A NEW MODEL FOR INDUCTIVE INFERENCE

(Extended Abstract)

Ronald L. Rivest*
MIT Lab. for Computer Science
Cambridge, Mass. 02139 USA

Robert Sloan†
MIT Lab. for Computer Science
Cambridge, Mass. 02139 USA

Abstract

We introduce a new model for inductive inference, by combining a Bayesian approach for representing the current state of knowledge with a simple model for the computational cost of making predictions from theories. We investigate the optimization problem: how should a scientist divide his time between doing experiments and deducing predictions for promising theories. We propose an answer to this question, as a function of the relative costs of making predictions versus performing experiments. We believe our model captures many of the qualitative characteristics of "real" science.

We believe that this model makes two important contributions. First, it allows us to study how a scientist might go about acquiring knowledge in a world where (as in real life) there are costs associated with both performing experiments and with computing the predictions of various theories.

This model also lays the groundwork for a rigorous treatment of a machine-implementable notion of "subjective probability". Subjective probability is at the heart of probability theory [5]. Previous treatments have not been able to handle the difficulty that subjective probabilities can change as the result of "pure thinking"; our model captures this (and other effects) in a realistic manner. In addition, we begin to provide an answer to the question of how to trade-off "thinking" versus "doing"—a question that is fundamental for computers that must exist in the world and learn from their experience.

Authors' net addresses: rivest@theory.lcs.mit.edu, sloan@theory.lcs.mit.edu

*This paper was prepared with support from NSF grant DCR-8607494, ARO Grant DAAL03-86-K-0171, and the Siemens Corporation.

†Supported by an NSF graduate fellowship.

1 Introduction

We examine the process of "inductive inference"—the process of drawing inferences from data. Angluin and Smith [1] provide an excellent introduction and overview of previous work in the field. Our work is distinguished by the following features:

- Our inference procedure begins with an *a priori* probability associated with each possible theory, and updates these probabilities in a Bayesian manner as evidence is gathered.
- Our inference procedure has two primitive actions available to it for gathering evidence, each of which has a cost (in terms of time taken):
  1. Using a theory to predict the result of a particular experiment.
  2. Running an experiment.
- Our inference procedure attempts to maximize the expected "rate of return", for example, in terms of the total probability of theories eliminated per unit time.

Our approach addresses the following three issues, which we feel are not always well handled by previous models.

(1) Induction is fundamentally different from deduction. Much previous work has tried to cast induction into the same mold as deduction: given some data (premises) to infer the correct theory (conclusion). This approach is philosophically wrong, since experimental data can only eliminate theories, not prove them. (See Feyerabend [3] and Kugel [7].) For similar reasons, we feel it is better to study inference procedures which represent the *set* of remaining theories (and perhaps their probabilities), rather than inference procedures which are constrained to return a *single* answer.

(2) The difficulty of making predictions is overemphasized. Much of the previous theoretical work in this area has been recursion-theoretic in nature, and the richness of the results obtained has been in large part due to the richness of the theories allowed; allowing partial recursive functions as theories makes inference very difficult. The resulting theory probably overemphasizes this recursion-theoretic aspect, compared to the ordinary practice of science. In this paper, all theories will be total (they predict a result for every experiment), and we assume that the cost of making such a prediction from a theory is a fixed constant c (time units), independent of the theory or the proposed experiment. (This is obviously an oversimplification, but serves our purposes well.)

(3) Experiments take time, and should be carefully chosen. Much of the previous work on inductive inference has assumed that the data (i.e. the list of all possible experimental results) is presented to the learner in some order (cf. [4,2]). However, the rate of progress in science clearly depends on which experiments are run next. (Consider experimental particle physics today.) Part of doing science well is choosing the right experiments to do.

A good scientist must decide how to allocate his time most effectively—should he next run some experiment (if so, which one?), or should he work with one of the more promising theories, computing what it would predict for some experiment (if so, which theory and which experiment?). These "natural" questions are not particularly well handled by previous models of the inductive inference problem, but our model will allow us to answer such

questions. Our results also shed some interesting light on related questions, such as when to run "crucial" experiments that distinguish between competing hypotheses.

Our model can perhaps be viewed as well as a contribution to the theory of subjective probability [5], which has traditionally had a problem with the fact that subjective probabilities can change as a result of "pure thinking". Various proposals, such as "evolving probabilities" [6] have been proposed, but these do not deal with the "thinking" aspect in a clean way.

