Coding of a source with unknown but ordered probabilities

Boris Ryabko

1979

Abstract

The article deals with the problem of optimum coding of a source for whose symbols it is known only that they are arranged in decreasing order of probability. On the basis of the resultant code, a design for a universal retrieval system is proposed and a hypothesis that accounts for Zipf’s law is advanced.

CODING OF A SOURCE WITH UNKNOWN BUT
ORDERED PROBABILITIES

B. Ya. Ryabko

The article deals with the problem of optimum coding of a source for whose symbols it is known
only that they are arranged in decreasing order of probability. On the basis of the resultant
code, a design for a universal retrieval system is proposed and a hypothesis that accounts for
Zipf's law is advanced.*

§ 1. INTRODUCTION

Let us consider the problem of coding of a source that generates a finite number of letters, regarding
which it is known only that they are arranged in order of decreasing probability. By reducing this problem to
one of computing the capacity of a discrete memoryless channel, we obtain a code F whose redundancy differs
from the minimum possible value by not more than one.

The problem of coding of a source with symbols ordered with respect to probability was considered in
[1], in which a Levenshtein code was employed [2]. This code, however, is constructed for a source with a
countable number of symbols [2, 3], and therefore it is inferior to the code proposed in this study for a source
with a finite alphabet. The difference between the redundancy of a Levenshtein code and code F increases with-
out limit as the number of symbols generated by the source increases.

Code F is used to construct a universal information retrieval system in computers.

In the course of investigating natural language, Zipf discovered [4] that if the words of a language are
ordered with respect to decreasing frequency of occurrence in text, then the frequency of the i-th word is
roughly proportional to 1/i^\gamma, where \gamma \approx 1. A number of models have been proposed to account for this be-
havior [5–8]. For instance, Mandelbrot showed that if the space between words is treated as a random sym-
bol, then Zipf's law will be satisfied [5]. He obtained this distribution on the basis of the assumption that the
evolutionary process of choice of word lengths can be described as a variety of random walk [6]. In this paper
we propose a model based on the assumption of universality of natural languages, that accounts for Zipf's law
by means of code F.

§ 2. STATEMENT OF THE PROBLEM AND NOTATION

For integer n > 1 let us consider class P_n of all sources that generate letters x_1, \ldots, x_n with prob-
abilities \( p_1, \ldots, p_n \) = p, respectively, where \( \sum_{i=1}^{n-1} p_i = 1 \) and \( p_i \geq p_{i+1} \) for \( i = 1, \ldots, n - 1 \). In what follows we
will identify the source with the probability vector of the letters generated by it. Assume that L_n is the set of
all decodable codes containing n words over an alphabet of r letters (r \geq 2). The redundancy of code L \in L_n
on source \( p = (p_1, \ldots, p_n) \) is the quantity \( \rho(L, p) = \sum_{i=1}^{n} p_i (l_i + \log_2 r_i) \), where \( l_i \) is the length of the word from L that
encodes letter \( x_i \). For the class of sources P and L \in L_n, assume that

*The main results of this paper were presented at the Seventh All-Union Symposium on Problems of Redund-
ancy in Information Systems, and were published in the Proceedings of the Symposium [Proceedings of Seventh
All-Union Symposium on Problems of Redundancy in Information Systems, Vol. 1: Abstracts of Reports, Lenin-
grad (1977), pp. 162-164].

article submitted July 11, 1977.
The quantity $R(P)$ is the lower bound of the redundancy of any code on class of sources $P$. The problem is to find a code $L \in L_n$ for which $R(L, P)$ is close to $R(P)$.

We denote by $S_n$ an $n$-dimensional simplex: $S_n = \{q = (q_1, \ldots, q_n): \sum_{i=1}^{n} q_i = 1$ and $q_i \geq 0$ for $i = 1, \ldots, n\}$; for finite set $M = \{m_1, \ldots, m_k\}$ we denote by $S(M)$ its convex hull: $S(M) = \{e = \sum_{i=1}^{k} \alpha_i m_i, (\alpha_1, \ldots, \alpha_k) \in S_k\}$. For arbitrary $\lambda, p \in S_n$, we define $\rho'((\lambda, p)) = \sum_{i=1}^{n} p_i \log \frac{p_i}{\lambda_i}$ [assuming that $0 \log(0/x) = 0$ for any $x$ and $y \log(y/0) = \infty$ for $y = 0$]. Note that for all $\lambda, p \in S_n$

$$\rho'((\lambda, p)) \geq 0, \rho'(p, p) = 0.$$ (1)

