Algorithms Seminar 1999–2000,
F. Chyzak (ed.), INRIA, (2000), pp. 133–136.

Available online at the URL
http://algo.inria.fr/seminars/.

Analytic Information Theory and the Redundancy Rate Problem

Wojciech Szpankowski
Department of Computer Sciences, Purdue University

February 13, 2000

Summary by Philippe Flajolet

1. Information, Entropy, and Codes

One of the most basic problems of information theory [1] is that of *source coding.* A source is by definition a mechanism that produces messages over a finite alphabet $\mathcal{A}$, a message of length n being conventionally denoted by $x_1^n = (x_1, \dots, x_n)$. A code C is a translation mechanism (an injective function, an algorithm) that, for each n, takes as input a message from $\mathcal{A}^n$ and transforms it into a binary sequence. Such a translation is thus a fixed-length to variable-length encoding.

Messages have some structure. For the English language source, the sequence 'Rzqxwa gkvzzxq wzd aaaaaaa rxbleurp' is much less likely than the sequence 'It rained yesterday over England'. Indeed, some letters are more frequent than others, certain letter combinations are impossible, etc. It is then customary to try and capture the principal features of the source by some probability distribution of sorts over $\mathcal{A}^n$. The main models considered in the talk are the following.

M1. A *memoryless model* considers letters as independent identically distributed random variables, with letter $i \in \mathcal{A}$ having probability p_i. (This is sometimes called the Bernoulli model.)

M2. A *Markov model* assumes an underlying finite set of states with transition probability $p_{i,j}$ between states i and j and a mapping from states to letters.

As discovered by Shannon around 1949, information is measured by entropy. The entropy of a probability distribution $P = \{p_s\}_{s\in S}$ over any finite set S is defined as

$$H(P) := -\sum_{s\in S} p_s \lg p_s,$$

where $\lg x = \log_2 x$. (Roughly, the definition extends the fact that an element in a set of cardinality m needs to be encoded by about $\lg m$ bits in order to be distinguished from its companions elements.) Most "reasonable" source models have an *entropy rate* h; namely, if x_1^n is randomly drawn according to the source model P, then the following limit exists,

$$h = \lim_{n\to\infty} -\frac{1}{n} \sum_{x_1^n \in \mathcal{A}^n} P(x_1^n) \lg P(x_1^n).$$

For instance, the entropy rate of a sequence drawn according to the memoryless model equals the entropy of the distribution of individual characters. For a Markov chain with transition probabilities $p_{i,j}$, the entropy rate is

$$h = \sum_i \pi_i \sum_j p_{i,j} \lg p_{i,j},$$

with π_i the stationary probability distribution of the chain. The entropy rate of written English is estimated to be about 1.3 bits per character.

Sources produce messages which are not uniformly random and this lies at the basis of data compression—the fact that one may find codes that tend to be shorter than the original message. (E.g., the present summary is compressed by `gzip` at a rate of about 3.5 bits per character.) We cannot compress arbitrarily however. The most fundamental theorem of information is due to Shannon and asserts the following: *You cannot beat entropy. In other words, any code has an expected length per character that is at least as large as the entropy rate of the source.*

Another famous theorem of Shannon goes the other direction and asserts: *The entropy rate is asymptotically achievable.* This leaves plenty of room for algorithmic design. As a matter of fact, coding algorithms separate into two groups: *(i)* codes that are designed for a specific (known) probability distribution over the inputs; *(ii)* universal codes that do not assume such a probabilistic distribution to be known *a priori* and do their best to come close to the optimum over an entire class of models. Amongst the first group, we find Huffman codes [3, pp. 402–406] and Shannon–Fano[1] codes [1, pp. 101–103]. Amongst the second group, the best known algorithms are the ones due to Lempel and Ziv[2] in 1977 and 1978.

2. Redundancy of Classical Codes

The codes normally considered are at least near-optimal with respect to the entropy lower bound. Define first the pointwise redundancy of a code C with respect to a model P as

$$R_n(C,P;x_1^n) := L\big(C(x_1^n)\big) + \lg P(x_1^n),$$

where L is length. Two critical parameters are then the *average redundancy* and the *maximal redundancy* defined by

$$\bar{R}_n(C,P) = E\big[R_n(C,P;x_1^n)\big], \qquad R_n^*(C,P) = \max_{x_1^n}\big[R_n(C,P;x_1^n)\big]. \tag{1}$$

where both average and maximum are meant with respect to x_1^n. In other words, the question asked is: *How far are we from the information theoretic optimum, either on average or in the worst case?* There, we assume the source distribution to be known and the code to be fixed, and analyse the redundancy parameters of the given code.

