Kolmogorov complexity, statistical regularization of inverse problems, and Birkhoff's formalization of beauty

V. Kreinovich, L. Longpré, and M. Koshelev

Department of Computer Science, University of Texas at El Paso, El Paso, TX 79968, USA

ABSTRACT

Most practical applications of statistical methods are based on the implicit assumption that if an event has a very small probability, then it cannot occur. For example, the probability that a kettle placed on a cold stove would start boiling by itself is not 0, it is positive, but it is so small, that physicists conclude that such an event is simply impossible.

This assumption is difficult to formalize in traditional probability theory, because this theory only describes measures on sets (e.g., for an inverse problem, on the set of all functions) and does not allow us to divide functions into "random" (possible) and non-random ("impossible") ones. This distinction was made possible by the idea of algorithmic randomness, introduced by Kolmogorov and his student Martin-Löf in the 1960s.

We show that this idea can also be used for inverse problems. In particular, we prove that for every probability measure, the corresponding set of random functions is compact, and, therefore, the correspondingly restricted inverse problem is well-defined.

The resulting techniques turns out to be interestingly related with the qualitative esthetic measure introduced by G. Birkhoff as order/complexity.

Keywords: inverse problem, Bayesian regularization, Kolmogorov complexity, Birkhoff's approach to esthetics

1. INTRODUCTION

What is an inverse problem: a brief reminder. In many applied problems (geophysics, medicine, astronomy, etc.), we cannot directly measure the state s of the system in which we are interested; to determine this state, we therefore measure some related characteristics y, and then use the measurement results $\widetilde{y}$ to reconstruct the desired state s. The problem of reconstructing the state s from the measurement results $\widetilde{y}$ is called the *inverse problem*. Let us give two examples:

- We are often interested in the *dynamics* of a system, e.g., in measuring the value $x(t)$ of the desired physical quantity x in different moments of time. If we cannot measure $x(t)$ directly, we measure some related quantity $y(t)$, and then try to reconstruct the desired values $x(t)$. For example, in case the dependency between $x(t)$ and $y(t)$ is linear, we arrive at a problem of reconstructing $x(t)$ from the equation $y(t) = \int k(t,s)x(s)\,ds + n(t)$, where $k(t,s)$ is a (known) function, and $n(t)$ denote the (unknown) errors of measuring $y(t)$.
- Another example of inverse problems is *image reconstruction* from a noisy raw data.

Inverse problems are often ill-posed. Usually, we know how the actual value y of the measured quantities depend on the state s of the system, i.e., we know a mapping $f : S \to Y$ from the set S of all possible states to the set Y of all possible values of y. Since a measurement is never 100% accurate, the actual measurement results $\widetilde{y}$ are (slightly) different from the actual value $y = f(s)$ of the measured quantity y.

Of course, to be able to reconstruct s from y, we must make sure that we are making sufficiently many measurements, so that from $f(s)$, we will be able to reconstruct s uniquely. In mathematical terms, we need the function f to be reversible (1-1). If this function is reversible, then in the ideal case, when the measurements are absolutely accurate (i.e., when $\widetilde{y} = y$), we will be able to reconstruct the state s uniquely, as $s = f^{-1}(y)$.

Due to the inevitable measurement inaccuracy, the measured value $\widetilde{y}$ is, in general, different from $y = f(s)$. Therefore, if we simply apply the inverse function f^{-1} to the measurement result $\widetilde{y}$, we get $\widetilde{s} = f^{-1}(\widetilde{y}) \neq s = f^{-1}(y)$.

E-mails: {vladik,longpre,mkosh}@cs.utep.edu

If the measurement error is large, i.e., if $\widetilde{y}$ is very distant from y, then, of course, the reconstructed state $\widetilde{s}$ may also be very different from the actual state s. However, it seems natural to expect that as the measurements become more and more accurate, i.e., as $\widetilde{y} \to y$, the reconstructed state $\widetilde{s}$ should also get closer and closer to the actual one: $\widetilde{s} \to s$.

To describe this expectation in precise terms, we need to find the metrics d_S and d_Y on the sets S and Y which characterize the closeness of the states or, correspondingly, of the measurement results; in terms of these metrics, the fact that $\widetilde{y}$ gets "closer and closer to y" can be written as $d_Y(\widetilde{y}, y) \to 0$, and the condition that $\widetilde{s} \to s$ means $d_S(\widetilde{s}, s) \to 0$. For example, to describe how close the two signals $x(t)$ and $x'(t)$ are, we may say that they are ε-close (for some real number $\varepsilon > 0$), if for every moment of time t, the difference between the two signals does not exceed ε, i.e., $|x(t) - x'(t)| \leq \varepsilon$. This description can be reformulated as $d_S(x, x') \leq \varepsilon$, where $d_S(x, x') = \sup_t |x(t) - x'(t)|$.

