

# The EL Theorem

Samuel Epstein  
samepst@jpttheorygroup.org

September 12, 2023

## Abstract

The combined universal probability  $\mathbf{m}(D)$  of strings  $x$  in sets  $D$  is close to  $\max \mathbf{m}(x)$  over  $x$  in  $D$ : their logs differ by at most  $D$ 's information  $\mathbf{I}(D : \mathcal{H})$  about the halting sequence  $\mathcal{H}$ .

## 1 Introduction

One common goal in computer science is to find the hidden part of the environment, this task has been called Inductive Inference, Extrapolation, Passive Learning, etc. The complete environment can be represented as a huge string  $x \in \Sigma^*$ . The known observations restrict it to a set  $D \subset \Sigma^*$ . For example in thermodynamics, the environment  $x$  can be seen as a record of every particle's position and velocity in a closed box. An observation of some macro parameters, such as pressure and temperature, restricting the possible environments to a set  $D$  of hypotheses consistent with the observation.

One method used to select a hypothesis (i.e. environment) is to leverage an *apriori* distribution over the environment space. This distribution  $p$  encodes any knowledge about the environment known before the observation is made. Then selection of the hypothesis is

$$\arg \max_{x \in D} p(x).$$

Note in AIT, for enumerable distributions (i.e. generatable as outputs of randomized algorithms), there is a universal apriori distribution  $\mathbf{m}(x)$ . This is because  $O(1)\mathbf{m} > p$ , for all enumerable  $p$ . Furthermore, for all  $x \in \Sigma^*$ ,  $\mathbf{d}(x|\mathbf{m}) = O(1)$ , where  $\mathbf{d}$  is deficiency of randomness; so there is no lower computable refutation to the statement: “ $x$  is generated from  $\mathbf{m}$ ”. Thus when the universal prior is used, inductive inference becomes an exercise of Occam's razor:

$$\arg \min_{x \in D} \mathbf{K}(x).$$

However there exists a potential complication. It could be there is a collection  $G \subset D$  of hypotheses representing a concept (such as a more detailed description of particles) where its combined apriori measure is greater than that of the simplest element  $x$ , with  $\mathbf{m}(x) \ll \mathbf{m}(G)$ . Or, making the endeavor more murkier, it could be that  $G$  is just the set of all complicated hypothesis and  $G$  has greater combined apriori measure than the simplest element. In this case, which explanation does one choose?

The EL Theorem shows that this dilemma is purely a mathematical construction. All the universal apriori measure of an observation  $D$  is concentrated on its simplest member. This is true for all non-exotic set  $D$  with low mutual information with the halting sequence,  $\mathbf{I}(D; \mathcal{H})$ . There are no (randomized) algorithmic means of creating  $D$  with arbitrarily high  $\mathbf{I}(D; \mathcal{H})$ .

## 2 Related Work

For information relating to the history of Algorithmic Information Theory and Kolmogorov complexity, we refer the readers to the textbooks [LV08] and [DH10]. A survey about the shared information between strings and the halting sequence is in the work [VV04]. Work on the deficiency of randomness can be found in [She83, KU87, VY87, She99]. Stochasticity of objects can be found in the works [She83, She99, VY87, VY99]. More information on stochasticity and algorithmic statistics are in the works [GTV01, VS17, VS15]. The EL Theorem is joint work between the author and L. A. Levin who published this result in [Lev16].

