

# FORMAL LEARNING THEORY IN CONTEXT

Daniel Osherson and Scott Weinstein

## INTRODUCTION

One version of the problem of induction is how to justify hypotheses in the face of data. Why advance hypothesis  $A$  rather than  $B$  — or in a probabilistic context, why attach greater probability to  $A$  than  $B$ ? If the data arrive as a stream of observations (distributed through time) then the problem is to justify the associated stream of hypotheses. Several perspectives on this problem have been developed including Bayesianism [Howson and Urbach, 1993] and belief-updating [Hansson, 1999]. These are broad families of approaches; the citations are meant just as portals.

Another approach is to attempt to justify the present choice of hypothesis by situating it in a strategy with good long-term prospects. Such is the idea behind *Formal Learning Theory*, which will be discussed in what follows. We'll see that it is naturally grouped with the somewhat older concept of a *confidence interval*.

## THE CHARACTER OF FORMAL LEARNING THEORY

Formal Learning Theory is a collection of theorems about games of the following character.

PLAYERS: You (the reader) and us (the authors).

GAME PIECES: The set  $\{0, 1, \dots\}$  of natural numbers, denoted  $\mathbb{N}$ .

YOUR GOAL: Guess which subset of  $\mathbb{N}$  we have in mind.

OUR GOAL: Pick a subset of  $\mathbb{N}$  that you'll fail to guess.

RULES: First, all parties agree to a family  $\mathcal{C}$  of nonempty subsets of  $\mathbb{N}$  that are legal choices. We then pick nonempty  $S \in \mathcal{C}$  and an  $\omega$ -sequence  $e$  that orders  $S$ .<sup>1</sup> We reveal  $e$  to you one member at a time. At stage  $n$  of this process (that is, once you've seen  $e_0, e_1 \dots e_n$ ), you announce a guess  $T_n$  about the identity of  $S$ .

WHO WINS: If your guess  $T_n = S$  for cofinitely many  $n$  then you win.<sup>2</sup> Otherwise, we win.

---

<sup>1</sup>The  $\omega$ -sequence is just a total function from  $\mathbb{N}$  onto  $S$ . It includes every member of  $S$  and no more (repetitions allowed).

<sup>2</sup>In other words, you win just in case there are only finitely many  $n \in \mathbb{N}$  such that  $T_n \neq S$ .

Come on. Let's play. Take  $\mathcal{C}_1 = \{\mathbb{N} - \{x\} \mid x \in \mathbb{N}\}$ . This is the family of subsets of  $\mathbb{N}$  that are missing just one number, e.g., the set of positive integers. We've chosen  $S \in \mathcal{C}_1$  and also an  $\omega$ -ordering  $e$  on  $S$ . We'll now begin to reveal  $e$ . You must guess after each number.

$e_0 = 2$ . Go ahead and guess.

$e_1 = 0$ . Go ahead.

$e_2 = 1$ . Guess.

$e_3 = 4$ . Again.

The game never stops so we interrupt it in order to make an observation, namely: *There is a winning strategy available to you.* If at each stage you guess  $\mathbb{N} - \{x_0\}$ , where  $x_0$  is the least number not yet revealed then you are sure to win no matter which  $S \in \mathcal{C}_1$  and which  $\omega$ -sequence  $e$  for  $S$  we choose. This is easy to verify.

On the other hand, suppose that  $\mathbb{N}$  is added to the game so that our legal choices at the start are  $\mathcal{C}_2 = \{\mathbb{N}\} \cup \{\mathbb{N} - \{x\} \mid x \in \mathbb{N}\}$ . Then it can be demonstrated that there is no winning strategy available to you. To clarify this claim, let  $\Sigma$  be the set of finite sequences of natural numbers (conceived as potential data available at any finite stage of the game). Call a mapping of  $\Sigma$  into  $\mathcal{C}_2$  a *strategy*. For example, the strategy described above is the mapping from  $\sigma \in \Sigma$  to  $\mathbb{N} - \{x_0\}$  where  $x_0$  is the least number that does not appear in  $\sigma$ . This strategy is not a guaranteed winner for  $\mathcal{C}_2$  since it fails if we choose  $\mathbb{N}$  at the outset. You would then keep changing hypotheses forever, never settling down to  $\mathbb{N}$  on any  $\omega$ -sequence over  $\mathbb{N}$  we use to generate data. More generally, it can be shown [Jain *et al.*, 1999] that in our expanded game *no* strategy is a guaranteed winner; every strategy can be led to failure by some choice of  $S \in \mathcal{C}_2$  and some  $\omega$ -sequence  $e$  for  $S$ .