In metric terms, we would like $\widetilde{y} \to y$ to imply $f^{-1}(\widetilde{y}) \to f^{-1}(y)$, i.e., in other words, we would like the inverse function f^{-1} to be continuous. Alas, in many applied problems, the inverse mapping f^{-1} is *not* continuous. As a result, arbitrarily small measurement errors can cause arbitrarily large differences between the actual and reconstructed states. Such problems are called *ill-posed* (see, e.g., Ref. 21).

For example, since all the measurement devices are inertial and thus suppress high frequencies, the functions $x(t)$ and $x(t) + \sin(\omega \cdot t)$, where ω is sufficiently big, lead to almost similar measured values $\widetilde{y}(t)$. Thus, one and the same measurement result $\widetilde{y}(t)$ can correspond to two different states: $x(t)$ and $x(t) + \sin(\omega \cdot t)$.

For ill-posed problems, additional knowledge is needed. The fact that a problem is ill-posed means the following: if the *only* information about the desired state s comes from the measurements, then we cannot reconstruct the state with any accuracy. Hence, to be able to reconstruct the state accurately, we need to have an *additional* (prior) knowledge about the state.

Ideal case: deterministic additional knowledge. In some cases, this knowledge consists of knowing which states from the set S are actually possible, and which are not. For example, we may know that not all signals $x(t)$, $0 \leq t \leq T$, are possible, but only smooth signals for which the signal itself is bounded by some value M (i.e., $|x(t)| \leq M$) and the rate with which the signal changes is bounded by some bound Δ (i.e., $|\dot{x}(t)| \leq \Delta$). For this type of knowledge, we, in effect, restrict possible states to a proper *subset* $K \subseteq S$ of the original set S. Then, instead of the original function $f : S \to Y$, we only have to consider its restriction $f_{|K} : K \to Y$ to the set. If this restriction has a continuous inverse, then the problem is solved – in the sense that the more accurate the measurements, the closer the reconstructed state to the original one.

It is known that if the set K if *compact*, then for any 1-1 continuous function $g : K \to Y$ its inverse is also continuous. It is also known that if a set K is not compact, then for some 1-1 continuous function $g : K \to Y$, its inverse is not continuous. So, one way to guarantee the continuity of the inverse function $f^{-1}_{|K}$ is to require that the set K is compact. For example, the above prior knowledge about the bounds M and Δ characterizes a set K that is compact in the above metric $d_S(x, x') = \sup |x(t) - x'(t)|$.

Alternative situations: probabilistic additional knowledge. In many real-life situations, we do not have enough prior information to restrict S to a compact set. In other words, even when we restrict ourselves to the states that we know to be possible, we still get a non-compact set of states S and an ill-posed inverse problem: we do not have enough prior information to claim that some of the states from the set S are absolutely impossible. In some of such situations, however, we have a (weaker) prior information: namely, for some states $s \in S$, although we cannot claim that s is *impossible*, but we know that s is *not* very *probable*. To be more precise, in addition to the set S of all states, we may know the probabilities of different states from this set. In other words, we know a *probability measure* μ on the set S of all possible states.

How can we combine this probabilistic knowledge with the knowledge coming from measurements? Bayes' formula allows us to use this *prior* probability measure and the measurement results to compute the *posterior* probability measure on the set of all possible states. This probability combines both types of knowledge and describes which states are more probable and which are less probable.

In practical applications, statistical knowledge leads to deterministic restrictions. From a strict statistical viewpoint, if an event has a positive probability (no matter how small), it is possible, rare but still possible. For example, if we know that the measurement error is characterized by Gaussian distribution with 0 average and standard deviation σ, then it is potentially possible to have the measurement error greater than 10σ. If we flip a coin, it is possible to get a sequence of 1,000 heads in a row; the probability of this event is small but still positive

(2^{-1000}). It is also possible for all the randomly moving molecules in a person's body to start moving in one and the same direction so that this person would rise up in the air; it is possible because the probability is positive. Similarly, it is possible that a kettle placed on a cold stove will start boiling by itself, while the stove would get even colder.

From the purely statistical viewpoint, all these events are theoretically possible. However, from the viewpoint of physics and other applications, such events are impossible. In general, most practical applications of statistical methods are based on the implicit assumption that if an event has a very small probability, then it cannot occur. For example, the probability that a kettle placed on a cold stove would start boiling by itself is not 0, it is positive, but it is so small, that physicists conclude that such an event is simply impossible.

In other words, in all these applications, if we know the probabilities of different events or different states, we somehow divide all possible events or states into two categories: the ones which are *possible* (*random*) and the ones which are *impossible* (*not random*).

The problem. The main problem is: how can we formalize this division? This distinction is difficult to formalize in traditional probability theory, because this theory only describes measures on sets (e.g., for an inverse problem, on the set of all functions) and does not allow us to divide functions into "random" (possible) and non-random ("impossible") ones.

