Systems That Learn

An Introduction to Learning Theory

Second Edition

Sanjay Jain

Daniel Osherson

James S. Royer

Arun Sharma

A Bradford Book

The MIT Press
Cambridge, Masssachusetts
London, England

Contents

chapters/front.tex

1 Introduction

The present chapter introduces the subject matter of this book, namely, formal models of empirical inquiry. We begin by indicating the issues that motivate our study. Next come illustrations of models, followed by discussion of their principal features.

§1.1 Empirical inquiry

Many people who have reflected about human understanding and its origins have noticed an apparent disparity. Bertrand Russell [164] (cited in Chomsky [40]) put the matter this way:

> How comes it that human beings, whose contacts with the world are brief and personal and limited, are nevertheless able to know as much as they do know?

Focusing attention on intellectual development, the disparity is between the information available to children about their environment, and the understanding they ultimately achieve about that environment. The former has a sparse and fleeting character whereas the latter is rich and systematic.

To better understand the issue, consider the acquisition of a first language.[1] A few years of casual contact with the ambient language suffices for the infant to master a grammatical system so complex that it still defies description by linguists. Within broad limits, the particular sample of language to which the infant is exposed does not seem to affect the grammatical principles induced, since children raised in different households within the same linguistic community are able to communicate effectively. Moreover, the child's learning mechanism is apparently built to acquire any human language, for children of different racial or ethnic backgrounds are able to acquire the same languages with the same facility. Evidently, some mental process (perhaps largely unconscious) allows children to convert the fragmentary information available about the ambient language into systematic principles that describe it generally.

The same kind of process underlies other tasks of childhood. By an early age the child is expected to master the moral code of his household and community, to assimilate its artistic conventions and its humor, and at the same time to begin to understand the physical principles that shape the material environment. In each case the child is

[1] For an overview and guide to the literature, see Pinker [148].

chapters/chap1.tex

required to convert data of a happenstance character into the understanding (implicit or explicit) that renders his world predictable and intelligible.

It is not surprising that so little is known about the mental processes responsible for children's remarkable intellectual achievements. Even elementary questions remain the subject of controversy and inconclusive findings. For example, there is little agreement about whether children use a general-purpose system to induce the varied principles bearing on language, social structure, etc., or whether different domains engage special-purpose mechanisms in the mind.[2]

The disparity just noted for intellectual development has also been observed in the acquisition of scientific knowledge by adults. Like the child, scientists typically have limited access to data about the environment, yet are sometimes able to convert this data into theories of astonishing generality and veracity. At an abstract level, the inquiries undertaken by child and adult may be conceived as a process of theory elaboration and test. From this perspective, both agents react to available data by formulating hypotheses, evaluating and revising old hypotheses as new data arrive. In the favorable case, the succession of hypotheses stabilizes to an accurate theory that reveals the nature of the surrounding environment. We shall use the term "empirical inquiry" to denote any enterprise that possesses roughly these features.

It is evident that both forms of empirical inquiry — achieved spontaneously in the early years of life, or more methodically later on — are central to human existence and cultural evolution. It is thus no accident that they have been the subject of speculation and inquiry for millenia.[3] The present book describes a set of conceptual and mathematical tools for analyzing empirical inquiry. Their purpose is to shed light on both intellectual development and scientific discovery. They may also be of use in guiding the development and evaluation of artificial systems of empirical inquiry (such as those described in Langley, Simon, Bradshaw, and Zytkow [123]).

Since the pioneering studies of Putnam [155], Solomonoff [184, 185], Gold [80], and the Blums [18] a large technical literature has been devoted to the development and use of the tools at issue here. Papers within this tradition are spread over journals and books in mathematics, computer science, linguistics, psychology, and philosophy. Our topic has variously been called "The Theory of Scientific Discovery," "Formal Learning Theory," "The Theory of Machine Inductive Inference," "Computational Learning Theory," and "The Theory of Empirical Inquiry." We shall use all these terms to describe the collection of definitions, examples, and theorems that emerge from the literature. Our goal is to

[2] For discussion, see Chomsky [40], Pinker [148], and Osherson and Wasow [142].

