Solomonoff Induction Violates Nicod's Criterion*

Jan Leike and Marcus Hutter

Australian National University
{jan.leike|marcus.hutter}@anu.edu.au

Abstract. *Nicod's criterion* states that observing a black raven is evidence for the hypothesis H that all ravens are black. We show that Solomonoff induction does not satisfy Nicod's criterion: there are time steps in which observing black ravens *decreases* the belief in H. Moreover, while observing any computable infinite string compatible with H, the belief in H decreases infinitely often when using the unnormalized Solomonoff prior, but only finitely often when using the normalized Solomonoff prior. We argue that the fault is not with Solomonoff induction; instead we should reject Nicod's criterion.

Keywords: Bayesian reasoning, confirmation, disconfirmation, Hempel's paradox, equivalence condition, Solomonoff normalization.

1 Introduction

Inductive inference, how to generalize from examples, is the cornerstone of scientific investigation. But we cannot justify the use of induction on the grounds that it has reliably worked before, because this argument presupposes induction. Instead, we need to give *deductive* (logical) arguments for the use of induction. Today we know a formal solution to the problem of induction: Solomonoff's theory of learning [16,17], also known as *universal induction* or *Solomonoff induction*. It is a method of induction based on Bayesian inference [9] and algorithmic probability [11]. Because it is solidly founded in abstract mathematics, it can be justified purely deductively.

Solomonoff defines a prior probability distribution M that assigns to a string x the probability that a universal monotone Turing machine prints something starting with x when fed with fair coin flips. Solomonoff's prior encompasses *Ockham's razor* by favoring simple explanations over complex ones: algorithmically simple strings have short programs and are thus assigned higher probability than complex strings that do not have short programs. Moreover, Solomonoff's prior respects *Epicurus' principle* of multiple explanation by never discarding possible explanations: any possible program that explains the string contributes to the probability [8].

For data drawn from a computable probability distribution μ, Solomonoff induction will converge to the correct belief about any hypothesis [1]. Moreover,

* The final publication is available at http://link.springer.com/.

this can be used to produce reliable predictions extremely fast: Solomonoff induction will make a total of at most $E + O(\sqrt{E})$ errors when predicting the next data points, where E is the number of errors of the informed predictor that knows μ [7]. In this sense, Solomonoff induction solves the induction problem [15]. It is incomputable, hence it can only serve as an ideal that any practical learning algorithm should strive to approximate.

But does Solomonoff induction live up to this ideal? Suppose we entertain the hypothesis H that all ravens are black. Since this is a universally quantified statement, it is refuted by observing one counterexample: a non-black raven. But at any time step, we have observed only a finite number of the potentially infinite number of possible cases. Nevertheless, Solomonoff induction maximally confirms the hypothesis H asymptotically.

This paper is motivated by a problem of inductive inference extensively discussed in the literature: the *paradox of confirmation*, also known as *Hempel's paradox* [5]. It relies on the following three principles.

- *Nicod's criterion* [14, p. 67]: observing an F that is a G increases our belief in the hypothesis that all Fs are Gs.
- *The equivalence condition*: logically equivalent hypothesis are confirmed or disconfirmed by the same evidence.
- *The paradoxical conclusion*: a green apple confirms H.

The argument goes as follows. The hypothesis H is logically equivalent to the hypothesis H' that all non-black objects are non-ravens. According to Nicod's criterion, any non-black non-raven, such as a green apple, confirms H'. But then the equivalence condition entails the paradoxical conclusion.

The paradox of confirmation has been discussed extensively in the literature on the philosophy of science [5,2,12,3,6,13,19]; see [18] for a survey. Support for Nicod's criterion is not uncommon [12,6,13] and no consensus is in sight.

Using results from algorithmic information theory we show that Solomonoff induction avoids the paradoxical conclusion because it does not fulfill Nicod's criterion. There are time steps when (counterfactually) observing a black raven disconfirms the hypothesis that all ravens are black (Theorem 7 and Corollary 12). In the deterministic setting Nicod's criterion is even violated infinitely often (Theorem 8 and Corollary 13). However, if we *normalize* Solomonoff's prior and observe a deterministic computable infinite string, Nicod's criterion is violated at most finitely many times (Theorem 11). Our results are independent of the choice of the universal Turing machine. A list of notation can be found on page 15.

2 Preliminaries

Let $\mathcal{X}$ be some finite set called *alphabet*. The set $\mathcal{X}^* := \bigcup_{n=0}^{\infty} \mathcal{X}^n$ is the set of all finite strings over the alphabet $\mathcal{X}$, and the set $\mathcal{X}^\infty$ is the set of all infinite strings over the alphabet $\mathcal{X}$. The empty string is denoted by ϵ, not to be confused with the small positive rational number ε. Given a string $x \in \mathcal{X}^*$, we denote its length

by $|x|$. For a (finite or infinite) string x of length $\geq k$, we denote with $x_{1:k}$ the first k characters of x, and with $x_{<k}$ the first $k-1$ characters of x. The notation $x_{1:\infty}$ stresses that x is an infinite string. We write $x \sqsubseteq y$ iff x is a prefix of y, i.e., $x = y_{1:|x|}$.

A *semimeasure* over the alphabet $\mathcal{X}$ is a probability measure on the probability space $\mathcal{X}^\sharp := \mathcal{X}^* \cup X^\infty$ whose σ-algebra is generated by the *cylinder sets* $\Gamma_x := \{xz \mid z \in \mathcal{X}^\sharp\}$ [11, Ch. 4.2]. If a semimeasure assigns zero probability to every finite string, then it is called a *measure*. Measures and semimeasures are uniquely defined by their values on cylinder sets. For convenience we identify a string $x \in \mathcal{X}^*$ with its cylinder set Γ_x.

