Title: Algorithmic information theory: a critical perspective

Date: Apr 10, 2018 03:30 PM

URL: http://pirsa.org/18040110

Abstract: Algorithmic information theory (AIT) delivers an objective quantification of simplicity-qua-compressibility,that was employed by Solomonoff (1964) to specify a gold standard of inductive inference. Or so runs the conventional account,that I will challenge in my talk.

Algorithmic information theory: a critical perspective

Tom Sterkenburg

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
FAKULTÄT FÜR PHILOSOPHIE, WISSENSCHAFTSTHEORIE UND RELIGIONSWISSENSCHAFT
MUNICH CENTER FOR MATHEMATICAL PHILOSOPHY

Algorithmic Information, Induction and Observers in Physics
Waterloo, April 2018


LMU
CAUTION

The Solomonoff-Levin definitions

- Solomonoff (1964): the algorithmic probability distribution Q_U.
  - A probability assignment based on universal description lengths.
  - An implementation of Occam's razor in prediction.

Solomonoff (1964). A formal theory of inductive inference. *Inform. Control*.
Zvonkin & Levin (1970). The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms. *Russ. Math. Surv.*

The Solomonoff-Levin definitions

- Solomonoff (1964): the algorithmic probability distribution Q_U.
  - A probability assignment based on universal description lengths.
  - An implementation of Occam's razor in prediction.
- Levin (1970): the universal a priori distribution ξ_W.
  - A weighted mean over a large class of effective probability distributions.
  - A universal prediction method.

Solomonoff (1964). A formal theory of inductive inference. *Inform. Control.*
Zvonkin & Levin (1970). The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms. *Russ. Math. Surv.*

A representation theorem

- The two definitions are equivalent.

Wood, Sunehag, & Hutter (2013). (Non-)equivalence of universal priors. *Proc. Solomonoff Memorial Conf.*

A representation theorem

- The two definitions are equivalent. That is,

$$\{Q_U\}_U = \{\xi_W\}_W.$$

Wood, Sunehag, & Hutter (2013). (Non-)equivalence of universal priors. *Proc. Solomonoff Memorial Conf.*

A representation theorem

- The two definitions are equivalent. That is,

$$\{Q_U\}_U = \{\xi_W\}_W.$$

  - The choice of universal Turing machine corresponds to the choice of universal weight function.

Wood, Sunehag, & Hutter (2013). (Non-)equivalence of universal priors. *Proc. Solomonoff Memorial Conf.*

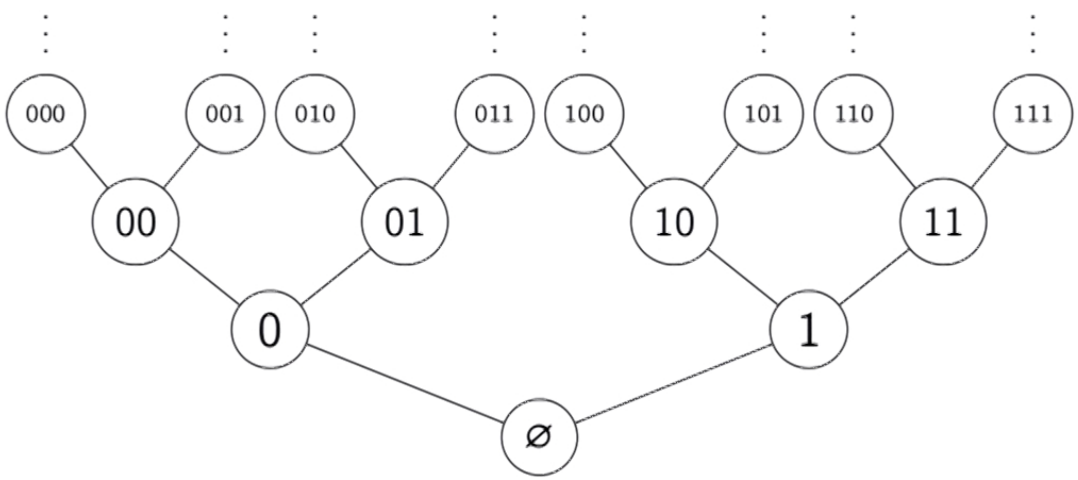
This talk

- Does the Solomonoff-Levin definition really give a convincing specification of a universal prediction method?
- Does the Solomonoff-Levin definition really give a convincing implementation of Occam's razor?