It is also known that $\rho'((\lambda, p))$ is convex in $p$, i.e., for arbitrary $p_1, \ldots, p_k \in S_n$ and $\alpha \in S_k$ the Jensen inequality is satisfied:

$$\rho'\left((\lambda, \sum_{i=1}^{k} \alpha_i p_i)\right) \leq \sum_{i=1}^{k} \alpha_i \rho'(p_i, p).$$ (2)

For each $P \subset S_n$ we determine $R'(P) = \sup\{\rho'((\lambda, p)): \rho((L, p)) = \inf_{k \in \mathbb{R}_+} R'(L, \lambda, p)\}$.

It is not difficult to show that $R'(P) = \inf_{L \in L_n} R'(L, \lambda, p)$, where $T_n = \{q = (q_1, \ldots, q_n): \sum_{i=1}^{n} q_i = 1$ and $q_i \geq 0$ for $i = 1, \ldots, n\}$. In view of the Kraft inequality, any $L \in L_n$ corresponds to a $\lambda \in T_n$ such that $\rho(L, p) = \rho'(\lambda, p)$ for any $p \in S_n$. On the other hand, for every $\lambda \in S_n$ we can construct a code $L \in L_n$ such that $i_j = [\log_{\lambda}(1/\lambda_i)]$ for $i = 1, \ldots, n$. This implies that

$$0 \leq R(P) - R'(P) \leq 1,$$ (3)

for any $P \subset S_n$, and therefore in what follows we will investigate the quantity $R'(P)$.

§ 3. RELATIONSHIP BETWEEN REDUNDANCY ON FINITE CLASS OF SOURCES AND CHANNEL CAPACITY

Consider a discrete memoryless channel (DMC) whose input alphabet contains $t$ letters and whose output alphabet contains $n$ letters ($t, n > 1$). The channel is specified by $t$ vectors of dimension $n$, $\{y_1, \ldots, y_t\} = Y$, where the $j$-th coordinate of the $i$-th vector $(y_{ij})$ is equal to the probability of reception of the $j$-th letter of the output alphabet under the condition that the $i$-th letter of the input alphabet has been transmitted.

If it is known that the $i$-th letter of the input alphabet is used with probability $\alpha_i (\alpha \in S_t)$, then the mean mutual information between the input and output, by definition, is $I(\alpha, Y) = \sum_{i=1}^{t} \alpha_i I(\alpha, Y) = \sum_{i=1}^{t} \alpha_i Y_{ij}$ for $j = 1, \ldots, n$. The capacity of the DMC $\{C(Y)\}$ is given by the expression $C(Y) = \sup_{\alpha \in S_t} I(\alpha, Y)$. It is known that for vector $\beta \in S_t$ the equality $I(\beta, Y) = C(Y)$ is satisfied if and only if there exists some constant $\Psi$ such that

$$\rho'(v(\beta, Y), y_i) = \Psi \quad \text{if and only if} \quad i = 1, \ldots, t,$$

where $\Psi = C(Y)$ [9]. Note that there is a known method for computing the capacity $C(Y)$ which in some cases enables us to compute it by solving two systems of linear equations [9].

LEMMA. If set $P \subset S_n$ is finite, then we have $R'(P) = C(P)$; if in this case $C(P) = I(\gamma, P)$ for some $\gamma$, then $R'(P) = R'(\nu(\gamma, P))$.

\*\*\* is the nearest integer not less than $x$. 

135
Proof. Let $R(P) = \inf_{\lambda \in S(P)} R'(\lambda, P)$, where $P \subseteq S$. It is shown in [9] (p. 535) that for finite $P$

$$R(P) = C(P).$$

(5)

To prove the first assertion of the lemma it is sufficient to show that no vector $\mu \in S \setminus S(P)$ exists such that

$$R'(\mu, P) < R(P).$$

(6)

We will employ indirect proof. Assume that there exists a $\mu \in S_n$ for which (6) is satisfied. It is understood that $C(P \cup \mu) > C(P)$, since the capacity of a DMLC does not decrease when the input alphabet is increased. From this and from (5) we obtain that

$$R(P \cup \mu) \geq R(P).$$

(7)

It follows from (1) that

$$R'(\mu, P \cup \mu) = R'(\mu, P).$$

(8)

Since $\mu \in S(P \cup \mu)$, we have $R(P \cup \mu) \leq R'(\mu, P \cup \mu)$. We obtain from (8) that $R(P \cup \mu) = R'(\mu, P)$. This and expression (7) yield the inequality $R'(\mu, P) \geq R(P)$, a contradiction with (6).

