Algorithms Seminar 2000–2001,
F. Chyzak (ed.), INRIA, (2002), pp. 135–138.

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

Information Theory by Analytic Methods: The Precise Minimax Redundancy

Wojciech Szpankowski
Department of Computer Science, Purdue University (USA)

March 5, 2001

Summary by Thomas Klausner

1. Introduction

The redundancy-rate problem of universal coding is concerned with determining by how much the actual code length (representation of a word in a code) exceeds the optimal code length. Revisiting the theme of his last year's seminar talk [1], Szpankowski went into more detail explaining different models for redundancy, and introduced the *generalized Shannon code* in order to solve the minimax redundancy problem for a single memoryless source.

A code is defined as follows:

Definition 1. A code C_n is a mapping from the set $\mathcal{A}^n$ of all sequences of length n over the alphabet $\mathcal{A}$ to the set $\{0,1\}^*$ of binary sequences.

Most of the time we use source models which specify probabilities for specific messages. For these, $\mathcal{P}(x_1^n)$ is the probability of the message x_1^n, the code length of a message $x_1^n = x_1 \dots x_n$, with $x_i \in \mathcal{A}$, in the code C_n will be denoted by $L(C_n, x_1^n)$, and $H_n(\mathcal{P}) = -\sum_{x_1^n} \mathcal{P}(x_1^n) \log \mathcal{P}(x_1^n)$ is the entropy of the probability distribution, where log is taken to base 2.

2. Basic Results

A *prefix code* or *instantaneous code* is a code in which no codeword is a prefix for another codeword; in other words, if you present the codewords as a binary trie, the valid codewords are only in the leaves (not in the internal nodes).

For prefix codes the following inequality holds:

Lemma 1 (Kraft's inequality). *For any prefix code (over a binary alphabet), the codeword lengths* $l_1, l_2, \dots, l_m$ *satisfy the inequality*

$$\sum_{i=1}^{m} 2^{-l_i} \le 1.$$

A related problem is to find out how many tuples $l_1, \dots, l_m$ exist where equality holds. This has been tackled and solved by Flajolet and Prodinger [2]. Asymptotically, it grows as $\alpha\phi^m$, where $\alpha \approx 0.254$ and $\phi \approx 1.794$.

Another important result is Shannon's classic lower bound on the average code length (see [3]):

Lemma 2 (Shannon). *For any code, the average code length* $\mathbf{E}\big[L(C_n, X_1^n)\big]$ *cannot be smaller than the* entropy *of the source* $H_n(\mathcal{P})$*:*

$$\mathbf{E}\big[L(C_n, X_1^n)\big] \ge H_n(\mathcal{P})$$

Trivially, one can see that there must exist at least one $\tilde{x}_1^n$ with

$$L(\tilde{x}_1^n) \geq -\log \mathcal{P}(\tilde{x}_1^n).$$

A lemma by Barron deals with the individual lengths of the code words:

Lemma 3 (Barron)**.** *Let $L(X_1^n)$ be the length of a codeword in a code satisfying Kraft's inequality, where X_1^n is generated by a stationary ergodic source. For any sequence of positive constants a_n satisfying $\sum 2^{-a_n} < \infty$, the following holds:*

$$\mathbf{P}\{L(X_1^n) \leq -\log \mathcal{P}(X_1^n) - a_n\} \leq 2^{-a_n}.$$

From this we immediately get

$$L(X_1^n) \geq -\log \mathcal{P}(X_1^n) - a_n \quad \textit{(almost surely)}.$$

3. Redundancy

Redundancy measures the distance to the optimal code state, reaching the lower bound given by the entropy. Since there are different ways to define the "worst case," we define three types of redundancy: pointwise $R_n(C_n, \mathcal{P}; x_1^n)$, average $\bar{R}_n(C_n, \mathcal{P})$ and maximal $R^*(C_n, \mathcal{P})$:

$$\begin{aligned} R_n(C_n, \mathcal{P}; x_1^n) &= L(C_n, x_1^n) + \log \mathcal{P}(x_1^n) \qquad (\geq -a_n(a.s.)), \\ \bar{R}_n(C_n, \mathcal{P}) &= \mathbf{E}_{X_1^n}\big[R_n(C_n, \mathcal{P}; X_1^n)\big] \\ &= \mathbf{E}\big[L(C_n, X_1^n)\big] - H_n(\mathcal{P}), \\ R^*(C_n, \mathcal{P}) &= \max_{x_1^n}\big[R_n(C_n, \mathcal{P}; x_1^n)\big]. \end{aligned}$$

