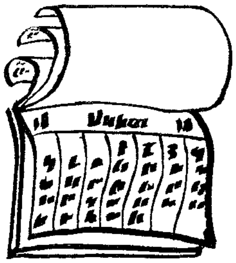
AFOSR 160

ZTB 139



PROGRESS REPORT

RESEARCH IN INDUCTIVE INFERENCE FOR THE PERIOD 1 APRIL 1959 TO 30 NOVEMBER 1960

R. J. Solomonoff

January 1961

CONTRACT AF49(638)-376

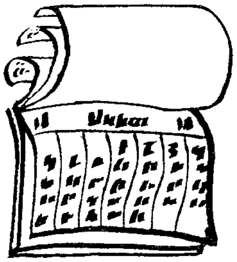
PREPARED FOR

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.

AFOSR 160

ZTB 139



PROGRESS REPORT

RESEARCH IN INDUCTIVE INFERENCE FOR THE PERIOD 1 APRIL 1959 TO 30 NOVEMBER 1960

R. J. Solomonoff

January 1961

CONTRACT AF49(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.

PROGRESS REPORT

RESEARCH IN INDUCTIVE INFERENCE FOR THE PERIOD 1 APRIL 1959 TO 30 NOVEMBER 1960

ABSTRACT

During the period covered by this report, the principal progress made was in the discovery of what are apparently several equivalent formal solutions to the general inductive inference problem. These solutions are applicable to numerical and/or non-numerical and/or analog and/or digital data. Any type of information that is available can be made part of the evidence upon which the inferences are made.

Much time was spent in attempts to verify the validity of and in finding valid counterexamples to these proposed solutions. No serious evidence of non-validity has been found, though the validity is as yet not entirely certain.

These general induction methods have been applied to several specific problems in non-numerical prediction. Some computer programs have been written for the discovery of regularities in English text and any other sequence of symbols. Some work has been done toward programming a computer to learn to assign descriptors to documents.

Beforé the new inference methods were discovered, much time was spent on the problem of discovering the grammars of phrase structure languages from a body of text alone. Since then, the general inference methods have cleared up a serious point of difficulty in this problem.

Another problem upon which considerable progress was made is the problem of programming a computer to improve its own inference methods. The general inductive inference solutions have not yet, however, been applied to this problem.

TABLE OF CONTENTS

PROGRESS REPORT

RESEARCH IN INDUCTIVE INFERENCE
FOR THE PERIOD 1 APRIL 1959 TO 30 NOVEMBER 1960

A. PROBLEMS UPON WHICH SIGNIFICANT PROGRESS HAS BEEN MADE

1. The General Inductive Inference Problem

1.1 The Nature of the Inductive Inference Problem and Some Applications of Its Solution.

The problem of inductive inference is to draw general conclusions from sets of isolated facts and to apply these conclusions to the making of predictions. Suppose we had a large number of documents and we had assigned to each one a set of catalog indices or descriptors. The problem of inductive inference would be to formulate general rules that relate the descriptors to the documents. These general rules could then be applied to the assignment of descriptors to new documents that are not in the original set that gave rise to the rules.

Ordinarily, the inductive inference process is performed by human beings. The process amounts to learning by numerous examples, and applying the learning to new situations. Mechanization of part of this process would enable us to assign descriptors to documents by machine.

In addition, the detailed understanding of the learning process in a machine may give some needed insight to the design of teaching machines. One of the most important problems in teaching is to decide just how large an intellectual jump should be made in presenting new material to be learned. The study of inductive inference gives us one way to evaluate quantitatively the size of "the conceptual jump" that is involved in the presentation of a particular new idea.

1. 2 Previous Work on the Inductive Inference Problem.

The problem of stating exactly the principles upon which valid inductive inferences may be performed is one of the oldest and most important problems in the philosophy of science and it has a large literature devoted to it.

Two of the most recent theories that have been proposed are those of R. A. Fisher and R. Carnap.

Fisher's theory is fairly general, but has not been rigorized to any great degree. It is quite distant from machine mechanization. It can be viewed as an application of Bayes' Theorem with a particular method of using "the principle of indifference." As with almost all applications of Bayes' Theorem, there are serious difficulties in obtaining the necessary a priori probabilities.

