Non-Omniscience, Probabilistic Inference, and Metamathematics

Paul Christiano[*]

June 22, 2014

Abstract

We suggest a tractable algorithm for assigning probabilities to sentences of first-order logic and updating those probabilities on the basis of observations. The core technical difficulty is relaxing the constraints of logical consistency in a way that is appropriate for bounded reasoners, without sacrificing the ability to make useful logical inferences or update correctly on evidence.

Using this framework, we discuss formalizations of some issues in the epistemology of mathematics. We show how mathematical theories can be understood as latent structure constraining physical observations, and consequently how realistic observations can provide evidence about abstract mathematical facts. We also discuss the relevance of these ideas to general intelligence.

Contents

[*] UC Berkeley. This work originated at a workshop organized by the Machine Intelligence Research Institute. Email: paulfchristiano@eecs.berkeley.edu

1 Introduction

1.1 Motivation

Probability theory provides a powerful framework for reasoning under uncertainty, and many aspects of human cognition can be understood as probabilistic reasoning. Unfortunately, the simplest models tend to be general but computationally intractable, while practically relevant algorithms tend to rely on ideas which are ad hoc and more narrowly applicable.

One challenge in bridging the gap between theoretically simple models and practically relevant algorithms is coping with *logical uncertainty*: any realistic agent is necessarily uncertain not only about its environment or about the future, but also about the logically necessary consequences of its beliefs. An agent might suspect that a particular physical theory is correct, yet be uncertain about the predictions of that theory until it performs some computation. To date there have been few conceptually clean proposals for algorithms which handle such uncertainty in a general and principled way.

Related to this is the rich structure of human deductive reasoning, which appears to play a central role in intellectual activity yet is typically either trivialized by or omitted from probabilistic accounts of general cognition.

In this work we provide simple, general, and potentially tractable algorithms for reasoning in the presence of logical uncertainty. These algorithms are probabilistic, not because the environment is uncertain (though this may also be true) but because confidence about logical propositions imposes unrealistic computational demands.

In addition to its potential algorithmic relevance, having a concrete yet general model of bounded reasoning provides a setting for considering a wide range of epistemological problems formally. We will pay particular attention to the epistemology of mathematics, and to the relationship between abstract mathematical knowledge and reasoning about concrete finite objects.

1.1.1 Metamathematics

Within the formal practice of mathematics, axioms are typically taken for granted and their consequences are explored. *Outside* of formal mathematics, mathematicians reason about those consequences and come to judgments about which axiom systems are reasonable, useful, or "true." Metamathematics has had considerable success in describing the formal process of mathematics (though many of its most impressive results have been limitative), but has largely stayed away from the "extra-formal" process by which mathematicians decide what axioms they ought to accept.

In this paper we consider Bayesian prior over "mathematical states of affairs," and view observations about the world as providing Bayesian evidence about underlying mathematical facts. For example, if we observe the behavior of a calculator we may infer the laws of arithmetic as an explanation for its behavior. Having made that inference, we can use deduction to infer that if we enter "3 + 4 - 3" we will see "4." But now the process can also be turned on its head: when we type "134 * 345" and see "46230," we can make inferences about the *mathematical* state of affairs underlying reality.

When we observe a few values of a function $f(0), f(1), \ldots$ we can make inductive generalizations about the behavior of f. After inferring many generalizations we can begin to make generalizations about generalizations, and come to have well-grounded beliefs about mathematical abstractions.

1.1.2 Bounded universal intelligence

There has been some recent interest in the idea of "universal" intelligence, i.e. single algorithms which yield adequate or optimal behavior in a wide variety of situations. For example, see [3, 5]. This work tends to bear little resemblance to practical work in AI or learning theory. One reason for this divergence is that existing proposals for universal intelligence tend to rest on on exhaustive searches over possible policies or explanations. The "finite" analogs that have been considered additionally rely on mathematical proof as a basis for judgments.

The algorithms we present here can be used as a basis for universal intelligence (see the discussion in section 5); though there is still a substantial gap between our techniques and directly practical algorithms, our work much more closely resembles practical techniques for learning than the brute force search that has characterized past efforts.

Moreover, past approaches to bounded universal intelligence have relied on the use of mathematical proofs of optimality as a guide for decision-making. These strike us as highly unsatisfactory: most empirically successful policies cannot be supported by proofs of good performance, and it is far from clear whether any policies even *have* provable bounds on performance which would be satisfactory[1]. In light of this, it seems unlikely that such proof-based approaches can properly be considered universal intelligence (even setting aside computational issues).

[1] Ironically, proof-based techniques themselves typically fall into this category, and so such proposals for universal intelligence are "optimal" only when compared to a class of agents which is too narrow to include themselves:

1. The validity of proofs in a particular system cannot be verified by any proof within that system. In practice, the soundness of human reasoning is accepted by humans either on the basis of inductive reasoning or (though it seems less likely to the author) on the basis of some as-yet poorly understood human capacity.

2. Even accepting that a theory T is sound, it appears to be essentially impossible to *provably* lower-bound the performance of an agent which takes an action only if it provably has good consequences.

An agent which makes decisions on the basis of probabilistic judgments can pursue the action which they believe to be best, regardless of how complex the situation is. Such agents are never “stuck” doing nothing because they cannot find a proof of any action-relevant statements.

1.2 Guide to the paper

The remainder of Section 1 briefly discusses related work, and makes some notes on the relationship between efficiency and computability.

In Section 2 we develop some logical preliminaries, and lay out notions of logical coherence which are suitable for bounded reasoners.

In Section 3 we provide an explicit construction of a coherent prior distribution which reflects a condition of ignorance.

In Section 4 we describe some examples of mathematical reasoning within our framework.

In Section 5 we describe how this prior can be extended and incorporated into a general system for goal-oriented behavior.

In Section 6 we point to some open problems suggested by the formalism developed here, and discuss the implications of this work.

1.3 Related work

The problem of logical non-omniscience has been considered at some length in formal epistemology; a central question in this area is under what conditions we might say that a reasoner’s beliefs are consistent with the evidence they have received, when the reasoner is not able to deduce all of the logical consequences of statements which they believe. Several notions have been proposed, some of which, particularly [2], are quite similar to our own. Our motivation is more algorithmic than philosophical, and our further contributions are to provide notions of coherence that are plausibly suitable for efficient reasoning, to examine the construction of priors under our coherence conditions, and to consider the application of the resulting systems to mathematical reasoning and to goal-oriented behavior more generally.

The problem of assigning prior probabilities to logical sentences has also received some recent attention [1, 4]. Our work differs primarily by offering constructions which are in closer concordance with existing techniques, and which are more appropriate for use by bounded reasoners. We also take more interest in the application of these systems to understanding mathematical reasoning (in this section we could just as well substitute our proposed algorithms for finite approximations to those of Hutter or Demski), and to goal-oriented behavior (in this section we make more use of the distinctive properties of our proposal).

There is a massive literature on practical probabilistic reasoning. This work typically considers the problem of constructing practically useful models and on finding algorithmic approaches to reasoning about models in which exact inference is intractable. We suspect that, properly understood, mathematical reasoning provides a rich source of challenges for researchers working on approximate inference. Our goal is to help fortify this connection by providing a formal model of the problem and an example of how algorithmic techniques might be applied.

Some philosophers working on mathematical epistemology have considered more explicitly the interaction between probabilistic reasoning and mathematical reasoning. We hope to contribute to this research program by providing clearer formal models for probabilistic reasoning about mathematics, which allow us to arbitrate between conflicting intuitions and to focus our attention on those aspects of mathematical reasoning which currently elude our formal models.

Finally, as alluded to in section 1.1.2, there has been a small amount of research on algorithms which can reproduce intelligent behavior (at least in theory) in as broad a range of domains as possible[5, 3]. In some sense our work can be seen as a continuation of this program, and an effort to design more realistic and efficient algorithms for universal intelligence. Our work differs from existing research by providing an explicit account of probabilistic reasoning given bounded resources, which seems likely to be a key ingredient in any approach to general intelligence.

1.4 Efficient, finite, and infinite

Throughout this paper we will consider three domains: learners which are *efficient*, learners which are computable but not necessarily efficient, and learners which make use of a halting oracle. We imagine each type of algorithm as observing and interacting with a world which is somewhat more complicated than itself: the finite learners interact with an environment which is finite but more computationally complex than they are, while the infinite learner interacts with an environment that is even more infinite than it (either by making more calls to the halting oracle, or by lying even farther up the arithmetical hierarchy).

As we move from infinite to finite to efficient algorithms, we are able to enforce weaker and weaker consistency conditions on the probability distributions we maintain. Each step seems to present significant additional difficulties, and so in general we will first present an idea in the context of an infinite learner and only later show how to scale it down to a finite or efficient learner.

Crudely speaking, we might see the infinite domain as analogous to traditional metamathematics or to recursion theory, while the efficient domain is analogous to proof complexity or computational complexity. Though the latter domains share much of the technical machinery from the former, the situation has proven to be qualitatively more complex.

The *role* of mathematical reasoning in the case of finite versus infinite reasoning is also conceptually different. The infinite agents we consider know all first-order consequences of anything they know—i.e., if $\varphi \vdash \psi$ and they believe φ, then they believe ψ. For them, the difficulty is that they are interested in infinite statements whose truth is not pinned down by any enumerable list of axioms, and they reason about the truth of stronger mathematical theories as an explanation for these complex statements.

The finite agents we consider are interested in strictly finite statements about the world. So in principle they may be able to infer everything they care about from a very short list of axioms—they have no intrinsic interest in statements like $\forall x : \varphi(x)$. But this inference might take an extremely long time. For these finite agents, the machinery of logic is useful as a *computational expedient.* (A similar situation obtains in the field of proof complexity, where in some sense cut-elimination and Herbrand's theorem show that the use of quantifiers is extraneous. Quantifiers still play an important role, however, in controlling the complexity of a proof. Our situation is similar in spirit though technically quite different.)

The importance and non-trivial structure of mathematical reasoning do not appear to be unique to any of these domains (finite, infinite, and efficient). Rather, it is characteristic of situations in which a learner's environment is *more complex* than the learner itself. We suggest that building any agent which successfully reasons about an environment more complex than itself is a useful "first step" for a formal account of epistemology or universal intelligence.

2 Coherence

2.1 Logical preliminaries

For concreteness, and to keep the exposition simple, we will consider a single first-order language L containing:

- An unlimited supply of variables: $x_1, x_2 \ldots$.
- An unlimited supply of constant symbols: $c_1, c_2, \ldots$,
- For each $k \geq 1$, an unlimited supply of k-ary predicate symbols: $A^k_1, A^k_2, \ldots$,
- For each $k \geq 1$, an unlimited supply of k-ary function symbols: $f^k_1, f^k_2, \ldots$.

We will work with classical logic, and choose a Hilbert-style deductive system in which modus ponens is the only rule of inference. This will greatly simplify our algorithms, since in our context modus ponens is a consequence of additivity for probability distributions.

In general, we are interested in learners who take no axioms for granted other than those of first-order logic with equality: everything else is to be learned from experience[2]. We will often

[2] In fact it is also possible to consider systems for which the axioms of first-order logic are themselves merely inductive generalizations, and the laws of probability theory are the only built-in epistemic principles. But taking such an extreme position at the outset would complicate the exposition considerably, and so we leave fleshing out this position to future work.

discuss agents who accept some stronger set of axioms, particularly Robinson arithmetic[3]. We are interested in the behavior of an agent who axiomatically accept Robinson arithmetic primarily as a simple approximation to the behavior of an agent who provisionally accepts Robinson arithmetic as an explanation for some observations (we will discuss this much more in section 4). However, the skeptical reader can just as well imagine that we sometimes work with an agent who accepts the axioms of Robinson arithmetic along with first-order logic.

Recall that Robinson Arithmetic Q is axiomatized by the following 8 axioms, where S is one of the unary function symbols, $+$ and $*$ are two of the binary function symbols (we write them in infix notation for convenience), and 0 is one of the constant symbols:

1. $\forall x : Sx \neq 0$.
2. $\forall x, y : Sx = Sy \to x = y$.
3. $\forall x : x \neq 0 \to \exists y : x = Sy$.
4. $\forall x : x + 0 = x$.
5. $\forall x, y : x + Sy = S(x + y)$.
6. $\forall x : x * 0 = 0$.
7. $\forall x, y : x * Sy = (x * y) + x$.

2.1.1 Equivalent sentences

Though we will talk about probabilities of sentences, we are really interested in the probabilities of events defined by sentences. We would like to define our probability distributions over such events rather than needing to pay attention to the details of the way in which an event is represented by a sentence. It will therefore be useful to have a notion of "trivial" equivalence between logical sentences, and to define our probability distributions on equivalence classes of sentences. This allows us to move freely between equivalent representations.

However, because we are ultimately interested in efficient algorithms, we need to ensure that there is an efficient algorithm to judge whether two sentences are equivalent. So for example, it would be inappropriate to consider two sentences equivalent if their equivalence is a propositional tautology, because identifying propositional tautologies is computationally intractable.

We opt for a fragment of logic which lacks the distributivity laws for conjunction and disjunction but which is otherwise complete. This is motivated by the observation that although $(\varphi \vee \psi) \wedge \xi$ is equivalent to $(\varphi \wedge \xi) \vee (\psi \wedge \xi)$, this operation increases the representational complexity of the sentence and so corresponds to a non-trivial representational transformation.

[3] A minimal set theory would serve an identical role; we have sacrificed some conceptual simplicity for greater familiarity.

Note that although this notion will allow us to more comfortably manipulate sentences, it is primarily a technical convenience and does not play an essential conceptual role.

Definition 1 (Trivial equivalence)**.** We define $\sim$ as the minimal equivalence relation satisfying the following conditions: For each ψ, φ, ξ:

$$(\psi \wedge \varphi) \sim (\varphi \wedge \psi). \qquad (\varphi \wedge \top) \sim \varphi$$

$$\big(\psi \wedge (\varphi \wedge \xi)\big) \sim \big((\psi \wedge \varphi) \wedge \xi\big). \qquad \neg\bot \sim \top$$

$$(\varphi \wedge \varphi) \sim \varphi. \qquad (\psi \vee \varphi) \sim \neg(\neg\psi \wedge \neg\varphi)$$

$$(\varphi \wedge \neg\varphi) \sim \bot. \qquad (\psi \rightarrow \varphi) \sim (\varphi \vee \neg\psi)$$

$$(\neg\neg\varphi) \sim \varphi. \qquad \exists x : \varphi(x) \sim \neg\forall x : \neg\varphi(x)$$

$$(\varphi \wedge \bot) \sim \bot$$

If x_j is not free in φ, then $(\forall x_i : \varphi) \sim \Big(\forall x_j : \varphi\left[x_i = x_j\right]\Big)$.

And whenever $\varphi \sim \psi$:

$$\neg\varphi \sim \neg\psi \qquad (\forall x : \varphi) \sim (\forall x : \psi) \qquad (\varphi \wedge \xi) \sim (\psi \wedge \xi).$$

If $\varphi \sim \psi$, we say that φ and ψ are trivially equivalent.

The important fact about trivial equivalence is that it is easy to determine whether two sentences are equivalent:

Theorem 1. *It is possible to determine whether $\varphi \sim \psi$ in $n \log^2 n$ time, where n is the total length of φ and ψ.*

Essentially, we can greedily apply the rules defining $\sim$ until we arrive at the simplest representation of some φ, which is unique. The proof is not difficult but involves a tedious structural induction and is deferred to the appendix. In fact, this proof technique shows that given a set of sentences with total length N, we can divide them into equivalence classes under trivial equivalence in time $N \log^2 N$.

2.2 Coherence

Now we will define what we mean by a probability distribution over logical facts. We follow the definition of Gaifman[2].

Rather than thinking in terms of models, we will think about a probability distribution as a map $\mathbb{P} : L \rightarrow \mathbb{R}$. There are coherence conditions imposed on this map by the logical relationships amongst sentences, together with the usual probabilistic laws.

Definition 2 (Coherence)**.** We say that a map $\mathbb{P} : S \to \mathbb{R}$ is *coherent* with respect to a set of sentences S if it satisfies the properties:

1. Normalization: $\mathbb{P}(\top) = 1$.
2. Preservation of axioms: For any axiom $\varphi \in S$, $\mathbb{P}(\varphi) = 1$.
3. Non-negativity: $\mathbb{P}(\varphi) \geq 0$.
4. Weak consistency: if $\varphi \sim \psi$, then $\mathbb{P}(\varphi) = \mathbb{P}(\psi)$.
5. Additivity: $\mathbb{P}(\varphi) = \mathbb{P}(\varphi \wedge \psi) + \mathbb{P}(\varphi \wedge \neg\psi)$.