For two functions $f, g : \mathcal{X}^* \to \mathbb{R}$ we use the notation $f \overset{\times}{\geq} g$ iff there is a constant $c > 0$ such that $f(x) \geq cg(x)$ for all $x \in \mathcal{X}^*$. Moreover, we define $f \overset{\times}{\leq} g$ iff $g \overset{\times}{\geq} f$ and we define $f \overset{\times}{=} g$ iff $f \overset{\times}{\leq} g$ and $f \overset{\times}{\geq} g$. Note that $f \overset{\times}{=} g$ does *not* imply that there is a constant c such that $f(x) = cg(x)$ for all x.

Let U denote some universal Turing machine. The *Kolmogorov complexity* $K(x)$ of a string x is the length of the shortest program on U that prints x and then halts. A string x is *incompressible* iff $K(x) \geq |x|$. We define $m(t) := \min_{n \geq t} K(n)$, the *monotone lower bound on* K. Note that m grows slower than any unbounded computable function. (Its inverse is a version of the *busy beaver* function.) We also use the same machine U as a monotone Turing machine by ignoring the halting state and using a write-only output tape. The *monotone Kolmogorov complexity* $Km(x)$ denotes the length of the shortest program on the monotone machine U that prints a string starting with x. Since monotone complexity does not require the machine to halt, there is a constant c such that $Km(x) \leq K(x) + c$ for all $x \in X^*$.

Solomonoff's prior M [16] is defined as the probability that the universal monotone Turing machine computes a string when fed with fair coin flips in the input tape. Formally,

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

Equivalently, the Solomonoff prior M can be defined as a mixture over all lower semicomputable semimeasures [20].

The function M is a lower semicomputable semimeasure, but not computable and not a measure [11, Lem. 4.5.3]. It can be turned into a measure M_{norm} using *Solomonoff normalization* [11, Sec. 4.5.3]: $M_{\text{norm}}(\epsilon) := 1$ and for all $x \in \mathcal{X}^*$ and $a \in \mathcal{X}$,

$$M_{\text{norm}}(xa) := M_{\text{norm}}(x) \frac{M(xa)}{\sum_{b \in \mathcal{X}} M(xb)} \tag{1}$$

since $M(x) > 0$ for all $x \in \mathcal{X}^*$.

Every program contributes to M, so we have that $M(x) \geq 2^{-Km(x)}$. However, the upper bound $M(x) \overset{\times}{\leq} 2^{-Km(x)}$ is generally false [4]. Instead, the following weaker statement holds.

Lemma 1 ([10] as cited in [4, p. 75]). *Let $E \subset \mathcal{X}^*$ be a recursively enumerable and prefix-free set. Then there is a constant $c_E \in \mathbb{N}$ such that $M(x) \leq 2^{-Km(x)+c_E}$ for all $x \in E$.*

Proof. Define

$$\nu(x) := \begin{cases} M(x), & \text{if } x \in E, \text{ and} \\ 0, & \text{otherwise.} \end{cases}$$

The semimeasure ν is lower semicomputable because E is recursively enumerable. Furthermore, $\sum_{x \in \mathcal{X}^*} \nu(x) \leq 1$ because M is a semimeasure and E is prefix-free. Therefore ν is a discrete semimeasure. Hence there are constant c and c' such that $Km(x) \leq K(x) + c \leq -\log \nu(x) + c + c' = -\log M(x) + c + c'$ [11, Cor. 4.3.1]. □

Lemma 2 ([11, Sec. 4.5.7]). *For any computable measure μ the set of μ-Martin-Löf-random sequences has μ-probability one:*

$$\mu(\{x \in \mathcal{X}^\infty \mid \exists c \forall t.\ M(x_{1:t}) \leq c\mu(x_{1:t})\}) = 1.$$

3 Solomonoff and the Black Ravens

Setup. In order to formalize the black raven problem (in line with [15, Sec. 7.4]), we define two predicates: blackness B and ravenness R. There are four possible observations: a black raven BR, a non-black raven $\overline{B}R$, a black non-raven $B\overline{R}$, and a non-black non-raven $\overline{BR}$. Therefore our alphabet consists of four symbols corresponding to each of the possible observations, $\mathcal{X} := \{BR, \overline{B}R, B\overline{R}, \overline{BR}\}$. We will not make the formal distinction between observations and the symbols that represent them, and simply use both interchangeably.

We are interested in the hypothesis 'all ravens are black'. Formally, it corresponds to the set

$$H \ := \ \{x \in \mathcal{X}^\sharp \mid x_t \neq \overline{B}R \ \forall t\} \ = \ \{BR, B\overline{R}, \overline{BR}\}^\sharp, \tag{2}$$

the set of all finite and infinite strings in which the symbol $\overline{B}R$ does not occur. Let $H^c := \mathcal{X}^\sharp \setminus H$ be the complement hypothesis 'there is at least one non-black raven'. We fix the definition of H and H^c for the rest of this paper.

