Modes of Convergence to the Truth: Steps toward a Better Epistemology of Induction

Word count:
273 (abstract), 5987 (body text,
excluding bibliography and appendix)

Abstract

Those who engage in normative or evaluative studies of induction, such as formal epistemologists, statisticians, and computer scientists, have provided many positive results for justifying (to a certain extent) various kinds of inductive inferences. But they all have said little about a very familiar kind of induction. I call it *full* enumerative induction, of which an instance is this: "We've seen this many ravens and they all are black, so all ravens are black"—without a stronger premise such as IID or a weaker conclusion such as "all the ravens observed in the future will be black". I explain why those theorists of induction all say little about full enumerative induction. To remedy this, I propose that Bayesians be *learning-theoretic* Bayesians and learning-theorists be *truly* learning-theoretic—in three steps. (i) Understand certain modes of convergence to the truth as *epistemic ideals* for an inquirer to achieve where possible. (ii) Require the norm that an inquirer ought to achieve the highest achievable epistemic ideal. (iii) See whether full enumerative induction can be justified as—that is, proved to be—a necessary means for achieving the highest epistemic ideal achievable for tackling the problem of whether all ravens are black. The answer is positive, thanks to a new theorem whose Bayesian version is proved as well. The technical breakthrough consists in introducing a mode of convergence slightly weaker than Gold's (1965) and Putnam's (1965) identification in the limit; I call it *almost everywhere* convergence to the truth, where the conception of "almost everywhere" is borrowed from geometry and topology. The result is also applied to justify a form of Ockham's razor.

Keywords: Enumerative Induction, Learning Theory, Bayesian Epistemology, Epistemic Norms, Ockham's Razor, Convergence, Topology.

1 Introduction

The *general* problem of induction is the problem of identifying (i) the range of the inductive inferences that we can justify and (ii) the extent to which we can justify them. Under the general problem there are several subproblems. For example, there is the special subproblem of how it is possible to reliably infer causal structures solely from observational data, which has attracted many statisticians, computer scientists, and philosophers.[1] And there is the more general subproblem of whether it is possible to escape Hume's dilemma—a dilemma that aims to undermine any justification of any kind of inductive inference.[2] In this paper I want to focus on a subproblem of induction that is more *normative and evaluative* in nature.

Here is the background. Normative/evaluative studies of inductive inference are pursued in many mathematical disciplines: formal epistemology, statistics, and theoretical computer science. But somewhat curiously, they all have said little about the most familiar kind of inductive inference, of which an instance is this:

> (Full Enumerative Induction) We have observed this many black ravens and they all are black; so all ravens are black.

To be sure, those disciplines have had much to say about enumerative induction, but typically only about a *restricted* version that weakens the conclusion or strengthens the premise. Here is an example:

> (Restricted Enumerative Induction) We have observed this many black ravens and they all are black; so all the ravens *observed in the future* will be black

This is the kind of enumerative induction studied by Bayesian confirmation theorists,[3] and also by formal/algorithmic learning theorists.[4] Classical statisticians and statistical learning theorists study an *even more* restricted version that adds a substantial assumption of IID (independent and identically distributed) random variables, where the randomness is due to objective chance.

So they all set aside a serious study of full enumerative induction. But why? The reason for statisticians is obvious: their primary job is to study inductive inferences

[1] For a groundbreaking treatment of this problem, see Spirtes, Glymour, and Scheines (1993/[2000]).

[2] See Hume (1777) for his formulation of the dilemma, and Reichenbach (1938, sec. 38) for a more powerful and popular formulation of the dilemma. A quite comprehensive list of attempted responses is provided in Salmon (1966: chap. 2).

[3] See, for example, Carnap (1950), Hintikka (1966), and Hintikka and Niiniluoto (1980). They all talk about the probabilistic inference from evidence $F(a_1) \wedge \ldots \wedge F(a_n)$ to the countable conjunction $\bigwedge_{i=1}^{\infty} F(a_i)$, where a_i means the i-th individual (or raven) observed in one's inquiry. If the index i is understood as enumerating all ravens in the universe in an order unbeknownst to the inquirer, then the evidence cannot be formalized the way they do.

[4] See, for example, Kelly (1996) and Schulte (1996).

under the IID assumption or the like. The reason for Bayesians and formal learning theorists is deeper. Let me explain.

It is logically possible that there are nonblack ravens but an inquirer never observes one throughout her entire life (even if she is immortal). Call such a possibility a *Cartesian scenario of induction*, for it can be (but need not be) realized by a Cartesian-like demon who always hides nonblack ravens from the inquirer's sight. Each Cartesian scenario of induction has a *normal counterpart*, in which the inquirer receives exactly the same data in time (without observing a nonblack raven) and, fortunately, all ravens are black. A Cartesian scenario of induction and its normal counterpart are empirically indistinguishable, but in the first it is false that all ravens are black, while in the second it is true. And that causes trouble for both formal learning theorists and Bayesians.

For Bayesians, justification of full enumerative induction requires justifying a prior probability distribution that strongly disfavors a Cartesian scenario of induction and favors its normal counterpart. Carnap and other Bayesian confirmation theorists seem to never mention such an anti-Cartesian prior in their papers, possibly because they cannot justify it or simply because they do not care about it. A subjective Bayesian would say that such a prior is epistemically permissible, but only because she thinks that any probabilistic prior is epistemically permissible—*even* including those that strongly favor a Cartesian scenario of induction and will thereby lead to counterinduction:

> (COUNTERINDUCTION) We have observed this many ravens and they all are black; so, *not* all ravens are black.

So a subjective Bayesian must concede that counterinduction is epistemically permissible as well. Similarly, by considering certain priors that equally favor Cartesian scenarios of induction and their normal counterparts, subjective Bayesian must also concede that it is epistemically permissible to have a degree of belief in "All ravens are black" very close to 0.5 regardless of how many black ravens have been observed in a row—namely, that it is permissible to follow the skeptical policy of "no induction". To make those concessions explicit is to invite worries and complaints from those who would like to have a justification for full enumerative induction, *against* counterinduction, and *against* the inductive policy. That is probably why we seldom hear subjective Bayesians talk about full enumerative induction.

Formal learning theorists fare no better. When they justify the use of an inductive principle for tackling a certain empirical problem, they justify it as a necessary means of achieving a certain epistemic ideal for tackling that problem. The epistemic ideal they have in mind is called *identification in the limit* (Gold 1965 and Putnam 1965), which is a logical guarantee to find the true hypothesis in *every* possible way for the inquiry to unfold indefinitely, by virtue of following a learning method. When applied to the problem of whether all ravens are black, identification in the limit sets an extremely high standard: finding the truth both in a Cartesian scenario and in its normal counterpart. So, that standard is too high to be achieved by any learning method (see proposition 3.5 below).[5] Where it is impossible to achieve identification in the limit, nothing can be justified as a necessary means for achieving that.

So those are the reasons why normative/evaluative studies of inductive inference have said little about full enumerative induction. And we are left with the problem of giving a justification for full enumerative induction, against counterinduction, and against the skeptical policy of "no induction". I call this problem the *Cartesian problem of induction* because of the role played by Cartesian scenarios of the sort mentioned above. The point of pursuing this problem is not to respond to every conceivable kind of inductive skeptic, but to push ourselves to the limit—to explore the extent to which we can justify full enumerative induction.

The present paper aims to take a first step toward the first positive solution to that problem. The key to my proposal is that Bayesians can and should be *learning-theoretic*, and learning theorists should have been *truly* learning-theoretic, true to what I identify as the spirit of learning theory. Here is the idea:

1. (Mathematics) There are various modes of convergence to the truth. For example, formal learning theory studies identification in the limit, which may be called *everywhere* convergence to the truth, for it quantifies over all possible ways for the inquiry to unfold indefinitely. Statistics studies modes of stochastic convergence, such as *almost sure* convergence to the truth.

2. (Epistemology) Certain modes of convergence to the truth are epistemic ideals for an inquirer to achieve where possible. Some modes or their combinations correspond to higher epistemic ideals than some others do. We, as inquirers, ought to *achieve the best we can* when tackling an empirical problem.[6]

3. (Crux) But for tackling the raven problem—i.e. the problem of whether all ravens are black—it is provably impossible to achieve everywhere convergence to the truth. So, to achieve the best we can, we should *look for what can be achieved*. How? Well, given that it is impossible to converge everywhere, let's try to see whether it is at least possible to converge "almost everywhere"—to converge in "almost all" possible ways for the inquiry to unfold indefinitely. And it should not be difficult to try, because topologists have worked out a geometric conception of "almost everywhere" and "almost all".

[5] The same impossibility result remains even if we resort to *partial* identification in the limit, a weakening of identification in the limit due to Osherson et al. (1986), which requires that, in each possible way for the inquiry to unfold indefinitely, the true hypothesis is output infinitely often and each false hypothesis is output at most finitely often.

[6] For precursors to this epistemological idea, see Reichenbach (1938), Kelly (1996), and Schulte (1996).

4. (Proposal) Let's try to define various modes of convergence to the truth, including "almost everywhere" convergence, and study their combinations, such as "almost everywhere" convergence plus "monotonic" convergence. Find the combinations that are achievable for tackling the raven problem. Then, from among the achievable combinations, identify the one that corresponds to the highest epistemic ideal. And check whether that ideal is achieved *only* by following full enumerative induction rather than counterinduction. If the answer is positive, then that methodological principle is justified as a necessary means of achieving the highest achievable epistemic ideal for tackling the raven problem. But is the answer positive?

5. (Result) Yes, according to corollary 6.6. And the story just told can be retold for Bayesians.

All this is done by holding on to the following guidelines of learning theory:

Look for what can be achieved.
Achieve the best you can.

Think and acting in the way just described is what I mean by being truly learning-theoretic. In fact, that was how Gold (1965) and Putnam (1965) created formal learning theory, one of the earliest branches of learning theory.[7] That is also how I am going to address the Cartesian problem of induction.

The philosophical view articulated above is what I call *learning-theoretic epistemology*, which has to be defended at greater length in a future work.[8] This paper is devoted to developing its logical and mathematical foundation, without which my presentation of that philosophical view would be mere hand-waving. That said, I will make several philosophical points to motivate the mathematical steps to be taken in this paper. I will also explain in detail how the mathematical results are meant to solve the Cartesian problem of induction, *assuming* learning-theoretic epistemology.

Roadmap: Section 2 implements the above proposal and provides a pictorial but very informative sketch of the main results. The minimal technical detailed for understanding the results are presented before the concluding sections, all in 6,000 words. Proofs, examples, open questions, and the details of the Bayesian results are presented in the very long appendix, which the reviewers need to read.

To declare the style in use: Emphasis is indicated by *italics*, while the terms to be defined are presented in **boldface**.

[7] They address certain mathematical or empirical problems that are impossible to solve with an effective procedure—or, in the terminology of this paper, impossible to solve by achieving *everywhere* convergence to the truth with *perfect monotonicity*. What they do is in effect to drop perfectly monotonic convergence and see whether it is at least possible to achieve everywhere convergence.

[8] But, for replies to some standard worries, see Kelly (2001, 2004) and Kelly and Glymour (2004), in which they respond to Keynes' worry that in the long run we are all dead, and Carnap's worry that no convergence criterion can constrain short-run inferential practices.

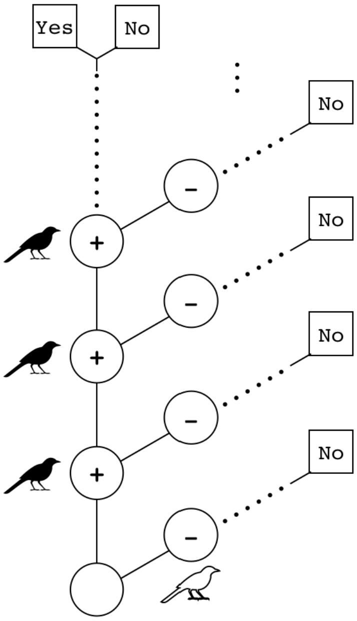

Figure 1: A partial representation of the hard raven problem

2 Pictorial Sketch of Main Results

This section sketches the key definitions and the main results, and presents their philosophical applications.

The problem of whether all ravens are black, which I call the **hard raven problem**, can be partially represented by the tree in figure 1. There are two competing hypotheses: `Yes` and `No`. The observation of a black raven is represented by a datum `+`; a nonblack raven, `-`. Observations of nonravens will be considered in subsequent sections but are omitted from the picture for simplicity. A data sequence (such as $(+, +, -)$) is a finite initial segment of a branch in the tree. A state of the world—i.e. a possible way for the inquiry to unfold indefinitely—is represented by an entire branch, which produces an infinite data stream (such as $(+, +, +, \ldots)$) and makes one of the competing hypotheses true (i.e. either `Yes` or `No`). The Cartesian scenarios of induction discussed in the previous section are represented by the vertical branch that makes No true. There are actually an infinity of Cartesian scenarios if observations of nonravens are considered.