2 The Model

2.1 Basic Notation and Assumptions

We assume the existence of some scientific domain of interest, defined by an (infinite) set of possible experiments. Performing the j-th experiment yields a datum χ_j; in this paper we assume for convenience that $\chi_j \in \{0,1\}$. We make the simplistic assumption that doing an experiment always takes precisely d units of time (independent of which experiment is performed).

We assume that there are an infinite (but enumerable) set of theories available about the given domain; we denote them as $\varphi_0, \varphi_1, \ldots$. Each theory is understood to be a total function from $\mathbf{N}$ into $\{0,1\}$; the value $\varphi_i(j) = \varphi_{ij}$ is the "prediction" theory φ_i makes about the result of experiment j. We assume there exists a *correct* theory, φ_r, such that $(\forall j)\varphi_{rj} = \chi_j$. We make the simplistic assumption that computing φ_{ij} from i and j always takes precisely c units of time (independent of i and j).

We assume that other operations, such as planning, take no time.

Our scientist begins with two kinds of *a priori* probabilities:

- The *a priori* probability that $\varphi_{ij} = 1$, for any i and j. We assume that $\Pr(\varphi_{ij} = 0) = \Pr(\varphi_{ij} = 1) = \frac{1}{2}$ *a priori*, for all i and j; the scientist has no reason to expect his theory to predict one way or the other, until he actually does the computation.
- The *a priori* probability p_i that theory $\varphi_i = \varphi_r$, (i.e. that φ_i is correct). We assume that the p_i's are computable, that $(\forall i)p_i > 0$ (all theories are possible at first), and that that $p_0 \geq p_1 \geq \ldots$.

We assume that these *a priori* probabilities are correct; the reader may imagine that "god" first determined all of the φ_{ij}'s by independent unbiased coin-flips, and then selected one of the theories at random to be correct (according to the probability distribution $p_0, p_1, \ldots$).

2.2 The Scientist Makes Progress

Our scientist begins in a state of total ignorance, and proceeds to enlighten himself by taking steps consisting of either doing an experiment (determining some χ_j) or making a prediction (computing some φ_{ij}). The scientist may choose which experiments and predictions he wishes to do or not to do, and can do these in any order (predictions may precede or follow corresponding experiments, for example).

We need notation to denote the scientist's state of knowledge at time t (after t steps have been taken).

- Let "$\perp$" denote "unknown".
- Let $\varphi_{ij}^t \in \{0, 1, \perp\}$ denote the scientist's knowledge of φ_{ij} at time t.
- Let $\chi_j^t \in \{0, 1, \perp\}$ denote the scientist's knowledge of χ_j at time t.

If at time t both $\varphi_{ij}^t = \varphi_{ij}$ and $\chi_j^t = \chi_j$ (i.e. both are known at time t), then there are two possibilities. Either $\varphi_{ij} \neq \chi_j$, in which case theory φ_i is *refuted*, or $\varphi_{ij} = \chi_j$, in which case theory φ_i is (to some extent) *confirmed*.

2.3 How Long Will Science Take?

Obviously, after a finite number of steps, our scientist will be able to refute only a finite number of theories, so at no point will he be able claim that he has discovered the complete "truth".

More realistically, he may ask "How long will it be before I have eliminated all theories with higher *a priori* probability than the correct theory?" The answer here depends on the set of *a priori* probabilities. A realistic "non-informative prior" attempts to have p_i decrease to zero as slowly as possible; for example we might have $p_i = C \cdot (i \ln(i) \ln \ln(i) ...)^{-1}$, where C is a normalizing constant and only the positive terms in the series of logarithms are included [8].

Note that at least one step is required to eliminate a theory, so that the expected number of steps required to eliminate all theories with higher *a priori* probability than the true one is at least equal to the expected number of such theories, i.e.

$$\sum_{r=0}^{\infty} r \cdot p_r \ ,$$

which is infinite. This result holds for many similar probability distributions which do not go to zero too quickly.

In fact, for a typical set of prior probabilities, our scientist expects to have an infinite amount of work to do before the true theory is even considered!

For this reason, among others, we will concentrate on the rate at which the scientist can refute false theories, rather than on the expected time taken before the scientist would assert that, on the basis of the evidence available to him, φ_r is the best available theory.

