On Choosing between Experimenting and Thinking when Learning

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

Robert Sloan†
EECS Department
University of Illinois
Chicago, Illinois 60680

Abstract

We introduce a model of inductive inference, or learning, that extends the conventional Bayesian approach by explicitly considering the computational cost of formulating predictions to be tested. We view the learner as a scientist who must divide her time between doing experiments and deducing predictions from promising theories, and we wish to know how she can do so most effectively. We explore several approaches based on the cost of making a prediction relative to the cost of performing an experiment. The resulting strategies share many qualitative characteristics with "real" science.

This model is significant for the following reasons.

- It allows us to study how a scientist might go about acquiring knowledge in a world where (as in real life) both performing experiments and making predictions from theories require time and effort.
- It lays the foundation for a rigorous machine-implementable notion of "subjective probability". Good (1959) argues persuasively that subjective probability is at the heart of probability theory. Previous treatments of subjective probability do not handle the complication that the learner's subjective probabilities may change as the result of pure thinking; our model captures this and other effects in a realistic manner. In addition, we begin to answer 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.

In: *Information and Computation*, Vol. 106, No. 1, September 1993, pp. 1–25

*This paper was prepared with support from NSF grant DCR-8607494, NSF grant CCR-8914428, ARO Grant DAAL03-86-K-0171, and the Siemens Corporation.

†Supported by an NSF graduate fellowship.

1 Introduction

We examine "inductive inference"—the process of drawing inferences from data. Angluin and Smith (1983) provide an excellent survey of previous work in the field. Our work is distinguished by the following features:

- Our inference procedure begins with prior probabilities for each possible theory, and updates these probabilities in a Bayesian manner as evidence is gathered.
- Our inference procedure may gather evidence in two ways, each of which has a cost (in terms of time taken):
  1. Use a theory to predict the result of a particular experiment.
  2. Run an experiment.
- Our inference procedure attempts to maximize the expected "rate of return" measured, for example, in terms of the total probability of theories eliminated per unit time.

Osherson, Stob, and Weinstein (1988) have examined the effectiveness of Bayesian approaches within a standard (recursion-theoretic) model of inductive inference. Their approach differs from ours, however, in that they define an efficient computation so as to permit any recursive function.

Our approach addresses the following three issues, which have not always been well handled by previous models.

(1) Induction is fundamentally different from deduction. Much previous work of a recursion-theoretic character has tried to cast induction into the same mold as deduction: given some data (premises) to infer the correct theory (conclusion). We feel that this approach is philosophically wrong, since experimental data can only eliminate theories, not prove them. (See Feyerabend (1981) and Kugel (1977).) We therefore prefer to study inference procedures that represent the *set* of remaining theories (and perhaps their probabilities), rather than inference procedures which are constrained to return a *single* answer. The "version space" approach suggested by Mitchell (1977) is more consistent with our point of view.

(2) The difficulty of making predictions is overemphasized. Previous theoretical work in this area has been largely recursion-theoretic, 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 are 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), $c > 0$, 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 previous work has assumed that the data (i.e., the list of all possible experimental results) is presented to the learner in some order (cf. Gold (1967), Blum and Blum (1975)). 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 her time most effectively—should she next run some experiment (if so, which one?), or should she 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 allows 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.

In this paper, each experiment takes a constant amount d of time to run. Again, this is also an oversimplification, but it allows us to explore the relevant issues without undue technical complication.

Our model can also be viewed as a contribution to the theory of subjective probability Good (1959) , 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" Good (1971) have been proposed, but these do not deal with the "thinking" aspect in a clean way.

2 The Model

We begin with a curious learner, or scientist, whom we'll call "Alice" for convenience. Alice wishes to understand some well-defined domain by running experiments and theorizing about the results she observes. At each point in time she face the "fundamental dilemma": is it better now to run an experiment, or to theorize some more?

2.1 Experiments

We assume that the domain is completely characterized by an infinite set of possible experiments, and that Alice can form a list

$$E_0, E_1, E_2, \ldots$$

of all possible experiments. If the plausibility of making such an enumeration seems dubious, we note that it is equivalent to the assumption that Alice has a technical vocabulary adequate for describing every conceivable experiment in the domain in a finite manner.

Performing experiment E_j yields a result, which we denote χ_j. We assume that each experiment is deterministic and well-defined, so that rerunning experiment E_j always yields result χ_j. We leave it as an open problem to generalize our results for experiments with probabilistically determined results.

For convenience we assume that $\chi_j \in \{0, 1\}$; each experiment is defined in such a way that it either succeeds, yielding $\chi_j = 1$, or fails, yielding $\chi_j = 0$. This assumption is made with little loss of generality, since an experiment yielding a nonbinary result can be viewed as a set of binary-valued experiments, each of which returns one bit of the result.

2.2 Theories

Alice's goal is to understand the domain perfectly, in the sense that she "knows" what the result of every conceivable experiment is. In the limit, of course, she can perform every conceivable experiment E_j and thus find out what each result χ_j is. She can do better, however, if she discovers some pattern or regularity to the results of the experiments, in which case she may, be able to correctly predict the results of all possible experiments after having performed only finitely many experiments. She can describe such patterns or regularities using a "theory."

We assume Alice has available an infinite list of theories about the domain; we let φ_i denote the ith possible theory, for $i \geq 0$. Each theory is a function from the natural numbers $\{0, 1, 2, \ldots\}$ into $\{0, 1\}$; the value $\varphi_i(j) = \varphi_{ij}$ is the "prediction" theory φ_i makes about the result of experiment E_j.