Using Solomonoff induction, our prior belief in the hypothesis H is

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

the cumulative weight of all programs that do not print any non-black ravens. In each time step t, we make one observation $x_t \in \mathcal{X}$. Our *history* $x_{<t} = x_1 x_2 \ldots x_{t-1}$ is the sequence of all previous observations. We update our belief with Bayes' rule in accordance with the Bayesian framework for learning [9]: our *posterior belief* in the hypothesis H is

$$M(H \mid x_{1:t}) \ = \ \frac{M(H \cap x_{1:t})}{M(x_{1:t})}.$$

We say that the observation x_t *confirms* the hypothesis H iff $M(H \mid x_{1:t}) > M(H \mid x_{<t})$ (the belief in H increases), and we say that the observation x_t *disconfirms* the hypothesis H iff $M(H \mid x_{1:t}) < M(H \mid x_{<t})$ (the belief in H decreases). If $M(H \mid x_{1:t}) = 0$, we say that H is *refuted*, and if $M(H \mid x_{1:t}) \to 1$ as $t \to \infty$, we say that H is *(maximally) confirmed asymptotically.*

Confirmation and Refutation. Let the sequence $x_{1:\infty}$ be sampled from a computable measure μ, the *true environment.* If we observe a non-black raven, $x_t = \overline{B}R$, the hypothesis H is refuted since $H \cap x_{1:t} = \emptyset$ and this implies $M(H \mid x_{1:t}) = 0$. In this case, our enquiry regarding H is settled. For the rest of this paper, we focus on the interesting case: we assume our hypothesis H is in fact true in μ ($\mu(H) = 1$), i.e., μ does not generate any non-black ravens. Since Solomonoff's prior M dominates all computable measures, there is a constant w_μ such that

$$\forall x \in \mathcal{X}^* \quad M(x) \geq w_\mu \mu(x). \tag{3}$$

Thus Blackwell and Dubins' famous merging of opinions theorem [1] implies

$$M(H \mid x_{1:t}) \to 1 \text{ as } t \to \infty \text{ with } \mu\text{-probability one.}^1 \tag{4}$$

Therefore our hypothesis H is confirmed asymptotically [15, Sec. 7.4]. However, convergence to 1 is extremely slow, slower than any unbounded computable function, since $1 - M(H \mid x_{1:t}) \overset{\times}{\geq} 2^{-m(t)}$ for all t.

In our setup, the equivalence condition holds trivially: a logically equivalent way of formulating a hypothesis yields the same set of infinite strings, therefore in our formalization it constitutes the same hypothesis. The central question of this paper is Nicod's criterion, which refers to the assertion that BR and $\overline{BR}$ confirm H, i.e., $M(H \mid x_{1:t}BR) > M(H \mid x_{<t})$ and $M(H \mid x_{1:t}\overline{BR}) > M(H \mid x_{<t})$ for all strings $x_{<t}$.

4 Disconfirming H

We first illustrate the violation of Nicod's criterion by defining a particular universal Turing machine.

Example 3 (Black Raven Disconfirms). The observation of a black raven can falsify a short program that supported the hypothesis H. Let $\varepsilon > 0$ be a small rational number. We define a semimeasure ρ as follows.

$$\rho(\overline{BR}^\infty) := \tfrac{1}{2} \quad \rho(BR^\infty) := \tfrac{1}{4} \quad \rho(BR\,\overline{BR}^\infty) := \tfrac{1}{4} - \varepsilon \quad \rho(x) := 0 \text{ otherwise.}$$

[1] Blackwell-Dubins' theorem refers to (probability) measures, but technically M is a semimeasure. However, we can view M as a measure by introducing an extra symbol to our alphabet [11, p. 264]. This preserves dominance (3), and hence absolute continuity, which is the precondition for Blackwell-Dubins' theorem.

$M(\cdot)$	H	H^c
$\bigcup_{a\neq x_t} \Gamma_{x_{<t}a}$	A	B
$\Gamma_{x_{1:t}}$	C	D
$\{x_{<t}\}$	E	0

$$A := \sum_{a\neq x_t} M(x_{<t}a \cap H)$$
$$B := \sum_{a\neq x_t} M(x_{<t}a \cap H^c)$$
$$C := M(x_{1:t} \cap H)$$
$$D := M(x_{1:t} \cap H^c)$$
$$E := M(x_{<t}) - \sum_{a\in\mathcal{X}} M(x_{<t}a)$$

Fig. 1: The definitions of the values A, B, C, D, and E. Note that by assumption, $x_{<t}$ does not contain non-black ravens, therefore $M(\{x_{<t}\} \cap H^c) = M(\emptyset) = 0$.

To get a universally dominant semimeasure ξ, we mix ρ with the universally dominant semimeasure M.

$$\xi(x) := \rho(x) + \varepsilon M(x).$$

For computable ε, the mixture ξ is a lower semicomputable semimeasure. Hence there is a universal monotone Turing machine whose Solomonoff prior is equal to ξ [20, Lem. 13]. Our a priori belief in H at time $t = 0$ is

$$\xi(H \mid \epsilon) = \xi(H) \geq \rho(\overline{BR}^\infty) + \rho(BR^\infty) = 75\%,$$

while our a posteriori belief in H after seeing a black raven is

$$\xi(H \mid BR) = \frac{\xi(H \cap BR)}{\xi(BR)} \leq \frac{\rho(BR^\infty) + \varepsilon}{\rho(BR^\infty) + \rho(BR\overline{B}R^\infty)} = \frac{\frac{1}{4} + \varepsilon}{\frac{1}{2} - \varepsilon} < 75\%$$

for $\varepsilon \leq 7\%$. Hence observing a black raven in the first time step disconfirms the hypothesis H. $\Diamond$

The rest of this section is dedicated to show that this effect occurs independent of the universal Turing machine U and on all computable infinite strings.

4.1 Setup

At time step t, we have seen the history $x_{<t}$ and now update our belief using the new symbol x_t. To understand what happens, we split all possible programs into five categories.