The second assertion of the lemma follows directly from condition (4).

§ 4. CODE FOR ORDERED SOURCE

Consider code $F \in L_n$ for which the length of the i-th word is

$$t_i = \log \left( (i-1)^{i-1}/i^i \right) \quad \text{for} \quad i = 1, \ldots, n;$$

$$t = \sum_{i=1}^{n} (i-1)^{i-1}/i^i. \tag{9}$$

**THEOREM.** The redundancy of code $F$ on class of sources $P_n$ does not exceed $\log t + 1$ and differs from the minimum possible redundancy by not more than 1:

$$\log t \leq R(P_n) \leq R(F, P_n) \leq \log t + 1. \tag{10}$$

**Proof.** For $i = 1, \ldots, n$ we denote by $q_i$ vectors whose first $i$ coordinates are equal to $1/i$, while the remaining ones are 0. Let $Q = \{ q_1, \ldots, q_n \}$. First we note that

$$P_n = S(Q), \tag{11}$$

since any $p \in P_n$ can be represented in the form $p = \sum_{i=1}^{n} (\tau_i) q_i$, where $\tau_i = p_i - p_{i+1}$ for $i = 1, \ldots, n$ ($p_{n+1} = 0$).

We obtain from (11) and (2) that for any $\lambda \in S_n$ we have $R'(\lambda, P_n) = R'(\lambda, Q)$. Consequently,

$$R'(P_n) = R'(Q). \tag{12}$$

Now let us compute, in accordance with [9], the quantity $C(Q)$, this being the capacity of a DMLC formed by vectors from $Q$. For this we first find $\phi$ and $(\varphi_1, \ldots, \varphi_n) = \varphi$, for which we have the expressions

$$\sum_{i=1}^{n} \varphi_i = 1, \quad \rho'(\varphi, q_i) = \psi \tag{13}$$

for $i = 1, \ldots, n$. Then we compute a vector $\xi = (\xi_1, \ldots, \xi_n)$ such that

$$I(\xi, Q) = C(Q). \tag{14}$$

For this we solve the system of equations

$$\sum_{j=1}^{n} z_{ij} \xi_j = \psi_i \quad \text{for} \quad i = 1, \ldots, n, \tag{15}$$

where $z_{ij}$ is the $i$-th coordinate of vector $q_j$. Using the fact that the matrices of the systems of equations formed by the last $n$ expressions in (13) and the expressions in (15) are triangular, we readily find that

*By definition $0^0 = 1.$
### TABLE 1

<table>
<thead>
<tr>
<th>Radius of code-word</th>
<th>Length of i-th code word</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1.062</td>
</tr>
<tr>
<td>10</td>
<td>1.386</td>
</tr>
<tr>
<td>50</td>
<td>2.082</td>
</tr>
<tr>
<td>100</td>
<td>2.082</td>
</tr>
<tr>
<td>500</td>
<td>2.348</td>
</tr>
<tr>
<td>10^4</td>
<td>2.350</td>
</tr>
<tr>
<td>10^8</td>
<td>3.962</td>
</tr>
</tbody>
</table>

Levenshtein code

\[
\psi = \log \frac{i}{1 + \frac{1}{(i-1)^{-1}}}
\]

Since both cofactors in the expression \[
\psi = \log \frac{i}{1 + \frac{1}{(i-1)^{-1}}}
\]
decrease monotonically as i increases from 1 to \(\infty\), we obtain \(i_0 > 0\) for \(i = 1, \ldots, n\). From this and from (13) and (15), as well as condition (4), we obtain (14). Using the lemma, we find that \(R'(\varphi, Q) = R'(Q)\). The assertion of the theorem follows from (12), (16), and (3).

**COROLLARY.** For large n, the quantity t from (9) is equal to \(\ln(n/e) + O(1)\). It follows from this and from (10) that \(R(F, P_n) = \log \log n + O(1)\).