In general, an empirical **problem** specifies a set of competing hypotheses and a set of possible finite data sequences (possibly with some presuppositions that come with the problem). A **learning method** for that problem is a mapping that sends each finite data sequence to one of the competing hypotheses or to the question mark that represents suspension of judgment. A learning method is evaluated in terms of its truth-finding performance in each state contained in a **state space** $\mathcal{S}$, which

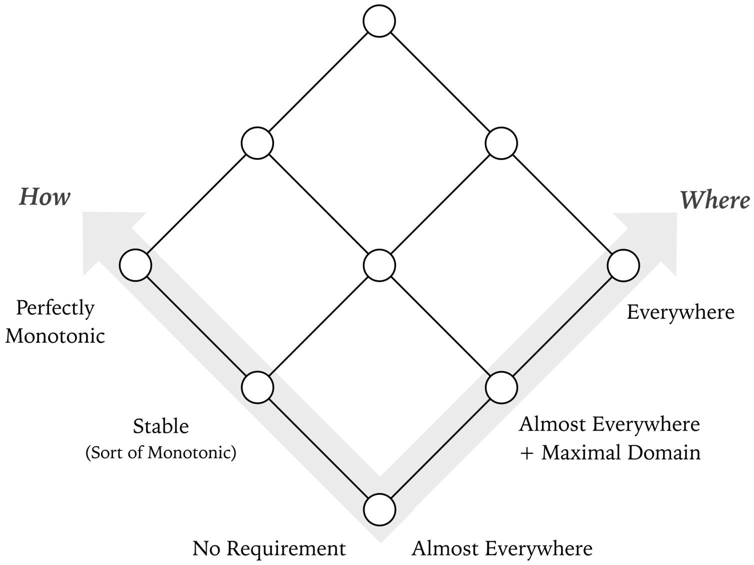

Figure 2: Modes of convergence to the truth, arranged by two dimensions

consists of all possible ways for the inquiry to unfold indefinitely (without violating the presuppositions of the problem under discussion). Each of those states makes exactly one of the competing hypotheses true and produces an infinite data stream $(e_1, e_2, e_3, \ldots)$, to be processed incrementally by a learning method M to output a sequence of conjectures $M(e_1), M(e_1, e_2), M(e_1, e_2, e_3), \ldots$ in time.

A learning method M is said to **converge to the truth** in a state $s \in \mathcal{S}$ if, in state s, method M will eventually output the true hypothesis and then always continue to do so. To achieve **everywhere** convergence to the truth is to achieve convergence to the truth in *all* states in state space $\mathcal{S}$. This convergence criterion is what formal learning theorists call *identification in the limit.* This is only one of the many convergence criteria that concern the question of *where convergence happens.*

Let's consider the question of *where convergence happens* together with that of *how convergence happens.* Have a look at figure 2, in which various modes of convergence to the truth are arranged by two dimensions. The dimension that stretches to the upper right concerns where. The other dimension, which stretches to the upper left, concerns how. I introduce three modes for each of the two dimensions, so in combination there are nine modes to be considered in this paper. (Modes of stochastic convergence will not be considered because they are irrelevant to full enumerative induction.) Now I turn to explaining those modes of convergence.

A learning method M for a problem is said to achieve **almost everywhere** convergence to the truth if, for every competing hypothesis h considered in the problem, M converges to the truth in "almost all" states that make h true—or speaking geometrically, M converges to the truth "almost everywhere" on the topo-

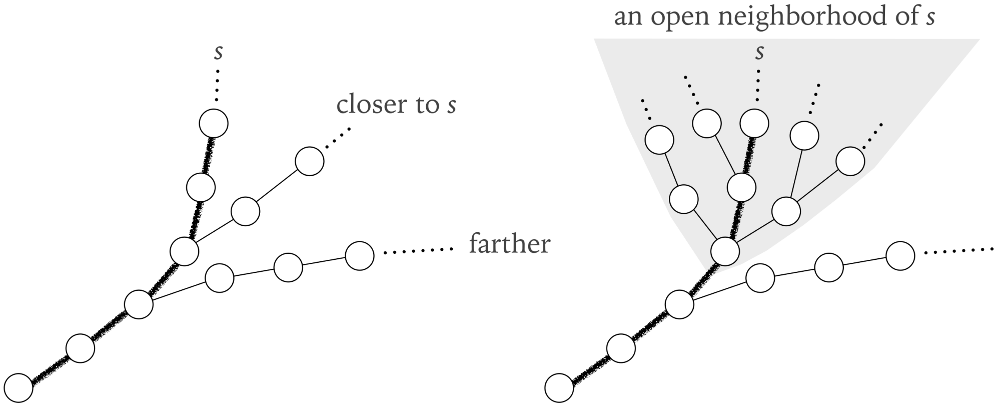

Figure 3: The empirical topology on a space of data streams or states

logical space of the states that make h true. “**Almost everywhere**” is defined in a quite standard way in geometry and topology; it means “everywhere except on a region that is negligible, i.e. nowhere dense.” A more satisfactory formal definition will be provided shortly. But, intuitively, we can think of this as converging everywhere except on a region that, like a slice of “hyper” Swiss cheese, is incredibly full of holes. But which topological structure to use? I adopt what may be called **empirical topology**,[9] according to which:

(i) an open neighborhood of a state s is characterized as the set of states that are “close” to s to a certain degree (as depicted on the righthand side of figure 3);

(ii) a state “closer” to s is one that requires more data to empirically distinguish from s (as depicted on the lefthand side of figure 3).

It is a lot easier to explain the other modes of convergence. A learning method is said to converge to the truth on a **maximal domain** if there exists no learning method that converges to the truth in the same states and in strictly more states.

A learning method is said to achieve **perfectly monotonic** convergence to the truth if, whenever it outputs one of the competing hypotheses, it outputs the truth and will continue to have the same output regardless of any further data received—it basically halts, just like an effective problem-solving method as studied in computability theory.

To be stable is to be “sort of” monotonic but not necessarily perfectly so. Specifically, a learning method is said to achieve **stable** convergence to the truth if, whenever it outputs a *true* competing hypothesis, it will continue to have the same out-

[9] Empirical topology is proposed by Vickers (1989) in computer science and by Kelly (1996) in epistemology.

put.[10] (What if it outputs a falsehood? Stable convergence is silent about this case.) So this mode of convergence corresponds to the intuitive condition that, whenever the inquirer forms a belief in the true competing hypothesis, this belief is not merely a true opinion but has been "stabilized" or "tethered" to the truth, attaining the epistemic status that Plato values in *Meno*. Finally, by "no requirement" I mean no requirement on how to converge.

The above finishes the sketch of the three modes of convergence on each of the two axes in figure 2. So there are nine "combined" modes of convergence arranged into a two-dimensional lattice structure, in which some modes are ordered higher than some others. A mode, if ordered higher, is mathematically stronger; it implies all the modes ordered lower in the lattice. That is mathematics. The following is epistemology. I make the evaluative assumption that:

> A mode of convergence to the truth, if ordered higher in the lattice in figure 2, corresponds to a higher epistemic ideal.

In fact, I even think that this assumption is obvious—or will become so after the definitions involved are rigorously stated and fully understood.[11]

The first main result is theorem 6.5 together with corollary 6.6, which says that, for tackling the hard raven problem, the achievable modes of convergence to the truth are the four in the shaded area of figure 4. So the highest achievable mode is the one marked by a star: "almost everywhere" + "maximal domain" + "stable". Furthermore, it can be achieved *only* by learning methods that implement full enumerative induction rather than counterinduction. This result relies on a crucial lemma—lemma 4.3—which says that, within the topological space of the states that make it false that all ravens are black, the set of the Cartesian scenarios is one of the (many) negligible regions.

Thanks to the above result, learning-theoretic epistemology can provide a justification for full enumerative induction, against counterinduction, and against the skeptical policy of "no induction". The justification consists in the following argument:

[10] Stable convergence to the truth is closely related to some properties that have been studied in learning theory, such as: Putnam's (1965) and Schulte's (1996) "mind-change"; Kelly and Glymour's (2004) "retraction"; Kelly, Genin, and Lin's (2016) "cycle". But these properties are defined only in terms of belief change without reference to truth. Stable convergence to the truth is a variant of the "no U-shaped learning" condition studied in Carlucci et al. (2005) and Carlucci et al. (2013), where U-shaped learning means the three-step process of believing in the truth, retracting it later, and then believing in the truth again.

[11] You might ask: When we compare two modes that are not ordered in the lattice—such that neither implies the other—how can we tell which corresponds to a higher epistemic ideal? See appendix D.5 for an interesting case study on this issue.

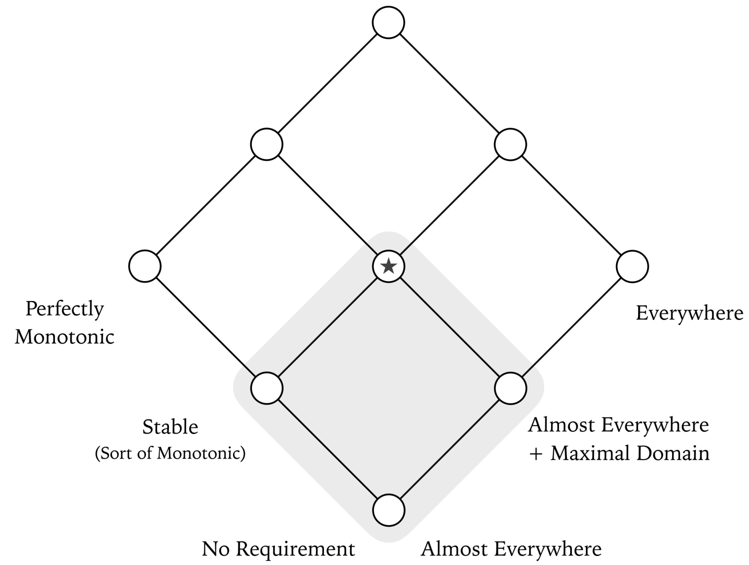

Figure 4: Modes of convergence achievable for the hard raven problem

Learning-Theoretic Argument

1. (Evaluative Premise) A mode of convergence to the truth, if ordered higher in the lattice in figure 2, corresponds to a higher epistemic ideal.

2. (Normative Premise) Given that an inquirer tackles a problem, she ought to follow a learning method that achieves, of the nine epistemic ideals in that lattice, the highest one achievable for tackling that problem—at least when such a highest one exists uniquely.

3. (Mathematical Premise) For tackling the hard raven problem, the achievable modes in that lattice are the four in the shaded area depicted in figure 4. (By corollary 6.6.)

4. So, given that an inquirer tackles the hard raven problem, she ought to follow a learning method that achieves the starred mode: "almost everywhere" + "maximal domain" + "stable". (By 1, 2, and 3.)

5. (Mathematical Premise) But that mode is achieved only by learning methods that implement full enumerative induction rather than counterinduction or the skeptical policy. (By corollary 6.6.)

6. *Therefore*, given that an inquirer tackles that problem, she ought to follow one of those inductive learning methods. (By 4 and 5.)

This is a *deductively valid* argument for a *normative conclusion* about induction, *only* with premises that are mathematical, evaluative, or normative, *free* from any empirical premise such as the principle saying that nature is uniform.[12] This argument applies a general normative framework—premises 1 and 2—to the inquirers who tackle the hard raven problem. For those inquirers, premise 3 finds what can be achieved, step 4 identifies the best that can be achieved, and premise 5 points to a necessary means for achieving that: following one of the learning methods that implement full enumerative induction rather than counterinduction or the skeptical policy.

But there are many such learning methods. Exactly which one(s) to follow? Any of them, or only some, or only a unique one? On this issue the above argument is silent, although we might be able to refine it and argue for a stronger norm if we introduce *additional* modes of convergence to the truth—a task that will not be attempted in the present paper. That said, the above argument seems to represent significant progress in justifying full enumerative induction.

The second main result is theorem A.3 (presented in appendix A), and one of its consequences is that, for tackling *any* problem, Ockham's razor of a certain kind is a necessary means for achieving any mode of convergence to the truth that implies "almost everywhere" + "stable", viz. any of those in the shaded area in figure 5.[13] This kind of Ockham's razor says: "Do not accept a competing hypothesis more complicated than necessary for fitting the data you have", where a competing hypothesis is simpler (less complicated) if it is more parsimonious in terms of the capacity to fit data. In light of this result, a learning-theoretic epistemologist can argue that, given that an inquirer tackles a problem for which the highest achievable mode of convergence to the truth implies "almost everywhere" + "stable", she ought to comply with the kind of Ockham's razor just mentioned.

The connection between the two main results is that, as we will see, any counterinductive inference violates that kind of Ockham's razor. In fact, a part of the first main result is proved as a corollary of the second main result.

The results sketched above all have Bayesian versions (presented in appendix B). This should not be surprising: just switch from "convergence of conjectured hypotheses to the truth" to "convergence of credences via conditionalization to full certainty in the truth", and modify the relevant definitions accordingly. Those who believe in radically subjective Bayesianism would not care, but some other

[12] I believe that these features of the argument (as indicated by italics) are key to escaping from Hume's dilemma. The strategy is to defend and develop Reichenbach's idea that, *pace* Hume, a justification of induction need not be an argument for an empirical thesis about the uniformity of nature but can be an argument for a normative/evaluative thesis about how to pursue an inquiry (Reichenbach 1938: sec. 39). See my unpublished manuscript BLIND, available upon request.