## 3 Conventions

As noted in the introduction,  $\mathbf{K}(x|y)$  is the conditional prefix free Kolmogorov complexity.  $\mathbf{m}(x)$  is the algorithmic probability.  $\mathbf{I}(x; \mathcal{H}) = \mathbf{K}(x) - \mathbf{K}(x|\mathcal{H})$  is the amount of information that the halting sequence  $\mathcal{H} \in \Sigma^\infty$  has about  $x$ . A probability is *elementary*, if it has finite support and rational values. The deficiency of randomness of  $x$  relative to a elementary probability measure  $Q$  is  $\mathbf{d}(x|Q) = -\log Q(x) - \mathbf{K}(x|Q)$ . We recall for a set  $D \subseteq \Sigma^*$ ,  $\mathbf{m}(D) = \sum_{x \in D} \mathbf{m}(x)$ . For the nonnegative real function  $f$ , we use  $<^+ f$ ,  $>^+ f$ , and  $=^+ f$  to denote  $< f + O(1)$ ,  $> f - O(1)$ , and  $= f \pm O(1)$ . We also use  $<^{\log} f$  and  $>^{\log} f$  to denote  $< f + O(\log(f+1))$  and  $> f - O(\log(f+1))$ , respectively.

## 4 The EL Theorem

**Definition 1 (Stochasticity)** *A string  $x$  is  $(\alpha, \beta)$ -stochastic if there exists an elementary probability measure  $Q$  such that*

$$\mathbf{K}(Q) \leq \alpha \text{ and } \mathbf{d}(x|Q) \leq \beta.$$

**Theorem 1 (Epstein, Levin)** *Let  $P$  be a lower-semicomputable semimeasure and  $c$  be a large constant. Every  $(\alpha, \beta)$ -stochastic set  $D$  with  $s = \lceil -\log P(D) \rceil$  contains an element  $x$  with*

$$\mathbf{K}(x) < s + \alpha + 2 \log \beta + \mathbf{K}(s) + 2 \log \mathbf{K}(s) + c.$$

The theorem is directly implied by the following lemma.

**Lemma 1** *Let  $P$  be a lower-semicomputable semimeasure and  $c$  be a large constant. If a set  $D$  is  $(\alpha, \beta)$ -stochastic relative to an integer  $s = \lceil -\log P(D) \rceil$ , then  $D$  contains an element  $x$  with*

$$\mathbf{K}(x) < s + \alpha + \log \beta + \mathbf{K}(\log \beta) + \mathbf{K}(s) + c.$$

*Note that if  $y$  is  $(\alpha, \beta)$ -stochastic relative to  $s$ , then it is  $(\alpha, \beta + \mathbf{K}(s))$ -stochastic. Hence the lemma implies the theorem.*

**Lemma 2** *Let  $P$  be a discrete measure and  $Q$  be a measure on sets. There exists a set  $S$  of size  $\lceil \beta/\gamma \rceil$  such that*

$$Q(\{D : P(D) \geq \gamma \text{ and } D \text{ is disjoint from } S\}) \leq \exp(-\beta).$$

**Proof.** We use the probabilistic method, and show that if we draw  $\lceil \beta/\gamma \rceil$  elements according to the distribution  $P$ , then the obtained set  $S$  satisfies the inequality with positive probability. The probability that a fixed set  $D$  with  $P(D) \geq \gamma$  is disjoint from  $S$  is

$$\leq (1 - \gamma)^{\beta/\gamma} \leq \exp(-\beta).$$

Hence the expected  $Q$ -measure of such a  $D$  is at most  $\exp(-\beta)$  and the required set  $S$  exists.  $\square$

**Proof of Lemma 1 for computable  $P$ .** Let  $Q$  be an elementary probability measure with  $\mathbf{K}(Q) \leq \alpha$  and  $\mathbf{d}(D|Q, s) \leq \beta$ . Without loss of generality, we assume that  $\beta$  is large positive power of 2. Fix a search procedure that on input  $Q$ ,  $\beta$ , and  $\gamma = 2^{-s}$  finds a set satisfying the conditions of Lemma 2.

