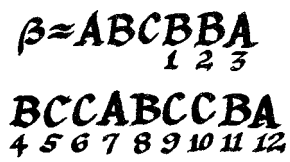
AFOSR 510

[redacted]

ZTB-140



A CODING METHOD FOR INDUCTIVE INFERENCE

R. J. Solomonoff

APRIL 1961

[redacted]

CONTRACT AF 49(638)-376

PREPARED FOR

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
AIR RESEARCH AND DEVELOPMENT COMMAND
UNITED STATES AIR FORCE

WASHINGTON 25, D. C.

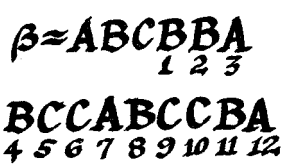
[redacted]

ZATOR COMPANY
140½ MOUNT AUBURN STREET, CAMBRIDGE 38, MASS.

[redacted]

AFOSR 510

ZTB-140



A CODING METHOD FOR INDUCTIVE INFERENCE

R. J. Solomonoff

APRIL 1961

CONTRACT AF 49(638)-376

PREPARED FOR

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
AIR RESEARCH AND DEVELOPMENT COMMAND
UNITED STATES AIR FORCE

WASHINGTON 25, D. C.

ZATOR COMPANY
140½ MOUNT AUBURN STREET, CAMBRIDGE 38, MASS.

A CODING METHOD FOR INDUCTIVE INFERENCE

R. J. Solomonoff

ABSTRACT

A new general inductive inference method has been described in which the a-priori probability of a sequence of symbols is computed on the basis of the lengths of various code strings that could be used to describe that sequence to a universal Turing machine. A coding method is displayed for a simple Bernoulli sequence and the inference technique is applied to the computation of probabilities of symbols in that sequence. The results obtained in this case are shown to be identical to Laplace's rule of succession. The probabilities correspond to Shannon's entropy if the Bernoulli sequence is a very long one.

TABLE OF CONTENTS

A CODING METHOD FOR INDUCTIVE INFERENCE

R. J. Solomonoff

I. INTRODUCTION

A very general inductive inference method has been described (References 1 and 2) in which an a-priori probability may be assigned to any long sequence of symbols.

The method consists of coding the sequence in a type of binary code, in all possible ways. The probability of any particular code is then 2^{-N}, N being the number of bits in its binary representation. The a-priori probability of the sequence is then the sum of the probabilities of all of its codes.

The conditional probabilities of various possible individual symbols following a given sequence of symbols may then be obtained by taking the normalized a-priori probabilities of sequences that consist of the given sequence, to which the various possible symbols have been catenated.

The present paper applies this general induction method to the problem of computing the probabilities of successive members of a Bernoulli sequence. This is about the simplest kind of inductive inference problem that exists, and has been the subject of much discussion.

The result, in the present case, is identical to one obtained by Laplace, which is called "Laplace's rule of succession." One set of assumptions that leads to his result is that the probabilities of the frequencies for each of the symbols in the Bernoulli sequence are initially uniformly distributed between zero and one. Then Laplace's rule gives the "expected value" of the frequency

of each type of symbol, after a certain initial subsequence of that entire Bernoulli sequence is known. The relative expected frequencies of the symbols A and B is

$$\frac{C_A + 1}{C_B + 1} ,$$

C_A and C_B being the number of occurrences of A and B, respectively, in the known subsequence.

The present analysis is used to illustrate a particularly important kind of coding method.

This coding method has been modified and generalized to apply to sequences of symbols in which Laplace's rule does not apply.

The present paper describes the simple coding method in some detail. It will be followed by papers dealing with modifications and generalizations of this basic coding technique.

II. CODING THE BERNOULLI SEQUENCE

A Bernoulli sequence is a Markov sequence in which the probability of each symbol is constant throughout the sequence, and is independent of the subsequence preceding that symbol.

For any Bernoulli sequence, we will give a method for assigning a set of numbers to that sequence. Each number will be a code for the sequence, so that given any integer, and an ordered list of the symbol types to be used in the sequence, the associated Bernoulli sequence can be uniquely determined.

Later, these code numbers will be used to compute a-priori probabilities of various sequences, and from these, in turn, an expression for conditional probabilities of successive symbols of the sequence will be obtained.

To assign a code number to a Bernoulli sequence, we will first assign an ordered sequence of ordered pairs on integers to the sequence. There will be a pair of integers for each symbol in the original sequence.

Consider the Bernoulli sequence:

$$\alpha \equiv \text{BBABCCABCCBA} \tag{1}$$

The only symbols used are A, B, and C.

We will then write the sequence of symbol types, A B C, followed by the original Bernoulli sequence. This gives:

$$\beta \equiv \underset{1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 10\ 11\ 12}{\text{ABCBBABCCABCCBA}} \tag{2}$$

The first symbol of α in Eq. (1) is B. The integer pair assigned to this B will be (3,2). The 3, because there are 3 symbols in β before the symbol to be coded. The 2, because the only previous occurrence of B is the second symbol.