(a) Programs that *never* print non-black ravens (compatible with H), but become falsified at time step t because they print a symbol other than x_t.
(b) Programs that eventually print a non-black raven (contradict H), but become falsified at time step t because they print a symbol other than x_t.
(c) Programs that *never* print non-black ravens (compatible with H), and predict x_t correctly.

(d) Programs that eventually print a non-black raven (contradict H), and predict x_t correctly.
(e) Programs that do not print additional symbols after printing $x_{<t}$ (because they go into an infinite loop).

Let A, B, C, D, and E denote the cumulative contributions of these five categories of programs to M. A formal definition is given in Figure 1, and implicitly depends on the current time step t and the observed string $x_{1:t}$. The values of A, B, C, D, and E are in the interval $[0,1]$ since they are probabilities. Moreover, the following holds.

$$M(x_{<t}) = A+B+C+D+E \qquad M(x_{1:t}) = C+D \tag{5}$$

$$M(x_{<t} \cap H) = A+C+E \qquad M(x_{1:t} \cap H) = C \tag{6}$$

$$M(H \mid x_{<t}) = \frac{A+C+E}{A+B+C+D+E} \qquad M(H \mid x_{1:t}) = \frac{C}{C+D} \tag{7}$$

We use results from algorithmic information theory to derive bounds on A, B, C, D, and E. This lets us apply the following lemma which states a necessary and sufficient condition for confirmation/disconfirmation at time step t.

Lemma 4 (Confirmation Criterion). *Observing x_t confirms (disconfirms) the hypothesis H if and only if $AD+DE < BC$ ($AD+DE > BC$).*

Proof. The hypothesis H is confirmed if and only if

$$M(H \mid x_{1:t}) - M(H \mid x_{<t}) \overset{(7)}{=} \tfrac{C}{C+D} - \tfrac{A+C+E}{A+B+C+D+E} = \tfrac{BC-AD-DE}{(A+B+C+D+E)(C+D)}$$

is positive. Since the denominator is positive, this is equivalent to $BC > AD + DE$. □

Example 5 (Confirmation Criterion Applied to Example 3). In Example 3 we picked a particular universal prior and $x_1 = BR$. In this case, the values for A, B, C, D, and E are

$$A \in [\tfrac{1}{2}, \tfrac{1}{2}+\varepsilon] \quad B \in [0,\varepsilon] \quad C \in [\tfrac{1}{4}, \tfrac{1}{4}+\varepsilon] \quad D \in [\tfrac{1}{4}-\varepsilon, \tfrac{1}{4}] \quad E \in [0,\varepsilon].$$

We invoke Lemma 4 with $\varepsilon := 7\%$ to get that $x_1 = BR$ disconfirms H:

$$AD + DE \geq \tfrac{1}{8} - \tfrac{\varepsilon}{2} = 0.09 > 0.0224 = \tfrac{\varepsilon}{4} + \varepsilon^2 \geq BC. \qquad \Diamond$$

Lemma 6 (Bounds on $ABCDE$). *Let $x_{1:\infty} \in H$ be some computable infinite string. The following statements hold for every time step t.*

(i) $0 < A, B, C, D, E < 1$
(ii) $A + B \overset{\times}{\leq} 2^{-K(t)}$
(iii) $A, B \overset{\times}{\geq} 2^{-K(t)}$
(iv) $C \overset{\times}{\geq} 1$
(v) $D \overset{\times}{\geq} 2^{-m(t)}$
(vi) $D \to 0$ *as* $t \to \infty$
(vii) $E \to 0$ *as* $t \to \infty$

Proof. Let p be a program that computes the infinite string $x_{1:\infty}$.

(i) Each of A, B, C, D, E is a probability value and hence bounded between 0 and 1. These bounds are strict because for any finite string there is a program that prints that string.

(ii) A proof is given in the appendix of [8]. Let $a \neq x_t$ and let q be the shortest program for the string $x_{<t}a$, i.e., $|q| = Km(x_{<t}a)$. We can reconstruct t by running p and q in parallel and counting the number of characters printed until their output differs. Therefore there is a constant c independent of t such that $K(t) \leq |p| + |q| + c = |p| + Km(x_{<t}a) + c$. Hence

$$2^{-Km(x_{<t}a)} \leq 2^{-K(t)+|p|+c} \tag{8}$$

The set $E := \{x_{<t}a \mid t \in \mathbb{N}, a \neq x_t\}$ is recursively enumerable and prefix-free, so Lemma 1 yields a constant c_E such that

$$M(x_{<t}a) \leq 2^{-Km(x_{<t}a)+c_E} \overset{(8)}{\leq} 2^{-K(t)+|p|+c+c_E}.$$

With $A + B \leq (\#\mathcal{X} - 1) \max_{a \neq x_t} M(x_{<t}a)$ follows the claim.

(iii) Let $a \neq x_t$ and let q be the shortest program to compute t, i.e., $|q| = K(t)$. We can construct a program that prints $x_{<t}a\overline{B}R$ by first running q to get t and then running p until it has produced a string of length $t-1$, and then printing $a\overline{B}R$. Hence there is a constant c independent of t such that $Km(x_{<t}a\overline{B}R) \leq |q| + |p| + c = K(t) + |p| + c$. Therefore

$$M(x_{<t}a \cap H^c) \geq M(x_{<t}a\overline{B}R) \geq 2^{-Km(x_{<t}a\overline{B}R)} \geq 2^{-K(t)-|p|-c}.$$

For the bound on $M(x_{<t}a \cap H)$ we proceed analogously except that instead of printing $\overline{B}R$ the program goes into an infinite loop.