In this paper, we show how this division can be described in precise terms, and how the corresponding formalization can help in solving ill-posed problems.

2. TOWARDS A FORMALIZATION OF RANDOMNESS

2.1. Kolmogorov-Martin-Löf formalization

Let us first formalize the simplest case: the (implicit) postulate of impossibility of events with probability 0. Our goal is to formalize a statement that *if an event has a very small probability then it cannot occur*. One clear problem with formalizing this statement is that we do not know what exactly "very small" means. The only case when we are absolutely sure that a probability is indeed "very small" is when this probability is actually equal to 0. Let us, therefore, start with this simpler case.

Many practical applications of probability and statistics use the implicit assumption that events with 0 probability cannot happen. Indeed, why, e.g., do we conclude that when we toss a fair coin many times, the frequency of heads in a sequence of results tends to 1/2? Because there is a theorem according to which this convergence happens with probability 1. Any other statistical law, be it the central limit theorem or special asymptotic properties of outliers, is only true with probability 1. However, when we prove that something is almost always true, we conclude that it is true for the actual ("random") sequence of observations.

First guess does not work. We want to formalize the following postulate: *If an event has probability 0, then it cannot occur.* So, if we have a set S with a probability measure μ, then a "random" element $s \in S$ cannot belong to any set $E \subseteq S$ of measure 0. In other words, it seems reasonable to define a "random" element as an element $s \in S$ that does not belong to any set E of measure 0.

This definition may sound natural at first glance, but it does not work even in the simplest cases. For example, let us consider a random variable which is uniformly distributed on the interval $[0,1]$. In this case, $S = [0,1]$, and μ is the usual (Lebesgue) measure. Whichever point $s \in S = [0,1]$ we take, the corresponding 1-point set $\{s\}$ has probability 0. According to our first-guess definition, a "random" element cannot belong to any of these sets and therefore, cannot be equal to any number at all. In other words, no element $s \in S$ is "random" in the sense of this definition. On the other hand, common sense (that we are trying to formalize) tells us that "random" numbers do exist.

Why did a seemingly natural formalization lead to a contradiction with common sense?

An improved definition: Kolmogorov-Martin-Löf randomness. When we formalized the above postulate – about the impossibility of 0-probability events – we interpreted the word "event" as meaning an arbitrary (measurable) set $E \subseteq S$. In the actual applications, however, we do not consider *arbitrary* sets, we are only interested in the events which are described by precise mathematical formulas, i.e., in the sets which are *defined* in some language (e.g., in the language of set theory). For example, we are interested in the set of all binary sequences for which the frequency of 1's tends to 1/2; we are interested in the set of all sequences for which the sample distribution tends to Gaussian, etc.

This distinction between arbitrary and definable sets may be viewed as nit-picking, but it actually solves the above contradiction. To show it, let us first describe this idea formally.

Definition 1. *Let a language L be fixed (e.g., the language of set theory). A set E is called definable if in the language L, there exists a formula $P(s)$ with one free variable s such that for every s, $s \in S$ if and only if $P(s)$ is true.*

Definition 2. *Let S be a set with a probability measure μ. An element $s \in S$ is called random (in the sense of Kolmogorov-Martin-Löf) if it does not belong to any definable set of measure 0.*

Now, we are ready to show that there exist elements which are random in the sense of this definition. Moreover, we will show an even stronger statement: that almost all elements are random:

Proposition. (Kolmogorov-Martin-Löf) *For every set S, almost all elements $s \in S$ are random.*

Idea of the proof. By Definition 2, an element $s \in S$ is random if and only if it does not belong to any definable set of measure 0. In other words, an element $s \in S$ is random if and only if it does not belong to a *union* E of all definable sets of measure 0. By Definition 1, to every definable set there corresponds a formula from the language L, and different sets correspond to different formulas. Every formula is a finite sequence of symbols in a finite alphabet. There are countably many such sequences, and therefore, there are at most countably many different formulas. Since different definable sets correspond to different formulas, there are no more than countably many definable sets. In particular, there are no more than countably many definable sets of measure 0. Therefore, the union E of all such sets also has measure 0. So, almost all element $s \in S$ do not belong to E and are, therefore, random in the sense of Definition 2. Q.E.D.

Similarly, one can prove that if some definable property is true for *almost all* elements $s \in S$, then it is true for every *random* element $s \in S$.

Relation to Kolmogorov complexity. The above definition was made possible by the idea of algorithmic randomness, introduced by Kolmogorov and his student Martin-Lof in the 1960s. Crudely speaking, *Kolmogorov complexity* $K(x)$ of a finite sequence x is the smallest length of a program that can generate this sequence, and a sequence is called *random* if its Kolmogorov complexity is almost equal to its length (i.e., if the shortest way to generate it is to just print it symbol after symbol; for detailed definitions, see Ref. 18)).