S. (2018). *Universal Prediction*. University of Groningen.

Part I:
A universal method of prediction?

- We assume the setting of binary sequential prediction.
  - A prediction method we define as a function $\mathrm{p} : \{0,1\}^* \to \mathcal{P}$ from finite data sequences to *predictions*, distributions over $\{0,1\}$.
  - Prediction methods correspond to *probability measures* μ over the whole Cantor space, by $\mathrm{p}_\mu(\boldsymbol{x}) = \mu^1(\cdot \mid \boldsymbol{x})$.



Dawid (1984). Statistical theory: The prequential approach. *J. R. Stat. Soc. A.*

A universal prediction method

- Universal **reliability**: to *always* converge on successful predictions.
  - This is quite impossible, at least without making inductive assumptions on what Nature can do.

Howson (2000). *Hume's Problem.*

A universal prediction method

- Universal **reliability**: to *always* converge on successful predictions.
  - This is quite impossible, at least without making inductive assumptions on what Nature can do.
- Alternatively, universal **optimality**: to converge on successful predictions whenever *some* prediction method would.
  - Rather than making assumptions about Nature, formulate reasonable restrictions on what *we* could ever do.

Howson (2000). *Hume's Problem*.

The restriction of effective computability

- Any prediction method we could possibly design may be captured in an algorithm.
  - Universal **optimality**: to converge on successful predictions whenever some *computable* prediction method would.

Mixture predictors

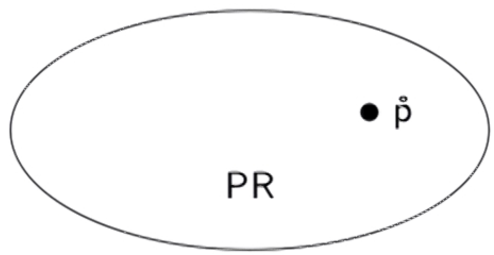
- Take the class $\mathcal{H}$ of all computable probability measures over Cantor space, corresponding to all computable prediction methods. A *mixture*, defined by

$$\xi_w(\cdot) := \sum_{\mu_i \in \mathcal{H}} w(\mu_i) \cdot \mu_i(\cdot),$$

corresponds to a prediction function that is optimal w.r.t. all computable prediction methods.

- End of story?

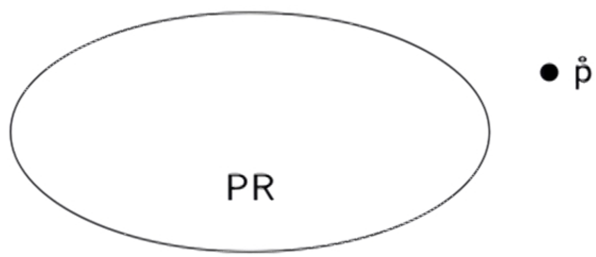
A diagonal argument



Putnam (1963). "Degree of confirmation" and inductive logic. *The Philosophy of Rudolf Carnap.*
Kelly (2016). Learning theory and epistemology. *Readings in Formal Epistemology.*

A diagonal argument

- The problem is that this mixture is *itself* no longer computable.
  - For any computable prediction method you propose, I can exhibit a sequence that your method *doesn't* converge on, but some other computable method *does*.



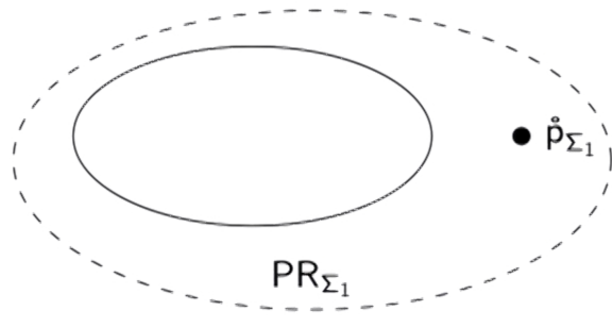
Putnam (1963). "Degree of confirmation" and inductive logic. *The Philosophy of Rudolf Carnap.*
Kelly (2016). Learning theory and epistemology. *Readings in Formal Epistemology.*

The Solomonoff-Levin definition

- Try to escape diagonalization by expanding to the class of "semi-computable" measures (on the space of infinite and *finite* sequences), that *does* contain universal elements.