Carnap's theory is quite mechanizable, but it applies (as yet) to only a small part of all induction methods that are used. It, too, can be viewed as an application of Bayes' Theorem, and has, in this respect, some of the same difficulties that Fisher's method has.

1. 3 The Present Proposed Solution to the Inductive Inference Problem.

The proposed method is also an application of Bayes' Theorem. However, the a priori probabilities involved are obtained in what appears to be an unambiguous, perfectly general manner. In particular, the method is completely mechanizable and makes it possible to assign weights to all inductive methods that are describable to a universal Turing machine. The problem of Geometric Probability, which is approached by neither Fisher's nor Carnap's methods, appears to obtain a reasonable solution.

The method can be described in several ways. Although all of these ways have not been proved to be equivalent, their equivalence is probable.

One of the most picturesque descriptions of the a priori assignment of probabilities to all describable universes consists of viewing each universe description as being produced as the output of a universal Turing machine which has random input. A more detailed description of this idea is given in Reference 1, Section 12. A somewhat different, but closely related technique of inference is described in Appendix I.

Another way to express this inductive method is by viewing optimum extrapolation operators as being constructed by maximally redundant networks of neurons. Appendix II has a more exact statement of this. There appears to be some relevance to the problem of creating reliable computers from unreliable components.

Still another (apparently equivalent) method consists in making predictions by using a weighted average of all describable prediction methods. The weight of each depends on the past success of that prediction method, as well as on the "complexity" of the description of that method. See Appendix III for a more exact formulation.

2. Some Tests and Applications of the Proposed Solution to the Induction Problem

If a general inductive inference method is valid, it is, in principle, impossible to prove this in the sense that a mathematical theorem can often be proved. Instead, it is only possible to gain positive evidence for validity by applying the method to various problems and seeing that the results seem intuitively reasonable. A non-valid inference method can, however, sometimes be shown to be invalid by either showing it to be inconsistent, or by presenting problems for which the solution that it gives is intuitively unreasonable.

In the present case much time has been spent in showing that apparent counter examples were, indeed, dealt with in an acceptable manner by the proposed method. Also, several cases of apparent ambiguity were shown to be not actually ambiguous.

In the direction of more positive verification, applications of the new inference methods have been made to several specific problems. In most cases, this consists of devising a method of encoding a given body of data in a "minimal" manner, so that when the code is presented as input to a Turing machine, its output is the original body of data.

2. 1 The Coding and Extrapolation of a Bernoulli Sequence.

A Bernoulli sequence is a sequence of symbols whose sole regularity is that each symbol occurs with a certain frequency. Through a coding method suggested by a formula of Carnap's, a very reasonable solution to this problem was obtained. Two necessary constraints on the form of the code were found and the formula used satisfies both of them, though it is probably not the only formula that does so. The inference rule obtained is the same as that given by Laplace's rule of succession, if the only data we have for prediction is part of the Bernoulli sequence. Usually, other data are available, and in such cases, we are able to obtain the effects of better a priori probability than is given by Laplace's rule.

The method of coding used for the Bernoulli sequences was made the basis for all other coding methods that were used for other problems.

2. 2 The Coding and Extrapolation of Markov Chains.

Using a modification of the coding method used for Bernoulli sequences, it was shown that if a Markov process is definable by a finite discrete transition matrix, then for a long sample sequence of the Markov chain, the proposed inductive inference method is likely to give the correct probabilities for the continuations of this sequence.

2. 3 The Coding and Extrapolation of Simple Languages That Use Definitions.

A significant increase in power of induction methods occurs when codes containing definitions are used. A simple example of such a coding technique was studied in much detail. A computer program has been written for applying this technique to the discovery of regularities in a sequence of symbols.

The program looks for significant subsequences of symbols, and defines new special symbols corresponding to these subsequences. These new symbols can also form significant subsequences resulting in the definitions of newer symbols. This process continues until no new significant subsequences can be found. The resultant sequence, along with the definitions of its symbols, is then coded much in the manner of an ordinary Bernoulli sequence. The inductive inference method gives an unambiguous interpretation to the term "significant subsequence."

This program applied to English text, or any other language, may find prefixes, roots, suffixes and words. It can also be used to extrapolate written music, or any other type of sequence of symbols.

The program has been written, using the Fortran II compiler for the IBM 709. It has not yet been debugged or run.