[13] This theorem, A.3, extends and strengthens some aspects of the main results of Kelly, Genin, and Lin (2016) in order to cover the hard raven problem and the like. But this theorem also simplifies and weakens some other aspects in order to highlight the core idea.

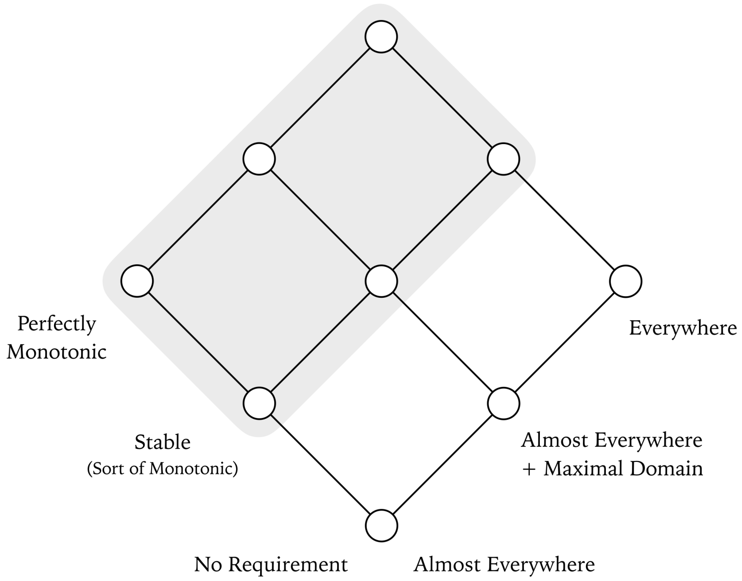

Figure 5: Those that can only be achieved by following Ockham's razor

Bayesians would and should. According to what I call *learning-theoretic Bayesianism*, the Bayesian inquirers who tackle a problem can be evaluated as good or bad inquirers for finding the truth among the competing hypotheses, depending partly on their probabilistic priors. For example, a Bayesian inquirer can be evaluated as a good inquirer for tackling a certain problem only if her probabilistic belief converges via conditionalization to full certainty in the true hypothesis everywhere (or almost everywhere)—provided that, of course, that epistemic ideal is achievable for tackling that problem. Learning-theoretic Bayesianism also evaluates inquirers (or their priors) in terms of other modes of convergence; for example, modes of stochastic convergence have to be considered in order to evaluate inquirers who tackle statistical problems. It seems to me that many Bayesians are at least sometimes learning-theoretic rather than radically subjective. Indeed, Bayesian statisticians caution against assigning a zero prior to a statistical hypothesis under discussion. And one of their typical reasons is that doing so will make it impossible to converge stochastically to full certainty in the truth if that statistical hypothesis is true. This reason is learning-theoretic in nature. It is just that, now, the modeling of belief is not qualitative but probabilistic.

The above summarizes the main results and the way they are used to solve the Cartesian problem of induction, *assuming* learning-theoretic epistemology. The rest of this paper is devoted to the mathematical details.

3 Preliminaries: Learning Theory

Definition 3.1. A **problem** is a triple $\mathcal{P} = \big(\mathcal{H}, \mathcal{E}, \mathcal{S}\big)$ consisting of:

- a hypothesis space $\mathcal{H}$, which is a set of competing **hypotheses**,
- an evidence space $\mathcal{E}$, which is a set of finite **data sequences** $(e_1, \ldots, e_n)$,
- a state space $\mathcal{S}$, which is a set of possible **states** of the world taking this form: $(h, \vec{e})$, where:
    - h, called the uniquely **true** hypothesis in this state, is an element of $\mathcal{H}$;
    - $\vec{e}$, called the **data stream** produced in this state, is an infinite sequence of data, written $\vec{e} = (e_1, e_2, e_3, \ldots)$, whose finite initial segments $(e_1, \ldots, e_n)$ are all in $\mathcal{E}$.

The state space $\mathcal{S}$ of a problem $\big(\mathcal{H}, \mathcal{E}, \mathcal{S}\big)$ is meant to capture the *presupposition* of that problem in this way: $\mathcal{S}$ consists of all possible ways for the inquiry to unfold indefinitely without violating the presupposition.

Definition 3.2. A **learning method** for a problem $\big(\mathcal{H}, \mathcal{E}, \mathcal{S}\big)$ is a function:

$$M : \mathcal{E} \to \mathcal{H} \cup \{\texttt{?}\},$$

where `?` represents suspension of judgment. Given each data sequence $(e_1, \ldots, e_n) \in \mathcal{E}$, the output of M is written as $M(e_1, \ldots, e_n)$.

Example 3.3. The **hard raven problem** poses this question: *"Are all ravens black?"* This problem is partially represented by the tree structure in figure 1 and is formally defined as follows.

- The hypothesis space $\mathcal{H}$ is $\{\texttt{Yes}, \texttt{No}\}$, where:
    - `Yes` means that all ravens are black,
    - `No` means that not all ravens are black.
- The evidence space $\mathcal{E}$ consists of all finite sequences of `+`, `0`, and/or `-`, where:
    - datum `+` denotes the observation of a black raven;
    - datum `-`, a nonblack raven;
    - datum `0`, a nonraven.
- The state space $\mathcal{S}$ consists of all states in one of the following three categories:[14]

[14] Well, there is a fourth category: the states $(\texttt{Yes}, \vec{e})$ in which $\vec{e}$ contains some occurrence of `-` (a nonblack raven). But such states are logically impossible, so they need not be considered.

(a) the states ($\texttt{Yes}, \vec{e}$) in which $\vec{e}$ is an infinite `+/0` sequence (namely, an infinite sequence containing only `+` and `0` observations).

(b) the states ($\texttt{No}, \vec{e}$) in which $\vec{e}$ is an infinite `+/0` sequence.

(c) the states ($\texttt{No}, \vec{e}$) in which $\vec{e}$ is an infinite `+/0/-` sequence that contains at least one occurrence of `-`.

The second category (b) contains the states in which there are nonblack ravens but the inquirer will never observe one, so they are the **Cartesian scenarios of induction**.

Definition 3.4. Let M be a method for a problem $\mathcal{P} = \big(\mathcal{H}, \mathcal{E}, \mathcal{S}\big)$. M is said to **converge to the truth** in a state $(h, \vec{e}) \in \mathcal{S}$ if

$$\lim_{n \to \infty} M(e_1, \ldots, e_n) \;=\; h\,,$$

namely, there exists a positive integer k such that, for each $n \geq k$, we have that $M(e_1, \ldots, e_n) = h$. M is said to converge to the truth **everywhere** for $\mathcal{P} = \big(\mathcal{H}, \mathcal{E}, \mathcal{S}\big)$ if it converges to the truth in every state contained in $\mathcal{S}$.

Proposition 3.5. *The hard raven problem has no learning method that converges to the truth everywhere.*

4 Preliminaries: Topology

Let a space X of data streams be given. Choose an arbitrary data stream $\vec{e}$ therein, and take it as the actual data stream to be received incrementally by the inquirer, as depicted on the lefthand side of figure 3. Consider an alternative data stream $\vec{e}\,'$ that is **identical** to $\vec{e}$ **up until** stage n; namely, $e'_i = e_i$ for each $i \leq n$ but $e'_{n+1} \neq e_{n+1}$. The larger n is, the later the point of departure is and the more data one needs to distinguish those two data streams. So, the larger n is, the harder it is to empirically distinguish those two data streams, and the "closer" $\vec{e}\,'$ is to the actual data stream $\vec{e}$—"closer" in an empirical sense. Consider the set of the data streams that are at least "that close" to $\vec{e}$:

$$N_n(\vec{e}) \;=\; \{\vec{e}\,' \in X : e'_i = e_i \text{ for each } i \leq n\}.$$

Take that as a *basic open neighborhood* of point $\vec{e}$ in space X, as depicted on the righthand side of figure 3. Such open neighborhoods provably form a *topological base* of X.[15]

[15] Here is the standard definition of topological bases. Given a set X of points, a family $\mathcal{B}$ of subsets of X is called a **topological base** if (i) every point in X is contained in some set in $\mathcal{B}$ and (ii) for any $B_1, B_2 \in \mathcal{B}$ and any point $x \in B_1 \cap B_2$, there exists $B_3 \in \mathcal{B}$ such that $x \in B_3 \subseteq B_1 \cap B_2$.

Similarly, given the space $|h|$ of states that make a certain hypothesis h true, two states in $|h|$ are close (hard to distinguish empirically) iff the data streams therein are close. So a basic open neighborhood of a state $s = (h, \vec{e})$ in $|h|$ takes the following form:

$$\begin{aligned} N_n\Big((h, \vec{e})\Big) &= \{(h, \vec{e}\,') \in |h| : e'_i = e_i \text{ for each } i \le n\} \\ &= |h| \cap |(e_1, \ldots, e_n)| \,. \end{aligned}$$

Such neighborhoods provably form a topological base of $|h|$, turning $|h|$ into a topological space. This is the topological base we will use. It is determined by the empirical distinguishability between states—distinguishability in terms of (finite) data sequences.

Definition 4.1. Given a problem $\mathcal{P} = \big(\mathcal{H}, \mathcal{E}, \mathcal{S}\big)$ and a hypothesis $h \in \mathcal{H}$, the **empirical topological base** of $|h|$ is the family of open neighborhoods constructed above; namely, it is defined as follows:

$$\begin{aligned} \mathcal{B}_{|h|} &= \Big\{N_n(s) : s \in |h| \text{ and } n \in \mathbb{N}^+\Big\} \\ &= \Big\{|h| \cap |(e_1, \ldots, e_n)| : (e_1, \ldots, e_n) \in \mathcal{E}\Big\} \smallsetminus \Big\{\varnothing\Big\} \,. \end{aligned}$$

I now turn to some concepts borrowed from general topology.

Definition 4.2. Let X be a topological space equipped with a topological base $\mathcal{B}_X$. A **negligible** (or **nowhere dense**) region within X is a subset R of X such that, for each nonempty open neighborhood $N \in \mathcal{B}_X$, there exists a nonempty open neighborhood $N' \in \mathcal{B}_X$ that is nested within N and disjoint from R.

A negligible region is like a slice of "hyper" Swiss cheese, incredibly full of holes: wherever you are in the ambient space X, say point x, and however small a basic open neighborhood of x is considered, say N, then within N you can always find an open "hole" N' of that slice of Swiss cheese. Here is an example:

Lemma 4.3. *In the hard raven problem, the Cartesian scenarios of induction (i.e. the states* $(\texttt{No}, \vec{e})$ *with* $\vec{e}$ *being a* +/0 *sequence) form a negligible region within the topological space* $|\texttt{No}|$*.*

Proof. Each nonempty basic open neighborhood in the topological space $|\texttt{No}|$, say $N = |\texttt{No}| \cap |(e_1, \ldots, e_n)|$, includes a nonempty basic open neighborhood, namely $N' = |\texttt{No}| \cap |(e_1, \ldots, e_n, \texttt{-})|$, which is disjoint from the set of the Cartesian scenarios of induction. □

With "negligible", we can define "almost everywhere" and "almost all" the standard way in topology:

Definition 4.4. Consider a property of points in a topological space X. That property is said to apply **almost everywhere** on space X if it applies to all points in $X \smallsetminus X'$, where X' is some negligible region within X. In that case, also say that it applies to **almost all** points in space X.

5 Mode (I): "Almost Everywhere"

We are finally in a position to define the key concept that kickstarts the new learning theory:

Definition 5.1. A learning method M for a problem $\mathcal{P} = \big(\mathcal{H}, \mathcal{E}, \mathcal{S}\big)$ is said to converge to the truth **almost everywhere** if, for each hypothesis $h \in \mathcal{H}$, M converges to the truth in almost all states that make h true—or speaking geometrically, M converges to the truth almost everywhere on topological space $|h|$.

This definition makes possible a series of positive results. Here is the first one:

Proposition 5.2. *The hard raven problem has a learning method that converges to the truth almost everywhere.*

Proof. By the preceding result, lemma 4.3, the topological space $|\texttt{No}|$ has the following negligible region:

$$C \;=\; \{(\texttt{No}, \vec{e}) : \vec{e} \text{ is a } \texttt{+/0} \text{ sequence}\},$$

which consists of the *C*artesian scenarios of induction. So it suffices to an example of a learning method that converges to the truth in:

- every state in $|\texttt{Yes}|$,
- every state in $|\texttt{No}| \smallsetminus C$.

The following learning method does the job:

M^*: "Output hypothesis `No` if you have observed a nonblack raven (`-`); otherwise output `Yes`."

This finishes the proof. □

Despite the above positive result, there are problems for which it is impossible to achieve almost everywhere convergence to the truth. Examples are provided in appendix D.2; their philosophical implications are briefly discussed in appendix D.3.