We do not allow "partial" theories here; each theory makes a prediction for each possible experiment. That is, each theory is a total function. While accepting partial functions as admissible theories would be an interesting direction to pursue, we feel that it would also greatly complicate our model and significantly distract attention from the issues we wish to focus on.

Alice hopes that one of her theories, say φ_r, is *correct* in the sense that $\varphi_{rj} = \chi_j$ for all j. If this is the case, and if she can determine r, then she will be able to correctly predict the result of any proposed experiment. If none of the theories is correct, then Alice's theorizing is in vain, and she has no way to converge upon an accurate means of predicting the results of new experiments. In this paper we assume that there is indeed a correct theory.

We assume that Alice's theories are ordered, in the sense that theory φ_i is judged to be "preferable," or "more likely", than theory $\varphi_{i'}$ if $i < i'$. Theory φ_0 is thus the one that Alice most prefers, or judges most likely.

2.3 Costs of thinking and experimentation

If there is a correct theory, Alice would like to identify it as efficiently as possible. This paper is primarily concerned with this issue of efficiency. We assume that the resources Alice uses can all be measured in common units.

The cost of thinking (cost of making a prediction) of doing (cost of running an experiment) need to be specified next. We assume that computing φ_{ij} from i and j always costs precisely c, independent of i and j, and that doing an experiment always costs precisely d, where $d > 0$, independent of which experiment is performed. Of course, these are not realistic assumptions, but they enable us to begin this study. It would be of interest to generalize these assumptions. We assume that other operations, such as planning, have zero cost. The reason for having separate parameters for the cost of predicting and the cost of experimentation is that Alice's decision as to whether it is better to predict or to experiment may depend upon the relative costs of these two operations.

2.4 Alice's State of Knowledge

Alice begins in a state of total ignorance, and proceeds to enlighten herself by taking steps consisting of either doing an experiment (determining some χ_j) or making a prediction (computing some φ_{ij}). Alice may choose which experiments and predictions she 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 Alice's state of knowledge at time t (after t steps have been taken).

- Let "**U**" denote "unknown".
- Let $\varphi_{ij}^t \in \{0, 1, \mathbf{U}\}$ denote Alice's knowledge of φ_{ij} at time t.
- Let $\chi_j^t \in \{0, 1, \mathbf{U}\}$ denote Alice's knowledge of χ_j at time t.

If at time t both χ_j and φ_{ij} are known, then either $\varphi_{ij} \neq \chi_j$, in which case theory φ_i has been *refuted*, or else $\varphi_{ij} = \chi_j$, in which case theory φ_i has been (partially) *confirmed.*

When a theory is refuted, we assume that it is removed from the list of available theories, and the remaining theories are renumbered, maintaining their relative order. By this convention, theories φ_0 and φ_1 are always Alice's "best" two theories: of the theories that have not yet been refuted, they are the two most preferred.

Subjective Probabilities

We assume in most of our development that Alice uses subjective probabilities to guide her choice of what to do next. Thus, Alice begins with two kinds of initial (or prior) subjective probabilities:

- The prior probability p_i that theory 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 $p_0 \geq p_1 \geq \ldots$ (the theories are listed in nonincreasing order of probability).
- The prior probability that $\varphi_{ij} = 1$, for any i and j. We assume that

  $$\Pr\{\varphi_{ij} = 0\} = \Pr\{\varphi_{ij} = 1\} = \frac{1}{2}$$

  initially, for all i and j; Alice has no reason to expect her theory to predict one way or the other, until she actually does the computation.

These probabilities are Alice's subjective prior probabilities, and are thus somewhat arbitrary. The p_i's do not necessarily represent any specific information Alice has about the theories; they may merely represent Alice's bias in favor of simpler theories.

Similarly, the prior belief that a theory is as likely to predict 0 as 1 merely represents Alice's ignorance of the prediction before she does the computation required

		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 Alice's State of Knowledge. Unknown entries are shown as blank.

to find out what the theory predicts. It is a subjective probability only, and doesn't preclude, for example, having theories which always predict 0 or which are highly structured predictions. Since "truth" is just one of the theories, nothing in the above assumptions precludes the true theory from making highly structured predictions. The 50–50 nature of Alice's prior beliefs about the predictions a theory might make are entirely reflective of her ignorance on the question before thinking about it, and not reflective of any necessary intrinsic randomness or unpredictability of the theories or of the true theory.

We note here for future reference 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 $\frac{1}{2} \geq p_1 \geq p_2 \geq \ldots$.

2.5 An Example

Consider Table 1, which illustrates a portion of Alice'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 she has run. Here she knows only $\chi_0 \ldots \chi_4$. The second column gives her current probabilities p_i.

The second part shows what predictions she 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, according to our convention. In this example, Alice has found out what her most probable theory predicts for experiments 0–5, and so on.

The table illustrates the convention of eliminating theories that have been found to be inconsistent with the data and renumbering the remaining theories; note that no theories are listed that are inconsistent with the known experimental results.

Running experiment 5 next has the potential of refuting φ_0. (It either refutes φ_0 or φ_3.) Making the prediction $\varphi_{1,5}$ can not (immediately) refute φ_1, but would affect

Alice's estimate of the likelihood that $\chi_5 = 0$. With the current state of knowledge, Alice would estimate that

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

We note that not only does making predictions affect Alice's subjective estimates of the likelihood of experimental outcomes, but also experimental outcomes can affect Alice's subjective estimates of the likelihood of a theory making a particular prediction. For example, if Alice has just discovered that $\chi_7 = 1$, and her current best theory has not yet predicted anything for this experiment (that is, $\varphi^t_{0,7} = \mathbf{U}$), then $\Pr(\varphi_{0,j} = 1)$ is at least p_0; indeed, this probability is now equal to $p_0 + (1-p_0)/2$. We note this effect as an observation only; our development does not require or depend on this effect.

