Predictions and algorithmic statistics for infinite sequences

Alexey Milovanov
National Research University Higher School of Economics
Moscow Institute of Physics and Technology
almas239@gmail.com

December 1, 2021

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.

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 $x = x_1 \ldots x_n$ 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 bit by
bit and may hang forever, so the output may be some finite string $x$; the probability of this event is $S(x) - S(x_0) - S(x_1)$. 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 [6, 3, 5], and in the sequel we denote by $M(x)$ one of the universal semi-measures.

Solomonoff suggested to use the ratio $M(1| x) := M(x_1)/M(x)$ to predict $P(1| x)$ for an unknown computable measure $P$. He proved the following bound for the prediction errors.

**Theorem 1** ([7]). For every computable distribution $P$ and for every $b \in \{0, 1\}$ the following sum over all binary strings is finite:

$$\sum_x P(x)[P(b|x) - M(b|x)]^2 < \infty.$$  

Moreover, this sum is bounded by $O(K(P))$, where $K(P)$ is the prefix complexity of the computable measure $P$.

Note that for a semimeasure the probabilities for 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 proof of this result goes as follows. We use that $P \leq cM$ to prove the bound $O(\log c)$; this is the only information we need about $M$. We can assume without loss of generality that $M$ is a measure (by allocating all the undefined cases to the output bit $1-b$). Then we consider the Kullback–Leibler divergence $d(P_k, M_k)$ between distributions on $k$-bit strings. It is at most $\log c$ for all $k$. On the other hand, when $k$ increases by 1, this divergence increases by the expected divergence between predicted distributions on the last bit, and this divergence due to Pinsker inequality has a lower bound that gives the required inequality.

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

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

This expectation is finite, therefore for $P$-almost all $\omega$ the value $D(\omega)$ is finite and, therefore, $P(b|x) - M(b|x) \to 0$.

---

1 One may even require that the probabilities of finite outputs, i.e., the differences $S(x) - S(x_0) - S(x_1)$ are maximal, but we do not need this.
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. The following example constructed by Hutter and An. Muchnik shows that this is indeed the case.

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

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

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

Indeed, let $P$ be the uniform distribution in $[0,1]$ (identified with the Cantor space). Consider the lower-semicomputable random real $\alpha \in [0,1]$ (say, the Chaitin’s number) and the restriction $Q$ of the uniform measure on $[0,1]$ to the interval $(0, \alpha)$. Since $\alpha$ is lower semicomputable, $Q$ is a lower semicomputable semimeasure. If we use $Q$ alone for prediction, all the zero bits of $\alpha$ are predicted with certainty (i.e., the ratio $Q(x_0)/Q(x)$ is 1). Still the $P$-probability of all the bits is $1/2$. So the $Q$-prediction is bad. The semimeasure $Q$ is not universal, but if it is included in the universal mix with weight close to 1, the prediction still will be bad. (Note that the rest of the mix will differ from $P$ on the prefixes of $\omega$ at most by a constant factor, since $\omega$ is $P$-random, so we can have the weight close enough to 1 and get bad predictions.)

Hutter and Lattimore generalized Theorem 2 by proving the same statement for a wide class of universal semi-measures [4].

Trying to overcome this problem and get a good prediction for all Martin-Löf random sequences, we suggest a new 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 the type considered in Theorem 3 is finite, 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 preference to simple hypotheses; the factor 2 is needed for technical reasons (in fact, any constant greater than 1 will work).

For a given $x$ we select the best explanation (that makes this quantity minimal). Then we predict the probability that the next bit after $x$ is $b$:

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

where $Q$ is the best explanation for string $x$ (or one of the best explanations if there are several).

In this paper we prove two main results:

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

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

Moreover, this sum is bounded by $O(K(P)^2 1.5K(P))$, where $K(P)$ is the prefix complexity of the computable measure $P$.

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

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

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.

Also we prove (see Theorem 7) that for every Martin-Löf random sequence the sum

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

is finite (thus improving Theorem 4).

In the next section we prove Theorem 4 (easier of the two). Then (Section 3) we prove Theorem 3 and the improved version of Theorem 4 mentioned above; finally, some open questions are mentioned.
2 Prediction on Martin-Löf random sequences

Recall the Schnorr–Levin theorem (see, e.g., [5, chapter 5] for the proof) 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 using the logarithmic scale. Let us denote by $KA(x)$ 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 $KA$). We have

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

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

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

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

Let us call

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

the randomness deficiency of a string $x$ with respect to a computable measure $P$. (There are several notions of randomness 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 (in one of the versions) characterizes Martin-Löf randomness in terms of this 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 $d(x|P) \to \infty$ as the length of a prefix $x$ of $\omega$ increases.

Note the dichotomy: the values $d_P(x)$ 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

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

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

Let us also recall the following result of Vovk that will be used to prove Theorem 4.
**Theorem 6** (P). Let $P$ and $Q$ be two computable distributions. Let $\omega$ be a sequence that is Martin-Löf random 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.