A more ambitious program, using a more complex coding method corresponding in part to V. Yngve's "left to right sentence analysis," has been partially completed. Some theoretical difficulties were encountered, however, and it was felt expedient to temporarily discontinue work on this particular problem.

2. 4 The Coding and Extrapolation of Context-Free Phrase Structure Languages.

A formal solution had been devised for the problem of discovering the grammar of a phrase structure language that does not employ context-dependent substitution, given only a finite set of acceptable sentences in in this language. This was described in Appendix II of Reference 2. This "formal solution" was, however, incomplete, in that no exact method was given for assigning a priori probabilities to various grammars. Using the new inductive inference methods, a method for assigning these a priori probabilities was found, as well as a more unified treatment of the entire grammar discovery problem.

2. 5 The Mechanized Learning of Descriptor Assignment.

A coding technique called "correlational coding" has been devised which corresponds to the normal use of correlations for prediction. The method is superior to the methods normally used in that an arbitrarily large number of classes of events may be examined to find ones that give good predictions through their correlation with the event class to be predicted. In normal correlation methods this cannot be done without the danger of coincidental correlations that will not extrapolate. This limitation is particularly important in the analysis of small samples.

Correlational coding appears to be appropriate for the design of a machine that will learn to classify new documents into or out of a descriptor class after being given a sample set of documents to which the descriptor applies and another class to which the descriptor does not apply. The same techniques are directly applicable to the learning of probabilistic medical diagnosis by the computer. In this case the machine devises various combinations of symptoms and computes their utilities in prediction. This is done on the basis of their a priori probability as well as their effectiveness in prediction over the known body of correct diagnoses.

2.6 The Utility Evaluation Problem.

The problem of assigning utilities to abstractions used in inductive inference had been worked on for a considerable period of time in 1957 and 1958. The progress made by May 1959 is discussed in Section 2.7 of Reference 3. At that time work on this problem was discontinued since its solution was no longer vital and an apparently workable (though not altogether general) solution had been obtained.

The new inductive inference method has been applied to this old problem with much success. A very general solution has been obtained which appears to be far more satisfactory than the tentative solution of April 1958, described in Reference 4.

2.7 Reduction of Size of Adequate Sample for Statistical Decisions.

Ordinarily, in fitting curves to empirical data, and often in selecting optimum hypotheses to extrapolate empirical data, only one half of the data is used to select the optimum hypothesis. The other half of the data is then used to determine how well this hypothesis fits the data.

Using the new inductive inference methods, it is unnecessary to divide the data in this way. The new technique gives the optimum set of hypotheses and the expected future accuracy of the set of hypotheses in a unified manner by treating all data points in the same way. The effect is to significantly reduce the sample size necessary for a given expected prediction accuracy. In the case of statistical studies where cost is largely proportional to sample size, a significant reduction of cost (in money, time or whatever measure is used) can result.

B. WORK OF THE FUTURE

1. The Immediate Future

The computer program of Section A 2.3 will be debugged and applied to English text, and its efficacy as a prediction method will be compared with that of more conventional methods. It may be worthwhile to employ the program for the extrapolation of music sequences.

The program of Section A 2.5 for prediction by correlational coding will be completed and applied to whatever data is most appropriate. This will probably be descriptor assignment and/or medical diagnosis. It is believed that raw data for either of these problems is readily available since other people have worked on very similar problems using different inference techniques.

Another application of correlational coding that seems very readily implementable is the recognition of hand printed characters. Further study will be made of its feasibility.

2. The More Protracted Future.

At the present time a better theoretical understanding of the new inductive inference method seems to be a very important goal. Work on diverse types of inference problems has clarified and will probably continue to clarify many important ideas.

The analysis of problems involving continuous rather than purely digital data has been done to some extent (Reference 1, Section 14). Further work on problems containing both qualitative and quantitative data would be very helpful for many difficult types of "character recognition" problems such as recognition of handwriting, spoken words, faces of people, etc.

Much work on inductive inference machines that learn to improve their inference methods was done before the new general inductive inference methods were discovered. The self-improving machine was a very promising approach at the time that work on it was temporarily discontinued, and it is felt that the new methods of inference may contribute very strongly to progress in this very important problem.

