



# Prediction and MDL for infinite sequences

Alexey Milovanov<sup>1</sup> 

Accepted: 4 March 2024  
© The Author(s) 2024

## Abstract

We combine Solomonoff's approach to universal prediction with algorithmic statistics and suggest to use the computable measure that provides the best "explanation" for the observed data (in the sense of algorithmic statistics) for prediction. In this way we keep the expected sum of squares of prediction errors bounded (as it was for the Solomonoff's predictor) and, moreover, guarantee that the sum of squares of prediction errors is bounded along any Martin-Löf random sequence. An extended abstract of this paper was presented at the 16th International Computer Science Symposium in Russia (CSR 2021) (Milovanov 2021).

## 1 Introduction

We consider probability distributions (or measures) on the binary tree, i.e., non-negative functions  $P : \{0, 1\}^* \rightarrow \mathbb{R}$  such that  $P(\text{empty word}) = 1$  and  $P(x0) + P(x1) = P(x)$  for every string  $x$ . We assume that all the values  $P(x)$  are rational;  $P$  is called *computable* if there exists an algorithm that on input  $x$  outputs  $P(x)$ .

Consider the following prediction problem. Imagine a black box that generates bits according to some unknown computable distribution  $P$  on the binary tree. Let  $x = x_1 \dots x_n$  be the current output of the black box. The predictor's goal is to guess the probability that the next bit is 1, i.e., the ratio  $P(1|x) = P(x1)/P(x)$ .

Ray Solomonoff suggested to use the universal semi-measure  $M$  (called also the *a priori probability*) for prediction. Recall that a semi-measure  $S$  on the binary tree (a *continuous semi-measure*) is a non-negative function  $S : \{0, 1\}^* \rightarrow \mathbb{R}$  such that  $S(\text{empty word}) \leq 1$  and  $S(x0) + S(x1) \leq S(x)$  for every string  $x$ . Semi-measures correspond to probabilistic processes that output a bit sequence but can hang forever, so an output may be some finite string  $x$ ; the probability of this event is

---

✉ Alexey Milovanov  
almas239@gmail.com

<sup>1</sup> LASIGE Faculdade de Ciências, Universidade de Lisboa, Lisboa, Portugal

$S(x) - S(x0) - S(x1)$ . A semi-measure  $S$  is called *lower semi-computable*, or *enumerable*, if the set  $\{(x, r) : r < S(x)\}$  is (computably) enumerable. Here  $x$  is a string and  $r$  is a rational number. Finally, a lower semi-computable semi-measure  $M$  is called *universal* if it is maximal among all semimeasures up to a constant factor, i.e., if for every lower semi-computable semi-measure  $S$  there exists  $c > 0$  such that  $M(x) \geq cS(x)$  for all  $x$ . Such a universal semi-measure exists [7, 11, 12].<sup>1</sup>

Solomonoff suggested to use the ratio  $M(1|x) := M(x1)/M(x)$  to predict  $P(1|x)$  for an unknown computable measure  $P$ . He proved the following bound for the prediction errors.

**Theorem 1** ([13]) *For every computable distribution  $P$  and for every  $b \in \{0, 1\}$  the following sum over all binary strings is finite:*

$$\sum_x P(x) \cdot (P(b|x) - M(b|x))^2 < \infty. \quad (1)$$

Moreover, this sum is bounded by  $O(K(P))$ , where  $K(P)$  is the prefix complexity of the computable measure  $P$  (the minimal length of a prefix-free program corresponding  $P$ ).

Note that for semi-measure the probabilities to predict 0 and 1 do not sum up to 1, so the statements for  $b = 0$  and  $b = 1$  are not equivalent (but both are true).

The sum from Theorem 1 can be rewritten as the expected value of the function  $D$  on the infinite binary sequences with respect to  $P$ , where  $D(\omega)$  is defined as

$$D(\omega) = \sum_{x \text{ is a prefix of } \omega} (P(b|x) - M(b|x))^2.$$

This expectation is finite, therefore for  $P$ -almost all  $\omega$  the value  $D(\omega)$  is finite and

$$P(b|x) - M(b|x) \rightarrow 0.$$

when  $x$  is an increasing prefix of  $\omega$ . One would like to have this convergence for all Martin-Löf random sequences  $\omega$  (with respect to measure  $P$ ), but this is not guaranteed, since the null set provided by the argument above may not be an *effectively* null set. An example from [5] shows that this is indeed the case.

**Theorem 2** ([5]) *There exist a specific universal semi-measure  $M$ , computable distribution  $P$  and Martin-Löf random (with respect to  $P$ ) sequence  $\omega$  such that*

$$P(b|x) - M(b|x) \not\rightarrow 0.$$

for increasing prefixes  $x$  of  $\omega$ .

