Unified Inductive Logic: From Formal Learning to Statistical Inference to Supervised Learning

Hanti Lin

University of California, Davis, CA 95616, USA
ika@ucdavis.edu

Abstract. While the traditional conception of inductive logic is Carnapian, I develop a Peircean alternative and use it to unify formal learning theory, statistics, and a significant part of machine learning: supervised learning. Some crucial standards for evaluating non-deductive inferences have been assumed separately in those areas, but can actually be justified by a unifying principle.

Keywords: Induction, Peirce, Formal Learning Theory, Machine Learning, Statistics.

1 Introduction

According to the Carnapian/Bayesian view, inductive logic is a matter of securing a "high proportion" of worlds that make the conclusion true within the domain of the worlds that make the premises true (Carnap 1945). An alternative view of inductive logic is developed here: it is a matter of being guaranteed to get a true conclusion given "enough" premises (or data, or evidence).

This idea can be traced back to C. S. Peirce (1902/[1994]: CP 2.780-1), and I propose that it be systematically developed along the following line of thought: We should be given a certain kind of guarantee *at least* when the amount of evidence is arbitrarily large. What kind of guarantee? A natural idea is to seek (i) a guarantee to *actually* get exactly the true answer to the question posed in one's context of inquiry. If that is unachievable, we should seek (ii) a guarantee to have a *high physical objective probability* of getting exactly the true answer. If even that is unachievable, we should seek, or settle with, (iii) a guarantee to have a high probability of getting *close* to the true answer. Note that those three guarantees, (i)-(iii), form a sequence of increasingly lower standards:

(i) a guarantee to *actually* get exactly the true answer,
↓ weaken
(ii) a guarantee to *have a high chance* of getting *exactly* the true answer,
↓ weaken again
(iii) a guarantee to have a high chance of getting *close to* the true answer.

When we tackle an empirical problem, we ought to strive for the highest achievable of such standards—achievable relative to the empirical problem undertaken.

In a slogan: *strive for the highest achievable!* This is a principle that, I claim, unifies multiple areas that study non-deductive inferences. In particular, many disciplines—including formal learning theory, statistical testing theory, statistical estimation theory, and supervised learning in machine learning—appear to *assume* distinct standards to evaluate the different types of inference methods that they each study. But, if I am right, those evaluative standards need not be *assumed*, let alone assumed *separately*; instead, they can be justified by the unifying principle I propose.

To make my claim precise, many familiar concepts (such as empirical problems and various modes of convergences) need to be reformulated in a uniform setting. This will be addressed in the first half of this paper (through section 3). Then, in the second half (beginning in section 4), the main mathematical results can be stated, and their philosophical significance will be explained. I will conclude by looking into both the history and the future: revisiting Peirce's original ideas and exploring the potential for further unification, including a version of Bayesianism.

2 Varieties of Empirical Problems

Examples first:

Example 1 (Enumerative Induction). The **easy raven problem** poses a question: Are all ravens black? Two potential answers: $\mathtt{Yes}$ vs. $\mathtt{No}$, which form the hypothesis space $\mathsf{H} = \{\mathtt{Yes}, \mathtt{No}\}$. Evidence is gathered by observing ravens one by one, and noting their colors as either black ($\mathtt{1}$) or nonblack ($\mathtt{0}$). So the space of the possible evidential states, E, is the tree of all finite binary sequences. A possible world for this problem takes the form:

$$w \;=\; (e_1 e_2 e_3 \cdots, h),$$

where $e_1 e_2 e_3 \cdots$ is an infinite binary sequence, and h is the competing hypothesis true in that world w. It is assumed in the background that either all ravens are black or, if not, a counterexample would be observed sooner or later if the evidence were to accumulate indefinitely. This assumption rules out only one possible world, $(\mathtt{111}\cdots, \mathtt{No})$, in which the true answer is $\mathtt{No}$ (not all ravens are black) and we would still always only observe black ravens $\mathtt{111}\cdots$. This background assumption is formalized by a set W of possible worlds—the set of all worlds of the form $(e_1 e_2 e_3 \cdots, h)$ except for $(\mathtt{111}\cdots, \mathtt{No})$.

If the above background assumption is relaxed to include the possible world $(\mathtt{111}\cdots, \mathtt{No})$, we obtain the *hard* raven problem, which was studied in formal learning theory only quite recently (Lin 2022).

The above example suggests that, in general, an empirical problem has at least three components: (i) competing hypotheses, (ii) data sequences as possible evidence, (iii) a background assumption. Indeed, these three components also figure in another classic empirical problem:

Example 2 (Statistical Testing). The **fair coin problem** poses a question: Is the coin fair? So the hypothesis space H is $\{\texttt{Fair}, \texttt{Unfair}\}$. Evidence is obtained by tossing the coin repeatedly, and observing the results, either landing heads ($\texttt{1}$) or landing tails ($\texttt{0}$). So the evidence space E for this problem is the tree of binary sequences (as in the easy raven problem). Assumed in the background is the standard IID assumption in statistics: that the bias θ of the coin, i.e., the probability of landing heads, stays constant through time and coin tosses are independent. Under the IID assumption, each possible bias θ in the unit interval $[0, 1]$ determines a probability function $\mathbb{P}_\theta$ defined over E (which is the binary tree). A possible world for this problem takes this form:

$$w \;=\; (e_1 e_2 e_3 \cdots, \theta, \mathbb{P}_\theta).$$

In this world, θ is the true bias of the coin, $\mathbb{P}_\theta$ is the true probability function that represents the data-generation mechanism, and it turns out to generate the data sequence $e_1 e_2 e_3 \cdots$. The background assumption is represented by a set of possible worlds, W, defined as the set of all worlds of the above form. The hypothesis $\texttt{Fair}$ asserts that $\theta = 0.5$, which is true in the worlds in which $\theta = 0.5$. Similarly for the hypothesis $\texttt{Unfair}$, which asserts the negation $\theta \neq 0.5$.

Some clarifications are in order. First, the background assumption is very weak. For example, it is logically compatible with this possible world:

$$w \;=\; (\texttt{1010}\cdots, 0.5, \mathbb{P}_{0.5}),$$

in which the coin is fair and alternates between landing heads ($\texttt{1}$) and tails ($\texttt{0}$). It is also logically compatible with this possible world:

$$w \;=\; (\texttt{1111}\cdots, 0.5, \mathbb{P}_{0.5}),$$

in which the coin is fair and it turns out to always land heads—yes, this is logically possible. In fact, the background assumption is even compatible with any worlds of the following form, where $e_1 e_2 e_3 \cdots$ is an arbitrary binary sequence:

$$w \;=\; (e_1 e_2 e_3 \cdots, 0.5, \mathbb{P}_{0.5}),$$

in which the coin is fair and it turns out to land heads or tails according to the pattern $e_1 e_2 e_3 \cdots$. All those worlds are ruled in, following the practice of classical statistics.

Second, note that probabilities are assigned to the data sequences in the evidence space. That is, each probability function $\mathbb{P}_\theta$ is defined over the evidence space, which represents a stochastic mechanism for generating data or evidence and conforms to the use of *objective physical probabilities*, or *chances*, in classical statistics. Bayesian statistics, and Bayesian epistemology in general, allow probabilities to be assigned to possible worlds, but those probabilities represent subjective degrees of belief. A comparison of the present work with Bayesianism will be provided in the concluding section.

The fair coin problem presupposes that there exist probabilities, so probabilities have to figure in the possible worlds in use. In contrast, for the easy raven problem, it suffices to use worlds of simpler forms without probabilities. So there should be flexibility in our designs of possible worlds:

Definition 1 (Possible World). *A possible world, or world for short, is an ordered pair or tuple of the form:*

$$w \;=\; (e_1 e_2 \cdots, \mathsf{other}),$$

where the first component $e_1 e_2 \cdots$ is an infinite data sequence, understood as the one produced in that world, and the second component other *specifies the other relevant elements of that world.*

Many different things can go into other, such as a distinguished statement true in world w, or the distribution of physical probabilities that exist objectively in that world.

I have mentioned three components of an empirical problem: (hypothesis, evidence, and assumption). There is a fourth component, to be motivated below:

Example 3 (Statistical Estimation). The **coin bias problem** is basically the same as the fair coin problem except that it poses a more fine-grained question: What is the bias of the coin? So the hypothesis space is $\mathsf{H} = [0, 1]$, the unit interval. The evidence space E is the same; so is the background assumption W. The possible worlds in use are the same, too, taking the form $w = (e_1 e_2 e_3 \cdots, \theta, \mathbb{P}_\theta)$. But our guess of the bias, as a hypothesis $h \in \mathsf{H}$, can be more or less accurate; the loss of accuracy can be measured by the difference between the guess h and the true bias. So the loss of accuracy of guessing $h \in [0, 1]$ in a world w, written $\mathsf{Loss}(h, w)$, can be defined as the difference between the guess h and the true bias in w.

So we need this additional definition:

Definition 2 (Loss Function). *A loss function is a function* $\mathsf{Loss} : \mathsf{H} \times \mathsf{W} \to \mathbb{R}$ *that maps any hypothesis h in* H *together with any world in* W *to a nonnegative real number, denoted by* $\mathsf{Loss}(h, w)$*, under the constraint that, in each world $w \in \mathsf{W}$, there is a unique hypothesis $h \in \mathsf{H}$ that attains zero loss of accuracy:* $\mathsf{Loss}(h, w) = 0$ *.*

The uniqueness constraint is adopted in this paper only for the sake of simplicity. It allows us to talk about convergence to *the truth* in a world w: the unique hypothesis that attains zero loss of accuracy in that world w. It is not hard to generalize, referring to *a* truth or truth*s* in a world, but let's opt for simplicity here in order to focus on more important issues.

When we were thinking about the first two examples (the easy raven problem and the fair coin problem), we were not compelled to think about the loss function only because the role of the loss function is already played by the talk of truth and falsity. That is, the loss of accuracy has only two values: 0 for getting

the truth, 1 for getting a falsehood. Or more precisely, in the easy raven and fair coin problems, the loss function is simple: $\mathsf{Loss}(h, w)$ is equal to 1 minus the truth value of hypothesis h in world w.