Appendix II discusses an extrapolation operator that is a network of suboperators, and is maximally resistant to mutations of certain kinds. Further study of such operators may suggest methods to construct reliable computers from unreliable components. These methods would be more "global" than any of the redundancy systems proposed up to the present time for increasing system reliability.

APPENDIX I

We shall describe a method for obtaining the a priori probability of a very long string of symbols. This string will be referred to as "the corpus." It will be convenient to let this string be something like a description (in almost any fairly consistent language) of all the things that a man could observe in his life.

Suppose the set, A, is the alphabet of symbols in the corpus. Then take a suitable "universal machine" (as described in Reference 1, Section 9) whose input alphabet is 0 and 1, and whose output alphabet in A.

Suppose A has N(A) different symbols in it, and the corpus has a total of k symbols.

Then choose some large number, R, such that $R >> k \log_2 N(A)$. Consider an arbitrary binary string S of length R bits.

Let M(S) be the output of our universal machine, when its input is the binary string, S. Now consider a number of such strings.

Let B be the set of strings, S_i, such that $M(S_i)$ exists, i.e., the machine computation eventually terminates if S_i is used as input.

Let N(B) be the number of strings in the set B.

Let C be the subset of B such that the first k symbols of $M(S_i)$ are the same as those of the corpus.

Let N(C) be the number of strings in the set C.

Then the a priori probability of the corpus is

$$\frac{N(C)}{N(B)} .$$

This a priori probability is the fraction of all meaningful binary string inputs of length R that give rise to the corpus as output. A "meaningful" binary string is one for which the computing process terminates in a finite number of steps.

ΓB

9

APPENDIX II

We shall consider an "operator" to be a device which can receive a string of binary symbols as input and present another or the same binary string as output.

Let $A \equiv [\ a_i, b_i\ ]$ be a set of ordered pairs of binary strings. a_i may be thought of as a stimulus, and b_i as the desired response to that stimulus. We want to extrapolate this list by constructing an operator that will respond in the "desired way" to new inputs.

To do this, consider a certain fixed large "adequate" set of "neurons." Here "neuron" is a mathematical device of the McCullough-Pitts type. It has several input channels and one output. Its output at time t is a Boolean function of its inputs at time $t-1$. An "adequate" set of such neurons will contain enough neurons and an adequate variety of neurons, so that by suitably interconnecting all of the neurons in the set it is possible to construct any operator.

To further define a specific operator constructed in this way, we will designate certain neuron input channels as "operator inputs" and certain neuron output channels as "operator outputs" for that operator. A different operator would be defined by different internal connections and/or different input-output designations.

Consider the set of all operators such that for any specific operator in the set and for all $[\ a_i, b_i\ ]$ in A, the string a_i presented to the specific operator inputs will result in the string b_i being excited at the specific operator outputs.

In general, if our pool of neurons is large, there will be many operators in this set. We will say that such operators are "of type G."

Consider a particular operator of type G. Let us select at random N neuron outputs that are not outputs of this operator, and short circuit them by fixing them permanently at zero. For certain operators of type G, this procedure will cause the operator to remain of type G, with very high probability. Such operators will be defined to be of "redundancy, N" with respect to the set A.

Of the set of all operators constructable from our fixed pool of neurons, consider the subset of type G. Those of this subset that have the maximum redundancy with respect to the set A will have the greatest likelihood of extrapolating properly from the set A.

APPENDIX III

A probability evaluation method (which we shall abbreviate "PEM") is an operator that accepts a string of symbols as input, and, as output, presents a set of fractions that give the probabilities of the next symbol in the string.

Let $\bar{A}(S)$ be the vector whose components give the probabilities of the next symbol of string S, in view of PEM, A.

If A is a PEM, and S is a string of k symbols, then let $A(S)$ represent the probability of S, in view of A. To obtain $A(S)$, first compute by A the probability P_n of the n^{th} symbol of S, in view of the previous $n - 1$ symbols of S. Then $A(S) \equiv \prod_{i=1}^{k} P_i$.

Next, we shall define the "description" of a PEM. Let M represent a universal Turing machine so that if string X is the input to the machine, then the string $M(X)$ is its output. Then the string $D(A)$ is "the description of PEM, A, with respect to Machine M," if $M(D^\frown S) = \bar{A}(S)$ for all strings, S. Here $\bar{A}(S)$ is a sequence of binary symbols that represents the required probabilities. $D^\frown S$ is the concatenation of D and S. In general, we will allow D to contain only the symbols zero and one.