[3] See Russell [163] for an historical overview.

organize part of this material, and to render it accessible to students and researchers interested in empirical inquiry. Along the way we shall provide pointers to areas given little coverage in these pages.

§1.2 Paradigms

The material to be presented facilitates the definition and investigation of precise models of empirical inquiry. Such models are often referred to as "paradigms." A paradigm offers formal reconstruction of the following concepts, each central to empirical inquiry.

1.1 (a) *a theoretically possible reality*

(b) *intelligible hypotheses*

(c) *the data available about any given reality, were it actual*

(d) *a scientist*

(e) *successful behavior by a scientist working in a given, possible reality*

The concepts figure in the following picture of scientific inquiry, conceived as a game between Nature and a scientist. First, a class of "possible worlds," or possible realities, is specified in advance; the class is known to both players of the game. Nature is conceived as choosing one member from the class, to be the actual world; her choice is initially unknown to the scientist. Nature then provides a series of clues about the actual reality. These clues constitute the data upon which the scientist will base his hypotheses. Each time Nature provides a new clue, the scientist may produce a new hypothesis. The scientist wins the game if his hypotheses ultimately become stable and accurate. Whether the scientist can win the game depends on the breadth of the set of possible worlds. The more constrained Nature's choice of actual world, the more likely the scientist is to discover it.

Different paradigms formalize this picture in different ways, resulting in different games. To fix our ideas, let us now examine some simple paradigms, without concern for rigor at this point.

§1.3 Some simple paradigms

Call a set of positive integers "describable" just in case it can be uniquely described using an English expression. For example, the set $\{2, 4, 6, 8, \ldots\}$ is one such set since it is uniquely described by the expression "all positive, even integers." The describable sets are the theoretically possible realities of the current paradigm (in the sense of 1.1a).

To play the game, it will help to focus on a proper subset of all these realities, namely, the subcollection $\mathcal{C}$ defined as follows. $\mathcal{C}$ contains all sets that consist of every positive integer with a sole exception. Plainly, every set in $\mathcal{C}$ is describable; the set $\{1, 3, 4, 5, 6, \ldots\}$, for example, is uniquely described by "all positive integers except for 2."

In what follows, we shall play the role of Nature; you play the role of scientist. In our role as Nature, we select one member of $\mathcal{C}$, and you (in your role as scientist) must discover the set that we have in mind. Clues about our choice will be provided in the following way. First, we shall order all the elements of the set in the form of a list; then the list will be presented one element at a time. There is no constraint on the list made from the chosen set, except that it must contain all the elements of the set, and only these. For example, one list of the set $\{2, 3, 4, 5, 6, 7, 8, 9, \ldots\}$ is: $3, 2, 5, 4, 7, 6, 9, 8, \ldots$. Aside from seeing the list's members presented one by one, you are provided no further information about it. A list of our set corresponds to 1.1c, the data made available about the possible reality chosen to be actual.

Each time a number is presented, you may announce a conjecture about the set chosen from $\mathcal{C}$ at the beginning of the game (guesses about how we listed the chosen set are not required). Your guesses must take the form of English expressions that uniquely describe a set of positive integers. It is these expressions that constitute the intelligible hypotheses of our paradigm (see 1.1b). Your conjectures at any given moment will be based exclusively on the data available to you, so for purposes of this game you may be construed as a system that translates data into hypotheses. Indeed, any such system is considered to be a "scientist" within the current paradigm, in the sense of 1.1d.

All of items 1.1a-d have now been specified. As for 1.1e, we stipulate that you win the game just in case you make only a finite number of conjectures, and the last one is correct.