Let $\Delta(S)$ be the set of all maps $\mathbb{P} : S \to \mathbb{R}$ which are coherent with respect to S.

The axioms may be taken to be only the axioms of first-order logic, or may include the axioms of some theory of interest T.

If S is the set of all sentences, then we simply say that $\mathbb{P}$ is *coherent.* Otherwise we say that $\mathbb{P}$ is locally coherent.

Many similar definitions have appeared recently all of which are essentially equivalent in the case of $S = L$ to Kolmogorov's formulation. Our presentation differs slightly from the traditional presentation for the sake of computational convenience: we use preservation of axioms and weak consistency in place of preservation of theorems because of its nicer computational properties, and adopt a purely syntactic formulation of additivity.

To justify our definition, we reproduce the following standard theorem which shows that these conditions are exhaustive, modified for our alternative axiomatization. It is worth noting that the result makes no use of the fact that $\mathbb{P}$ assigns probability 1 to axioms of first-order logic, except in the conclusion that the theories T are theories of first-order logic. If we weakened the preservation of axioms condition, we would obtain a similar theorem but with respect to some broader class of assignments $L \to \{\bot, \top\}$ than consistent theories of first-order logic.

Theorem 2. *A distribution $\mathbb{P}$ is coherent if and only if there is some probability measure μ on the space of complete consistent theories such that for all φ, $\mathbb{P}(\varphi) = \mu\left(\left\{T \,|T \vdash \varphi\right\}\right)$.*

Proof. It is easy to verify that for any μ, the function $\mathbb{P} : L \to \mathbb{R}$ defined by μ is coherent. It remains to show that for any coherent $\mathbb{P}$, we can find a measure μ such that $\mathbb{P}(\varphi) = \mu\left(\left\{T \,|T \vdash \varphi\right\}\right)$. We will describe a process which generates a theory, such that the distribution of theories generated by the process reproduces $\mathbb{P}$ in the appropriate way.

The first step is showing that if $\vdash \varphi$, then $\mathbb{P}(\varphi) = 1$. Let S be the set of sentences that are assigned probability 1. Since $\mathbb{P}$ preserves axioms, each axiom is in S. So it suffices to show that S is closed under modus ponens. First we make some preliminary observations:

- By additivity and non-negativity, $\mathbb{P}(\varphi \wedge \psi) \leq \mathbb{P}(\varphi)$.
- By weak consistency $\mathbb{P}(\top \wedge \bot) = \mathbb{P}(\bot)$, $\mathbb{P}(\top \wedge \top) = \mathbb{P}(\top)$. By normalization and additivity, $\mathbb{P}(\bot) = 0$.
- By weak consistency, $\mathbb{P}(\varphi \wedge \neg\varphi) = 0$. By additivity and normalization, it follows that $\mathbb{P}(\varphi) = 1 - \mathbb{P}(\neg\varphi)$.
- By weak consistency and the previous observation, $\mathbb{P}(\varphi \to \psi) = 1$ if and only if $\mathbb{P}(\varphi \wedge \neg\psi) = 0$.

Now suppose that $\varphi \in S$ and $\varphi \to \psi \in S$. Then by the preceding observations and additivity:

$$\mathbb{P}(\psi) \geq \mathbb{P}(\varphi \wedge \psi) = \mathbb{P}(\varphi \wedge \neg\psi) + \mathbb{P}(\varphi \wedge \psi) = \mathbb{P}(\varphi) = 1,$$

hence S is closed under modus ponens and contains all theorems of first-order logic. We conclude that if $\vdash \varphi$, then $\mathbb{P}(\varphi) = 1$.

Now, fix some enumeration $\varphi_1, \varphi_2, \ldots$ of all of the sentences of L. Let $T_0 = \emptyset$ and iteratively define T_{i+1} in terms of T_i as follows. If T_i is complete, we set $T_{i+1} = T_i$. Otherwise, let φ_j be the first statement in our enumeration which is independent of T_i.

Let $T_{i+1} = T_i \cup \varphi_j$ with probability $\mathbb{P}\left(\varphi_j \;\middle|\; T_i\right)$[4] and $T_{i+1} = T_i \cup \neg\varphi_j$ with probability $\mathbb{P}\left(\neg\varphi_j \;\middle|\; T_i\right)$. Because φ_j was independent of T_i, the resulting system remains consistent in either case. Define $T = \cup_i T_i$. Since each T_i is consistent, T is consistent by compactness. For each i, φ_i or $\neg\varphi_i$ will eventually be included in the theory so T is complete.

By additivity and weak consistency

$$\mathbb{P}(\varphi \mid T_i) = \mathbb{P}\left(\varphi \;\middle|\; T_i \wedge \varphi_j\right) \mathbb{P}\left(\varphi_j \;\middle|\; T_i\right) + \mathbb{P}\left(\varphi \;\middle|\; T_i \wedge \neg\varphi_j\right) \mathbb{P}\left(\neg\varphi_j \;\middle|\; T_i\right),$$

thus the sequence $\mathbb{P}(\varphi \mid T_i)$ is a martingale.

Since $\mathbb{P}(T) = 1$ if T is a theorem, if $T \vdash \varphi$ then $\mathbb{P}(\varphi \mid T) = 1$. Since $T_i \vdash \varphi$ or $T_i \vdash \neg\varphi$ for large enough i, $\mathbb{P}(\varphi \mid T_i)$ stabilizes at either 0 or 1. Moreover, $T \vdash \varphi$ iff this sequence stabilizes at 1. The martingale property then implies that $T \vdash \varphi$ with probability $\mathbb{P}(\varphi \mid T_0) = \mathbb{P}(\varphi)$. □

2.2.1 Impossibility

Ultimately we are interested in understanding finite agents. Unfortunately, a finite agent cannot find any coherent mapping $\mathbb{P} \in \Delta(S)$, even implicitly.

Theorem 3. *There is no recursively approximable coherent map $\mathbb{P} : L \to \mathbb{R}$ which assigns non-negligible probability to the axioms of Q.*

[4] By definition $\mathbb{P}(T) = \mathbb{P}\left(\wedge_{\varphi \in T}\varphi\right)$, which is defined uniquely by weak consistency, and $\mathbb{P}\left(\varphi_j \;\middle|\; T_i\right) = \frac{\mathbb{P}\left(\varphi_j \wedge T_i\right)}{\mathbb{P}(T_i)}$. It is easy to verify by induction that $\mathbb{P}(T_i) > 0$ with probability 1, over the random choices of our process.

Proof. Suppose $\mathbb{P}'$ was such a map. Then we can define a new coherent map $\mathbb{P}$ which assigns probability 1 to Q, via

$$\mathbb{P}(\varphi) = \mathbb{P}'(\varphi \mid \mathrm{Q}).$$

Note that $\mathbb{P}$ is also recursively approximable. It is coherent, because it is obtained by conditioning the distribution over theories corresponding to $\mathbb{P}'$ on the event that the theory entails Q.

Thus there is a Turing machine M such that M halts on input $\ulcorner\varphi\urcorner$ and outputs 0 if $\mathbb{P}(\varphi) = 0$ and 1 if $\mathbb{P}(\varphi) = 1$—M can simply compute increasingly accurate approximations to φ until it either proves $\mathbb{P}(\varphi) < \frac{2}{3}$, in which case it outputs 0, or $\mathbb{P}(\varphi) > \frac{1}{3}$, in which case it outputs 1. One of these is guaranteed to happen eventually.

Since $\mathbb{P}(\mathrm{Q}) = 1$, by diagonalization we can construct a sentence φ such that $\mathbb{P}(\varphi) = \mathbb{P}\big(M(\ulcorner\varphi\urcorner) = 0\big)$. Since M always halts, it is either a theorem of Q that $M(\ulcorner\varphi\urcorner) = 0$ or $M(\ulcorner\varphi\urcorner) = 1$. Thus one of these statements is true and has probability 1. Suppose that $M(\ulcorner\varphi\urcorner) = 1$ and hence $\mathbb{P}\big(M(\ulcorner\varphi\urcorner) = 1\big) = 1$. Then $\mathbb{P}(\varphi) = 0$, and so by construction of M $M(\ulcorner\varphi\urcorner) = 0$, a contradiction. If instead $M(\ulcorner\varphi\urcorner) = 0$, then $\mathbb{P}\big(M(\ulcorner\varphi\urcorner) = 0\big) = 1$. Then $\mathbb{P}(\varphi) = 1$, so by construction of M $M(\ulcorner\varphi\urcorner) = 1$, a contradiction. □

In light of this impossibility result (and the even more serious difficulties when we try to design *efficient* algorithms), we rely on a weaker notion of coherence.

2.3 Local coherence

For any finite set S, it is possible to find locally coherent distributions $\mathbb{P} \in \Delta(S)$ in an amount of time polynomial in the size of S.

One remaining question is what class of sentences to take. Representing a larger class introduces significant computational complexity, but allows us to represent more complex hypotheses and perform more complex deductive reasoning. This is simply the traditional tension between accuracy and complexity.

For now, we will assume that we have already identified a reasonably-sized set S_0 containing some sentences of interest to us—descriptions of what we may observe or decide, statements about what we value, explanatory hypotheses which might account for our observations, and so on.

Two considerations seem important when selecting S:

1. Even if we are only interested in S_0, we may want to consider a distribution which is coherent over a larger set in order to constrain our beliefs about S_0 further.

2. If $\varphi, \psi \in S_0$ and $\mathbb{P} \in \Delta(S_0)$, it isn't clear how to define $\mathbb{P}(\varphi \mid \psi)$. Typically this would be defined in terms of $\mathbb{P}(\varphi \wedge \psi)$, but if S_0 isn't closed under conjunctions then $\varphi \wedge \psi$ need not be in the domain of S_0. Requiring S to be large enough to form arbitrary

conditional probabilities for observations in S_0 is typically prohibitive (it requires S to be as large as all possible *subsets* of S_0).

In this section we describe *finite* but radically impractical notions of coherence, which are comparable with those that have been presented in the literature. In the following section, we turn our attention to *tractable* notions of coherence, which prove to be substantially more challenging.

2.3.1 Preserving propositional tautologies

One approach to relaxing logical omniscience is to assume that *propositional* tautologies are assigned probability 1, while allowing tautologies of first order logic to be uncertain [2].

Define $\mathrm{cl}\,(S_0)$ to be the *closure* of S_0 under conjunctions and negations: $\mathrm{cl}\,(S_0)$ consists of every sentence of the form $\bigwedge_i \varphi_i$, where each φ_i is either a sentence of S_0 or its negation.

Coherence with respect to $\mathrm{cl}\,(S_0)$ is the minimal criterion necessary for being able to condition on arbitrary sentences of S_0 arbitrarily many times. Indeed, we have:

$$\mathbb{P}\left(\varphi \mid \varphi_1 \wedge \cdots \wedge \varphi_k\right) = \frac{\mathbb{P}\left(\varphi \wedge \varphi_1 \wedge \cdots \wedge \varphi_k\right)}{\mathbb{P}\left(\varphi_1 \wedge \cdots \wedge \varphi_k\right)},$$

and so if a probability distribution is not coherent with respect to $\mathrm{cl}\,(S_0)$, there is no clear approach to conditioning on observations in S_0 while remaining coherent.

Although $\mathrm{cl}\,(S_0)$ contains only conjunctions of sentences in S_0, if $\mathbb{P}$ is coherent with respect to $\mathrm{cl}\,(S_0)$ we can straightforwardly extend it to sentences of the form $\varphi \vee \psi$ by using the principle of inclusion and exclusion.

Because the size of $\mathrm{cl}\,(S_0)$ is $2^{|S_0|}$, coherence with respect to $\mathrm{cl}\,(S_0)$ is typically too demanding for a tractable algorithm. This is not a rectifiable deficiency; coherence with respect to $\mathrm{cl}\,(S_0)$ implies that propositional tautologies receive probability 1, and identifying propositional tautologies is NP-hard and generally believed to require exponential time.

2.3.2 Bounded quantifier rank

If we are only interested in producing finite algorithms without concern for computational efficiency, we can enlarge the set S_0 far beyond $\mathrm{cl}\,(S)$. In this section we will lay out a considerably stronger notion of coherence. For simplicity, the reader uninterested in computational complexity can use this as a model for our "idealized mathematical reasoner" in our discussions of mathematical epistemology.

Coherence with respect to $\mathrm{cl}\,(S_0)$ allows us to make any purely logical arguments concerning the sentences in S_0, but it causes us to treat quantified statements as if they were atoms, constrained only by whatever relationships appear in the set S_0 itself.

Unfortunately, if we allow ourselves to construct new terms arbitrarily, it is not clear what stops our problem from again becoming uncomputable. For example, if we accept the axioms of arithmetic and are able to reason about the term $x+1$ whenever we are able to reason about the term x, then we must assign probability 1 to every true Σ_1 arithmetical sentence. This is impossible, as per the argument in 3.

Our approach is to allow ourselves to construct new terms of limited complexity. This yields a set of sentences which has strong closure properties, yet is still finite. This notion is very closely related to quantifier rank, and indeed we could make use of quantifier rank directly if we worked with a purely relational language. We present the definition here for convenience:

Definition 3 (Functional quantifier rank)**.** The *functional quantifier rank* $\mathrm{qr}\left(\cdot\right)$ of a formula φ or term t is defined inductively as follows:

- $\mathrm{qr}\left(t\right) = 0$ if t is a constant symbol or variable.
- $\mathrm{qr}\left(f(t_1, \ldots, t_k)\right) = \max\left(\mathrm{qr}\left(t_1\right), \ldots, \mathrm{qr}\left(t_k\right)\right) + 1.$
- $\mathrm{qr}\left(A(t_1, \ldots, t_k)\right) = \max\left(\mathrm{qr}\left(t_1\right), \ldots, \mathrm{qr}\left(t_k\right)\right).$
- $\mathrm{qr}\left(\varphi \wedge \psi\right) = \mathrm{qr}\left(\varphi \vee \psi\right) = \max\left(\mathrm{qr}\left(\varphi\right), \mathrm{qr}\left(\psi\right)\right).$
- $\mathrm{qr}\left(\neg\varphi\right) = \mathrm{qr}\left(\varphi\right)$
- $\mathrm{qr}\left(\forall x : \varphi\left(x\right)\right) = \mathrm{qr}\left(\exists x : \varphi\left(x\right)\right) = \mathrm{qr}\left(\varphi\left(x\right)\right) + 1$

Write $L\left[k\right]$ for the set of $\varphi \in L$ with $\mathrm{qr}\left(\varphi\right) \leq k$.

It is easily verified by induction on k that each $L\left[k\right]$ contains only finitely many non-equivalent sentences.

So given a set of sentences of interest S_0, we can work with $\mathbb{P} \in \Delta\left(L\left[k\right]\right)$, where k is much larger than the maximum quantifier rank of any $\varphi \in S_0$. These distributions not only assign probability 1 to propositional tautologies, they respect any deduction in first order logic which requires manipulating terms of bounded complexity. It is possible to interpret these distributions as distributions over stratified models, in which each quantifier ranges over elements of some specified complexity $\leq k$ and every nested quantifier must range over a larger set.

2.4 Tractable notions of coherence

When moving to the domain of *tractable* notions of coherence, we face the two difficulties mentioned earlier:

1. An efficient reasoner cannot infer all consequences of their beliefs over propositional logic, and so must make use of some approximate scheme to draw as many useful inferences as possible.

2. It is unclear how to update a compactly represented probability distribution on a sequence of observations.

The first problem is a standard challenge in algorithms; for example, it is identical to the problem faced in constraint satisfaction. We propose addressing it with standard techniques from these domains. In particular, we propose applying the well-understand positive-semidefinite relaxation of the marginal polytope to obtain a relaxation of $\Delta\left(\mathrm{cl}\left(S_0\right)\right)$.

The second problem is more technical, but also appears to be quite important. We propose considering $\mathbb{P}\left(\cdot \mid \varphi\right)$ as the distribution which assigns probability 1 to φ and which has minimal KL divergence to $\mathbb{P}$. We can then estimate the KL divergence as the Bregman divergence associated to a certain estimate of the entropy of $\mathbb{P}$. The resulting algorithm is conceptually simple, has a simple interpretation, and can be proven to be effective in simple cases.

We will explain both these ideas in more depth over the following sections.