2.4 How the Scientist Updates His Knowledge

To model the evolution of the scientist's knowledge more carefully, we show how his subjective probabilities associated with the various theories change as a result of the steps he has taken, using Bayes' Rule.

What happens to the probabilities maintained by the scientist after step t is performed? Let p_i^t denote the probabilities after step t (here $p_i^0 = p_i$). We consider the effect of step t on the probability that theory φ_i is correct. That is, we look at how p_i^{t-1} is updated to become p_i^t.

The process of updating these probabilities according to the result of the last step, can be performed by executing the following operations in order:

1. For all i,
   - Set p_i^t to 0 if φ_i has just been refuted.
   - Set p_i^t to $2 * p_i^{t-1}$ if φ_i has just been confirmed.
   - Otherwise set p_i^t to p_i^{t-1}.
2. Normalize the p_i^t's so that they add up to 1.

The above procedure follows directly from Bayes' Rule, since it is equally likely for a prediction to be a 0 or a 1.

We note that if the scientist just sits and "thinks" about an experiment (i.e. he just computes the predictions of various theories for this experiment), his subjective probability that $\Pr(\chi_j = 0)$ will *evolve*, since

$$\Pr(\chi_j = 0) = \sum_{\varphi_{ij}^t = 0} p_i^t + \frac{1}{2} \sum_{\varphi_{ij}^t = \perp} p_i^t.$$

It would also not be unreasonable to treat this probability as an interval, since one knows the upper and lower limits that it could evolve to.

2.5 An Example

Consider Table 1, which illustrates a portion of a particular scientist's knowledge at some point in time. (Here unknown values are shown as blanks, and only a portion of the actual infinite table is shown.)

The second row of the table shows which experiments he has run. (Here he knows only $\chi_0 \ldots \chi_4$.) The second column gives his current probabilities p_i^t.

The second part shows what predictions he has made. Each row of this table corresponds to one theory. Theories which have been refuted have current probability zero and are not shown here; *it is convenient from here on to assume that φ_0 is the most probable theory, φ_1 is the next most probable theory, and so on.* In this example, the scientist has found out what his most probable theory predicts for experiments 0–5, and so on.

Running experiment 5 next has the potential of refuting φ_0. (It will either refute φ_0 or φ_3.) Making the prediction $\varphi_{1,5}$ can not (immediately) refute φ_1, but would affect the scientist's estimate of the likelihood that $\chi_5 = 0$. With the current state of knowledge, the scientist would estimate that

$$\Pr(\chi_5 = 0) = 0.04 + \frac{1}{2}(1 - 0.60 - 0.04) = 0.22.$$

Note, however, that $\Pr(\varphi_{1,5} = 0)$ remains $\frac{1}{2}$, independent of anything else, until it is computed.

3 Our Inference Procedures

The approach taken by a scientist will depend upon the relative costs of making predictions versus doing experiments, his initial probabilities for the theories, and exactly how he wishes to "optimize" his rate of progress.

			j						
			0	1	2	3	4	5	6
i	p_i	$\chi_j \rightarrow$	0	1	1	0	0		
0	0.60		0	1	1	0	0	1	
1	0.10		0	1	1	0			
2	0.05		0	1	1				
3	0.04	$\varphi_{ij} \rightarrow$	0					0	
4	0.03			1	1				
5	0.02		0						
6	0.01								

Table 1: Partial View of Scientist's State of Knowledge

3.1 General Assumptions

At each step, the scientist must decide what to do next. Although this choice is, and always remains, a choice among an infinite number of alternatives, it is reasonable to restrict this to a finite set by adopting the following rules:

- When running or predicting the result of an experiment which has neither been previously run nor had predictions made for it, without loss of generality choose the least-numbered such experiment available.
- When making a prediction for a theory for which *no* previous predictions have been made, choose the most probable such theory (in the case of ties, choose the least-numbered such theory).

3.2 Optimization Criteria

The scientist will choose what actions to take according to some optimization criteria. For example, he may wish to:

1. Maximize the expected total probability currently associated with theories which are refuted by the action chosen.
2. Minimize the entropy $-\sum_{i=0}^{\infty} p_i^t \log(p_i^t)$ of his assignment of probabilities to theories.
3. Maximize the probability assigned to the theory he currently believes to be the most likely.
4. Maximize the highest probability assigned to any theory.
5. Minimize the expected total probability assigned to *incorrect theories.*

More generally, he may wish to maximize his "rate of progress" by dividing his progress (measured by one of the above criteria) by the time taken by the action chosen.