Thus, for the sake of uniformity, every empirical problem is required to include a loss function as a fourth component. In fact, as will become clear in the next section, this fourth component is even a necessity in order to produce a uniform treatment of various standards for evaluating inference methods. Hence the following definition:

Definition 3 (Empirical Problem). *An empirical problem, or problem for short, is a quadruple* $(\mathsf{H}, \mathsf{E}, \mathsf{W}, \mathsf{Loss})$ *with the following interpretations and requirements:*

- H *is a set of hypotheses, understood as the hypotheses under consideration.*
- E *is a tree, called the evidence tree, in which the nodes are finite data sequences* $e_1 \cdots e_n$ *ordered by sequence extension, and every branch is infinite.*
- W *is a set of possible worlds such that the infinite data sequences that appear in the worlds therein are exactly the branches of the evidence tree* E*. This set* W *is meant to contain all and only the possible worlds compatible with one's background assumption.*
- Loss *is a loss function on* $\mathsf{H} \times \mathsf{W}$*.*

This definition is also general enough to cover problems in supervised learning, from binary classification to nonparametric regression, to be discussed in section 6.

The components of an empirical problem have their own roles to play. The evidence space E and the hypothesis space are used to define inference methods as functions from the former to the latter (with a minor refinement to be formally stated below). The other two components are used to define various standards for assessing inference methods, as we will see below. Those standards examine each inference method by considering how that method performs for truth seeking across a range of the possible worlds—the worlds in W. So, W represents the background assumption against which inference methods are evaluated. These ideas will be fleshed out in the formal definitions provided in the next section.

3 Modes of Convergence as Evaluative Standards

Here are the objects of evaluation:

Definition 4 (Inference Method). *An inference method for an empirical problem* $(\mathsf{H}, \mathsf{E}, \mathsf{W}, \mathsf{Loss})$ *is a function* $M : \mathsf{E} \to \mathsf{H} \cup \{?\}$*; that is, M can receive any finite data sequence that figures as a node of the evidence tree* E*, and then output one of the hypotheses in* H *or a question mark* $?$ *that represents judgment suspension.*

Only the first two components of a problem, H and E, are needed to define the inference methods for that problem. The remaining two components, W and Loss, are used to define standards for assessing inference methods. But preliminaries first:

Definition 5. *Here are some notation conventions:*

- *Let $\mathbb{P}_w$ denote the probability measure true in w if w contains such a thing.*
- *Let $h^{M,n,w}$ denote the hypothesis or output of method M at stage n in world w, that is, $M(e_1e_2\cdots e_n)$, where $e_1e_2\cdots e_n$ is the sequence of the first n data points received in world w.*
- *Similarly, let $\hat{h}^{M,n}$ be the random variable that maps each possible data sequence of length n to the output of M on that sequence; so $\hat{h}^{M,n}$ can be intuitively understood to denote the random hypothesis or output of method M at stage n (leaving worlds and data sequences unspecified), following the standard use of variables in statistics.*

The random variable notation $\hat{h}^{M,n}$ is particularly convenient for expressing probabilities like the following:

$$\mathbb{P}_w\Big[\,\mathsf{Loss}\big(\hat{h}^{M,n},w\big)<\epsilon\,\Big] \;=_{\mathrm{df}}\; \mathbb{P}_w\Big\{\,e_1e_2\cdots e_n : \mathsf{Loss}\big(M(e_1e_2\cdots e_n),w\big)<\epsilon\,\Big\}.$$

The right side means the probability, in world w, of observing a data sequence $e_1e_2\cdots e_n$ of length n such that the loss of M's output is below the threshold ϵ. This phrase can be more concisely expressed using the notation on the left side: it denotes the probability, in world w, for the loss of M's output to be bellow ϵ given sample size n.

Here is the real thing:

Definition 6 (Three Basic Modes of Convergence). *An inference method M for a problem $(\mathsf{H},\mathsf{E},\mathsf{W},\mathsf{Loss})$ can be said to achieve one or another mode of convergence to the truth. The following defines three modes.*

- **Convergence with Nonstochastic Identification**

  for any world $w\in\mathsf{W}$,
  there exists sample size N such that,
  for all $n\geq N$,

  $\mathsf{Loss}\big(h^{M,n,w},w\big)=0$;

  that is, given sample size n, M outputs exactly the truth in w.

- **Convergence with Stochastic Identification:**

  for any probability threshold $1-\delta<1$,
  for any world $w\in\mathsf{W}$,
  there exists sample size N such that,
  for all $n\geq N$,

  $\mathbb{P}_w$ *exists, and*
  $\mathbb{P}_w\big[\,\mathsf{Loss}\big(\hat{h}^{M,n},w\big)\;=\;0\,\big]\;>\;1-\delta$;

  that is, given sample size n, the probability for M to output exactly the truth is high (greater than $1-\delta$) in w.

