1^n)\right)$) the following holds: $\overline{R}_n(\mathcal{S}) \sim R_n^*(\mathcal{S})$. We also provide asymptotics of the average redundancy for memoryless sources (cf. Theorem 6).

2 The Maximal Minimax Redundancy

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 $L(C_n^S, x_1^n) = \lceil \lg 1/P(x_1^n) \rceil$ 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. Note that Kraft's inequality for generalized Shannon codes reads as

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

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

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

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*

$$Q^*(x_1^n) := \frac{1}{c_n(\mathcal{S})} \sup_{P \in \mathcal{S}} P(x_1^n),$$

where

$$c_n(\mathcal{S}) = \sum_{y_1^n \in \mathcal{A}^n} \sup_{P \in \mathcal{S}} P(y_1^n).$$

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

$$R_n^*(\mathcal{S}) = \lg c_n(\mathcal{S}). \tag{7}$$

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

$$\sum_{i=1}^{j-1} q_i 2^{\langle -\lg q_i \rangle} + \frac{1}{2} \sum_{i=j}^{|\mathcal{A}|^n} q_i 2^{\langle -\lg q_i \rangle} \le 1. \tag{8}$$

Then

$$R_n^*(\mathcal{S}) = \lg c_n(\mathcal{S}) + R_n^*(Q^*), \tag{9}$$

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

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

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^*. ∎

3 The Average Minimax Redundancy

Now, we study the *average* minimax redundancy $\overline{R}_n$ which is defined in (1). If $\mathcal{S}$ consist of just one probability measure P one recognizes that the optimal code is the Huffman code. Our ultimate goal is to establish a general precise result for the average minimax redundancy $\overline{R}_n(\mathcal{S})$ for non-trivial $\mathcal{S}$. From known results (cf. [1, 4, 5, 6, 7, 8, 9, 10, 11, 13, 14]) we conclude that (in these particular cases) $\overline{R}_n(\mathcal{S}) \sim R_n^*(\mathcal{S})$. So our aim is to provide a quite general result in this direction. We can prove the following result.

Theorem 3 *Suppose that $\mathcal{S}$ is a system of probability distributions P on $\mathcal{A}^n$. Then*

$$\overline{R}_n(\mathcal{S}) \le \lg c_n(\mathcal{S}) - \inf_{P\in\mathcal{S}} \left(\sum_{x_1^n} P(x_1^n) \lg \frac{\sup_{P\in\mathcal{S}} P(x_1^n)}{P(x_1^n)} \right) + O(1). \tag{10}$$

Furthermore, suppose that there exists a probability distribution $\widetilde{Q}$ in the convex hull† of $\mathcal{S}$ such that

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

then

$$\overline{R}_n(\mathcal{S}) \ge \lg c_n(\mathcal{S}) - \sup_{P\in\mathcal{S}} \left(\sum_{x_1^n} P(x_1^n) \lg \frac{\sup_{P\in\mathcal{S}} P(x_1^n)}{P(x_1^n)} \right) - C + O(1). \tag{11}$$

We will show in Section 4 that these lower and upper bounds fit together quite well.

The proof of the upper bound is easy. The lower bounds are a little bit more involved. It mainly relies on the following interesting fact (we are not aware of any relevant reference).

Lemma 1 *Suppose that $\mathcal{S}$ is a subset of probability distributions P on a finite set X. Then for all probability distributions $\tilde{Q}$ contained in the convex hull of $\mathcal{S}$ we have*

$$\inf_{Q} \sup_{P\in\mathcal{S}} \left(\sum_{x\in X} P(x) \lg \frac{\tilde{Q}(x)}{Q(x)} \right) = 0. \tag{12}$$

†We assume no topology on the set of all probability measures on X. Therefore the convex hull of $\mathcal{S}$ is just the set of all finite convex combinations of elements of $\mathcal{S}$.

4 Memoryless Sources

Finally, we illustrate our findings for memoryless sources. We 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".

It should be mentioned that all subsequent results can be generalized to memoryless sources and to Markov sources with an arbitrary finite alphabet. However, for the sake of brevity and transparency of we have decided to treat just memoryless sources with the binary alphabet.