In this paper we will discuss all of the above optimization criteria; some very briefly, and some at length. In the remainder of this section we discuss the general form that all our inference procedures take, regardless of the particular optimization criterion they use.

3.3 Menus of Options

We propose that the scientist organize his strategy as a "greedy" strategy of the following form:

- He organizes his decision at each step into a finite number of options. Each such option is a *program* specifying a sequence of predictions and/or experiments to run, which terminates with probability 1.
- At a given step, for each available option, the scientist computes the expected "rate of return" of that option, defined as the expected total gain of that option (where gain is measured by some optimization criterion) divided by the expected cost of that option.
- The scientist then chooses to execute an option having highest expected rate of return, breaking ties arbitrarily.

The reason for introducing the notion of an "option", rather than just concentrating on the elementary possibilities for a given step, is that certain steps have *no* expected rate of return in and of themselves. For example, making a prediction when the corresponding experiment has not yet been run has zero expected rate of return, as does running an experiment when no prediction regarding that experiment has yet been made.

From now on, we let q_i^t denote $1 - p_i^t$. We also observe that if our set of probabilities satisfies $p_0 \geq p_1 \geq \ldots$ then it also satisfies $p_0 q_0 \geq p_1 q_1 \geq \ldots$, since p_0 is no further from $\frac{1}{2}$ than p_1 is and $\frac{1}{2} \geq p_1 \geq p_2 \geq \ldots$.

4 Inference procedure 1: Maximizing the weight of refuted theories

We begin by studying an inference procedure which tries to refute wrong theories as quickly as possible. Specifically, the scientist will choose an action which maximizes the quotient of the expected total probability of theories eliminated by that action, divided by the cost of that action. The reason for this choice is its simplicity, and the ease with which the scientist can implement such a strategy. Furthermore, if our *a priori* probability happens to be one of the ones for which infinite expected time is required simply to eliminate all wrong theories (See section 2.3.), then this measure probably makes the most sense.

4.1 A Simple Menu of Options

In this subsection and the following subsection, we will spell out a particular menu of options and analyze our scientist's strategy when he uses this menu and the "maximizing the weight of refuted theories" optimization criterion. In later sections we will analyze our scientist's strategy when he uses the same menu but different optimization criteria.

We first consider the following two options, each of which will always have non-zero expected rate of return:

- *Prediction/Experiment Pair:* Make a prediction φ_{0j} for the least j for which no predictions yet exist, and then run the corresponding experiment. Here, as usual, φ_0 denotes the theory which is currently most probable. Our expected rate of return is

$$\frac{p_0^t q_0^t}{2(c+d)}.$$

We aren't compelled to restrict the prediction/experiment pairs to using the most probable theory, but do so because it is convenient to limit our options, and also because the expected return from other theories will not be as good.

- *Prediction:* Compute a prediction φ_{ij}, given that the corresponding experiment determining χ_j has already been run. The expected rate of return for this prediction is
$$\frac{p_i^t q_i^t}{2c}.$$
Here again it is clear that we should choose the least i possible, so as to maximize the rate of return.

If we stick to options in this simple menu, then the opportunity to make a prediction only arises after the simple prediction/experiment pair has already been run for that experiment.

4.2 An Expanded Menu of Options

An expanded menu can be obtained by adding the following two options to the simple menu:

- *Simple Experiment:* Run experiment j, given that at least one prediction has been made for this experiment. The expected rate of return is
$$\frac{p_0^t q_0^t}{2d},$$
since the probability that "truth" differs from φ_0 is $q_0^t/2$, and (as argued below), in this case we must have only the prediction φ_{0j}.

- *Crucial Two-Way Experiment:* Determine the least j such that the two most probable theories make differing prediction for χ_j. Then run experiment j. The expected rate of return is
$$\frac{p_0^t + p_1^t - (p_0^t - p_1^t)^2}{2(4c+d)} = \frac{p_0^t q_0^t + p_1^t q_1^t + 2p_0^t p_1^t}{2(4c+d)}. \quad (1)$$

We note that in the expanded menu, the only way an opportunity can arise to run a simple experiment is by having the search for a crucial experiment generate predictions for the first two theories, without running the corresponding experiment since the predictions were identical. This is the only way we can obtain a situation where predictions have been made for experiments that haven't been run. Furthermore, additional predictions won't be made for this experiment until after this experiment has been run. Since the crucial experiment will eliminate one of the top two theories, we will be left in a situation where (after renumbering of theories as usual) there is a j for which we know φ_{0j} but have not yet run experiment j.