This theorem says that if a sequence $\omega$ is random with respect to two measures, then the predictions provided by these two measures are close to each other. It is a consequence of Theorem 4 since the latter provides some prediction method that does not depend on $P$ at all, and gives predictions close to $P$. In this section we use Vovk’s result as a black box to prove Theorem 4; in the next section we revisit the proof of Vovk’s result and see how the inequality it provides can be used to prove Theorem 3.

**Proof of Theorem 2**. Assume that a sequence $\omega$ 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 equals $KA(x) + O_P(1)$ where constant in $O_P$ depends on $P$ (so it can absorb $K(P)$) but not on $x$. On the other hand, the inequality $KA(x) \leq K(Q) - \log Q(x) + O(1)$ implies that

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

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

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

if $x$ is a prefix of $\omega$ of increasing length. So, for a bad $Q$, the difference between two sides of $(*)$ 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 3. \qed
On the sum of squares of errors

In this section we prove Theorem 3. For that, we revisit the proof of Vovk’s result (Theorem 6) and get an explicit bound for the differences of the prediction probabilities according to two measures:

Lemma 1. Let $P$ and $Q$ be computable distributions. Assume that $P(x), Q(x) \geq M(x)/C$ for some string $x = x_1 \ldots x_n$ and some $C > 0$; here $M(x)$ is the universal semimeasure. Then

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

Note that in this bound we can replace $x_i$ by 0 since both $P$ and $Q$ are measures and the difference in predicted probabilities is the same for 0 and 1.

This bound implies Theorem 6: if $\omega$ is Martin-Löf random with respect to two computable measures $P$ and $Q$, then due to Schnorr–Levin theorem the ratios $P(x)/M(x)$ and $Q(x)/M(x)$ are bounded by some constant $C$ for all prefixes $x$ of $\omega$. Then we use Lemma 1 to conclude that $\sum_x [P(0|x) - Q(0|x)]^2$ over all prefixes $x$ of $\omega$ is finite.

Proof of Lemma 1: Let

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

Note that

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

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

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

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

Probability distribution $R$ is computable and $K(R) \leq K(P, Q) + O(1)$. Hence,

$$R(x) \leq 2^{K(P, Q)} M(x) \leq 2^{K(P, Q)} \cdot C \cdot P(x).$$

The similar inequality holds for distribution $Q$. Therefore,

$$r_1 r_2 \ldots r_n \leq C \cdot 2^{K(P, Q)} \cdot p_1 p_2 \ldots p_n,$$
and
\[ r_1 r_2 \cdots r_n \leq C \cdot 2^{K(P, Q)} \cdot q_1 q_2 \cdots q_n. \]

Multiplying these two bounds, we see that
\[ \frac{p_1 + q_1}{2} \cdot \cdots \cdot \frac{p_n + q_n}{2} \leq C^2 \cdot 2^{2K(P, Q)} \cdot p_1 \cdots p_n \cdot q_1 \cdots q_n. \]

These two inequality 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 \( h \in (0, 1) \), i.e., to prove that
\[ - \frac{\ln(1 - h) + \ln(1 + h)}{2} \geq \frac{1}{2} h^2; \]
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 \( p_1, \ldots, p_n, q_1, \ldots, q_n \in (0, 1] \) we have
\[ \left( \frac{p_1 + q_1}{2} \cdots \frac{p_n + q_n}{2} \right)^2 \leq c p_1 \cdots p_n q_1 \cdots q_n, \]

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. \( \square \)
To complete the proof of Lemma 1 it remains to let $c = C^2 \cdot 2^{2K(P,Q)}$ and apply Lemma 3.

Now we are ready to prove the following result (that immediately implies Theorem 4).

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

Then

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

**Proof.** Assume that distribution $Q$ is the best explanation for some $x = x_1 \ldots x_n$.

Then

$$3K(Q) - \log Q(x) \leq 3K(P) - \log P(x). \quad (*)$$

Since $d(\omega|P) = D$, we have

$$- \log P(x) \leq KA(x) + D.$$

Therefore (note that $K(Q) \geq 0$) we have

$$- \log Q(x) \leq 3K(P) - \log P(x) \leq 3K(P) + KA(x) + D,$$

so we have

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

We want to estimate

$$\sum_{i=1}^{n} (Q(0|x_1 \ldots x_i) - P(0|x_1 \ldots x_i))^2$$

using Lemma 1. As we have seen, we may use this lemma for $C = 2^{-3K(P) - D}$, but then we will need an upper bound for $K(Q)$. This bound can be obtained as follows.

From (*) we get that

$$2K(Q) + [K(Q) - \log Q(x)] \leq 3K(P) - \log P(x),$$

so

$$2K(Q) + KA(x) \leq 3K(P) - \log P(x) + O(1),$$

or

$$2K(Q) \leq 3K(P) + d(x|P) + O(1).$$
Now Lemma 1 gives
\[
\sum_{i=1}^{n-1} [Q(0|x_1 \ldots x_i) - P(0|x_1 \ldots x_i)]^2 = O(K(P) + D),
\]
since \(\log C\) and \(K(P, Q)\) are \(O(K(P) + D)\) and \(d(x|P) \leq d(\omega | P) = D\).