The second symbol of α is also B. Its integer pair can be either (4,2) or (4,4). The reason is that in β, both the second and fourth symbols are B.

The integer pair for A, the third symbol of α, is (5,1).

The integer pair for B, the fourth symbol of α, is either (6,2), (6,4) or (6,5).

The integer pair for C, the fifth symbol of α, is (7,3).

One permissible intermediate code for the first five symbols of α is then

$$a \equiv (3,2),\ (4,2),\ (5,1),\ (6,5),\ (7,3). \tag{3}$$

Since there are two possible choices for the representation of the second symbol of α, and three possible choices for the fourth symbol of α, there are $2 \times 3 = 6$ permissible intermediate codes for the subsequence consisting of the first five symbols of α.

To change the intermediate code,

$$(a_1,b_1),(a_2,b_2),(a_3,b_3),\ldots,(a_n,b_n) \tag{4}$$

into an integer, we use the formula

$$k = ((\ldots(((b_n a_{n-1} + b_{n-1})\ a_{n-2} + b_{n-2})\ a_{n-3} + b_{n-3}).\ldots.)\ a_2 + b_2)\ a_1 + b_1. \tag{5}$$

k is then the integer code number for the sequence, α.

In the case of intermediate code, a, this gives

$$\begin{aligned} k &= (((3\cdot 6 + 5)\cdot 5 + 1)\cdot 4 + 2)\cdot 3 + 2 \\ &= 1400 \end{aligned} \tag{6}$$

It is possible to reverse this process and go from k back to the original intermediate code. To do this:

Divide k by a_1. The remainder is b_1.
Divide the quotient by a_2. The remainder is b_2.
Divide the resulting quotient by a_3. The remainder is b_3.
Continue until no more b_i's are obtainable.

In the present case, all a_i's are known.

$$a_i = i + 2. \tag{7}$$

More generally

$$a_i = i + d - 1 \tag{8}$$

where d is the number of symbol types in the Bernoulli sequence.

As an example, we will show how to go from the code number, 1400, in Eq. (6) back to Eq. (3).

Since from Eq. (8), $a_1 = 1 + d - 1 = 3$, we divide 1400 by 3 to obtain 466 with remainder 2, so $b_1 = 2$.

We divide 466 by 4 (which is a_2) to obtain 116 and remainder 2, which is b_2.

We divide 116 by 5 (which is a_3) to obtain 23 and remainder 1, which is b_3.

We divide 23 by 6 (which is a_4) to obtain 3 and remainder 5, which is b_4.

Since 3 is smaller than 7 (which is a_5) we stop at this point and set b_5 equal to 3.

Thus we have reproduced the original intermediate code of Eq. (3). To go from Eq. (3) back to the original Bernoulli sequence, α, is then trivial.

It is possible to obtain a very simple approximation for the value of k. Expanding Eq. (5), we obtain

$$\begin{aligned} k = {} & b_n a_{n-1} a_{n-2} \cdot \cdots \cdot a_2 a_1 \\ & + b_{n-1} a_{n-2} \cdot \cdots \cdot a_2 a_1 \\ & + \cdots \cdots \\ & + b_3 a_2 a_1 \\ & + b_2 a_1 \\ & + b_1 \end{aligned} \tag{9}$$

Since $a_{n-1} = n + d - 2$, and n will be very large in all cases of interest, we may neglect all but the first term, and write

$$k \approx b_n a_{n-1} a_{n-2} \ldots . a_2 a_1 . \tag{10}$$

III. CONDITIONAL PROBABILITIES IN THE BERNOULLI SEQUENCE

Now let us consider the problem of determining the relative probabilities of various possible symbols following the n symbol sequence α. We will use Eq. (5) of Reference 1 (page 18) to obtain this. (The notation is modified for the present context.)

$$\lim_{\epsilon \to 0} \lim_{m \to \infty} \frac{\sum_{j=1}^{r^{n+m}} \sum_{i=1}^{\infty} \left(\frac{1-\epsilon}{2} \right)^{N(S_{T A C_{n+m,j}})_i}}{\sum_{j=1}^{r^{n+m}} \sum_{i=1}^{\infty} \left(\frac{1-\epsilon}{2} \right)^{N(S_{T B C_{n+m,j}})_i}}$$

Essentially, this equation says that we should consider all possible continuations of the sequence to a distance of m symbols into the future.

Then the relative probability of the symbol A following the sequence α, rather than the symbol B following α, will be approximated by the total a-priori probabilities of all sequences of length $n + m$ that start out with the sequence α A, divided by the corresponding expression for sequences that start out with α B.

The approximation approaches exactness as m approaches infinity.

In the present case, it will become clear that the value of the approximation is constant for $m \geq 2$, so we will consider only $m = 2$, which is the simplest case.

For the a-priori probability of a particular sequence of $m + n$ symbols, we will use the approximation,

$$\sum_i \frac{1}{k_i} \tag{11}$$

k_i is the i^{th} code number that is a legal representation of that sequence.

