Asymptotic Logical Uncertainty and the Benford Test

Scott Garrabrant[1,2], Tsvi Benson-Tilsen[1,3], Siddharth Bhaskar[2], Abram Demski[1,4], Joanna Garrabrant, George Koleszarik, and Evan Lloyd[2]

[1] Machine Intelligence Research Institute
[2] University of California, Los Angeles
[3] University of California, Berkeley
[4] University of Southern California

The Ninth Conference on Artificial General Intelligence, 2016

Outline

- Logical Uncertainty
  - Motivation
  - Related work
- Our approach
  - Operationalizing pseudo-randomness
  - Generalized Benford test
  - Computable algorithm

Motivation

- "Probability" Over Logical Statements
  - State of belief for a conjecture
  - Guessing the outcome of a long computation
- $\mathbb{P}(\text{"P=NP"}) = 0.1$?
  - Measure of surprise on seeing (dis)proof?
  - Measure of calibration on similar statements?
- $\mathbb{P}(\text{"The } 10^{100} \text{ digit of } \pi \text{ is a 3"}) = 0.1$?
  - Element of a pseudo-random sequence

Motivation

- Standard probability theory requires *logical omniscience*[1]
  - Coherence requires knowledge of all logical consequences of current beliefs
    - (*eg*: same probability on equivalent sentences)
    - (can relax in some ways[2, 3])
- Other approaches converge to coherent distribution eventually[4, 5]
  - Generally not computable
- (see sections 1 and 2 of paper)

---

[1]: Parikh: Knowledge and the Problem of Logical Omniscience
[2]: Cozic: Impossible States at Work: Logical Omniscience and Rational Choice
[3]: Halpern and Pucella: Dealing with Logical Omniscience
[4]: Hutter et al: Probabilities on Sentences in an Expressive Logic
[5]: Demski: Logical Prior Probability.

Operationalizing Pseudo-randomness in Logic

- Fix an enumeration of sentences, $\phi_1, \phi_2, \ldots$
- Pick a finite time bound $T(N) \geq N$
- Point to an infinite subset of logic with a Turing machine, Z:

$$S = \{\phi_i | Z \text{ halts within } T(i) \text{ steps, pointing at a } 1\}$$

- "Pseudo-random" sentences are decidable, but in more than $T(N)$ steps
  - *ie,* there is a binary sequence $\{b_i\}$ where
    $$b_i = \begin{cases} 1 & \text{if the } i^{th} \text{ element of } S \text{ is provable} \\ 0 & \text{if disprovable} \end{cases}$$
  - No simple Turing machine can predict $\{b_i\}$ with odds better than chance after T(N) steps

Operationalizing Pseudo-randomness in Logic

- More formally, for a given S and description length $K(W)$, consider all Turing machines W :
  - Run W with time limit of $T(N)$ steps
  - Interpret as selecting a subset of S :

$$S' = \{\phi_i \in S | W \text{ halts within } T(i) \text{ steps, pointing at a } 1\}$$

- "Empirical" frequency of provable sentences as a function of sample size m:

$$r(m, W) \equiv \frac{\left| \{s \in \text{smallest m elements of } S' | \phi_s \text{ is provable}\} \right|}{m}$$

Irreducible Patterns

- The law of the iterated logarithm gives a bound that holds almost surely for any random sequence, as a function of sample size and the generating frequency p:

$$|r(m, W) - p| < \frac{c \cdot K(W) \cdot \sqrt{\log \log m}}{\sqrt{m}}$$

- We call a set of decidable sentences to be an *irreducible pattern* with respect to p and $k = K(W)$ if this bound holds for *all* machines of description length k

Example Irreducible Patterns

- We could construct a machine Z that chooses the sentences $\{\phi_i |$"The $f(i)$ digit of π is a 3"$\}$
  - (where $f(i)$ grows faster than the best π-digit calculator can manage)
  - Conjecture: this is an irreducible pattern with $p = 1/10$
- Or, $\{\phi_i |$"The first digit of $3 \uparrow^i 3$ is a 1"$\}$
  - (where $x \uparrow^1 y = x^y$, $x \uparrow^n 1 = x$, and $x \uparrow^n y = x \uparrow^{n-1} (x \uparrow^n (y-1)))$
  - Conjecture: since Benford's Law holds for powers of 3, we expect this to be an irreducible pattern with $p = \log_{10}(2)$
- We'd like to have a general way to find all such patterns...