The effectiveness of this approach in general remains an open question. Indeed, we expect that more sophisticated approaches are possible and will have more desirable properties. However, this does appear to be a plausible, tractable algorithm for assigning probabilities to mathematical claims, and we believe it might serve as a useful model for how this problem can be approached.

In this section we will concern ourselves only with defining a notion of coherence which is appropriate for bounded reasoners. In future sections we will turn our attention to actually computing probability distributions which satisfy this notion of coherence.

2.4.1 Relaxing $\Delta\left(\mathrm{cl}\left(S_0\right)\right)$

Since S_0 includes all of the sentences that we are interested in, our primary concern is obtaining estimates for the probabilities of sentences in S_0; it is only for the sake of accuracy that we would like these probabilities to be extensible to a distribution in $\Delta\left(\mathrm{cl}\left(S_0\right)\right)$. Write M for the set of distributions in $\Delta\left(S_0\right)$ which can be extended in this way, i.e. the set of *projections* of distributions in $\Delta\left(\mathrm{cl}\left(S_0\right)\right)$ to the coordinates in S_0. Essentially, we are interested in finding and updating distributions in M.

Unfortunately, it is easily seen to be impossible to optimize functions over M, or to determine whether M is compatible wfrom ith some constraints of the form $\mathbb{P}\left(\varphi_i\right) = 1$ (this is equivalent to Boolean satisfiability). So our task amounts to finding some set $\widetilde{M}$ which *approximates* M well.

This is a problem which has been well-studied in the literature on probabilistic inference, as well as the literature on constraint satisfaction and approximation algorithms. We will consider the strongest relaxation typically considered, which appears to have some characteristics that make it suitable for our setting. More powerful relaxations are possible and are sometimes considered, but little can yet be said about their general usefulness.

As usual, we stress that our work serves more as a demonstration of how these problems could be addressed, and it is extremely unlikely that the actual implementations we provide are the final say on the subject.

2.4.2 Sum-of-squares relaxations

One of the most widely successful paradigms in optimization has been the use of semidefinite programming relaxations. In this section we describe their application to the present problem.

Write $S_0 \times S_0$ for the set of sentences of the form $\varphi \wedge \psi$, where φ, ψ are sentences in S_0 or their negations. Consider some $\mathbb{P} \in \Delta(S_0 \times S_0)$. If $\mathbb{P}$ can be extended to a coherent distribution on L, or even on $\mathrm{cl}(S_0)$, then its restriction to $S_0 \times S_0$ must satisfy some simple positivity conditions. In particular, let $\alpha_\varphi \in \mathbb{R}$ be a variable for each $\varphi \in S_0$, and consider the sum

$$\sum_{\varphi,\psi \in S_0} \alpha_\varphi \alpha_\psi \mathbb{P}(\varphi \wedge \psi).$$

This sum is the expectation of a certain random variable, $\mathbb{E}\left[f^2\right]$, where f is defined by

$$f = \sum_{\varphi \in S_0} \begin{cases} \alpha_\varphi & \text{if } \varphi \\ 0 & \text{else} \end{cases}$$

Since f^2 is always non-negative, its expectation should also be non-negative. Thus for any choice of α, we should have

$$\sum_{\varphi,\psi \in S_0} \alpha_\varphi \alpha_\psi \mathbb{P}(\varphi \wedge \psi) \geq 0.$$

If we let $\Sigma_\mathbb{P}$ be the matrix with rows and columns indexed by S_0 and with (φ, ψ) entry given by $\mathbb{P}(\varphi \wedge \psi)$, then this condition is precisely equivalent to the positive semi-definiteness of $\Sigma_\mathbb{P}$ (written $\Sigma_\mathbb{P} \geq 0$). Write $\Delta^+(S_0)$ for the set of $\mathbb{P} \in \Delta(S_0 \times S_0)$ satisfying this constraint.

Example 1. Suppose S_0 consists of three propositions p, q, r. Then there is a $\mathbb{P} \in \Delta(S_0 \times S_0)$ such that

$$\mathbb{P}(p) = \mathbb{P}(q) = \mathbb{P}(r) = \frac{1}{2}$$
$$\mathbb{P}(p \wedge q) = \mathbb{P}(q \wedge r) = \frac{1}{2}$$
$$\mathbb{P}(p \wedge r) = 0$$

Of course this distribution cannot be extended to a coherent distribution over $S_0 \times S_0 \times S_0$. We can detect this failure of extensibility by taking $\alpha_p = \alpha_r = 1$ and $\alpha_q = -1$, and noticing

$$\sum \alpha_\varphi \alpha_\psi \mathbb{P}(\varphi \wedge \psi) = -\frac{1}{2} < 0.$$

Thus $\mathbb{P} \notin \Delta^+(S_0)$. This provides a simple example of how $\Delta^+(S_0)$ can impose "global" constraints on $\mathbb{P}$. It is well known that there are distributions $\mathbb{P} : S_0^2 \to [0, 1]$ which are not in $\Delta^+(S_0)$ but which nevertheless have extensions in $\Delta\left(S_0^k\right)$ for $k = \theta(n)$, where $S_0^k = S_0 \times S_0 \times \cdots \times S_0$.

Of course, there are also distributions $\mathbb{P} \in \Delta^+(S_0)$ which have no extension to $\mathrm{cl}(S_0)$, or even to $\Delta\left(S_0^4\right)$. So one must take care when describing such an objection as a "distribution" (the expression "psuedodistribution" is sometimes used in the literature).

In addition to being testable in polynomial time (for example by computing the smallest eigenvalue of $\Sigma_\mathbb{P}$), this constraint is particularly interesting for the following reason: if $\Sigma_\mathbb{P} \geq 0$, then there are random variables A_φ such that $\mathbb{P}(\varphi \wedge \psi) = \mathbb{E}\left[A_\varphi A_\psi\right]$. For example, we can take A to be a multivariate normal distribution with covariance matrix $\Sigma_\mathbb{P}$. This provides us with a way to understand $\mathbb{P}$ as a *global* description of S_0, even though it only assigns probabilities to sentences in $S_0 \times S_0$. This is typically the property of interest in constraint satisfaction, where this global description can be used to extract a solution to a constraint satisfaction problem.

2.4.3 Entropy and updating

One special consequence of this global description of distributions in $\Delta^+(S_0)$ is that we can use it to obtain a measure of the information content of a $\mathbb{P} \in \Delta^+(S_0)$. Namely, we can consider the differential entropy of some continuous distribution which has the same moments as $\mathbb{P}$. One natural candidate is the Gaussian with covariance matrix $\Sigma_\mathbb{P}$. This is also the maximum achievable variance, and it has the very simple form $\log\left(|\Sigma_\mathbb{P}|\right)$.

Inspired by the understanding of Bayesian inference in learning theory, there is a very close connection between notions of entropy and Bayesian updating. The *relative entropy* (or KL divergence) of one distribution $\mathbb{P}$ from another $\mathbb{Q}$ is given by

$$D(\mathbb{P} \| \mathbb{Q}) = \sum_{\text{outcomes } x} \mathbb{P}(x) \log \frac{\mathbb{P}(x)}{\mathbb{Q}(x)}.$$

Intuitively, this measures how well $\mathbb{Q}$ approximates $\mathbb{P}$.

An equivalent description of the distribution $\mathbb{P}(\cdot \mid \varphi)$ is as the distribution which assigns probability 1 to φ and *minimizes* the KL divergence from $\mathbb{P}$. In fact, in the setting of online learning theory this is the *more* natural way to describe a Bayesian update, which can be more readily generalized and analyzed. So as long as we have an appropriate notion of KL divergence, we can define an analog of Bayesian updating.

In fact the KL divergence can be calculated directly from the entropy as a Bergman divergence, and so an analog of entropy naturally yields an analog of the KL divergence. In this case, the resulting quantity is the KL divergence between the Gaussians with covariance matrices $\Sigma_\mathbb{P}$ and $\Sigma_\mathbb{Q}$, which is given by:

$$D\left(\Sigma_\mathbb{P} \middle\| \Sigma_\mathbb{Q}\right) = \frac{1}{2}\left(\log\left|\Sigma_\mathbb{Q}\right| - \log\left|\Sigma_\mathbb{P}\right| + \mathrm{tr}\left(\Sigma_\mathbb{P}\Sigma_\mathbb{Q}^{-1} - I\right)\right).$$

This quantity is also called the *logdet divergence* between $\Sigma_{\mathbb{P}}$ and $\Sigma_{\mathbb{Q}}$, and it has been used successfully for matrix online learning in a number of contexts, most notably kernel and metric learning.

Now we would like to define the distribution $\mathbb{P}(\cdot \mid \varphi)$ as the distribution in $\Delta^+(S)$ which assigns probability 1 to φ and minimizes $D\left(\Sigma_{\mathbb{P}(\cdot \mid \varphi)} \middle\| \Sigma_{\mathbb{P}}\right)$. There is one more subtlety before we can complete our definition. If $\mathbb{P}$ is coherent, then $\Sigma_{\mathbb{P}}$ will necessarily not be of full rank (for example, if φ is the negation of an axiom then $\mathbb{P}(\varphi \wedge \psi) = 0$ for every ψ, so $\Sigma_{\mathbb{P}}$ has a zero row). So we will typically have $\log|\Sigma_{\mathbb{P}}| = \log\left|\Sigma_{\mathbb{Q}}\right| = -\infty$. We can extend the definition of the logdet divergence to these cases as follows: let (u_i, θ_i) be the eigenvectors and eigenvalues of $\Sigma_{\mathbb{P}}$, and (v_i, λ_i) be the eigenvectors and eigenvalues of $\Sigma_{\mathbb{Q}}$, sorted in decreasing order. Then we can define

$$D\left(\Sigma_{\mathbb{P}} \middle\| \Sigma_{\mathbb{Q}}\right) = \sum_{i,j} \frac{\theta_i}{\lambda_j} u_i \cdot v_j + \sum_i \log \frac{\theta_i}{\lambda_j} - r$$

where r is the rank of $\Sigma_{\mathbb{Q}}$, $\log \frac{0}{0}$ is taken to be 0 by convention in the second sum, and $\frac{0}{0} = 0$ in the first sum.

Now for $\mathbb{P}, \mathbb{Q} \in \Delta^+(S)$, define $D(\mathbb{P} \| \mathbb{Q}) = D\left(\Sigma_{\mathbb{P}} \middle\| \Sigma_{\mathbb{Q}}\right)$, and for $\varphi \in S$ define $\mathbb{P}(\cdot \mid \varphi)$ as the distribution which assigns probability 1 to φ and which has minimum KL divergence from $\mathbb{P}$. It is easy to check that this function is concave in the first argument, and that it achieves its minimum at $D(\mathbb{Q} \| \mathbb{Q}) = 0$. Hence this minimization is well-defined.

Moreover, we can implement perform convex optimization over $\Delta^+(S)$. Thus we can compute $\mathbb{P}(\cdot \mid \varphi)$ efficiently.

It is easy to verify the following theorem:

Theorem 4. *If $\psi \in S$ and $\mathbb{P}(\psi) = 0$, then $\mathbb{P}(\psi \mid \varphi) = 0$ as well. Thus conditioning on a sequence of observations $\varphi_i \in S$ results in a distribution which assigns probability 1 to each of them.*

Proof. If the range of $\Sigma_{\mathbb{P}}$ is not equal to the range of $\Sigma_{\mathbb{Q}}$, then there is some v with $v^T \Sigma_{\mathbb{Q}} v = 0$ but $v^T \Sigma_{\mathbb{Q}} v \neq 0$, so $D\left(\Sigma_{\mathbb{P}} \middle\| \Sigma_{\mathbb{Q}}\right) = \infty$.

So if $\Sigma_{\mathbb{P}}$ has a range contained in the subspace with $\mathbb{P}\varphi = 0$, then $\Sigma_{\mathbb{P}(\cdot \mid \psi)}$ must as well, i.e. we must have $\mathbb{P}(\varphi \mid \psi) = 0$. □

Unfortunately, though the analogy with learning theory suggests that this notion of updating might have desirable features, it lacks some of the familiar characteristics of updating. For example, $\mathbb{P}(\varphi \mid \psi)\,\mathbb{P}(\psi) \neq \mathbb{P}(\varphi)$ in general. It seems very unlikely that it will be the final say on the subject, but again we hope primarily to give a sense of how it might even be *possible* for a bounded agent to accept Bayesian epistemology in full generality. Of course, we could also enforce further properties of $\mathbb{P}(\cdot \mid \psi)$, and only perform the minimization of KL divergence over $\mathbb{P}$'s that satisfied those extra properties. For example, we could enforce $\mathbb{P}(\varphi \mid \psi) = \frac{\mathbb{P}(\varphi \wedge \psi)}{\mathbb{P}(\psi)}$ whenever $\varphi, \psi \in S$.

2.4.4 Lift and project

To define $\Delta^+(S)$, we first extended $\mathbb{P}$ to S^2 and then imposed a constraint on the extension. Similarly, we could extend $\mathbb{P}$ to S^{2k}, and consider $\Delta^+\left(S^k\right)$. Even if we are only interested in sentences in S, this extension imposes additional constraints on $\mathbb{P}$ and may allow it to capture additional inferences. This is the lift and project method. The set $\Delta^+\left(S^k\right)$ corresponds to k rounds of the Lasserre hierarchy of relaxations for $\Delta\left(\mathrm{cl}\left(S_0\right)\right)$. This approach allows us to spend more computational resources in exchange for a "more consistent" assignment of probabilities.

3 Logical priors

3.1 Motivation

We are interested in understanding and designing agents which *make good predictions*. Our goal will be to show that if a learner can express an assumption, then the learner will eventually make predictions as well as if it accepted that assumption. Since first order logic is an extremely expressive language, this implies that our learner can make good predictions in a wide variety of situations.

We will typically measure prediction performance by the log score, i.e. the logarithm of the probability that a learner assigns to an outcome. Given a sequence of observations $\varphi_1, \varphi_2, \ldots$, the total score received by the agent is

$$\log\left(\mathbb{P}\left(\varphi_1\right)\right) + \log\left(\mathbb{P}\left(\varphi_2 \mid \varphi_1\right)\right) + \log\left(\mathbb{P}\left(\varphi_3 \mid \varphi_1 \wedge \varphi_2\right)\right) + \cdots$$

This scoring rule has many nice properties and is often used in statistical learning theory. For example, it doesn't matter whether we score the *total* performance of a learner over a sequence of data, or whether we score the learner on *each* prediction and then add up the results.

To construct a distribution which makes "good" predictions (as measured by the log score), we make extensive use of the following simple observation:

Theorem 5. *Let $\varphi_1, \varphi_2, \ldots$ be a sequence of sentences such that $T \vdash \varphi_i$ for each i, and let $\mathbb{P}$ be any coherent probability distribution. Then*

$$\sum_{i=1}^{\infty} \log\left(\mathbb{P}\left(\varphi_i \mid \varphi_1 \wedge \cdots \wedge \varphi_{i-1}\right)\right) \geq \log\left(\mathbb{P}\left(T\right)\right).$$

Proof. For every k, $T \to \bigwedge_{i=1}^{k} \varphi_k$. Thus

$$\mathbb{P}\left(T\right) \leq \mathbb{P}\left(\bigwedge_{i=1}^{k} \varphi_k\right) = \prod_{i=1}^{k} \mathbb{P}\left(\varphi_i \mid \varphi_1 \wedge \cdots \varphi_{i-1}\right)$$

as desired. □

This theorem says that if we want to be able to predict almost as well as if we assumed T, it is sufficient to assign T a high prior probability. Thus our goal will be to define a distribution $\mathbb{P}$ which simultaneously assigns a reasonably high probability to many theories T. A probability distribution which assigns a very low (or zero) probability to a sentence φ is said to be *dogmatic*, while a prior which avoids this characteristic is said to be non-dogmatic.

3.2 Parsimony

Unfortunately, it is not possible to assign *all* consistent sentences a reasonable probability. Indeed, if I supply an uneducated conjecture about the first 100 digits of π, you should assign this conjecture a prior probability which is on the order of 10^{-100}—after all, there are about that many similar yet mutually exclusive alternatives. So if a formal notion of non-dogmatism is to be satisfiable, it needs to avoid judging such low probabilities as dogmatic.

In order to define our prior, we will first introduce a function $\mu : L \to [0,1]$ indicating the least probability which would be reasonable to assign to a sentence $\varphi \in L$. This gives a measure of how quick we should be to infer φ given some evidence in its favor, i.e. μ encodes a choice about which explanations should be quickly inferred. Our view is that the *simple* explanations are the ones that should be quickly learnt; we will not provide philosophical justification for this view here, but note that the issue is a common topic of discussion in formal epistemology.

