Predicting Non-Stationary Processes

Daniil Ryabko[a*] Marcus Hutter[b]
[a]IDSIA, Galleria 2, CH-6928 Manno, Switzerland, daniil@ryabko.net
[b] RSISE@ANU and SML@NICTA, Canberra, Australia, marcus@hutter1.net

May 2008

Abstract

[1] Suppose we are given two probability measures on the set of one-way infinite finite-alphabet sequences. Consider the question when one of the measures predicts the other, that is, when conditional probabilities converge (in a certain sense), if one of the measures is chosen to generate the sequence. This question may be considered a refinement of the problem of sequence prediction in its most general formulation: for a given class of probability measures, does there exist a measure which predicts all of the measures in the class? To address this problem, we find some conditions on local absolute continuity which are sufficient for prediction and generalize several different notions that are known to be sufficient for prediction. We also formulate some open questions to outline a direction for finding the conditions on classes of measures for which prediction is possible.

Keywords

Sequence prediction, local absolute continuity, non-stationary measures, average/expected criteria, absolute/KL divergence, mixtures of measures.

*This research was supported by the Swiss NSF grants 200020-107616 and 200021-113364.

[1]Parts of the results were reported at Dagstuhl Seminar on Combinatorial and Algorithmic Foundations of Pattern and Association Discovery, Dagstuhl, Germany, 2006, see [1].

1 Introduction

Let a sequence x_t, $t \in \mathbb{N}$ of letters from some finite alphabet $\mathcal{X}$ be generated by some probability measure μ on $\mathcal{X}^\infty$. Having observed the first n letters we want to predict what is the probability of the next letter being x, for each $x \in \mathcal{X}$. This task is motivated by numerous applications — from weather forecasting and stock market prediction to source coding and data compression.

If the measure μ is known then the best forecasts one can make for the $(n+1)$st outcome is μ-conditional probabilities of x_{n+1} being $x \in \mathcal{X}$ given $x_1, \dots, x_n$. However, it is clear that if nothing is known about the distribution μ then no prediction is possible, since for any predictor there is a measure on which it errs (gives grossly wrong probability forecasts) on every step. Thus one has to restrict the attention to some class of measures. Laplace was perhaps the first to address the question of sequence prediction, asking the question of what is the probability that the Sun will rise tomorrow given that it has risen every day for 5000 years. He suggested to assume that the probability that the Sun rises is the same every day and the trials are independent of each other. Thus Laplace considered the task of sequence prediction when the true generating measure belongs to the family of Bernoulli i.i.d. measures with binary alphabet $\mathcal{X} = \{0, 1\}$. The predicting measure hi suggested was $\rho_L(x_{n+1} = 1|x_1, \dots, x_n) = \frac{k+1}{n+2}$ where k is the number of 1s in $x_1, \dots, x_n$. The conditional probabilities of ρ_L converge to the true conditional probabilities μ-a.s. under any Bernoulli i.i.d measure μ. This approach generalizes to the problem of predicting any finite-memory (e.g. Markovian) measure. Moreover, in [9] a measure ρ_R was constructed for predicting an arbitrary stationary measure. The conditional probabilities of ρ_R converge to the true ones *on average*, where the average is taken over time steps μ-a.s. for any stationary measure μ. However, as it was shown in the same work, there is no measure for which conditional probabilities converge to the true ones μ-a.s. for every stationary μ. Thus already for the problem of predicting outcomes of a stationary measure two criteria of prediction arise: prediction in the average (or in Cesaro sense) and prediction on each step, and the solution exists only for the former problem.

What if the measure generating the sequence is not stationary? Another possible assumption is that the measure μ generating the sequence is computable. Solomonoff [11, Eq.(13)] suggested a measure ξ for predicting any computable probability measure. Observe that the class of all computable probability measures is countable; denote it by $(\nu_i)_{i \in \mathbb{N}}$. A Bayesian predictor ξ for such a class is given by $\xi(A) = \sum_{i=1}^\infty w_i \nu_i(A)$ for any measurable set A, where the weights w_i are positive and sum to one[2]. It was shown in [12] that ξ-conditional probabilities converge to μ-conditional probabilities almost surely for any computable measure μ. In fact this is a special case of a more general (though without convergence rate) result of

[2]It is not necessary for prediction that the weights sum to one. In [12] and [13] $w_i = 2^{-K(i)}$ where K stands for the prefix Kolmogorov complexity, and so the weights do not sum to 1. Further, the ν and ξ are only semi-measures.