We refer the reader to a fascinating article by Peebles and Silk (1990) on various theories for the origin of galaxies and the large-scale structure of the universe; this article contains a chart comparing the five best theories with 38 critical observables (experimental results) that is quite similar to our chart. In their case, however, experimental results are seldom sufficient to completely refute a theory, but they do estimate a measure of the fit of the theory to the observed result.

3 A global strategy

We begin our study with a "global" strategy, whose goal is to minimize the total cost incurred in eliminating all r incorrect theories that precede the correct theory φ_r in Alice's list of theories.

3.1 How Long Does Science Take?

Obviously, after a finite number of steps, Alice can refute only a finite number of theories, so at no point can she be certain to have discovered the "truth".

More realistically, she may ask "How long should it take me to eliminate all theories with higher prior probability than the correct theory?" Her answer depends on her set of prior probabilities. For example, she might have a "non-informative prior" that attempts to have p_i decrease to zero as slowly as possible, as in

$$p_i = \alpha \cdot (i \ln(i) \ln\ln(i) \ldots)^{-1} , \tag{1}$$

where α is a normalizing constant and only the positive terms in the series of logarithms are included (see Rissanen (1983)).

Since at least one step is required to eliminate a theory, the expected number of steps required to eliminate all theories before the correct one is at least the expected number of such incorrect theories,

$$\sum_{r \geq 0} r \cdot p_r ,$$

which is infinite for the prior (1) and for any distribution that decreases slowly enough. Thus, for a typical set of initial probabilities, Alice expects to have an infinite amount of work to do before the true theory is even considered!

3.2 Cost of eliminating r theories

In order to overcome the anomaly of the previous paragraphs, we now discuss how much work must be done to eliminate the first r incorrect theories, *as a function of* r. We begin by calculating a lower bound on this quantity. Let $f(c,d,r)$ denote the expected cost of refuting $\varphi_0, \varphi_1, \ldots, \varphi_{r-1}$. Note that this "expectation" is based on Alice's subjective probabilities, and does not correspond to the "true" cost of eliminating the first r theories, which depends on the exact nature of the theories and of the true state of the world.

We define the "standard (global) strategy" as follows. The strategy works exclusively on the first unrefuted theory until it is refuted. When the correct theory becomes the first unrefuted theory, the strategy thus enters an infinite loop attempting to refute it. For a given theory, the standard strategy first checks the predictions the theory makes against known experimental results. Only if the theory agrees with all new experimental results does the standard strategy run new prediction/experiment pairs in an attempt to refute the theory.

Theorem 1 *For any inference procedure,*

$$f(c,d,r) \geq 2cr + d\,\Omega\left(\lg r\right).$$

Proof : We argue that the standard strategy minimizes the expected cost of eliminating the first r theories, to within constant factors.

To refute φ_i Alice must compute values φ_{ij} for different values j until $\varphi_{ij} = 0$ and $\chi_j = 1$ or vice versa. If φ_i is not the right theory, she expects to have to make two predictions until she finds a prediction φ_{ij} that is contradicted by an experiment. Hence her expected computation cost for eliminating r theories is at least $2cr$.

Now for the cost of doing experiments. Since for wrong theories the φ_{ij} are all assumed to be independent by Alice, she has no reason not to reuse the same experimental χ_j's in refuting each φ_i. Thus, she might as well follow the standard strategy. There are r theories to refute. The expected number of experiments necessary is equal to the expected maximum number of agreements between any φ_i and χ over all r theories φ_i for $i = 0, 1, \ldots, r-1$. 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? (Here a coin flip corresponds to making a prediction, and a "head" corresponds to a refutation.) We show that the answer is $\Omega\left(\lg r\right)$.

More formally, let X_i be the number of experiments required to refute (wrong) theory φ_i; it is easy to check that $\Pr\left\{X_i = j\right\} = 2^{-j}$ for $j = 1, 2, \ldots$. Let $X = \max_{i=0}^{r-1} X_i$. We want to show $\mathrm{E}\left[X\right] = \Omega\left(\lg r\right)$; that is, the expected number of experiments run to eliminate all the theories preceding φ_r is $\Omega(\lg r)$. We have

$$\begin{aligned} \mathrm{E}\left[X\right] &= \sum_{k \geq 1} k \Pr\left\{X = k\right\} \\ &= \sum_{k \geq 1} \Pr\left\{\exists i : X_i \geq k\right\} \end{aligned}$$

$$
\begin{aligned}
&= \sum_{k\geq 1}\left(1-(1-2^{-k+1})^r\right) \\
&\geq \sum_{k=1}^{\lfloor(\lg r)/2\rfloor}(1-(1-2/\sqrt{r})^r) \\
&\geq (1-(1-2/\sqrt{r})^r)\lfloor(\lg r)/2\rfloor \\
&= \Omega\left(\left(1-e^{-2\sqrt{r}}\right)\lg r\right) \\
&= \Omega(\lg r)\ .
\end{aligned}
$$

Therefore the expected number of experiments run in order to eliminate all theories preceding φ_r is $\Omega(\lg r)$. (We note for the record that a corresponding upper bound can also be shown, so that $\mathrm{E}[X] = \Theta(\lg r)$.) ∎

The standard global strategy has been analyzed with respect to Alice's assumed prior distribution for the values φ_{ij} predicted by the theories and the values χ_j resulting from the experiments. In truth, however, these assumptions may be wildly wrong.