Let's play. We have selected a set and ordered it. Here is the first member of the list: 1. Guess, if you like. Next member: 3. Guess again, if you like. To abbreviate, here are the next ten members of the list: $4, 5, 6, 7, 8, 9, 10, 11, 12, 13$. Perhaps your latest conjecture is "all positive integers except for 2." That is a reasonable conjecture. However, it is wrong since according to our list the next number is 2. So go ahead and guess again. Here are the subsequent ten members: $15, 16, 17, 18, 19, 20, 21, 22, 23, 24$. Perhaps now your latest conjecture is "all positive integers except for 14."

The game goes on forever, so we interrupt it at this point to consider the paradigm in more general terms. Let us say that a “guessing rule” is a set of instructions for converting the clues received up to a given point into a conjecture about the chosen set. Your own guesses may well have been chosen according to some guessing rule, and you might take a moment to attempt to articulate it.

Now consider the following guessing rule:

1.2 **Guessing rule:** Suppose that S is the set of numbers that have been presented so far. Let m be the least positive integer that is not a member of S. (S must be finite, so such a number certainly exists.) Emit the conjecture “all positive integers except for m” unless this was your last conjecture (in which case make no conjecture at all).

To illustrate, if the numbers presented so far were $\{4, 5, 8, 1\}$, then rule 1.2 would direct you to conjecture “all positive integers except for 2” (unless you had just made this conjecture, in which case you would not do it again). You should be able to convince yourself of the following fact:

1.3 **Fact:** No matter which set was chosen from $\mathcal{C}$ at the start of the game, and no matter what list was made from that set, consistent application of guessing rule 1.2 is a winning strategy; that is, if you use rule 1.2, then you win in all cases.

Now let us modify the game by adding the set of all positive integers (without exception) to the initial collection $\mathcal{C}$. So our choice of set as “actual reality” is expanded to include one new possibility, namely, $\{1, 2, 3, 4, 5, \ldots\}$. This changes matters quite a bit. For example, guessing rule 1.2 is no longer guaranteed to succeed at the game. Indeed, it is clear that, faced with any listing for the new set $\{1, 2, 3, 4, 5, \ldots\}$, the rule changes its guess infinitely often, and hence never produces a last, accurate conjecture. A more significant fact is the following.

1.4 **Fact:** No guessing rule is guaranteed to win the new game. That is, for every guessing rule R there is a set in the (expanded) collection $\mathcal{C}$ and some way to list the set such that R fails to produce a last, correct conjecture on the list.

The techniques needed to prove Fact 1.4 will be presented in Chapter 3. You can grasp the matter intuitively, however, by playing the new game with a friend. This time you play the role of Nature, and try to defeat your opponent with the following tactic. Begin with the list $2, 3, 4, 5, 6, 7, \ldots$, extending it until your friend announces the hypothesis “all positive integers except 1.” Suppose that your list must be extended to 33 for this to happen. Then continue your list with $1, 35, 36, 37, 38, 39, 40, \ldots$ until you have extracted the hypothesis "all positive integers except 34." Suppose that the list has reached 61 at this point. Then continue with $34, 63, 64, 65, 66, 67, \ldots$ until you hear "all positive integers except 62." If you continue in this devious way, one of two things will happen. Either:

(a) your friend will go for the bait each time, and thereby change her hypothesis infinitely often, or

(b) she will at some point refuse to adopt the conjecture that you intended for her.

In both cases your friend will fail to make a last, correct conjecture about the list you have made. Moreover, in both cases the list you make belongs to the game. To see this, consider the two cases. In case (a), you will end up listing every positive integer. Since this set is a member of the initial collection $\mathcal{C}$, your list represents a legitimate choice for Nature at the start of the game. In case (b), you will end up listing some set consisting of every positive integer with a sole exception. This set is also in $\mathcal{C}$. Thus, in both cases your friend's guessing rule fails on some list for a set that might have been Nature's initial choice. Hence her guessing rule is not guaranteed to win the new game, which proves 1.4. (A more rigorous version of the proof is given in Chapter 3.)