Note that the expected cost of *finding* a crucial experiment is exactly $4c$, since if we pick a j and compute φ_{0j} and φ_{1j}, we have a $\frac{1}{2}$ chance of finding j to be crucial.[1]

We claim that, using either the simple or expanded menu, the *relative* order of two theories will not change, except when a theory is refuted, if an optimal greedy strategy is

[1] Note also, that there is no special reason to restrict ourselves to crucial two-way experiments. We could also run crucial n-way experiments, where we find the least j such that the n most probable theories split as evenly as possible (in terms of probability weight). Now the expected cost of finding such a j increases from $4c$ to $(2^n + 2^{n-1} - 2)c$.

used. This follows since it is always preferable to work with the more probable theories, given a particular option, and this work will tend to enhance the probability of that theory if it is not refuted.

Having given our menu of options, we can now make one simple definition. When we speak of *checking* or *testing* φ_i, we are talking about either doing a prediction/experiment pair involving φ_i or doing simple experiment j for some j for which φ_i has already made a prediction. In short, testing φ_i means to take some action that could potentially refute φ_i.

4.3 Behavior of this Inference Procedure

4.3.1 For the Simple Menu

For the simple menu, clearly we begin with a prediction/experiment pair. After that, the scientist will oscillate between further testing of his best theory (using prediction/experiment pairs), and testing of his other theories (using predictions).

The ratio $c/(c+d)$ will affect the relative amount of time spent on prediction/experiment pairs. We will typically see all theories down to some probability threshold (depending on c, d, and p_0) fully checked out against existing experimental data, before proceeding with the next prediction/experiment pair.

4.3.2 For the Expanded Menu

Given our assumption that it is more expensive to perform an experiment than to compute a theory's prediction, our scientist will at least want to consider whether he should get his experimental data from crucial experiments rather than from prediction/experiment pairs.

Let's consider whether at the beginning of time, the scientist is better off running a prediction/experiment pair, or running a crucial two-way experiment. The crucial experiment will have a higher expected rate of return if

$$\frac{p_0 + p_1 - (p_0 - p_1)^2}{2(4c + d)} > \frac{p_0 q_0}{2(c + d)} \tag{2}$$

or

$$\frac{d}{c} \geq \frac{3p_0 q_0}{p_1(2p_0 + q_1)} - 1.$$

It is sufficient for equation 2 to hold if

$$\frac{c + d}{3c} > \frac{p_0 q_0}{p_1 q_1}.$$

We see that for any ratio d/c, it is possible to have a crucial experiment be advantageous over a prediction/experiment pair; consider what happens when $p_0 = p_1 = \frac{1}{2}$.

No matter how cheap experiments get, relative to the cost of making predictions, it is possible to find a probability distribution where it is advantageous to find an experiment which will be crucial, before running any experiments.

Thus in general, it may pay to use the expanded menu, for any values of d and c.

5 Inference procedure 2: A minimum entropy approach

The entropy of a probability distribution P,

$$H_2(P) = \sum_{i=1}^{\infty} -p_i \log_2 p_i^{\dagger}, \tag{3}$$

is considered to be a good measure of the information contained in that probability distribution. Maximizing entropy corresponds to maximizing uncertainty; minimizing entropy corresponds to minimizing uncertainty. Thus a reasonable optimization criterion for our scientist would be minimizing the entropy of the *a posteriori* probability distribution.

Unfortunately, for some probability distributions, the entropy will be infinite. Consider, for instance, the previously mentioned distribution due to Rissanen [8],

$$p_i = C \cdot (i \ln(i) \ln\ln(i) ...)^{-1} , \tag{4}$$

where C is a normalizing constant and only the positive terms in the series of logarithms are included. Wyner [9] shows that the entropy series, equation 3, converges only if the series $\sum_{i=1}^{\infty} p_i \log i$ is convergent, but this series is clearly diverges for the distribution given in equation 4.

However, any particular experiment or prediction made by our scientist only causes him to alter a finite number of his *a posteriori* probabilities for theories. Thus, while the total entropy for the probability distribution may well be infinite, the *change* in entropy caused by any action will be a fixed finite amount.