<sup>1</sup> One may even require that the probabilities for finite outputs, i.e., the differences  $S(x) - S(x0) - S(x1)$  are maximal, but we do not require this.

Lattimore and Hutter generalized Theorem 2 by proving the same statement for a wide class of universal semi-measures [8].

Trying to overcome this problem and get a good prediction for all Martin-Löf random sequences, we suggest the following approach to prediction. For a finite string  $x$  we find a distribution  $Q$  on the binary tree that is the best (in some sense) explanation for  $x$ . The probabilities of the next bits are then predicted as  $Q(0|x)$  and  $Q(1|x)$ .

This approach combines two advantages. The first is that the series of type (1) also converges, though the upper bound for it (at least the one that we are able to prove) is much greater than  $O(K(P))$ . The second property is that the prediction error (defined as in Theorem 2) converges to zero for every Martin-Löf random sequence.

Let us give formal definitions. The quality of the computable distribution  $Q$  on the binary tree, considered as an “explanation” for a given string  $x$ , is measured by the value  $3K(Q) - \log Q(x)$ : the smaller this quantity is, the better is the explanation. One can rewrite this expression as the sum

$$2K(Q) + [K(Q) - \log Q(x)].$$

Here the expression in the square brackets can be interpreted as the length of the two-part description of  $x$  using  $Q$  (first, we specify the hypothesis  $Q$  using its shortest prefix-free program, and then, knowing  $Q$ , we specify  $x$  using arithmetic coding; the second part requires about  $-\log Q(x)$  bits). The first term  $2K(Q)$  is added to give extra preference to simple hypotheses; the factor 2 is needed for technical reasons (in fact, any constant greater than 1 will work<sup>2</sup>).

For a given  $x$  we select the best explanation that makes this quality minimal. Formally, we find

$$Q_x = \operatorname{argmin}_Q \{3K(Q) - \log Q(x)\}.$$

So,  $Q_x$  is the best explanation for string  $x$  (or one of the best explanations if there are several).

Then we predict the probability that the next bit after  $x$  is  $b$ :

$$H(b|x) := \frac{Q_x(xb)}{Q_x(x)},$$

In this paper we prove the following results:

**Theorem 3** *For every computable distribution  $P$  the following sum over all binary strings  $x$  is finite:*

$$\sum_x P(x)(P(0|x) - H(0|x))^2 < \infty.$$

<sup>2</sup> One can consider a similar prediction method with factor  $1 + \alpha$  for arbitrary positive  $\alpha$  instead of factor 2. However it does not give an significant improvement so we do not do it.

**Theorem 4** Let  $P$  be a computable measure and let  $\omega$  be a Martin-Löf random sequence with respect to  $P$ . Then

$$H(0|x) - P(0|x) \rightarrow 0$$

for prefixes  $x$  of  $\omega$  as the length of prefix goes to infinity.

We speak about the probabilities of zeros, but both  $P$  and  $Q$  are measures, so this implies the same results for the probabilities of ones.

We prove that

$$\sum_{x \text{ is a prefix of } \omega} (H(0|x) - P(0|x))^2 < \infty$$

(Theorem 7) that is the strengthening of Theorem 4.

**Remark 1** Note that function  $H$  as well as Solomonoff's predictor is uncomputable. Indeed, consider some computable predictor  $H$ . Then there exists a computable distribution  $P$  such that  $|H(0|x) - P(0|x)| \geq \frac{1}{2}$  for every  $x$ . Of course this predictor does not satisfy Theorem 3.

In [3] Hutter suggested a similar approach but without coefficient 3 for  $K(Q)$  (see also [2, 4]). For this approach he proved an analogue of Theorem 3 with different proof technique.

In [5] the existence of a semi-computable measure satisfying Theorem 4 was proved. In [6] authors prove an analogue of Theorem 3 for the Hellinger distance with double exponential in  $K(P)$  bound.

In the next section we prove Theorem 4.

In Section 3 we prove Theorem 3.

Finally, in Section 4 we consider the case when we know some information about  $P$ . More precisely, we know that  $P$  belongs to some enumerable set of computable measures. We suggest a similar approach for prediction in this case. We prove analogues of Theorems 4 and 3 (Theorems 9 and 10) for this prediction method. We achieved better (polynomial in complexity of  $P$ ) error estimations in these theorems.

## 2 Prediction on Martin-Löf Random Sequences

Recall the Schnorr–Levin theorem [11, ch.5] that says that a sequence  $\omega$  is random with respect to a computable probability measure  $P$  if and only if the ratio  $M(x)/P(x)$  is bounded for  $x$  that are prefixes of  $\omega$ .