The Solomonoff-Levin definition

- Try to escape diagonalization by expanding to the class of "semi-computable" measures (on the space of infinite and *finite* sequences), that *does* contain universal elements.



A failed escape

- However, we are not so much interested in the underlying measures as in the actual *prediction methods*—the *conditional* measures.
  - In the case of *computable* measures, this doesn't make a difference: the computable measures correspond precisely to the computable conditional measures.

Leike & Hutter (2015). On the computability of Solomonoff induction and knowledge-seeking. *ALT '15*.
Putnam (1963). "Degree of confirmation" and inductive logic. *The Philosophy of Rudolf Carnap*.

A failed escape

- However, we are not so much interested in the underlying measures as in the actual *prediction methods*—the *conditional* measures.
  - In the case of *computable* measures, this doesn't make a difference: the computable measures correspond precisely to the computable conditional measures.
  - But in the case of *semi-computable* measures, this *does* make a difference. In particular, the Solomonoff-Levin *predictor* is no longer semi-computable!

Leike & Hutter (2015). On the computability of Solomonoff induction and knowledge-seeking. *ALT '15*.
Putnam (1963). "Degree of confirmation" and inductive logic. *The Philosophy of Rudolf Carnap*.

Addendum: the funny notion of a semi/limit-computable method

- Consider the notion of a *partially computable* method for categorical prediction.

Kelly, Juhl, & Glymour (1994). Reliability, realism, and relativism. *Reading Putnam.*

Addendum: the funny notion of a semi/limit-computable method

- Consider the notion of a *partially computable* method for categorical prediction. It doesn't seem very adequate for this purpose, because at each trial it might be undefined and we have to either
  - resign to waiting forever (actually losing universality!); or
  - stop waiting and issue a default prediction at some point (actually losing universality—or else computability!).

Kelly, Juhl, & Glymour (1994). Reliability, realism, and relativism. *Reading Putnam*.

Addendum: the funny notion of a semi/limit-computable method

- Consider the notion of a *partially computable* method for categorical prediction. It doesn't seem very adequate for this purpose, because at each trial it might be undefined and we have to either
  - resign to waiting forever (actually losing universality!); or
  - stop waiting and issue a default prediction at some point (actually losing universality—or else computability!).
- With a *semi*-computable prediction method we superficially seem to be in a better place—but are we really?

Kelly, Juhl, & Glymour (1994). Reliability, realism, and relativism. *Reading Putnam*.

Part II:
An implementation of Occam's razor?

- The (modern) definition of Solomonoff's algorithmic probability distribution, via monotone Turing machine U, is given by

$$Q_U(\boldsymbol{y}) := \sum_{\boldsymbol{x} \in A_U(\boldsymbol{y})} 2^{-|\boldsymbol{x}|},$$

with

$$A_U(\boldsymbol{y}) = \lfloor \{U(\boldsymbol{x}) \succcurlyeq \boldsymbol{y}\} \rfloor$$

the prefix-free set of shortest *U-descriptions* of $\boldsymbol{y}$.

Solomonoff (1964). A formal theory of inductive inference. *Inform. Control.*
Ortner & Leitgeb (2011). Mechanizing induction. *Handbook of the History of Logic: Inductive Logic.*

Part II:
An implementation of Occam's razor?

- The (modern) definition of Solomonoff's algorithmic probability distribution, via monotone Turing machine U, is given by

$$Q_U(\boldsymbol{y}) := \sum_{\boldsymbol{x} \in A_U(\boldsymbol{y})} 2^{-|\boldsymbol{x}|},$$

  with

$$A_U(\boldsymbol{y}) = \lfloor \{ U(\boldsymbol{x}) \succcurlyeq \boldsymbol{y} \} \rfloor$$

  the prefix-free set of shortest *U-descriptions* of $\boldsymbol{y}$.

- The algorithmic probability of $\boldsymbol{y}$ is higher as it is more *compressible*.
- Hence the predictive probability

$$Q(y \mid \boldsymbol{y}) = \frac{Q(\boldsymbol{y}y)}{Q(\boldsymbol{y})}$$

  is greatest for the y such that $\boldsymbol{y}y$ is more compressible, which is "evidently an implementation of Occam's razor that identifies simplicity with compressibility."

Solomonoff (1964). A formal theory of inductive inference. *Inform. Control.*
Ortner & Leitgeb (2011). Mechanizing induction. *Handbook of the History of Logic: Inductive Logic.*

Coding systems and compressibility (1)

- Let's investigate the relevant notion of compressibility in some more detail.