First we will provide a formal notion of complexity.

Define the complexity of a non-negative integer n as $\mathcal{K}(n) = n + 1$.[5] Extend $\mathcal{K}(\cdot)$ to terms via:

- $\mathcal{K}(x_i) = 2 + \mathcal{K}(i)$
- $\mathcal{K}(c_i) = 2 + \mathcal{K}(i)$
- $\mathcal{K}\left(f_i^k(t_1, t_2, \ldots, t_k)\right) = 1 + \mathcal{K}(k) + \mathcal{K}(i) + \sum_j \mathcal{K}\left(t_j\right)$

Finally, extend $\mathcal{K}(\cdot)$ to formulas via:

- $\mathcal{K}(t_1 = t_2) = \mathcal{K}(t_1 \neq t_2) = 3 + \mathcal{K}(t_1) + \mathcal{K}(t_2)$
- $\mathcal{K}(\forall x_i : \varphi) = \mathcal{K}(\exists x_i : \varphi) = 3 + \mathcal{K}(i) + \mathcal{K}(\varphi)$
- $\mathcal{K}\left(A_i^k(t_1, t_2, \ldots, t_k)\right) = \mathcal{K}\left(\neg A_i^k(t_1, t_2, \ldots, t_k)\right) = 3 + \mathcal{K}(i) + \mathcal{K}(k) + \sum_j \mathcal{K}\left(t_j\right)$

[5] We've made this choice for simplicity, but the notion would be improved by taking the complexity of n to be the length of n in some more efficient encoding than unary. In this case, we would want to change the definitions of $\mathcal{K}\left(f_i^k(t_1, \ldots, t_k)\right)$ and $\mathcal{K}\left(A_i^k(t_1, \ldots, t_k)\right)$ depend on $k + \mathcal{K}(i)$.

- $\mathcal{K}(\varphi \wedge \psi) = \mathcal{K}(\varphi \vee \psi) = 3 + \mathcal{K}(\varphi) + \mathcal{K}(\psi)$

Amongst the sentences with complexity k, there are families of about 2^k mutually inconsistent sentences, and so $2^{-\theta(k)}$ is the strongest lower bound we can give for the probability of k bit sentences using their complexity alone.

Theorem 6. *For every k there are 2^k inconsistent sentences φ_i with $\mathcal{K}(\varphi_i) = \theta(k)$.*

Essentially, we can consider the theory of k bit strings; there are 2^k distinct models each of which can be pinned down by a set of axioms of complexity $\theta(k)$. The proof is in the appendix.

In light of this, we will take $\mu(\varphi) = 2^{-\mathcal{K}(\varphi)}$. It is straightforward to verify that $\sum_{\varphi} \mu(\varphi) = 1$[6], and this will be important in the sequel.

3.3 A simple prior

A number of very direct approaches to this problem have been proposed, each of which aims to ensure $\mathbb{P}(\varphi) \geq \mu(\varphi)$ for some distribution $\mu(\varphi)$ over sentences. For example, Hutter et al. define a prior by choosing a model $\mathcal{M}_\varphi$ of each sentence φ, and then setting $\mathbb{P}(\psi) = \mu\left(\left\{\varphi : \mathcal{M}_\varphi \models \psi\right\}\right)$. This approach is satisfactory, but rests on arbitrary choices and is needlessly computationally inefficient in the finite case. They also have a philosophically different motivation and so restrict attention to separable models, thereby obtaining a distribution which cannot even be approximated with a halting oracle (they also are not concerned with the *speed* of learning). of how quickly a generalization is learned).

Similarly, Demski has proposed generating a compete theory T, starting from $T_0 = \emptyset$, iteratively forming T_{i+1} from T_i by adjoining a random sentence φ consistent with T_i (sampled with probability proportional to $\mu(\varphi)$). This proposal appears to some theoretically desirable properties and to be less arbitrary, but does not lead to tractable algorithms in the finite case.

Our approach is to simply consider the (convex) set of coherent distributions $\mathbb{P}$ and to optimize an appropriate convex function over this set. This more closely mirrors methods which have proven to be practically successful, in particular maximum-entropy priors, and is extremely straightforward to generalize to the finite case. Our function will be chosen to ensure that an optimum is sufficiently far from any boundary, i.e. such that $\mathbb{P}(\varphi)$ is reasonably large for each φ.

For any $\mathbb{P} : L \to [0, 1]$, define

$$\Psi(\mathbb{P}) = \sum_{\varphi \in L} \mu(\varphi) \log\left(\mathbb{P}(\varphi)\right).$$

[6] The sum here is taken over formulas, though we could consider each formula as a sentence by considering the unbound variables as implicitly universally quantified

We would like to take $\mathbb{P}$ to be the coherent probability distribution which maximizes $\Psi(\mathbb{P})$.

Unfortunately, for any coherent $\mathbb{P}$ we will have $\Psi_0(\mathbb{P}) = -\infty$. This is because some φ are contradictions in first order logic, and so will be assigned probability 0 by $\mathbb{P}$. We could exclude these sentences, and in the case of bounded reasoners discussed below this will be adequate. But this doesn't fix the problem in the case of unbounded reasoners; if μ has infinite entropy it might still be the case that all $\mathbb{P}$ have $\Psi(\mathbb{P}) = -\infty$.

Nevertheless, we can compare different candidates $\mathbb{P}$ to see which are better or worse. Namely, for any $\mathbb{P}$ and $\mathbb{Q}$, consider the sum:

$$\Psi(\mathbb{P}\|\mathbb{Q}) = \Psi(\mathbb{P}) - \Psi(\mathbb{Q}) = \sum_{\varphi\in L} \mu(\varphi) \log\left(\frac{\mathbb{P}(\varphi)}{\mathbb{Q}(\varphi)}\right)$$

where we consider $\log\left(\frac{0}{0}\right) = 0$. We say that $\mathbb{P} > \mathbb{Q}$ if the set of negative terms in this sum is absolutely convergent, and sums to less than the set of positive terms. If neither $\mathbb{P} > \mathbb{Q}$ nor $\mathbb{Q} > \mathbb{P}$, we say that $\mathbb{P}$ and $\mathbb{Q}$ are incomparable. The key observation is that there is a unique maximal $\mathbb{P}$, and that $\mathbb{P} > \mathbb{Q}$ for all $\mathbb{Q} \neq \mathbb{P}$.

Theorem 7. *There exists a coherent distribution $\mathbb{P}$ such that for every coherent distribution $\mathbb{Q}$, $\mathbb{P} \geq \mathbb{Q}$.*

We show in the appendix that it is possible to compute $\mathbb{P}$ using a halting oracle. The key observation is that we can use a halting oracle to optimize Ψ over the set of restrictions of coherent $\mathbb{P}$ to any $S \subset L$. As $S \to L$, these approximations converge to the global optimum very quickly, and so we can approximate the global optimum by taking a large sets S.

It seems unlikely this is the final say on a choice of logical priors, but it is simple, does the job, and illustrates a general approach towards constructing logical priors which helps the close the gap between this setting and contemporary work in machine learning. This approach also suggests a path forward for several of the open problems we raise in section 6.1.

3.4 The Gaifman condition

A common coherence condition for probability distributions is the so-called Gaifman condition:

Definition 4. $\mathbb{P}$ satisfies the *Gaifman condition* if for any φ,

$$\mathbb{P}\big(\forall n : \varphi(n)\big) = \inf_k \varphi(0) \wedge \cdots \wedge \varphi(k).$$

This is an intuitive condition for a probability distribution $\mathbb{P}$ which we take to represesent the *truth* of the matter, but it is not a natural condition for a subjective probability distribution.

The distribution we have described does not satisfy the Gaifman condition. This is inevitable: the Gaifman condition is *stronger* than ω-completeness, which cannot be achieved by any definable distribution.

Theorem 8. *For any definable, coherent distribution $\mathbb{P}$ over L which assigns non-zero probability to Q, there is a formula $\varphi(n)$ such that*

$$\mathbb{P}\big(\forall n : \varphi(n)\big) \neq \lim_{k\to\infty} \mathbb{P}\big(\varphi(0) \wedge \cdots \wedge \varphi(k)\big)$$

In particular, this holds for any $\mathbb{P}$ which lies in the arithmetical hierarchy.

Proof. Suppose we have any $\mathbb{P}$ which satisfies the Gaifman condition. Consider the set S of sentences φ such that $\mathbb{P}(\varphi \mid \mathrm{Q}) = 1$. If $\mathbb{P}$ is definable, so is this set. We claim that this is the set of all true statements about the integers, which is not definable.

Note that all theorems of Q, and hence all Σ_1 sentences that are true of the integers, are necessarily in S by coherence.

By applying the Gaifman condition to the sentences $\psi(n) = \varphi(n) \wedge \mathrm{Q}$, we see that if $\mathbb{P}\big(\varphi(k) \bigm| \mathrm{Q}\big) = 1$ for each k, then $\mathbb{P}\big(\forall n : \varphi(n) \bigm| \mathrm{Q}\big) = 1$. Thus all Π_2 sentences that are true of the integers are necessarily in S. Continuing by induction we find that all Σ_k sentences that are true of the integers are in S, and hence S contains all true sentences, contradicting its definability. □

This impossibility result can be considerably expanded, to deal with narrow classes of φ (in fact it has been proved even for Π_1 sentences, at least for distributions which don't assign probability 0 to true Σ_2 sentences) or to deal with weaker convergence conditions.

Moreover, our results on the learnability of universal generalizations seem like an acceptable practical substitute—our algorithm predicts *as well as if* it had learned the universal generalization in the long run.

3.5 Scaling down

Extending the prior we have described to the finite case requires essentially no modification. That is, for any set S, we can define a prior $\mathbb{P}_S : S \to [0,1]$ as the locally coherent map which maximizes

$$\Psi_S(\mathbb{P}_S) = \sum_{\varphi \in S} \mu(\varphi) \log \mathbb{P}_S(\varphi).$$

Because the set of locally coherent distributions is defined by a polynomial list of linear inequalities, we can compute this $\mathbb{P}_S$ in time polynomial in the size of S.

Using standard techniques for semidefinite programming, we can similarly find the $\mathbb{P}_S$ which is sum-of-squares coherent and maximizes this regularizer. Note that in this setting we need not be concerned with convergence issues, as long as we excluded sentences which are contradictions from the sum (since those sentences will necessarily receive probability 0).

3.6 Beyond non-dogmatism

In reality "ignorance" means more than non-dogmatism, and we would like our prior to satisfy further intuitive properties. For example, when a hypothesis *isn't* in conflict with every other conceivable hypothesis, we should be able to assign it a higher probability than $\mu(\varphi)$. Statements which have little logical bearing on each other ought to have little mutual information (and if two statements *are* logically related, conditioning on a natural interpolant ought to reduce the mutual information between them). And so on.

It is easy to check that none of the distributions that have been proposed so far (ours included) matches all or even many of our strong intuitions about an ignorance prior. The recent proposal of Abram Demski appears to come closest; especially interesting is its reproduction of an intuitive conditional independence structure. Unfortunately, as mentioned, his proposal is fundamentally prohibitively computationally expensive. We feel that approaches based on entropy maximization are particularly likely to have such desirable characteristics, and we have presented our solution largely to provide an indication of how a future weighted entropy-maximization approach might work.

4 Learning mathematical facts

We have described a prior over mathematical states of affairs, but haven't yet said much (aside implicitly in theorem 5) about what happens when we use this prior to actually reason about mathematics. In this section we will introduce a model for the process of mathematical reasoning, and then make some observations about the results of such reasoning.

All of the discussion in this section is intended to apply equally to any of the priors discussed in sections 3 and 2. But the arguments are easiest to make and the conclusions are most intuitively plausible in the setting of section 2.3.2, and so for concreteness the reader may want to keep this example in mind. (Extending some of the results to the setting of section 2.4.2 would require proving some further facts about the conditioning process proposed there.)

4.1 Modeling reasoning

In this section we will turn our attention to *passive* reasoning. That is, we consider a reasoner who is exposed to a sequence of *observations* $\varphi_0, \varphi_1, \ldots$, and conditions on each one in turn. We will be interested qualitative statements about what an agent would come to believe after observing "enough" statements of a certain type. The implicit analogy with human reasoning is normally clear, though we are far from being able to pin down the quantitative aspects of the situation (nor even model the situation for humans well enough that such quantitative statements would be meaningful).

The sentences $\varphi_0, \varphi_1, \ldots$ (as well as the agent's beliefs) don't have any free variables, but they can make arbitrary use of the symbols A_i^k, f_i^k, c_i^k. These symbols might be used to encode information about sense experiences (for example $f_0^1(t)$ might return the data perceived by the agent at time t, so that the agent continuously updates on expressions of the form $f_0^1(t) = s_t$) or to present some mathematical facts to the agent (for example, $f_0^1(x)$ might return the successor of x, and the agent might update on facts like $\exists x : \forall y : y \neq x \to \exists z : z = f_0^1(y)$) or for some other purpose.

When we say that an agent would "come to believe φ," all we really mean is that it would come to make predictions as well as if it had come to believe φ. For example, if we say that an agent has "come to believe Peano arithmetic" we do not mean to suggest that it believes the literal axioms themselves, because we generally have no way to rule out the possibility that (for example) the agent has assigned different interpretations to the symbols (or that it has learned a more powerful theory which can interpret Peano arithmetic). All we mean to say is that the agent has formed *some* internal model of the situation which allows them to make predictions as well as if they believed Peano arithmetic, in the sense of theorem 5.

This entire section considers an "open loop:" the agent's beliefs have no effect on the observations it receives. Realistic reasoning, mathematical or otherwise, often involves interaction between a learner and an environment (indeed, as we discuss in section 5.4, even events within a single mind might be best modeled as interactions between e.g. a goal-oriented agent and a memory subsystem which it can consult as a resource). We will turn our attention to these more complex cases in section 5. Many of the examples in this section have greater practical relevance once we consider interaction.

4.2 Learning with unbounded resources

The examples in this section apply both to finite agents and to infinite agents with globally coherent beliefs. Subsequent sections will focus on unique characteristics of learning with bounded resources, which is of greater interest primarily because of the analogy with human reasoning. We will focus on learning arithmetic, under varying conditions, only for simplicity of presenetation..

4.2.1 Learning arithmetic

In a very simple case, a learner might simply be told true facts about arithmetic, where $+, *, 0, S$ are assigned function and constant symbols from L. For example, the reasoner might condition on many facts of the form

$$SSS0 + SS0 = SSSSS0.$$

In this case, by theorem 5 we can guarantee that the agent will eventually make predictions as well as if it knew the definitions of $+, *, 0, S$.

4.2.2 Learning numerals

We can imagine a more realistic task, in which the agent is still given true arithmetic facts, but they now involve some unidentified constant symbols c_i. For example, the agent might be told:

$$c_2 * c_3 = c_5,$$

or so on. We imagine that there is some real correspondence between the c_i and integers. The question is: can the learner simultaneously learn this correspondence and arithmetic?

The answer is essentially that the learner can if and only if they receive enough information to determine the correspondence in principle. That is, consider the theory T' which includes not only the axioms of arithmetic but also axioms pinning down the c_i. For example, we might have $T' \vdash c_1 = S0, c_2 = Sc_1, c_3 = c_2 * (c_2 * c_2 + c_1)$, in addition to the axioms of arithmetic. The complexity of this theory (in the sense of $\mathcal{K}(T')$ defined in section 3.2) is roughly $\sum_i \log(c_i) + \mathcal{K}(\text{PA})$. So in the long run, the learner needs to get about $\log(c_i)$ predictions wrong about each c_i before it has pinned down its value (which is necessary information-theoretically), and it can do this in parallel with learning the definition of $+$ and $*$.

4.2.3 Arithmetic as an explanation

In the cases we have considered so far, an agent has inferred the laws of arithmetic to account for observations about the objects of arithmetic, i.e. about the truth of arithmetic facts. Much more interesting is the case in which an agent infers mathematical structure to explain not-obviously-mathematical phenomena.

For simplicity, we'll consider an agent which already believes axioms which allow it to reason about binary strings (for example, it has learned an axiomatic characterization of string concatenation and string equality).