In this perspective, the talk first reviews results relative to the classical Huffman code and to a version of Fano–Shannon codes, this in the case of a memoryless source. Redundancy is then $O(1)$ but with fluctuations that depend on the fine arithmetic structure of the parameters of the model under consideration; see Figure 1. The methods use Fourier analysis and *Gleichverteilung* mod 1.

Louchard and Szpankowski (1997), Savari (1997), Wyner (1998), and Jacquet–Szpankowski (1995) proved that the Lempel–Ziv algorithms under either a memoryless or a Markov model have rates that are $\Theta(n/\log n)$ for LZ'78[3] and $\Theta(n \log\log n/\log n)$ for LZ'77. The proofs provide detailed asymptotic information on the redundancy. The results again involve subtle fluctuations. The analysis is close to that of digital tries, with Mellin transforms playing a prominent rôle.

[1] To design a Shannon–Fano code for the distribution P on S, partition S as $S = S_0 \cup S_1$ in such a way that the probabilities of S_0 and S_1 differ by as little as possible from 1/2. All elements of S_j are assigned a code that starts with j. Proceed recursively.

[2] Roughly, the LZ algorithms recognize, as characters flow, the frequently repeated blocks of letter and avoid copying these over and over again, but instead output pointers to the location of the first occurrence of such a block.

[3] LZ'78 parses a sequence into "phrases" and outputs a pointer to the longest phrase already encountered; LZ'77 outputs a pointer to the longest factor already encountered.

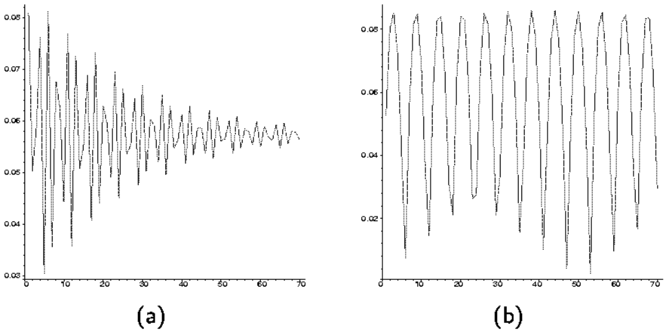

FIGURE 1. Huffman code redundancy for a memoryless source with control parameter $\alpha = \lg(1/p - 1)$: (a) irrational case ($p = 1/\pi$); (b) rational case ($p = 1/9$).

3. Minimax Redundancy for Classes of Source Models

The *strong redundancy-rate problem* asks what can be achieved when the source model ranges over a whole class of sources $\mathcal{S}$. Thus, the source model is a bit constrained but basically unknown and the question becomes information-theoretic rather than algorithmic (no coding algorithm is fixed any more). Consider redundancies in the sense of (1) and define the *minimax redundancies*,

$$\bar{R}_n(\mathcal{S}) = \min_C \max_{P\in\mathcal{S}} \bar{R}_n(C,P), \qquad R^*_n(\mathcal{S}) = \min_C \max_{P\in\mathcal{S}} \bar{R}^*_n(C,P), \tag{2}$$

corresponding to an average-case or a worst-case scenario, respectively. By their definitions, these quantities represent the additional cost on top of entropy incurred (at least) by any code (this is $\min_C$) in order to be able to cope with *all* sources (this is $\max_{P\in\mathcal{S}}$).

It would seem that the minimax problem of estimating the quantities in (2) is intractable. However, Shtarkov proved in 1978 that the (worst-case) minimax redundancy is narrowly bounded by the (Shtarkov) inequalities

$$\lg \sum_{x_1^n\in\mathcal{A}^n} \sup_{P\in S} P(x_1^n) \le R^*_n(\mathcal{S}) \le 1 + \lg \sum_{x_1^n\in\mathcal{A}^n} \sup_{P\in S} P(x_1^n). \tag{3}$$

There the quantity $\sup P(x_1^n)$ could be termed a "maximum likelihood coefficient" since it describes the probability of any individual realization x_1^n under the model $P \in \mathcal{S}$ that assigns to it the highest probability. Take for instance a binary word $x_1^n \in \{0,1\}^n$ comprising k letters 0 and $n-k$ letters 1, and $\mathcal{S}$ the class of all memoryless models with $\mathbf{P}(0) = p$ and $\mathbf{P}(1) = 1-p$. Clearly, the maximum likelihood coefficient is given by the Bernoulli distribution whose parameter is $p = k/n$ (maximum likelihood probabilities equal frequencies), and its value is $(k/n)^k\big((n-k)/n\big)^{n-k}$. The sum appearing in (3) then evaluates to

$$A_n := \sum_{x_1^n\in\mathcal{A}^n} \sup_{P\in S} P(x_1^n) = \sum_{k=0}^{n} \binom{n}{k} \left(\frac{k}{n}\right)^k \left(\frac{n-k}{n}\right)^{n-k}.$$