The above discussion leads us to a precise description of the optimization criterion for our second inference procedure. The scientist chooses an action which maximizes the quotient of the expected decrease in the entropy of the probability distribution resulting from that action, divided by the cost of that action.

5.1 Behavior of this Inference Procedure

Let's begin by calculating the expected change in entropy for each of our action in our (expanded) menu.

- For computing the prediction φ_{ij} (assuming that χ_j is already known), we get,

$$\mathrm{E}[\Delta(H(P))] = -p_i + .5(1-p_i)\log(1-p_i) + .5(1+p_i)\log(1+p_i). \tag{5}$$

- For running a two way experiment between φ_0 and φ_1 we get

$$\mathrm{E}[\Delta(H(P))] = -p_0 - p_1 + .5(1+p_0-p_1)\log(1+p_0-p_1) + .5(1-p_0+p_1)\log(1-p_0+p_1). \tag{6}$$

- In fact, in general, for running χ_j where the total probability weight of theories which predict that χ_j will be zero is r_0 and the total probability weight of theories which predict that χ_j will be one is r_1 we get

$$\mathrm{E}[\Delta(H(P))] = -r_0 - r_1 + .5(1+r_0-r_1)\log(1+r_0-r_1) + .5(1-r_0+r_1)\log(1-r_0+r_1). \tag{7}$$

[†]Throughout this section we will discuss entropy in bits, and will henceforth assume all logarithms without an explicit base to be base 2.

Consider the probability distribution, R, that has only two outcomes, one with probability $r_0 + .5(1 - r_0 - r_1)$, the other with probability $r_1 + .5(1 - r_0 - r_1)$. We can rewrite equation 7 in terms of the entropy of R,

$$\mathrm{E}[\Delta(H(P))] = -r_0 - r_1 + H(R). \tag{8}$$

Equations 5 and 6 can be rewritten in a similar manner (since really they're just special cases of equation 7).

In fact, the calculations for this entropy driven inference procedure and the previous, "Kill wrong theories" driven procedure yield very similar results. Equation 8 and equation 1 could both be written as

$$\text{PROGRESS} = k(r_0 + r_1 - \text{penalty}(|r_0 - r_1|). \tag{9}$$

(The difference in signs between equation 8 and equation 9 arises because in equation 8 we're trying to *minimize* entropy, so our progress is negative, and our penalty is positive.)

Let $\delta = |r_0 - r_1|$. For the entropy approach, $k = 1$ in equation 9, and $\text{penalty}(\delta) = H(.5+\delta/2, .5-\delta/2)$. (In terms of r_0 and r_1 that probability distribution is $r_0+u/2, r_1+u/2$, where $u = 1 - r_0 - r_1$ is the *undecided* probability weight—the total probability weight of those theories i such that $\varphi_i(j) = \perp$.) For the kill wrong theories approach of the previous section, $k = 1/2$ in equation 9, and $\text{penalty}(\delta) = \delta^2$.

As one might expect given this strong similarity between the two optimization criteria, the inference procedures behave in a roughly similar manner.

6 Inference procedure 3: Making the best theory good

Our scientist might decide that he would like to at all times have a theory that's "pretty good." There are several approaches he might take.

In the extreme, he might simply decide that his goal would be to always increase the *a posteriori* probability assigned to the current best theory. Such a cynical strategy turns out to be impossible. No actions lead to an *expected* increase in the probability assigned to the best theory. If we check the best theory with any kind of action, then with probability $p_0 + .5(1 - p_0)$ it is confirmed, and its probability goes up to $2p_0/(1 + p_0)$. However, with probability $1 - p_0$ it is refuted and its probability goes to zero. Thus its expected probability after any action is $[(p_0 + 1)/2] \cdot 2p_0/(1 + p_0) = p_0$. If we check other theories, they may be either refuted, which would increase the probability assigned to φ_0, or confirmed, which would decrease the probability assigned to φ_0, and it again works out that the expected value of the *a posteriori* probability weight assigned to φ_0 is p_0.

Since our scientist cannot steadily increase the probability assigned to the best theory, he might settle for a strategy which always keeps the current best theory best. To accomplish this goal, the scientist should never test φ_0 against any theory. He should simply test the other theories, making sure to stop testing φ_i as soon as $p_i \geq .5p_0$ (otherwise φ_i might replace φ_0 as best). This procedure is obviously uninteresting.