Well, the above is only a first step toward justifying full enumerative induction. For there remains a subproblem, which may be called the *problem of counterinduction*: Almost everywhere convergence to the truth, alone, is so liberal that it is witnessed also by some crazy learning methods that apply counterinduction. Here is an example:

$M^\dagger$: “Output hypothesis `No` if you have observed a nonblack raven (`-`) or everything you have observed is a black raven (`+`); output `Yes` if every raven you have observed is black and you have observed a nonraven (`0`).”

This method converges to the truth almost everywhere for the hard raven problem because it converges to the truth in:

- every state in $|\texttt{Yes}| \smallsetminus \{s\}$,
  where $s = (\texttt{Yes}, \text{the constant sequence of } \texttt{+})$,
- every state in $|\texttt{No}| \smallsetminus (C \smallsetminus \{s'\})$,
  where $s' = (\texttt{No}, \text{the constant sequence of } \texttt{+})$.

$\{s\}$ is a negligible region within $|\texttt{Yes}|$; $C \smallsetminus \{s'\}$, within $|\texttt{No}|$.

The learning method $M^\dagger$ defined above applies counterinduction occassionally. It will be ruled out by the modes of convergence to be introduced below.

6 Modes (II) and (III): “Stable” and “Maximal”

Definition 6.1. A learning method M is said to **have converged** to the truth given the n-th stage of inquiry in a state $s = (h, \vec{e})$ if

$$M(e_1, \ldots, e_k) \quad = \quad h \quad \text{for each } k \geq n\,.$$

With the above concept we can define the following two epistemic ideals:

Definition 6.2. A learning method M for a problem $\mathcal{P} = \big(\mathcal{H}, \mathcal{E}, \mathcal{S}\big)$ is said to converge to the truth with **perfect monotonicity** if

> in any state $s = (h, \vec{e})$, given any stage n such that $M(e_1, \ldots, e_n) \neq \texttt{?}$, M has converged to the truth.

Say that M converges to the truth with **stability** if the following (weaker) condition holds:

> in any state $s = (h, \vec{e})$, given any stage n such that $M(e_1, \ldots, e_n) = h$ (i.e. the truth in s), M has converged to the truth.

Stable convergence is sort of monotonic but not necessarily perfectly so, while perfect monotonicity can be very demanding. We have this negative result:

Proposition 6.3. *For the hard raven problem, it is impossible to simultaneously achieve the following two modes of convergence to the truth: “almost everywhere” and “perfectly monotonic”.*

And the following is the last mode of convergence needed to state the first main result:

Definition 6.4. A learning method M for a problem is said to converge to the truth on a **maximal domain** if there is no learning method for the same problem that converges to the truth in all states where M does and in strictly more states.[16]

Then we have the first main result:

Theorem 6.5. *The hard raven problem has a learning method that converges to the truth (i) almost everywhere, (ii) on a maximal domain, and (iii) with stability. Every such learning method M has the following properties:*

1. *M is* **never counterinductive** *in that, for any data sequence $(e_1, \ldots, e_n)$ that has not witnessed a nonblack raven, $M(e_1, \ldots, e_n) \neq$* `No`*;*

2. *M is* **enumeratively inductive** *in that, for any data stream $\vec{e}$ that never witnesses a nonblack raven, $M(e_1, \ldots, e_n)$ converges to* `Yes` *as $n \to \infty$.*

The idea that underlies the proof of the above theorem is explained in appendix C.1, which I have tried to make as instructive as possible. You do not want to miss it if you are interested in how exactly stable convergence helps to argue against counterinduction. The proof itself is in appendix C.3.

Corollary 6.6. *Consider the modes of convergence to the truth arranged in the lattice in figure 4. The four modes in the shaded area are exactly those achievable for the hard raven problem. To achieve the strongest of those four, namely "almost everywhere" + "maximal" + "stable", a necessary means is to follow one of the learning methods that implement full enumerative induction, rather than counterinduction or the skeptical policy.*

This corollary follows immediately from preceding results: propositions 3.5 and 6.3 and theorem 6.5.

7 Conclusion

I articulate what I take to be truly learning-theoretic and formulate the core ideas of learning-theoretic epistemology, which I employ to develop a solution to the Cartesian problem of induction. This seems to me the only thoroughly developed solution we have at the time of writing this paper. I hope the present work will help to spark the development of competing solutions for future comparison. But, hey Bayesians, you do not have to work out a new solution—please have a look at appendix B for a solution in learning-theoretic Bayesianism.

[16] I am indebted to BLIND for bringing this concept to my attention.

8 References

Baltag, A., Gierasimczuk, N., and Smets, S. (2015) "On the Solvability of Inductive Problems: A Study in Epistemic Topology", in Ramanujam, R. (Ed.) *Proceedings of the 15th Conference on Theoretical Aspects of Rationality and Knowledge (TARK-2015)*, ACM 2015 (ILLC Prepublication Series PP-2015-13).

Carlucci, L. and Case, J. (2013) "On the Necessity of U-Shaped Learning", *Topics in Cognitive Science*, 5(1): 56-88.

Carlucci, L., Case, J., Jain, S., and Stephan, F. (2005) "Non U-Shaped Vacillatory and Team Learning", *Algorithmic Learning Theory*, 241-255, Springer Berlin Heidelberg.

Gold, E. M. (1965) "Limiting Recursion", *Journal of Symbolic Logic*, 30(1): 27-48.

Gold, E. M. (1967) "Language identification in the limit", *Information and Control*, 10(5): 447-474.

Hintikka, J. (1966), "A Two-Dimensional Continuum of Inductive Methods", in J. Hintikka and P. Suppes (eds.) *Aspects of Inductive Logic*, Amsterdam: North-Holland.

Hintikka, J., and I. Niiniluoto (1980), "An Axiomatic Foundation for the Logic of Inductive Generalization", in Jeffrey, R. (ed.) *Studies in Inductive Logic and Probability*, vol. 2. Berkeley and Los Angeles: University of California Press.

Hume, D. (1777) *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*, reprinted and edited with introduction, comparative table of contents, and analytical index by Bigge, L.A. Selby (1975), MA. Third edition with text revised and notes by P. H. Nidditch. Oxford, Clarendon Press.

Kelly, K. T. (1996) *The Logic of Reliable Inquiry*, Oxford: Oxford University Press.

Kelly, K. T. (2001) "The Logic of Success", *the British Journal for the Philosophy of Science*, Special Millennium Issue 51: 639-666.

Kelly, K. T. (2004) "Learning Theory and Epistemology", in *Handbook of Epistemology*, I. Niiniluoto, M. Sintonen, and J. Smolenski, (eds.) Dordrecht: Kluwer.

Kelly, K. T. and C. Glymour (2004) "Why Probability Does Not Capture the Logic of Scientific Justification", in C. Hitchcock (ed.) *Contemporary Debates in the Philosophy of Science*, London: Blackwell.

Kelly, T. K, K. Genin, and H. Lin (2016) "Realism, Rhetoric, and Reliability", *Synthese* 193(4): 1191-1223.

Osherson, D., S. Micheal, and S. Weinstein (1986) *Systems that Learn: An Introduction to Learning Theory for Cognitive and Computer Scientists*, MIT Press.

Oxtoby, J. C. (1996) *Measure and Category: A Survey of the Analogies between Topological and Measure Spaces*, 2nd Edition, Springer.

Putnam, H. (1963) "Degree of Confirmation and Inductive Logic", in Schilpp, P. A. (ed.) *The Philosophy of Rudolf Carnap*, La Salle, Ill: Open Court.

Putnam, H. (1965) "Trial and Error Predicates and a Solution to a Problem of Mostowski", *Journal of Symbolic Logic*, 30(1): 49-57.

Reichenbach, H. (1938) *Experience and Prediction: An Analysis of the Foundation and the Structure of Knowledge*, Chicago, University of Chicago Press.

Schulte, O. (1996) "Means-Ends Epistemology", *The British Journal for the Philosophy of Science* 79(1), 141147.

Spirtes, P., C. Glymour, and R. Scheines (2000) *Causation, Prediction, and Search*, 2nd Edition, the M.I.T. Press.

Vickers, S. (1989) *Topology via Logic*, Cambridge University Press, Cambridge.

A "Almost Everywhere" + "Stable" $\Longrightarrow$ "Ockham"

I have presented a result (corollary 6.6) that can be used to justify a norm that says when to follow this methodological principle:

Do not accept a counterinductive hypothesis.

(When? At least when tackling the hard raven problem.) This section presents a similar result, which can be used to justify a norm that says when to follow this methodological principle:

Do not accept a hypothesis if it is more
complicated than necessary for fitting data.

This is Ockham's razor of a certain kind, where a hypothesis is simpler iff it is parsimonious in terms of the capacity to fit data.

To be more precise:

Definition A.1. Let a problem $\mathcal{P} = \big(\mathcal{H}, \mathcal{E}, \mathcal{S}\big)$ be given. A data sequence $(e_1, \ldots, e_n)$ and a hypothesis h therein are said to be **compatible** if the propositions they express have a nonempty overlap, which also means that there exists a state in $\mathcal{S}$ that makes hypothesis h true and produces data sequence $(e_1, \ldots, e_n)$. For each hypothesis $h \in \mathcal{H}$, let $\mathcal{E}(h)$ denote the set of data sequences in $\mathcal{E}$ that are compatible with h (so $\mathcal{E}(h)$ captures the data-fitting capacity of h). The **empirical simplicity order**, written $\prec$, is defined on $\mathcal{H}$ as follows: for all hypotheses h and $h' \in \mathcal{H}$,

$$h \prec h' \quad \text{iff} \quad \mathcal{E}(h) \subset \mathcal{E}(h').$$

Or in words, h is **simpler** then h' iff h "fits" strictly less data sequences than h' does. Say that h is **no more complex** than h' if $h' \not\prec h$.

In the hard raven problem, for example, the inductive hypothesis `Yes` is simpler than the counterinductive hypothesis `No`.

Definition A.2. A learning method M for a problem $\mathcal{P} = \big(\mathcal{H}, \mathcal{E}, \mathcal{S}\big)$ is said to follow **Ockham's tenacious razor** just in case, for each hypothesis $h \in \mathcal{H}$ and each data sequence $(e_1, \ldots, e_n) \in \mathcal{E}$, h is the output of M given $(e_1, \ldots, e_n)$ only if

- (*Razor Condition*) h is no more complex than any hypothesis in $\mathcal{H}$ that is compatible with $(e_1, \ldots, e_n)$;
- (*Tenacity Condition*) h continues to be the output of M given any data sequence in $\mathcal{E}$ that extends $(e_1, \ldots, e_n)$ and is compatible with h.

In other words, a learning method M follows Ockham's tenacious razor just in case, whenever M outputs a hypothesis h, h is no more complex than necessary for fitting the available data and h will continue to be the output until it is refuted by the accumulated data. In the hard raven problem, to comply with the razor condition is exactly to be never counterinductive—to never infer `No` whenever one has not observed a nonblack raven.

Then we have the second main result:

Theorem A.3 (Ockham Stability Theorem)**.** *Let M be a learning method for a problem. Suppose that M converges to the truth almost everywhere. Then following two conditions are equivalent:*

1. *M converges to the truth with stability.*
2. *M follows Ockham's tenacious razor.*

I call it the *Ockham stability theorem.* I understand its epistemological significance as follows. Almost everywhere convergence to the truth is a fundamental epistemic ideal to strive for whenever it is achievable. Convergence with stability is also good epistemically, but without almost everywhere convergence, it is not clear what value there is in achieving only stability. So, almost everywhere convergence first, stable convergence second. Given almost everywhere convergence, we might want to strive (further) for stable convergence (if that is possible), and to achieve that is *exactly* to follow Ockham's tenacious razor, as stated in the above theorem.

An immediate application of the *1* ⇒ *2* side of the Ockham stability theorem is to prove, as a corollary, the "never be counterinductive" part of theorem 6.5. For, when tackling the hard raven problem, to be never counterinductive is exactly to comply with the razor condition. So the Ockham stability theorem helps to justify a local norm of Ockham's razor: Given that an inquirer tackles the hard raven problem, she ought to always follow Ockham's tenacious razor and, hence, never apply counterinduction.

We can use the preceding theorem to justify the use of Ockham's razor for tackling other problems, such as curve-fitting problems. See appendix D.4 for an example.