4 "Local" (or "Greedy") Strategies

In many situations, a "local" or "greedy" strategy can perform as well or better than a global strategy. For example, there are greedy strategies for finding minimum-cost spanning trees that always find the optimal solution. One of the motivations for the research reported in this paper is to begin a study of how such local, or greedy, strategies perform in the inference task introduced in this paper.

The local strategies we will study all have a common structure: at each step they choose an action that maximizes some measure of "rate of progress", taking into account the cost of the action chosen. The notion of "rate of progress" can be defined (in several ways) in terms of Alice's subjective probabilities; the rate of progress is intended to measure her increase in certainty of knowing what the correct theory is. We shall explore several different notions of "rate of progress."

To take in account the cost of taking actions, we can consider the expected rate of progress per unit cost. This is a natural way of defining a "greedy" strategy.

However, we note that maximizing the expected ratio of gain to cost can have somewhat unusual properties, if the actions taking have costs that may vary. For example, suppose Alice has two options, option A and option B. Option A always achieves a gain of 49 and a cost of 10, and thus has an expected ratio of gain to cost of 4.9. Option B always achieves a gain of 24, but has a cost of 4 with probability 1/2 and a cost of 6 with probability 1/2. Option B thus has an expected ratio of gain to cost of 5.0, and thus looks preferable to option A. But executing option B a large number k times in succession yields a total gain of $24k$ and a total cost of almost exactly $5k$ (by the central limit theorem), so re-executing option B has an effective ratio of gain to cost of only 4.8; option A is thus preferable to option B as a way to make progress if these options are to be executed repeatedly. In spite of this anomaly, we shall continue to explore local strategies that attempt to maximize the expected rate of progress per unit cost.

5 Alice's Options

We find it convenient to organize Alice's actions into "options". She organizes her 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. Each option is capable of refuting some theory. Her menu has the following four options:

- *Prediction/Experiment Pair:* Make a prediction φ_{0j} for the least j for which no predictions yet exist, and then run the corresponding experiment. Alice is not compelled to restrict the prediction/experiment pairs to using the most probable theory, φ_0, but we make this restriction because it is convenient to limit her options, and also because the expected return from other theories is not as good.

  The cost of making a prediction/experiment pair is just $(c + d)$.

- *Prediction:* Compute a prediction φ_{ij}, given that the corresponding experiment determining χ_j has already been run. Here again it is clear that Alice should choose the least i possible, so as to maximize her rate of return.

  The cost of making a prediction is just c.

- *Simple Experiment:* Run experiment j, given that at least one prediction has been made for this experiment.

  The cost of running an experiment is just d.

- *Crucial Experiment:* Determine the least j such that the two most probable theories make differing predictions for χ_j. Then run experiment j.

  The cost of running a crucial experiment is $d + 2cx$, where x is the number of j's that are examined before finding one such that $\varphi_{0j} \neq \varphi_{1j}$.

Having given our menu of options, we can now make one simple definition. When we speak of *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.

6 How Alice Updates Her Subjective Probabilities

Alice will measure her "rate of progress" in terms of her subjective probabilities, which evolve as she runs experiments and makes predictions. To model the evolution of Alice's knowledge more carefully, we show how her subjective probabilities associated with the various theories change as a result of the steps she has taken, using Bayes' Rule.

What happens to the probabilities maintained by Alice 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; that is, if $\chi_j \neq \varphi_{ij}$.
   - Set p_i^t to $2p_i^{t-1}$ if φ_i has just been confirmed; that is, if $\chi_j = \varphi_{ij}$.
   - 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 judged initially to be equally likely for a prediction to be a 0 or a 1.

In the rest of this paper, superscripts t are generally dropped; we assume that Alice is, at a particular point of time t, deliberating on what she should do next. We also let q_i denote $1 - p_i$.

We note that if Alice just sits and "thinks" about an experiment (i.e., she just computes the predictions of various theories for this experiment), her subjective probability $\Pr\{\chi_j = 0\}$ *evolves*, since at time t

$$\Pr\{\chi_j = 0\} = \sum_{i:\varphi_{ij}^t = 0} p_i + \frac{1}{2} \sum_{i:\varphi_{ij}^t = \mathbf{U}} p_i. \tag{2}$$

It would also be reasonable to treat this probability as an interval, since one knows the upper and lower limits that it could evolve to. Taking this approach, we would represent the probability specified in equation (2) by the interval

$$\left[\sum_{i:\varphi_{ij}^t = 0} p_i \ , \sum_{i:\varphi_{ij}^t \in \{0, \mathbf{U}\}} p_i \right] .$$

7 Specific Greedy Strategies

The approach taken by Alice depends upon the relative costs of making predictions versus doing experiments, her initial probabilities for the theories, and how she wishes to "optimize" her rate of progress.

7.1 General Assumptions

At each step, Alice 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 least-numbered (i.e., most probable) such theory.

7.2 Optimization Criteria

Alice chooses what actions to take according to some optimization criteria. For example, she may choose one of the following strategies:

1. `Refute-most-weight`:[1] Maximize the expected total probability of the theories refuted by the action chosen.