Suppose our agent is given observations of the form $f(x_0 x_1 \cdots x_k) = y_0 y_1 \cdots y_k$. The rule f may be a complicated one which the agent cannot hope to learn exactly. For example, this would occur if f was a computation which used more memory than the agent could represent. Or, f might simply be stochastic. In either case, the learner can discover structure in the values of f which made prediction easier. Representing this structure might require introducing arithmetic as an abstraction. This is the route by which we suggest that a realistic learner could come to believe the axioms of arithmetic.

For example, suppose that whenever $f(x) = y$, we have $\sum i x_i = \sum i y_i$. If the learner was able to represent this fact and condition on it, the probability they assigned to the true judgments of the form $f(x) = y$ would be (roughly) k^2 times higher, a significant advantage. But formalizing this constraint without the use of arithmetic is a non-trivial challenge.

We cannot rule out the possibility that $\mathbb{P}$ would find some clever alternative representation of this constraint, but we can show that in general it will find one way or another to make predictions *as well as if* it had learned this constraint.

Moreover, if an agent is tasked with learning *many* such relationships, which can be concisely expressed within a single theory T, then the *total* loss on all of those prediction problems is given by $\mathcal{K}(T)$. This suggests that if there is a mathematical theory which has explanatory power in many domains, the agent will either come to believe it or come to find an equivalently powerful method for making predictions.

4.3 Learning with bounded resources

In the previous sections we have focused on cases where an agent infers arithmetic because the agent's observations are *in fact* constrained by arithmetic. In general we are interested in a more subtle phenomenon, which emerges when we consider bounded reasoners.

For example, consider a human reasoner who already accepts the axioms of Peano Arithmetic, and is concerned exclusively with statements whose truth or falsity will be determined in a finite amount of time—ie., which might actually affect their experiences. In particular, every arithmetical statement of material interest to humans involves only bounded quantifiers. Moreover, every mathematical observation that humans ever make also contains only bounded quantifiers—that a particular theorem has a proof of at most so many lines, that a particular computation returns a particular value after a particular amount of time, and so on. Neglecting uncertainty about physics, we imagine that an account of all true Δ_0^0 sentences would allow us to make any physical prediction whose truth or falsity can be determined by physically possible operations in finite time (this is, in essence, the Church-Turing hypothesis).

All of the facts of interest to this human, or which this human can observe, are already determined by her belief in the axioms of Peano arithmetic. So in what sense can any of these observations shed light on more complicated claims? On the existence of infinities, or the truth of Σ_2^0 sentences?

We suggest two possible routes, and give examples of each in the sequel:

1. First, a theory concerning infinite objects or complex axioms might be a simple explanation for certain finite data. For example, the existence of a "set of all natural numbers" may be a simpler explanation, at least to a certain way of thinking, than the axiom schema of induction. The real existence of a continuum might be a simpler explanation for our physical observations than the existence of a computation which approximates the solutions of certain equations on discrete approximations to a continuum.

   In this case a reasoner using the approach we have described might discriminate between theories with no testable distinctions on the basis of simplicity, and the theories that concern themselves with finite objects need not win.

2. Much more fundamentally, although all truths about finite objects might be determined by the axioms of Robinson Arithmetic the implications might be too computationally expensive to examine. A theory which simply explains many observations which would

otherwise require laborious computation may be considered to be confirmed by those observations.

4.4 Computational expedience

4.4.1 Generalization

Consider a learner which knows Robinson arithmetic, and has made some observations of an unknown function f. It has inferred that $f(0) = 1$ and in general $f(Sx) = 2f(x)$, i.e. that $f(x) = 2^x$. (We'll write 2^x for $f(x)$ from here on out.)

Suppose the agent is now asked to predict whether $2^y 2^x = 2^x 2^y$ for some large values of x and y. In some cases the agent can derive this equality using sentences within S, simply by applying the inductive definitions of exponentiation and multiplication until each side has been reduced to a numeral. In these cases the agent will clearly assign probability 1 to $2^y 2^x = 2^x 2^y$. But in general the intermediate steps need no lie in S, because they involve very large numbers.

Of course, if the agent knew the generalization that $\forall a, b : ab = ba$, it could deduce as a special case that $2^y 2^x = 2^x 2^y$. For each x and y, $2^y 2^x = 2^x 2^y$ follows from the axioms of Robinson arithmetic. Nevertheless, a bounded agent who believed Robinson arithmetic might not be able to deduce the fact, and so might find the generalization useful.

However we can reason identically as before to show that an agent will quickly *learn* this generalization: $\mathbb{P}(\forall a, b : ab = ba)$ is lower-bounded by the complexity of the assertion, and so after observing many consequences which it cannot deduce without the generalization, eventually the learner will start predicting those consequences correctly (very likely by assigning high probability to $ab = ba$ itself).

But suppose this game continues, and the learner is now asked to determine whether $2^{x+y} = 2^x 2^y$ for some large integers x and y (for example, $x = 2^a$ and $y = 2^b$). Again, the learner cannot deduce this fact directly, but with enough examples will come to believe the universal generalization.

It may seem as though these examples are slightly artificial because they rely on such large integers. In fact the large sizes of the numbers involved is an artifact of an arithmetic encoding (which we have chosen only to make the examples simple). In a more realistic application, rather than very large numbers we might have modestly sized objects, such that the combinatorial explosion keeps the objects from being amenable to reasoning about directly. A typical case might be one in which an agent reasons about the result of a brute force search, where the brute force search covers more items than those the agent can reason about directly.

Another realistic class of examples might occur when the agent is reasoning about some object x has some special *physical* significance—for example, "the number I've just written down on a sheet of paper"—rather than having some special mathematical significance. This could simply be represented by an additional predicate A, rather than requiring any new machinery to cope with physical facts. We will explore this setting in the section 5.4 below.

4.4.2 Induction

In the last section we described why an agent might infer $\forall a, b : ab = ba$ or $\forall x, y : 2^{x+y} = 2^x 2^y$ from observations about finite objects. And of course there could be a long litany of such generalizations which might be inferred for the sake of computational expedience. Inferring such a long list of computational shortcuts is more efficient than simply guessing about the result of individual computations. But there is a still more efficient approach.

If the agent assumes the axiom of induction, then it would immediately conclude that both of these assertions are true in general (along with many more), since they are both easy theorems of PA.[7] Thus after observing many such generalizations, the agent would come to infer induction (or at least, it would come to make subsequent generalizations *as well as if* it had learned induction).

4.4.3 And beyond

Since few natural statements are independent of Peano arithmetic, it is natural to conjecture that PA is as far as this process will take us. In fact, since almost all "natural" statements can be derived in even weaker systems, we might suspect that we can't even learn PA unless we consider some very pathological observations, unless PA happens to be simpler (and thereby receive a higher prior probability) than weaker theories like RCA_0.

In fact this impression is misleading. For example consider an agent checking whether each integer n is a perfect number, one at a time. It would quickly form the conjecture that odd numbers are not perfect, whether or not it could actually find a proof of the fact. This would occur even if this statement is a theorem of PA; thus even theorems of PA might prove to be *useful* generalizations, which might be accepted as additional axioms.

Continuing in this vein, basic theorems of analysis might be most easily derived from assumptions about the existence of infinities or continua (which can be characterized by simple sets of axioms). Any theorem about strictly finite objects which can be proven by analysis can *also* be proven by a brute-force approach (or proven from the axioms of Peano Arithmetic), but such a proof might be considerably more difficult than one that proceeded from the existence of a continuum. Even in very simple cases, for example when we are evaluating computable functions at rational points to finite precision, the machinery of calculus rests on relatively few axioms and may be a useful aid for some agents who are not sophisticated enough to rebuild the same machinery starting from basic arithmetic.

[7] The axiom of induction is not itself a single axiom, but rather an axiom schema. The system we have described cannot learn an axiom schema in a straightforward way, but this is primarily a technical obstruction. We can work with any finitely axiomatizable extension of PA, such as NBG set theory, and conclude the same results.

Even if the agent believes that Peano arithmetic is sound and eventually discovers that everything it knows about real analysis can be derived within Peano arithmetic by alternative means, this won't eliminate the agent's confidence in Peano arithmetic. The existence of a continuum gives a simple account of *why* facts about analysis should be true, and (assuming the soundness of Peano arithmetic) therefore gives an account of why those facts are consistent with Peano arithmetic.

Of course it is very difficult to say what the end result of such a learning process would be; our point is merely that even if very weak theories suffice to prove all ordinary theorems, there is no reason why our learner would stop there. Indeed, its reasons for accepting stronger theories would be quite similar to the historical motivation for developing such theories: to provide an elegant unifying framework, and to allow for simpler or more direct arguments than would otherwise be possible.

4.5 Metamathematical examples

4.5.1 Consistency of a theory

In order to reason about its own reasoning, an agent might be particularly interested in understanding the consistency of an axiom system which it currently accepts. It is well known that no powerful theories can answer this question by deductive means; but by reasoning *inductively*, an agent might be able to arrive at high confidence in the consistency of a certain axiom system.

Suppose that an agent is interested in the consistency of PA, and makes observations of a mathematical community which is generating theorems about PA. For simplicity, suppose these observations take the form of "at time T a proof of φ is published, which has length k." The agent can infer some properties of publications; perhaps it will infer that there is a notion of mathematical importance, and that the community searches for proofs in PA of important statements and publishes them. In reality the agent would make a great many additional observations, for example about the structure of the proofs *etc.*

If PA is inconsistent, then a reasonable model for the mathematical community's output is expected assign a non-negligible probability to a proof of falsehood being published (or at least to proofs of some contradictory propositions being published). Thus the consistency of PA makes the bold prediction "the published proofs will not demonstrate contradictory facts." Given publications verifying φ and ψ, the agent can check for itself that φ and ψ are not (easily) shown to be inconsistent. Thus the agent can make repeated tests of this bold prediction, which will tend to provide support for the consistency of PA.

Of course, there are other explanations for this observation: for example, it may be that the shortest proof of inconsistency is quite long, or that the search process used by mathematicians is unlikely to turn it up for some other reason, or simply that mathematicians have a norm against publishing any proof that is easily seen to contradict a known result. There are a number of reasons that the consistency of PA might stand out as a particularly promising hypothesis:

1. "There are no proofs of $\perp$" is a simpler assertion than "Every proof of $\perp$ has length at least k" for some large integer k, or other statements of the form "Every proof of $\perp$ has property P". The total prior probability of statements of the latter type is likely to be comparable with the prior probability of the consistency of PA; and most statements of the latter type fail to make useful predictions, because the property P does not explain why such proofs are not published by the mathematical community.
2. The consistency of PA is a consequence of other, potentially more natural statements, such as the axioms of set theory. Some of these premises may be simpler than the consistency of PA and thereby receive higher probability. Other premises (such as the validity of ϵ_0 induction) might be supported by other deductive arguments or other useful generalizations, and thereby receive high probability despite being harder to describe than the consistency of PA.
3. The agent has access to other lines of evidence, for example its own experimentation with simple theorems of PA, which might discriminate between the consistency of PA and other explanations for similar phenomena.
4. Many other useful generalizations may be contingent on the consistency of PA. For example, other generalizations about which statements are "hard" to prove, or coherent simple models for mathematical importance (or mathematicians' decisions about which proofs to publish) might only be coherent if PA is consistent. At least, the complexity of these statements might need to be increased to accommodate the inconsistency of PA. So to the extent that these generalizations are useful, they also provide evidence for the consistency of PA.

Similar arguments might cause an agent to suspect not only the consistency of PA, but the unprovability of more complex assertions φ like the twin prime conjecture. However, this suspicion would naturally be much weaker:

1. Many of the arguments given above are particular to consistency and do not apply in the case of the unprovability of a general φ, and therefore the strength of evidence in favor of consistency is greater than the strength of evidence in favor of unprovability of φ.
2. The consistency of PA is a simpler theory (and hence is likely to have a higher prior probability) than the unprovability of φ (which also requires specifying the statement φ).
3. The probability of a proof of φ appearing given the provability of φ is much lower than the probability of a proof of $\perp$ appearing given the provability of $\perp$. This is due both to the increased simplicity of $\perp$, and due to the fact that a proof of $\perp$ would lead to a proof for all other propositions (and hence would be expected to have other effects on the community's output, even if the proof of $\perp$ itself wasn't published). As a result, the non-appearance of a proof of φ is weaker evidence for the unprovability of φ than the non-appearance of a proof of $\perp$ is for the consistency of PA.
4. Even if a proof of φ exists, induction suggests that in general it will be quite long; if there is a proof of $\perp$, there is little *a priori* reason to think the same thing. If the

> community's output is more likely to find short than long proofs, this again makes the non-appearance of a proof of φ less evidence than the non-appearance of a proof of $\bot$.

As in the last section, we must stress that it is difficult to predict the actual results of induction when applied to this case. Nevertheless, we feel we have given some indication that this framework is capable of "learning" the consistency of a theory. Moreover, we feel that the considerations which are relevant to this framework are quite similar to those which are relevant to an informed human's judgment about the consistency of a theory.

4.5.2 Large cardinals

The current status of large cardinal axioms strikes us as another domain where "inductive" mathematical reasoning is currently playing a significant role. Deductively, it is clear that large cardinal axioms represent further assumptions: assuming consistency of the theories involved, there is no way to move from PA to ZFC, or from ZFC to stronger large cardinal axioms. Nevertheless, it seems that mathematicians are able to form beliefs about such axioms based purely on observations of finite objects—the success or failure of certain searches for proofs.

It seems likely that the kind of inductive reasoning we describe here plays an important role in forming these beliefs. Our framework seems unlikely to be expressive enough to capture all of this reasoning, but once again we can work through the *kinds* of considerations that would lead our framework to come to a tentative conclusion about large cardinal axioms. To the extent that these considerations seem to capture the considerations that are relevant for humans reasoning about large cardinals, this provides some evidence that it will be possible to make headway on formalizing "informal" mathematical reasoning in this domain.

The first key observation about large cardinal axioms is that they explain the consistency of large cardinal axioms.[8] The consistency of a large cardinal axiom in turn explains why attempts to derive inconsistencies have failed—see the discussion in the preceding section.

But this does not capture the full strength of the evidence in favor of large cardinal axioms. Another source of evidence is inductive generalization from the consistency of weaker large cardinal axioms to the general consistency of large cardinal axioms.

To be more precise, a reasoner who accepts many particular large cardinal axioms $\varphi_1, \varphi_2, \ldots$ will by nature gravitate towards some explanation for why all of these axioms have proven to be consistent. If there was some grand theory T which explained this fact, then our reasoner would eventually assign each large cardinal axiom a high probability given its predecessor. On the other hand, the existence of such a grand theory would also lend evidence to each individual assertion φ_i, because now T itself constitutes a particularly parsimonious explanation for the failure to deduce inconsistencies from particular large cardinal axioms.

[8] Given an appropriate account of self-verification, they might explain their own consistency. But until such an account is available, we can content ourselves with the observation that each such axiom explains the consistency of all *weaker* large cardinal axioms.

Human mathematicians are quite uncertain about what such a theory T might look like, and moreover it may be that there is no finitely axiomatizable theory which has the desired characteristics. But whether or not there is such a theory T, the existence of syntactic and mathematical relationships between the axioms $\varphi_1, \varphi_2, \ldots$ suffices to make generalizations over them; indeed, we can posit the existence of an unknown grand explanation, and infer some of the characteristics of this explanation (and therefore use it to make judgments) from our observations. This works even if we cannot pin down the grand explanation itself.

5 Interaction

5.1 Modeling interaction

Rather than considering a passive learner receiving evidence from a teacher, we can consider an active learner interacting with a more general environment.

To keep the exposition simple, for now we'll assume that the agent and environment interact via a binary channel; at each time step the agent specifies a bit corresponding to its action and the environment responds with a bit (the agent's observation). Formally, the environment is represented by a special function symbol $\mathcal{O} : \{0,1\}^* \to \{0,1\}$ and the history of the agents actions is represented by $x : \mathbb{N} \to \{0,1\}$. At time t the learner chooses $x(t) \in \{0,1\}$ and then is "told" $\mathcal{O}\left(x(0) \cdots x(t)\right)$. Formally, if the agent selects $x(t) = x$ and the environment responds with $\mathcal{O}\left(x(0) \cdots x(t)\right) = y$, then the agent updates on the sentences $x(t) = \underline{x}$ and $\mathcal{O}\left(x(0) \cdots x(t)\right) = \underline{y}$, where $\underline{x}$ and $\underline{y}$ are terms in the language representing x and y.