This has the same flavour as Abel's identities. Indeed, we have

$$A_n = \frac{n!}{n^n}[z^n]\frac{1}{\big(1-T(z)\big)^2} \quad \text{where} \quad T(z) = ze^{T(z)}$$

is the tree function. It is then a simple matter, by singularity analysis of the tree function, to get

$$A_n \sim \frac{1}{2}\frac{n!e^n}{n^n} \sim \sqrt{\frac{\pi n}{2}} \quad \text{and} \quad \lg A_n = \frac{1}{2}\lg n + \lg\sqrt{\frac{\pi}{2}} + o(1).$$

The quantity $\lg A_n$ is at most 1 from the minimax redundancy as results from inequalities (3).

Renewal sources. Another topic of the talk is to analyse redundancy for the class of renewal sources defined as follows.

M3. A renewal model starts with a random sequence $(x_i)_{-\infty}^{+\infty}$ of '0's and '1's, infinite in both directions and such that the spacings between the '1's are independent identically distributed random variables. Then extract the window corresponding to $x_1^n = x_1 \ldots x_n$. (You're sitting under a bus shelter and record every minute whether you're seeing a bus passing or not.)

This class of sources makes for an interesting study since minimax redundancy turns out to be $O(\sqrt{n})$; see [2] for a complete analysis.

The maximal likelihood approach leads to the consideration of the sum

$$r_n = \sum_k \sum_{\mathcal{P}(n,k)} \binom{k}{k_0, \ldots, k_{n-1}} \left(\frac{k_0}{k}\right)^{k_0} \left(\frac{k_1}{k}\right)^{k_1} \cdots \left(\frac{k_{n-1}}{k}\right)^{k_{n-1}}.$$

There, the summation condition $\mathcal{P}(n,k)$ is $n = k_0 + 2k_1 + \cdots$, $k = k_0 + k_1 + \cdots$. The computation heavily involves the tree function $T(z)$ and proceeds in several steps.

First, one disposes of the normalizing factor of $k!/k^k \sim e^{-k}\sqrt{2\pi k}$ by introducing as an artefact a random variable K_n and relating r_n to $E\left[\sqrt{2\pi K_n}\,\right]$. Second, the distribution of K_n is described by the bivariate generating function

$$S(z,u) := \prod_{i=1}^{\infty} \beta(z^i u) \quad \text{where} \quad \beta(z) = \frac{1}{1 - T(ze^{-1})}.$$

This has roughly the character of (the square root of) a partition generating function with u marking the number of parts. Third, the saddle-point method is applied to extract coefficients. Fourth, the saddle-point analysis conduces to a local analysis near 1 that is solved by Mellin transform techniques. The eventual result is that

$$\lg r_n = \frac{2}{\log 2}\sqrt{\left(\frac{\pi^2}{6} - 1\right) n} - \frac{5}{8}\lg n + \frac{1}{2}\lg \log n + O(1),$$

and this quantity closely approximates the minimax redundancy of renewal sources by Shtarkov's inequalities. Note the asymptotic form $r_n \approx e^{\sqrt{n}}$ that is typical of partition estimates.

Conclusion. The redundancy problem is typical of situations where second-order asymptotics are essential. Such problems of information theory are thus candidates *par excellence* for the methods of *analytic information theory*. By this, it is meant the study of randomness in words and codes by means of the classical methods of analytic combinatorics. The reader interested in these questions will be well-advised to consult the forthcoming book by Szpankowski [4] and references therein.

Bibliography

[1] Cover (Thomas M.) and Thomas (Joy A.). – *Elements of information theory.* – John Wiley & Sons Inc., New York, 1991, xxiv+542p. A Wiley-Interscience Publication.

[2] Flajolet (Philippe) and Szpankowski (Wojtek). – *Analytic Variations on Redundancy Rates of Renewal Processes.* – Research Report n° 3553, Institut National de Recherche en Informatique et en Automatique, 1998. 10 pages. Submitted to IEEE Transactions on Information Theory.

[3] Knuth (Donald E.). – *The art of computer programming.* – Addison-Wesley Publishing Co., Reading, Mass.-London-Amsterdam, 1997, third edition, xx+650 pp.p. Volume 1: Fundamental algorithms.

[4] Szpankowski (Wojciech). – *Average-case analysis of algorithms on sequences.* – John Wiley & Sons Inc. To appear, preliminary version available from `http://www.cs.purdue.edu/people/spa`.