(iv) Since by assumption the program p computes $x_{1:\infty} \in H$, we have that $M(x_{1:t} \cap H) \geq 2^{-|p|}$.

(v) Let n be an integer such that $K(n) = m(t)$. We proceed analogously to (iii) with a program q that prints n such that $|q| = m(t)$. Next, we write a program that produces the output $x_{1:n}\overline{B}R$, which yields a constant c independent of t such that

$$M(x_{1:t} \cap H^c) \geq M(x_{1:n}\overline{B}R) \geq 2^{-Km(x_{1:n}\overline{B}R)} \geq 2^{-|q|-|p|-c} = 2^{-m(t)-|p|-c}.$$

(vi) This follows from Blackwell and Dubins' result (4):

$$D = (C + D)\left(1 - \tfrac{C}{C+D}\right) \leq (1+1)(1 - M(H \mid x_{1:t})) \to 0 \text{ as } t \to \infty.$$

(vii) $\sum_{t=1}^{\infty} M(\{x_{<t}\}) = M(\{x_{<t} \mid t \in \mathbb{N}\}) \leq 1$, thus $E = M(\{x_{<t}\}) \to 0$. □

Lemma 6 states the bounds that illustrate the ideas to our results informally: From $A \overset{\times}{=} B \overset{\times}{=} 2^{-K(t)}$ (ii,iii) and $C \overset{\times}{=} 1$ (iv) we get

$$AD \overset{\times}{=} 2^{-K(t)}D, \qquad BC \overset{\times}{=} 2^{-K(t)}.$$

According to Lemma 4, the sign of $AD + DE - BC$ tells us whether our belief in H increases (negative) or decreases (positive).

Since $D \to 0$ (vi), the term $AD \stackrel{\times}{=} 2^{-K(t)}D$ will eventually be smaller than $BC \stackrel{\times}{=} 2^{-K(t)}$. Therefore it is crucial how fast $E \to 0$ (vii). If we use M, then $E \to 0$ slower than $D \to 0$ (v), therefore $AD+DE-BC$ is positive infinitely often (Theorem 8). If we use M_{norm} instead of M, then $E = 0$ and hence $AD + DE - BC = AD - BC$ is negative except for a finite number of steps (Theorem 11).

4.2 Unnormalized Solomonoff Prior

Theorem 7 (Counterfactual Black Raven Disconfirms H). *Let $x_{1:\infty}$ be a computable infinite string such that $x_{1:\infty} \in H$ ($x_{1:\infty}$ does not contain any non-black ravens) and $x_t \neq BR$ infinitely often. Then there is a time step $t \in \mathbb{N}$ (with $x_t \neq BR$) such that $M(H \mid x_{<t}BR) < M(H \mid x_{<t})$.*

Proof. Let t be time step such that $x_t \neq BR$. From the proof of Lemma 6 (iii) we get $M(H^c \cap x_{<t}BR) \geq 2^{-K(t)-c}$ and thus

$$\begin{aligned} M(H \mid x_{<t}BR) &\leq \frac{M(H \cap x_{<t}BR) + M(H^c \cap x_{<t}BR) - 2^{-K(t)-c}}{M(x_{<t}BR)} \\ &= 1 - \frac{2^{-K(t)-c}}{M(x_{<t}BR)} \leq 1 - \frac{2^{-K(t)-c}}{A+B} \stackrel{(ii)}{\leq} 1 - 2^{-c-c'}. \end{aligned}$$

From (4) there is a t_0 such that for all $t \geq t_0$ we have $M(H \mid x_{<t}) > 1 - 2^{-c-c'} \geq M(H \mid x_{<t}BR)$. Since $x_t \neq BR$ infinitely often according to the assumption, there is a $x_t \neq BR$ for $t \geq t_0$. □

Note that the black raven in Theorem 7 that we observe at time t is *counterfactual*, i.e., not part of the sequence $x_{1:\infty}$. If we picked the binary alphabet $\{BR, \overline{BR}\}$ and denoted only observations of ravens, then Theorem 7 would not apply: the only infinite string in H is BR^∞ and the only counterfactual observation is $\overline{BR}$, which immediately falsifies the hypothesis H. The following theorem gives an on-sequence result.

Theorem 8 (Disconfirmation Infinitely Often for M). *Let $x_{1:\infty}$ be a computable infinite string such that $x_{1:\infty} \in H$ ($x_{1:\infty}$ does not contain any non-black ravens). Then $M(H \mid x_{1:t}) < M(H \mid x_{<t})$ for infinitely many time steps $t \in \mathbb{N}$.*

Proof. We show that there are infinitely many $n \in \mathbb{N}$ such that for each n there is a time step $t > n$ where the belief in H decreases. The ns are picked to have low Kolmogorov complexity, while the ts are incompressible. The crucial insight is that a program that goes into an infinite loop at time t only needs to know n and not t, thus making this program much smaller than $K(t) \geq \log t$.