To actually specify a model, we need to describe the behavior of the learner. In a realistic application we might imagine that the learner has some goals unrelated to learning, and chooses an action to satisfy those goals (which might incidentally require learning about mathematics). Below we sketch a few simple rules of action:

- **The greedy learner:** The greedy learner is interested in some proposition φ (or family of propositions, for example concerning the value of a function f), and greedily selects the query x_t for which its distribution over $\mathcal{O}\left(x_1 x_2 \cdots c_t\right)$ has maximum mutual entropy with its distribution φ. Equivalently, it chooses the query such that the expected entropy of its beliefs about φ after learning the result of the query is minimized.
- **The patient learner:** The patient learner has a fixed lifetime T and a question of interest φ. It choose a policy which minimizes the expected entropy of its beliefs about φ at time T.

  This can be done using dynamic programming, essentially considering a brute force search over all sequences of observations and actions. For any sequence of observations at time $T-1$, the agent can use the greedy learner's policy to evaluate the optimal action. This allows the agent to estimate the value it will secure if it makes any particular sequence of $T-1$ observations, which can be used to determine the optimal action for any sequence of $T-2$ actions. Proceeding in this way, the agent can determine the optimal initial action and then make it.

- **The utility maximizer:** the utility maximizer has a fixed lifetime T and a function term $U : \{0,1\}^T \times \{0,1\}^T \to [0,1]$ representing its utility function.

  Using a similar dynamic programming approach, it selects the policy which maximizes $\mathbb{E}\left[U\left(x_1 x_2 \cdots x_T, y_1 y_2 \cdots y_T\right)\right]$.

  This definition depends on considerably more subtleties than those preceding it. For example, note that the obvious generalization of the agent described in the preceding section would choose its actions to maximize its *expectation* of the value of U, and so a priori it might take actions which caused it to believe that U was large rather than actions which caused U to be large. This is only non-problematic because of the martingale property of Bayesian beliefs—a Bayesian can never expect an observation to change their estimate of U in a particular direction.

  Relatedly, the agent will eventually learn a model for the observations $x(t)$ themselves. This causes the agent's actions to constitute evidence to itself (namely, if the agent outputs 0 and then updates on this fact, it has now learned that a certain algorithm defined with respect to a halting oracle outputs 0). This evidence could influence the value of U, and would cause the simplest definition of this agent to behave as an "evidential" decision-theorist.

  Discussing these subtleties would take us outside the scope of this paper. As usual, we merely pause to observe that having a formal model in hand for such goal-oriented behavior seems to open up a wide range of previously philosophical questions to a more technical analysis.

5.2 Planning

All of the agents described in the last section plan for the future by explicitly considering all possible plans. Computationally bounded agents, or agents which have infinite time horizons, must take a different approach.

In principle, we can reduce the planning problem to the epistemic problem of evaluating the quality of a state. If we had access to particularly accurate evaluations of the quality of intermediate states, we could make good decisions by looking only one step ahead. Moreover, the quality of a state has a simple mathematical form in principle (regardless of how difficult it is to reason about): the quality is simply the expected utility obtained conditioned on that state occurring (if we take a state to also include the history leading to that state).

Reasoning successfully about this quantity is quite difficult, and in practice a range of domain-specific heuristics are often used. However, these heuristics can be given a unifying account as effective approximations to the "real" value of a state, which a sophisticated agent might be able to learn by a combination of inductive and deductive reasoning (in the same way that humans originally learned these functions). For example, evaluation functions for chess positions can (in principle) be learned as predictors of who will win a chess game given reasonable play between the two sides. The only significance of our formalism here is the observation that the "real" value of a state can be given a precise mathematical characterization, and so an agent capable of human-level mathematical reasoning would be able to use this very general formulation.

Reasoning about this quantity naturally encapsulates planning, heuristic evaluation functions, and exploration vs. exploitation. For example, an agent might discover a simple plan beginning with action a, and then take action a based on the knowledge that its future-self will do something *at least as good as* the simple plan it has identified. Or an agent might take an action based on the belief that it will be better-informed in the resulting state, and so will make better choices. Note that this formalism also immediately leads the agent to make use of computational aids in its environment in order to help construct plans—there is no distinction between its knowledge about the environment and its knowledge about the value of intermediate states, and plans are evidence that clarifies the value of intermediate states.

Formally (again restricting attention to infinite agents), we can define $x(0), x(2), \ldots$ and $y(t) = \mathcal{O}\big(x(0) \cdots x(t)\big)$. By quining, we can ensure that the agent knows a description of its own decision procedure (for more realistic approaches which are applicable to feasible agents, see the section on introspection below), and we can add the additional axiom that x_t is defined from $x(0), \cdots, y(t-1)$ and $y(0), y(1), \ldots, y(t-1)$ by using this decision procedure. The agent has utility function U, which is a real-valued term depending on x and y, then its decision procedure is given by:

$$x_i = \underset{b \in \{0,1\}}{\operatorname{argmax}} \mathbb{E}\left[U \middle| x_i = b \wedge \bigwedge_{j<i} \Big(x_j = \underline{x_j} \wedge y_j = \underline{y_j} \Big) \right],$$

where $\underline{x_j}$ is either the symbol 0 or the symbol 1 depending on whether x_j was 0 or 1.

This proposed algorithm can be straightforwardly instantiated using a halting oracle, and in the following sections we will discuss elaborations which allow it to be instantiated by a potentially tractable algorithm.

Unfortunately it seems unlikely to yield good behavior, for (at least) the following reason: in order to reason about the value of a state, the agent must reason about the value they will be able to obtain starting from that state, which in turn requires reasoning about the accuracy of the judgments that they will make. This may be straightforward for *particular* judgments—if the agent believes that φ has probability 2/3 and can determine by introspection that it assigns φ probability 2/3, then it will typically believe that its judgment about φ is reasonable. But the entire point of this approach to planning was to prevent the agent from needing to consider every particular contingencies that would arise in the future. Doing so requires reasoning about the quality of a *generic* future judgment. For example, in order for the agent to believe that it will take actions *at least as good* as a simple plan that it has identified, it must believe that its future self won't predictably make errors in judgment. In order for the agent to believe that acquiring more information is useful, it must believe that its future self will make better judgments given more evidence, ideally without needing to consider every possible piece of evidence that might be received.

Unfortunately, our system does not necessarily believe any of these self-confidence assertions. Indeed, in the classical case of deductive logic, such self-verification axioms typically lead to self-referential paradoxes. It seems quite plausible that the probabilistic reasoners will not have these difficulties; a probabilistic reasoner can acquire inductive evidence in favor of their own reliability, and it is still possible that there are "approximately self-verifying" logical priors.

See the section on self-verification below for further discussion of these issues.

5.3 Limited memory

In the case of unbounded computation, we considered agents who remember the entire history of their interaction with the environment, and reason about the outcomes of all future interactions. In the case of bounded computation this becomes impractical: we are interested in considering agents whose lifetimes are longer than they can directly reason about. For example, a human can reason directly about the *length* of their life, but cannot simultaneously hold all of the experiences of their life in their head.

In this case it may be useful to consider agents which have a *limited memory.* In addition to giving our agents a distinguished symbol for the environment $\mathcal{O}$, we can provide them with a distinguished symbol t for the current time. Rather than having beliefs about all values x_i, y_i, the agent might keep track only of those which have occurred recently or will occur soon: $x_{t-k}, y_{t-k}, x_{t-k+1}, y_{t-k+1}, \ldots, y_{t+k}, x_{t+k}$. For for $i < t - k$, such an agent no longer has any belief about the value of y_i (though they may still have beliefs of the form "the last time I saw Alice it was in Detroit," *etc*). Note that this occurs automatically if we make use of the bounded quantifier depth agent in section 2.3.2.

The main difficulty with such proposals is updating the agent's beliefs as time passes. This can be done by first shifting the agent's beliefs to reflect the change in the value of t, updating on the observation, and using the framework of KL divergence discussed in section 2.4.2 to extend the agent's beliefs to the new sentences introduced by the increase in t. Formally, if at a certain time the agent takes action x^* and observes y^*, then we update the agents' beliefs as follows:

1. First, we form a new set S', by adjoining a new symbol t'. For each $\varphi \in S$, we include both φ and $\varphi\,[t = t']$ in S'. We also include the statement $t = t' + 1$. We may also include some additional statements in S' to help pin down the relationship between t and t'. For example, we might include all of the sentences with some fixed quantifier rank, or we might take the closure of S'.

2. Now we do a preliminary transfer of the agent's beliefs to S': for each $\varphi \in S$ the agent's beliefs are unchanged, the sentences $t' = t + 1, x_{t'} = x_{T+1}, y_{t'} = y_{T+1}$ are assigned probability 1, and the agent's beliefs about other sentences are left undefined.

3. Now we take the coherent distribution over S' minimizing the KL divergence from these preliminary beliefs. This is not completely straightforward, because the preliminary beliefs do not assign probabilities to some statements. We would like to consider the "relative" entropy only for those statements where both distributions assign probabilities, and consider the absolute entropy everywhere else. We can emulate this by defining the relative entropy from $\mathbb{Q}$ to $\mathbb{P}$ as the smallest relative entropy from any coherent extension of $\mathbb{Q}$ to $\mathbb{P}$ (which turns out to have a simple algebraic form).

These definitions can be applied either in the setting either of section 2.3.1 or section 2.4.2.

4. We discard every sentence including t from the agent's beliefs, and then replace every instance of t' with t.

It is easily proven that such an agent's beliefs are correct about the observations which it still "remembers." After the observations have been forgotten, their consequences might still be remembered (for example, if I observed a proof of an interesting fact in the environment, I would remember the fact even after the observation faded).

5.4 External memory

Although we have described this constraint as "limited memory" it is worth mentioning that it also arises from a limitation on computational resources. In the framework we have described, if an agent can reason about any sentence $\varphi(x)$, it will typically *automatically* be able to do a brute force search over all y with $|y| \sim |x|$.

In many realistic situations this is not the case. For example, it is quite natural for me to write down the number 2860486313 without having the time to determine whether this number is prime.

Modeling this in our framework (or building an agent which is capable of such reasoning using our framework) requires formalizing the concept of external memory. In fact this is an automatic consequence of the notion of limited memory defined above, but it is important enough that it is worth calling out separately.

For illustration, consider an environment $\mathcal{O}$ which implements a Turing machine tape. That is, at each step the agent observes the contents of the current tape cell, and can write new contents (from some alphabet Σ) as well as move to the next or previous cell.

We'll imagine the case where the agent is given a description of the environment axiomatically, but in principle an agent could also infer such a description from observations (for example, it will quickly learn that if it writes, moves back, and then moves forward, it will observe the same string it just wrote—eventually it will develop the Turing machine model to explain these observations, and potentially even develop a theory of arithmetic in order to accommodate the Turing machine model).

Introducing such an environment allows an agent to reason about objects *indirectly*. For example, an agent might have beliefs about the integer encoded in a certain region of the tape, despite the fact that this integer is too large for the agent to reason about directly.

By reasoning in this way, the agent can perform computations on these objects and reason about the results of those computations (and infer things from observing the results of the computations) without ever being able to directly reason about the objects in question. For example, the agent can simulate a Turing machine M by maintaining the simple belief "There is a time t such that the environment's state is the result of running M for t steps, and my current location is the location of the Turing machine tape head at time t." If the agent then follows the next step of the policy defining M, then this belief will be preserved. If eventually the agent discovers that the policy defining M represents accepting or rejecting, then the agent will (correctly) believe that M accepts or rejects—without being able to represent any of the intermediate steps taken by M.

Similarly, an agent could "write down" some numbers and reason about their properties even though it cannot hold them in memory. It could verify that a number written in a certain place is prime by carrying out the steps of a primality-testing algorithm, even without being able to write down the number itself. And so on.

This provides a more realistic source of examples for the computational utility of sophisticated mathematical theories. For example, an agent might be able to perform some series of steps which resulted in it believing "There exist numbers a and b such ab is written down in register 1 and ba is written down in register 2," even if a and b are much too large for the agent to reason about directly. The commutativity of multiplication would then lead to the prediction that the two registers contain the same numbers, and in particular that every observation of one register will yield the same result as the same observation of the other register. (In fact the agent could come to suspect the commutativity of addition before it even finished looking at the registers, after it observed that the two numbers had many bits in common.)

Although we've described external environments that provide such scratch space, and it is natural to think in terms of the analogy with external computational aids, the "environment" need not be external. For example, we could design a brain as a system with two components: an agent, and a useful computational environment that the agent interacts with. The non-agent part of the brain could be used to read and write memories, to perform specialized processing tasks, or whatever else.

5.5 Introspection

We briefly mentioned the possibility that an agent using a halting oracle can be made aware of their own decision procedure via quining. The reason for this is that the agents we have considered have a compact description, and they have conditioned on a finite set of observations. There are some additional subtleties, for example we must quine not only the agent but also all observations, and this actually changes the content of the agent's updates (when a self-aware agent updates on φ it not only updates on φ, but on the fact that it updated on φ, *etc.*):

This approach is not satisfactory for bounded agents with limited memory. In particular, in order to describe the current state of an agent whose beliefs are a map in $\Delta^+(S_0)$, we need to either represent an entire map $S_0 \to [0,1]$ (which certainly cannot be described by a sentence in S_0!) or represent all of the observations $\varphi_1, \varphi_2, \ldots$ on which the agent has conditioned (which requires too much memory after more than a few observations).

Of course, an agent can still manipulate a mathematical representation of their current beliefs: they are the beliefs at time t of an agent who begins with a certain prior and then forms their beliefs by conditioning on observations. However, the agent is now ignorant of its own characteristics. To the extent that understanding its future behavior is important for planning, we need to attend to the agent's self-knowledge.

There are at least three plausible approaches by which the agent could come to know facts about itself:

1. The agent can use deduction to infer a limited set of characteristics about itself. For example, it knows that it has *some* beliefs and takes optimal actions with respect to those beliefs, even if it doesn't know all of its beliefs, and this knowledge can be quite useful for planning.
2. After making a decision $x(t)$, the agent conditions on the fact that it took this action. This allows the agent to build up a simplified self-model in the same way that it builds a model of the environment, and to inductively infer generalizations about its own behavior which can then be used for planning.
3. We can provide the agent with an environment which allows *explicit* introspective access. For example, we could allow the agent to write down a proposition φ in external memory, and then ask the environment to tell it the agent's current beliefs about φ. This might correspond to the agent imagining a hypothetical situation and then inferring their own response to that hypothetical situation, which of course could be an important input into planning.

5.5.1 Avoiding quining

Even the discussion of "bounded introspection" above relies on quining: the agent is given a compact description of its own initial dynamics, and defines its current state as the result of evolving those dynamics up until time t. Though this imposes only a "constant" requirement on the agent's beliefs (they must be complex enough to describe the agent itself), this requirement might still be prohibitively difficult to meet. Methods based on quining also appear to be *particularly* and egregiously psychologically implausible.

An alternative approach is to simply provide the agent with special symbols which refer to its own characteristics. For example, we might provide the agent with a symbol $\mathbb{P}$ referring to its own prior distribution, or a symbol $\mathbb{P}_t$ referring to its beliefs at time t, or a symbol A referring to the decision of the agent when it has beliefs t.

We can then provide rules bridging between these special symbols and the agent's observations. For example, we can enforce the axiom $x_{t+1} = A(\mathbb{P}_t)$, and $\mathbb{P}_{t+1}(\varphi) = \mathbb{P}_t\left(\varphi \,\middle|\, \ulcorner x_t = \underline{x_t} \wedge y_t = \underline{y_t} \urcorner\right)$. With these rules in place, updating on x_t gives the agent information about A and $\mathbb{P}$. (And as usual, the agent can proceed to form inductive generalizations about these functions even if their exact form is too complex for it to understand.)

We can also add some axioms constraining A and $\mathbb{P}$ which are not as complex as their full specification. And we can provide opportunities for introspection, as long as the agent is given axioms relating the output of introspection to the symbols A and $\mathbb{P}$ (or whatever other symbols we supply).

6 Conclusion

6.1 Further work

We have painted a very crude picture of probabilistic reasoning about mathematics. On almost every front there are plausible directions for improvement:

6.1.1 Choice of priors

Perhaps the most pressing question is what prior distribution over states of affairs is appropriate, especially when subject to computational limitations. We have described an ad hoc distribution which seems to be efficient and to support some nice properties, but which has little theoretical justification and does not perform sensibly in all cases. It seems quite likely that there is a more compelling answer to this question.