The same result can be reformulated in the logarithmic scale. Let us denote by  $KM(x)$  the *a priori complexity* of  $x$ , i.e.,  $\lceil -\log M(x) \rceil$  (the rounding is chosen in this way to ensure upper semicomputability of  $KM$ ). We have

$$KM(x) \leq -\log P(x) + O(1)$$

for every computable probability measure  $P$ , where  $O(1)$  depends on  $P$  but not on  $x$ . Indeed, since  $M$  is maximal, the ratio  $P(x)/M(x)$  is bounded. Moreover, since  $P(x)$

can be included in the mix for  $M(x)$  with coefficient  $2^{-K(P)}$ , we have

$$KM(x) \leq -\log P(x) + K(P) + O(1)$$

with some constant in  $O(1)$  that does not depend on  $P$  (and on  $x$ ). As we have discussed in the previous section, the right-hand side includes the length of the two-part description of  $x$ .

Let us call

$$d(x|P) := -\log P(x) - KM(x)$$

the *randomness deficiency* of a string  $x$  with respect to a computable measure  $P$ . (There are several notions of deficiency, but we need only this one.) Then we get

$$d(x|P) \geq -K(P) - O(1)$$

so the deficiency is almost non-negative. The Schnorr–Levin theorem characterizes Martin-Löf randomness in terms of deficiency:

**Theorem 5** (Schnorr–Levin)

(a) *If a sequence  $\omega$  is Martin-Löf random with respect to a computable distribution  $P$ , then  $d(x|P)$  is bounded for all prefixes  $x$  of  $\omega$ .*

(b) *Otherwise (if  $\omega$  is not random with respect to  $P$ ), then  $d(x|P) \rightarrow \infty$  as the length of a prefix  $x$  of  $\omega$  increases.*

Note that there is a dichotomy: the values  $d(x|P)$  for prefixes  $x$  of  $\omega$  either are bounded or converge to infinity (as the length of  $x$  goes to infinity). We can define randomness deficiency for infinite sequence  $\omega$  as

$$d(\omega|P) := \sup_{x \text{ is prefix of } \omega} d(x|P);$$

it is finite if and only if  $\omega$  is random with respect to  $P$ .

Let us also recall the following result of Vovk:

**Theorem 6** ([15]) *Let  $P$  and  $Q$  be two computable distributions. Let  $\omega$  be a Martin-Löf random sequence with respect both to  $P$  and  $Q$ . Then*

$$P(0|x) - Q(0|x) \rightarrow 0$$

for prefixes  $x$  of  $\omega$  as the length of prefix goes to infinity.

We will prove this theorem (and even more exact statement) in the next section.

**Proof of Theorem 4** Now we have a sequence  $\omega$  that is Martin-Löf random with respect to some computable measure  $P$ , so  $D = d(\omega|P)$  is finite. For each prefix  $x$  of  $\omega$  we take the best explanation  $Q$  that makes the expression

$$3K(Q) - \log Q(x)$$

minimal. Note that  $P$  is among the candidates for  $Q$ , so this expression should not exceed

$$3K(P) - \log P(x).$$

Since  $\omega$  is random with respect to  $P$  and  $x$  is a prefix of  $\omega$ , Schnorr–Levin theorem guarantees that the latter expression

$$3K(P) - \log P(x) = KM(x) + O(1)$$

where constant in  $O$  depends on  $P$  and  $\omega$  but not on  $x$ . On the other hand, the inequality  $KM(x) \leq K(Q) - \log Q(x) + O(1)$  implies that

$$3K(Q) - \log Q(x) = 2K(Q) + K(Q) - \log Q(x) \geq 2K(Q) + KM(x) - O(1). \quad (2)$$

So measures  $Q$  with large  $K(Q)$  cannot compete with  $P$ , and there is only a finite list of candidate measures for the best explanation  $Q$ . For some of these  $Q$  the sequence  $\omega$  is  $Q$ -random with respect to  $Q$ , so one can use Vovk's theorem to get the convergence of predicted probabilities when these measures are used.

Still we may have some "bad"  $Q$  in the list of candidates for which  $\omega$  is not  $Q$ -random. However, the Schnorr–Levin theorem guarantees that for a bad  $Q$  we have

$$-\log Q(x) - KM(x) \rightarrow \infty$$

if  $x$  is a prefix of  $\omega$  of increasing length. So the difference between two sides of (2) goes to infinity as the length of  $x$  increases, so  $Q$  loses to  $P$  for large enough  $x$  (is worse as an explanation of  $x$ ). Therefore, only good  $Q$  will be used for prediction after sufficiently long prefixes, and this finishes the proof of Theorem 4.  $\square$

### 3 On the Expectation of Squares of Errors

In this section we prove Theorem 3. First we will prove some strengthening of Theorem 6

