Solomonoff Induction Violates Nicod's Criterion

Jan Leike and Marcus Hutter

http://jan.leike.name/

Australian National University

ALT'15 — 6 October 2015

Outline

Motivation

What does this green apple tell you about black ravens?

The Paradox of Confirmation

Proposed by [Hempel, 1945].
$H =$ all ravens are black

The Paradox of Confirmation

Proposed by [Hempel, 1945].
$H =$ all ravens are black
$H' =$ all nonblack objects are nonravens

The Paradox of Confirmation

Proposed by [Hempel, 1945].
$H =$ all ravens are black
$H' =$ all nonblack objects are nonravens

- Nicod's criterion:
  Something that is F and G confirms "all Fs are Gs"
  $\Longrightarrow$ A nonblack nonraven confirms H'

The Paradox of Confirmation

Proposed by [Hempel, 1945].
$H =$ all ravens are black
$H' =$ all nonblack objects are nonravens

- Nicod's criterion:
  Something that is F and G confirms "all Fs are Gs"
  $\Longrightarrow$ A nonblack nonraven confirms H'
- Equivalence condition:
  Logically equivalent hypotheses are confirmed by the same evidence
  $\Longrightarrow$ A nonblack nonraven confirms H

The Paradox of Confirmation

Proposed by [Hempel, 1945].
$H =$ all ravens are black
$H' =$ all nonblack objects are nonravens

- Nicod's criterion:
  Something that is F and G confirms "all Fs are Gs"
  $\Longrightarrow$ A nonblack nonraven confirms H'
- Equivalence condition:
  Logically equivalent hypotheses are confirmed by the same evidence
  $\Longrightarrow$ A nonblack nonraven confirms H

Paradox?

Outline

Solomonoff Induction

Let U be a universal monotone Turing machine.

Solomonoff's universal prior [Solomonoff, 1964]:

$$M(x) := \sum_{p:\, U(p) = x\dots} 2^{-|p|}$$

M is a probability distribution on $\mathcal{X}^\infty \cup \mathcal{X}^*$

Solomonoff Induction

Let U be a universal monotone Turing machine.

Solomonoff's universal prior [Solomonoff, 1964]:

$$M(x) \;:=\; \sum_{p:\, U(p)=x\ldots} 2^{-|p|}$$

M is a probability distribution on $\mathcal{X}^\infty \cup \mathcal{X}^*$

Solomonoff normalization: $M_{\mathrm{norm}}(\epsilon) := 1$ and

$$M_{\mathrm{norm}}(xa) := M_{\mathrm{norm}}(x) \frac{M(xa)}{\sum_{b \in \mathcal{X}} M(xb)}$$

M_{norm} is a probability distribution on $\mathcal{X}^\infty$

Properties of Solomonoff Induction

Observe (non-iid) data $x_{<t} := x_1 x_2 \dots x_{t-1} \in \mathcal{X}^*$,
predict

$$\arg\max_{a \in \mathcal{X}} M(a \mid x_{<t})$$

Properties of Solomonoff Induction

Observe (non-iid) data $x_{<t} := x_1 x_2 \dots x_{t-1} \in \mathcal{X}^*$, predict

$$\arg\max_{a \in \mathcal{X}} M(a \mid x_{<t})$$

- At most $E + O(\sqrt{E})$ errors when observing data from a computable measure μ ($E =$ errors of the predictor that knows μ) [Hutter, 2001]

Properties of Solomonoff Induction

Observe (non-iid) data $x_{<t} := x_1 x_2 \dots x_{t-1} \in \mathcal{X}^*$,
predict

$$\arg\max_{a \in \mathcal{X}} M(a \mid x_{<t})$$

- At most $E + O(\sqrt{E})$ errors when observing data from a computable measure μ ($E =$ errors of the predictor that knows μ) [Hutter, 2001]
- M *merges* with any computable measure μ [Blackwell and Dubins, 1962]:

$$\sup_H |M(H \mid x_{<t}) - \mu(H \mid x_{<t})| \to 0 \ \mu\text{-a.s. as } t \to \infty$$

Properties of Solomonoff Induction

Observe (non-iid) data $x_{<t} := x_1 x_2 \dots x_{t-1} \in \mathcal{X}^*$,
predict

$$\arg\max_{a \in \mathcal{X}} M(a \mid x_{<t})$$

- At most $E + O(\sqrt{E})$ errors when observing data from a computable measure μ ($E =$ errors of the predictor that knows μ) [Hutter, 2001]
- M *merges* with any computable measure μ [Blackwell and Dubins, 1962]:

$$\sup_H |M(H \mid x_{<t}) - \mu(H \mid x_{<t})| \to 0 \ \mu\text{-a.s. as } t \to \infty$$

