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

This talk is based on our paper,

http://arXiv.org/abs/1609.03543/

which will be updated more frequently at

https://intelligence.org/files/LogicalInduction.pdf

These slides will be available at:

https://intelligence.org/seminar-f2016/

and possibly in a more updated form at:

http:/acritch.com/research/

Overview

1. Formalizing logical induction
   - Definitions
   - Basic properties
2. Properties of Garrabrant Inductors / LIA2016
   - Conservatism
   - (definition: efficiently computable)
   - Provability induction
   - Learning pseudorandom frequencies
   - Learning provable relationships
   - (definition: timely manner)
   - Self-reflective properties
   - Other properties
3. The Garrabrant induction criterion
4. LIA2016
5. Conclusions (PowerPoint)

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.

Basic properties

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

0. **(computability)** There should be a Turing machine which computes $\mathbb{P}_n(\phi)$ for any input (n, ϕ).
1. **(convergence)** The limit $\mathbb{P}_\infty(\phi) := \lim_{n\to\infty} \mathbb{P}_n(\phi)$ should exist for all sentences ϕ.
2. **(coherent limit)** $\mathbb{P}_\infty$ should be a coherent probability distribution, i.e. obey laws like $\mathbb{P}_\infty(A \wedge B) + \mathbb{P}_\infty(A \vee B) = \mathbb{P}_\infty(A) + \mathbb{P}_\infty(B)$
3. **(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 paper (http://arXiv.org/abs/1609.03543/), 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 computably 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 $\kappa(\phi)$ in a prefix-free encoding, if $\Gamma \nvdash \neg\phi$, then

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

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

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

(definition: efficiently computable)

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

An e.c. 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:

e.c. statements

$$\leftrightarrow$$

easy to state, hard to verify

Henceforth, $\overline{\phi}$ will always denote an e.c. sequence of sentences.

(definition: efficiently computable)

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...

Provability induction

- **(provability induction)** For any e.c. sequence $\overline{\phi}$ of provable statements ϕ_n,

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

  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.

Learning pseudorandom frequencies

In the paper, we define a notion of *pseudorandom* with respect to a particular runtime class $\mathcal{O}(r(n))$ depending on the runtime of $\overline{\mathbb{P}}$. Black-boxing those for now, we have:

- **(Learning pseudorandom frequencies)** For any e.c. sequence of decidable sentences $\overline{\phi}$ that is pseudorandom with frequency p over the class of $\mathcal{O}(r(n))$-time divergent weightings,

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

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

Learning pseudorandom frequencies

Note that learning pseudorandom frequencies

- **is not that hard** to satisfy on its own, but
- **is trickier to get along with coherence** (i.e., $\mathbb{P}_\infty$ being a probability distribution).

Learning provable relationships

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

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

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

(definition: timely manner)

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

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

Given e.c. 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

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

Self-reflective properties

- **(introspection)** For any efficiently computable sequence of statements ϕ_n, any interval (a, b), any e.c. sequence of positive rationals $\delta_n \to 0$, there exists a sequence $\varepsilon_n \to 0$ such that for all n:

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

- **(paradox resistance)** Fix a rational $p \in (0,1)$, and use Gödels 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

$$\overline{\mathbb{P}}_n(L_n) \eqsim_n p.$$

Self-reflective properties

- **(introspection)** For any efficiently computable sequence of statements ϕ_n, any interval (a, b), any e.c. sequence of positive rationals $\delta_n \to 0$, there exists a sequence $\varepsilon_n \to 0$ such that for all n:

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

- **(paradox resistance)** Fix a rational $p \in (0,1)$, and use Gödels 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

$$\overline{\mathbb{P}}_n(L_n) \eqsim_n p.$$

Self-reflective properties

- **(belief in consistency)** Let $\mathrm{con}(n)$ be the sentence ⌜There is no proof of contradiction ($\bot$) from Γ using n 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 $\mathrm{Ack}(n, n)$.

Self-reflective properties

- **(belief in consistency)** Let $\mathrm{con}(n)$ be the sentence ⌜There is no proof of contradiction ($\bot$) from Γ using n 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 $\mathrm{Ack}(n, n)$.