**Lemma 1** *Let  $P$  and  $Q$  be computable distributions. and let  $M$  be a universal semi-measure. Assume that for string  $x = x_1 \dots x_n$  and  $C > 0$  it holds that  $P(x), Q(x) \geq M(x)/C$ . Then:*

$$\sum_{i=1}^{n-1} (P(x_i|x_1 \dots x_{i-1}) - Q(x_i|x_1 \dots x_{i-1}))^2 = O(\log C + K(P, Q)).$$

**Proof of Theorem 6 from Lemma 1** According to one of definitions of Martin-Löf randomness the values  $M(x)/P(x)$  and  $M(x)/Q(x)$  are bounded by a constant. It reminds to use Lemma 1.  $\square$

**Proof of Lemma 1** Denote

$$p_i = P(x_i | x_1 \dots x_{i-1}), \quad q_i = Q(x_i | x_1 \dots x_{i-1}).$$

Note that

$$P(x_1 \dots x_n) = p_1 p_2 \dots p_n, \quad Q(x_1 \dots x_n) = q_1 q_2 \dots q_n.$$

Now consider the “intermediate” measure  $R$  for which the probability of 0 (or 1) after some  $x$  is the average of the same conditional probabilities for  $P$  and  $Q$ :

$$R(0 | x_1 \dots x_{i-1}) = \frac{P(0 | x_1 \dots x_{i-1}) + Q(0 | x_1 \dots x_{i-1})}{2}.$$

The corresponding  $r_i = R(x_i | x_1 \dots x_{i-1})$  are equal to  $(p_i + q_i)/2$ .

Probability distribution  $R$  is computable and  $K(R) \leq K(P, Q) + O(1)$ . Hence, it holds that  $R(x) \leq 2^{K(P,Q)} M(x) \leq 2^{K(P,Q)} \cdot C \cdot P(x)$ . The similar inequality holds for distribution  $Q$ . Therefore:

$$r_1 \dots r_n \leq C \cdot 2^{K(P,Q)} \cdot p_1 \dots p_n$$

and

$$r_1 \dots r_n \leq C \cdot 2^{K(P,Q)} \cdot q_1 \dots q_n.$$

Multiplying we obtain:

$$\left(\frac{p_1 + q_1}{2} \dots \frac{p_n + q_n}{2}\right)^2 \leq C^2 \cdot 2^{2K(P,Q)} \cdot p_1 \dots p_n \cdot q_1 \dots q_n. \tag{3}$$

These two inequalities show that the product of arithmetical means of  $p_i$  and  $q_i$  is not much bigger than the product of their geometrical means, and this is only possible if  $p_i$  is close to  $q_i$  (logarithm is a strictly convex function).

To make the argument precise, recall the bound for the logarithm function:

**Lemma 2** For  $p, q \in (0, 1]$  we have

$$\log \frac{p + q}{2} - \frac{\log p + \log q}{2} \geq \frac{1}{8 \ln 2} (p - q)^2$$

**Proof** Let us replace the binary logarithms by the natural ones; then the factor  $\ln 2$  disappears. Note that the left hand side remains the same if  $p$  and  $q$  are multiplied by some factor  $c \geq 1$  while the right side can only increase. So it is enough to prove this for  $p = 1 - h$  and  $q = 1 + h$  for some  $h \in (0, 1)$ , and this gives

$$-\frac{\ln(1 - h) + \ln(1 + h)}{2} \geq \frac{1}{2} h^2;$$

and this happens because  $\ln(1 - h) + \ln(1 + h) = \ln(1 - h^2) \leq -h^2$ . □

For the product of  $n$  terms we get the following bound:

**Lemma 3** *If for  $p_1, \dots, p_n, q_1, \dots, q_n \in (0, 1]$  we have*

$$\left(\frac{p_1 + q_1}{2} \cdot \dots \cdot \frac{p_n + q_n}{2}\right)^2 \leq c p_1 \dots p_n q_1 \dots q_n,$$

*then  $\sum_i (p_i - q_i)^2 \leq O(\log c)$ , with some absolute constant hidden in  $O(\cdot)$ -notation.*

**Proof** Taking logarithms, we get

$$2 \sum_i \log \frac{p_i + q_i}{2} \leq \log c + \sum_i \log p_i + \sum_i \log q_i,$$

and therefore

$$\sum_i \left(\log \frac{p_i + q_i}{2} - \frac{\log p_i + \log q_i}{2}\right) \leq \frac{1}{2} \log c.$$

It remains to use Lemma 2 to get the desired inequality.

To complete the proof of Lemma 1 it remains to use in inequality (3) Lemma 3 for  $c := C^2 \cdot 2^{2K(P \cdot Q)}$ . □

Now we prove a strengthening of Theorem 4.