An intuitive approach to the problem is based on maximum entropy methods. One problem with this approach is that it does not result in reasonable probabilities for sentences like $\forall x : \varphi(x)$, and so such generalizations will never be learnt. To address this we can try to define a weighted notion of entropy, representing an ensemble of variables (with some logical constraints) which are of different levels of interest.

Carrying out this definition is not straightforward. One plausible rendition is

$$H_\mu(\mathbb{P}) = \sum_i \mu(\varphi_i) H(\varphi_i \mid \varphi_1, \varphi_2, \ldots, \varphi_{i-1})$$

where the φ_i are arranged in increasing order of $\mu(\varphi_i)$, and $H(\varphi_i \mid \varphi_1, \varphi_2, \ldots, \varphi_{i-1})$ is the expected (under $\mathbb{P}$) entropy of $\mathbb{P}(\psi)$ after conditioning on the truth values of $\varphi_1, \varphi_2, \ldots, \varphi_{i-1}$. This has the downside that the maximizing $\mathbb{P}$ may have $\mathbb{P}(\varphi) = 2^{-\mu(\varphi)}$, which means that learning a generalization may take an amount of time which is exponentially large in the complexity of that generalization.

This choice of H_μ has the interesting characteristic that the maximizing $\mathbb{P}$ satisfies the following minimax property: $\mathbb{P}$ *maximizes* the *worst-case* total log-score on a series of prediction problems, where the intended answers to those problems must be consistent and the cost assigned to a prediction of the truth of φ is weighted by $\mu(\varphi)$.

Unfortunately, this choice of H_μ has the undesirable property that for the maximizing $\mathbb{P}$, $\mathbb{P}(\varphi)$ can be as low as $2^{-\mu(\varphi)}$. This implies that the time required to learn a generalization might be exponential in the complexity of that generalization, which we consider unacceptable.

In the setting of section 2.4.2, an alternative generalization is the quantity $\operatorname{tr}(\mu \log \Sigma_{\mathbb{P}})$, where μ is a trace 1 diagonal matrix whose diagonal entries are $\mu(\varphi)$ (where $\operatorname{tr}(\log \Sigma_{\mathbb{P}})$ would be analog of the unweighted entropy). We do not yet have an understanding of this function or its characteristics.

6.1.2 Probabilistic generalizations

The system we have described is able to make universal generalizations, of the form $\forall x : \varphi(x)$, given enough positive examples of φ (or at least, it is able to predict subsequent values of $\varphi(x)$ *as well as if* it had made such a generalization). Most realistic generalizations are not of this form. This is quite clear in everyday experience, but is also plausible in a mathematical setting; even if the actual output of mathematical reasoning is proofs, such heuristics and probabilistic generalizations may play a central role in the actual practice of mathematics.

So we would like to be able to build a system which can learn probabilistic generalizations, of the form "the probability of $\varphi(x)$ is 2/3, for a generic x" or "a generic k digit number has a $1/k$ probability of being prime."

We can make some crude steps towards this goal by simple heuristics. For example, we can use a number quantifier to learn that the number of x satisfying $\varphi(x)$ is 2/3 of the total number of x (and even without a number quantifier we could explicitly code such a sentence in set theory, for example). But most of these heuristics seem to perform badly even in simple cases, and to fail to capture exactly what we want.

The approach we would find most satisfying would be one in which this constraint entered into the cost function used to determine $\mathbb{P}$, rather than appearing as a logical constraint. To motivate this hope, we consider the case of ordinary Bayesian reasoning. When a Bayesian using a maximum entropy prior receives a piece of evidence for a proposition X which suggests that X is twice as likely as they had previously supposed, their beliefs $\mathbb{P}$ now maximize the function:

$$H(\mathbb{P}) + \mathbb{P}(X)$$

(if H is measured in bits).

Similarly, if we are able to pick $\mathbb{P}$ as the maximizer of some payoff function motivated by entropy, we might be able to impose a linear term which caused $\mathbb{P}$ to assign a higher probability to each expression $\varphi(x)$ (though this tendency could then be overruled by conflicting generalizations and logical constraints). That is, if we accept the generalization "For each x, $\varphi(x)$ is twice as likely as we would otherwise suppose" then our beliefs could be determined by maximizing

$$H(\mathbb{P}) + \sum_x \mathbb{P}(\varphi(x))$$

(as in Theorem 7, we would need to take care in defining the maximum, since the sum would probably be infinite). Then we can recover universal generalizations via "For each x, $\varphi(x)$ is infinitely more likely than we would otherwise suppose."

There are a few challenges with carrying out this program, however. First, we would like to impose this linear term only if $\mathbb{P}$ thinks it is true. So instead of a linear term, we would like to include an interaction term between $\varphi(x)$ and the event "for each x, $\varphi(x)$ is more likely than we would otherwise think." But we need to take care to do this in a way that doesn't create an infinite incentive or disincentive for $\mathbb{P}$ to assign high probability to this soft generalization.

Second, we don't yet have any working formulation of our prior as a suitably modified maximum entropy distribution, and imposing such linear costs in our current framework does not achieve the desired functionality. If this approach is to be successful, it will probably require a different choice of prior.

6.1.3 Self-verification

Our ultimate objective concerning reflective reasoning is the notion of *reflective consistency.* Intuitively, we would like to build systems which consider their own output to be evidence about a claim.

Formally, we would like to write down some algorithm $\mathbb{P}$ such that $\mathbb{P}\left(\varphi \,\middle|\, \mathbb{P}(\ulcorner\varphi\urcorner) = p\right)$ is, if not equal to p, at least pulled towards p. That is, $\mathbb{P}$ should treat the observation of $\mathbb{P}(\ulcorner\varphi\urcorner) = p$ in the same way that it treats the testimony of an informed and wise teacher about φ. Of course, there might be some pathological sentences, such as "the wise teacher thinks this sentence is false," for which updating on the teacher's testimony that the sentence is false actually causes you to believe that the sentence is true.

In fact we may want to go somewhat further; we may imagine that $\mathbb{P}$ is evaluating the quality of its own judgment *in general* rather than regarding a particular proposition. Formally, suppose that c_q is a constant symbol indicating that the agent will need to make a decision about $\varphi\left(c_q\right)$. Then we would like to say that the agent's beliefs about $\varphi\left(c_q\right)$ are accurate in general, given that it had to make a decision about c_q. In symbols, if $\underline{x}$ is the numeral representing x, then

$$\mathbb{P}\left(\varphi\left(c_q\right) \,\middle|\, \mathbb{P}\left(\ulcorner \varphi\left(\underline{c_q}\right) \urcorner \,\middle|\, \ulcorner c_q = \underline{c_q} \urcorner\right) = p\right)$$

should be, if not equal to p, at least strongly pushed towards p.[9] (As before, there may be pathological sentences for which the effect is reversed, and in general we should never expect the probability to be *exactly* p.)

The analogous problem in the case of deductive reasoning is to build a system which can prove that anything it proves is correct. This project is in some sense straightforward: if you can build a system which believes "φ is true" for each of its axioms φ, it can conclude by an inductive argument that everything it proves is true. Unfortunately, it turns out that

[9] It may seem strange to condition on $c_q = \underline{c_q}$, i.e. we might hope that $\mathbb{P}\left(\varphi\left(c_q\right) \,\middle|\, \mathbb{P}\left(\ulcorner \varphi\left(\underline{c_q}\right) \urcorner\right) = p\right)$ is close to p. But this is unlikely to be true in general. For example, if c_q is defined to satisfy $\varphi\left(c_q\right)$ then the agent might assign probability 1 to $\varphi\left(c_q\right)$ but low probability to $\varphi\left(\underline{c_q}\right)$, unless it conditions on $c_q = \underline{c_q}$.

both steps of this plan—defining a notion of truth, and building a system which trusts its own axioms—are essentially impossible.

In the probabilistic setting, we no longer have to contend with these impossibility results (at least if we are willing to allow arbitrarily small uncertainty in our system's self-evaluation). But the project is no longer as straightforward as in the deductive case: what should our system believe about itself, in order to treat its own judgments as evidence? There was a clear rationale for proofs, in that they can be justified by assuming the truth of the axioms. But for probabilities, it is no longer adequate to believe that your absolute assumptions are correct, you need to believe that your prior is correct.

This appears to be a much higher bar. Moreover, a demonstration of a natural self-verifying system would provide some indication that the chosen prior is a good one, in exactly the same way that the fact that proofs provably preserve truth provides an indication that proofs are epistemically solid.

An agent may also come to trust its own judgments on the basis of *inductive* evidence, and this may be in closer accordance with an intuitive picture of our own self-trust. That is, a system might be able to observe its own reliability in a number of cases and infer that it is likely to be reliable in analogous future cases.

This approach also requires further technical work, in order to demonstrate that it is possible for the system to learn the appropriate kind of generalization. So far, proposed systems have not been able to learn these generalizations, either because they lack sufficient expressive power, or because these generalizations would be inconsistent and are hence assigned zero probability. The severity of these difficulties is not yet clear; discussing them at more length would be outside of the scope for this paper.

6.1.4 Shifting attention?

Our finite systems all rely on a fixed set of sentences S with respect to which probabilistic judgments are coherent. Combined with externalizing memory (and the ability to carry out proofs in externalized memory), this may be an adequate basis for reasoning. But it seems plausible that the set S itself should be subject to change as different facts become important to the agent.

Suppose the agent wishes to change its focus from the set S to a set T. *Removing* elements from S is straightforward, we can simply restrict our original coherent distribution to a smaller set of sentences. The key problem is determining what probability to assign to sentences $\varphi \in T \backslash S$. So for simplicity, assume $T \supset S$.

One approach is to simply fix the probability of each sentence of S and to find the completion to T which minimizes entropy. However, this fails to allow an agent to update on the constraints implied by any new sentences that have been included. For example, if I previously assigned $\exists x : \varphi(x)$ a probability of $\frac{1}{2}$ but T contains all of the steps of a proof of $\varphi(17)$, then my probability for $\exists x : \varphi(x)$ should go to 1, not remain at $\frac{1}{2}$. In this case retaining a probability of $\frac{1}{2}$ would lead to incoherence, but in milder cases the sentences in

T might contain evidence about an assertion ψ.

An alternative approach is to extend a set of beliefs $\mathbb{P}$ defined on S to a set of beliefs $\mathbb{P}'$ defined on T so as to minimize the KL divergence from $\mathbb{P}$ to $\mathbb{P}'$ (but not enforcing any other temporal consistency condition).

A further difficulty when considering approaches of this type is handling the introduction of the new statement as an observation. For example, if T includes a proof of $\varphi(17)$, it seems intuitive that this should increase our probability assignment to $\forall x : \varphi(x)$ as if we had conditioned on $\varphi(17)$. But this does not seem to be true for natural resolutions. (this is closely related to the challenges described in the next section).

6.1.5 Paradox of ignorance

One potentially problematic aspect of our approach is the benefit that prior ignorance appears to confer upon a learner. That is, we might expect that an agent operating under a more stringent logical consistency condition, which forced their prior to assign probability 1 to a larger number of (true) statements, would only perform better than its less-informed peer. But in fact we can see that this is not generically true.

Consider a pair of agents, one powerful and one limited, trying to determine the truth of $\forall x : \varphi(x)$ for some Δ_0 formula φ. Each of them has access to an environment which can evaluate $\varphi(x)$, though it takes longer to evaluate a sentence with a more complicated argument x.

To the limited agent, it may be that each new value of $\varphi(x)$ is a surprise, and so constitutes evidence about the generalization $\forall x : \varphi(x)$. If so, it may be able to query nature about some very simple inputs x and thereby obtain a reasonable view of the universal generalization. At the same time, it may be that the more powerful agent is able to deduce $\varphi(x)$ directly for simple x (i.e., it considers a set of sentences S large enough to prove each $\varphi(x)$, and therefore its prior necessarily assigns these statements probability 1). Casually it seems that this should be to its advantage, since it should be able to jump to the same conclusion as the limited agent but without needing to consult the environment. But in fact, because the agent has assigned these statements *prior probability* 1, they don't act as evidence at all—they would be true whether or not the universal generalization $\forall x : \varphi(x)$ were true—and the prior probability of the universal generalization is still roughly $\mu\left(\forall x : \varphi(x)\right)$ Thus the more powerful agent must consult the environment regarding more complex examples, and pay a larger cost, in order to begin to form reasonable beliefs about the universal generalization.

We could respond to this challenge in a number of ways:

- We could define a prior in which the existence of logical constraints influences prior probabilities in the desired way. For example, the existence of a proof of $\varphi(17)$ which forces us to assign prior probability 1 to $\varphi(17)$ might cause us to increase our prior probability of $\forall x : \varphi(x)$, just as if we had observed $\varphi(17)$.
- We could provide an explicit mechanism by which agents can "update" on facts that

they assign prior probability 1. For example, we might imagine an agent's beliefs as being formed in stages, subject to increasingly stringent logical constraints. Then the complex agent might assign prior probability 1 to the simple sentences $\varphi(x)$, yet update on them in an earlier "stage" during which it was ignorant.

- We could conclude that it is not a problem at all, and that it is correct that ignorance can be an advantage in certain situations; we find this problematic but not completely implausible.
- Even if $\mathbb{P}\big(\varphi(x)\big) = 1$ is guaranteed by logical coherence, the *fact that* $\mathbb{P}\big(\varphi(x)\big) = 1$ is guaranteed by logical coherence may not be assigned probability 1, and so we could try to build the agent such that updating on this fact would have a similar effect to updating on $\varphi(x)$ itself.

6.1.6 Implementation

Most of our discussion has been highly theoretical, and our primary interest has been in understanding the nature of mathematical reasoning. Nevertheless, it may be possible to implement the system described in section 2.4.2 in practice, and to apply it to simple problems. There is little doubt that experience with a working implementation would provide a new perspective on these results.

Moreover, the computational time required to deal with a set of n sentences (and their pairwise conjunctions) is $O(n^3)$, which remains easily manageable up through $n = 10^3$ (at which point we may have enough expressive power to be interesting, given a careful choice of 10^3 sentences).

6.1.7 Practical questions

As discussed in section 4.5, there are a number of contemporary mathematical discussions in which inductive reasoning plays a role, and it may be interesting to try to apply formal frameworks like this one to these discussions:

- Applying formal frameworks to current discussions may help us understand the current state of evidence, and clarify discussion.
- Understanding the reasoning which is used in practice may help highlight gaps in a formal framework.

References

[1] Abram Demski. Logical prior probability. In Joscha Bach, Ben Goertzel, and Matthew Iklé, editors, *AGI*, volume 7716 of *Lecture Notes in Computer Science*, pages 50–59. Springer, 2012.

[2] Haim Gaifman. Reasoning with limited resources and assigning probabilities to arithmetical statements, 2004.

[3] Marcus Hutter. A theory of universal artificial intelligence based on algorithmic complexity. Technical report, April 2000.

[4] Marcus Hutter, John W. Lloyd, Kee Siong Ng, and William T. B. Uther. Probabilities on sentences in an expressive logic. *CoRR*, abs/1209.2620, 2012.

[5] Bas R. Steunebrink and Jürgen Schmidhuber. A family of Gödel machine implementations. In Jürgen Schmidhuber, Kristinn R. Thórisson, and Moshe Looks, editors, *AGI*, volume 6830 of *Lecture Notes in Computer Science*, pages 275–280. Springer, 2011.

7 Appendix

Theorem 1. *It is possible to determine whether $\varphi \sim \psi$ in $n \log^2 n$ time, where n is the total length of φ and ψ.*

Proof. First, observe that $\varphi \sim \psi$ iff there is a series of operations, of the sort described in the definition of $\sim$, which transforms φ into ψ. (These operations might be applied to arbitrary subexpressions).

By using a sequence of such operations, we will describe how to transform any expression φ into a canonical expression φ^* in a new language where $\wedge$ is modified to take a set rather than a pair of arguments. We will show that if $\varphi \sim \psi$ then $\varphi^* = \psi^*$, so that computing φ^* provides an algorithm to test for trivial equivalence.

First, we remove all occurrences of $\to, \vee, \exists$. Then we inductively define φ^* as follows:

Case 1: $\varphi \in \{\bot, \top\}$. $\top^* = \wedge_\emptyset, \bot^* = \neg\wedge_\emptyset$, where $\wedge_\emptyset$ is $\wedge$ applied to no arguments.

Case 2: φ is an atom. $\varphi^* = \varphi$.

Case 3: $\varphi = \neg\psi$. If $\psi^* = \neg\xi^*$ then $\varphi^* = \xi^*$. Otherwise, $\varphi^* = \neg\psi^*$.