Our games look a little like scientific inquiry. Nature chooses a reality  $S$  from a collection  $\mathcal{C}$  that is constrained by established theory. The data are revealed to the scientist in some order  $e$ . Success consists in ultimately stabilizing on  $S$ . More realism comes from limiting members of  $\mathcal{C}$  to effectively enumerable subsets of  $\mathbb{N}$ , named via the programs that enumerate them. Scientists can then be interpreted as computable functions from data to such names. In the same spirit, data-acquisition may be rendered a less passive affair by allowing the scientist to query Nature about particular members of  $\mathbb{N}$ . Also, the success criterion can be relaxed or tightened, the data can be partially corrupted in various ways, the computational power of the scientist can be bounded, efficient inquiry can be required, scientists can be allowed to work in teams, and so forth for a great variety of paradigms that have been analyzed. In each case, at issue is the existence of strategies that guarantee success. [Jain *et al.*, 1999] offer a survey of results. Instead of subsets of  $\mathbb{N}$ , the objects of inquiry can be grammars over a formal language (as in [Gold, 1967]), relational structures [Martin and Osherson, 1998], or arbitrary collections of data-streams [Kelly, 1996]. These developments often appear under the rubric *Formal Learning Theory*.

The entire field stems from three remarkable papers. [Putnam, 1979] introduced the idea of a computable strategy for converting data into conjectures about a hidden recursive function (the data are increasing initial segments of the function's graph). He proved the non-existence of strategies that guarantee success, and contrasted his results with the ambitious goals for inductive logic announced in [Carnap, 1950]. For example, Putnam demonstrated that no recursive function  $T$  extrapolates every recursive, zero-one valued function.  $T$  extrapolates such a function  $b$  just in case for cofinitely many  $n$ ,  $T(b[n]) = b_{n+1}$ . Here,  $b[n]$  is the initial segment of  $b$  of length  $n$  and  $b_{n+1}$  is the next value. Putnam deploys a diagonal argument to establish this limitative result. In particular, given a recursive extrapolator  $T$ , define a recursive function  $b$  by course-of-values recursion as follows:

$$b_{n+1} = 1 - T(b[n]).$$

It is clear that  $T$  fails to correctly extrapolate the next value of  $b$  on any initial segment whatsoever. On the other hand, as Putnam observes, given any recursive extrapolator  $T$  and any recursive function  $b$ , there is a recursive extrapolator  $T'$  which correctly extrapolates every function  $T$  does and  $b$  as well. In particular, for every natural number  $n$  and every finite sequence  $s$  of length  $n$ , we let

$$T'(s) = \begin{cases} b_{n+1} & \text{if } s = b[n] \\ T(s) & \text{otherwise.} \end{cases}$$

It is clear that  $T'$  is recursive, and extrapolates  $b$  along with all the sequences  $T$  extrapolates.

The relevance of Formal Learning Theory to the acquisition of language by infants was revealed by [Gold, 1967]. Among other facts, Gold demonstrated that no strategy guarantees success in stabilizing to an arbitrarily chosen finite-state ("Type 3") grammar on the basis of a presentation of its strings. It follows immediately that the same is true for all levels of the Chomsky hierarchy [Chomsky, 1959] of increasingly inclusive formal grammars. In particular, Gold showed that no strategy is successful on a collection that includes an infinite language  $I = \{w_0, w_1, \dots\}$  and all of its finite subsets. For, suppose we are given a strategy  $L$  that succeeds on each of the finite languages  $J_n = \{w_0, \dots, w_n\}$ . We may then construct an enumeration  $e$  of  $I$  on which  $L$  fails to stabilize. Our construction proceeds by stages; at each stage  $n$ , we specify a finite initial segment  $s^n$  of  $e$ . We begin the construction by letting  $s^0$  be the empty sequence. At each stage  $n + 1$  we recursively specify a finite sequence  $s^{n+1}$  with the following properties:

- $s^{n+1}$  extends  $s^n$ ,
- every item in  $s^{n+1}$  is a member of  $J_n$ ,
- $w_n$  appears in  $s^{n+1}$ , and
- $L(s_{n+1})$  is a grammar for  $J_n$ .