**Theorem 7** *Let  $P$  be a computable measure, let  $\omega$  be a Martin-Löf random sequence with respect to  $P$  such that  $d(\omega|P) = D$ .*

*Then*

$$\sum_{x \text{ is a prefix of } \omega} (H(0|x) - P(0|x))^2 = O((K(P) + D) \cdot 2^{\frac{3K(P)+D+O(1)}{2}}).$$

**Proof** Consider  $Q_x$  (the best distribution) for some  $x = x_1 \dots x_n$ . Then

$$3K(Q_x) - \log Q_x(x) \leq 3K(P) - \log P(x). \tag{4}$$

Since  $d(\omega|P) = D$  we obtain that

$$-\log P(x) \leq KM(x) + D. \tag{5}$$

Therefore,

$$-\log Q_x(x) \leq 3K(P) - \log P(x) \leq 3K(P) + KM(x) + D, \text{ so}$$

$$Q_x(x) \geq M(x) \cdot 2^{-3K(P)-D} \text{ and} \\ P(x) \geq M(x) \cdot 2^{-D}.$$

We want to estimate  $\sum_{i=1}^n (Q(0|x_1 \dots x_i) - P(0|x_1 \dots x_i))^2$  by Lemma 1. We can use this lemma for  $C = 2^{3K(P)+D}$ .



From (4), (5) and  $KM(x) \leq -\log Q(x) + K(Q)$  it follows that

$$K(Q_x) \leq \frac{3K(P) + D + O(1)}{2}. \tag{6}$$

Therefore by Lemma 1 we obtain

$$\sum_{i=1}^{n-1} (Q(0|x_1 \dots x_i) - P(0|x_1 \dots x_i))^2 = O(K(P) + D).$$

In fact, we can not use this lemma for the last term  $(Q(0|x) - P(0|x))^2$ . This term we just bound by 1.

So, every probability distribution that is the best for some  $x$  “contributes”  $O(K(P) + D)$  in the sum  $\sum_{x \text{ is a prefix of } \omega} (H(0|x) - P(0|x))^2$ .

There are at most  $2^{\frac{3K(P)+D+O(1)}{2}}$  such distribution by (6), so we obtain the required estimation.  $\square$

Recall the following well-known statement

**Proposition 8** *Let  $P$  be a computable distribution. Then the  $P$ -measure of all sequences  $\omega$  such that  $d(\omega|P) \geq D$  is not greater than  $2^{-D}$ .*

**Proof of Theorem 3** Denote by  $\Omega$  the set of all infinite sequences with zeros and ones. Note that

$$\sum_x P(x)(P(0|x) - H(0|x))^2 = \int_{(\Omega, P)} \sum_{x \text{ is a prefix of } \omega} (H(0|x) - P(0|x))^2.$$

By Theorem 7 we can estimate the sum in the integral for sequence  $\omega$  with  $d(x|\omega) = D$  as  $O((K(P) + D) \cdot 2^{\frac{3K(P)+D+O(1)}{2}})$ . By Proposition 8 the measure of sequences with such randomness deficiency is at most  $2^{-D}$ . So we can estimate the integral as

$$\sum_{D=0}^{\infty} O((K(P) + D) \cdot 2^{\frac{3K(P)+D+O(1)}{2}})2^{-D} = O(K(P)2^{\frac{3K(P)}{2}}).$$

(Recall that the  $P$ -measure of sequences that are not Martin-Löf random with respect to  $P$  is equal to 0, so they do not affect to the integral.)  $\square$

### 4 Prediction for Enumerable Classes of Hypotheses

Assume that we have some information about distribution  $P$ . We know that  $P$  belongs to some enumerable set  $\mathcal{A}$  of computable distributions, (i.e. there is an algorithm that enumerate programs that generate distributions from  $\mathcal{A}$ ). For this case it is natural to consider the following measure of complexity in  $\mathcal{A}$ :

$$K_{\mathcal{A}}(P) := K(i_P), \text{ where } i_P \text{ is the number of } P \text{ in a computable enumeration of } \mathcal{A}.$$

If  $P$  appears multiple times in an enumeration, we select  $i_P$  with the lowest complexity. (This definition depends on the choice of a computable enumeration, but the influence of this dependence is constrained by an additive constant.) Clearly,  $K_{\mathcal{A}}(P) \geq K(P) + O(1)$ .

We can now extend our prediction method. To predict the next bit of  $x$ , we choose  $Q \in \mathcal{A}$  with the lowest value of  $3K_{\mathcal{A}}(Q) - \log Q(x)$  and base our next-bit prediction on  $Q$  :

$$H_{\mathcal{A}}(x) := \frac{Q(xb)}{Q(x)}.$$

In this section, we demonstrate that if set  $\mathcal{A}$  possesses certain favorable properties, analogous results to previous theorems emerge. Furthermore, we can achieve an improved error estimation.