The Generalized Benford Test

- Inspired by Benford's Law (first digit follows $p(d) = \log_{10}(1 + 1/d)$)
- We'll design an algorithm $A_{L,T}$ that on every input $N \in \mathbb{N}$ outputs a value $\mathbb{P}(\phi_N) \in [0,1]$
  - Within time bound "close" to $O(T(N))$,
  - $R(N) = T(N) \cdot N^4 \cdot \log(T(N))$
- $A_{L,T}$ passes *the generalized Benford test* if for all irreducible patterns S and their respective probabilities p,

$$\lim_{\substack{N \to \infty \\ N \in S}} A_{L,T}(N) = p$$

Finding Irreducible Patterns

- For a single sentence ϕ_N, find a "reference class" containing it
  - *eg*: all digits of π, first digit of powers of 3
- Use a theorem prover L to test patterns
  - $L(N)$ halts pointing at a 1 if ZFC proves ϕ_N,
  - Halts pointing at a 0 if ZFC disproves ϕ_N,
  - Otherwise doesn't halt
- Strategy: iterate over pairs of Turing machines X and Y
  - X : best irreducible pattern, S_X, that contains N
  - Y : worst case subsequence, $S_Y \subseteq S_X$

Finding Irreducible Patterns

- Let $S = \{i \in [0 \ldots N] | X \text{ and } Y \text{ accept } i \text{ within time } T(i)\}$
  - Simulate L with time limit $T(N)$ on each $i \in S$; stop at N or first time-out
  - $Q_N(X, Y)$ is number of sentences that were decided in time
  - $F_N(X, Y)$ is the fraction (out of Q_N) that were true
- Define an objective B_N measuring the deviation in subset S_Y from the putative irreducible pattern S_X with probability approximately P

$$B_N(X, Y, P) = \max\left(K(X), \frac{|F_N(X, Y) - P|\sqrt{Q_N(X, Y)}}{K(Y)\sqrt{\log\log Q_N(X, Y)}}\right)$$

Properties of Our Algorithm

- Algorithm computes (see paper for fuller sketch)

$$\underset{P \in J_N}{\operatorname{argmin}} \max_{Y \in TM(N)} \min_{X \in TM(N)} B_N(X, Y, P),$$

  - $J_N = \left\{ \frac{0}{N}, \frac{1}{N}, \dots \frac{N}{N} \right\}$
  - $TM(N)$ is set of Turing machines that accept N within $T(N)$ steps
- Passes the Generalized Benford Test
  - When X enumerates an irreducible pattern, B_N has a constant upper bound
  - B_N having a constant upper bound implies that for sufficiently large N, P will be driven arbitrarily close to p

Summary

- Benford's Law points at logical uncertainty motivated by hard-to-compute sequences of logical sentences.
- The law of the iterated logarithm yields an "empirical" test of randomness that can be used to locate a "reference class" for a single sentence.
- Although this method yields a fully-specified logical uncertainty, we don't yet know how to combine it with notions of coherence.

References I

[1] Rohit Parikh.
Knowledge and the Problem of Logical Omniscience.
In *ISMIS*, volume 87, pages 432–349, 1987.

[2] Mikaël Cozic.
Impossible States at Work: Logical Omniscience and Rational Choice.
In *Contributions to Economic Analysis*, volume 280, pages 47–68, 2006.

[3] Joseph Y Halpern and Riccardo Pucella.
Dealing with Logical Omniscience.
In *Proceedings of the 11th Conference on Theoretical Aspects of Rationality and Knowledge*, pages 169–176. ACM, 2007.

References II

[4] Marcus Hutter, John W. Lloyd, Kee Siong Ng, and William T. B. Uther.
Probabilities on sentences in an expressive logic.
Journal of Applied Logic, 11(4):386–420, 2013.

[5] Abram Demski.
Logical prior probability.
Artificial General Intelligence. 5th International Conference, AGI 2012, Oxford, UK, December 8–11, 2012. Proceedings, (7716):50–59, 2012.

[6] Haim Gaifman.
Concerning measures in first order calculi.
Israel Journal of Mathematics, 2(1):1–18, 1964.