It is not difficult to show that the maximum redundancy of the code proposed in [1] (based on Levenshtein code [2]) is greater than \(\log \log n + c \log \log \log n\), where \(c = 0\) is a constant. Table 1 gives redundancies and lengths of some code words of F and of Levenshtein code for some n.

The resultant code can be used for universal block coding (in the sense of [10]). Assume that the block length is k, and that the source generates two symbols whose probabilities are independent and equal to p and 1 − p, respectively. We are to encode \(2^k\) blocks by words over an alphabet of r letters. Let us assume that \(p > 1/2\) and we order the blocks on the basis of nonincreasing probability. Then we place block number j in correspondence with a word of length (9), where

\[
i = \begin{cases} 
2^j & \text{for } j \leq 2^{k-1}, \\
2^k (2^k - j) & \text{for } 2^{k-1} < j \leq 2^k.
\end{cases}
\]

It is not difficult to show that for any \(p \in [0, 1]\) the redundancy of this coding does not exceed \((1/k) \log_r k\) in order of magnitude.

This method can be generalized to the case of a Bernoulli source in which the number of generated letters is greater than 2, and also to the case of a Markov source of finite connectedness.

§ 5. UNIVERSAL RETRIEVAL SYSTEM AND ZIPF'S LAW

Consider a set of words \(L = \{l_1, \ldots, l_n\}\) over an r-letter alphabet the set possessing the prefix property. A retrieval system is a program that effects a one-to-one correspondence between L and the set of words \(X = \{x_1, \ldots, x_n\}\). In an alphabetic retrieval system L is a set of keys; in a system in which retrieval is conducted by comparison of elements of X on the basis of a linear ordering specified on X, the elements of L correspond to branches in the search tree [11]. In these systems the retrieval time for an element from X is proportional to the length of the corresponding word from L.

The problem of constructing an L that minimizes the mean search time for specified frequencies of address \(p_1, \ldots, p_n\) is analogous to the problem of letter-by-letter coding of a source. When it is known only that \((p_1, \ldots, p_n) \in P_n\), a retrieval system with set of code words (9) will be quasi-optimal.

This retrieval system is also of interest in that there exists an extensive class of retrieval problems in which the elements can be ordered approximately in terms of frequency of address, although the frequencies themselves are unknown. For example, if the elements of X are ordered on the basis of time of appearance (abstracts of papers, reports, and so forth), then it is frequently sensible to assume that the frequency of address to the object is greater, the more recently it was "produced" (since fresher material is more frequently called upon).
The hypothesis that accounts for Zipf's law is based on the assumption that in perception and "production" of text (speech) the time required to determine the meaning of a word is proportional to its length (the assumption that the "value" of a word is proportional to its length is quite natural and has been expressed by several authors, e.g., [5, 8]).

The stock of words and frequencies with which they are used vary considerably over the lifetime of each individual, and therefore a person's retrieval system must be fairly universal. We can also assume that initially people come to recognize the more widespread words, i.e., that the stock of words increases with time, but that the order of words with respect to decreasing frequency changes little. If this is the case, then the retrieval system described above will be quasi-optimal. As can be seen from (7), the length of the i-th word (in terms of number of letters) in the vocabulary of this system is approximately equal to \( c + \log_i (i - 1) \) for large i (c is independent of i). Indeed, in English, the distribution of words with respect to length is roughly like this [12]. It is easy to show that the redundancy of texts with such word lengths will be minimal if the frequency of the i-th word (with respect to increasing length) is proportional to \( 1/i^p \), which agrees with the frequency distribution of words in Zipf's law.

In concluding, the author wishes to thank R. E. Krichefskii for his attention and assistance in all phases of this paper.

Remark. While this paper was being readied for press, paper [13] was published, this paper being devoted to the problem of constructing a code that minimizes the maximum ratio of the mean code length and entropy subject to the condition that the source belongs to \( P_n \). But this minimum is equal to infinity, and therefore the author introduces an additional constraint: \( p_1 \geq 1/m \) for \( i = 1, \ldots, n \), where \( m > 1 \) is a parameter. The resultant code turns out to be similar to \( P \); specifically, the lengths of the first \( m \) words are \( c \log m \), while the length of the \( i \)-th word for \( i > m \) is \( c(i \log i - (i - 1) \log (i - 1)) \).

**LITERATURE CITED**


---

*We assume that words which are unknown to a particular individual are used by him with zero frequency.
†Here L are words of actual language, while X are concepts or "meanings" of words.