We assume that enumerable set  $\mathcal{A}$  has the following property: if  $P_1, \dots, P_k \in \mathcal{A}$  then their mixture  $\frac{P_1 + \dots + P_k}{k}$  belongs to  $\mathcal{A}$ . Moreover there exists an algorithm that for given numbers of  $P_1, \dots, P_k$  (in some enumeration of  $\mathcal{A}$ ) outputs the number of their mixture.

Further everywhere  $\mathcal{A}$  is an enumerable set of computable distributions with this property.

**Remark 2** Consider the following example of set  $\mathcal{A}$ : the set of all *provable* (in some proof system) computable distributions on the binary tree. For every program  $p \in \mathcal{A}$ , there exists a proof that  $p(x)$  halts for every  $x$ ,  $p(x) = p(x0) + p(x1)$ , and  $p(\text{empty word}) = 1$ . We guess that all using in practice computable distributions are provable computable, so, in some sense we get better error estimation “almost free”. Our discussion about practice might look unsuitable because our prediction method is not computable. However, it can be considered as the ideal version of the MDL principle prediction which is approximated in practice.

**Theorem 9** *Let  $P \in \mathcal{A}$  be a computable measure, let  $\omega$  be a Martin-Löf random sequence with respect to  $P$  such that  $d(\omega|P) = D$ .*

*Then*

$$\sum_{x \text{ is a prefix of } \omega} (H_{\mathcal{A}}(0|x) - P(0|x))^2 = O((K_{\mathcal{A}}(P) + D) \cdot \text{poly}(K_{\mathcal{A}}(P) + D)).$$

**Theorem 10** *For every computable distribution  $P \in \mathcal{A}$  the following sum over all binary strings  $x$  is finite:*

$$\sum_x P(x)(P(0|x) - H(0|x))^2 < \text{poly}(K_{\mathcal{A}}(P)).$$

The proofs of these theorems are in general the same as the Proofs of Theorems 7 and 3, however some new tools are added. The difference is that we can get better estimation on the number of possible best explanations for prefixes of some sequence.

**Lemma 4** *Let  $x$  be a finite string. Assume that there are  $2^k$  probability distributions  $Q_1, \dots, Q_{2^k} \in \mathcal{A}$  such that for every  $i$  it holds  $K_{\mathcal{A}}(Q_i) \leq a$  and  $Q_i(x) \geq 2^{-b}$ . Then there is probability distribution  $Q \in \mathcal{A}$  such that  $K_{\mathcal{A}}(Q) \leq a - k + O(\log(a + k))$  and  $Q(x) \geq 2^{-b-k}$ .*

Note that  $3K_{\mathcal{A}}(Q) - \log Q(x) \leq 3 \cdot (a - k) + O(\log(a + k)) + b + k \leq 3 \cdot a - b$  for big enough  $k$ . This means that string  $x$  can not has many “best” explanations.

**Proof of Lemma 4** Let enumerate all distributions of  $\mathcal{A}$  with complexity at most  $a$  by groups of size  $2^{k-1}$  (the last group can be incomplete). The number of such groups is  $O(2^{a-k})$ . The complexity of every group is at most  $a - k + O(\log(a + k))$ . Indeed, to describe a group we need its ordinal number in an enumeration and describe this enumeration (we need to know  $k$ ,  $a$  and some enumeration of  $\mathcal{A}$ ).

One of these complete group contains some  $Q_i$ . Define  $Q$  as the mixture of the distributions in this group. Since the group has complexity at most  $a - k + O(\log a + k)$  the same estimation holds for the complexity of  $Q$ . Since some  $Q_i$  belongs to the mixture it holds that  $Q(x) \geq 2^{-b-k+1}$ . Recall that  $Q$  belongs to  $\mathcal{A}$  because every mixture of distributions from  $\mathcal{A}$  belongs to  $\mathcal{A}$ . □

Also we need the following lemma.

**Lemma 5** *Let string  $s$  be a prefix of string  $h$  and let  $P$  be a computable distribution such that  $d(s|P) = D$ . Then  $d(h|P) \geq D - 2 \log D + O(1)$ .*

(So, a prefix oconf a string that has small deficiency, has (almost as) small deficiency).

In fact the proof of this lemma is the same as the proof of Theorem 124 in [11].