- M is lower semicomputable, but $M(xy \mid x)$ is incomputable

Properties of Solomonoff Induction

Observe (non-iid) data $x_{<t} := x_1 x_2 \ldots x_{t-1} \in \mathcal{X}^*$,
predict

$$\arg\max_{a \in \mathcal{X}} M(a \mid x_{<t})$$

- At most $E + O(\sqrt{E})$ errors when observing data from a computable measure μ ($E =$ errors of the predictor that knows μ) [Hutter, 2001]
- M *merges* with any computable measure μ [Blackwell and Dubins, 1962]:

$$\sup_H |M(H \mid x_{<t}) - \mu(H \mid x_{<t})| \to 0 \ \mu\text{-a.s. as } t \to \infty$$

- M is lower semicomputable, but $M(xy \mid x)$ is incomputable

$\Longrightarrow$ M is really good at learning

Outline

Setup

- Alphabet $=$ observations:

$$\mathcal{X} := \{BR, \overline{B}R, B\overline{R}, \overline{BR}\}$$

Setup

- Alphabet = observations:

$$\mathcal{X} := \{BR, \overline{B}R, B\overline{R}, \overline{BR}\}$$

- Hypothesis H = "all ravens are black":

$$H := \{x \in \mathcal{X}^\infty \cup \mathcal{X}^* \mid x \text{ does not contain } \overline{B}R\}$$

Setup

- Alphabet = observations:

$$\mathcal{X} := \{BR, \overline{B}R, B\overline{R}, \overline{BR}\}$$

- Hypothesis $H =$ "all ravens are black":

$$H := \{x \in \mathcal{X}^\infty \cup \mathcal{X}^* \mid x \text{ does not contain } \overline{B}R\}$$

- Data $x_{<t}$ drawn from a computable measure μ for $t = 1, 2, \ldots$

Setup

- Alphabet = observations:

$$\mathcal{X} := \{BR, \overline{B}R, B\overline{R}, \overline{BR}\}$$

- Hypothesis $H =$ "all ravens are black":

$$H := \{x \in \mathcal{X}^\infty \cup \mathcal{X}^* \mid x \text{ does not contain } \overline{B}R\}$$

- Data $x_{<t}$ drawn from a computable measure μ for $t = 1, 2, \ldots$
- $M(H \mid x_{<t})$ is subjective belief in H at time step t

Setup

- Alphabet = observations:

$$\mathcal{X} := \{BR, \overline{B}R, B\overline{R}, \overline{BR}\}$$

- Hypothesis $H =$ "all ravens are black":

$$H := \{x \in \mathcal{X}^\infty \cup \mathcal{X}^* \mid x \text{ does not contain } \overline{B}R\}$$

- Data $x_{<t}$ drawn from a computable measure μ for $t = 1, 2, \ldots$
- $M(H \mid x_{<t})$ is subjective belief in H at time step t
- Confirmation and disconfirmation:

$$\mu(H) = 0 \implies \exists t.\ M(H \mid x_{<t}) = 0 \ \mu\text{-a.s.}$$
$$\mu(H) = 1 \implies M(H \mid x_{<t}) \to 1 \ \mu\text{-a.s.}$$

Setup

- Alphabet = observations:

$$\mathcal{X} := \{BR, \overline{B}R, B\overline{R}, \overline{BR}\}$$

- Hypothesis H = "all ravens are black":

$$H := \{x \in \mathcal{X}^\infty \cup \mathcal{X}^* \mid x \text{ does not contain } \overline{B}R\}$$

- Data $x_{<t}$ drawn from a computable measure μ for $t = 1, 2, \ldots$
- $M(H \mid x_{<t})$ is subjective belief in H at time step t
- Confirmation and disconfirmation:

$$\mu(H) = 0 \implies \exists t.\ M(H \mid x_{<t}) = 0 \ \mu\text{-a.s.}$$
$$\mu(H) = 1 \implies M(H \mid x_{<t}) \to 1 \ \mu\text{-a.s.}$$

- Equivalence condition is satisfied.

Nicod's Criterion

Question: Does a black raven *confirm* H:

$$M(H \mid x_{<t}) < M(H \mid x_{<t}BR)?$$

Nicod's Criterion

Question: Does a black raven *confirm* H:

$$M(H \mid x_{<t}) < M(H \mid x_{<t}BR)?$$

Question: Does a nonblack nonraven *confirm* H:

$$M(H \mid x_{<t}) < M(H \mid x_{<t}\overline{BR})?$$

Nicod's Criterion

Question: Does a black raven *confirm* H:

$$M(H \mid x_{<t}) < M(H \mid x_{<t} BR)?$$

Question: Does a nonblack nonraven *confirm* H:

$$M(H \mid x_{<t}) < M(H \mid x_{<t} \overline{BR})?$$

Answer: Not always.