Blackwell and Dubins [2]: if a measure μ is absolutely continuous w.r. to a measure ρ then the conditional measure ρ given $x_1, \dots, x_n$ converges to μ given $x_1, \dots, x_n$ in total variation μ-a.s.

Thus the problem of sequence prediction for certain classes of measures was often addressed in the literature. Although the mentioned classes of measures are sufficiently interesting, it is often hard to decide in applications with which assumptions does a problem at hand comply; not to mention such practical issues as that a predicting measure for all computable measures is necessarily non-computable itself. Also the general approach may be easier to extend to the problems of active learning, which is a rather hard problem itself (see e.g. [7]).

In this work we start to address the following **general questions**: For which classes of measures is sequence prediction possible? Under which conditions does a measure ρ predict a measure μ?

Extensive as the literature on sequence prediction is, these questions have not been formulated, and so in the general problem posed has not received much attention. One line of research which exhibits this kind of generality consists in extending the result of Blackwell and Dubins mentioned above, which states that if μ is absolutely continuous with respect to ρ, then ρ predicts μ in total variation distance. In [5] a question of whether, given a class of measures $\mathcal{C}$ and a prior ("meta"-measure) λ over this class of measures, the conditional probabilities of a Bayesian mixture of the class $\mathcal{C}$ w.r.t. λ converge to the true μ-probabilities (weakly merge, in terminology of [5]) for λ-almost any measure μ in $\mathcal{C}$. This question can be considered solved, since the authors provide necessary and sufficient conditions on the measure given by the mixture of the class $\mathcal{C}$ w.r.t. λ under which prediction is possible. The major difference from the general questions we posed above is that we do not wish to assume that we have a measure on our class of measures. For large (non-parametric) classes of measures it may not be intuitive which measure over it is natural; rather, the question is whether a "natural" measure which can be used for prediction exists.

We start with the following observation. For a Bayesian mixture ξ of a countable class of measures ν_i, $i \in \mathbb{N}$, we have $\xi(A) \geq w_i \nu_i(A)$ for any i and any measurable set A, where w_i is a constant. This condition is stronger than the assumption of absolute continuity and is sufficient for prediction in a very strong sense (in total variation). Since we are willing to be satisfied with prediction in a weaker sense (e.g. convergence of conditional probabilities), we make a weaker assumption: Say that *a measure ρ dominates a measure μ with coefficients $c_n > 0$* if

$$\rho(x_1, \dots, x_n) \;\geq\; c_n \mu(x_1, \dots, x_n) \tag{1}$$

for all $x_1, \dots, x_n$.

The concrete question we pose is, under what conditions on c_n does (1) imply that ρ predicts μ? Observe that if $\rho(x_1, \dots, x_n) > 0$ for any $x_1, \dots, x_n$ then any measure μ is *locally* absolutely continuous with respect to ρ, and moreover, for any measure μ some constants c_n can be found that satisfy (1). For example, if ρ is Bernoulli i.i.d. measure with parameter $\frac{1}{2}$ and μ is any other measure, then (1) is

(trivially) satisfied with $c_n = 2^{-n}$. Thus if $c_n \equiv c$ then ρ predicts μ in a very strong sense, whereas exponentially decreasing c_n are not enough for prediction. We will show that dominance with any subexponentially decreasing coefficients is sufficient for prediction, in a weak sense of convergence of expected averages. Dominance with any polynomially decreasing coefficients (and some others),is sufficient for (almost sure) prediction on time-average. However, for prediction on every step we have a negative result: for any dominance coefficients that go to zero there exists a pair of measures ρ and μ which satisfy (1) but ρ does not predict μ in this sense. Thus the situation is similar to that for predicting any stationary measure: prediction is possible in the average but not on every step.

Note also that for Laplace's measure ρ_L it can be shown that ρ_L dominates any i.i.d. measure μ with linearly decreasing coefficients $c_n = \frac{1}{n+1}$. Thus dominance with decreasing coefficients generalizes (in a sense) predicting countable classes of measures (where we have dominance with a constant), absolute continuity (via local absolute continuity), and predicting i.i.d. and finite-memory measures.