Self-reflective properties

- **(Trust in future beliefs)** For any computable function $f(n) > n$ and efficiently computable 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:

$$\mathbb{P}_n(\phi_n \mid \text{“}\underline{\mathbb{P}}_{\underline{f(n)}}(\underline{\phi_n}) \geq \underline{p_n}\text{”}) \gtrsim_n p_n.$$

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

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

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)

The Garrabrant induction criterion

A *market* $\overline{\mathbb{P}}$ is said to satisfy the **Garrabrant induction criterion** relative to a *deductive process* $\overline{D}$ if there is no efficiently computable *trader* $\overline{T}$ that *(plausibly) exploits* $\overline{\mathbb{P}}$ relative to $\overline{D}$. A market $\overline{\mathbb{P}}$ that meets this criterion is called a **Garrabrant inductor**.

A **deductive process** $\overline{D}$ is a computable nested sequence $D_1 \subseteq D_2 \subseteq D_3 \ldots$ of finite sets of sentences $D_n \subset \mathcal{S}$, interpreted as theorems that have been proven by day n. We write D_∞ for the union $\bigcup_n D_n$.

A **trader** $\overline{T}$ is a sequence of things called n-strategies T_n, each of which is a formula for buying and selling a linear combination of "shares" of sentences $T_n(\mathbb{P}_{\leq n})$ in response to the history of market prices $\mathbb{P}_{\leq n}$ on day n.

The Garrabrant induction criterion

A *market* $\overline{\mathbb{P}}$ is said to satisfy the **Garrabrant induction criterion** relative to a *deductive process* $\overline{D}$ if there is no efficiently computable *trader* $\overline{T}$ that *(plausibly) exploits* $\overline{\mathbb{P}}$ relative to $\overline{D}$. A market $\overline{\mathbb{P}}$ that meets this criterion is called a **Garrabrant inductor**.

A **deductive process** $\overline{D}$ is a computable nested sequence $D_1 \subseteq D_2 \subseteq D_3 \ldots$ of finite sets of sentences $D_n \subset \mathcal{S}$, interpreted as theorems that have been proven by day n. We write D_∞ for the union $\bigcup_n D_n$.

A **trader** $\overline{T}$ is a sequence of things called n-strategies T_n, each of which is a formula for buying and selling a linear combination of "shares" of sentences $T_n(\mathbb{P}_{\leq n})$ in response to the history of market prices $\mathbb{P}_{\leq n}$ on day n.

The Garrabrant induction criterion

A *market* $\overline{\mathbb{P}}$ is said to satisfy the **Garrabrant induction criterion** relative to a *deductive process* $\overline{D}$ if there is no efficiently computable *trader* $\overline{T}$ that *(plausibly) exploits* $\overline{\mathbb{P}}$ relative to $\overline{D}$. A market $\overline{\mathbb{P}}$ that meets this criterion is called a **Garrabrant inductor**.

A trader's (cash and stock) holdings on day n from trading against $\overline{\mathbb{P}}$ is the sum $H_n := \sum_{i \leq n} T_n(\mathbb{P}_{\leq n})$.

A trader $\overline{T}$ **(plausibly) exploits** a market $\overline{\mathbb{P}}$ if, as $n \to \infty$, the bounds on the value of its holdings H_n determinable from D_n via *boolean logic only* are bounded below but not bounded above.

The Garrabrant induction criterion

A *market* $\overline{\mathbb{P}}$ is said to satisfy the **Garrabrant induction criterion** relative to a *deductive process* $\overline{D}$ if there is no efficiently computable *trader* $\overline{T}$ that *(plausibly) exploits* $\overline{\mathbb{P}}$ relative to $\overline{D}$. A market $\overline{\mathbb{P}}$ that meets this criterion is called a **Garrabrant inductor**.

Example. Say $\phi =$ "$1+1=2$" and $\chi =$ "$2+2=4$", and suppose you're a trader whose your holdings on day 5 are

$$-\mathbf{1}+\phi+\chi$$

representing -\$1 of cash, one share of ϕ and one share of χ.