It will be noted that this expression differs a bit from those used in Reference 1. There, the expression used was something like

$$\lim_{\epsilon \to 0} \sum_i \left(\tfrac{1}{2}(1-\epsilon)\right)^{N(k_i)} \tag{12}$$

Here, $N(k_i)$ is the number of bits in the binary expression for the integer k_i. We will approximate $N(k_i)$ by $\log_2 k_i$. This then gives for a-priori probability,

$$\lim_{\epsilon \to 0} \sum_i \left(\tfrac{1}{2}\right)^{\log_2 k_i} (1-\epsilon)^{\log_2 k_i} \tag{13}$$

If we set $1 - \epsilon \equiv 2^{-\delta}$, we obtain

$$\lim_{\delta \to 0} \sum_i \tfrac{1}{2}^{\log_2 k_i} \cdot 2^{-\delta \log_2 k_i} = \lim_{\delta \to 0} \sum_i \frac{1}{k_i^{(1+\delta)}} \tag{14}$$

In the present coding problem, it can be shown that the expressions for relative conditional probability that are obtained are independent of δ, so we will use

$$\sum_i \frac{1}{k_i}$$

for the total a-priori probability of the sequence being coded.

In order to obtain the necessary total a-priori probabilities, let us first consider the intermediate code expressions for $\alpha A \cdot$ followed by various single symbols. Such a code will have approximately the number

$$k = 3 \cdot 4 \cdot 5 \ldots\ldots (n+d-1)(n+d) \cdot b_{n+2} \tag{15}$$

assigned to it, and the a-priori probability of this code will be approximately

$$\frac{1}{k} = \frac{(d-1)!}{(n+d)!\, b_{n+2}} \tag{16}$$

The value of b_{n+2} can be any integer from 1 to $n+d$. It will be seen, that if we fix the value of b_{n+2}, and consider the $n+1^{th}$ sequence symbol to

be A, then there will be just $C_A!\, C_B!\, C_C!\, (C_A+1)$ different possible intermediate codes that start out with α. Here C_A, C_B, and C_C are the number of times that A, B, and C, respectively, occur in α. The reason for this particular number is that when the r^{th} occurrence of A, say, is being coded as a pair of integers, there will be only one possible choice for a_i, the first integer, but there will be just r possible choices for b_i, the second integer. Each such choice of b_i results in an acceptable intermediate code for the same sequence. The a_i and b_i are those of Eq. (4).

The total a-priori probability of all sequences of $n+2$ symbols that begin with α A, and have the same value of b_{n+2}, will be

$$\frac{(d-1)!\, C_A!\, C_B!\, C_C!\, (C_A+1)}{(n+d)!\, b_{n+2}} \tag{17}$$

Then, summing over all possible values of b_{n+2}, we obtain

$$\frac{(d-1)!\, C_A!\, C_B!\, C_C!}{(n+d)!} \left(\sum_{i=1}^{n+d} \frac{1}{i} \right) \left(C_A + 1 \right) \tag{18}$$

for the desired total a-priori probability.

The corresponding expression, in which the $n+1^{th}$ symbol must be B, is

$$\frac{(d-1)!\, C_A!\, C_B!\, C_C!}{(n+d)!} \left(\sum_{i=1}^{n+d} \frac{1}{i} \right) \left(C_B + 1 \right) \tag{19}$$

The relative probability of α being continued with A rather than B, is the ratio of these two expressions,

$$\frac{C_A+1}{C_B+1} \tag{20}$$

It will be noted that this expression is identical to that obtained by "Laplace's rule of succession."

It may seem unreasonable to go through this rather arduous process to obtain such a simple result – one that could be otherwise obtained from far simpler assumptions. However, the present demonstration is used to illustrate a very simple case of a certain coding method. It is well that this coding method should give reasonable results in this particular case. Later, we shall generalize the coding method that has been described, so that it may deal with descriptions that utilize definitions of subsequences that occur with significant frequencies.

The mean values of expressions (18) and (19) are, for long sequences, equivalent to Shannon's $\Sigma\, p_i \log_2 p_i$. If we take $\log_2$ of expression (18) and divide this by n we obtain the mean bit cost of coding a symbol of the original sequence. Using Sterling's approximation, it can be easily shown that this approaches

$$\frac{C_A}{n} \log_2\left(\frac{C_A}{n}\right) + \frac{C_B}{n} \log_2\left(\frac{C_B}{n}\right) + \frac{C_C}{n} \log_2\left(\frac{C_C}{n}\right) \equiv \sum p_i \log_2 p_i \qquad (21)$$

as n approaches infinity.

REFERENCES

1. R. J. Solomonoff, "A Preliminary Report on a General Theory of Inductive Inference," Zator Technical Bulletin No. 138, AFOSR TN-60-1459; Zator Company, November 1960. (Revision of Report V-131, February 1960.)

2. R. J. Solomonoff, "Progress Report: Research in Inductive Inference for the Period 1 April 1959 to November 1960," Zator Technical Bulletin No. 139, AFOSR 160; Zator Company, January 1961.