2. `Minimize-entropy`: Minimize the entropy
$$-\sum_{i\geq 0} p_i \lg(p_i) \ .$$

3. `Maximize-leader`: Maximize the highest probability assigned to any theory.

4. `Minimize-error`: Minimize the expected total probability assigned to *incorrect theories*,
$$\sum_{i\geq 0} p_i(1-p_i) \ .$$

More generally, we assume that she wishes to maximize her "rate of progress" by dividing her progress (measured by the change in one of the above criteria) by the cost of the action chosen.

There are, of course, other strategies that Alice might adopt. Some of these are not entirely rational. As an example, she might simply decide that her 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.[2]

As another example of an irrational strategy, Alice might try to keep the current best theory best. To accomplish this goal, Alice should never test φ_0 against any theory. She should simply test the other theories, making sure to stop testing φ_i as soon as $p_i \geq p_0/2$ (otherwise φ_i might replace φ_0 as best). This procedure is obviously uninteresting, since any strategy that refuses to put its best candidate theory to the test is not "science," but dogmatism.

[1] One stragegy which we do *not* consider is `Refute-most-theories`. It would be very different from all other stragies, because it would give us incentive to examine lower probability theories ahead of higher probability theories.

[2] If Alice tests the best theory with any kind of action, then with probability $p_0 + (1-p_0)/2$ it is confirmed, and its probability goes up to $2p_0/(1+p_0)$. However, with probability $(1-p_0)/2$ it is refuted and its probability goes to 0. Thus its expected probability after any action is $[(p_0+1)/2] \cdot 2p_0/(1+p_0) = p_0$. If Alice tests 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 assigned to φ_0 is p_0.

In this paper we discuss all of the "sensible" optimization criteria listed above; 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.

7.3 Menus of Options

We propose that Alice organize her strategy as a "greedy" strategy of the following form:

- At a given step, for each available option, Alice computes the expected "rate of return" of that option, defined as the expected ratio of the total gain of that option (where gain is measured by some optimization criterion) to the cost of that option.
- Alice 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.

8 The "`refute-most-weight`" greedy strategy

We begin by studying an inference procedure, "`refute-most-weight`", which tries to refute wrong theories as quickly as possible. Specifically, Alice chooses an action which maximizes the expected value of the ratio of the total probability of theories eliminated by that action to the cost of that action. The reason for this choice is its simplicity, and the ease with which Alice can implement such a strategy. Furthermore, if Alice's prior probability distribution happens to be one of the ones for which infinite expected time is required simply to eliminate all wrong theories (see section 3.1), then this measure probably makes the most sense.

We now analyze the expected rate of return for Alice's four options.

Prediction/Experiment Pair:

The expected "reward" for this action is p_0 times the probability that φ_0 will, in fact, be refuted. Theory φ_0 is never refuted if it is the true theory, and is refuted with probability $1/2$ otherwise; therefore the probability that φ_0 is refuted is $q_0/2$. Our expected rate of return is thus

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

Prediction:

Compute a prediction φ_{ij}, given that the corresponding experiment determining χ_j has already been run. The expected reward is the same as for the prediction/experiment pair, and the cost is simply c. Thus, the expected rate of return for this prediction is

$$\frac{p_i q_i}{2c} .$$

If we stick to prediction/experiment pairs and predictions, then the opportunity to make a prediction only arises after a simple prediction/experiment pair has already been run for that experiment.

Simple experiment:

The expected reward of a simple experiment is the same as for a prediction or a prediction/experiment pair, so the rate of return is

$$\frac{p_0 q_0}{2d} .$$

Crucial Experiment:

To calculate the expected reward of a crucial experiment, we must consider three cases.

1. If φ_0 is the true theory, then Alice refutes φ_1, for a reward of p_1. The probability that φ_0 is the true theory is p_0, so this case contributes $p_0 p_1$ to the total expected reward.

2. Similarly, the case where φ_1 is the true theory contributes $p_0 p_1$ to the total final reward.

3. If neither φ_0 nor φ_1 is the true theory, then it is equally likely that φ_0 or φ_1 is refuted. Since this case has probability $1 - p_0 - p_1$, its contribution to the expected reward is $(1 - p_0 - p_1)(p_0 + p_1)/2$.

Thus the expected reward is

$$\begin{aligned} 2p_0 p_1 + \frac{(1 - p_0 - p_1)(p_0 + p_1)}{2} &= \frac{p_0(1 - p_0) + p_1(1 - p_1) + 2p_0 p_1}{2} \\ &= \frac{p_0 + p_1 - (p_0 - p_1)^2}{2}. \end{aligned}$$

To compute the expected rate of return, we must multiply the expected reward by $\mathrm{E}[1/C]$, where C is the random variable denoting the cost of finding and running a crucial experiment. Note that $C = C(c, d)$ is a function of c and d. The expected cost of *finding* a crucial experiment is easily seen to be $4c$, since if Alice picks a j and compute φ_{0j} and φ_{1j}, she has a $1/2$ chance of finding j to be crucial. Thus

$$\mathrm{E}[C] = 4c + d .$$

In general, however, it is *not* necessarily true that $\mathrm{E}[1/C] = 1/\mathrm{E}[C]$, unless C is a constant.[3] This fact is relevant for analyzing crucial experiments, because the cost

[3] An earlier version of this paper (Rivest and Sloan (1988)) contained an error here based on the assumption that $\mathrm{E}[1/C] = 1/\mathrm{E}[C]$.