- If $D_5 = \emptyset$, the current bounds on your worth are $[-\mathbf{1},\mathbf{1}]$.
- If $D_5 = \{\phi\}$, your bounds are $[\mathbf{0},\mathbf{1}]$.
- If $D_5 = \{\phi \wedge \chi\}$, your bounds are $[\mathbf{1},\mathbf{1}]$ (the $\wedge$ is respected)
- If $D_5 = \{\forall \mathbf{x} : \phi\}$, your bounds are only $[-\mathbf{1},\mathbf{1}]$ (the quantifier $\forall$ is not respected)

The Garrabrant induction criterion

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

LIA2016

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

- We fix a (redundant) computable enumeration of all e.c. traders, and define two functions:
- `TradingFirm` watches a market $\mathbb{P}_{\leq n}$ and assembles performance-budgeted versions of those traders together, yielding a non-e.c. "supertrader" $\overline{T}$ who exploits $\overline{\mathbb{P}}$ iff $\overline{\mathbb{P}}$ is exploitable.
- `MarketMaker` looks at any trading strategy T_n and sets prices so that strategy can't make more than 2^{-n} from trading with them (no matter how stocks are valued).
- `LIA` pits `MarketMaker` and `TradingFirm` against each other in a recursion, which builds a market $\overline{\mathbb{P}}$ not exploitable by the output of `TradingFirm` applied to it, and hence not by and e.c. trader.

LIA2016

Given the deductive process $\overline{D}$, the shape of the recursion looks like this: $\mathtt{LIA}_{\leq 0} := ()$, and

$$\mathtt{LIA_n} := \mathtt{MarketMaker_n}(\mathtt{TradingFirm}_{\mathbf{n}}^{\overline{\mathbf{D}}}(\mathtt{LIA}_{\leq \mathbf{n-1}}), \mathtt{LIA}_{\leq \mathbf{n-1}}),$$

After enough lemmas and definitions, the main existence result looks like this:

Theorem ($\overline{\mathtt{LIA}}$ is a Logical Inductor)

The sequence of belief states $\overline{\mathtt{LIA}}$ satisfies the **Garrabrant induction criterion** *relative to $\overline{D}$, i.e., $\overline{\mathtt{LIA}}$ is not exploitable by any e.c. trader relative to the deductive process $\overline{D}$.*

Proof.

If any e.c. trader exploits $\overline{\mathtt{LIA}}$ (relative to $\overline{D}$), then so does the trader $\overline{F} := (\mathtt{TradingFirm}_n^{\overline{D}}(\mathtt{LIA}_{\leq n-1}))_{n \in \mathbb{N}^+}$. But $\overline{F}$ does not exploit $\overline{\mathtt{LIA}}$. Therefore no e.c. trader exploits $\overline{\mathtt{LIA}}$. □

LIA2016

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

LIA2016

The proofs of all our nice properties involve cooking up some e.c. trader that would exploit you otherwise. E.g.:

Proof sketch of Convergence.

Suppose for a contradiction that the limit

$$\mathbb{P}_\infty(\phi) := \lim_{n\to\infty} \mathbb{P}(\phi)$$

does not exist. Then for some rationals $p \in [0,1]$ and $\varepsilon > 0$, we have $\mathbb{P}_n(\phi) < p - \varepsilon$ and $\mathbb{P}_n(\phi) > p + \varepsilon$ infinitely often, so a trader can make $\$\infty$ buy buying shares for less than $p - \varepsilon$, waiting for a chance to sell then for $p + \varepsilon$, and repeating (details in paper). □

LIA2016

Proof sketch of Non-dogmatism.

Suppose for a contradiction that $\Gamma \nvdash \neg\phi$, but $\mathbb{P}_\infty(\phi) = 0$. (The other case is similar.) A trader can buy one share of ϕ at or below every price point 2^{-k}, never spending more than \$1, but accruing an even growing number of ϕ-shares $k \cdot \phi$. Since we never have $D_n \vdash \phi$, those shares are plausibly worth $\$k$, which $\to \infty$ as $n \to \infty$, contradicting the *GIC*. Hence $\mathbb{P}_\infty(\phi)$ must be bounded away from zero. □

See the paper for more rigorous details, and many more properties/proofs:

http://arXiv.org/abs/1609.03543/
https://intelligence.org/files/LogicalInduction.pdf

(The latter is being updated more frequently.)

Conclusions

Beamer $\rightarrow$ PowerPoint