The redundancy-rate problem consists in finding the rate of growth of the corresponding minimax quantities

$$\begin{aligned} \bar{R}_n(\mathcal{S}) &= \min_{C_n} \sup_{\mathcal{P}\in\mathcal{S}} \mathbf{E}\big[R_n(C_n, \mathcal{P}; x_1^n)\big], \\ R_n^*(\mathcal{S}) &= \min_{C_n} \sup_{\mathcal{P}\in\mathcal{S}} \max_{x_1^n}\big[R_n(C_n, \mathcal{P}; x_1^n)\big], \end{aligned}$$

as $n \to \infty$ for a class $\mathcal{S}$ of source models.

There are also other measures of optimality, e.g. for coding, gambling, or predictions. For these, the following functions, called minimax regret functions, are used:

$$\begin{aligned} \bar{r}_n &= \min_{C_n} \sup_{\mathcal{P}\in\mathcal{S}} \sum_{x_1^n} \mathcal{P}(x_1^n)\big[L_i + \log \sup_{\mathcal{P}} \mathcal{P}(x_1^n)\big], \\ r_n^* &= \min_{C_n} \max_{x_1^n}\big[L_i + \log \sup_{\mathcal{P}} \mathcal{P}(x_1^n)\big]. \end{aligned}$$

Note that $r_n^* = R_n^*$. Sometimes, the maximin regret is of interest:

$$\tilde{r}_n = \sup_{\mathcal{P}\in\mathcal{S}} \min_{C_n} \sum_{x_1^n} \mathcal{P}(x_1^n)\big[L_i + \log \sup_{\mathcal{P}} \mathcal{P}(x_1^n)\big].$$

These functions are sometimes called the average minimax regret ($\bar{r}_n$), the maximal minimax regret (r_n^*), and the average maxmin regret ($\tilde{r}_n$). One can interpret these functions as target functions for the game theoretical problem of choosing L so that for all x_1^n, the value of the function gets as good as possible, that is, $-\log \sup \mathcal{P}(x_1^n)$.

In the following, we will only look at the redundancy functions.

4. Precise Maximal Redundancy

In 1978, Shtarkov proved the following bounds for the minimax redundancy:

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

We want to find a precise result for $R_n^*(\mathcal{S})$. We start with the easier problem of finding the optimal code for maximal redundancy for a known source $\mathcal{P}$

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

We already know that for the average redundancy of one known source

$$\bar{R}_n(\mathcal{P}) = \min_{C_n\in C} \mathbf{E}_{x_1^n}\left[R_n(C_n, \mathcal{P}; x_1^n)\right],$$

the Huffmann code is optimal—indeed, it is designed so as to solve this optimization problem. For the maximal redundancy problem we introduce a new code, the *generalized Shannon code.*

In the ordinary *Shannon code*, the length of its symbol in the code for a given $\mathcal{P}$ is $\lceil 1/\mathcal{P}(x_1^n)\rceil$. In the generalized Shannon code, on the other hand, we set the length to be $\lfloor 1/\mathcal{P}(x_1^n)\rfloor$ for some symbols $x_1^n \in \mathcal{L}$ and $\lceil 1/\mathcal{P}(x_1^n)\rceil$ for the others in such a way that Kraft's inequality holds. For non-dyadic codes (dyadic ones fulfill $R_n^*(\mathcal{P}) = 0$), we sort the probabilities $\mathcal{P}(x_1^n)$:

$$0 \le \langle -\log p_1\rangle \le \langle -\log p_2\rangle \le \cdots \le \left\langle -\log p_{|\mathcal{A}|^n}\right\rangle \le 1 \quad (\text{where } \langle x\rangle = x - \lfloor x\rfloor)$$

and choose j_0 to be the maximal j such that Kraft's inequality still holds:

$$\sum_{i=0}^{j-1} p_i 2^{\langle -\log p_i\rangle} + \sum_{i=j}^{|\mathcal{A}|^n} p_i 2^{\langle -\log p_i\rangle - 1} \le 1.$$