Coding systems and compressibility (1)

- Let's investigate the relevant notion of compressibility in some more detail.
  - A coding system or simply *code* is a function $C : \{0,1\}^* \to \{0,1\}^*$ from *source sequences* to their *description sequences*, in such a way that no description is a prefix of another.
  - A code comes with a *code length function* $L_C : \{0,1\}^* \to \mathbb{N}$, that returns the length of a given source sequence's description.

Coding systems and compressibility (1)

- Let's investigate the relevant notion of compressibility in some more detail.
  - A coding system or simply *code* is a function $C : \{0,1\}^* \to \{0,1\}^*$ from *source sequences* to their *description sequences*, in such a way that no description is a prefix of another.
  - A code comes with a *code length function* $L_C : \{0,1\}^* \to \mathbb{N}$, that returns the length of a given source sequence's description.
- Codes and probability distributions on finite sequences can be treated as *equivalent*. Namely, for every code C the function 2^{-L_C} gives a probability assignment; conversely, for every probability assignment there is some code that thus (approximately) corresponds to it.

Coding systems and compressibility (2)

- If $\boldsymbol{y}$ has a small code length $L_C(\boldsymbol{y})$ then one can say that C *compresses* $\boldsymbol{y}$ well, or even that $\boldsymbol{y}$ is *simple* to C.

Universal coding sytems

- Given a class $\mathcal{C}$ of codes. A *universal* code $C^{\mathcal{C}}$ for this class is "almost as good" as any code in it: for every $C \in \mathcal{C}$ there is an *overhead constant* such that for every source sequence $\boldsymbol{y}$, the universal description length of $\boldsymbol{y}$ via $C^{\mathcal{C}}$ does not exceed the description length $L_C(\boldsymbol{y})$ more than this overhead.

Universal coding sytems

- Given a class $\mathcal{C}$ of codes. A *universal* code $C^{\mathcal{C}}$ for this class is "almost as good" as any code in it: for every $C \in \mathcal{C}$ there is an *overhead constant* such that for every source sequence $\boldsymbol{y}$, the universal description length of $\boldsymbol{y}$ via $C^{\mathcal{C}}$ does not exceed the description length $L_C(\boldsymbol{y})$ more than this overhead.
  - A universal code for $\mathcal{C}$ represents the full class $\mathcal{C}$ in the sense that if some $C \in \mathcal{C}$ assigns a particular sequence a short description, then the universal code does too—up to the overhead constant.
  - But the corresponding "universal compressibility" is again really a relative measure of how well sequences are fit by this particular class, equivalent to the goodness-of-fit of the corresponding mixture over the class $\mathcal{P}$ of distributions corresponding to $\mathcal{C}$.
  - A mixture ξ over $\mathcal{P}$ represents the full class $\mathcal{P}$ in the sense that if some $P \in \mathcal{P}$ assigns a particular sequence a high probability, then the mixture does too—up to the weight.

Universal coding sytems

- Given a class $\mathcal{C}$ of codes. A *universal* code $C^{\mathcal{C}}$ for this class is "almost as good" as any code in it: for every $C \in \mathcal{C}$ there is an *overhead constant* such that for every source sequence $\boldsymbol{y}$, the universal description length of $\boldsymbol{y}$ via $C^{\mathcal{C}}$ does not exceed the description length $L_C(\boldsymbol{y})$ more than this overhead.
  - A universal code for $\mathcal{C}$ represents the full class $\mathcal{C}$ in the sense that if some $C \in \mathcal{C}$ assigns a particular sequence a short description, then the universal code does too—up to the overhead constant.
  - But the corresponding "universal compressibility" is again really a relative measure of how well sequences are fit by this particular class, equivalent to the goodness-of-fit of the corresponding mixture over the class $\mathcal{P}$ of distributions corresponding to $\mathcal{C}$.
  - A mixture ξ over $\mathcal{P}$ represents the full class $\mathcal{P}$ in the sense that if some $P \in \mathcal{P}$ assigns a particular sequence a high probability, then the mixture does too—up to the weight.
- Arguably, *truly* universal compressibility must again be found in the class of all *effectively computable* elements.

The issue of variance

- The choice of overhead constants.
  - ... Or the choice of universal machine in the algorithmic probability distribution.
  - ... Or the choice of weights in the universal mixture.
- If any choice of overhead constants gives a universal code (algorithmic probability distribution, universal mixture) that is as valid as the next one, does this not make such a choice and thereby the definition rather arbitrary?