Suppose we have constructed  $s^n$  and we are at stage  $n + 1$ . For  $m > 0$ , let  $r_m$  be the finite sequence obtained by extending  $s^n$  with a sequence of  $w_n$ 's of length  $m$ . Since  $L$  succeeds on  $J_n$ , there is  $m > 0$  such that  $L(r_m)$  is a grammar for  $J_n$ . We let  $s^{n+1}$  be  $r_m$  for the least such  $m$ . Now, if we let  $e_n = s_n^{n+1}$ , then it is not hard to verify that  $e$  is an enumeration of  $I$  on which  $L$  fails to stabilize.

Finally, [Blum and Blum, 1975] introduced new techniques to prove unexpected theorems about paradigms close to Putnam's and Gold's. The concepts presented in their work were subsequently mined by a variety of investigators. Among their discoveries is the surprising fact that there is a collection  $F$  of total recursive 0-1 valued functions such that a computable strategy can achieve Gold-style success on  $F$ , but no computable strategy can successfully extrapolate all functions in  $F$ .<sup>3</sup> An example of such a collection is the set of "self-describing" 0-1 valued total recursive functions. A total recursive function  $b$  is *self-describing* just in case the least  $n$  such that  $b(n) = 1$  is the index of a Turing machine that computes  $b$ . Gold-style success on a total recursive function  $b$  consists in stabilization to an index for (a Turing machine that computes)  $b$  when presented with an enumeration of the argument value pairs  $\langle 0, b(0) \rangle, \langle 1, b(1) \rangle, \dots$ . It is easy to see that there is a computable strategy guaranteeing Gold-style success on every self-describing total recursive function. It is a remarkable consequence of Kleene's Recursion Theorem<sup>4</sup> that for every total recursive 0-1 valued function  $a$ , there is a self-describing total recursive 0-1 valued function  $b$  such that  $a$  and  $b$  differ on only finitely many arguments. It follows at once that there is no recursively enumerable set of natural numbers  $X$  such that

- for every self-describing total recursive 0-1 valued function  $b$ , there is an  $n \in X$  such that  $n$  is an index for  $b$  and
- for every  $n \in X$ ,  $n$  is an index for a total recursive function.

On the other hand, [Blum and Blum, 1975] show that if  $F$  is a collection of total recursive functions and there is a computable strategy that successfully extrapolates every function in  $F$ , then there is a recursively enumerable set of natural numbers  $X$  such that

- for every  $b \in F$ , there is an  $n \in X$  such that  $n$  is an index for  $b$  and
- for every  $n \in X$ ,  $n$  is an index for a total recursive function.

It follows immediately that the self-describing functions can be learned Gold-style by a single recursive strategy but not extrapolated by any single recursive function.

Rather than present Formal Learning Theory in further detail, we rely on the examples given above to communicate its flavor. They illustrate a fundamental feature of virtually all paradigms embraced by the theory. Even when success can be guaranteed, *at no stage do the data imply the correctness of the latest*

<sup>3</sup>This result was also obtained independently by [Barzdin and Freivalds, 1972].

<sup>4</sup>See [Rogers, 1967] for a proof and applications of this theorem.

*hypothesis*. In the case of  $\mathcal{C}_1$ , for every data-set  $e_0 \cdots e_n$  and every hypothesis  $T$  that could be issued in response, there is  $T' \in \mathcal{C}_1$  distinct from  $T$  and  $\omega$ -sequence  $e$  for  $T'$  that begins with  $e_0 \cdots e_n$ . Intuitively, your current data don't exclude the possibility that the "hole" in  $\mathbb{N}$  occurs beyond the one cited in your current hypothesis. In this sense, Formal Learning Theory concerns methods for arriving at truth non-demonstratively, and thus belongs to Inductive Logic.