For large  $\beta$ , the set  $D$  must intersect the obtained set  $S$ . Indeed, consider the  $Q$ -test  $g(X|Q, s)$  that is equal to  $\exp(\beta)$  if  $X$  is disjoint from  $S$ , and is zero otherwise. This is indeed a test, because the above lemma implies that its expected value for  $X \sim Q$  is bounded by 1. Since the test is also computable, it is a lower bound to the optimal test  $\mathbf{t}(X|Q, s)$ , up to a constant factor. By stochasticity of the set  $D$ ,  $g(D|Q, s) < O(1)\mathbf{t}(D|Q, s) < O(2^\beta)$ , because  $2^{\mathbf{d}(X|Q, s)}$  is an optimal  $Q$  test relative to  $s$ . Thus for large enough  $\beta$ ,  $D$  intersects  $Q$ .

It remains to construct a description of each element in  $S$  of the size given in the proposition. We construct a special decompressor that assigns short description to each element in  $S$ . On input of a string, the decompressor interprets the string as a concatenation of 4 parts:

1. A prefix-free description of  $Q$  of size at most  $\alpha$ .
2. A prefix-free description of  $\log \beta$  of size  $\mathbf{K}(\log \beta)$ .
3. A prefix-free description of  $s$  of size  $\mathbf{K}(s)$ .
4. An integer of bitsize  $\log(\beta/\gamma) = s + \log \beta$ .

It interprets the last integer as the index of an element in the set  $S$  of size  $\lceil \beta/\gamma \rceil$  that is computed by the search procedure on input  $Q$ ,  $\beta$ , and  $\gamma$ . The element is the output of the decompressor. The proposition is proven for computable  $P$ .  $\square$

**Remark 1** *If  $P$  is computable, a set  $S$  satisfying the conditions of the lemma can be easily searched. But if  $P$  is not computable, then the collection of sets  $D$  with  $P(D) \geq \gamma$  grows over time. Thus after constructing a good  $S$ , it can happen that a large  $Q$ -measure of sets  $D$  appears that does not contain an element from  $S$ , and that new elements to  $S$  need to be added. This type of interactive construction leads to an equivalent characterization of the problem in terms of a game which is shown in [She12]. Below, another proof is presented.*

**Proof of Lemma 1 for lower-semicomputable  $P$ .** We still assume that  $\beta$  is a large power of 2. Let  $\gamma = 2^{-s}/2$ . We can rewrite  $P = \frac{\gamma}{\beta}(P_1 + \dots + P_f + P_*)$ , with  $f \leq \beta/\gamma$ , such that  $P_1, \dots, P_f$  are probability measures with finite support obtained by a lower semi-computable approximation of  $P$ , and  $P_*$  is a lower-semicomputable semimeasure.

*Construction of a lower-semicomputable test  $g$  over sets.* We first construct tests  $g_1, \dots, g_f$  together with a list of strings  $z_1, \dots, z_f$ . Let  $g_0(X) = 1$ . Assume we already constructed  $z_1, \dots, z_{i-1}$  and  $g_{i-1}$  for some  $i = 1, \dots, f$ . Choose  $z_i$  such that the test

$$g_i(X) = \begin{cases} g_{i-1}(X) & \text{if } g_{i-1}(X) \geq \exp(\beta) \\ \exp(P_i(X))g_{i-1}(X) & \text{if } g_{i-1}(X) < \exp(\beta) \text{ and } X \text{ is disjoint from } \{z_1, \dots, z_i\} \\ 0 & \text{otherwise.} \end{cases}$$

satisfies  $\mathbf{E}g_i(X) \leq \mathbf{E}g_{i-1}(X)$  where the expectations are taken for  $X \sim Q$ . Let  $g(X)$  be equal to  $\exp(\beta)$  if there exists an  $i$  such that  $g_i(X) \geq \exp \beta$ , otherwise let  $g(X) = 0$ . *End of construction*