This definition has been very successful in data processing. This idea has lead to a methodology of Minimum Length Description that has been successfully used in data processing (Rissanen, Cheeseman, etc.; for a detailed survey, see Ref. 18).

2.2. Beyond Kolmogorov-Martin-Löf

Kolmogorov-Martin-Löf definition does not fully capture the intuitive meaning of randomness. The above definition captures some aspects of randomness, but not all of them. For example, if we take a random sequence, and add 10^6 zeros in front of it, the resulting sequence will still be random in the sense of their definition. This definitely does not bode well with the intuitive understanding of randomness: e.g., if a roulette stops at 0 for 10^6 times in a row, everyone would agree that this roulette is rigged and does not produce a truly random sequence.

The reason for this conflict with intuition is that Kolmogorov-Martin-Löf's definition only formalizes the postulate that *events with 0 probability cannot occur*, while our intuition prompts a stronger statement: *An event with a very small probability cannot occur.* How can we formalize this stronger postulate?

First guess does not work. Similarly to the case of 0-probability events, we can formulate a seemingly natural first guess on how to formalize this postulate. Namely, we can fix some small number p_0, so small that all probabilities below p_0 would be, by all means, considered "very small", and thus, formalize the above intuitive principle as follows: *If an event E has probability $\mu(E) < p_0$, then it cannot occur.*

Alas, this seemingly natural first guess does not work (just like the seemingly natural first guess did not work for 0-probability events). To show that this first guess does not work, we can take the same example of a uniformly distributed random variable, for which $S = [0, 1]$ and μ is the Lebesgue measure. No matter how small p_0 is, we can always find an integer N for which $1/N < p_0$, and then divide the interval $S = [0, 1]$ into N subintervals of equal length $1/N$. The measure $\mu(E)$ of each of these subintervals is $1/N < p_0$ and therefore, according to the above formalization, none of these events can occur. In other words, a "random" element cannot belong to any of these intervals and thus, there cannot be any "random" elements at all.

This paradox is even worse than for 0-probability events: in the case of 0-probability events, we could easily resolve this paradox by restricting ourselves to definable events; here, all the events $[0, 1/N]$, $[1/N, 2/N]$, etc., are clearly definable. How can we resolve this paradox?

New definition. The paradox emerged because we assumed that the words "very small probability" had a universal numerical meaning, independent on the event. Since this assumption leads to a paradox, we must therefore conclude that which value is "very small" should depend on the event.

In other words, we cannot assume that there is a universal value p_0 such that if $\mu(E) < p_0$, then the event E cannot occur. We can, however, conclude that if we have a definable *sequence* of decreasing events (measurable sets) $E_1 \supseteq E_2 \supseteq \ldots \supseteq E_n \supseteq \ldots$ for which $\mu(E_n) \to 0$, then for some element E_N of this sequence, the probability will be sufficiently small, and therefore, this event E_N will be impossible. In other words, for some N, no "random" element $s \in S$ can belong to this set E_N. This idea can be easily formalized (see, e.g., Ref. 12):

Definition 3. *Let μ be a probability measure on a set S. We say that a set $R \subseteq S$ is a a set of random elements if it has the following property: for every definable decreasing sequence of measurable sets) $E_1 \supseteq E_2 \supseteq \ldots \supseteq E_n \supseteq \ldots$ for which $\mu(E_n) \to 0$, there exists an N for which $R \cap E_N = \emptyset$.*

Let us show that this definition is indeed consistent. Similarly to the case of 0-probability events, we will show that not only that random elements exist, but that "almost all" elements are random in this sense:

Theorem 1. *For every probability measure μ on a set S, and for every $\varepsilon > 0$, there exists a set R of random elements for which $\mu(R) > 1 - \varepsilon$.*

Proof. Similarly to the proof of Proposition 1, we can show that there exist at most countably many definable sequences, and therefore, at most countably many definable decreasing sequences $e = \{E_n\}$ for which $\mu(E_n) \to 0$. Therefore, we can order all such sequences into a sequence of sequences: $e^{(1)} = \{E_n^{(1)}\}$, $e^{(2)} = \{E_n^{(2)}\}$, ... Since for each of these sequences $e^{(k)}$, we have $\mu(E_n^{(k)}) \to 0$ as $n \to \infty$, there exists an N_k for which $\mu(E_{N_k}^{(k)}) < \varepsilon/2^k$.

Let us show that as R, we can take the complement $S \setminus E$ to the union E of all the sets $E_{N_k}^{(k)}$. Indeed, by our choice of R, for every definable decreasing sequence $e^{(k)} = \{E_n^{(k)}\}$, there exists an N, namely $N = N_k$, for which $R \cap E_N^{(k)} = \emptyset$.

