Logical Induction

Scott Garrabrant, Tsvi Benson-Tilsen, Andrew Critch,
Nate Soares, Jessica Taylor

Machine Intelligence Research Institute

(scott|tsvi|critch|nate|jessica)@intelligence.org

Aug 6, 2016

Overview

Definitions

- $\mathcal{L}$:= a **language** of propositional logic, including connectives $\neg$, $\wedge$, $\vee$, $\rightarrow$, $\leftrightarrow$, for constructing proofs using modus ponens.
- $\mathcal{S}$:= all **sentences** expressible in $\mathcal{L}$.
- Γ := a set of **axioms** in $\mathcal{S}$ for encoding and proving statements about variables and computer programs (e.g. First Order Logic + Peano Arithmetic).
- a **belief state** := a map $\mathbb{P} : \mathcal{S} \rightarrow [0,1]$ that is constant outside some finite subset of $\mathcal{S}$.
- a **reasoning process** $\overline{\mathbb{P}}$:= a computable sequence of belief states $\{\mathbb{P}_n : L \rightarrow [0,1]\}$.

We can now state some properties that we think a "good reasoning process" should satisfy.

Desirable properties

A "good" reasoning process $\overline{\mathbb{P}}$ should satisfy:

1. **(computability)** There should be a Turing machine which computes $\mathbb{P}_n(\phi)$ for any input (n, ϕ).
2. **(convergence)** The limit $\mathbb{P}_\infty(\phi) := \lim_{n\to\infty} \mathbb{P}_n(\phi)$ should exist for all sentences ϕ.
3. **(coherent limit)** $\mathbb{P}_\infty$ should be a coherent probability distribution, i.e. obey laws like $\mathbb{P}_\infty(A \wedge B) + \mathbb{P}_\infty(A \vee B) = \mathbb{P}_\infty(A) + \mathbb{P}_\infty(B)$
4. **(non-dogmatism)** If $\Gamma \nvdash \phi$ then $\mathbb{P}_\infty(\phi) < 1$, and if $\Gamma \nvdash \neg\phi$ then $\mathbb{P}_\infty(\phi) > 0$.

Progress

Our forthcoming paper, "Logical Induction" (Garrabrant et al, 2016), shows that these properties are:

Related: A single property, the **Garrabrant Induction Criterion** (GIC), implies them all.

Feasible: We have a logical induction algorithm, "**LIA2016**", that satisfies the GIC.

Extensible: Many further desirable properties follow from **GIC**, and are hence satisfied by **LIA2016**.

Conservatism

- **(uniform non-dogmatism)** For any recursively enumerable sequence of sentences $\{\phi_n\}_{n\in\mathbb{N}}$ such that $\Gamma \cup \{\phi_n\}_{n\in\mathbb{N}}$ is consistent, there is a constant $\varepsilon > 0$ such that for all n,

$$\mathbb{P}_\infty(\phi_n) \geq \varepsilon.$$

- **(Occam bounds)** There exists a fixed positive constant C such that for any sentence ϕ with Kolmogorov complexity $\kappa(\phi)$, if $\Gamma \nvdash \neg\phi$, then

$$\mathbb{P}_\infty(\phi) \geq C2^{-\kappa(\phi)},$$

and if $\Gamma \nvdash \phi$, then

$$\mathbb{P}_\infty(\phi) \leq 1 - C2^{-\kappa(\phi)}.$$

Self-reflection

- **(belief in consistency)** Let $\mathrm{con}(t)$ be the sentence ⌜There is no proof of contradiction ($\perp$) from Γ using t or fewer symbols⌝. Then
$$\lim_{n\to\infty} \overline{\mathbb{P}}_n(\mathrm{con}(n)) = 1.$$

- **(belief in future consistency)** In fact, for any encoding $\underline{f}$ of a computable function $f : \mathbb{N} \to \mathbb{N}$,
$$\lim_{n\to\infty} \overline{\mathbb{P}}_n(\mathrm{con}(\underline{f}(n))) = 1.$$
For example, $f(n)$ could be $n^{n^{n^n}}$, or even $\mathrm{Ackermann}(n, n)$.

Self-reflection

- **(belief in consistency)** Let $\mathrm{con}(t)$ be the sentence ⌜There is no proof of contradiction ($\perp$) from Γ using t or fewer symbols⌝. Then
$$\lim_{n\to\infty} \overline{\mathbb{P}}_n(\mathrm{con}(n)) = 1.$$

- **(belief in future consistency)** In fact, for any encoding $\underline{f}$ of a computable function $f : \mathbb{N} \to \mathbb{N}$,
$$\lim_{n\to\infty} \overline{\mathbb{P}}_n(\mathrm{con}(\underline{f}(n))) = 1.$$

For example, $f(n)$ could be $n^{n^{n^n}}$, or even $\mathrm{Ackermann}(n, n)$.

Important concept: polytime generable

We say that a sequence of statements (or other objects) $\overline{\phi}$ is **polytime generable (p.g.)** if there exists a Turing machine M such that $M(n)$ generates the output ϕ_n in time polynomial in n.