Let's play the last game again (with the extended collection $\mathcal{C}$), but this time within a slightly different paradigm. Instead of being able to arbitrarily order the chosen set, Nature is now required to present the set in increasing order. So there is just one possible listing of any given set in $\mathcal{C}$. For this paradigm, it is easy to formulate a guessing rule that wins in all cases (try stating such a rule).

The foregoing variations point to a basic question about any, well-defined paradigm. The question is: For what collections of realities can winning guessing rules be formulated? This question is a dominant theme of our book.

§1.4 Discussion of the paradigms

We now comment on various aspects of the paradigms just introduced. In fact, our remarks are relevant to almost all of the paradigms discussed in this book.

§1.4.1 Possible realities as sets of numbers

Limiting the possible realities to sets of positive integers is not as austere as it might seem at first. This is because integers may be conceived as codes for objects and events found in scientific or developmental contexts. For example, the sentences to which children are exposed in the course of language acquisition (like all sentences of human language) are complex structures involving phonetic, syntactic, and semantic levels of representation. Their complexity notwithstanding, it may nonetheless be possible to enumerate all possible sentences in a kind of alphabetical order in something like the way pairs, triples, or quadruples of integers can be enumerated.[4] If the enumeration can be carried out by a computable process, then it yields a useful correspondence between sentences and integers, and the latter can be used as codes for the former. In this case, a set of integers corresponds to a language, namely, the language whose sentences are coded by the integers in the set.[5]

Notice that a correspondence of this kind requires that the set of coded entities be denumerable, i.e., have the same cardinality as the integers serving as codes. It might be thought that the restriction to denumerable domains excludes scientific contexts bearing on physical quantities whose values are arbitrary real numbers. However, the rational numbers provide sufficient precision in scientific practice, and the rationals are a denumerable set. So integers can also be used to code many situations involving physical quantities.

For mathematical simplicity, the possible realities figuring in this book are taken to be sets of numbers, or else functions from numbers to numbers. We limit ourselves thereby to studying scientific or developmental contexts in which the relevant objects of inquiry (like sentences) can be coded as integers. It is our belief that much insight into empirical investigation can be achieved within this limitation, a claim that the reader will ultimately have to evaluate for him- or herself.[6]

There is an additional, noteworthy property of the sets and functions that play the role of possible realities in most of what follows. They are "computable" in the sense of being manipulable and recognizable by computer programs (this will be made precise in the next chapter). It is important to recognize that most sets of numbers and most numerical functions are not computable. In fact, from the point of view of their respective cardinalities, the computable functions stand in the same relation to the class of all numerical functions as do the integers to the real line. It follows that by limiting attention to possible realities of a computable nature our theory does not embrace every conceivable scientific situation (we return to this point in Chapter 3). Once again, we believe that this restriction leaves a large and important class of scientific contexts within the purview

[4] For exposition of this kind of enumeration, see Boolos and Jeffery [21, Chapter 1].

[5] For more extended discussion see Weihrauch [192, Chapter 3.3].

[6] Paradigms involving more expressive scientific languages are discussed in Martin and Osherson [127, 128].

of the theory, although we admit to having no proof of this claim.

Another concession to mathematical simplicity can be noted here. Starting in the next chapter, the natural numbers $\{0, 1, 2, 3, \ldots\}$ will be used to construct possible realities, rather than the positive integers $\{1, 2, 3, 4, \ldots\}$. This choice facilitates the use of techniques and results from the theory of recursive functions.

§1.4.2 Intelligible hypotheses

We take hypotheses to be symbolic representations of a real or fictitious world. For example, most hypotheses announced in scientific journals are written in the symbols of the Roman alphabet, supplemented with mathematical notation. Alternatively, the alphabet might be drawn from some system of neural notation used by the brain to represent the structure of the ambient language.