Let q_n be a program that starting with $t = n + 1$ incrementally outputs $x_{1:t}$ as long as $K(t) < \log t$. Formally, let $\phi(y, k)$ be a computable function such that $\phi(y, k + 1) \leq \phi(y, k)$ and $\lim_{k\to\infty} \phi(y, k) = K(y)$.

```
program q_n :
    t := n + 1
    output x_{<t}
    while true :
        k := 0
        while φ(t, k) ≥ log t :
            k := k + 1
        output x_t
        t := t + 1
```

The program q_n only needs to know p and n, so we have that $|q_n| \leq K(n)+c$ for some constant c independent of n and t. For the smallest $t > n$ with $K(t) \geq \log t$, the program q_n will go into an infinite loop and thus fail to print a t-th character. Therefore

$$E = M(\{x_{<t}\}) \geq 2^{-|q_n|} \geq 2^{-K(n)-c}. \tag{9}$$

Incompressible numbers are very dense, and a simple counting argument shows that there must be one between n and $4n$ [11, Thm. 3.3.1 (i)]. Furthermore, we can assume that n is large enough such that $m(4n) \leq m(n)+1$ (since m grows slower than the logarithm). Then

$$m(t) \leq m(4n) \leq m(n) + 1 \leq K(n) + 1. \tag{10}$$

Since the function m grows slower than any unbounded computable function, we find infinitely many n such that

$$K(n) \leq \tfrac{1}{2}(\log n - c - c' - c'' - 1), \tag{11}$$

where c' and c'' are the constants from Lemma 6 (ii,v). For each such n, there is a $t > n$ with $K(t) \geq \log t$, as discussed above. This entails

$$m(t)+K(n)+c+c'' \overset{(10)}{\leq} 2K(n)+1+c+c'' \overset{(11)}{\leq} \log n - c' \leq \log t - c' \leq K(t) - c'. \tag{12}$$

From Lemma 6 we get

$$AD + DE \overset{(i)}{>} DE \overset{(9),(v)}{\geq} 2^{-m(t)-c-K(n)-c''} \overset{(12)}{\geq} 2^{-K(t)+c'} \overset{(i,ii)}{\geq} BC.$$

With Lemma 4 we conclude that x_t disconfirms H. □

To get that M violates Nicod's criterion infinitely often, we apply Theorem 8 to the computable infinite string BR^∞.

4.3 Normalized Solomonoff Prior

In this section we show that for computable infinite strings, our belief in the hypothesis H is non-increasing at most finitely many times if we normalize M.

For this section we define A', B', C', D', and E' analogous to A, B, C, D, and E as given in Figure 1 with M_{norm} instead of M.

Lemma 9 **($M_{\text{norm}} \geq M$).** *$M_{\text{norm}}(x) \geq M(x)$ for all $x \in \mathcal{X}^*$.*

Proof. We use induction on the length of x: $M_{\text{norm}}(\epsilon) = 1 = M(\epsilon)$ and

$$M_{\text{norm}}(xa) = \frac{M_{\text{norm}}(x)M(xa)}{\sum_{b\in\mathcal{X}} M(xb)} \geq \frac{M(x)M(xa)}{\sum_{b\in\mathcal{X}} M(xb)} \geq \frac{M(x)M(xa)}{M(x)} = M(xa).$$

The first inequality holds by induction hypothesis and the second inequality uses the fact that M is a semimeasure. □

The following lemma states the same bounds for M_{norm} as given in Lemma 6 except for (i) and (vii).

Lemma 10 (Bounds on $A'B'C'D'E'$). *Let $x_{1:\infty} \in H$ be some infinite string computed by program p. The following statements hold for all time steps t.*

(i) $A \leq A'$, $B \leq B'$, $C \leq C'$, $D \leq D'$	*(iv)* $C' \overset{\times}{\geq} 1$
(ii) $A' + B' \overset{\times}{\leq} 2^{-K(t)}$	*(v)* $D' \overset{\times}{\geq} 2^{-m(t)}$
(iii) $A', B' \overset{\times}{\geq} 2^{-K(t)}$	*(vi)* $D' \to 0$ *as* $t \to \infty$
	(vii) $E' = 0$

Proof. (i) Follows from Lemma 9.

(ii) Let $a \neq x_t$. From Lemma 6 (ii) we have $M(x_{<t}a) \overset{\times}{\leq} 2^{-K(t)}$. Thus

$$M_{\text{norm}}(x_{<t}a) \overset{(1)}{=} \frac{M_{\text{norm}}(x_{<t})M(x_{<t}a)}{\sum_{b\in\mathcal{X}} M(x_{<t}b)} \overset{\times}{\leq} \frac{M_{\text{norm}}(x_{<t})2^{-K(t)}}{\sum_{b\in\mathcal{X}} M(x_{<t}b)} \overset{\times}{\leq} 2^{-K(t)}.$$

The last inequality follows from $\sum_{b\in\mathcal{X}} M(x_{<t}b) \geq M(x_{1:t}) \overset{\times}{\geq} 1$ (Lemma 6 (iv)) and $M_{\text{norm}}(x_{<t}) \leq 1$.

(iii-v) This is a consequence of (i) and Lemma 6 (iii-v).

(vi) Blackwell and Dubins' result also applies to M_{norm}, therefore the proof of Lemma 6 (vi) goes through unchanged.

(vii) Since M_{norm} is a measure, it assigns zero probability to finite strings, i.e., $M_{\text{norm}}(\{x_{<t}\}) = 0$, hence $E' = 0$. □

Theorem 11 (Disconfirmation Finitely Often for M_{norm}). *Let $x_{1:\infty}$ be a computable infinite string such that $x_{1:\infty} \in H$ ($x_{1:\infty}$ does not contain any non-black ravens). Then there is a time step t_0 such that $M_{\text{norm}}(H \mid x_{1:t}) > M_{\text{norm}}(H \mid x_{<t})$ for all $t \geq t_0$.*

Intuitively, at time step t_0, M_{norm} has learned that it is observing the infinite string $x_{1:\infty}$ and there are no short programs remaining that support the hypothesis H but predict something other than $x_{1:\infty}$.