A polytime generable sequence ϕ_n can be thought of as a sequence of T/F questions that is relatively easy to generate, but which can be arbitrarily difficult to answer deductively as n grows. In other words, think:

p.g. statements

$$\leftrightarrow$$

easy to state, hard to verify

Important concept: polytime generable

Example (statements that are hard to verify). Say f is any computable function. Fix an encoding $\underline{f}$ of f. By the parametric diagonal lemma [Boolos, 1993; p.53], there is a sentence $G(-)$ with one free variable such that for all n, Γ proves

$$G(\underline{n}) \leftrightarrow \text{“There is no proof of } \underline{G}(\underline{n}) \text{ in} \leq \underline{f}(\underline{n}) \text{ characters.”}$$

Then the sequence $\phi_n := G(\underline{n})$ is log-time generable: writing down ϕ_n only requires substituting the string $\underline{n}$ into $G(-)$, which takes $\mathcal{O}(\log(n))$ time. But if Γ is consistent, the length of the shortest proof of ϕ_n is at least $f(n)$. Nonetheless, we have...

Timely learning

- **(provability induction)** Any p.g. sequence of theorems ϕ_n will eventually be believed by $\mathbb{P}_n$ as soon as they are generated, i.e.

$$\lim_{n\to\infty} \mathbb{P}_n(\phi_n) = 1.$$

  In particular, $\overline{\mathbb{P}}$ can be seen to "outpace deduction" by a factor of f for any computable function f.

An analogy: Ramanujan vs Hardy. Imagine the ϕ_n are output by a heuristic algorithm that generates mathematical facts without proofs, similar in style to S. Ramanujan. Then $\overline{\mathbb{P}}_n$ resembles G.H. Hardy: he can only verify those results very slowly using the proof system Γ, but after enough examples, he begins to trust Ramanujan as soon as he speaks, even if the proofs of Ramanujan's later conjectures are impossibly long.

Important concept: timely manner

Given any sequences $\overline{x}$ and $\overline{y}$, we write

$$x_n \simeq_n y_n \quad \text{for} \quad \left(\lim_{n\to\infty} x_n - y_n = 0\right),$$
$$x_n \gtrsim_n y_n \quad \text{for} \quad \left(\liminf_{n\to\infty} x_n - y_n \geq 0\right), \text{ and}$$
$$x_n \lesssim_n y_n \quad \text{for} \quad \left(\limsup_{n\to\infty} x_n - y_n \leq 0\right).$$

Given p.g. sequences of statements $\overline{\phi}$ and probabilities $\overline{p}$, we say that $\overline{\mathbb{P}}$ assigns $\overline{p}$ to $\overline{\phi}$ in a **timely manner** if

$$\mathbb{P}_n(\phi_n) \simeq_n p_n$$

Timely learning

Henceforth, $\overline{\phi}$ will always denote a p.g. sequence of sentences.

- **(timely adoption of limits)** Let $\overline{p}$ be a p.g. sequence of rational probabilities. If

$$\mathbb{P}_{\infty}(\phi_n) \simeq_n p_n.$$

then

$$\mathbb{P}_{n}(\phi_n) \simeq_n p_n.$$

The same implication holds with $\lesssim$ or $\gtrsim$ in place of $\simeq$.

Hence, any p.g. assignment of probabilities that $\overline{\mathbb{P}}$ will learn, it learns in a timely manner.

Timely learning

- **(introspection)** A Garrabrant inductor $\mathbb{P}$ roughly knows what its own beliefs are at the time that it has them. Formally, for any polytime generable sequence of statements ϕ_n, any interval (a, b) and any $\varepsilon > 0$, for sufficiently large n:

$$\mathbb{P}_n(\phi_n) \in (a+\varepsilon, b-\varepsilon) \implies \mathbb{P}_n \ulcorner \mathbb{P}_n(\phi_n) \in (a, b) \urcorner > 1-\varepsilon$$
$$\mathbb{P}_n(\phi_n) \notin (a-\varepsilon, b+\varepsilon) \implies \mathbb{P}_n\left(\ulcorner \mathbb{P}_n(\phi_n) \in (a, b) \urcorner \right) < \varepsilon$$

- **(Liar's Paradox resistance)** Fix a rational $p \in (0, 1)$, and use Cantor's Diagonal Lemma to define a sequence of "liar sentences" L_n satisfying

$$\Gamma \vdash L_n \leftrightarrow \ulcorner \mathbb{P}_n(L_n) \le p \urcorner .$$

Then

$$\lim_{n\to\infty} \overline{\mathbb{P}}_n(L_n) = p.$$

Timely learning

- **(introspection)** A Garrabrant inductor $\mathbb{P}$ roughly knows what its own beliefs are at the time that it has them. Formally, for any polytime generable sequence of statements ϕ_n, any interval (a, b) and any $\varepsilon > 0$, for sufficiently large n:

$$\mathbb{P}_n(\phi_n) \in (a+\varepsilon, b-\varepsilon) \implies \mathbb{P}_n \ulcorner \mathbb{P}_n(\phi_n) \in (a,b) \urcorner > 1-\varepsilon$$

$$\mathbb{P}_n(\phi_n) \notin (a-\varepsilon, b+\varepsilon) \implies \mathbb{P}_n\left(\ulcorner \mathbb{P}_n(\phi_n) \in (a,b) \urcorner \right) < \varepsilon$$