- **Convergence with Stochastic Approximation**

  for any upper bound on loss $\epsilon > 0$,
  for any probability threshold $1 - \delta < 1$,
  for any world $w \in \mathsf{W}$,
  there exists sample size N *such that,*
  for all $n \geq N$,

  $\mathbb{P}_w$ *exists, and*
  $\mathbb{P}_w\big[\,\mathsf{Loss}\big(\hat{h}^{M,n}, w\big) \,<\, \epsilon\,\big] \,>\, 1 - \delta$;

  that is, given sample size n, *the probability that* M *outputs a hypothesis* ϵ*-close to the truth is high* (*greater than* $1 - \delta$) *in* w.

These three basic modes of convergence have been mostly developed and studied separately in different areas, couched in very different languages, and employed to talk about apparently different subjects. The formalism developed here provides a uniform reformulation of those modes. Moreover, the plain English glosses accompanying the above reformulated definitions (the clauses following 'that is') make it clear that the old, familiar modes of convergence are covered:

(i) **Convergence w/ Nonstochastic Identification**: It goes by the name *identification/decidability in the limit* in formal learning theory (Kelly 1996: ch. 3), originally designed for theory choice in a deterministic setting.

(ii) **Convergence w/ Stochastic Identification**: This captures the conjunction of *consistency in significance level* and *consistency in power* as studied in statistical hypothesis testing (Lehmann 1999: ch. 3). It also captures the so-called *model selection consistency* in statistical model selection (Claeskens et al. 2008: ch. 4).

(iii) **Convergence w/ Stochastic Approximation**: This captures so-called *estimation consistency* in the statistical theory of estimation (Lehmann 1999: ch. 2). As you will see below (section 6), it also captures the consistency of learning algorithms in supervised learning, which includes classification (Shalev-Shwartz et al. 2014: Part I) and nonparametric regression (Györfi et al. 2002: ch. 1).

It is not hard to develop variants of the above modes of convergence. Let me briefly outline some notable ones. For each of the three modes defined above, exchanging the quantifiers 'for any world' and 'there exists sample size' gives us a higher standard, a mode of *uniform* convergence. There are other variants, which can be obtained by thinking about: *rates* of convergence (Györfi et al. 2002: ch. 1), *monotonic* convergence or somewhat *stable* convergence if not perfectly monotonic (Lin 2022), and convergence for *almost all* worlds in W—almost all in a topological sense (Lin 2019). Those modes of convergence and their combinations have been studied in one or another area. From an epistemological point of view, they correspond to higher or lower standards for evaluating inference methods. However, I will focus on the three basic modes defined above, which suffice for making the philosophical point I want to make below.

4 Towards Unification

The three basic modes of convergence have been reformulated in a uniform notation to clarify their close connection. Although they are developed in distinct areas, they set standards for assessing *all* inference methods in *any* empirical problems. The standards range from high to low—from actually getting the truth, to probably getting exactly the truth, to probably getting close to the truth. Hence the hierarchy depicted in Table 1.

Table 1. *The Hierarchy of the Three Basic Modes of Convergence*

(i) **Convergence w/ Nonstochastic Identification**
|
(ii) **Convergence w/ Stochastic Identification**
|
(iii) **Convergence w/ Stochastic Approximation**

We need just one more definition before the first result can be stated:

Definition 7 (Achievability). *A mode of convergence is said to be achievable for an empirical problem if some inference method for that problem satisfies that mode of convergence.*

Then we have:

Proposition 1. *Consider the hierarchy of the three modes of convergence depicted in Table* 1*.*

- *For the easy raven problem, the highest achievable is mode* (i)*.*
- *For the fair coin problem, it is mode* (ii)*.*
- *For the coin bias problem, it is mode* (iii)*.*

See the appendix for the proof. The novelty of this result consists in what is *not* achievable. The claims of achievability are already obtained with greater generality in formal learning theory and statistics. But I still provide elementary proofs of those achievability claims, considering that readers familiar with one of the two areas might not be so with the other.

Since the novelty is in the claims of *un*achievability, let me explain the proof strategy. Mode (i) is unachievable for the two statistical problems due to a specific kind of *underdetermination by data*: those two problems allow that one and the same infinite data sequence can be generated under different hypotheses. Mode (ii) is unachievable for the coin bias problem because of considerations about *cardinality*: the number of competing hypotheses (all the real numbers in the unit interval) is strictly greater than the number of possible evidential inputs (all finite binary sequences). So, for any inference method, there is at least one hypothesis doomed to be never an output.

I suspect that the proof strategy just sketched allows us to generalize the result to cover wide classes of empirical problems. However, I will focus on the three paradigm problems—easy raven, fair coin, and coin bias—to streamline the discussion and emphasize more pressing points: a picture of different areas unified into a cohesive whole.

Although statistics and formal learning theory may appear to be very different, they operate with the same guiding principle in the unifying picture I propose:

Strive for the Highest Achievable!

Statisticians design inference methods with a minimum qualification in mind, which is standardly called consistency in statistics but is just the stochastic mode (ii) or (iii) defined above. Statisticians do not aim at the higher mode (i). The *cause* underlying statisticians' practices may lie in how their textbooks are written and their PhD programs structured. But the *reason* that justifies what statisticians do, I propose, is based on this fact: the empirical problems addressed by statisticians are too hard to make the higher mode (i) achievable. Similarly, formal learning theorists do not aim at the lower modes (ii) and (iii), because they should not: since the empirical problems they study make the higher mode (i) achievable, they should use that higher standard to evaluate inference methods.