is not constant. We have $C = 2ck + d$, if k pairs of values $\varphi_{0j}, \varphi_{1j}$ are computed before Alice finds a j such that $\varphi_{0j} \neq \varphi_{1j}$. The probability that k pairs need to be computed is 2^{-k}, since there is a 1/2 chance for each j that $\varphi_{0j} \neq \varphi_{1j}$. Thus,

$$\mathrm{E}\,[1/C] = \sum_{k \geq 1} 2^{-k} \frac{1}{2ck + d} \ .$$

This formula is nearly impossible to evaluate in any simple closed form (see Gonnet (1984), equation (II.1.11), for a starting point). It is, however, easy to approximately evaluate this formula in practice for any given values of c and d, because of the exponentially decreasing powers of 2. We can also derive the bounds:

$$\frac{1}{4c + d} \leq \mathrm{E}\,[1/C] \leq \frac{1}{2.885c + d} \tag{3}$$

where the constant $2.885\ldots = 1/\ln 2$. The lower bound follows from the inequality $\mathrm{E}\,[1/C] \geq 1/\mathrm{E}\,[C]$, since $C > 0$ and the function $1/C$ is convex. The formula for the upper bound is derived from an exact analysis of the case $d = 0$ and numerical experimentation; it is actually a good approximation, correct to within 7% for all positive values of c and d. We omit the justification here.

Notice that $\mathrm{E}\,[1/C]$ remains the same if we condition it by any one of the following three events: theory φ_0 is correct, theory φ_1 is correct, or neither φ_0 nor φ_1 is correct. Therefore, the expected rate of return is

$$\frac{p_0 + p_1 - (p_0 - p_1)^2}{2} \ \mathrm{E}\,[1/C] \ . \tag{4}$$

We note that using all four options, 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 Alice 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 eliminates one of the top two theories, Alice is 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.

We claim that, using any of the above options, the *relative* order of two theories does not change, except when a theory is refuted, if an optimal greedy strategy is used. This follows since it is always preferable to work with the more probable theories, given a particular option, and this work tends to enhance the probability of that theory if it is not refuted.

8.1 Analysis of the "`Refute-Most-Weight`" strategy

If Alice is has a choice between starting with a prediction/experiment pair or making a prediction clearly Alice begins with a prediction/experiment pair. After that, Alice oscillates between further testing of her best theory (using prediction/experiment pairs), and testing of her other theories (using predictions).

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

If it is more expensive to perform an experiment than to compute a theory's prediction, then Alice should consider whether she should get her experimental data from crucial experiments rather than from prediction/experiment pairs.

Let's consider whether at the beginning of time, Alice is better off running a prediction/experiment pair or running a crucial experiment. The crucial experiment has a higher expected rate of return if

$$\frac{p_0 + p_1 - (p_0 - p_1)^2}{2} \mathrm{E}\left[1/C\right] > \frac{p_0 q_0}{2(c+d)} \tag{5}$$

or, substituting the lower bound from equation (3), if

$$\frac{c+d}{3c} > \frac{p_0 q_0}{p_1 q_1 + 2 p_0 p_1}.$$

Since the right-hand side of this inequality is $1/3$ if $p_0 = p_1 = 1/2$, it is possible to have a crucial experiment be advantageous over a prediction/experiment pair for any values of c and d.

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 is crucial, before running any experiments.

9 The "`Minimize-entropy`" greedy strategy

The (binary) entropy $H(P)$ of a probability distribution P is defined

$$H(P) = \sum_{i \geq 1} -p_i \lg p_i \ , \tag{6}$$

where $\lg x$ denotes the binary logarithm of x. The quantity $H(P)$ is considered to be a good measure of the information contained in probability distribution P. Maximizing entropy corresponds to maximizing uncertainty; minimizing entropy corresponds to minimizing uncertainty. Thus a reasonable optimization criterion for Alice would be minimizing the entropy of the a posteriori probability distribution.

Unfortunately, for some probability distributions, the entropy is infinite. Consider, for instance, the previously mentioned distribution due to Rissanen (1983),

$$p_i = \alpha \cdot (i \ln(i) \ln\ln(i) \ldots)^{-1} \ , \tag{7}$$

where α is a normalizing constant and only the positive terms in the series of logarithms are included. Wyner (1972) shows that the entropy series, equation (6), converges only if the series $\sum_{i \geq 1} p_i \ln i$ is convergent, but this series clearly diverges for the distribution given in equation (7).

However, any particular experiment or prediction Alice makes only causes her to alter a finite number of her a posteriori probabilities for theories, excluding the effect of renormalizing. It happens, as we shall see below, that even with renormalization, the expected *change* in entropy for any of the four options is finite.

The above discussion leads us to a precise description of the optimization criterion for our second inference procedure. Alice 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.