2 Main results

We consider processes on the set of one-way infinite sequences $\mathcal{X}^\infty$ where $\mathcal{X}$ is a finite set (alphabet). We use $x_{1:n}$ for $x_1, \dots, x_n$ and $x_{<n}$ for $x_1, \dots, x_{n-1}$, $x_t \in \mathcal{X}$. The symbol μ is reserved for the "true" measure generating examples. The symbol $\mathbf{E}_\nu$ stands for expectation with respect to a measure ν and $\mathbf{E}$ is for $\mathbf{E}_\mu$ (expectation with respect to the "true" measure).

For two measures μ and ρ define the following measures of divergence.

- (d) Kullback-Leibler (KL) divergence $d_n(\mu, \rho|x_{<n}) = \sum_{x\in\mathcal{X}} \mu(x_n = x|x_{<n}) \log \frac{\mu(x_n=x|x_{<n})}{\rho(x_n=x|x_{<n})}$,
- ($\bar{d}$) average KL divergence $\bar{d}_n(\mu, \rho|x_{1:n}) = \frac{1}{n}\sum_{t=1}^n d_t(\mu, \rho|x_{<t})$,
- (a) absolute distance $a_n(\mu, \rho|x_{<n}) = \sum_{x\in\mathcal{X}} |\mu(x_n = x|x_{<n}) - \rho(x_n = x|x_{<n})|$,
- ($\bar{a}$) average absolute distance $\bar{a}_n(\mu, \rho|x_{1:n}) = \frac{1}{n}\sum_{t=1}^n a_t(\mu, \rho|x_{<t})$.

The argument $x_{1:n}$ will be often omitted. The following implications hold (and are complete):

$$\begin{array}{ccccc} d & \Rightarrow & \bar{d} & & \mathbf{E}\bar{d} \\ \Downarrow & & \Downarrow & & \Downarrow \\ a & \Rightarrow & \bar{a} & \Rightarrow & \mathbf{E}\bar{a} \end{array}$$

to be understood as e.g.: if $\bar{d}_n \to 0$ a.s. then $\bar{a}_n \to 0$ a.s, or, if $\mathbf{E}\bar{d}_n \to 0$ then $\mathbf{E}\bar{a}_n \to 0$. The horizontal implications $\Rightarrow$ follow immediately from the definitions, and the $\Downarrow$ follow from the following Lemma:

Lemma 1 ($\boldsymbol{a^2 \le 2d}$)**.** *For all measures ρ and μ and sequences $x_{1:\infty}$ we have: $a_t^2 \le 2d_t$ and $\bar{a}_n^2 \le 2\bar{d}_n$ and $(\mathbf{E}\bar{a}_n)^2 \le 2\mathbf{E}\bar{d}_n$.*

Proof. Pinsker's inequality [3, Lem.3.11a] implies $a_t^2 \le 2d_t$. Using this and Jensen's inequality for the average $\frac{1}{n}\sum_{t=1}^n[...]$ we get

$$2\bar{d}_n \;=\; \frac{1}{n}\sum_{t=1}^n 2d_t \;\ge\; \frac{1}{n}\sum_{t=1}^n a_t^2 \;\ge\; \left(\frac{1}{n}\sum_{t=1}^n a_t\right)^2 \;=\; \bar{a}_n^2$$

Using this and Jensen's inequality for the expectation $\mathbf{E}$ we get $2\mathbf{E}\bar{d}_n \ge \mathbf{E}\bar{a}_n^2 \ge (\mathbf{E}\bar{a}_n)^2$. □

The main concept we introduce is the following.

Definition 2. *We say that a measure ρ dominates a measure μ with coefficients $c_n > 0$ iff $\rho(x_{1:n}) \;\ge\; c_n\mu(x_{1:n})$. for all $x_{1:n}$.*

Suppose that ρ dominates μ with decreasing coefficients c_n. Does ρ predict μ in (expected, expected average) KL divergence (absolute distance)? First let us give an example.

Proposition 3. *Let ρ_L be the Laplace measure $\rho_L(x_{n+1}=a|x_{1:n}) = \frac{k+1}{n+|\mathcal{X}|}$ for any $a\in\mathcal{X}$ and any $x_{1:n}\in\mathcal{X}^n$, where k is the number of occurrences of a in $x_{1:n}$. Then $\rho_L(x_{1:n}) \;\ge\; \frac{n!}{(n+|\mathcal{X}|-1)!}\,\mu(x_{1:n})$ for any Bernoulli i.i.d. μ. This bound is sharp.*

The proof is only technical and can be found in [8]. Thus for ρ_L and binary $\mathcal{X}$ we have $c_n = \mathcal{O}(\frac{1}{n})$. As mentioned above, in general, exponentially decreasing coefficients c_n are not sufficient for prediction. On the other hand, in a weak sense of convergence in expected average KL divergence (or absolute distance) the property (1) with subexponentially decreasing c_n is sufficient. We also remind that if c_n are bounded from below then prediction in the strong sense of total variation is possible.