There is, however, at least one interesting way for the scientist to always have a "pretty good" best theory. The scientist chooses an action to maximize the quotient of the expected value of the probability weight assigned to the best theory not yet refuted after that action, divided by the cost of that action.

6.1 Behavior of this Inference Procedure

The first thing we do is calculate the expected value of the weight assigned to the best theory for each action.

- If we test φ_0 (with any kind of action), then with probability $p_0 + .5(1 - p_0)$ it will be confirmed, and the probability weight for the best theory will become $2p_0/(1 + p_0)$. With probability $.5(1 - p_0)$, φ_0 will be refuted, and the probability weight for the best theory will become $p_1/(1 - p_0)$. The expected value of the probability weight for the best theory is therefore $p_0 + p_1/2$.
- If $p_i \leq .5p_0$ (so if even we test and confirm φ_i it will still have a lower *a posteriori* probability weight than φ_0), then testing φ_i does not lead to an increase in the expected value of the probability weight of the best theory.
- If $p_i > .5p_0$, and we test φ_i, then the expected value of the probability weight of the best theory after the test is $p_i + p_0/2$.

  Note however, that this situation is of no practical importance. If χ_j is known and both φ_{0j} and φ_{ij} are unknown, then it will be more profitable to compute φ_{0j} than to compute φ_{ij}. Consider now the case where there is some j such that $\chi_j^t = \varphi_{0j}^t$ but $\varphi_{ij}^t = \perp$. Whichever theory is now numbered zero began with an initial probability weight greater than or equal to the initial probability weight of of the theory now numbered i. Moreover, since at time t φ_0 has been confirmed more than φ_i, it must be that $p_0^t \geq 2p_i^t$.
- If we run a crucial experiment for the two best theories, then the expected value of the probability weight of the best theory is $p_0 + p_1$.[2]

Having listed the payoffs for each action, we can now give the payoff/cost ratios for the actions we might take:

- A simple pair with the best theory: $\frac{p_0+.5p_1}{c+d}$.
- Prediction for φ_{0j} if χ_j known: $\frac{p_0+.5p_1}{d}$
- Simple experiment χ_j where φ_{0j} is known: $\frac{p_0+.5p_1}{d}$.
- Crucial two way experiment: $\frac{p_0+p_1}{4c+d}$.
- We might consider running a two way experiment when we have some leftover predictions (say from an earlier two way experiment) for one of the two theories. If we have k such predictions, then the expected cost decreases from $d+4c$ to $d+(3-\sum_{i=1}^{k-1} 2^{-i})c$.

All our scientist needs to do is pick the maximum reward/cost action from the above list, but we'll make a few qualitative observations here: If there is a j for which χ_j is known but φ_{0j} is not, then it's always best to compute φ_{0j}. It's better to do a crucial experiment instead of a simple pair if $d/c > 6p_0 + 2p_1$; otherwise it is better to do the simple pair.

[2] In this case there we gain nothing by running a crucial experiment for the best n theories for $n > 2$.

7 An optimality result

There are a number of ways one might measure the efficiency of our inference procedures. Here we consider the question, "How efficiently do these procedures eliminate wrong theories?" This measure seems especially appropriate since all of these inference procedures have the qualitative behavior that early on they are busy refuting lots of wrong theories. It turns out that all our procedures do this refuting of wrong theories well; we will show that all of our procedures perform within a constant factor of the optimum.

We begin by calculating the best possible refutation rate.

7.1 The optimal refutation rate

Assume that the right theory has index at least r. Define $f(c, d, r)$ to be the expected cost of refuting $\varphi_0, \varphi_1, \ldots, \varphi_{r-1}$.

Theorem: *For any inference procedure,* $f(c, d, r) \geq 2cr + d\Theta(\log r)$.

Proof: To refute φ_i we must keep on computing values of φ_{ij} until we get one where $\varphi_{ij} = 0$ and $\chi_j = 1$ or vice versa. Given that φ_i is not the right theory, we expect we will on average have to try two φ_{ij} until we get one that is refuted by χ. Hence our expected computation cost for eliminating r theories must be at least $2cr$.