So, statistics and formal learning theory need not be regarded as two separate areas that *assume* different standards to evaluate different kinds of inference methods. Instead, there is a unifying principle: strive for the highest achievable. The use of different standards in different areas—stochastic vs. nonstochastic—need not be assumed and can be *justified* by the proposed principle.

Similarly, hypothesis testing and parameter estimation need not be viewed as distinct subareas of statistics with separate evaluative standards. The use of different standards—identification vs. approximation—need not be assumed and can be *justified* by the proposed principle.

5 Deterministic vs. Stochastic?

Although formal learning theory is often presented in a way that gives the impression that it only concerns deterministic hypotheses (as in Kelly 1996), let's be more careful. The hypothesis "all ravens are black" is neutral regarding whether the actual world is deterministic or indeterministic. This hypothesis can be true in a deterministic world, but it can also be true in an indeterministic world—a world in which all ravens turn out to be black by chance. We have represented this world without probabilities by:

$$w \;=\; (\texttt{111}\cdots, \texttt{Yes})\,.$$

Strictly speaking, this w only represents a very coarse-grained world, for it says nothing whatsoever about whether determinism is true or false. The coarse-

grained world w can be realized by two different, more fine-grained worlds:

$$\begin{aligned} u &= (\mathtt{111}\cdots, \mathtt{Yes}, \mathbb{P}_{\text{trivial}}), \\ u' &= (\mathtt{111}\cdots, \mathtt{Yes}, \mathbb{P}'), \end{aligned}$$

where, $\mathbb{P}_{\text{trivial}}$ is a probability function that only assigns trivial probabilities 0 and 1—in particular, it assigns probability 1 to the constant sequence $\mathtt{111}\cdots$. But $\mathbb{P}'$ is a nontrivial probability function, assigning (say) .7, to the constant sequence $\mathtt{111}\cdots$. In both of those fine-grained worlds, it is true that all ravens are black. However, it is true deterministically in the first world u, and true only by chance in the second world u'. Hence the following definition:

Definition 8 (Probabilistic Extension). *Suppose that an evidence tree* E *is given. Consider a world* $w = (e_1 e_2 \cdots, h)$ *that contains no probability measure. This world* w *is said to have another world* w' *as a probabilistic extension if* w' *takes the form* $w' = (e_1 e_2 \cdots, h, \mathbb{P})$*, which differs from* w *only by adding a* (*trivial or nontrivial*) *probability measure* $\mathbb{P}$ *over the evidence tree* E*.*

Definition 9 (Fine-Grained Version). *Fine-grained versions of a problem* $(\mathsf{H}, \mathsf{E}, \mathsf{W}, \mathsf{Loss})$ *are problems of the form:* $(\mathsf{H}, \mathsf{E}, \mathsf{W}', \mathsf{Loss}')$*, with the same* H*, the same* E*, but a different* W' *and a different* Loss'*, satisfying the following constraints:*

- W' *can be obtained from the original* W *by, first, removing each world* w *that contains no probability measure and, then, replacing* w *with one or multiple probabilistic extensions of* w*.*
- $\mathsf{Loss}'(h, w') = \mathsf{Loss}(h, w)$ *if* w' *is a probabilistic extension of* w*.*

Then we have:

Proposition 2. *Consider the three modes of convergence in Table* 1*. If the highest mode* (i) *is achievable for a problem* $\mathcal{P}$*, then the three modes* (i)-(iii) *are all achievable for any fine-grained version of* $\mathcal{P}$ *that only involve countably additive probability measures.*

See the appendix for the proof.

Corollary 1. *Mode* (i) *implies mode* (ii) *in the precise sense described by the preceding proposition. Mode* (ii) *in turn implies mode* (iii) *in the standard sense.*

The easy raven problem fails to make (ii) and (iii) achievable only because of its somewhat misleading mathematical representation. The fine-grained versions of that problem do a better representational job, being explicitly neutral between determinism and indeterminism, making all the three modes provably achievable.

So it is misleading to say that formal learning theory differs from statistics in that the former presupposes determinism. This claim is incorrect; there is no such presupposition. The distinction between the two areas is best understood as a division of labor guided by a common principle, as explained earlier.