Then, if S is a suitably long corpus, the unnormalized probability distribution vector for the next symbol of S is

$$\sum_i 2^{-L(D(A_i))} A_i(S)\bar{A}_i(S)$$

where $L(D(A_i))$ is the number of digits in the string $D(A_i)$, and the summation on i is to be over all conceivable PEM's, A_i.

This probability distribution is to an important extent independent of just which universal Turing machine is used.

REFERENCES

1. R. J. Solomonoff, "A Preliminary Report on a General Theory of Inductive Inference," Zator Technical Bulletin No. 138; AFOSR TN-50-1459. Zator Company, November 1960. (Circulated in an earlier form as publication V 131, dated February 4, 1959.)
2. R. J. Solomonoff, "A Progress Report on Machines to Learn to Translate Languages and Retrieve Information," Zator Technical Bulletin No. 134; AFOSR TN-59-646. Zator Company, October 1959.
3. R. J. Solomonoff, "Progress Report: Research in Inductive Inference for the Year Ending 31 March 1959," Zator Technical Bulletin No. 130; AFOSR TN-59-219. Zator Company, May 1959.
4. R. J. Solomonoff, "Utility Evaluation," Publication V 123, Zator Company, April 1958.

Invited Formal Talks

1. 19 January 1959, M. I. T., Research Seminar on Artificial Intelligence; talk on "Applications of Generalized Languages to Pattern Discovery."
2. 27 February 1959, M. I. T., Research Laboratory for Electronics, Mechanical Translation Seminar; talk on "Some New Parsing Routines, Mechanical Translation, and Theoretical Gap Analysis."
3. 12 May 1959, IBM Research Laboratories, Lamb Estate, Yorktown Heights, N. Y.; talk on "Discovery Methods for Phrase Structure Grammars and Applications to Inductive Inference."
4. 15 and 18 May 1959, two guest lectures at M. I. T. class on "Artificial Intelligence;" talks on "Use of Formal Languages for Inductive Inference."
5. 19 August 1959, M. I. T., Research Seminar on Artificial Intelligence; talk on "Some Generalizations of Phrase Structure Lanuages and Their Applications."
6. 25 September 1959, Radio Corporation of America, Sarnoff Laboratories, Princeton, N. J.; talk on "Phrase Structure Languages, Properties and Applications."
7. 2 March 1960, University of Pennsylvania, Philadelphia, Pa.; talk on "The Use of Formal Languages for Mechanized Inductive Inference."
8. 4 March 1960, National Bureau of Standards, Washington, D. C.; talk on "The Use of Formal Languages for Mechanized Inductive Inference."

Unpublished Studies

1. Expansions to the Appendices to the Cleveland paper (AFOSR-TN-59-646):
   (a) "Stochastic Languages," 2 pp.
   (b) "Approximation Languages, with Applications to Information Retrieval; a Formal Solution to One Aspect of Descriptor Assignment Learning," 6 pp.
2. "Generalized Pattern Discovery by Linguistic Methods" (draft), 47 pp. with sections:
   (a) Definitions of Pattern, Language, Grammar.
   (b) How certain kinds of MT, several examples of arithmetic are examples of phrase structure languages.
   (c) Some very general kinds of phrase structure languages.
   (d) How previous methods of grammar discovery are specifically applicable to MT discovery, as well as to discovery of some more general phrase structure grammars.
   (e) Application to examples from English to French translation.
   (f) Discussion of the possible advantages of phrase structure methods to MT.
   (g) Discussion of some higher order languages.

3. Preliminary FORTRAN computer program for the discovery of regularities in a sequence of symbols (see Section 2. 3, page 5).

Meetings Attended in Addition to Those at Which Papers Were Presented

1. Western Joint Computer Conference, San Francisco, 3-5 March 1959.
2. Interdisciplinary Conference on Self-Organizing Systems, sponsored by ONR and University of Illinois, Chicago, 5-6 May 1959.
3. Eastern Joint Computer Conference, Boston, Mass., 13-15 December 1959.
4. American Mathematical Society, New York City, 14-16 April 1960.