**Proof of Lemma 5** For each  $k$  consider the enumerable set of all finite sequences that have deficiency greater than  $k$ . All the infinite continuations of these sequences form an open set  $S_k$ , and  $P$ -measure of this set does not exceed  $2^{-k}$ . Now consider the measure  $P_k$  on  $\Omega$  that is zero outside  $S_k$  and is equal to  $2^k P$  inside  $S_k$ . That means that for every set  $U$  the value  $P_k(U)$  is defined as  $2^k P(U \cap S_k)$ . Actually,  $P_k$  is not a probability distribution according to our definition, since  $P_k(\Omega)$  is not equal to 1. However,  $P_k$  can be considered as a lower semi-computable semi-measure, if we change it a bit and let  $P_k(\Omega) = 1$  (this means that the difference between 1 and the former value of  $P_k(\Omega)$  is assigned to the empty string).

It is clear that  $P_k$  is lower semi-computable since  $P$  is computable and  $S_k$  is enumerable.

Now consider the sum

$$S = \sum_k \frac{1}{k(k-1)} P_k$$

It is a lower semi-computable semi-measure again Then we have

$$-\log S(x) \leq -\log P(x) - k + 2 \log k + O(1)$$

for every string  $x$  that has a prefix with deficiency greater than  $k$ . Since  $S$  does not exceed a priori probability (up to  $O(1)$ -factor), we get the desired inequality. □

**Proof of Theorem 9 Part 1.** We claim that there are only  $\text{poly}(D + K_{\mathcal{A}}(P))$  different distributions that are the best for some prefix of  $\omega$ .

Let  $x$  be a prefix of  $\omega$  and  $Q$  is the best distribution for  $x$ . As in the proof of Theorem 7 (see (6)) we obtain

$$K_{\mathcal{A}}(Q) \leq \frac{3K_{\mathcal{A}}(P) + D + O(1)}{2}, \tag{7}$$

$$Q(x) \geq M(x) \cdot 2^{-3K_{\mathcal{A}}(P)-D}$$

and hence

$$d(x|Q) \leq 3K_{\mathcal{A}}(P) + D. \tag{8}$$

Let  $Q_1, \dots, Q_m$  be different and the best distribution for prefixes  $x_1, \dots, x_m$  of  $\omega$ . We need to prove that  $m = \text{poly}(D + K_{\mathcal{A}}(P))$ .

Fix some natural  $a$  and  $b$ . Consider all  $Q_i$  such that  $K(Q_i) = a$  and

$$b \leq d(x_i|Q_i) < 2b.$$

It is enough to we prove that there are only  $\text{poly}(D + K(P))$  best distributions with the such complexity and randomness deficiency. Indeed, the honest estimation of  $m$  will be multiplied by  $\text{poly}(D + K(P))$  because these parameters are bounded by polynomials by (7) and (8). So, further we consider only  $Q_i$  with such parameters.

Let  $x_i$  be the shortest prefix among  $x_1, \dots, x_m$ .

By Lemma 5 every  $Q_j$  is “rather good” distribution for  $x_i$ :  $d(x_i|Q_j) \leq 2b + O(\log b)$  and hence  $Q_j(x_i) \geq Q_i(x_i) \cdot 2^{-O(\log b)}$ . By Lemma 4 there exists a distribution from  $R \in \mathcal{A}$  such that

$$K_{\mathcal{A}}(R) \leq a - \log m + O(\log a + \log m) \text{ and}$$

$$R(x) \geq Q_i(x) \cdot 2^{-\log m - O(\log b)}.$$

Since  $Q_j$  is not worse distribution then  $R$  for  $x$  we have:

$$3 \cdot K_{\mathcal{A}}(Q_i) - \log Q_i(x) \leq 3 \cdot K_{\mathcal{A}}(R) - \log R(x).$$

Therefore:

$$3a \leq 3a - 2 \log m + O(\log b) \text{ and hence}$$

$$\log m \leq O(\log b) = O(\log(K_{\mathcal{A}}(P) + D)).$$

That is proved our claim.

**Part 2.** To complete the proof we do the same things as in the proof of Theorem 7.

If  $x = x_1 \dots x_n$  is a prefix of  $\omega$  and  $Q$  is the best distribution for  $x$  then by Lemma 1

$$\sum_{i=1}^{n-1} (Q(0|x_1 \dots x_i) - P(0|x_1 \dots x_i))^2 = O(K_{\mathcal{A}}(P) + D + K(P, Q)) = O(K_{\mathcal{A}}(P) + D).$$

(In the last equation we use  $K(P, Q) = O(K(P) + K(Q)) = O(K_{\mathcal{A}}(P) + K_{\mathcal{A}}(Q))$ .) So, every probability distribution that is the best for some  $x$  “contributes”  $O(K_{\mathcal{A}}(P) + D)$  in the sum  $\sum_{x \text{ is a prefix of } \omega} (H_{\mathcal{A}}(0|x) - P(0|x))^2$ . There are  $\text{poly}(D + K_{\mathcal{A}}(P))$  such distributions, so we obtain the required estimation.  $\square$