Now for the cost of doing experiments. Since for wrong theories the φ_{ij} are all independent, we might as well reuse the same experimental χ_j's in refuting each φ_i. However, we have r such φ_i's to refute. What is the expected maximum number of agreements between any φ_i and χ over all r φ_i's? Equivalently, if we play a game where we toss a coin until we've seen a total of r heads, what is the expected length of the longest consecutive run of tails? We will show that the answer is $\Theta(\log r)$.

More formally, let X_i be the number of experiments required to refute (wrong) φ_i; it is easy to check that $\Pr[X_i = j] = 2^{-j}$ for $j = 1, 2, \ldots$. Let $X = \max_{i=1}^{r} X_i$. We want to show $\mathrm{E}[X] = \Theta(\log r)$.

$$\begin{aligned}
\mathrm{E}[X] &= \sum_{k=1}^{\infty} k \Pr[X = k] \\
&= \sum_{k=1}^{\infty} k(\Pr[X \geq k] - \Pr[X \geq k+1]) \\
&= \sum_{k=1}^{\infty} \Pr[\exists i : X_i \geq k] \\
&\leq \sum_{k=1}^{\lfloor \log r \rfloor} \Pr[\exists i : X_i \geq k] + \sum_{k=\lfloor \log r \rfloor}^{\infty} r 2^{-k+1} \\
&\leq \log r + 1. \qquad (10)
\end{aligned}$$

In the other direction we have

$$\mathrm{E}[X] = \sum_{k=1}^{\infty} \Pr[\exists i : X_i \geq k]$$

$$
\begin{aligned}
&= \sum_{k=1}^{\infty} 1-(1-2^{-k+1})^r \\
&\geq \sum_{k=1}^{\lfloor .5 \log r \rfloor} 1-(1-2/\sqrt{r})^r \\
&\geq (1-(1-2/\sqrt{r})^r)\frac{1}{2}\log r \\
&\sim (1-e^{-2\sqrt{r}})\frac{1}{2}\log r. \qquad (11)
\end{aligned}
$$

7.2 How our procedures compare to the optimum

The three inference procedures we discussed in the preceding three sections all perform within a constant factor of the optimum in refuting wrong theories.

None of them ever actually does an experiment when there are known experimental values against which the best theory has not yet been tested. Thus, until the right theory has become φ_0, we never do any more experiments than the optimum theory refutation strategy.

We do sometimes perform more computations than the optimum theory refutation strategy. In particular, we sometimes perform "wasted" computations as part of a crucial two way experiment. In such an experiment we might compute φ_{0j} and φ_{1j} for some j and find them to be equal. By the definition of a crucial experiment, we will refute one of those two theories before ever doing experiment χ_j; hence one of those computations was "wasted." However, we only perform crucial experiments when we're going to do an experiment, and we only do $O(\log r)$ experiments, so we only miss the optimum of $2cr$ by $cO(\log r)$.

8 Conclusions

We have introduced a new model for the process of inductive inference, which

1. is relatively simple, yet
2. captures a number of the qualitative characteristics of "real" science,
3. provides a crisp model for evolving or dynamic subjective probabilities, and
4. demonstrates that crucial experiments are of interest for *any* relative cost of experiments and making predictions.

References

[1] Dana Angluin and Carl H. Smith. Inductive inference: theory and methods. *Computing Surveys*, 15(3):237–269, September 1983.

[2] Lenore Blum and Manuel Blum. Toward a mathematical theory of inductive inference. *Information and Control*, 28(2):125–155, June 1975.

[3] P. K. Feyerabend. *Philosophical Papers: Realism, Rationalism, & Scientific Method.* Volume 1, Cambridge University Press, 1981.

[4] E. Mark Gold. Language identification in the limit. *Information and Control*, 10:447–474, 1967.

[5] I. J. Good. Kinds of probability. *Science*, 129(3347):443–447, February 1959.

[6] I. J. Good. The probabilistic explication of information, evidence, surprise, causality, explanation, and utility. In V. P. Godame and D. A. Sprott, editors, *Foundations of Statistical Inference*, pages 108–141, Holt, Reinhart, and Winston, 1971.

[7] Peter Kugel. Induction, pure and simple. *Information and Control*, 35:276–336, 1977.

[8] Jorma Rissanen. A universal prior for integers and estimation by minimum description length. *The Annals of Statistics*, 11(2):416–431, 1983.

[9] A. D. Wyner. An upper bound on the entropy series. *Information and Control*, 20:176–181, 1972.