9.1 Analysis of the "`Minimize-entropy`" strategy

We need to calculate the expected change in entropy for each of Alice's four options. To begin with, we calculate the change in entropy caused by refuting φ_0. (The case for φ_i is similar, but the notation is simpler for $i = 0$.) Let P^{t_0} be the initial probability distribution , and let $\Delta(H(P))$ denote the change in entropy that occurs when φ_0 is refuted. Then

$$\begin{aligned}
\Delta(H(P)) &= -\sum_{i\geq 1} \frac{p_i}{1-p_0} \lg\left(\frac{p_i}{1-p_0}\right) + \sum_{i\geq 0} p_i \lg p_i \\
&= \frac{1}{1-p_0}\left[(1-p_0)\lg(1-p_0) - \sum_{i\geq 1} p_i \lg p_i\right] + \sum_{i\geq 0} p_i \lg p_i \\
&= \lg(1-p_0) + \frac{1}{1-p_0}\left(p_0 \lg p_0 + \sum_{i\geq 0} p_i \lg p_i\right) + \sum_{i\geq 0} p_i \lg p_i \\
&= \lg(1-p_0) + \frac{p_0 \lg p_0}{1-p_0} + \frac{p_0}{1-p_0} H(P^{t_0}). \qquad (8)
\end{aligned}$$

Now we calculate the change in entropy that occurs if φ_0 is instead partially confirmed. In the case we have

$$\begin{aligned}
\Delta(H(P)) &= -\frac{2p_0}{1+p_0} \lg\left(\frac{2p_0}{1+p_0}\right) - \sum_{i\geq 1} \frac{p_i}{1+p_0} \lg\left(\frac{p_i}{1+p_0}\right) + \sum_{i\geq 0} p_i \lg p_i \\
&= \frac{1}{1+p_0}\left[-2p_0 \lg\left(\frac{2p_0}{1+p_0}\right) + p_0 \lg p_0 - \sum_{i\geq 0} p_i \lg p_i\right] + \sum_{i\geq 0} p_i \lg p_i \\
&= \lg(1+p_0) - \frac{p_0}{1+p_0}\left(2 + \lg p_0 + H(P^{t_0})\right). \qquad (9)
\end{aligned}$$

Now we are ready to compute the expected change in entropy for each possible option.

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

$$\mathrm{E}\left[\Delta(H(P))\right] = -p_i + \frac{(1-p_i)}{2} \lg(1-p_i) + \frac{(1+p_i)}{2} \lg(1+p_i). \qquad (10)$$

  Equation (10) comes from taking $(1-p_i)/2$ (the probability that φ_i is refuted) times the quantity specified by equation (8) plus $(1+p_i)/2$ (the probability that φ_i is confirmed) times the quantity specified by equation (9).

- For running a crucial experiment between φ_0 and φ_1 we get

$$\mathrm{E}\left[\Delta(H(P))\right] = -p_0 - p_1 + \frac{(1+p_0-p_1)}{2}\lg(1+p_0-p_1) + \frac{(1-p_0+p_1)}{2}\lg(1-p_0+p_1). \tag{11}$$

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

$$\mathrm{E}\left[\Delta(H(P))\right] = -r_0 - r_1 + \frac{(1+r_0-r_1)}{2}\lg(1+r_0-r_1) + \frac{(1-r_0+r_1)}{2}\lg(1-r_0+r_1). \tag{12}$$

Consider the probability distribution, R, that has only two outcomes, one with probability $r_0 + (1 - r_0 - r_1)/2$, the other with probability $r_1 + (1 - r_0 - r_1)/2$. We can rewrite equation (12) in terms of the entropy of R,

$$\mathrm{E}\left[\Delta(H(P))\right] = -r_0 - r_1 + H(R). \tag{13}$$

Equations (10) and (11) can be rewritten in a similar manner (since really they're just special cases of equation (12)).

In fact, the calculations for this entropy driven inference procedure and the previous "`refute-most-weight`" strategy yield very similar results. Equation (13) and equation (3) could both be written as

$$\text{PROGRESS} = k(r_0 + r_1 - \text{penalty}(|r_0 - r_1|)). \tag{14}$$

(The difference in signs between equation (13) and equation (14) arises because in equation (13) Alice is trying to *minimize* entropy, so her progress is negative, and her penalty is positive.)

Let $\delta = |r_0 - r_1|$. For the entropy approach, $k = 1$ in equation (14), and $\text{penalty}(\delta) = H(1/2+\delta/2, 1/2-\delta/2)$. (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—the total probability of those theories i such that $\varphi_i(j) = \mathbf{U}$.) For the `refute-most-weight` strategy, $k = 1/2$ in equation (14), 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.

10 The "`Maximize-leader`" greedy strategy

As we pointed out in Section 7.2, there is no strategy which leads to an expected increase in the probability assigned to the current best theory. There is, however, at least one interesting way for Alice to always have a "pretty good" best theory. Alice chooses an action to maximize the quotient of the expected value of the probability assigned to the best theory not yet refuted after that action, and the cost of that action.

10.1 Analysis of the "Maximize-leader" strategy

The first thing we do is calculate the expected value of the probability assigned to the best theory for each of Alice's options.

- If Alice tests φ_0 (with any kind of action), then with probability $p_0+(1-p_0)/2$ it is confirmed, and the probability for the best theory becomes $2p_0/(1+p_0)$. With probability $.5(1-p_0)$, φ_0 is refuted, and the probability for the best theory becomes $p_1/(1-p_0)$. The expected value of the probability for the best theory is therefore $p_0+p_1/2$.

- If $p_i \leq p_0/2$ (so if even Alice tests and confirm φ_i it still has a lower a posteriori probability than φ_0), then testing φ_i does not lead to an increase in the expected value of the probability of the best theory.

- If $p_i > p_0/2$, and Alice tests φ_i, then the expected value of the probability 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 is more profitable to compute φ_{0j} than to compute φ_{ij}. Consider now the case where there is some j such that $\chi_j = \varphi_{0j}$ but $\varphi_{ij} = \mathbf{U}$. Whichever theory is now numbered zero began with an initial probability greater than or equal to the initial probability of the theory now numbered i. If at time t theory φ_0 has been confirmed more than φ_i, it must be that $p_0 \geq 2p_i$. Hence if φ_j is known, the only $\varphi_{i,j}$ for which it can be worth while to make a prediction is $\varphi_{0,j}$.