To be intelligible, a symbolic system must provide finite representations of the reality it is designed to depict, even if that reality is infinite in size. For example, the English expressions like "all positive integers except 5" is a finite string of letters that uniquely describes an infinite set. Computer programs can also be conceived as finite descriptions of sets of numbers. Specifically, program P can be taken as specifying the set of all numbers n such that P given input n eventually stops running. This is the approach described in the next chapter and used throughout the sequel. The emphasis on computer programs as hypotheses stems in part from the desire for technological applications. Moreover, it is felt that programs stand in a particularly intimate relation to the sets they describe, inasmuch as they provide a means for recognizing the members of the set. In contrast, the English description "all positive integers that Gauss ever wrote down" uniquely describes a set of numbers, but gives little access to its members.

§1.4.3 Scientists

In our sample paradigm above, scientists were conceived as systems that convert finite sequences of numbers into hypotheses. The scientist may thus be pictured as traveling down an infinite list of numbers, examining the finite amount of data available at any point in the voyage, and emitting hypotheses from time to time about the contents of the entire list, including the infinite, unseen portion. For most of the book, it will be assumed that scientists are *mechanical,* that is, simulable by a computer. Indeed, we shall usually equate scientists with computer programs. It will sometimes prove helpful, however, to remove the assumption of computability from our conception of scientists, in which case they will be conceived as arbitrary functions mapping finite data-sets into conjectures. This liberal attitude will allow us to separate information-theoretic from computability-theoretic aspects of scientific discovery, as will become clearer in Chapter 3.

On the other hand, much of our attention will also be devoted to scientists drawn from narrow subsets of the class of computable processes. That is, we shall consider scientists who operate under various constraints concerning the time devoted to processing data, available memory, selection of hypotheses, ability to change hypotheses, etc. Study of such restrictions will shed light on several issues, including:

(a) the impact of various design features on the performance of computers as scientists, for example, the feature that prevents a computer from abandoning an hypothesis that is consistent with all available data;

(b) the prospects for success by scientists who possess human characteristics, such as time and memory limitations; and

(c) the wisdom of conforming to "rational policies" such as never producing an hypothesis falsified by current data, or never producing an hypothesis that describes a theoretical possibility ruled out in advance.

It will be seen that exploration of such issues sometimes leads to unexpected conclusions.

§1.4.4 Success versus confidence about success

To be successful on a list of numbers, the scientist must produce a final, correct conjecture about the contents of the entire list. She is not required, however, to "know" that any specific conjecture is final. To see what is at issue, consider the first paradigm introduced above in which $\mathcal{C}$ contains just the sets of positive integers with a sole exception. Upon seeing $2, 3, 4, \ldots, 1000$, a scientist might be confident that the list contains the set of positive integers except for 1. But her confidence does not prevent the list from continuing this way: $1, 1002, 1003, 1004 \ldots, 2000$. Confidence at 2000 that the list holds all positive integers except for 1001 is equally unfounded, since the list may continue: $1001, 2002, 2003, 2004, \ldots$. Thus, the scientist is never justified in feeling certain that her latest conjecture will be her last.

On the other hand, Fact 1.3 does warrant a different kind of confidence, namely, that systematic application of guessing rule 1.2 will eventually lead to an accurate, last conjecture on any list generated from a member of $\mathcal{C}$. The relevant distinction may be put this way: If we know that the actual world is drawn from $\mathcal{C}$, then we can be certain that our inquiry will ultimately succeed (if the right guessing rule is applied). But we cannot be certain at any given stage of our inquiry that success has finally arrived.

This asymmetry is a fundamental characteristic of empirical inquiry. In the usual case, scientists can never feel completely confident that their current theory will remain uncontradicted by tomorrow's data. They can only hope that the mental system by which they select hypotheses is adapted to the reality they face. Distinguishing these two issues allows us to focus on scientific success itself, rather than on the secondary question of warranted belief that success has been obtained. Thus, our question will typically be:

> What kind of scientist reliably succeeds on a given class of problems?

rather than:

> What kind of scientist "knows" when she is successful on a given class of problems?

Clarity about this distinction was one of the central insights that led to the mathematical study of empirical discovery (see Gold [80, pp. 465-6]).

§1.4.5 Criteria of success

Compare guessing rule 1.2 to the following, revised version.