Then $R_n^*(\mathcal{P}) = 1 - \langle -\log p_{j_0}\rangle$ and the generalized Shannon code with $\mathcal{L} = \{1, \dots, j_0\}$ is optimal.

Now we generalize to systems of probability distributions $\mathcal{S}$. Let

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

Then

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

with

$$R_n^*(Q^*) = 1 - \langle -\log q_{j_0}\rangle$$

as above.

If we now take the generalized Shannon code that minimizes the maximal redundancy, we get for a sequence generated by a single memoryless source, for $n \to \infty$, and $\alpha = \log\frac{1-p}{p}$ irrational:

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

5. Average Minimax Redundancy

In the simple case where $\mathcal{S}$ consists of one distribution $\mathcal{P}$, the computation of $\bar{R}_n^H$ is the Huffman problem:

$$\bar{R}_n^H(\mathcal{P}) = \min_{C_n \in C} \sum_{x_1^n} \mathcal{P}(x_1^n) R_n(C_n, \mathcal{P}; x_1^n).$$

From known results (where we have $\bar{R}_n^H \approx R_n^*$), we conjecture:

Conjecture 1. *Under certain additional conditions, we have, as $n \to \infty$,*

$$\bar{R}_n = R_n^* + \Theta(1) = \log\Bigl(\sum_{x_1^n \in \mathcal{A}^n} \sup_{\mathcal{P} \in \mathcal{S}} \mathcal{P}(x_1^n)\Bigr) + \Theta(1).$$

6. Average Redundancy for Particular Codes

For single memoryless sources, we have explicit results for $n \to \infty$ for some codes. In particular, we have for the Huffman code

$$\bar{R}_n = \begin{cases} \frac{3}{2} - \frac{1}{\ln 2} & \text{if } \alpha \text{ irrational}, \\ \frac{3}{2} - \frac{1}{M}\left(\langle Mn\beta\rangle - \frac{1}{2}\right) - \left(M(1-2^{-1/M})\right)^{-1} 2^{-\langle Mn\beta\rangle/M} & \text{if } \alpha = \frac{N}{M}, \end{cases}$$

for the Shannon code

$$\bar{R}_n = \begin{cases} \frac{1}{2} & \text{if } \alpha \text{ irrational}, \\ \frac{1}{2} - \frac{1}{M}\left(\langle Mn\beta\rangle - \frac{1}{2}\right) & \text{if } \alpha = \frac{N}{M}, \end{cases}$$

and for the generalized Shannon code

$$\bar{R}_n = \frac{3}{2} - 2\ln 2 + o(1) \approx 0.113705639.$$

For more basics and in-depth knowledge regarding analytic information theory, the interested reader is referred to Szpankowski's book [4].

Bibliography

[1] Flajolet (Philippe). – Analytic information theory and the redundancy rate problem [summary of a talk by Wojciech Szpankowski]. In Chyzak (Frédéric) (editor), *Algorithms Seminar, 1999–2000*, pp. 133–136. – Institut National de Recherche en Informatique et en Automatique, November 2000. Research Report n° 4056.

[2] Flajolet (Philippe) and Prodinger (Helmut). – Level number sequences for trees. *Discrete Mathematics*, vol. 65, n° 2, 1987, pp. 149–156.

[3] Shannon (C. E.). – A mathematical theory of communication. *Bell System Technical Journal*, vol. 27, 1948, pp. 379–423 and 623–656.

[4] Szpankowski (Wojciech). – *Average-case analysis of algorithms on sequences.* – John Wiley & Sons, Chichester, New York, March 2001, *Wiley-Interscience Series in Discrete Mathematics.*