- If Alice runs a crucial experiment for the two best theories,[4] then the expected value of the probability of the best theory is

$$
\begin{aligned}
&\sum_{i=0}^{1} \Pr\left[\varphi_i \text{ is confirmed and } \varphi_{\bar{i}} \text{ is refuted}\right] (\text{weight of } \varphi_i \text{ after that}) \\
&= \left(p_0 + \frac{1-p_0-p_1}{2}\right)\frac{2p_0}{1+p_0-p_1} + \left(p_1 + \frac{1-p_0-p_1}{2}\right)\frac{2p_1}{1+p_0-p_1} \\
&= \left(\frac{1+p_0-p_1}{2}\right)\frac{2p_0}{1+p_0-p_1} + \left(\frac{1-p_0+p_1}{2}\right)\frac{2p_0}{1-p_0+p_1} \\
&= p_0+p_1.
\end{aligned}
$$

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

- A simple pair with the best theory: $(p_0+p_1/2)/(c+d)$.

- Prediction for φ_{0j} if χ_j known: $(p_0+p_1/2)/c$.

- Simple experiment χ_j where φ_{0j} is known: $(p_0+p_1/2)/d$.

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

- Crucial experiment: $(p_0 + p_1)\mathrm{E}[1/C]$.
- Alice might consider running a crucial experiment when we have some leftover predictions (say from an earlier crucial experiment) for one of the two theories. If she has k such predictions, then the expected cost $\mathrm{E}[C]$, used in our lower bound for $\mathrm{E}[1/C]$, decreases from $d + 4c$ to $d + (3 - \sum_{i=1}^{k-1} 2^{-i})c$.

All Alice 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 is always best to compute φ_{0j}. It is 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.

11 The "Minimize-error" greedy strategy

We have

$$\mathrm{E}[\text{weight of wrong theories}] = \sum_{i \geq 0} p_i(1 - p_i) = 1 - \sum_{i \geq 0} p_i^2 .$$

Hence the goal is to maximize

$$\sum_{i \geq 0} p_i^2 \tag{15}$$

which looks similar to the entropy strategy, only nicer, since the sum of the squares is guaranteed to converge.

Unfortunately, however, when we look at any action that might refute a given theory, that infinite sum is in its expected value. For the entropy case, it dropped out.

In the sum of squares case we get:

If we test φ_j against a known experiment and it is confirmed, the change in the value given in equation (15) is

$$\frac{p_j}{(1 + p_j)^2} \left[3p_j - (p_j + 2) \sum_{i \geq 0} p_i^2 \right] . \tag{16}$$

If instead it is refuted, then the change is

$$\frac{p_j}{(1 - p_j)^2} \left[-p_j + (2 - p_j) \sum_{i \geq 0} p_i^2 \right] . \tag{17}$$

Taking $(1 + p_j)/2$ times (16) plus $(1 - p_j)/2$ gives us the expected change in (15) when we test theory j. This quantity is

$$\frac{p_j^2}{1 - p_j^2} \left(1 - 2p_j + \sum_{i \geq 0} p_i^2 \right) . \tag{18}$$

It seems likely that (18) is maximized for p_j as big as possible (i.e., $j = 0$), but to compare this action's expected return versus, say, a two way experiment, Alice must compute an infinite sum, which violates the spirit of our procedures. Hence, we do not consider this strategy further.

12 How the greedy strategies compare to the global strategy

The greedy strategies we discussed above all perform within a constant factor of the optimum in refuting wrong theories, in terms of the expected cost required to refute the first r theories. We argue this point as follows.

Using any of these strategies, Alice never 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, she never does any more experiments than the optimum theory refutation strategy.

Alice sometimes performs more computations than the optimum theory refutation strategy. In particular, she sometimes performs "wasted" computations as part of a crucial experiment. In such an experiment she might compute φ_{0j} and φ_{1j} for some j and find them to be equal. By the definition of a crucial experiment, she will refute one of those two theories before ever doing experiment χ_j; hence one of those computations was "wasted." She only perform crucial experiments, however, when she is going to do an experiment, and she only does $O(\lg r)$ experiments, so she only misses the optimum of $2cr$ by $O(c \lg r)$.

13 Conclusions

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

- is relatively simple, yet
- captures a number of the qualitative characteristics of "real" science,
- provides a crisp model for evolving or dynamic subjective probabilities, and
- 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] G. H. Gonnet. *Handbook of Algorithms and Data Structures*. Addison-Wesley, 1984.

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

[7] 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.

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

[9] Thomas M. Mitchell. Version spaces: A candidate elimination approach to rule learning. In *Proceedings IJCAI-77*, pages 305–310, Cambridge, Mass., August 1977. International Joint Committee for Artificial Intelligence.

[10] Daniel N. Osherson, Michael Stob, and Scott Weinstein. Mathematical/mechanical? learners pay a price for Bayesianism. *The Journal of Symbolic Logic*, 53(4):1245–1251, 1988.

[11] P. J. E. Peebles and Joseph Silk. A cosmic book of phenonmena. *Nature*, 346:233–239, July 1990.

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

[13] Ronald L. Rivest and Robert Sloan. A new model for inductive inference. In Moshe Vardi, editor, *Proceedings of the Second Conference on Theoretical Aspects of Reasoning about Knowledge*, pages 13–27. Morgan Kaufmann, March 1988.

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