1.5 **Guessing rule:** Suppose that S is the set of numbers that have been presented so far. Let m be the least positive integer that is not a member of S. If 2^m is not a member of S, make no conjecture. Otherwise, emit the conjecture "all positive integers except for m" unless this was your last conjecture (in which case make no conjecture at all).

Thus, 1.5 is just like 1.2 except that it imposes a possible delay in producing the conjecture "all positive integers except for m." Although the delay is pointless, it is easy to see that systematic use of 1.5, as with use of 1.2, guarantees success in the first paradigm introduced above.

Rule 1.5 highlights the liberal attitude embodied in our current definition of scientific success. We require that the scientist produce a final conjecture that is correct, but there is no requirement that this final conjecture come as quickly as possible. Let us admit, however, that scientists who examine extravagant amounts of data before making a final, correct conjecture might be considered as useless as scientists who never guess correctly at all. So we might be led to formulate more stringent criteria of scientific success that impose standards of efficiency. Indeed, this is the topic of Chapter 12, below.

Although liberal with respect to efficiency, our present success criterion is stringent about accuracy. To succeed on a given list, the scientist must produce an English description for exactly the numbers in the list, with no omissions or additions. Such accuracy is required in order to make our games interesting, since the sets in play differ by only a few numbers. But what if the initial collection $\mathcal{C}$ were such that every pair of members was infinitely different? In this case, a small error in the scientist's final conjecture might be tolerable. Success in this approximate sense is the topic of Chapters 6 and 7.

More generally, we shall investigate success criteria that are liberal and stringent in a wide variety of ways. In addition to efficiency and accuracy, we shall be concerned with tolerance for noisy data, with different senses of "last conjecture on a list," with behavior on lists outside of a given scientist's competence, and so forth. Of two, distinct criteria we often ask: What kinds of scientific problems can be solved under one criterion that cannot be solved under the other? In this way we hope to shed light on the limits of empirical inquiry that emerge from different kinds of scientific ambition.[7]

Beyond what has been mentioned so far, many additional topics will occupy the chapters of the present book. But enough preliminaries! To get started in a serious way, it is now necessary to review a body of notation and conventions drawn from the theory of computation. This is the topic of the next chapter. Subsequently, in the remaining two chapters of Part I, we formally define and investigate some elementary paradigms.

Summary

The Theory of Machine Inductive Inference (or "Computational Learning Theory," etc.) attempts to clarify the process by which a child or adult discovers systematic generalizations about her environment. The clarification is achieved through the analysis of formal models — called *paradigms* — of scientific inquiry. Each paradigm specifies five concepts central to empirical inquiry, namely:

(a) *a theoretically possible reality*

(b) *intelligible hypotheses*

(c) *the data available about any given reality, were it actual*

[7] As mentioned in the preface, however, the book is far from exhaustive in treating paradigms with claim to illuminating aspects of scientific discovery. In particular, we do not discuss PAC models of learning (see Kearns and Vazirani [105] for an excellent overview).

(d) *a scientist*

(e) *successful behavior by a scientist working in a given, possible reality*

One important question about a given paradigm is this: For what classes of possible realities do there exist scientists who are guaranteed to succeed within any reality drawn from the class?

In order to apply the resources of computational theory to the problem of inductive inference, possible realities are often conceived as sets of integers. In turn, the integers can be conceived as codes for complex objects such as sentences or experimental data. Hypotheses within our theory are usually taken to be computer programs operating over integer inputs. A scientist is any system that converts the finite data sets generated by an environment into hypotheses about the totality of that environment. We shall mainly be concerned with scientists whose behavior is simulable by computer. Much of our inquiry will be devoted to scientists who possess special properties, such as efficiency, ability to cope with noisy data, etc.

Through a simple game, we illustrated one criterion of successful scientific behavior; subsequent chapters will investigate a variety of alternatives. It was noted that our theory distinguishes scientific success from the confidence that a scientist might feel about such success. Only the former will be at issue here.