Solomonoff Induction and Nicod's Criterion

Theorem (Counterfactual Black Raven Disconfirms H)

Let $x_{1:\infty} \in H \subset \mathcal{X}^\infty$ *be computable and* $x_t \neq BR$ *infinitely often.*
$\Longrightarrow \exists t \in \mathbb{N}$ *(with* $x_t \neq BR$*) s.t.* $M(H \mid x_{<t}BR) < M(H \mid x_{<t})$

Solomonoff Induction and Nicod's Criterion

Theorem (Counterfactual Black Raven Disconfirms H)

Let $x_{1:\infty} \in H \subset \mathcal{X}^\infty$ be computable and $x_t \neq BR$ infinitely often.
$\Longrightarrow \exists t \in \mathbb{N}$ *(with $x_t \neq BR$) s.t.* $M(H \mid x_{<t}BR) < M(H \mid x_{<t})$

Theorem (Disconfirmation Infinitely Often for M)

Let $x_{1:\infty} \in H$ be computable.
$\Longrightarrow M(H \mid x_{1:t}) < M(H \mid x_{<t})$ *infinitely often.*

Solomonoff Induction and Nicod's Criterion

Theorem (Counterfactual Black Raven Disconfirms H)

Let $x_{1:\infty} \in H \subset \mathcal{X}^\infty$ *be computable and* $x_t \neq BR$ *infinitely often.*
$\Longrightarrow \exists t \in \mathbb{N}$ *(with* $x_t \neq BR$*) s.t.* $M(H \mid x_{<t}BR) < M(H \mid x_{<t})$

Theorem (Disconfirmation Infinitely Often for M)

Let $x_{1:\infty} \in H$ *be computable.*
$\Longrightarrow M(H \mid x_{1:t}) < M(H \mid x_{<t})$ *infinitely often.*

Theorem (Disconfirmation Finitely Often for M_{norm})

Let $x_{1:\infty} \in H$ *be computable.*
$\Longrightarrow \exists t_0 \forall t > t_0.\ M_{\mathrm{norm}}(H \mid x_{1:t}) > M_{\mathrm{norm}}(H \mid x_{<t})$.

Solomonoff Induction and Nicod's Criterion

Theorem (Counterfactual Black Raven Disconfirms H)

Let $x_{1:\infty} \in H \subset \mathcal{X}^\infty$ *be computable and* $x_t \neq BR$ *infinitely often.*
$\Longrightarrow \exists t \in \mathbb{N}$ *(with* $x_t \neq BR$*) s.t.* $M(H \mid x_{<t}BR) < M(H \mid x_{<t})$

Theorem (Disconfirmation Infinitely Often for M)

Let $x_{1:\infty} \in H$ *be computable.*
$\Longrightarrow M(H \mid x_{1:t}) < M(H \mid x_{<t})$ *infinitely often.*

Theorem (Disconfirmation Finitely Often for M_{norm})

Let $x_{1:\infty} \in H$ *be computable.*
$\Longrightarrow \exists t_0 \forall t > t_0.\ M_{\text{norm}}(H \mid x_{1:t}) > M_{\text{norm}}(H \mid x_{<t})$.

Theorem (Disconfirmation Infinitely Often for M_{norm})

There is an (incomputable) $x_{1:\infty} \in H$ *s.t.*
$M_{\text{norm}}(H \mid x_{1:t}) < M_{\text{norm}}(H \mid x_{<t})$ *infinitely often.*

Outline

Resolving the Paradox of Confirmation I

Solution: Reject Nicod's criterion!
[Good, 1967, Jaynes, 2003, Vranas, 2004]

Not all black ravens confirm H.

Resolving the Paradox of Confirmation II

In the literature there are perhaps 100 'paradoxes' and controversies which are like this, in that they arise from faulty intuition rather than faulty mathematics. Someone asserts a general principle that seems to him intuitively right. Then, when probability analysis reveals the error, instead of taking this opportunity to educate his intuition, he reacts by rejecting the probability analysis.

[Jaynes, 2003, p. 144]

Outline

References I

Blackwell, D. and Dubins, L. (1962).
Merging of opinions with increasing information.
The Annals of Mathematical Statistics, pages 882–886.

Good, I. J. (1967).
The white shoe is a red herring.
The British Journal for the Philosophy of Science, 17(4):322–322.

Hempel, C. G. (1945).
Studies in the logic of confirmation (I.).
Mind, pages 1–26.

Hutter, M. (2001).
New error bounds for Solomonoff prediction.
Journal of Computer and System Sciences, 62(4):653–667.

Jaynes, E. T. (2003).
Probability Theory: The Logic of Science.
Cambridge University Press.

References II

Solomonoff, R. (1964).
A formal theory of inductive inference. Parts 1 and 2.
Information and Control, 7(1):1–22 and 224–254.

Vranas, P. B. (2004).
Hempel's raven paradox: A lacuna in the standard Bayesian solution.
The British Journal for the Philosophy of Science, 55(3):545–560.