6 Supervised Learning

The present setting also covers problems studied in supervised learning, whose simplest examples concern binary classification:

Example 4 (Binary Classification). Suppose that we want to determine whether a given object is a cat by examining a (pixelated) picture of it. Or suppose that we want to determine whether a given watermelon is tasty by examining properties such as its skin color distribution and the sound it produces when tapped. Or suppose, in general, that we want to classify an object of a certain kind into category 0 or category 1, and do it on the basis of the feature that it has in a countable set of mutually exclusive features: $\mathsf{X} = \{x_1, x_2, x_3 \ldots\}$. A **(binary) classifier** is an indicator function $h : \mathsf{X} \to \{0, 1\}$. Suppose, further, that we are given a set H of (some or all) such classifiers. We would like to pick a good classifier from H in light of examples. An **example** is an object with a specified feature $x \in \mathsf{X}$ and a specified category 0 or 1; so an example is formally an ordered pair $(x, 0)$ or $(x, 1)$. Such examples form an **example space**: $\mathsf{X} \times \{0, 1\}$. A data sequence $(e_1 \cdots e_n)$ is a finite sequence of such ordered pairs. All such data sequences form an evidence tree $\mathsf{E} = (\mathsf{X} \times \{0, 1\})^{<\infty}$. A **learning algorithm** is just an inference method $M : \mathsf{E} \to \mathsf{H} \cup \{?\}$, although people in machine learning typically only consider $M : \mathsf{E} \to \mathsf{H}$. To sum up: a **task of binary classification** can be formally identified as an ordered pair (X, H) consisting of two elements:

- a countable feature space $\mathsf{X} = \{x_1, x_2, x_3 \ldots\}$,
- a set $\mathsf{H} \subseteq \{0, 1\}^{\mathsf{X}}$, as a candidate pool of classifiers.

Those two elements suffice to determine the other elements of such a task, including classifiers, examples, an evidence tree E, and learning algorithms. *But what counts as a good classifier—good for predictive purposes?* The quality of a classifier depends on the state of the world. Assume that examples are generated by a probabilistic mechanism, represented by a probability distribution D over the example space $\mathsf{X} \times \{0, 1\}$, where $D(x, y)$ denotes the probability that the next example has feature x and belongs to category y. The **predictive risk** of a classifier h with respect to the (true but unknown) distribution D is given by the probability of *misclassification*:

$$\mathsf{Risk}(h, D) \;=\; D\big\{(x, y) : h(x) \neq y\big\} \;=\; \sum_{(x,y):\, h(x) \neq y} D(x, y)$$

The smaller, the better. All this points to an empirical problem $(\mathsf{H}, \mathsf{E}, \mathsf{W}, \mathsf{Loss})$, called a **(binary) classification problem**:

- Question: Which classifier is the best in class (in H) for predictive purposes? So the hypothesis space is H, the given candidate pool of classifiers.
- E, the evidence tree, is the tree of sequences of examples (x, y) taken from the example space $\mathsf{X} \times \{0, 1\}$.

- W, which represents the background assumption, is the set of all possible worlds of the form $w = (e_1 e_2 \cdots, D, \mathbb{P}_D)$, where D is an arbitrary probability distribution over the example space, and $\mathbb{P}_D$ is the IID probability measure (on E) generated from distribution D.
- $\mathsf{Loss}(h, w) = \mathsf{Risk}(h, D_w) - \min_{h' \in \mathcal{H}} \mathsf{Risk}(h', D_w)$, where D_w denotes the distribution in world w

This concludes the last and longest of the four examples examined in this paper. It is no coincidence that textbooks in machine learning use the symbol h to denote classifiers, referring to them as hypotheses (Shalev-Shwartz et al. 2014).

Once we see what binary classification is, generalizations are straightforward. When the category set $\mathsf{Y} = \{0, 1\}$ is generalized to a set of finite categories, $\mathsf{Y} = \{1, 2, \ldots, k\}$, we obtain problems of *multiclass classification*. When both the feature space X and the category space Y become continuous, say Euclidean spaces $\mathbb{R}^n$, we have problems of *nonparametric regression*. These generalizations encompass nearly the entire area of supervised learning.

In supervised learning, the minimum qualification for good learning algorithms is called *consistency*, which is essentially the mode of convergence (iii) as defined in this paper—convergence with stochastic approximation. However, consistency does not need to be *assumed* as a minimum qualification in supervised learning. It can be justified as follows: By the principle of striving for the highest achievable, the achievability of at least mode (iii) in the hierarchy in Table 1 implies that mode (iii) or a higher mode has to be achieved, which in turn implies, thanks to corollary 1, that mode (iii) has to be achieved at the very least—as a minimum qualification. The same argument also justifies mode (iii) as a minimum qualification for good estimators in statistical estimation.

7 Closing

This paper reformulates three basic modes of convergence to the truth using a uniform notation, and explores their potential to create a unified picture of inductive logic. Additional modes of convergence (as mentioned at the end of section 3) should be investigated to enrich the simple hierarchy in Table 1 and to determine whether more areas can be incorporated into the unified framework. In this closing note, let me explain how we arrived at this point and where we might go from here.

In hindsight, Peirce already did a lot for us over a century ago. As pointed out in Lin's (forthcoming) discussion of Peirce's convergentism, Peirce sought to justify enumerative induction by appeal to the mode of convergence with nonstochastic identification (Peirce 1994: CP 2.775, 7.125), and he *also* studied statistical estimation using the mode of convergence with stochastic approximation (Peirce 1994: CP 2.669-93). Unfortunate, after Peirce, those two modes of convergence took diverging paths. The stochastic one helped establish the estimation theory in statistics (Fisher 1925), and the nonstochastic one helped create formal learning theory (Putnam 1965, Gold 1967)—but those two areas

have long been regarded as largely unrelated. I propose that we return to Peirce's idea and reinforce it with the principle developed above "Strive for the Highest Achievable!"