Because the data never imply the correctness of the current conjecture, a successful guessing strategy warrants confidence in the strategy but not in any hypothesis produced by it. Using the rule described earlier, for example, you are justified in expecting to stabilize to the correct member of  $\mathcal{C}_1$ . But Formal Learning Theory offers no warrant for ever suspecting that stabilization is underway. There might be *external reasons* for such a feeling, e.g., information about an upper bound on the "hole" that your opponent (in this case, the authors) is likely to choose. But information of this kind is foreign to Formal Learning Theory, which only offers various kinds of reliable methods for some games — along with proofs of the nonexistence of reliable methods for other games.

To underline the separation between long term performance and warrant for individual conjectures, consider the following guessing policy for  $\mathcal{C}_1$ .

- On the first datum, guess  $\mathbb{N} - \{0\}$ .
- Never change an hypothesis unless it is contradicted by your data.
- When contradiction arrives, if the number of times you've changed hypotheses is even, guess  $\mathbb{N} - \{x_0\}$  where  $x_0$  is the least number not yet encountered; otherwise, guess  $\mathbb{N} - \{x_1\}$  where  $x_1$  is the *second* least number not yet encountered.

The new and old guessing policies enjoy the same guarantee of correct stabilization in the  $\mathcal{C}_1$  game. But if the common guarantee provided warrant for the conjectures of one strategy, it would seem to provide equal warrant for the conjectures of the other, yet the conjectures will often be different! Indeed, for any potential conjecture at any stage of the game, there is a guessing strategy with guaranteed correct stabilization that issues the conjecture in question. They can't all be warranted.

The remaining discussion compares Formal Learning Theory to the statistical theory of *confidence intervals* initiated by [Neyman, 1937] before the advent of Formal Learning Theory. (See [Salsburg, 2002, Ch. 12] for the history of Neyman's idea.) To begin, we rehearse well-known arguments that confidence intervals offer global performance guarantees without provision for evaluating specific hypotheses — in much the sense just indicated for Formal Learning Theory. Then we'll attempt to situate both theories in the logic of *hypothesis acceptance*.

## CONFIDENCE INTERVALS

The theory of confidence intervals shares with Formal Learning Theory the goal of revealing a hidden reality on the basis of data that do not deductively imply the correct answer. Let us attempt to isolate the kind of performance guarantee associated with confidence intervals, and distinguish such guarantees from “confidence” about the specific reality behind one’s current data. We focus on a simple case, building on the discussion in [Baird, 1992, §10.5].

- (1) URN PROBLEM: Suppose that an urn is composed of balls numbered from 1 to  $L$  (no gaps, no repeats), with  $L \geq 2$ . The urn is sampled with replacement  $k \geq 2$  times. What can be inferred about  $L$  on this basis?

Let  $X_{L,k}$  be the set of possible samples of  $k$  balls that can be drawn from the urn with  $L$  balls. We think of such samples as ordered sequences. It is clear that  $X_{L,k}$  is finite, and that its members have uniform probability of appearing [namely  $(1/L) \exp k$ ]. Let  $f$  be a mapping of  $\bigcup\{X_{L,k} \mid L, k \geq 2\}$  into the set of intervals of the form  $[i, j]$ , where  $i, j$  are positive integers ( $i \leq j$ ). We think of  $f$  as attempting to construct an interval containing  $L$  on the basis of a sample. Success at this enterprise is quantified as follows.

- (2) DEFINITION: Let  $r \in (0, 1)$  be given. Call  $f$  *r-reliable* just in case for every  $L, k \geq 2$ , there are at least  $100 \times r\%$  of  $x \in X_{L,k}$  with  $L \in f(x)$ .

The definition embodies the kind of performance guarantee that we hope to associate with a given mapping  $f$  from data  $\bigcup\{X_{L,k} \mid L, k \geq 2\}$  into finite intervals (hypotheses about  $L$ ). Of course, for a given level of reliability, narrower intervals are more informative than wider ones.