To complete the proof, we must show that $\mu(R) > 1 - \varepsilon$. Indeed, from $\mu(E_{N_k}^{(k)}) < \varepsilon/2^k$, we conclude that $\mu(E) = \mu(\cup E_{N_k}^{(k)}) \le \sum \mu(E_{N_k}^{(k)}) < \sum \varepsilon/2^k = \varepsilon$, and therefore, $\mu(R) = \mu(S \setminus E) = 1 - \mu(E) > 1 - \varepsilon$. Q.E.D.

3. RANDOMNESS MAKES INVERSE PROBLEMS WELL-DEFINED

We will show that for any definable probability measure μ on a set S, if we restrict ourselves to *random* elements $s \in S$ (i.e., to element $s \in R$), then the inverse problem becomes well-defined, i.e., for any continuous 1-1 function $f : S \to Y$, the inverse mapping $f_{|R}^{-1}$ is continuous as well.

To prove this, we will show that the closure $\overline{R}$ is a compact set, and therefore, even the mapping $f_{|\overline{R}}^{-1}$ is continuous, and therefore, its restriction $f_{|R}^{-1}$ is continuous too. This result was first announced in Ref. 15.

Definition 4. *A definable metric space (S, d) is called definably separable if there exists a definable everywhere dense sequence $s_n \in S$.*

Definition 5. *We say that a probability measure μ on a metric space S with a metric d is consistent with the metric d if for every point $s \in S$ and for the every $r > 0$, the ball $D_r(s) = \{s' \mid d(s, s') \le r\}$ is measurable with respect to μ.*

Theorem 2. *Let S be a definably separable definable metric space, let μ be a probability measure μ on S which is consistent with the metric, and let R be a set of random elements of S. Then, the closure $\overline{R}$ is a compact set.*

Corollary. *Let S be a definably separable definable metric space, let μ be a probability measure μ on S which is consistent with the metric, and let R be a set of random elements of S, and let $f : S \to Y$ be a continuous 1-1 function. Then, the inverse mapping $f_{|R}^{-1} : Y \to R$ is continuous.*

In other words, if we assume that the actual state s is random, then, as the measurement error become smaller and smaller, the reconstructed element $\widetilde{s}$ gets closer and closer to the actual state s.

Proof of Theorem 2. A set K in a metric space S is compact if and only it is closed, and for every positive real number $\varepsilon > 0$, it has a finite $\varepsilon-$net, i.e., a finite set $K(\varepsilon)$ with the property that every $s \in K$, there exists an element $s(\varepsilon) \in K(\varepsilon)$ that is $\varepsilon-$close to s.

The closure $K = \overline{R}$ is clearly closed, so, to prove that this closure is compact, it is sufficient to prove that it has a finite $\varepsilon-$set for all $\varepsilon > 0$. For that, it is sufficient to prove that for every $\varepsilon > 0$, there exists a finite $\varepsilon-$net for the set R.

If a set R has a $\varepsilon-$net $R(\varepsilon)$, and $\varepsilon' > \varepsilon$, then, as one can easily see, this same set $R(\varepsilon)$ is also a $\varepsilon'-$net for R. Therefore, it is sufficient to show that finite $\varepsilon-$nets for R exist for $\varepsilon = 2^{-k}, k = 0, 1, 2, \ldots$

Let us fix $\varepsilon = 2^{-k}$. Since the set S is definably separable, there exists a definable sequence $s_1, \ldots, s_i, \ldots$ which is everywhere dense in S. As E_n, we will now take the complement to the union U_n of n closed balls $D_\varepsilon(s_1), \ldots, D_\varepsilon(s_n)$ of radius ε with centers in $s_1, \ldots, s_n$. Since each ball is measurable, the set E_n is measurable too.

Clearly, $E_n \supseteq E_{n+1}$. Since s_i is an everywhere dense sequence, for every $s \in S$, there exists an n for which $s \in D_\varepsilon(s_n)$ and for which, therefore, $s \in U_n$ and $x \notin E_n = S \setminus U_n$. Hence, the intersection of all the sets E_n is empty, and therefore, $\mu(E_n) \to 0$.

Therefore, according to the definition of a set of random elements, there exists an N for which $R \cap E_N = \emptyset$. This means that $R \subseteq U_N$. This, in its turn, means that the elements $s_1, \ldots, s_N$ form an $\varepsilon-$net for R.

So, the set R has a finite $\varepsilon-$net for $\varepsilon = 2^{-k}, k = 0, 1, 2, \ldots$ Hence, its closure $\overline{R}$ is compact. Q.E.D.