B The Story Retold in Bayesian Terms

Definition B.1. Let a problem $\mathcal{P} = \big(\mathcal{H}, \mathcal{E}, \mathcal{S}\big)$ be given. Subsets of state space $\mathcal{S}$ are called **propositions**. Hypothesis $h \in \mathcal{H}$ and data sequence $(e_1, \ldots, e_n) \in \mathcal{E}$ are understood to express the following propositions:

$$\begin{aligned} |h| &= \{(h', \vec{e}\,') \in \mathcal{S} : h' = h\}; \\ |(e_1, \ldots, e_n)| &= \{(h', \vec{e}\,') \in \mathcal{S} : \vec{e}\,' \text{ extends } (e_1, \ldots, e_n)\}. \end{aligned}$$

That is, $|h|$ is the set of states in $\mathcal{S}$ that make hypothesis h true, and $|(e_1, \ldots, e_n)|$ is the set of states in $\mathcal{S}$ that produce data sequence $(e_1, \ldots, e_n)$.[17] Let $\mathcal{A}_\mathcal{P}$ denote the smallest σ-algebra that contains the above propositions for all $h \in \mathcal{H}$ and all $(e_1, \ldots, e_n) \in \mathcal{E}$. Given a probability function $\mathbb{P}$ defined on that algebra, I will write $\mathbb{P}(h)$ as a shorthand for $\mathbb{P}(|h|)$. Similarly, I will write $\mathbb{P}(e_1, \ldots, e_n)$ and $\mathbb{P}(h \,|\, e_1, \ldots, e_n)$, where the latter stands for conditional probability as defined the standard way.[18]

Definition B.2. A **probabilistic prior** for a problem $\mathcal{P} = \big(\mathcal{H}, \mathcal{E}, \mathcal{S}\big)$ is a probability function $\mathbb{P}$ defined on σ-algebra $\mathcal{A}_\mathcal{P}$ with $\mathbb{P}(e_1, \ldots, e_n) > 0$ for each data stream $(e_1, \ldots, e_n) \in \mathcal{E}$. $\mathbb{P}$ is said to (have its posteriors) **converge to the truth** in a state $s = (h, \vec{e}) \in \mathcal{S}$ if

$$\lim_{n \to \infty} \mathbb{P}(h \,|\, e_1, \ldots, e_n) \;=\; 1\,,$$

[17] If you like, the concept of problems and other learning-theoretic concepts can be defined purely in terms of propositions, as done in Baltag et al. (2015) and Kelly et al. (2016).

[18] Namely, $\mathbb{P}(A \,|\, B) = \mathbb{P}(A \cap B)/\mathbb{P}(B)$.

that is, for any $\epsilon > 0$, there exists a positive integer k such that, for each $n \geq k$, $\mathbb{P}(h \,|\, e_1, \ldots, e_n) > 1 - \epsilon$. $\mathbb{P}$ is said to converge to the truth **everywhere** for problem $\big(\mathcal{H}, \mathcal{E}, \mathcal{S}\big)$ if it converges to the truth in each state in $\mathcal{S}$.

Proposition B.3. *The hard raven problem has no probabilistic prior that converges to the truth everywhere.*

Proof. Copy the proof of proposition 3.5, paste it here, replace 'learning method' by 'probabilistic prior', and replace 'M' by '$\mathbb{P}$'. □

Definition B.4. A probabilistic prior $\mathbb{P}$ is said to converge to the truth **almost everywhere** for a problem $\mathcal{P} = \big(\mathcal{H}, \mathcal{E}, \mathcal{S}\big)$ if, for each hypothesis $h \in \mathcal{H}$, $\mathbb{P}$ converges to the truth almost everywhere on topological space $|h|$.

Proposition B.5. *The hard raven problem has a probabilistic prior that converges to the truth almost everywhere.*

Proof. Immediate from theorem B.7, to be presented and proved below. □

Definition B.6. Let $\mathbb{P}$ be a probabilistic prior for a problem $\mathcal{P} = \big(\mathcal{H}, \mathcal{E}, \mathcal{S}\big)$. $\mathbb{P}$ is said to **have started to stably converge** to the truth given stage n in state $s = (h, \vec{e}) \in \mathcal{S}$ if

1. $\mathbb{P}(h \,|\, e_1, \ldots, e_n, \ldots, e_{n+i})$ is monotonically increasing as a function of i defined on $\mathbb{N}$,
2. $\mathbb{P}(h \,|\, e_1, \ldots, e_n, \ldots, e_{n+i})$ converges to 1 as $i \to \infty$.

$\mathbb{P}$ is said to converge to the truth with **stability** if, for each hypothesis $h \in \mathcal{H}$, for each state $s = (h, \vec{e}) \in \mathcal{S}$ that makes h true, and for each stage n as a positive integer, if $\mathbb{P}(h \,|\, e_1, \ldots, e_n) > 1/2$, then $\mathbb{P}$ has started to converge to the truth given stage n in state s.

Theorem B.7. *The hard raven problem has a probabilistic prior that converges to the truth (i) almost everywhere, (ii) on a maximal domain, and (iii) with stability. Every such probabilistic prior $\mathbb{P}$ has the following properties:*

1. *$\mathbb{P}$ is* **never counterinductive** *in that, for any data sequence $(e_1, \ldots, e_n)$ that has not witnessed a nonblack raven, $\mathbb{P}(\texttt{No} \,|\, e_1, \ldots, e_n) \leq 1/2$;*
2. *$\mathbb{P}$ is* **enumeratively inductive** *in that, for any data stream $\vec{e}$ that never witnesses a nonblack raven, $\mathbb{P}(\texttt{Yes} \,|\, e_1, \ldots, e_n)$ converges to 1 as $n \to \infty$.*

Definition B.8. A probabilistic prior $\mathbb{P}$ for a problem $\big(\mathcal{H}, \mathcal{E}, \mathcal{S}\big)$ is said to follow **Ockham's tenacious razor** just in case, for each hypothesis $h \in \mathcal{H}$ and each data sequence $(e_1, \ldots, e_n) \in \mathcal{E}$, $\mathbb{P}(h \,|\, e_1, \ldots, e_n) > 1/2$ only if

- (*Razor Condition*) h is no more complex than any hypothesis in $\mathcal{H}$ that is compatible with the given data sequence $(e_1, \ldots, e_n)$;
- (*Tenacity Condition*) for any data sequence $(e_1, \ldots, e_n, \ldots, e_{n+n'})$ in $\mathcal{E}$ that extends $(e_1, \ldots, e_n)$ and is compatible with h, $\mathbb{P}(h \,|\, e_1, \ldots, e_{n+i})$ is monotonically increasing as a function of $i \in \{0, \ldots, n'\}$.

Theorem B.9 (Ockham Stability Theorem, Bayesian Version)**.** *Let $\mathbb{P}$ be a probabilistic prior for a problem that converges to the truth almost everywhere. Then, condition 1 below implies condition 2 below (but the converse does not hold):*

1. $\mathbb{P}$ converges to the truth with stability.

2. $\mathbb{P}$ follows Ockham's tenacious razor.

Although the converse does not hold,[19] we might be able to formulate a stronger version of tenacity or a weaker version of stability in Bayesian terms in order to restore the equivalence between conditions *1* and *2*. But that will not be attempted here. For there is no loss in application to epistemology: to justify Ockham's razor, what is really needed is just the implication relation from the epistemic ideal expressed by *1* to the methodological principle expressed by *2*, which shows that the latter is a necessary means for achieving the former. The converse does no justificatory work. Showing that Ockham's razor achieves an epistemic ideal does not suffice to argue that one *has to* follow Ockham's razor, for there might be other means that also achieves the epistemic ideal.

C Proofs

C.1 The Idea of Proof of Theorem 6.5

Theorem 6.5 has three parts. The existential claim and the universal claim about "be enumeratively inductive" are two easier parts. The crucial part is the universal claim about "never be counterinductive". The reason why it is crucial is two-fold: first, it is a very instructive special case of the 1 ⇒ 2 side of the Ockham stability theorem A.3; second, it is a lemma for proving the part about "be enumeratively inductive". So let me separate the crucial part for a closer examination:

Proposition C.1 (Never Be Counterinductive)**.** *Let M be a learning method for the hard raven problem that converges to the truth almost everywhere with stability. Then M is never counterinductive.*

[19] Here is the reason why the converse does not hold: the tenacity condition—as defined in the Bayesian framework—only requires posterior probability to remain the same or go up as data accumulate, but does not require it to go up high enough to ensure convergence to 1, let along convergence with stability.

Note that we do not need convergence on a maximal domain here. It will be convenient to have the following concept:

Definition C.2. Given a problem $\big(\mathcal{H}, \mathcal{E}, \mathcal{S}\big)$, a data sequence $(e_1, \ldots, e_n) \in \mathcal{E}$ is said to be **compatible** with a hypothesis $h \in \mathcal{H}$ if the propositions they express have a nonempty overlap, namely:

$$|h| \cap |(e_1, \ldots, e_n)| \;\; \neq \;\; \varnothing\,,$$

which also means that $(e_1, \ldots, e_n)$ can be extended into a data stream $\vec{e}$ such that $(h, \vec{e})$ is a state in $\mathcal{S}$.

The proof of the above proposition proceeds as follows. Let M be a learning method for the hard raven problem that converges to the truth almost everywhere. Suppose that M is sometimes counterinductive, namely, for some `+/0` sequence $(e_1, \ldots, e_n)$, we have that:

$$M(e_1, \ldots, e_n) \;\; = \;\; \texttt{No}\,. \tag{1}$$

It suffices to show that M fails to converge to the truth with stability. Since $(e_1, \ldots, e_n)$ is a `+/0` sequence, it is compatible with `Yes`. To summarize, we have had:

- M converges to the truth almost everywhere.
- $(e_1, \ldots, e_n)$ is compatible with `Yes`.

Given these two conditions, we can apply the so-called *forcing lemma* (to be stated soon) in order to "force" M to output `Yes` by extending $(e_1, \ldots, e_n)$ into a certain `+/0` sequence $(e_1, \ldots, e_n, \ldots, e_{n'})$ such that:

$$M(e_1, \ldots, e_n, \ldots, e_{n'}) \;\; = \;\; \texttt{Yes}\,. \tag{2}$$

Now, choose a state s such that:

$$s \;\; \in \;\; |\texttt{No}| \cap |(e_1, \ldots, e_n, \ldots, e_{n'})|\,. \tag{3}$$

We can always make this choice because every data sequence is compatible with `No`. By (1)-(3), we have: given the earlier stage n in state s, M outputs the truth `No` but fails to have converged to the truth. So M does not converge to the truth with stability. This finishes the proof of the part "never be counterinductive" in theorem 6.5—as soon as the forcing lemma is stated and established.

Let me state the forcing lemma here, remark on its importance, and leave its proof to appendix C.2:

Lemma (**Forcing Lemma**). *Let $(\mathcal{H}, \mathcal{E}, \mathcal{S})$ be an arbitrary problem. Suppose that M is a learning method for it that converges to the truth almost everywhere, and that $(e_1, \ldots, e_n) \in \mathcal{E}$ is compatible with $h \in \mathcal{H}$. Then the above data sequence can be extended into a data sequence $(e_1, \ldots, e_n, \ldots, e_{n'}) \in \mathcal{E}$ such that:*

1. *$(e_1, \ldots, e_n, \ldots, e_{n'})$ is still compatible with h,*
2. *$M(e_1, \ldots, e_n, \ldots, e_{n'}) = h$.*

This lemma has a weaker and classic version, which deletes 'almost' and applies only to learning methods that converge to the truth everywhere. The weaker version has played an important role in proving many results in formal learning theory. Now, with the forcing lemma strengthened to cover almost everywhere convergence, many old proof techniques can be carried over to the learning theory developed here.[20] In fact, most of the results of this paper—positive or negative—are proved with the help of the forcing lemma.

Now let me turn to sketching the proof of the part "be enumeratively inductive". Suppose that learning method M converges to the truth almost everywhere with stability (and we are going to suppose that M converges on a maximal domain only when we really need to). Then, by the preceding result, M is never counterinductive, and hence it fails to converge to the truth in every Cartesian scenario of induction, say $(\texttt{No}, \vec{e})$, where $\vec{e}$ contains no occurrence of a nonblack raven. This failure of convergence in the Cartesian state $(\texttt{No}, \vec{e})$ opens the possibility for M to converge to the truth in its normal counterpart $(\texttt{Yes}, \vec{e})$. To turn this possibility into a reality, it suffices to invoke the last supposition of the theorem, that M converges to the truth on a maximal domain. It can be shown that, in order for M to converge to the truth on a maximal domain, the domain of convergence of M has to be so comprehensive that it contains all states that make hypothesis `Yes` true, which implies that M is enumeratively inductive.

As to the proof of the existential claim, it is almost routine to verify that it is witnessed by the method M^* constructed above, which says: "Output hypothesis `No` if you have observed a nonblack raven (`-`); otherwise output `Yes`."

This finishes the proof sketch of theorem 6.5.

C.2 Proof of the Forcing Lemma

The forcing lemma has two versions, one for (qualitative) learning methods and the other for probabilistic priors.

[20] In case you are interested: the forcing lemma can even be strengthened further to apply to *convergence to the truth on a dense set*. I wonder whether such a weak convergence criterion is interesting epistemologically, but I will not address this question here.

Lemma C.3 (**Forcing Lemma, Qualitative Version**). *Let $\big(\mathcal{H}, \mathcal{E}, \mathcal{S}\big)$ be an arbitrary problem. Suppose that M is a learning method for it that converges to the truth almost everywhere, and that $(e_1, \ldots, e_n) \in \mathcal{E}$ is compatible with $h \in \mathcal{H}$. Then the above data sequence can be extended into a data sequence $(e_1, \ldots, e_n, \ldots, e_{n'}) \in \mathcal{E}$ such that:*

1. *$(e_1, \ldots, e_n, \ldots, e_{n'})$ is still compatible with h,*
2. *$M(e_1, \ldots, e_n, \ldots, e_{n'}) = h$.*

Proof. Suppose that $(e_1, \ldots, e_n)$ is compatible with h. Namely,

$$|h| \cap |(e_1, \ldots, e_n)|$$

is a nonempty basic open set of topological space $|h|$. We are going to make use of the following characterization of "almost everywhere" in general topology:

> A property applies almost everywhere on a topological space (with a distinguished topological base) if, and only if, each nonempty (basic) open set U has a nonempty (basic) open subset U' such that the property applies everywhere on U'.