*One method for constructing confidence intervals for  $L$* 

Let us fix  $r \in (0, 1)$  and consider a specific  $r$ -reliable function  $f_r$ . Let  $X_k = \bigcup_{2 \leq L} X_{L,k}$  (this is the set of all potential samples of size  $k$ ). For  $x \in X_k$ , we write  $\max(x)$  for the largest member of  $x$ . It is easy to verify:

- (3) FACT: Let  $L_0 \geq 2$  and  $1 \leq m \leq L_0$  be given. For all  $L > L_0$ , the proportion of samples  $y$  from  $X_{L,k}$  with  $\max(y) \leq m$  is less than or equal to  $(m/L_0)^k$ .

For  $x \in X_k$ , define  $f_r(x) = [\max(x), L_0]$  where:

- (4)  $L_0$  is the least integer greater than or equal to  $\max(x)$  such that for all  $L > L_0$ , the proportion of samples  $y$  from  $X_{L,k}$  with  $\max(y) \leq \max(x)$  is less than or equal to  $1 - r$ .

That such an  $L_0$  exists for each  $x \in X_k$  (and can be calculated from  $x$ ) follows from Fact (3). To show that  $f_r$  is  $r$ -reliable, let  $L, k \geq 2$  and  $x \in X_{L,k}$  be given. Then  $L \in f_r(x)$  iff  $L \in [\max(x), L_0]$  where  $L_0$  satisfies (4). Since  $L \geq \max(x)$ ,  $L \notin [\max(x), L_0]$  iff  $L > L_0$ , which implies that  $x \in A = \{y \in X_{L,k} \mid \max(y) \leq \max(x)\}$ . But by (4),  $\text{Prob}(A) \leq 1 - r$ .

- (5) EXAMPLE: Suppose  $r = .95$ , and let  $D$  symbolize the appearance of balls numbered 61 through 90 in thirty consecutive draws. Then,  $\max(x) = 90$ . To form a confidence interval using the .95-reliable function  $f_{.95}$  defined by (4), we seek the least  $L_0$  such that the probability of drawing thirty balls labeled 90 or less from an  $L_0$ -urn is no greater than 5%. By Fact (3),  $L_0$  is the least integer satisfying:

$$\left(\frac{90}{L_0}\right)^{30} < .05.$$

Calculation reveals that  $L_0 = 100$ . Hence  $f_{.95}(D) = [90, 100]$ .

### *Confidence and confidence intervals*

Parallel to our discussion of Formal Learning Theory, we now consider the relation between  $r$ -reliability and the warrant for particular intervals. Suppose the urn confronts us with  $D$  of Example (5). Does the fact that  $f_{.95}(D) = [90, 100]$  justify 95% *confidence* that  $L \in [90, 100]$ ? In other words, if  $Prob_{\text{subj}}$  represents a person's subjective assessment of chance, should the following hold?

$$(6) \quad Prob_{\text{subj}}(L \in f_{.95}(D) \mid D \text{ is the draw}) \geq .95.$$

Inasmuch as reliability in the sense of Definition (2) concerns fractions of  $X_{L,k}$  (the finite set of  $k$ -length data that can emerge from an  $L$ -urn) whereas (6) concerns justified belief about a particular urn and draw, it is not evident how the former impinges on the latter [Hacking, 2001]. Indeed, (6) seems impossible to defend in light of the existence of a different .95-reliable function  $h_{.95}$  with the property that

$$(7) \quad h_{.95}(D) \cap f_{.95}(D) = \emptyset.$$

We exhibit  $h_{.95}$  shortly. To see the relevance of (7), observe that whatever reason  $r$ -reliability provides for embracing (6) extends equally to:

$$(8) \quad Prob_{\text{subj}}(L \in h_{.95}(D) \mid D \text{ is the draw}) \geq .95.$$

If  $Prob_{\text{subj}}$  is coherent, (6) and (8) entail:<sup>5</sup>

$$Prob_{\text{subj}}(L \in f_{.95}(D) \wedge L \in h_{.95}(D) \mid D \text{ is the draw}) \geq .90.$$

But the latter judgment is incoherent in light of (7). At least one of (6), (8) must therefore be abandoned; by symmetry, it seems that both should.