4.1 The Maximal Redundancy of the Generalized Shannon Code

We start with the following result of the maximal redundancy for the optimal generalized Shannon code. We give a detailed proof of this result in the Appendix.

Theorem 4 *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* $n \to \infty$,

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

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

The next step is to consider memoryless sources P_p such that p is contained in an interval $[a, b]$, i.e. we restrict on a quite special (but natural) case for $\mathcal{S}$. Here the result reads as follows.

Theorem 5 *Let* $0 \le a < b \le 1$ *be given and let* $\mathcal{S}_{a,b} = \{P_p : a \le p \le b\}$. *Then, as* $n \to \infty$.

$$R_n^*(\mathcal{S}_{a,b}) = \frac{1}{2}\lg n + \lg C_{a,b} - \frac{\log\log 2}{\log 2} + o(1), \tag{13}$$

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. Expression (13) is the first asymptotic expansion with the correct constant term (i.e., containing the term $R_n^*(Q^*)$). This is of some importance since it is proposed (cf. [13]) to design optimal codes that optimize the constant term.

4.2 The Average Redundancy of Memoryless Sources

By applying our general Theorem 3 we obtain the following slightly generalized result for the average redundancy of memoryless sources:

Theorem 6 *Let* $0 \le a < b \le 1$ *be given and let* $\mathcal{S}_{a,b} = \{P_p : a \le p \le b\}$. *Then, as* $n \to \infty$.

$$\overline{R}_n(\mathcal{S}_{a,b}) = \frac{1}{2}\lg n + O(1). \tag{14}$$

Remark: Note that this result has been known for $\mathcal{S}_{0,1}$ (cf. [1, 6, 10, 13]).
Sketch of Proof. There are mainly two things which have to be checked. First we note that

$$\sup_{P\in\mathcal{S}_{a,b}}\left(\sum_{x_1^n} P(x_1^n)\lg\frac{\sup_{P\in\mathcal{S}_{a,b}} P(x_1^n)}{P(x_1^n)}\right) = O(1),$$

which is implied by

$$\sum_{k=0}^{n}\binom{n}{k}p^k(1-p)^{n-k}\lg\frac{\left(\frac{k}{n}\right)^k\left(1-\frac{k}{n}\right)^{n-k}}{p^k(1-p)^k} \le \frac{1}{\log 2}. \tag{15}$$

But (15) follows directly by applying the inequality $\log x \le x-1$ for

$$\log\frac{k/n}{p} \le \frac{k/n}{p} - 1 \quad \text{and} \quad \log\frac{1-k/n}{1-p} \le \frac{1-k/n}{1-p} - 1.$$

Second we have to show that there exists a convex combination $\tilde{Q}$ of the probability distributions $P_{k/n}$ $(an \le k \le bn)$ such that, as $n\to\infty$

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

For example, if $0<a<b<1$ we can use

$$\tilde{Q} = \frac{1}{B}\sum_{an\le k\le bn}\beta_k P_{k/n},$$

where

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

and $B=\sum_k\beta_k$. ∎

References

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

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

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

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

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

[6] J. Rissanen, Complexity of Strings in the Class of Markov Sources, *IEEE Trans. Information Theory*, 30, 526–532, 1984.

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

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

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

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

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

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

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

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

Appendix: Proof of Theorem 4

Set $\alpha_p = \lg \frac{1-p}{p}$ and $\beta_p = \lg \frac{1}{1-p}$. 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 [11] that for every Riemann integrable function $f : [0,1] \to \mathbf{R}$ we have

$$\lim_{n\to\infty} \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} f(\langle \alpha_p k + \beta_p n \rangle) = \int_0^1 f(x)\, dx. \tag{16}$$

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. [11] or [12] Chap. 8)

$$\begin{aligned} \lim_{n\to\infty} \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} f(\langle \alpha_p k + \beta_p n \rangle) &= \frac{1}{M} \sum_{m=0}^{M-1} f\left(\left\langle \frac{mN}{M} + \beta_p n \right\rangle \right) & (17) \\ &= \frac{1}{M} \sum_{m=0}^{M-1} f\left(\frac{m + \langle M\beta_p n \rangle}{M} \right). & (18) \end{aligned}$$

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. ∎