The result is a conception of logic that unifies formal learning theory, statistics, and a significant part of machine learning: supervised learning. But can it be extended to cover other areas of machine learning, such as reinforcement learning and unsupervised learning? I am optimistic for reinforcement learning, for it has been given a largely uniform foundation as supervised learning (Mohri et al. 2018). Of course, a detailed argument is needed nonetheless, but that has to be left to future work.

However, the prospect for incorporating unsupervised learning into the unified picture remains unclear. In fact, even theorists of machine learning have difficulty evaluating algorithms of unsupervised learning by a rigorous standard—let alone a standard defined as a mode of convergence. To talk about convergence to the truth, there needs to be a truth to begin with, but such a truth is typically missing in unsupervised learning, as observed in a standard textbook on the theoretical foundation of machine learning: "[A] basic problem is the lack of "ground truth" for clustering, which is a common problem in unsupervised learning" (Shalev-Shwartz & Ben-David 2014: 308).

The most influential approach to inductive logic to date in philosophy is the Bayesian one. I suspect that it can be developed in a manner that fits into the unified picture I have painted. On the Bayesian approach, inductive logic should be founded on the idea that probabilities as rational degrees of belief ought to be updated by conditionalization on the new evidence (Carnap 1945). But is it that all priors—initial assignments of probabilistic degrees of belief—are equally permissible? The answer is "yes" according to the *subjectivists* in Bayesian epistemology. But the answer is "no" according to anti-subjectivists, including the *objectivists*, who maintain that permissible priors must be "flat", conforming to the principle of indifference or a variant of it. But the objectivists are not the only people who would like to constraint the candidate pool of permissible prior by something more than the axioms of probability. For a group of Bayesians who are less known to the philosophical community but work in a branch of statistics called *Bayesian nonparametric statistics*, it is customary to rule out some prior assignments of degrees of belief by appeal to considerations about convergence (see Rousseau 2016 for a review). The idea is simple: a Bayesian prior is permissible only if updating it successively via conditionalization ensures a sequence of posteriors that achieves a specific mode of convergence to the truth, commonly called *Bayesian consistency*. Thus, even Bayesians may find their place within the unifying framework I have outlined for inductive logic; however, the details remain to be addressed in future work.

I hope all this offers initial reasons to be optimistic about the prospect of a unified inductive logic.

Acknowledgements

I thank Konstantin Genin for his patient and inspiring discussions. I also thank two anonymous reviewers for their detailed and constructive comments on earlier versions of this paper.

A Appendix: Proofs

A.1 Proof of Proposition 1

To establish the first part for the easy raven problem, it suffices to verify that mode (i) is achieved in the easy raven problem by the inference method that outputs `Yes` exactly when the input contains no `0`, and outputs `No` otherwise.

Now, establish the second part for the fair coin problem as follows. Mode (i), convergence with nonstochastic identification, is unachievable because the fair coin problem features this kind of underdetermination by data: The background assumption W in the fair coin problem is so weak that it contains a pair of worlds in W sharing the *same* indefinite data sequence (say, heads, tails, heads, tails, and so on), while one world makes `Fair` true and the other makes `Unfair` true. So, to converge to the truth in one of those two worlds is to converge to a falsehood in the other. Mode (i) is thus unachievable. As to the achievability of mode (ii) for the fair coin problem, it can be proved by Bernoulli's law of large numbers as follows. Let M be the inference method that outputs `Fair` exactly when the observed frequency $\bar{X}_n$ of heads in the sample is *close enough* to 0.5 in the sense that $|\bar{X}_n - 0.5| < 1/\sqrt[4]{n}$, where n is the sample size. Now recall Bernoulli's law of large numbers:

$$\mathbb{P}_\theta(|\bar{X}_n - \theta| < \epsilon) \;\geq\; 1 - \frac{1}{4n\epsilon^2}\,.$$

In any world in which the truth is `Fair`, the probability for M to get the truth is obtained by letting $\theta = 0.5$ and $\epsilon = 1/\sqrt[4]{n}$ in Bernoulli's law of large numbers:

$$\begin{aligned} & \mathbb{P}_{.5}\left(|\bar{X}_n - 0.5| < 1/\sqrt[4]{n}\right) \\ \geq\; & 1 - \frac{1}{4n(1/\sqrt[4]{n})^2} \\ =\; & 1 - \frac{1}{4\sqrt{n}}\,, \end{aligned}$$

which converges to 1 as n tends to infinity. Moreover, in any world where the truth is `Unfair` with true bias $\theta \neq 0.5$, when the sample size n is large enough to ensure that $1/\sqrt[4]{n} < \frac{1}{2}|\theta - 0.5|$, the probability for M to get the truth is

$$\begin{aligned} & \mathbb{P}_\theta\left(|\bar{X}_n - 0.5| \geq 1/\sqrt[4]{n}\right) \\ \geq\; & \mathbb{P}_\theta\left(|\bar{X}_n - \theta| < 1/\sqrt[4]{n}\right) \\ \geq\; & 1 - \frac{1}{4\sqrt{n}}\,, \end{aligned}$$

which converges to 1 as n tends to infinity. So M achieves mode (ii).