---

<sup>5</sup>Here we use: for any two statements  $p$  and  $q$ , coherence entails that  $Prob_{\text{subj}}(p \wedge q) \geq Prob_{\text{subj}}(p) + Prob_{\text{subj}}(q) - 1$ . The proof is elementary.

*An alternative method for constructing confidence intervals for  $L$*

To specify  $h_r$ , we rely on the following well known facts [Ross, 1988], writing  $\bar{X}$  for the arithmetical mean of sample  $X$ .

- (a) For a uniform distribution over  $\{1 \cdots L\}$ , the mean  $\mu = (L + 1)/2$ , and the variance  $\sigma^2 = (L^2 - 1)/12$ .
- (b) (Chebyshev) For any sample  $X$  of size  $n$  drawn independently and identically from a distribution with mean  $\mu$  and variance  $\sigma^2$ ,

$$\text{Prob}(|\bar{X} - \mu| \geq k\sigma/\sqrt{n}) \leq k^{-2}, \text{ for all } k > 0.$$

It follows that if an independent sample  $X$  of size  $n$  is drawn from an urn with highest number ball  $L$  then:

$$\text{Prob}\left(\left|\bar{X} - \frac{L+1}{2}\right| \geq \frac{k\sqrt{L^2-1}}{\sqrt{12n}}\right) \leq k^{-2}, \text{ for all } k > 0$$

hence,

$$\text{Prob}\left(\left|\bar{X} - \frac{L+1}{2}\right| \geq \frac{kL}{\sqrt{12n}}\right) \leq k^{-2}, \text{ for all } k > 0,$$

so,

$$\text{Prob}\left(\left|\bar{X} - \frac{L+1}{2}\right| < \frac{kL}{\sqrt{12n}}\right) > 1 - k^{-2}, \text{ for all } k > 0.$$

Algebraic manipulation with  $k = \sqrt{1/.05}$  yields:

$$(9) \quad \text{Prob}\left(\frac{\sqrt{3n}(2\bar{X}-1)}{\sqrt{3n}+4.47} < L < \frac{\sqrt{3n}(2\bar{X}-1)}{\sqrt{3n}-4.47}\right) > .95.$$

Define  $h_{.95}(X)$  to be the interval specified by (9) (with integer values). Then,  $h_{.95}$  is .95-reliable. Setting  $X = D = [61, 90]$ , as in Example (5), we have  $h_{.95}(D) = [101, 284]$ , verifying (7) inasmuch as  $f_{.95}(D) = [90, 100]$ .

The method based on (9) allows us to articulate another argument against allowing confidence intervals to determine subjective probability. Even though  $h_{.95}$  is .95-reliable, calculation confirms the following fact.

- (10) Let  $E = 61 \cdots 90, 400$  (that is,  $E$  is  $D$  with 400 added to the end). Then  $h_{.95}(E) = [116, 319]$ . That is,  $h_{.95}(E)$  does not include the highest ball observed (400).

Thus,  $E$  visibly belongs to the small set of samples on which  $h_{.95}$  is inaccurate, and it would be folly to believe:

$$\text{Prob}_{\text{subj}}(L \in h_{.95}(E) \mid E \text{ is the draw}) \geq .95.$$

## COMPARISON

Consider the urn problem (1) from the perspective of Formal Learning Theory. Let  $S$  be the strategy of guessing  $L$  to be the highest numbered ball seen in the data so far. Then, if balls are drawn forever,  $S$  will stabilize to the correct conjecture with unit probability. How does  $S$  compare with the function  $f_{.95}$  defined earlier for confidence intervals?

$S$  appears to ignore useful information that is exploited by  $f_{.95}$ , namely, the independent and uniform character of the draws composing the current data set. But once a given draw is made, it is hard to see the relevance of this information to current belief. For fixed  $L$ , all samples of the same size have the same probability prior to the draw, and just one sample has all the probability afterwards (namely, the sample observed). The interval produced by  $f_{.95}$  is then either correct or incorrect, thus has probability either 1 or 0, just as for  $S$ 's conjecture. To equate subjective confidence in the interval's accuracy with the fraction of potential samples on which  $f_{.95}$  succeeds is to overlook the change in sampling distribution consequent to the draw; it was once uniform but now is concentrated on the observed data. (This point is elaborated in [Hacking, 2001]). Ignoring the transition when forging personal confidence leads to incoherent judgment, as seen in the previous section. But once the current data decouple belief about  $L$  from the performance guarantees of  $f_{.95}$  and  $S$ , it seems just as legitimate to issue conjectures based on one as the other.