Proof. We use Lemma 10 (ii,iii,iv,vii) to conclude

$$A'D' + D'E' - B'C' \leq 2^{-K(t)+c}D' + 0 - 2^{-K(t)-c'-c''} \leq 2^{-K(t)+c}(D' - 2^{-c-c'-c''}).$$

From Lemma 10 (vi) we have that $D' \to 0$, so there is a t_0 such that for all $t \geq t_0$ we have $D' < 2^{-c-c'-c''}$. Thus $A'D' + D'E' - B'C'$ is negative for $t \geq t_0$. Now Lemma 4 entails that the belief in H increases. □

Interestingly, Theorem 11 does not hold for M since that would contradict Theorem 8. The reason is that there are quite short programs that produce $x_{<t}$, but do not halt after that. However, from p and $x_{<t}$ we cannot reconstruct t, hence a program for $x_{<t}$ does not give us a bound on $K(t)$.

Since we get the same bounds for M_{norm} as in Lemma 6, the result of Theorem 7 transfers to M_{norm}:

Corollary 12 (Counterfactual Black Raven Disconfirms H). *Let $x_{1:\infty}$ be a computable infinite string such that $x_{1:\infty} \in H$ ($x_{1:\infty}$ does not contain any non-black ravens) and $x_t \neq BR$ infinitely often. Then there is a time step $t \in \mathbb{N}$ (with $x_t \neq BR$) such that $M_{\text{norm}}(H \mid x_{<t}BR) < M_{\text{norm}}(H \mid x_{<t})$.*

For incomputable infinite strings the belief in H can decrease infinitely often:

Corollary 13 (Disconfirmation Infinitely Often for M_{norm}). *There is an (incomputable) infinite string $x_{1:\infty} \in H$ such that $M_{\text{norm}}(H \mid x_{1:t}) < M_{\text{norm}}(H \mid x_{<t})$ infinitely often as $t \to \infty$.*

Proof. We iterate Corollary 12: starting with $\overline{BR}^{\infty}$, we get a time step t_1 such that observing BR at time t_1 disconfirms H. We set $x_{1:t_1} := \overline{BR}^{t_1-1}BR$ and apply Corollary 12 to $x_{1:t_1}\overline{BR}^{\infty}$ to get a time step t_2 such that observing BR at time t_2 disconfirms H. Then we set $x_{1:t_2} := x_{1:t_1}\overline{BR}^{t_2-t_1-1}BR$, and so on. □

4.4 Stochastically Sampled Strings

The proof techniques from the previous subsections do not generalize to strings that are sampled stochastically. The main obstacle is the complexity of counterfactual observations $x_{<t}a$ with $a \neq x_t$: for deterministic strings $Km(x_{<t}a) \to 0$, while for stochastically sampled strings $Km(x_{<t}a) \nrightarrow 0$. Consider the following example.

Example 14 (Uniform IID Observations). Let λ_H be a measure that generates uniform i.i.d. symbols from $\{BR, B\overline{R}, \overline{BR}\}$. Formally,

$$\lambda_H(x) := \begin{cases} 0 & \text{if } \overline{B}R \in x, \text{ and} \\ 3^{-|x|} & \text{otherwise.} \end{cases}$$

By construction, $\lambda_H(H) = 1$. By Lemma 2 we have $A, C, E \stackrel{\times}{=} 3^{-t}$ and $B, D \stackrel{\times}{=} 3^{-t}2^{-m(t)}$ with λ_H-probability one. According to Lemma 4, the sign of $AD + DE - BC$ is indicative for the change in belief in H. But this is inconclusive both for M and M_{norm} since each of the summands AD, BC, and DE (in case $E \neq 0$) go to zero at the same rate:

$$AD \stackrel{\times}{=} DE \stackrel{\times}{=} BC \stackrel{\times}{=} 3^{-2t}2^{-m(t)}.$$

Whether H gets confirmed or disconfirmed thus depends on the universal Turing machine and/or the probabilistic outcome of the string drawn from λ_H. ◊

5 Discussion

We chose to present our results in the setting of the black raven problem to make them more accessible to intuition and more relatable to existing literature. But these results hold more generally: our proofs follow from the bounds on A, B, C, D, and E given in Lemma 6 and Lemma 10. These bounds rely on the fact that we are observing a computable infinite string and that at any time step t there are programs consistent with the observation history that contradict the hypothesis and there are programs consistent with the observation history that are compatible with the hypothesis. No further assumptions on the alphabet, the hypothesis H, or the universal Turing machine are necessary.

In our formalization of the raven problem given in Section 3, we used an alphabet with four symbols. Each symbol indicates one of four possible types of observations according to the two binary predicates blackness and ravenness. One could object that this formalization discards important structure from the problem: BR and $\overline{B}R$ have more in common than BR and $\overline{BR}$, yet as symbols they are all the same. Instead, we could use the latin alphabet and spell out 'black', 'non-black', 'raven', and 'non-raven'. The results given in this paper would still apply analogously.

Our result that Solomonoff induction does not satisfy Nicod's criterion is not true for every time step, only for some of them. Generally, whether Nicod's criterion should be adhered to depends on whether the paradoxical conclusion is acceptable. A different Bayesian reasoner might be tempted to argue that a green apple *does* confirm the hypothesis H, but only to a small degree, since there are vastly more non-black objects than ravens [2]. This leads to the acceptance of the paradoxical conclusion, and this solution to the confirmation paradox is known as the *standard Bayesian solution*. It is equivalent to the assertion that blackness is equally probable regardless of whether H holds: $P(\text{black}|H) \approx P(\text{black})$ [19]. Whether or not this holds depends on our prior beliefs.