4. WHAT IF WE DO NOT KNOW PROBABILITIES?

4.1. Formulation of the problem and a new definition of "typicality"

If we *know* the *probabilities* of different states, i.e., if we know the probability measure μ on the set S of possible states, then it is natural to assume that the actual state s is random with respect to this measure. In this case, as we have shown, the inverse problem becomes well-defined. However, in many real-life situation, we may assume that the actual state is random with respect to *some* probability distribution, but we do not know the corresponding probabilities (or, at least, we do not know the exact values of these probabilities). It turns out that even when we do not know the exact probabilities, when we only know that s is random with respect to *some* (unknown) probability measure, the inverse problem is still well-defined.

Indeed, our definition of the set R of random elements is based on considering decreasing definable sequences of measurable sets for which $\mu(E_n) \to 0$. Let us first comment on the word "measurable". For Lebesgue measure, it is known that there exist non-measurable sets, but none of them is definable (to be more precise, proving the existence of non-measurable sets requires the use of the Axiom of Choice). Therefore, for Lebesgue measure, every definable set is measurable and hence, for this measure, we can omit the word "measurable" from the definition of the set of random elements. A similar omission is possible for other reasonable measures. Therefore, we can re-formulate the above definition by saying that R is a set of random elements if for every definable decreasing sequence for which $\mu(E_n) \to 0$, there exists an N for which $E_N \cap R = \emptyset$. For a decreasing sequence of sets, the limit $\lim \mu(E_n)$ is equal to the measure $\mu(E)$ of the intersection $E = \cap E_n$, so we only need to consider sequences for which this intersection has measure 0.

If for some sequence $\{E_n\}$, the intersection $E = \cap E_n$ is non-empty, then for some measures μ, we can have $\mu(E) > 0$. Therefore, if we do not have any information about the probability measure, we cannot tell whether the above property should be applicable to this sequence.

However, if the intersection is empty, then we can absolutely sure that whatever measure μ we consider, the above property should be applied, i.e., there should exist an N for which $E_N \cap R = \emptyset$. In other words, whatever measure μ we take, the set of all random elements must satisfy the following property: for every definable decreasing sequence E_n with an empty intersection, there exists an N for which $E_n \cap R = \emptyset$. It turns out that this property is sufficient to make the inverse problem well-defined.

Definition 6. *Let S be a definable set. We say that a subset $R \subseteq S$ is a set of typical elements if for every definable decreasing sequence $E_1 \supseteq E_2 \supseteq \ldots$ with an empty intersection, there exists an N for which $E_N \cap R = \emptyset$.*

This condition is weaker than the condition to be a set of random elements with respect to a certain probability measure; in particular, any set of *random* elements with respect to Lebesgue measure is the set of typical elements in this sense. Thus, the above existence result (Theorem 1) also proves the existence of a set of typical elements.

4.2. If we only consider "typical" states, then the inverse problem is well-defined

Let us show that if we only consider typical states, then the inverse problem is well-defined.

Theorem 3. *Let S be a definably separable definable metric space, and let R be a set of typical elements of S. Then, the closure $\overline{R}$ is a compact set.*

We have, in effect, proved this result when we proved Theorem 2.

Corollary. *Let S be a definably separable definable metric space, let R be a set of typical elements of S, and let $f : S \to Y$ be a continuous 1-1 function. Then, the inverse mapping $f^{-1}_{|R} : Y \to R$ is continuous.*

4.3. How can we actually use this result to get guaranteed estimates?

To actually use this result, we need an *expert* who will tell us what is typical and what is not. We will show that if we use such an expert, then for every computable function $f : S \to Y$, if we know that $s \in R$, then sufficiently accurate knowledge of $f(s)$ will enable us to reconstruct s with any given accuracy.

Definition 7.

- *By an expert, we mean a mapping $\mathcal{E} : \{E_n\} \to Z$ that transforms a definable decreasing sequence with an empty intersection into an integer N for which $E_N \cap R = \emptyset$ (i.e., for which all elements from E_N are not typical).*
- *We say that an output is computable with an expert if it is computable on a computer that can consult an expert (i.e., that sends an expert a formula defining $\{E_n\}$ and gets N).*

The following definitions are standard in constructive analysis.[9,11,10,20]

Definition 8.

- *We say that an algorithm $\mathcal{U}$ computes a real number x if for every natural number k, it generates a rational number r_k such that $|r_k - x| \le 2^{-k}$. We say that we have a computable real number if we have an algorithm $\mathcal{U}$ that computes it.*
- *By a computable separable metric space, we understand a separable metric space (S, d) with an everywhere dense sequence $\{s_n\}$ for which there exists an algorithm, that transforms a pair of positive integers n, m into a computable real number $d(s_n, s_m)$.*
- *By a computable element of a computable space we understand a pair consisting of an element $s \in S$ and an algorithm that given n, returns an integer $m(n)$ for which $s_{m(n)}$ is a $2^{-n}-$approximation to s.*
- *Let S and Y be computable separable metric spaces. We say that an algorithm $\mathcal{V}$ computes a function $f : S \to Y$ if $\mathcal{V}$ includes calls to an (unspecified) algorithm $\mathcal{U}$ so that when we take as $\mathcal{U}$ an algorithm that computes an element $s \in S$, then $\mathcal{V}$ will compute an element $f(s) \in Y$.*
- *We say that a computable function f is constructively continuous on a set S if there exists an algorithm, that for every $\varepsilon > 0$, generates $\delta > 0$ such that if $d(s, s') \le \delta$, then $d(f(s), f(s')) \le \varepsilon$.*