Instead of attempting to interpret the merit of  $f_{.95}$  and  $S$  in terms of personal probabilities (or the utility of announcing their counsel), let us explore the idea that they both embody attractive policies for *hypothesis acceptance*. The distinction between acceptance and personal probabilities (belief) has been explained diversely by different authors [Popper, 1959; Cohen, 1992; Maher, 1993; Levi, 1967; Kaplan, 1998]. We rely here on just a few observations. Belief appears to be largely involuntary at any given moment. Although you can influence the beliefs that enter your mind over time (e.g., by reading certain newspapers rather than others), you cannot alter your current beliefs by a mere act of will. Acceptance, on the other hand, requires choice. Thus, it might be difficult to refrain from strong belief that a recently purchased lottery ticket will lose. But you haven't thereby accepted this thesis (if you did, you would throw the ticket away).<sup>6</sup>

Acceptance is assumed to be categorical inasmuch as it rests on a definite selection among alternative theories even if confidence in each is graded and the choice is provisional. Naturally, beliefs are often a factor in theory acceptance. But they are not always decisive since there may be other factors, such as the impression that one theory is more interesting than another or the suspicion that announcing a certain theory will promote inquiry leading to a better one.

We suggest that Formal Learning Theory helps to evaluate *policies for accepting*

---

<sup>6</sup>This kind of example is discussed more thoroughly in [Maher, 1993, §6.2.1]. A revealing analogy (offered in [Cohen, 1992, Ch. 2]) compares belief to (mere) desire, and acceptance to consent or acquiescence.

*hypotheses.* For example, it informs us that the strategy described earlier for  $\mathcal{C}_1$  stabilizes to the truth on every presentation of data whereas no strategy has this property for  $\mathcal{C}_2$ . More refined criteria allow comparison of the plurality of successful policies for  $\mathcal{C}_1$  — in terms of efficient use of data, for example, or the computational resources needed for implementation. Even so, Formal Learning Theory does not designate a uniquely best strategy, and it is cold comfort in defending any particular conjecture in response to data. The theory nevertheless seems relevant to the orderly adoption of hypotheses.

Neyman defended the theory of confidence intervals from the same perspective albeit without the “acceptance” terminology (see [Baird, 1992, §10.6] for illuminating discussion). Similarly to Formal Learning Theory, alternative strategies for constructing intervals can be shown to be reliable (as seen in the last section), so the theory is silent about the correct response to a particular sample. And as before, supplementary criteria may be evoked to compare different strategies, for example, concerning the width of constructed intervals or the possibility of issuing an interval contradicted by the available data [illustrated by (10)].

Both theories clarify the options of an agent who seeks method in her acceptance of hypotheses. In the case of the urn, it remains an extra-systematic choice whether to guess  $L$  exactly or just bracket it (or both in parallel). Likewise, no uniquely best choice emerges from the set of recommended strategies of each kind. But it may be hoped that new criteria will eventually come to light, allowing increasingly refined evaluation of policies for hypothesis selection. The logic of acceptance would be thereby extended beyond the contributions already due to confidence intervals and Formal Learning Theory.

## CONCLUSION

All this leaves untouched a fundamental question: Since we’re each destined to issue but finitely many hypotheses, why rely on a guessing strategy whose performance is guaranteed only in the limit? Formal Learning Theory does not resolve the matter, but perhaps it facilitates articulation of the issue, preparing the way for subsequent clarification.