This bound is valid for every measure \(Q\) that was used at least once for prediction on a prefix of \(\omega\), and provide the bound for the square of prediction errors that would appear if \(Q\) were used for prediction on all the previous steps (while in fact other measures could be optimal at these steps). For use only the cases when \(Q\) was used for prediction, are important.

Then we add all these bounds for all measures \(Q\), there are at most \(2^{3K(P) + O(1)}\) of them, due to the bound for \(K(Q)\) we have proven. In this way we cover all prediction errors, except for the last prediction for each measure \(Q\) (since the sum in Lemma 1 has the upper limit \(n - 1\)). This does not matter, since it gives additional term that is at most 1 for every \(Q\).

In total, we get
\[
\sum_{x \text{ is a prefix of } \omega} (H(0|x) - P(0|x))^2 = O(K(P) + D) \cdot 2^{3K(P) + O(1)}.
\]
as required.

Now, having a bound for every \(P\)-random sequence in terms of its randomness deficiency, we get a bound for the mathematical expectation (integral); recall that
\[
\sum_x P(x) [P(0|x) - H(0|x)]^2 = \int_{(\Omega, P)} \sum_x [H(0|x) - P(0|x)]^2.
\]
For this integration, the set of non-random sequences can be ignored since it has measure 0 and does not matter for the expectation. The set of random sequences where the randomness deficiency is between \(d\) and \(d + 1\) (for some integer \(d\)) has measure at most \(2^{-d}\) (since \(M(x)\) exceeds \(P(x)\) at least by factor \(2^d\)), and the function that we integrate is bounded by
\[
O(K(P) + d) \cdot 2^{3K(P) + d + O(1)}.
\]
So the integral does not exceed
\[
\sum_d 2^{-d} \cdot O(K(P) + d) \cdot 2^{3K(P) + d + O(1)} = O(1) \cdot \left[ K(P) 2^{1.5K(P)} \sum_d 2^{-d/2} + 2^{1.5K(P)} \sum_d d 2^{-d/2} \right] = O(K(P) 2^{1.5K(P)}),
\]
and is finite. This finishes the proof of Theorem 3.
4 Open questions

A natural question arises: can we get a better bound in Theorem 7 (and therefore Theorem 3) than $O(K(P)^2 K(P)^2)$? We had an exponential (in $K(P)$) bound in Theorem 7 because the bound for the number of distributions used for predictions after different prefixes of $\omega$ was exponential. However, the author does not know an example of $P$-random sequence $\omega$ such that there are exponentially many (in terms of $K(P)$ and $d(\omega|P)$) different best explanation for prefixes of $\omega$. So, there is a hope that there are only $\text{poly}(K(P) + d(\omega|P))$ such distributions, and this would give a better bound in Theorem 7 (and Theorem 3).

Algorithmic statistics [1, 8, 5] studies good descriptions (or models) for individual strings; such a model is a probability distribution on strings (as isolated discrete objects). One can also take into account the tree structure on the set of strings and consider measures on the infinite binary tree as statistical models. It would be interesting to develop such a version of algorithmic statistics. In particular, one may look for the counterpart of “standard models”, a restricted class of models that are enough to provide models with parameters close to optimal ones.

Acknowledgments

I would like to thank Alexander Shen and Nikolay Vereshchagin and for useful discussions, advice and remarks. This work is supported by 19-01-00563 RFBR grant and by RaCAF ANR-15-CE40-0016-01 grant.

References

[1] Gács, P., Tromp, J., Vitányi, P.M.B., Algorithmic statistics, IEEE Transactions on Information Theory, 47(6), 2443–2463 (2001)

[2] Hutter, M., Muchnik, An., Universal convergence of semimeasures on individual random sequences. In: Ben-David, S., Case, J., Maruoka, A. (eds.), Proceeding of ALT 2004. Lecture Notes in Computer Science, vol. 3244, pp. 234–248. Springer (2004)

[3] 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.

[4] 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, pp. 212–223. Springer (2013)

[5] Shen, A., Uspensky V. and Vereshchagin N., *Kolmogorov Complexity and Algorithmic Randomness*, AMS (2017)

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

[7] Solomonoff R. J., Complexity-based induction systems: Comparisons and convergence theorems. *IEEE Transactions on Information Theory*, IT-24:422–432 (1978)

[8] Vereshchagin N., Shen A., Algorithmic Statistics: Forty Years Later. *Computability and Complexity. Essays Dedicated to Rodney G. Downey on the Occasion of His 60th Birthday*. Lecture Notes in Computer Science, v. 10010: 669–737, Springer (2017)

[9] Vovk V. G., On a randomness criterion. *Soviet Mathematics Doklady*, 35(3):656–660 (1987). See also: [vovk.net/criterion.pdf](vovk.net/criterion.pdf)