Definition 9. *Assume that we are given the following information:*

- *computable separable metric spaces S and Y;*
- *a computably continuous computable 1-1 function $f : S \to Y$;*
- *an algorithm that, given an integer k, returns a $2^{-k}-$approximation to $f(s)$, where $s \in R$ is an (unknown) typical element;*
- *a positive integer l.*

We say that algorithm solves the inverse problem if, given the above information, this algorithm returns a $2^{-l}-$approximation to s. If such an algorithm exists, then we will say that an inverse problem is computable.

Theorem 4. *The inverse problem is computable with an expert.*

Comment. The algorithm described in the proof is general and therefore (as many general algorithms), when applied to simple problems, it may require unnecessarily many computation steps. There are cases when simpler methods are possible: e.g., if the signal that we are trying to reconstruct is a smooth function, then we can ask an expert what is the upper bound for the signal's energy $\int (\dot{x}(t))^2\,dt$, and then use known regularization techniques.[21]

Proof of Theorem 4. Due to the above Corollary, the inverse function f^{-1} is continuous on $f(R)$. In particular, for every l, there exists a $\delta > 0$ such that if $s, s' \in R$ and $d(f(s), f(s')) \le \delta$, then $d(s, s') \le 2^{-l}$. If we know δ, then we can compute the desired approximation to s as follows. Since S is definably separable, there exists a definable sequence s_n that is everywhere dense in S. Using this sequence, we:

- Compute $f(s)$ with accuracy $\delta/8$; the result of this computation (one of the elements of the everywhere dense sequence y_m) will be denoted by $\widetilde{f}(s)$.
- For $n = 1, 2, \ldots$, compute $f(s_n)$ with accuracy $\delta/8$, and for the result $\widetilde{f}(s_n)$ of this computation, we compute the distance $d(\widetilde{f}(s), \widetilde{f}(s_n))$ with an accuracy $\delta/8$. When this estimate $\widetilde{d}$ is $\le \delta/2$, we stop, and produce s_n as the desired result.

Let us show that this algorithm will work.

- First, let us prove that this algorithm will stop. Indeed, since the sequence s_n is everywhere dense in S, we have a subsequence s_{n_k} that tends to s. Since f is continuous, we have $f(s_{n_k}) \to f(s)$. So, there exists a k for which $d(f(s_{n_k}), f(s)) \le \delta/8$. Since $\widetilde{f}(s)$ and $\widetilde{f}(s_{n_k})$ are $(\delta/8)-$approximations to $f(s)$ and $f(s_{n_k})$, we can conclude that $d(\widetilde{f}(s_{n_k}), \widetilde{f}(s)) \le d(\widetilde{f}(s_{n_k}), f(s_{n_k})) + d(f(s_{n_k}), f(s)) + d(f(s), \widetilde{f}(s)) \le \delta/8 + \delta/8 + \delta/8 = (3/8)\delta$. Hence $\widetilde{d} \le d(\widetilde{f}(s_{n_k}), \widetilde{f}(s)) + \delta/8 \le \delta/2$. So, if the algorithm did not stop before the value n_k, it will stop at this point.
- Let us now show that the algorithm produces the desired value. Indeed, if $\widetilde{d} \le \delta/2$, then $d(\widetilde{f}(s_n), \widetilde{f}(s)) \le \widetilde{d} + \delta/8 \le \delta/2 + \delta/8$, and $d(f(s_n), f(s)) \le d(\widetilde{f}(s_{n_k}), \widetilde{f}(s)) + d(\widetilde{f}(s_n), f(s_n)) + d(f(s), \widetilde{f}(s)) \le \delta/2 + \delta/8 + \delta/8 + \delta/8 < \delta$. Hence, due to our choice of δ, we have $d(s, s_n) \le 2^{-l}$.

So, to complete the description of the algorithm, we must describe how to compute δ.