- **(Liar's Paradox resistance)** Fix a rational $p \in (0,1)$, and use Cantor's Diagonal Lemma to define a sequence of "liar sentences" L_n satisfying

$$\Gamma \vdash L_n \leftrightarrow \ulcorner \mathbb{P}_n(L_n) \leq p \urcorner .$$

Then

$$\lim_{n\to\infty} \overline{\mathbb{P}}_n(L_n) = p.$$

Self-trust

- **(Trust in future beliefs)** For any computable function $f(n) > n$ and polytime generable sentences ϕ_n, we have a result roughly interpretable as saying that a GI's current beliefs about the sequence, conditioned on its future beliefs, agree with its future beliefs:

$$\text{“}\mathbb{P}(\phi_n \mid \ulcorner \mathbb{P}_{f(n)}(\phi_n) \geq p_n \urcorner) \gtrsim_n p_n\text{''}.$$

The precise statement (see paper for definitions) looks like this:

$$\mathbb{E}_n([\phi_n] \cdot \mathrm{Ind}_{\delta_n}(\ulcorner \mathbb{P}_{f(n)}(\phi_n) \geq p_n \urcorner)) \gtrsim_n p_n \cdot \mathbb{E}_n(\ulcorner \mathbb{P}_{f(n)}(\phi_n) \urcorner).$$

Learning statistical patterns

- **(Learning pseudorandom frequencies)** Let $\overline{\phi}$ be a p.g. sequence of Γ-decidable sentences. If $\overline{\phi}$ is *pseudorandom over* $\mathcal{O}(\overline{\mathbb{P}})$ *with frequency* p (defined in paper), then

$$\lim_{n\to\infty} \mathbb{P}_n(\phi_n) = p.$$

- **(Learning pseudorandom trends)** A stronger version of the above, where the frequencies vary over time.

Learning provable relationships

- **(Learning case breakdowns)** Let $\overline{\phi}^1, \ldots, \overline{\phi}^k$ be k p.g. sequences of sentences such that for each n, Γ proves that $\phi_n^1, \ldots, \phi_n^k$ are exclusive and exhaustive (i.e. exactly one of them is true). Then

$$\lim_{n\to\infty} \left(\mathbb{P}_n(\phi_n^1) + \cdots + \mathbb{P}_n(\phi_n^k)\right) = 1$$

- **(Learning affine relations)** A stronger version of the above, holding for every coherence relationship expressible as an affine combination of probabilities.

Other properties

- Well-behaved conditional credences, the analog of conditional probabilities;
- Well-behaved *logically uncertain variables*, the analogues of classical random variables;
- Well-behaved expected value operators for logically uncertain variables;
- Relationship to universal semi-measures;
- $\cdots$ (check out the paper)

Formalizing the Garrabrant induction criterion

Intuitively, **GIC** will say that you cannot easily make $\$\infty$ betting against a Garrabrant inductor unless you risk plausibly going arbitrarily into debt.

After enough definitions, GIC looks like this:

A *market* $\overline{\mathbb{P}}$ is said to satisfy the **Garrabrant induction criterion** with respect to a *deductive process* $\overline{D}$ if it cannot be *unboundedly exploited* by an *polytime trader* T with *bounded loss tolerance*:

$$\forall T \in \mathrm{Traders}: \quad \mathrm{MinPWorth}(T, \overline{\mathbb{P}}, \overline{D}) > -\infty \quad \Rightarrow \quad \mathrm{MaxPWorth}(T, \overline{\mathbb{P}}, \overline{D}) < +\infty.$$

A market $\overline{\mathbb{P}}$ which meets this criterion is called a **Garrabrant inductor**.

Formalizing the Garrabrant induction criterion

Informally, Garrabrant induction is “a financial solution to the computer science problem of metamathematics.”

Formalizing the Garrabrant induction criterion

$$\left\langle \text{Time permitting, use whiteboard to elaborate and/or field questions.} \right\rangle$$

LIA2016

The basic ideas behind **LIA2016** are these:

- Against finitely many traders, you can use Brouwer's fxed point theorem to balance your prices with your anticipation of their trades, so that they mostly trade with each other and don't get much of your money.
- Against all traders, you can balance your prices against an ever-expanding finite pool of traders that every trader eventually winds up in before it earns too much money, so that the total complexity-weighted wealth of the trader pool is bounded.

LIA2016

$$\left\langle \begin{array}{c}\text{Time permitting, use whiteboard} \\ \text{to elaborate and/or field questions.}\end{array} \right\rangle$$

Conclusions

Beamer $\rightarrow$ PowerPoint