So, by the "only if" side and the hypothesis that M converges to the truth almost everywhere, it follows that $|h| \cap |(e_1, \ldots, e_n)|$ has a nonempty basic open subset:

$$|h| \cap |(e_1, \ldots, e_n, \ldots, e_k)|$$

on which M converges to the truth everywhere. Now, within this nonempty set, choose an arbitrary state $(h, \vec{e})$. So, in that state, M converges to the truth. Then there exists a positive integer $n' \geq k$ such that M outputs the truth h given the n'-th stage along data stream $\vec{e}$. That is:

$$M(e_1, \ldots, e_n, \ldots, e_k, \ldots, e_{n'}) = h\,.$$

It is not hard to see that the input is still compatible with h. □

Lemma C.4 (**Forcing Lemma, Bayesian Version**). *Let $\big(\mathcal{H}, \mathcal{E}, \mathcal{S}\big)$ be an arbitrary problem. Suppose that $\mathbb{P}$ is a probabilistic prior for it that converges to the truth almost everywhere, and that $(e_1, \ldots, e_n) \in \mathcal{E}$ is compatible with $h \in \mathcal{H}$. Then the above data sequence can be extended into a data sequence $(e_1, \ldots, e_n, \ldots, e_{n'}) \in \mathcal{E}$ such that:*

1. *$(e_1, \ldots, e_n, \ldots, e_{n'})$ is still compatible with h,*
2. *$\mathbb{P}(h \,|\, e_1, \ldots, e_n, \ldots, e_{n'}) > 1/2$.*

Proof. Copy the proof of the qualitative version of the forcing lemma, and paste it here. Now, replace the only occurrence of $M(e_1, \ldots, e_n, \ldots, e_k, \ldots, e_{n'}) = h$ by $\mathbb{P}(h \,|\, e_1, \ldots, e_n, \ldots, e_k, \ldots, e_{n'}) > 1/2$. As the last step, replace each occurrence of M by $\mathbb{P}$. □

C.3 Proofs for Enumerative Induction

The proofs presented in this section rely on the forcing lemma proved in section C.2.

Proof of Proposition 6.3. Suppose that a learning method M for the hard raven problem achieves perfectly monotonic convergence. Then M is a "non-inductive" method in that it never outputs `Yes`, so it fails to converge to the truth in every state in $|\texttt{Yes}|$. So M fails to converge to the truth almost everywhere in the topological space $|\texttt{Yes}|$. So M fails to converge to the truth almost everywhere. □

Proof of Theorem 6.5. To establish the existential claim, it suffices to show that it is witnessed by the learning method M^* we have discussed: "Output hypothesis `No` if you have observed a nonblack raven (`-`); otherwise output `Yes`." Proposition 5.2 has established that M^* converges to the truth almost everywhere. It is routine to verify that M^* converges to the truth with stability. To show that M^* has a maximal domain of convergence, note that it converges to the truth in all states in $|\texttt{Yes}|$ and in all states in $|\texttt{No}|$ except the Cartesian scenarios of induction. No learning method converges to the truth in strictly more states. For to do so is is to converge to the truth both in a normal state $(\texttt{Yes}, \vec{e})$ and its Cartesian counterpart $(\texttt{No}, \vec{e})$, which is impossible. This establishes maximal convergence for M^*, and finishes the proof of the existential claim.

To establish the first part of the universal claim "never be counterinductive", it suffices to invoke the proof that has already been detailed in appendix C.1, or simply to note that it is a corollary of theorem A.3. Note that the proof relies only on the two modes of convergence to the truth, "almost everywhere" and "stable". To establish the second part "be enumeratively inductive", suppose that M is a learning method for the hard raven problem that converges to the truth on a maximal domain, and that M is never counterinductive (making use of the first part). It suffices to show that M is enumeratively inductive, as follows. Since M is never counterinductive, M fails to converge to the truth in each Cartesian scenario of induction. So the domain of convergence of M is included in that of M^*, which has been proved to be a maximal domain of convergence. But M converges on a maximal domain, so M must have the same domain of convergence as M^*. Then M converges to the truth in every state $(\texttt{Yes}, \vec{e})$ contained in $|\texttt{Yes}|$. It follows that M is enumeratively inductive. □

Proof of Theorem B.7. The proof of the existential claim is the crux, so let me first present the proof of the easy part, the universal claim. Just copy the proof of the universal claim in theorem 6.5 (i.e. the preceding paragraph), paste it here, and apply the following replacements: First, replace the reference to theorem A.3 by the reference to its Bayesian counterpart, theorem B.9. Second, replace 'learning method' by 'probabilistic prior'. Third, replace M by $\mathbb{P}$. As the last step, replace

M^* by $\mathbb{P}^*$, which is the probabilistic prior to be constructed below for proving the existential claim.

To prove the existential claim, construct a witness $\mathbb{P}^*$ as a linear combination of two other probabilistic priors:

$$\mathbb{P}^* \;=\; \frac{1}{2}\,\mathbb{P}_0 + \frac{1}{2}\,\mathbb{P}_1\,,$$

where $\mathbb{P}_0$ and $\mathbb{P}_1$ are defined as follows. Let $\mathbb{P}_0$ be the probability function generated by, so to speak, assuming that `Yes` is true and observations of `+`, `0`, `-` are i.i.d. (independent and identically distributed) random variables, with equal probability 1/2 for `+` and for `0`, and with probability 0 for `-`. So:

$$\begin{aligned}
\mathbb{P}_0(\texttt{Yes}) &= 1\,.\\
\mathbb{P}_0(e_1, \ldots, e_n) &= \begin{cases} \left(\frac{1}{2}\right)^n & \text{if } e_i \neq \texttt{-} \text{ for each } i \leq n,\\ 0 & \text{otherwise.}\end{cases}
\end{aligned}$$

Similarly, let $\mathbb{P}_1$ be the probability function generated by, so to speak, assuming that `No` is true and observations of `+`, `0`, `-` are i.i.d. random variables with equal probability 1/3 for `+`, for `0`, and for `-`. So:

$$\begin{aligned}
\mathbb{P}_1(\texttt{No}) &= 1\,.\\
\mathbb{P}_1(e_1, \ldots, e_n) &= \left(\frac{1}{3}\right)^n\,.
\end{aligned}$$

It suffices to show that $\mathbb{P}^*$, defined as the half-and-half mixture of $\mathbb{P}_0$ and $\mathbb{P}_1$, converges to the truth with all the three modes mentioned in the existential claim. By the construction of $\mathbb{P}^*$, we have:

$$\begin{aligned}
\mathbb{P}^*(\texttt{Yes}) &= 1/2\,.\\
\mathbb{P}^*(\texttt{No}) &= 1/2\,.\\
\mathbb{P}^*(e_1, \ldots, e_n \,|\, \texttt{Yes}) &= \mathbb{P}_0(e_1, \ldots, e_n) \;=\; \begin{cases} \left(\frac{1}{2}\right)^n & \text{if } e_i \neq \texttt{-} \text{ for each } i \leq n,\\ 0 & \text{otherwise.}\end{cases}\\
\mathbb{P}^*(e_1, \ldots, e_n \,|\, \texttt{No}) &= \mathbb{P}_1(e_1, \ldots, e_n) \;=\; \left(\frac{1}{3}\right)^n\,.
\end{aligned}$$

Now, calculate conditional probability $\mathbb{P}^*(\texttt{Yes} \,|\, e_1, \ldots, e_n)$ by plugging the above probability values into the following instance of Bayes' theorem:

$$\begin{aligned}
&\mathbb{P}^*(\texttt{Yes} \,|\, e_1, \ldots, e_n)\\
=\;& \frac{\mathbb{P}^*(e_1, \ldots, e_n \,|\, \texttt{Yes})\,\mathbb{P}^*(\texttt{Yes})}{\mathbb{P}^*(e_1, \ldots, e_n \,|\, \texttt{Yes})\,\mathbb{P}^*(\texttt{Yes}) + \mathbb{P}^*(e_1, \ldots, e_n \,|\, \texttt{No})\,\mathbb{P}^*(\texttt{No})}\,.
\end{aligned}$$

Then we have:

$$\begin{aligned}
\mathbb{P}^*(\texttt{Yes}\,|\,e_1,\ldots,e_n) &= \begin{cases} \frac{1}{1+(2/3)^n} & \text{if } e_i \neq \texttt{-} \text{ for each } i \leq n, \\ 0 & \text{otherwise.} \end{cases} && (4) \\
\mathbb{P}^*(\texttt{No}\,|\,e_1,\ldots,e_n) &= 1 - \mathbb{P}^*(\texttt{Yes}\,|\,e_1,\ldots,e_n)\,. && (5) \\
\lim_{n\to\infty} \frac{1}{1+(2/3)^n} &= 1\,. && (6)
\end{aligned}$$

By the above three equations, (4)-(6), it follows that $\mathbb{P}^*$ converges to the truth in all states in $|\texttt{Yes}|$, and in all states in $|\texttt{No}|$ except the Cartesian scenarios of induction. But recall that, by lemma 4.3, the set of the Cartesian scenarios of induction is negligible within the topological space $|\texttt{No}|$. So $\mathbb{P}^*$ converges to the truth almost everywhere. Argue for stable convergence by considering the following two cases.

Case (i): suppose that $\mathbb{P}^*(\texttt{Yes}\,|\,e_1,\ldots,e_n) > 1/2$ and state $s \in |\texttt{Yes}| \cap |(e_1,\ldots,e_n)|$. So $s = (\texttt{Yes}, \vec{e})$, where $\vec{e}$ is an infinite $\texttt{+}/\texttt{0}$ sequence. By equation (4) and the fact that $1/\big(1+(2/3)^n\big)$ is a monotonically increasing function of n that converges to 1 as $n \to \infty$, we have: $\mathbb{P}^*$ converges to the truth in s and $\mathbb{P}^*(\texttt{Yes}\,|\,e_1,\ldots,e_n) \geq \mathbb{P}^*(\texttt{Yes}\,|\,e_1,\ldots,e_n,\ldots,e_{n'})$ for any $n' \geq n$.

Case (ii): suppose that $\mathbb{P}^*(\texttt{No}\,|\,e_1,\ldots,e_n) > 1/2$ and state $s \in |\texttt{No}| \cap |(e_1,\ldots,e_n)|$. So, by equations (4) and (5), $(e_1,\ldots,e_n)$ contains an occurrence of $\texttt{-}$. Then $s = (\texttt{No}, \vec{e})$, where $e_i = \texttt{-}$ for some $i \leq n$. So $\mathbb{P}^*(\texttt{No}\,|\,e_1,\ldots,e_n,\ldots,e_{n'}) = 1$ for all $n' \geq n$. So we have: $\mathbb{P}^*$ converges to the truth in s and $\mathbb{P}^*(\texttt{No}\,|\,e_1,\ldots,e_n) \geq \mathbb{P}^*(\texttt{No}\,|\,e_1,\ldots,e_n,\ldots,e_{n'})$ for any $n' \geq n$.

By the results of cases (i) and (ii), $\mathbb{P}^*$ converges to the truth with stability. To establish maximal domain of convergence, suppose for *reductio* that there is a probability function $\mathbb{P}$ that has a strictly more inclusive domain of convergence than $\mathbb{P}^*$ does. But $\mathbb{P}^*$ converges to the truth in all states except the Cartesian scenarios of induction. So $\mathbb{P}$ must converge to the truth in a certain normal state and in its Cartesian counterpart, which is impossible. So $\mathbb{P}^*$ converges to the truth on a maximal domain. □

C.4 Proofs for Ockham's Razor

The proofs presented in this section rely on the forcing lemma proved in section C.2.

Proof of Theorem A.3. Suppose that learning method M converges to the truth almost everywhere for problem $\big(\mathcal{H}, \mathcal{E}, \mathcal{S}\big)$. To prove the side $1 \Rightarrow 2$ by contraposition, suppose that M does not follow Ockham's tenacious razor. It suffices to show that M does not converge to the truth with stability. Discuss the following two exhaustive cases.

Case (i): Suppose that M violates the tenacity condition. That is, $M(e_1,\ldots,e_n) = h$ and $M(e_1,\ldots,e_n,\ldots,e_{n'}) \neq h$, where $(e_1,\ldots,e_n,\ldots,e_{n'})$ is compatible with h. By that compatibility, choose a state s in the nonempty set $|h| \cap (e_1, \ldots, e_n, \ldots, e_{n'})$. It follows that, given stage n in state s, M outputs the truth h but it has not converged to the truth. So M fails to converge to the truth with stability.