Theorem 4. *Let μ and ρ be two measures on $\mathcal{X}^\infty$ and suppose that $\rho(x_{1:n}) \ge c_n\mu(x_{1:n})$ for any $x_{1:n}$, where c_n are positive constants satisfying $\frac{1}{n}\log c_n^{-1} \to 0$. Then ρ predicts μ in expected average KL divergence $\mathbf{E}_\mu\bar{d}_n(\mu,\rho)\to 0$ and in expected average absolute distance $\mathbf{E}_\mu\bar{a}_n(\mu,\rho)\to 0$.*

The proof can be found in [8]; it is based on the same idea as the proof of convergence of Solomonoff predictor to any of its summands in [9], see also [3].

With a stronger condition on c_n prediction in average KL divergence can be established.

Theorem 5 ($\bar{\boldsymbol{d}}\boldsymbol{\to}\mathbf{0}$ and $\bar{\boldsymbol{a}}\boldsymbol{\to}\mathbf{0}$). *Let μ and ρ be two measures on $\mathcal{X}^\infty$ and suppose that $\rho(x_{1:n}) \ge c_n\mu(x_{1:n})$ for every $x_{1:n}$, where c_n are positive constants satisfying*

$$\sum_{n=1}^\infty \frac{(\log c_n^{-1})^2}{n^2} \;<\; \infty. \tag{2}$$

Then ρ predicts μ in average KL divergence $\bar{d}_n(\mu,\rho)\to 0$ μ-a.s. and in average absolute distance $\bar{a}_n(\mu,\rho)\to 0$ μ-a.s.

In particular, the condition (2) on the coefficients is satisfied for polynomially decreasing coefficients, or for $c_n = \exp(-\sqrt{n}/\log n)$.

Proof. Again the second statement (about absolute distance) follows from the first one and Lemma 1, so that we only have to prove the statement about KL divergence.

Introduce the symbol $\mathbf{E}^n$ for μ-expectation over x_n conditional on $x_{<n}$. Consider random variables $l_n = \log \frac{\mu(x_n|x_{<n})}{\rho(x_n|x_{<n})}$ and $\bar{l}_n = \frac{1}{n}\sum_{t=1}^n l_t$. Observe that $d_n = \mathbf{E}^n l_n$, so that the random variables $m_n = l_n - d_n$ form a martingale difference sequence (that is, $\mathbf{E}^n m_n = 0$) with respect to the standard filtration defined by $x_1, \ldots, x_n, \ldots$. Let also $\bar{m}_n = \frac{1}{n}\sum_{t=1}^n m_t$. We will show that $\bar{m}_n \to 0$ μ-a.s. and $\bar{l}_n \to 0$ μ-a.s. which implies $\bar{d}_n \to 0$ μ-a.s.

Note that

$$\bar{l}_n \;=\; \frac{1}{n}\log\frac{\mu(x_{1:n})}{\rho(x_{1:n})} \;\leq\; \frac{\log c_n^{-1}}{n} \;\to\; 0.$$

Thus to show that $\bar{l}_n$ goes to 0 we need to bound it from below. It is easy to see that $n\bar{l}_n$ is (μ-a.s.) bounded from below by a constant, since $\frac{\rho(x_{1:n})}{\mu(x_{1:n})}$ is a positive μ-martingale whose expectation is 1, and so it converges to a finite limit μ-a.s. by Doob's submartingale convergence theorem, see e.g. [10, p.508]. Next we will show that $\bar{m}_n \to 0$ μ-a.s. We have

$$\begin{aligned} m_n \;&=\; \log\frac{\mu(x_{1:n})}{\rho(x_{1:n})} - \log\frac{\mu(x_{<n})}{\rho(x_{<n})} - \mathbf{E}^n\log\frac{\mu(x_{1:n})}{\rho(x_{1:n})} + \mathbf{E}^n\log\frac{\mu(x_{<n})}{\rho(x_{<n})} \\ &=\; \log\frac{\mu(x_{1:n})}{\rho(x_{1:n})} - \mathbf{E}^n\log\frac{\mu(x_{1:n})}{\rho(x_{1:n})}. \end{aligned}$$