Case 4: $\varphi = \forall x_j : \psi$. Let x_i be the lexicographically first variable not appearing in ψ^*. Then $\varphi^* = \forall x_i : \psi^* \left[x_j = x_i\right]$.

Case 5: $\varphi = \psi \wedge \xi$. For any expression ζ^*, define

$$S\left(\zeta^*\right) = \begin{cases} S & \text{if } \zeta^* = \bigwedge\limits_{\theta^* \in S} \theta^* \\ \{\zeta^*\} & \text{otherwise} \end{cases}$$

Define $S = S\left(\psi^*\right) \cup S\left(\xi^*\right)$. If there is some $\neg\theta^* \in S$ such that $S\left(\theta^*\right) \subset S$, then $\varphi^* = \neg\wedge_\emptyset$. In this case, we say that φ is false by virtue of noncontradiction.

If $S = \{\theta^*\}$, then $\varphi^* = \theta^*$. Otherwise, $\varphi^* = \bigwedge\limits_{\theta^* \in S} \theta^*$.

It is straightforward to verify by induction that if $\varphi^* = \psi^*$ then $\varphi \sim \psi$. The only challenging step is showing that if φ is false by noncontradiction, then $\varphi \sim \bot$, but this can be done by rearranging the conjuncts of φ appropriately until it is of the form $\psi \wedge (\theta \wedge \neg\theta) \sim \psi \wedge \bot \sim \bot$. By using standard data structures for sets, we can compute $A \cup B$ or $A \subset B$ in time $\log\left(|A|\right)|B|$. This yields an $O\left(n \log^2 n\right)$ time algorithm for computing φ^*, where n is the length of φ (as long as we always merge the smaller set into the larger set).

It remains to show that if $\varphi \sim \psi$, then $\varphi^* = \psi^*$. It is sufficient to check each of the transformations in the definition of $\sim$, and then we can induct on the number of transformations transforming φ into ψ. Moreover, since φ^* depends only on the canonical form of its subexpressions, it suffices to consider the case where $\varphi \sim \psi$ pattern matches exactly with one of the defining transformations, and then we can induct on the structure of φ.) It is routine to check almost all of these rules. Only two present difficulty:

1. $\varphi \wedge \neg\varphi \sim \bot$. In this case, observe that $S\left(\varphi^*\right) \subset S\left(\varphi^*\right)$ regardless of the form of φ^*, and so the expression on the left is false by virtue of noncontradiction and has a canonical form of $\neg\wedge_\emptyset$.
2. $\varphi \wedge (\psi \wedge \xi) \sim (\varphi \wedge \psi) \wedge \xi$. The canonical forms of these expressions are usually equal by virtue of the associativity of set unions. The only possible failure is if one of them is false by virtue of noncontradiction. Let $S = S\left(\varphi^*\right) \cup S\left(\psi^*\right) \cup S\left(\xi^*\right)$. It is straightforward to check that each side of this expression is false by noncontradiction if and only if S contains some $\neg\theta^*$ such that $S\left(\theta^*\right) \subset S$. Since this is the same condition on both sides, these are equivalent.

□

Theorem 6. *For every k there are 2^k inconsistent sentences φ_i with $\mathcal{K}\left(\varphi_i\right) = \theta\left(k\right)$.*

Proof. Consider the theory T_x specified by a binary vector $x = x_0 x_1 \cdots x_k$, defined by the conjunction of the axioms:

$$\begin{aligned}
\forall x, y, z, w :& f^2(x, y) = f^2(z, w) \to x = z \wedge y = w \\
\forall x, y :& A^1(x) \wedge A^1(y) \to x = y \\
& c_0 \neq c_1 \\
& A^1\left(f^2\left(c_{x_0}, f^2\left(c_{x_1}, \cdots f^2\left(c_{x_{k-1}}, c_{x_k}\right)\cdots\right)\right)\right)
\end{aligned}$$

For simplicity, write φ_x for the last axiom. The first three axioms have complexity $\theta\left(1\right)$ since they have no dependence on k. By induction and the definition of $\mathcal{K}\left(\cdot\right)$, we have $\mathcal{K}\left(\varphi\left(x\right)\right) = \theta\left(k\right)$. Moreover, it is easy to verify that any ψ_x and ψ_y are incompatible given the first three axioms. Thus T_x and T_y are incompatible given $x \neq y$. □

Theorem 7. *There exists a coherent distribution $\mathbb{P}$ such that for every coherent distribution $\mathbb{Q}$, $\mathbb{P} \geq \mathbb{Q}$.*

Proof. Observe that $\Psi(\mathbb{P})$ is continuous in the product topology and that the space of coherent distributions is a compact set. Thus if $\mathbb{P}_1 < \mathbb{P}_2 < \cdots$ (or in general if the $\mathbb{P}_i$ form a chain), then $\mathbb{P} = \lim_i \mathbb{P}_i$ (relative to some ultrafilter) satisfies $\Psi(\mathbb{P}) > \Psi(\mathbb{P}_i)$ for all i.

Thus we can find a maximal $\mathbb{P}$. For any $\mathbb{P}' \neq \mathbb{P}$, we must have either $\mathbb{P} > \mathbb{P}'$ or $\mathbb{P}$ and $\mathbb{P}'$ are incompatible. We will show that if $\mathbb{P}'$ is incomparable with $\mathbb{P}$, then $\mathbb{Q} = \frac{1}{2}(\mathbb{P} + \mathbb{P}')$ dominates $\mathbb{P}$, contradicting maximality of $\mathbb{P}$. This implies that $\mathbb{P}$ satisfies the conditions of the theorem.

To see that $\mathbb{Q} > \mathbb{P}$, let S be the set of sentences where $\mathbb{P} > \mathbb{P}'$, and let T be the set of sentences where $\mathbb{P}' \geq \mathbb{P}$. Then

$$\begin{aligned}
\Psi(\mathbb{Q}\|\mathbb{P}) &= \sum_{\varphi\in L} \mu(\varphi)\Big(\log(\mathbb{Q}(\varphi)) - \log(\mathbb{P}(\varphi))\Big) \\
&= \sum_{\varphi\in S} \mu(\varphi)\Big(\log(\mathbb{Q}(\varphi)) - \log(\mathbb{P}(\varphi))\Big) + \sum_{\varphi\in T} \mu(\varphi)\Big(\log(\mathbb{Q}(\varphi)) - \log(\mathbb{P}(\varphi))\Big) \\
&\geq \sum_{\varphi\in S} \mu(\varphi)\Big(\log(\mathbb{Q}(\varphi)) - \log(\mathbb{P}(\varphi))\Big) + \sum_{\varphi\in T} \mu(\varphi)\left(\log\left(\frac{\mathbb{P}(\varphi)}{2}\right) - \log(\mathbb{P}(\varphi))\right) \\
&\geq \sum_{\varphi\in S} \mu(\varphi)\Big(\log(\mathbb{Q}(\varphi)) - \log(\mathbb{P}(\varphi))\Big) - 1 \\
&= +\infty > 0
\end{aligned}$$

as desired. □

7.1 Computing $\mathbb{P}$

In this section we show how to compute the distribution $\mathbb{P}$ defined in section 3 using a halting oracle.

The first observation is that if S is a finite set of sentences, then a distribution $\mathbb{P} : S \to [0,1]$ can be extended to a coherent probability distribution if and only if $\mathbb{P}$ satisfies the coherence axioms when restricted to S and assigns probability 1 to each theorem of PA (this is a straightforward modification of theorem 2). Using a halting oracle, we can straightforwardly test whether a given $\mathbb{P} : S \to [0,1]$ satisfies these properties.

Given a finite set S, we can restrict our potential function $\Psi(\mathbb{P})$ by considering

$$\Psi_S(\mathbb{P}) = \sum_{\varphi\in S} \mu(\varphi)\log\big(\mathbb{P}(\varphi)\big).$$

For any S, let $\mathbb{P}_S : S \to [0,1]$ be the coherent distribution maximizing $\Psi_S(\mathbb{P}_S)$. We will show that for any φ, ϵ we can find a set S such that $\big|\mathbb{P}(\varphi) - \mathbb{P}_S(\varphi)\big| \leq \epsilon$; this implies that we can approximate $\mathbb{P}$ simply by computing $\mathbb{P}_S$ for a sufficiently large S.

Let $\mu(S) = \sum_{\varphi\in S} \mu(\varphi)$, and $\mu\left(\overline{S}\right) = \sum_{\varphi\in L\setminus S} \mu(\varphi)$. To measure the distance between two distributions $\mathbb{P}$ and $\mathbb{Q}$ over the set S, we define

$$\|\mathbb{P}-\mathbb{Q}\|_2^2 = \sum_{\varphi\in S}\mu(\varphi)\left(\mathbb{P}(\varphi)-\mathbb{Q}(\varphi)\right)^2.$$

Theorem 8. *For any coherent $\mathbb{P}_S : S \to [0,1]$, any $\mathbb{P} : L \to [0,1]$, and any $\epsilon > 0$, there is a $\overline{\mathbb{P}_S} : L \to [0,1]$ such that:*

1. *For each $\varphi \in S$, $\left|\overline{\mathbb{P}_S}(\varphi) - \mathbb{P}_S(\varphi)\right| \le \epsilon$.*
2. *$\Psi\left(\overline{\mathbb{P}_S}\middle\|\mathbb{P}\right) \ge \Psi_S(\mathbb{P}_S) - \Psi_S(\mathbb{P}) - \epsilon + \log(\epsilon)\,\mu\left(\overline{S}\right)$.*

In particular, taking $\epsilon = \log\mu\left(\overline{S}\right)$, we have

$$\Psi\left(\overline{\mathbb{P}_S}\middle\|\mathbb{P}\right) \ge \Psi_S(\mathbb{P}_S) - \Psi_S(\mathbb{P}) - \mu\left(\overline{S}\right)\left(\log\mu\left(\overline{S}\right)+1\right).$$

Proof. Let $\mathbb{Q}$ be any extension of $\mathbb{P}_S$ to a coherent probability distribution. Define

$$\overline{\mathbb{P}_S} = (1-\epsilon)\mathbb{Q} + \epsilon\mathbb{P}.$$

Then it is trivial to verify that condition (1) above holds.

Moreover,

$$\begin{aligned}
\Psi\left(\overline{\mathbb{P}_S}\middle\|\mathbb{P}\right) &= \sum_{\varphi}\mu(\varphi)\left(\log\left(\overline{\mathbb{P}_S}(\varphi)\right) - \log\mathbb{P}(\varphi)\right)\\
&= \sum_{\varphi\in S}\mu(\varphi)\left(\log\left(\overline{\mathbb{P}_S}(\varphi)\right) - \log\mathbb{P}(\varphi)\right) + \sum_{\varphi\in\overline{S}}\mu(\varphi)\left(\log\left(\overline{\mathbb{P}_S}(\varphi)\right) - \log\mathbb{P}(\varphi)\right)\\
&\ge \sum_{\varphi\in S}\mu(\varphi)\left(\log\left((1-\epsilon)\mathbb{P}_S(\varphi)\right) - \log\mathbb{P}(\varphi)\right) + \sum_{\varphi\in\overline{S}}\mu(\varphi)\left(\log\left(\epsilon\mathbb{P}(\varphi)\right) - \log\mathbb{P}(\varphi)\right)\\
&= \sum_{\varphi\in S}\mu(\varphi)\left(\log(1-\epsilon) + \log\left(\mathbb{P}_S(\varphi)\right) - \log\mathbb{P}(\varphi)\right) + \sum_{\varphi\in\overline{S}}\mu(\varphi)\Big(\log(\epsilon) + \log\left(\mathbb{P}(\varphi)\right) - \log\mathbb{P}(\varphi\\
&= \mu(S)\log(1-\epsilon) + \Psi_S(\mathbb{P}_S) - \Psi_S(\mathbb{P}) + \mu\left(\overline{S}\right)\log\epsilon\\
&\ge -\epsilon + \Psi_S(\mathbb{P}_S) - \Psi_S(\mathbb{P}) + \mu\left(\overline{S}\right)\log\epsilon
\end{aligned}$$

as desired. □

Theorem 9 (Strong concavity)**.** *For any $\mathbb{P}, \mathbb{Q} : S \to [0,1]$,*

$$\Psi_S\left(\delta\mathbb{P}+(1-\delta)\mathbb{Q}\right) \ge \delta\Psi_S(\mathbb{P}) + (1-\delta)\Psi_S(\mathbb{Q}) + \frac{1}{2}\delta(1-\delta)\sum_{\varphi\in S}\mu(\varphi)\left(\mathbb{P}(\varphi)-\mathbb{Q}(\varphi)\right)^2.$$

Proof. The key ingredient is the strong concavity of the logarithm, which can be deduced directly from the fact that its second derivative is $-\frac{1}{x^2} \leq -1$:

$$\log\left(\delta x + (1-\delta)y\right) \geq \delta \log x + (1-\delta)\log y + \frac{1}{2}\delta(1-\delta)(x-y)^2.$$

Using this inequality, we can compute (where all sums are over S):

$$\begin{aligned}
\Psi_S\left(\delta\mathbb{P} + (1-\delta)\mathbb{Q}\right) &= \sum_{\varphi} \mu(\varphi)\log\left(\delta\mathbb{P}(\varphi) + (1-\delta)\mathbb{Q}(\varphi)\right) \\
&\geq \sum_{\varphi} \mu(\varphi)\left(\delta\log\left(\mathbb{P}(\varphi)\right) + (1-\delta)\log\mathbb{Q}(\varphi)\right) + \frac{1}{2}\delta(1-\delta)\left(\mathbb{P}(\varphi) - \mathbb{Q}(\varphi)\right)^2 \\
&= \delta\Psi_S(\mathbb{P}) + (1-\delta)\Psi_S(\mathbb{Q}) + \frac{1}{2}\delta(1-\delta)\sum_{\varphi}\left(\mathbb{P}(\varphi) - \mathbb{Q}(\varphi)\right)^2,
\end{aligned}$$

as desired. □

Corollary 1. *If $\mathbb{P}_S$ is the coherent distribution maximizing $\Psi_S(\mathbb{P}_S)$, then for any coherent $\mathbb{P} : S \to [0,1]$:*

$$\Psi_S(\mathbb{P}) \leq \Psi_S(\mathbb{P}_S) - \frac{1}{2}\|\mathbb{P} - \mathbb{Q}\|_2^2$$

Proof. Consider the distributions $(1-\delta)\mathbb{P}_S + \delta\mathbb{P}$. By theorem 9:

$$\begin{aligned}
\Psi_S\left((1-\delta)\mathbb{P}_S + \delta\mathbb{P}\right) &\geq (1-\delta)\Psi_S(\mathbb{P}_S) + \delta\Psi_S(\mathbb{P}) + \delta(1-\delta)\|\mathbb{P} - \mathbb{Q}\|_2^2 \\
&= \Psi_S(\mathbb{P}_S) + \delta\left(\Psi_S(\mathbb{P}) - \Psi_S(\mathbb{P}_S) + \|\mathbb{P} - \mathbb{Q}\|_2^2\right) + \mathcal{O}\left(\delta^2\right)
\end{aligned}$$

And taking $\delta \to 0$, the corollary follows by the optimality of $\mathbb{P}_S$. □

Putting these two theorems together, we conclude:

Theorem 10. *If $\mathbb{P}_S$ is the coherent distribution maximizing $\Psi_S(\mathbb{P}_S)$ and $\mathbb{P}$ is the coherent distribution maximizing $\Psi(\mathbb{P})$ in the sense of Theorem 7, then for any $\varphi \in S$ we have*

$$\left|\mathbb{P}_S(\varphi) - \mathbb{P}(\varphi)\right| \leq \sqrt{\mu(\varphi)\,\mu\left(\overline{S}\right)\log\frac{2}{\mu\left(\overline{S}\right)}}$$

Proof. By Theorem 8 we can find an extension $\overline{\mathbb{P}_S}$ such that

$$0 \geq \Psi\left(\overline{\mathbb{P}_S}\middle\|\mathbb{P}\right) \geq \Psi_S(\mathbb{P}_S) - \Psi_S(\mathbb{P}) - \mu\left(\overline{S}\right)\log\frac{2}{\mu\left(\overline{S}\right)}$$

Then applying corollary 1, together with the observation

$$\|\mathbb{P} - \mathbb{Q}\|_2^2 \geq \mu(\varphi)\left(\mathbb{P}(\varphi) - \mathbb{Q}(\varphi)\right)^2$$

yields the desired result. □