Case (ii): Suppose that M violates Ockham's razor. Then $M(e_1, \ldots, e_n) = h$, for some $(e_1, \ldots, e_n) \in \mathcal{E}$ and some $h \in \mathcal{H}$, but there exists another hypothesis $h' \in \mathcal{H}$ that is compatible with $(e_1, \ldots, e_n)$ and simpler than h. Since $(e_1, \ldots, e_n)$ is compatible with h', by the forcing lemma C.3 and the almost everywhere convergence of M, we have: $(e_1, \ldots, e_n)$ can be extended into a data sequence $(e_1, \ldots, e_n, \ldots, e_{n'}) \in \mathcal{E}$ such that, first, $M(e_1, \ldots, e_n, \ldots, e_{n'}) = h'$ and, second, $(e_1, \ldots, e_n, \ldots, e_{n'})$ is compatible with h'. Since $(e_1, \ldots, e_n, \ldots, e_{n'})$ is compatible with h' and since h' is simpler than h, it follows that $(e_1, \ldots, e_n, \ldots, e_{n'})$ is also compatible with h. By that compatibility, choose a state $s \in |h| \cap |(e_1, \ldots, e_n, \ldots, e_{n'})|$. So, given the earlier stage n in state s, M outputs the truth, h, but has not converged to the truth, for $M(e_1, \ldots, e_n, \ldots, e_{n'}) \neq h$. It follows that M fails to converge to the truth with stability.

To prove the side 2 ⇒ 1, it suffices to show that the tenacity condition (alone) implies convergence to the truth with stability. Suppose that M has the tenacity property, and that M outputs the truth h given stage n in state $s = (h, \vec{e})$. It suffices to show that M has converged to the truth given the same stage n in the same state s. Note that, for any natural number i, the data sequence $(e_1, \ldots, e_{n+i})$ extends $(e_1, \ldots, e_n)$ and is compatible with h. So, by the tenacity condition, $M(e_1, \ldots, e_{n+i}) = h$, for all $i \geq 0$. It follows that M has converged to the truth, h, given stage n in state s. □

Proof of Theorem B.9. Copy the proof of the 1 ⇒ 2 side of theorem A.3, and paste it here. Then apply the following replacements. For case (i):

- First, replace $M(e_1, \ldots, e_n) = h$ and $M(e_1, \ldots, e_n, \ldots, e_{n'}) \neq h$
  by $1/2 < \mathbb{P}(h \,|\, e_1, \ldots, e_n) > \mathbb{P}(h \,|\, e_1, \ldots, e_n, \ldots, e_{n'})$.
- And then replace each occurrence of M by $\mathbb{P}$.

For case (ii):

- First, replace $M(e_1, \ldots, e_n) = h$
  by $\mathbb{P}(h \,|\, e_1, \ldots, e_n) > 1/2$.
- Then replace $M(e_1, \ldots, e_n, \ldots, e_{n'}) = h'$
  by $\mathbb{P}(h' \,|\, e_1, \ldots, e_n, \ldots, e_{n'}) > 1/2$.
- Then replace $M(e_1, \ldots, e_n, \ldots, e_{n'}) \neq h$
  by $\mathbb{P}(h \,|\, e_1, \ldots, e_n, \ldots, e_{n'}) \not> 1/2$
- And, as the last step, replace each occurrence of M by $\mathbb{P}$.

This finishes the proof of the 1 ⇒ 2 side.

To prove that the converse 2 ⇒ 1 does *not* hold, construct a problem $\mathcal{P} = \big(\mathcal{H}, \mathcal{E}, \mathcal{S}\big)$ as follows. Consider only the following data streams, where m and n are arbitrary natural numbers:

$$\begin{aligned} s_\omega &= 0^\omega \,. \\ s_m &= 0^m 1^\omega \,. \\ s_{mn} &= 0^m 1^n 2^\omega \,. \end{aligned}$$

Their initial segments form the evidence space $\mathcal{E}$. The hypothesis space $\mathcal{H}$ consists of

- h = "The actual sequence will not end with occurrences of 2."
- h' = "It will."

The state space $\mathcal{S}$ consists of (h, s_ω), (h, s_m), and (h', s_{mn}), for all natural numbers m and n. Construct a countably additive probability function $\mathbb{P}$ that assigns the following probabilities to singletons of states:

$$\begin{aligned} \mathbb{P}\{s_\omega\} &= 0\,. \\ \mathbb{P}\{s_m\} &= \left(\frac{1}{2}\right)^{m+1} \times 60\%\,. \\ \mathbb{P}\{s_{mn}\} &= \left(\frac{1}{2}\right)^{m+1} \times 40\% \times \left(\frac{1}{2}\right)^{n+1}\,. \end{aligned}$$

Those assignments of probabilities are designed to ensure the following:

$$\begin{aligned} \mathbb{P}\{s_m\} &= \left(\frac{1}{2}\right)^{m+1} \times 60\%\,. \\ \mathbb{P}\{s_{m0}, s_{m1}, s_{m2}, \ldots\} &= \left(\frac{1}{2}\right)^{m+1} \times 40\%\,. \\ \mathbb{P}\{s_m, s_{m0}, s_{m1}, s_{m2}, \ldots\} &= \left(\frac{1}{2}\right)^{m+1}\,. \\ \sum_{m=0}^{\infty} \mathbb{P}\{s_m, s_{m0}, s_{m1}, s_{m2}, \ldots\} &= 1\,. \end{aligned}$$

It follows that, for each natural number m, we have:

$$\mathbb{P}(h \,|\, 0^m) = 60\%\,.$$

So $\mathbb{P}$ fails to converge to the truth H in state 0^ω. It is routine to verify that $\mathbb{P}$ converges to the truth in all the other states and, hence, does so almost everywhere.

It is also routine to verify that $\mathbb{P}$ follows Ockham's tenacious razor. In state (h, s_ω) and given information 0^m, $\mathbb{P}$ assigns a probability greater than 1/2 (namely 60%) to the truth (namely h) but fails to have started to stably converge to the truth, because it even fails to converge to the truth in that state. So $\mathbb{P}$ fails to converge to the truth with stability. This finishes the proof. □

D Examples, Discussions, and Open Questions

This section contains materials that might be of interest to some but not all readers.

D.1 Review of "Almost Everywhere" in Topology

Let X be a topological space (equipped with a distinguished topological base). Let π be a property that may or may not apply to points in X. The following conditions are equivalent:

1. π applies to almost all points in X—or speaking geometrically, π applies almost everywhere on X.

2. Every nonempty (basic) open set of X has a nonempty (basic) open subset on which π applies everywhere.

3. The set of points to which π applies is comprehensive enough to include a dense open subset of X.

The equivalence between conditions 1 and 2 is used in some of the proofs in this paper. Condition 3 emphasizes the fact that "being a dense subset of X" alone does not suffice for "containing almost all points in X". For example, the set of rationals is dense in the set of reals, but the former is too small to include an open subset of the latter. So the property of being a rational does not apply almost everywhere on the real line.

Sometimes topologists adopt a more lenient criterion of "almost all", according to which a property π is said to apply to almost all points in X just in case π applies to all points in $X \smallsetminus X'$, where X' is a countable union of negligible (i.e. nowhere dense) subsets of X. This more lenient criterion is used for proving the well-known theorem that almost all continuous functions defined on the unit interval are nowhere differentiable.[21]

The present paper adopts the more stringent criterion of "almost everywhere", requiring that X' be a negligible subset of X. This choice is made for a reason that is both epistemological and exploratory. The more stringent convergence criterion corresponds to a higher epistemic ideal. I propose to see whether the higher ideal is

[21] See Oxtoby (1996).

achievable for the hard raven problem, and the answer is positive. If that were too high to be achievable, I would try to see whether the lower ideal is achievable.

D.2 Too Hard to Achieve "Almost Everywhere"

A problem can be too hard to allow for the achievement of almost everywhere convergence to the truth. There are multiple ways of generating such problems. A Cartesian skeptic has one way to offer, making use of two empirically equivalent hypotheses:

Example D.1. The **very hard raven problem** poses the following joint question:

> *Are all ravens black? If not, will all the ravens observed in the future be black?*

There are three potential answers: `Yes`, `NoYes`, and `NoNo`. Note that `NoYes` is a Cartesian skeptical hypothesis, a hypothesis that is akin to (but not as terrible as) the proposition that one is a brain in a vat. Hypotheses `Yes` and `NoYes` are empirically equivalent—they are compatible with exactly the same data sequences. This problem can be formally defined as follows:

- the hypothesis space $\mathcal{H}$ is $\{\texttt{Yes}, \texttt{NoYes}, \texttt{NoNo}\}$,
- the evidence space $\mathcal{E}$ consists of all finite sequences of `+`, `0`, and/or `-`,
- the state space $\mathcal{S}$ consists of all states in the following three categories:
  - (a) the states $(\texttt{Yes}, \vec{e})$ in which $\vec{e}$ is an infinite `+`/`0` sequence,
  - (b) the states $(\texttt{NoYes}, \vec{e})$ in which $\vec{e}$ is an infinite `+`/`0` sequence.
  - (c) the states $(\texttt{NoNo}, \vec{e})$ in which $\vec{e}$ is an infinite `+`/`0`/`-` sequence that contains at least one occurrence of `-`.

Then we have this negative result:

Proposition D.2. *For the very hard raven problem, it is impossible to achieve almost everywhere convergence to the truth.*

Sketch of Proof. Suppose for *reductio* that some learning method M converges to the truth almost everywhere for the very hard raven problem. By almost everywhere convergence on the space $|\texttt{Yes}|$, there exists a `+`/`0` sequence $(e_1, \dots, e_n)$ such that M converges to the truth everywhere on $|\texttt{Yes}| \cap |(e_1, \dots, e_n)|$. By almost everywhere convergence on the space $|\texttt{NoYes}|$, $(e_1, \dots, e_n)$ can be extended to some `+`/`0` sequence $(e_1, \dots, e_n, \dots, e'_n)$ such that M converges to the truth everywhere on $|\texttt{NoYes}| \cap |(e_1, \dots, e_n, \dots, e'_n)|$. Choose an (infinite) data stream $\vec{e} \in |(e_1, \dots, e_n, \dots, e'_n)|$. So:

$$\begin{aligned} (\texttt{Yes}, \vec{e}) &\in |\texttt{Yes}| \cap |(e_1, \dots, e_n)|\,, \\ (\texttt{NoYes}, \vec{e}) &\in |\texttt{NoYes}| \cap |(e_1, \dots, e_n, \dots, e'_n)|\,. \end{aligned}$$

It follows that M converges to the truth both in state $(\texttt{Yes}, \vec{e})$ and in state $(\texttt{NoYes}, \vec{e})$. But that is impossible because those two states are empirically indistinguishable and make distinct hypotheses true. □

Due to that negative result, learning-theoretic epistemologists make no normative recommendation as to how to tackle the very hard raven problem. This raises some philosophical worries and questions, especially about the nature and purpose of learning-theoretic epistemology—see the next subsection, D.3, for a short philosophical discussion.

Here I would like to give more examples to show that, to construct a problem for which almost everywhere convergence is unachievable, it is not necessary to invoke two empirically indistinguishable *states*, and it is not sufficient to invoke two empirically equivalent *hypotheses*.

Example D.3. The **cardinality problem** poses the following question:

> *Given that the incoming data stream will be a* $0/1$ *sequence, how many occurrences of* 1 *will there be? Zero, one, two, ..., or infinite?*

This problem can be formally defined as follows:

- the hypothesis space $\mathcal{H}$ is $\{\texttt{0}, \texttt{1}, \texttt{2}, \ldots, \infty\}$,
- the evidence space $\mathcal{E}$ consists of all finite sequences of 0 and/or 1,
- the state space $\mathcal{S}$ consists of all states of the following form:
    - the states $(\texttt{n}, \vec{e})$ in which $\texttt{n}$ is a natural number and $\vec{e}$ is a $0/1$ sequence that contains exactly $\texttt{n}$ occurrences of 1;
    - the states $(\infty, \vec{e})$ in which $\vec{e}$ is a $0/1$ sequence that contains infinitely many occurrences of 1.

In the above problem, any two states are empirically distinguishable, but we still have the following negative result:

Proposition D.4. *For the cardinality problem, it is impossible to achieve almost everywhere convergence.*

Sketch of Proof. Suppose for *reductio* that there exists a learning method M that converges to the truth almost everywhere for the cardinality problem. So, in particular, M converges to the truth ∞ almost everywhere in topological space $|\infty|$. It follows that, for some finite data sequence σ_*, M converges to the truth ∞ everywhere on basic open set $|\infty| \cap |\sigma_*|$. Let $\texttt{k}$ be the least hypothesis compatible with σ_*. By the forcing lemma (in appendix C.2), there exists a data sequence σ_k that extends σ_* and is compatible with hypothesis $\texttt{k}$ such that $M(\sigma_k) = \texttt{k}$. Continue applying the forcing lemma to obtain this result: for each $n \geq k$, data sequence σ_n is extended into data sequence σ_{n+1} compatible with hypothesis $\mathtt{n+1}$ such that $M(\sigma_{n+1}) = \mathtt{n+1}$. Let σ be the infinite data sequence that extends σ_n for all natural numbers $n \geq k$. Then it is not hard to argue that M fails to converge to the truth in state $s = (\infty, \sigma)$. But this state s is in basic open set $|\infty| \cap |\sigma_*|$. Contradiction. □

The presence of two empirically equivalent hypotheses, alone, does not imply the impossibility of achieving almost everywhere convergence. Here is a counterexample:

Example D.5. The **even-vs-odd problem** poses the following question:

> *Given that the incoming data stream will be a* 0/1 *sequence with finitely many occurrences of* 1*, will there be evenly many or oddly many?*

This problem can be formally defined as follows:

- the hypothesis space $\mathcal{H}$ is $\{\mathtt{Even}, \mathtt{Odd}\}$,
- the evidence space $\mathcal{E}$ consists of all finite 0/1 sequences,
- the state space $\mathcal{S}$ consists of all states of the following form:
    - the states $(\mathtt{Even}, \vec{e})$ in which $\vec{e}$ is a 0/1 sequence that contains evenly many occurrences of 1;
    - the states $(\mathtt{Odd}, \vec{e})$ in which $\vec{e}$ is a 0/1 sequence that contains oddly many occurrence of 1.