Let $f(n)$ be some function monotonically increasing to infinity such that

$$\sum_{n=1}^{\infty}\frac{(\log c_n^{-1} + f(n))^2}{n^2} \;<\; \infty. \tag{3}$$

For a sequence of random variables λ_n define

$$(\lambda_n)^{+(f)} \;=\; \begin{cases} \lambda_n & \text{if } \lambda_n \geq -f(n) \\ 0 & \text{otherwise} \end{cases}$$

and $\lambda_n^{-(f)} = \lambda_n - \lambda_n^{+(f)}$. Introduce also $m_n^+ \;=\; \left(\log\frac{\mu(x_{1:n})}{\rho(x_{1:n})}\right)^{+(f)} - \mathbf{E}^n\left(\log\frac{\mu(x_{1:n})}{\rho(x_{1:n})}\right)^{+(f)}$, $m_n^- = m_n - m_n^+$ and the averages $\bar{m}_n^+$ and $\bar{m}_n^-$. Observe that m_n^+ is a martingale difference sequence. Hence to establish the convergence $\bar{m}_n^+ \to 0$ we can use the martingale strong law of large numbers [10, p.501], which states that, for a martingale difference sequence γ_n, if $\mathbf{E}(n\bar{\gamma}_n)^2 < \infty$ and $\sum_{n=1}^{\infty}\mathbf{E}\gamma_n^2/n^2 < \infty$ then $\bar{\gamma}_n \to 0$ a.s. Indeed, for m_n^+ the first condition is trivially satisfied (since the expectation in question is a finite sum of finite numbers), and the second follows from the fact that $|m_n^+| \leq \log c_n^{-1} + f(n)$ and (3).

Furthermore, we have $m_n^- = \left(\log \frac{\mu(x_{1:n})}{\rho(x_{1:n})}\right)^{-(f)} - \mathbf{E}^n \left(\log \frac{\mu(x_{1:n})}{\rho(x_{1:n})}\right)^{-(f)}$. As it was mentioned before, $\log \frac{\mu(x_{1:n})}{\rho(x_{1:n})}$ converges μ-a.s. either to (positive) infinity or to a finite number. Hence $\left(\log \frac{\mu(x_{1:n})}{\rho(x_{1:n})}\right)^{-(f)}$ is non-zero only a finite number of times, and so its average goes to zero. To see that $\mathbf{E}^n \left(\log \frac{\mu(x_{1:n})}{\rho(x_{1:n})}\right)^{-(f)} \to 0$ we write

$$\mathbf{E}^n \left(\log \frac{\mu(x_{1:n})}{\rho(x_{1:n})}\right)^{-(f)} = \sum_{x_n \in \mathcal{X}} \mu(x_n|x_{<n}) \left(\log \frac{\mu(x_{<n})}{\rho(x_{<n})} + \log \frac{\mu(x_n|x_{<n})}{\rho(x_n|x_{<n})}\right)^{-(f)}$$
$$\geq \sum_{x_n \in \mathcal{X}} \mu(x_n|x_{<n}) \left(\log \frac{\mu(x_{<n})}{\rho(x_{<n})} + \log \mu(x_n|x_{<n})\right)^{-(f)}$$

and note that the first term in brackets is bounded from below, and so for the sum in brackets to be less than $-f(n)$ (which is unbounded) the second term $\log \mu(x_n|x_{<n})$ has to go to $-\infty$, but then the expectation goes to zero since $\lim_{u\to 0} u \log u = 0$.

Thus $\bar{m}_n^- \to 0$ μ-a.s., which together with $\bar{m}_n^+ \to 0$ μ-a.s. implies $\bar{m}_n \to 0$ μ-a.s., which, finally, together with $\bar{l}_n \to 0$ μ-a.s. implies $\bar{d}_n \to 0$ μ-a.s. □

However, no form of dominance with decreasing coefficients is sufficient for prediction in absolute distance or KL divergence:

Proposition 6 (***d*** $\not\to$ **0** and ***a*** $\not\to$ **0**)**.** *For each sequence of positive numbers c_n that goes to 0 there exist measures μ and ρ and a number $\epsilon > 0$ such that $\rho(x_{1:n}) \geq c_n \mu(x_{1:n})$ for all $x_{1:n}$, yet $a_n(\mu, \rho|x_{1:n}) > \epsilon$ and $d_n(\mu, \rho|x_{1:n}) > \epsilon$ infinitely often μ-a.s.*

Proof. Let μ be concentrated on the sequence $11111\ldots$ (that is $\mu(x_n = 1) = 1$ for all n), and let $\rho(x_n = 1) = 1$ for all n except for a subsequence of steps $n = n_k$, $k \in \mathbb{N}$ on which $\rho(x_{n_k} = 1) = 1/2$ independently of each other. It is easy to see that choosing n_k sparse enough we can make $\rho(1_1 \ldots 1_n)$ decrease to 0 arbitrary slowly; yet $|\mu(x_{n_k}) - \rho(x_{n_k})| = 1/2$ for all k. □

Following is the table of conditions on dominance coefficients and answers to the questions whether these conditions are sufficient for prediction (coefficients bounded from below are included for the sake of completeness).

	$\mathbf{E}\bar{d}_n$	$\bar{d}_n$	d_n	$\mathbf{E}\bar{a}_n$	$\bar{a}_n$	a_n
$\log c_n^{-1} = o(n)$	+	?	−	+	?	−
$\sum_{n=1}^{\infty} \frac{\log c_n^{-1}}{n^2} < \infty$	+	+	−	+	+	−
$c_n \geq c > 0$	+	+	+	+	+	+

An open question is to find whether $\log c_n^{-1} = o(n)$ is sufficient for prediction in $\bar{d}_n$ or at least in $\bar{a}_n$. Another problem is to find out whether any conditions on dominance

coefficients are necessary for prediction; so far we only have some sufficient conditions. On the one hand, the obtained results suggest that some form of dominance with decreasing coefficients may be necessary for prediction, at least in the sense of convergence of averages. On the other hand, the condition (1) is uniform over all sequences which probably is not necessary for prediction. As for prediction in the sense of almost sure convergence, perhaps more subtle behavior of the ratio $\frac{\mu(x_{1:n})}{\rho(x_{1:n})}$ should be analyzed, since dominance with decreasing coefficients is not sufficient for prediction in this sense.

References

[1] R. Ahlswede, A. Apostolico, V. I. Levenshtein (Editors), Combinatorial and Algorithmic Foundations of Pattern and Association Discovery, Dagstuhl Seminar Proceedings 06201, 2006.

[2] D. Blackwell and L. Dubins. Merging of opinions with increasing information. *Annals of Mathematical Statistics*, 33:882–887, 1962.

[3] M. Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability.* Springer, Berlin, 2005. 300 pages, http://www.idsia.ch/~marcus/ai/uaibook.htm.

[4] M. Hutter. On the foundations of universal sequence prediction. In *Proc. 3rd Annual Conference on Theory and Applications of Models of Computation (TAMC'06)*, volume 3959 of *LNCS*, pages 408–420. Springer, 2006.

[5] M. Jackson, E. Kalai, and R. Smorodinsky. Bayesian representation of stochastic processes under learning: de Finetti revisited. *Econometrica*, 67(4):875–794, 1999.

[6] B. Ryabko and J. Astola. Universal codes as a basis for time series testing. *Statistical Methodology*, To appear, available online, 2006.

[7] D. Ryabko and M. Hutter. Asymptotic learnability of reinforcement problems with arbitrary dependence. In *Proc. The 17th International Conference on Algorithmic Learning Theory*, LNCS vol. 4264, pp. 334–347.

[8] D. Ryabko and M. Hutter. On Sequence Prediction for arbitrary measures. http://arxiv.org/abs/cs.LG/0606077

[9] B. Ryabko. Prediction of random sequences and universal coding. *Problems of Information Transmission*, 24:87–96, 1988.

[10] A. N. Shiryaev. *Probability.* Springer, 1996.

[11] R. J. Solomonoff. A formal theory of inductive inference: Part 1 and 2. *Inform. Control*, 7:1–22, 224–254, 1964.

[12] R. J. Solomonoff. Complexity-based induction systems: comparisons and convergence theorems. *IEEE Trans. Information Theory*, IT-24:422–432, 1978.

[13] A. K. Zvonkin and L. A. Levin. The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms. *Russian Mathematical Surveys*, 25(6):83–124, 1970.