The following is a very concise example against the standard Bayesian solution [3]: There are two possible worlds, the first has 100 black ravens and a million other birds, while the second has 1000 black ravens, one white raven, and a million other birds. Now we draw a bird uniformly at random, and it turns out to be a black raven. Contrary to what Nicod's criterion claims, this is strong evidence that we are in fact in the second world, and in this world non-black ravens exist.

For another, more intuitive example: Suppose you do not know anything about ravens and you have a friend who collects atypical objects. If you see a black raven in her collection, surely this would not increase your belief in the hypothesis that all ravens are black.

We must conclude that violating Nicod's criterion is not a fault of Solomonoff induction. Instead, we should accept that for Bayesian reasoning Nicod's criterion, in its generality, is false! Quoting the great Bayesian master E. T. Jaynes [9, p. 144]:

> 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.

Acknowledgement. This work was supported by ARC grant DP150104590.

References

1. D. Blackwell and L. Dubins. Merging of opinions with increasing information. *The Annals of Mathematical Statistics*, pages 882–886, 1962.
2. I. J. Good. The paradox of confirmation. *British Journal for the Philosophy of Science*, pages 145–149, 1960.
3. I. J. Good. The white shoe is a red herring. *The British Journal for the Philosophy of Science*, 17(4):322–322, 1967.
4. P. Gács. On the relation between descriptional complexity and algorithmic probability. *Theoretical Computer Science*, 22(1–2):71 – 93, 1983.
5. C. G. Hempel. Studies in the logic of confirmation (I.). *Mind*, pages 1–26, 1945.
6. C. G. Hempel. The white shoe: No red herring. *The British Journal for the Philosophy of Science*, 18(3):239–240, 1967.
7. M. Hutter. New error bounds for Solomonoff prediction. *Journal of Computer and System Sciences*, 62(4):653–667, 2001.
8. M. Hutter. On universal prediction and Bayesian confirmation. *Theoretical Computer Science*, 384(1):33–48, 2007.
9. E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.
10. L. A. Levin. Laws of information conservation (nongrowth) and aspects of the foundation of probability theory. *Problemy Peredachi Informatsii*, 10(3):30–35, 1974.
11. M. Li and P. M. B. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Texts in Computer Science. Springer, 3rd edition, 2008.
12. J. L. Mackie. The paradox of confirmation. *British Journal for the Philosophy of Science*, pages 265–277, 1963.
13. P. Maher. Inductive logic and the ravens paradox. *Philosophy of Science*, pages 50–70, 1999.
14. J. Nicod. *Le Problème Logique de L'Induction*. Presses Universitaires de France, 1961.
15. S. Rathmanner and M. Hutter. A philosophical treatise of universal induction. *Entropy*, 13(6):1076–1136, 2011.
16. R. Solomonoff. A formal theory of inductive inference. Parts 1 and 2. *Information and Control*, 7(1):1–22 and 224–254, 1964.
17. R. Solomonoff. Complexity-based induction systems: Comparisons and convergence theorems. *IEEE Transactions on Information Theory*, 24(4):422–432, 1978.
18. R. G. Swinburne. The paradoxes of confirmation: A survey. *American Philosophical Quarterly*, pages 318–330, 1971.
19. P. B. Vranas. Hempel's raven paradox: A lacuna in the standard Bayesian solution. *The British Journal for the Philosophy of Science*, 55(3):545–560, 2004.
20. I. Wood, P. Sunehag, and M. Hutter. (Non-)equivalence of universal priors. In *Solomonoff 85th Memorial Conference*, pages 417–425. Springer, 2011.

List of Notation

$:=$	defined to be equal
$\#A$	the cardinality of the set A, i.e., the number of elements
$\mathcal{X}$	a finite alphabet
$\mathcal{X}^*$	the set of all finite strings over the alphabet $\mathcal{X}$
$\mathcal{X}^\infty$	the set of all infinite strings over the alphabet $\mathcal{X}$
$\mathcal{X}^\sharp$	$\mathcal{X}^\sharp := \mathcal{X}^* \cup \mathcal{X}^\infty$, the set of all finite and infinite strings over the alphabet $\mathcal{X}$
Γ_x	the set of all finite and infinite strings that start with x
x, y	finite or infinite strings, $x, y \in \mathcal{X}^\sharp$
$x \sqsubseteq y$	the string x is a prefix of the string y
ϵ	the empty string
ε	a small positive rational number
t	(current) time step
n	natural number
$K(x)$	Kolmogorov complexity of the string x: the length of the shortest program that prints x and halts
$m(t)$	the monotone lower bound on K, formally $m(t) := \min_{n \geq t} K(n)$
$Km(x)$	monotone Kolmogorov complexity of the string x: the length of the shortest program on the monotone universal Turing machine that prints something starting with x
BR	a symbol corresponding to the observation of a black raven
$\overline{B}R$	a symbol corresponding to the observation of a non-black raven
$B\overline{R}$	a symbol corresponding to the observation of a black non-raven
$\overline{BR}$	a symbol corresponding to the observation of a non-black non-raven
H	the hypothesis 'all ravens are black', formally defined in (2)
U	the universal (monotone) Turing machine
M	the Solomonoff prior
M_{norm}	the normalized Solomonoff prior, defined according to (1)
p, q	programs on the universal (monotone) Turing machine