The two competing hypotheses, `Even` and `Odd`, are empirically equivalent because no data sequence refutes one and saves the other. But we still have the following positive result:

Proposition D.6. *For the even-vs-odd problem, it is possible to achieve convergence to the truth everywhere—and, a fortiori, almost everywhere.*

Sketch of Proof. Everywhere convergence is achievable for this problem, as witnessed by this method: "Output `Even` if you have observed evenly many occurrences of 1; otherwise output `Odd`." □

D.3 Sensitivity to the Chosen Set of Hypotheses

It was remarked earlier that, for the very hard raven problem, it is even impossible to achieve almost everywhere convergence. As a consequence, learning-theoretic epistemologists have been unable to make a normative recommendation for an inquirer tackling that problem. Let me say why they have nothing to apologize.

The very hard raven problem embodies not just the philosophical problem of responding to the inductive skeptic, but also the problem of responding to the

Cartesian skeptic, as highlighted by the two empirically equivalent hypotheses put on the table:

- `Yes`: "Yes, all ravens are black."
- `NoYes`: "No, not all ravens are black; and yes, all ravens to be observed are black."

Learning-theoretic epistemology is not designed to respond to the Cartesian skeptic, and we may conjoin it with a good, independent reply to the Cartesian skeptic. To be sure, learning-theoretic epistemologists can, and should, insist that when an inquirer tackles the hard raven problem *rather than* the very hard one, she ought to be inductive and never be counterinductive, thanks to the argument and the results provided above.

So learning-theoretic epistemology typically makes a normative recommendation of this form: "*If* one tackles such and such a problem, one ought to follow a learning method having such and such properties." Such a normative recommendation is *sensitive* to, or *conditional* upon, the problem pursued by the inquirer.

In fact, learning theorists have long recognized that epistemology needs such sensitivity, for a reason that is not tied to Cartesian skepticism but can be traced back to the genesis of learning theory. The development of learning theory was historically motivated by the observation that it is mathematically impossible for us learn everything by meeting the epistemic ideal of everywhere convergence to the truth. That is, it is provably impossible to design a learning machine that is so powerful as to be capable of convergently solving the "ultimate" problem, the problem that entertains all hypotheses that human beings can understand (Putnam 1963). The cardinality problem corresponds to one such example, which makes the point even without invoking two empirically indistinguishable states or a Cartesian-like demon who is always hiding some observable items from the inquirer.

Given that it is impossible to design a learning machine for learning everything that one can understand, one has to prioritize certain things to learn. That is, in a context of inquiry, an inquirer has to identify the hypotheses whose truth values she really wants to learn, and to pursue the problem consisting of those hypotheses. When she switches to a different context of inquiry, she might need to reconsider the priority and decide to pursue a different problem. For example, an inquirer in a philosophy seminar on Cartesian skepticism might take the very hard raven problem to be of the utmost importance and decide to pursue it. But when she returns to the laboratory, the only important problem to pursue might just be the hard raven problem, rather than the very hard one. Here I only claim that she might switch that way. As to whether she ought to switch that way or is at least epistemically permitted to switch that way, the positive answer has to be defended elsewhere.

To sum up, learning-theoretic epistemologists recognize two groups of important issues to address:

> (*Mathematical Issues*) What can be learned? Which set of hypotheses can be learned in the limit? Which problem can be solved with which combinations of modes of convergence to the truth?
>
> (*Normative Issues*) One has no alternative but to prioritize certain things to learn. But which to prioritize? Which hypotheses and which problem are the things that one really cares about, or ought to care about, in which context of inquiry, such as a laboratory or a philosophy seminar?

While the mathematical issues have driven the development of learning theory, learning-theoretic epistemologists still have a lot to do to address the normative issues.

D.4 How to Justify Ockham's Razor: One More Example

Here is one more example that illustrates the application of theorem A.3 to justification of Ockham's razor:

Example D.7. Let x and y be real-valued variables, and suppose that y depends functionally on x. The **hard polynomial degree problem** poses the following question:

> *Given that y is a polynomial function of x, what is the degree of that polynomial function?*

This problem considers "rectangular" data on the x-y plane. A rectangular datum e_i is an open rectangle on the x-y plane that is axis-aligned and has only rational endpoints. Understand e_i to say: "The true polynomial function passes through rectangle e_i." A (finite or infinite) sequence of such rectangles is said to be compatible with a polynomial function if that polynomial function passes through all rectangles therein. This problem can be formally defined as follows:

- the hypothesis space $\mathcal{H}$ is the set of possible polynomial degrees, $\{\mathtt{0}, \mathtt{1}, \mathtt{2}, \ldots\}$;
- the evidence space $\mathcal{E}$ is the set of finite sequences of rectangular data that are compatible with at least some polynomial function;
- the state space $\mathcal{S}$ is the set of states taking the following form:

  $$(\mathtt{d}, \vec{e})\,,$$

  where $\mathtt{d}$ is a polynomial degree in $\mathcal{H}$ and $\vec{e}$ is an infinite sequence of rectangular data that is compatible with at least one polynomial function of degree $\mathtt{d}$.

A hypothesis of a lower polynomial degree is simpler:

$$\texttt{0} \prec \texttt{1} \prec \texttt{2} \prec \ldots \texttt{n} \prec \texttt{n+1} \ldots$$

Here is an example of a learning method that follows Ockham's tenacious razor:

$M^*_{\texttt{ock}}$ "Output degree `d` whenever `d` is the lowest polynomial degree that can fit the data you have (namely, whenever the data sequence you have is compatible with some polynomial function of degree `d` but with no polynomial function of any lower degree)."

This method never suspends judgment. There are other methods that also follow Ockham's tenacious razor, and they differ from the previous one by being less opinionated, willing to suspend judgment occasionally before jumping to a conclusion.

Before we apply theorem A.3 to the hard polynomial degree problem, we have to figure out what can be achieved for that problem:

Proposition D.8. *For the hard polynomial degree problem, it is possible to achieve convergence to the truth almost everywhere with stability on a maximal domain.*

Sketch of Proof. The existential claim is witnessed by the method $M^*_{\texttt{ock}}$ defined above, and can be proved in a way that mimics the proof of the existential claim of theorem 6.5. □

For the hard polynomial problem, is it possible achieve a higher mode, such as one that implies everywhere convergence or perfectly monotonic convergence? The answer is negative. To secure at least almost everywhere convergence, perfectly monotonic convergence is impossible because the problem in question is essentially an inductive problem, with a hypothesis that goes beyond the logical consequences of data. Everywhere convergence is unachievable because the state space is liberal enough to allow for two empirically indistinguishable states that make distinct hypotheses true. For example, consider a data stream $\vec{e}$ being so unspecific that it is compatible with some polynomial function of degree `d` and also with some other polynomial function of degree `d+1`. So there are (at least) two empirically indistinguishable states that make distinct hypotheses true, namely $(\texttt{d}, \vec{e})$ and $(\texttt{d+1}, \vec{e})$. Therefore, this problem does not have a learning method that converges to the truth everywhere. This is why it is called a "hard" problem, which suggests that we can obtain an "easy" version if we are willing to make a sufficiently strong presupposition to constrain the state space.[22]

[22] To be more specific, the **easy polynomial degree problem** is the same as the hard one except that it has a more constrained state space, in which each state $(\texttt{d}, \vec{e})$ is required to be such that its data stream $\vec{e}$ is compatible with exactly one polynomial function and that unique polynomial function has degree `d`. For this problem, everywhere convergence to the truth is achievable.

So here is what we have: When tackling the hard polynomial problem, an inquirer ought to achieve the highest achievable epistemic ideal among those in the lattice in figure 4, and that is the joint mode of convergence to the truth "almost everywhere" + "stable" + "maximal". A necessary means for achieving that is to follow Ockham's tenacious razor, thanks to theorem A.3. This is why an inquirer tackling the hard polynomial degree problem ought to follow Ockham's tenacious razor—or so I submit.

You might wonder whether Ockham's razor can be justified in a simpler way than I just did. Suppose that we have proved that a problem $\mathcal{P}$ is easy enough to make it possible to achieve at least "almost everywhere" + "stable", setting aside all other modes of convergence. Then it is tempting to quickly conclude that $\mathcal{P}$ ought to be tackled with a method that achieves "almost everywhere" + "stable", and then immediately apply theorem A.3 to conclude that $\mathcal{P}$ ought to be tackled with a method that follows Ockham's tenacious razor. It is tempting to do all this and rush to justify Ockham's razor, without considering higher epistemic ideals, such as one that adds convergence on a maximal domain. Is it OK to rush to justify Ockham's razor that way?

The answer is negative, and it is important to know why:[23] Some problems involve a trade-off between "stable" and "maximal" that forces the inquirer to sacrifice one in order to secure the other—see appendix D.5 for an example. In that case, it may not be immediately clear as to whether the inquirer should opt for the package "almost everywhere" + "stable" or side with "almost everywhere" + "maximal". Only the former requires following Ockham's tenacious razor; the latter does not.

D.5 Trade-off Between "Stable" and "Maximal"

There are situations in which the inquirer is, in a sense, forced to make a trade-off between two desirable modes of convergence, such as stability and maximality. Here is an example:

Example D.9. The **bounded even-vs-odd problem** poses the following question:[24]

> *Given that the incoming data stream will be a* 0/1 *sequence with at most two occurrences of* 1*, will there be evenly or oddly many occurrences of* 1*?*

Then, this problem can be formally defined as follows:

- the hypothesis space $\mathcal{H}$ is $\{\texttt{Even}, \texttt{Odd}\}$,

[23] The point made in this paragraph is a supplement to the way that Ockham's razor is justified in Kelly, Genin, and Lin (2016). In that earlier work, the achievability of "everywhere convergence" plus "cycle-free" (a variant of stability) is taken to be sufficient for justifying the use of Ockham's razor.

[24] I thank Konstantin Genin for bringing this problem to my attention.

- the evidence space $\mathcal{E}$ consists of all finite 0/1 sequences that have at most two occurrences of 1,
- the state space $\mathcal{S}$ consists of all states of the following form:
  - the states $(\texttt{Even}, \vec{e})$ in which $\vec{e}$ is a 0/1 sequence that contains exactly zero or two occurrences of 1;
  - the states $(\texttt{Odd}, \vec{e})$ in which $\vec{e}$ is a 0/1 sequence that contains exactly one occurrence of 1.

In this problem, `Odd` is simpler than `Even`.

Then we have the following trade-off result:

Proposition D.10. *Consider the following two modes of convergence:*

(1) convergence to the truth with stability,

(2) convergence to the truth on a maximal domain.

For the bounded even-vs-odd problem, each of those modes is achievable, but they are not jointly achievable.

Sketch of Proof. Everywhere convergence is achievable for this problem, as witness by this method: "Output `Even` if you have observed evenly many occurrences of 1; otherwise output `Odd`." As a consequence, convergence on a maximal domain is equivalent to everywhere convergence. Suppose that M converges to the truth on a maximal domain. So M converges to the truth everywhere and, hence, in the states $(\texttt{Even}, \vec{e})$ with $\vec{e}$ containing no occurrence of 1. But convergence in those states can be argued to violate the razor condition in Ockham's tenacious razor. Then, by the Ockham stability theorem A.3, M fails to achieve stable convergence. □

Fortunately, for the bounded even-vs-odd problem, the inquirer is forced to choose between stability and maximality only in a weak sense: she could have easily fine-grain the hypothesis space and ask instead: "How many occurrences of 1 will there be? Zero, one, or two?" Call this fine-grained problem the **zero-vs-one-vs-two problem**. For this problem, stability and maximality are jointly achievable together with everywhere convergence. This observation leads to the following open questions:

> (*Open Questions*) Are there problems for which it is possible to achieve stability, possible to achieve maximality, impossible to achieve both simultaneously, and even impossible to achieve both no matter how the hypothesis space is fine-grained? If there are such problems, which mode of convergence should one sacrifice in exchange for the other?

I tend to think that, in such an unfortunate problem, stability should be sacrificed in exchange for maximality—in general, the consideration about "where to converge" should be prioritized over the consideration about "how to converge". But this normative claim will have to be defended in another paper.