The issue of variance

- The choice of overhead constants.
  - ... Or the choice of universal machine in the algorithmic probability distribution.
  - ... Or the choice of weights in the universal mixture.
- If any choice of overhead constants gives a universal code (algorithmic probability distribution, universal mixture) that is as valid as the next one, does this not make such a choice and thereby the definition rather arbitrary?
  - Perhaps we can identify a privileged, *objective* such choice?

The invariance theorem

- Any two choices are equivalent up to an additive/multiplicative constant.
- "The bearing of the invariance theorem is that "from an asymptotic perspective, the complexity ... does not depend on accidental peculiarities of the chosen optimal method."
- I fix some universal code, you fix another; then for any sequence we investigate the description lengths will not differ more than a constant.
- An alternative perspective: I fix some universal code, and for any sequence I investigate, you can choose another universal code such that the two description lengths for this sequence *diverge arbitrarily much*.

Kolmogorov (1965). Three approaches to the quantitative definition of information. *Probl. Inf. Transm.*
Chaitin (1969). On the length of programs for computing finite binary sequences: statistical considerations. *J. ACM.*
Kolmogorov (1983). Combinatprobabilities. *Russ. Math. Surv.*

The invariance theorem

- ▶ Any two choices are equivalent up to an additive/multiplicative constant.
- ▷ "The bearing of the invariance theorem is that "from an asymptotic perspective, the complexity . . . does not depend on accidental peculiarities of the chosen optimal method."
- ▷ I fix some universal code, you fix another; then for any sequence we investigate the description lengths will not differ more than a constant.
- ▷ An alternative perspective: I fix some universal code, and for any sequence I investigate, you can choose another universal code such that the two description lengths for this sequence *diverge arbitrarily much*.
- ▷ Yet another perspective: we only care about the *order* of complexity. We can distinguish, for instance, data streams of complexity order $O(\log t)$ from those of order $O(1)$.

Kolmogorov (1965). Three approaches to the quantitative definition of information. *Probl. Inf. Transm.*
Chaitin (1969). On the length of programs for computing finite binary sequences: statistical considerations. *J. ACM.*
Kolmogorov (1983). Combinatprobabilities. *Russ. Math. Surv.*

The invariance theorem

- Any two choices are equivalent up to an additive/multiplicative constant.
  - "The bearing of the invariance theorem is that "from an asymptotic perspective, the complexity ... does not depend on accidental peculiarities of the chosen optimal method."
  - I fix some universal code, you fix another; then for any sequence we investigate the description lengths will not differ more than a constant.
  - An alternative perspective: I fix some universal code, and for any sequence I investigate, you can choose another universal code such that the two description lengths for this sequence *diverge arbitrarily much*.
  - Yet another perspective: we only care about the *order* of complexity. We can distinguish, for instance, data streams of complexity order $O(\log t)$ from those of order $O(1)$.
- Is this enough?

Kolmogorov (1965). Three approaches to the quantitative definition of information. *Probl. Inf. Transm.*
Chaitin (1969). On the length of programs for computing finite binary sequences: statistical considerations. *J. ACM.*
Kolmogorov (1983). Combinatprobabilities. *Russ. Math. Surv.*

The permissiveness of universality

- Intuition: universality just is an *extremely permissive* notion.
- Consider again the definition of the algorithmic probability distribution,

$$Q_U(\boldsymbol{y}) := \sum_{x \in A_U(\boldsymbol{y})} 2^{-|\boldsymbol{x}|},$$

which we can write as

$$Q_U(\boldsymbol{y}) := \sum_{x \in A_U(\boldsymbol{y})} \lambda(\boldsymbol{x}),$$

for the *uniform* distribution λ.

S. (2017). A generalized characterization of algorithmic probability. *Theor. Comput. Sys.*

A so(m)ber conclusion

- The Solomonoff-Levin definition really doesn't give a convincing specification of a universal prediction method.
- The Solomonoff-Levin definition doesn't really give a convincing implementation of Occam's razor.

A so(m)ber conclusion

- The Solomonoff-Levin definition really doesn't give a convincing specification of a universal prediction method.
- The Solomonoff-Levin definition doesn't really give a convincing implementation of Occam's razor.

tom.sterkenburg@lmu.de
www.cwi.nl/~tom