**Proof of Theorem 10** The proof is the same as the proof of Theorem 3 but with using Theorem 9 instead of Theorem 7.  $\square$

## 5 Open Questions

A natural question arises: can we get a better estimation in the last theorem than  $O(K(P)2^{\frac{3K(P)}{2}})$ ? The exponential (in  $K(P)$ ) estimation arises from our attempt to estimate the number of distributions that are optimal for some  $x$ . However, the author is not aware of an example of  $P$ -random sequence  $\omega$  such that there are exponentially many (in terms of  $K(P)$  and  $d(\omega|P)$ ) different best distributions for prefixes of  $\omega$ .

Algorithmic statistics [1, 11, 14] studies good distributions for strings among distributions on finite sets. There exists a family of “standard statistics” that cover all the best distributions for finite strings. It is interesting: are these the same for distributions on the binary tree?

**Acknowledgements** I would like to thank Alexander Shen and Nikolay Vereshchagin discussions, advice and remarks. I am thankful to the anonymous referees for their thorough reading and numerous suggestions for improving the paper.

**Funding** Open access funding provided by FCTIFCCN (b-on). Funded by the European Union (ERC, HOFGA, 101041696). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. Also supported by FCT through the LASIGE Research Unit, ref. UIDB/00408/2020 and ref. UIDP/00408/2020.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

1. Gács, P., Tromp, J., Vitányi, P.M.B.: Algorithmic statistics. *IEEE Trans. Inf. Theory* **47**(6), 2443–2463 (2001)
2. Hutter, M., Poland, J.: Asymptotics of Discrete MDL for Online Prediction. *IEEE Trans. Inf. Theory* **51**(11), 3780–3795 (2005)
3. Hutter, M.: Discrete MDL Predicts in Total Variation Advances in Neural Information Processing Systems 22 (NIPS–2009) 817–825
4. Hutter, M.: Sequential predictions based on algorithmic complexity. *J. Comput. Syst. Sci.* **72**, 95–117 (2006)

5. Hutter, M., Muchnik, A.: Universal convergence of semimeasures on individual random sequences. In: Ben-David, S., Case, J., Maruoka, A. (eds.) ALT 2004. LNCS (LNAI), vol. 3244, pp. 234–248. Springer, Heidelberg (2004)
6. Hutter, M., Muchnik, A.: On semimeasures predicting Martin-Löf random sequences. In *Theoretical Computer Science Volume 382, Issue 3*, 6 September 2007, Pages 247–261. <https://www.sciencedirect.com/science/article/pii/S0304397507002393>
7. Li M., Vitányi, P.: An Introduction to Kolmogorov complexity and its applications, 3rd ed., Springer, 2008 (1 ed., 1993; 2 ed., 1997), xxiii+790 pp. ISBN 978-0-387-49820-1
8. Lattimore, T., Hutter, M.: On Martin-Löf Convergence of Solomonoff’s Mixture. In: Chan TH.H., Lau L.C., Trevisan L. (eds.) *Theory and Applications of Models of Computation. TAMC 2013. Lecture Notes in Computer Science*, vol 7876. Springer, Berlin, Heidelberg (2013)
9. Milovanov, A.: Algorithmic Statistics and Prediction for Polynomial Time-Bounded Algorithms, in: *Sailing Routes in the World of Computation*. Springer, 2018 p. 287–296. [https://link.springer.com/chapter/10.1007/978-3-319-94418-0\\_29](https://link.springer.com/chapter/10.1007/978-3-319-94418-0_29)
10. Milovanov, A.: Predictions and Algorithmic Statistics for Infinite Sequences. In: *Computer Science - Theory and Applications: 16th International Computer Science Symposium in Russia (CSR 2021) Proceedings*, pp. 283–295. Springer, (2021)
11. Shen, A., Uspensky, V., Vereshchagin, N.: *Kolmogorov Complexity and Algorithmic Randomness*, ACM, (2017)
12. Solomonoff, R.J.: A formal theory of inductive inference: Parts 1 and 2. *Information and Control*, 7(1)–22, 224–254 (1964)
13. Solomonoff, R.J.: Complexity-based induction systems: Comparisons and convergence theorems. *IEEE Transactions on Information Theory*, IT-(24), 422–432, (1978)
14. Vereshchagin, N.,K., Shen, A.: *Algorithmic Statistics: Forty Years Later*. *Computability and Complexity*: 669-737 (2017)
15. Vovk, V.G.: On a criterion for randomness. *Dokl. Akad. Nauk SSSR* **294**(6), 1298–1302 (1987)

**Publisher’s Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.