We first show that each required string  $z_i$  in the construction exists. Suppose  $z_1, \dots, z_{i-1}$  and  $g_{i-1}$  have already been constructed. We show the existence of  $z_i$  using the probabilistic method. If we draw  $z_i$  according to  $P_i$ , then for each set  $X$  for which the second condition of  $g_i$  is satisfied, we have

$$\mathbf{E}_{z_i \sim P_i} g_i(X) \leq (1 - P_i(X))g_{i-1}(X) \exp P_i(X) \leq g_{i-1}(X),$$

because of the inequality  $1 + r \leq \exp(r)$  for all reals  $r$ . If  $X$  satisfies the first or third condition, then  $\mathbf{E}g_i(X) \leq \mathbf{E}g_{i-1}(X)$  is trivially true. So

$$\begin{aligned} \mathbf{E}_{X \sim Q} \mathbf{E}_{z_i \sim P_i} g_i(X) &\leq \mathbf{E}_{X \sim Q} g_{i-1}(X), \\ \mathbf{E}_{z_i \sim P_i} \mathbf{E}_{X \sim Q} g_i(X) &\leq \mathbf{E}_{X \sim Q} g_{i-1}(X), \end{aligned}$$

and the required  $z_i$  exists.

We have  $G(x) \leq O(\mathbf{t}(X|Q, (\gamma, \beta)))$ , where  $\mathbf{t}$  is the optimal test because the construction implies  $\mathbf{E}g \leq 1$  and is effective, thus  $g$  is lower semicomputable. Every set  $X$  with  $P(X) \geq 2^{-s} = 2\gamma$  satisfies  $P_1(X) + \dots + P_f(X) \geq \frac{\beta}{\gamma}P(D) - 1 \geq 2\beta - 1 \geq \beta$  by choice of  $P_i$ . Any such  $X$  that is disjoint from the set  $\{z_1, \dots, z_f\}$  satisfies

$$g_f(X) = \exp(P_1(X)) \exp(P_2(X)) \dots \exp(P_f(X)) \geq \exp(\beta).$$

This implies  $\mathbf{d}(X|Q, s) > \beta$  for large  $\beta$ , because up to  $O(1)$  constants, we have

$$1.44\beta \leq \log g(X) \leq \mathbf{d}(X|Q, (\beta, \gamma)) \leq \mathbf{d}(X|Q, s) + 2 \log \beta.$$

By the assumption on  $(\alpha, \beta)$ -stochasticity of  $D$ , we have  $\mathbf{d}(D|Q, s) \leq \beta$  and hence  $D$  must contain some  $z_j$ . The theorem follows by constructing a description for each string  $z_i$  of bitsize  $s + \alpha + \log \beta + \mathbf{K}(\log \beta) + \mathbf{K}(s)$  in a similar way as above.  $\square$

## 4.1 Non-Stochastic Objects

It is well known in the literature that non-stochastic objects have high mutual information with the halting sequence [VS17]. In the following lemma, we reprove this fact, without using left-total machines, which was used in the original proof.

**Lemma 3**  $\Lambda(x) <^{\log} \mathbf{I}(x; \mathcal{H})$ .

**Proof.** We dovetail all programs to the universal Turing machine  $U$ . For  $p \in \text{Domain}(U)$ ,  $n(p) \in \mathbb{N}$  is the position in which the program  $p \in \Sigma^*$  terminates. Let  $\Omega^n = \sum_{p: n(p) < n} 2^{-\|p\|}$  and  $\Omega = \Omega^\infty$  be Chaitin's Omega. Let  $\Omega_t^n$  be  $\Omega^n$  restricted to the first  $t$  digits. Let  $x^* \in \Sigma^{\mathbf{K}(x)}$ , with  $U(x^*) = x$  with minimum  $n(x^*)$ . Let  $k(p) = \max\{\ell : \Omega_\ell^{n(p)} = \Omega_\ell\}$  and  $k = k(x^*)$ . We define the elementary probability measure  $Q(x) = \max\{2^{-\|p\|+k} : k(p) = k, U(p) = x\}$ ,  $Q(\emptyset) = 1 - Q(\Sigma^* \setminus \{\emptyset\})$ .