Now, establish the third part for the coin bias problem as follows. The set of all possible (evidential) inputs is the set of all nodes in the binary tree. So there are countably many possible inputs. But there are uncountably many hypotheses. Let M be an arbitrary inference method. It follows that there is some hypothesis h_M that M is doomed to never output in any world. So, the probability for M to output (exactly) the truth always stays zero in the worlds in which the truth is h_M. So M fails to achieve mode (ii). But mode (iii) is achievable, for a well-known reason: the inference method that always outputs the observed frequency of heads achieves mode (iii), convergence with stochastic approximation, thanks to Bernoulli's law of large numbers (again). Q.E.D.

A.2 Proof of Proposition 2

Suppose that mode (i) is achievable for a problem $(\mathsf{H}, \mathsf{E}, \mathsf{W}, \mathsf{Loss})$, which has a refined version $(\mathsf{H}, \mathsf{E}, \mathsf{W}', \mathsf{Loss}')$. Let M be an inference method for the original problem that achieves mode (i), which implies that there is a one-to-one correspondence between the worlds in W and the infinite branches of the evidence tree E:

$$w = (e_1 e_2 \cdots, -) \quad \mapsto \quad e_1 e_2 \cdots \in \mathsf{E}\,.$$

So, any probability measure defined over E can be equivalently construed as a probability measure defined over W. Let $\mathsf{Success}(M, n)$ be the set of the worlds $w \in \mathsf{W}$, or equivalently the branches $b \in \mathsf{E}$, such that, by stage n, M has output the truth and would never drop it in b. By definition, $\mathsf{Success}$ is monotonic in this sense: whenever $n \leq n'$,

$$\mathsf{Success}(M, n) \;\subseteq\; \mathsf{Success}(M, n')\,.$$

Since M achieves mode of convergence (i), it follows that

$$\textstyle\bigcup_{n\in\mathbb{N}} \mathsf{Success}(M, n) \;=\; \mathsf{E}\,.$$

Let w' be an arbitrary world in W', and let $\mathbb{P}_{w'}$ denote be the probability measure over E in w'. Then, by countable additivity, we have:

$$\begin{aligned} & \lim_{n\to\infty} \mathbb{P}_{w'}\big(\mathsf{Success}(n)\big) \\ =\; & \mathbb{P}_{w'}\big(\textstyle\bigcup_{n\in\mathbb{N}} \mathsf{Success}(n)\big) \\ =\; & \mathbb{P}_{w'}(\mathsf{E}) \\ =\; & 1 \end{aligned}$$

Now, note that the probability for M to get the truth by stage n in world w' is greater than or equal to $\mathbb{P}_{w'}\big(\mathsf{Success}(n)\big)$, which approaches 1 as $n \to \infty$ thanks to the above calculation. So M achieves mode (ii), convergence with stochastic identification. Achieving mode (ii) immediately implies achieving mode (iii). Q.E.D.

References

1. Carnap, R. (1945) "On Inductive Logic", *Philosophy of science*, 12(2), 72-97.
2. Claeskens, G., & Hjort, N. L. (2008) *Model Selection and Model Averaging*, Cambridge University Press.
3. Lehmann, E. L. (Ed.). (1999) *Elements of Large-sample Theory*, New York, NY: Springer New York.
4. Fisher, R.A. (1925) *Statistical Methods for Research Workers*, Oliver & Boyd.
5. Gold, E.M. (1967) "Language Identification in the Limit", *Information and Control*, 10(5): 447-474.
6. Györfi, L., Kohler, M., Krzyzak, A., & Walk, H. (2002) *A Distribution-free Theory of Nonparametric Regression*, New York: Springer.
7. Kelly, K.T. (1996) *The Logic of Reliable Inquiry*, Oxford University Press.
8. Lin, H. (2019) "The Hard Problem of Theory Choice: A Case Study on Causal Inference and Its Faithfulness Assumption", *Philosophy of Science*, 86(5), 967-980.
9. Lin, H. (2022) "Modes of Convergence to the Truth: Steps toward a Better Epistemology of Induction", *The Review of Symbolic Logic*, 15(2), 277-310.
10. Lin, H. (forthcoming) "Convergence to the Truth", in Sylvan, K., Sosa, E., Dancy, J. & Steup, M. (eds.) *The Blackwell Companion to Epistemology*, 3rd edition, Wiley Blackwell.
11. Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018) *Foundations of Machine Learning*, MIT Press.
12. Peirce, C. S. (1994) *Collected Papers of Charles Sanders Peirce* (Volumes I-VIII), InteLex Corp.
13. Putnam, H. (1965) "Trial and Error Predicates and a Solution to a Problem of Mostowski", *The Journal of Symbolic Logic*, 30(1): 49-57.
14. Rousseau, J. (2016) "On the Frequentist Properties of Bayesian Nonparametric Methods", *Annual Review of Statistics and Its Application*, 3: 211-231.
15. Shalev-Shwartz, S. & Ben-David, S. (2014) *Understanding Machine Learning: From Theory to Algorithms*, Cambridge University Press.