Whatever its normative contribution may turn out to be, Formal Learning Theory also has a potential *descriptive* role in characterizing inductive practice. A given person (suitably idealized) implements a function from finite data sets to hypotheses about the data’s provenance. Characterizing the functions that people typically implement could provide insight into the scope and limits of human cognition by revealing the class of empirical problems that such a system can solve in principle. Focussing the issue on children could likewise yield fresh perspectives on development, for example, by delineating the collection of *natural languages* (i.e., systems of communication learnable by infants).<sup>7</sup>

---

<sup>7</sup>To illustrate, the infant is unlikely to remember at any given time more than a few of the utterances that she has encountered. This design feature can be shown to influence the collection

Formal Learning Theory serves as catalyst to this enterprise. It connects classes of inductive strategies to the empirical problems for which they are adapted. It thereby also suggests the kinds of problems that might pose insuperable obstacles to human inquiry.

## BIBLIOGRAPHY

- [Baird, 1992] D. Baird. *Inductive Logic: Probability and Statistics*. Prentice Hall, Englewood Cliffs NJ, 1992.
- [Barzdin and Freivalds, 1972] Ja. M. Barzdin and R. V. Freivalds. On the prediction of general recursive functions. *Soviet Math. Dokl.*, 13:1224–1228, 1972.
- [Blum and Blum, 1975] L. Blum and M. Blum. Toward a mathematical theory of inductive inference. *Information and Control*, 28:125–155, 1975.
- [Carnap, 1950] Rudolph Carnap. *The Logical Foundations of Probability*. University of Chicago Press, Chicago IL, 1950.
- [Chomsky, 1959] Noam Chomsky. On certain formal properties of grammars. *Information and Control*, 2:137–167, 1959.
- [Cohen, 1992] L. J. Cohen. *Belief & Acceptance*. Oxford University Press, Oxford, UK, 1992.
- [Gold, 1967] E. M. Gold. Language identification in the limit. *Information and Control*, 10: 447–474, 1967.
- [Hacking, 2001] I. Hacking. *An Introduction to Probability and Inductive Logic*. Cambridge University Press, Cambridge UK, 2001.
- [Hansson, 1999] S. O. Hansson. *A Textbook of Belief Updating*. Kluwer, Dordrecht, 1999.
- [Howson and Urbach, 1993] C. Howson and P. Urbach. *Scientific Reasoning: The Bayesian Approach*. Open Court Publishing Company, Peru, Illinois, 1993.
- [Jain *et al.*, 1999] Sanjay Jain, Daniel Osherson, James Royer, and Arun Sharma. *Systems that Learn*. M.I.T. Press, Cambridge MA, 2nd edition, 1999.
- [Kaplan, 1998] M. Kaplan. *Decision Theory as Philosophy*. Cambridge University Press, Cambridge UK, 1998.
- [Kelly, 1996] Kevin T. Kelly. *The Logic of Reliable Inquiry*. Oxford University Press, 1996.
- [Levi, 1967] I. Levi. *Gambling with Truth*. MIT Press, Cambridge MA, 1967.
- [Maher, 1993] P. Maher. *Betting on Theories*. Cambridge University Press, Cambridge, UK, 1993.
- [Martin and Osherson, 1998] Eric Martin and Daniel Osherson. *Elements of Scientific Inquiry*. MIT Press, Cambridge MA, 1998. Revised edition: [www.princeton.edu/~osherson/IL/ILpage.htm](http://www.princeton.edu/~osherson/IL/ILpage.htm).
- [Neyman, 1937] Jerzy Neyman. Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society*, CCXXXVI(A): 333–380, 1937.
- [Osherson *et al.*, 1986] D. Osherson, M. Stob, and S. Weinstein. *Systems that Learn*. MIT Press, 1986.
- [Popper, 1959] K. Popper. *The Logic of Scientific Discovery*. Hutchinson, London, 1959.
- [Putnam, 1979] H. Putnam. Probability and confirmation. In *Mathematics, Matter, and Method: Philosophical Papers, Volume I*. Cambridge University Press, Cambridge, 1979.
- [Rogers, 1967] H. Rogers. *Theory of Recursive Functions and Effective Computability*. McGraw-Hill, New York, 1967.
- [Ross, 1988] S. Ross. *A First Course in Probability, 3rd Edition*. Macmillan, New York City, 1988.
- [Salsburg, 2002] David Salsburg. *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*. Owl Books, 2002.

---

of languages that are potentially learnable, and to interact with other design features (such as the computer simulability of the infant's guessing strategy). For discussion, see [Osherson *et al.*, 1986].