$$\begin{aligned} \mathbf{d}(x|Q) &= -\log Q(x) - \mathbf{K}(x|Q) <^+ (\mathbf{K}(x) - k) - \mathbf{K}(x|\Omega_k) \\ &<^+ (\mathbf{K}(x|\Omega_k) + \mathbf{K}(\Omega_k) - k) - \mathbf{K}(x|\Omega_k) <^+ (k + \mathbf{K}(k)) - k \\ &<^+ \mathbf{K}(k). \end{aligned}$$

$$\begin{aligned} \mathbf{K}(x|\mathcal{H}) &<^+ \mathbf{K}(x|Q) + \mathbf{K}(Q|\mathcal{H}) <^+ \mathbf{K}(x|Q) + \mathbf{K}(\Omega_k|\mathcal{H}) \\ &<^+ -\log Q(x) + \mathbf{K}(k) <^+ (\mathbf{K}(x) - k) + \mathbf{K}(k) \\ k &<^{\log} \mathbf{K}(x) - \mathbf{K}(x|\mathcal{H}) \end{aligned}$$

$$\Lambda(x) <^+ \mathbf{K}(Q) + O(\log \max\{\mathbf{d}(x|P), 1\}) <^+ k + O(\mathbf{K}(k)) <^{\log} \mathbf{I}(x; \mathcal{H}).$$

□

**Corollary 1 (EL Theorem)**

For finite  $D \subset \Sigma^*$ ,  $\min_{x \in D} \mathbf{K}(x) <^{\log} -\log \mathbf{m}(D) + \mathbf{I}(D; \mathcal{H})$ .

**Proof.** This follows from Theorem 1 and Lemma 3

□

## References

- [DH10] R. G. Downey and D.R. Hirschfeldt. *Algorithmic Randomness and Complexity*. Theory and Applications of Computability. Springer New York, 2010.
- [GTV01] P. Gács, J. Tromp, and P. Vitányi. Algorithmic Statistics. *IEEE Transactions on Information Theory*, 47(6):2443–2463, 2001.
- [KU87] A. N. Kolmogorov and V. A. Uspensky. Algorithms and Randomness. *SIAM Theory of Probability and Its Applications*, 32(3):389–412, 1987.
- [Lev16] L. A. Levin. Occam bound on lowest complexity of elements. *Annals of Pure and Applied Logic*, 167(10):897–900, 2016. And also: S. Epstein and L.A. Levin, Sets have simple members, arXiv preprint arXiv:1107.1458, 2011.
- [LV08] M. Li and P. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer Publishing Company, Incorporated, 3 edition, 2008.
- [She83] A. Shen. The concept of (alpha,beta)-stochasticity in the Kolmogorov sense, and its properties. *Soviet Mathematics Doklady*, 28(1):295–299, 1983.
- [She99] A. Shen. Discussion on Kolmogorov Complexity and Statistical Analysis. *The Computer Journal*, 42(4):340–342, 1999.

- [She12] A. Shen. Game arguments in computability theory and algorithmic information theory. *ArXiv e-prints*, 2012. <http://http://arxiv.org/abs/1204.0198>.
- [VS15] Nikolai K. Vereshchagin and Alexander Shen. Algorithmic statistics revisited. *CoRR*, abs/1504.04950, 2015.
- [VS17] Nikolai K. Vereshchagin and Alexander Shen. Algorithmic statistics: Forty years later. In *Computability and Complexity*, pages 669–737, 2017.
- [VV04] N. Vereshchagin and P. Vitányi. Kolmogorov’s Structure Functions and Model Selection. *IEEE Transactions on Information Theory*, 50(12):3265 – 3290, 2004.
- [V’Y87] V.V. V’Yugin. On Randomness Defect of a Finite Object Relative to Measures with Given Complexity Bounds. *SIAM Theory of Probability and Its Applications*, 32:558–563, 1987.
- [V’Y99] V.V. V’Yugin. Algorithmic complexity and stochastic properties of finite binary sequences, 1999.