We must find δ such that if $d(f(s), f(s')) \le \delta$, then $d(s, s') \le 2^{-l}$. To find this δ, let us choose an integer p. Since f is constructively continuous, we can compute the value η such that if $d(s, s') \le \eta$, then $d(f(s), f(s')) \le 2^{-p}$. Let us take $\beta = \min(\eta, 2^{-p})$. For this choice of β, if $d(s, s') \le \beta$, then $d(s, s') \le 2^{-p}$ and $d(f(s), f(s')) \le 2^{-p}$. Let us find a $\beta-$net $s^{(1)}, \ldots, s^{(m)}$ for S. This can be done similarly to the proof of Lemma 1, only instead of referring to existence of the desired N, we use the expert to produce such an N. For this $\beta-$net, we take all pairs $s^{(i)}, s^{(j)}$ for which $d(s^{(i)}, s^{(j)}) \ge 2^{-l} - 2\beta$, and find the smallest value M of $d(f(s^{(i)}), f(s^{(j)}))$ for all such pairs. If $M > 2\beta$, then we return $\delta = M - 2\beta$. Else, we increase p by 1, and repeat the process again and again. Let us prove that this part of the algorithm does produce the correct value of δ (and thus, that the entire algorithm is correct). Indeed:

- Let us first show that this algorithm will stop. Indeed, due to the above corollary, there exists a value $\delta' > 0$ for which if $d(f(s), f(s')) \le \delta'$, then $d(s, s') \le (1/2) \cdot 2^{-l}$. So, if $d(s^{(i)}, s^{(j)}) > (1/2) \cdot 2^{-l}$, then $d(f(s^{(i)}), f(s^{(j)})) > \delta'$. Hence, if we take p so big that $2^{-p} < \min(\delta'/2, (1/4) \cdot 2^{-l})$, then from $d(s^{(i)}, s^{(j)}) \ge 2^{-l} - 2\beta$, and from $\beta \le 2^{-p} < (1/4) \cdot 2^{-l}$, we can conclude that $d(s^{(i)}, s^{(j)}) > (1/2) \cdot 2^{-l}$, and therefore, that $M \ge \delta' > 2 \cdot 2^{-p} \ge 2\beta$.
- Let us now show that if the algorithm did stop, then we got the desired δ. Indeed, let $d(f(s), f(s')) \le M - 2\beta$. Since elements $s^{(i)}$ form a $\beta-$net for S, there exist elements $s^{(i)}$ and $s^{(j)}$ that are $\beta-$close to s and s' correspondingly. Sue to the choice of β, we can conclude that $d(f(s), f(s^{(i)}) \le \beta$ and $d(f(s'), f(s^{(j)}) \le \beta$. Hence, $d(f(s^{(i)}), f(s^{(j)})) \le d(f(s), f(s')) + 2\beta \le M$. By definition of M, this means that $d(s^{(i)}, s^{(j)}) \le 2^{-l} - 2\beta$. Therefore, $d(s, s') \le d(s^{(i)}, s^{(j)}) + 2\beta \le 2^{-l}$.

So, the second part of the algorithm produces correct δ. Q.E.D.

5. RELATION TO BIRKHOFF'S ESTHETIC MEASURE

5.1. Formulation of the problem

Beauty as a particular case of prior information. One of the applications of the inverse problem is to "clean" (filter) musical recordings, photos of old paintings, etc. In such applications, we know that the reconstructed state s is *beautiful*. How can we take this additional information into consideration?

How can we formalize beauty? Birkhoff's formula. In the 1930s, G. D. Birkhoff, one of the world leading mathematicians, has proposed a formula that described beauty in terms of "order" O and "complexity" C (see Refs. 2–6; see also Ref. 8). Namely, according to his formula, the beauty B of an object is equal to $B = O/C$. How can we describe "order" and "complexity"?

In the simplest cases, Birkhoff formalized these notions and showed that his formula is indeed working. Namely, he showed that the beauty of simple geometric patterns, of simple melodies, and even of simple verses can be well described by his formula. However, since there was no general notion of complexity, he was unable to formalize his idea in the general case. This is what we are planning to do in this section.

5.2. Towards formalization of Birkhoff's formula

Relation to formalized complexity. In our formalization, we will use the general computer-based notion of object complexity, which is widely used in computer science. For example, we can define the *complexity* $C(x)$ of an object x as the length $l(p)$ of the shortest program p (in a certain language) which generates this object. This notion of object complexity was originally proposed by G. Chaitin, A. Kolmogorov, and R. Solomonoff, and it is usually called *Kolmogorov complexity* (see, e.g., Ref. 18). Alternatively, we can use a *modification* of this notion which takes into consideration not only the *length* $l(p)$ of the program p (i.e., the number of bits in its computer description), but also the *time* $t(p)$ that the program p takes to generate the desired object x.

In order to choose an appropriate formalization, let us start with an informal discussion of Birkhoff's ideas.

Informal motivations: the ideas behind Birkhoff's notions. In Birkhoff's description, *complexity* of an object looks like time which is necessary to generate this object. For example, he defines the complexity of a polygon as the number of its vertices, etc. Intuitively, it is clear that beauty must be reasonably simple, so, all other characteristics being equal, the more over-complicated the object is, the less beautiful it is.

Similarly, Birkhoff's *order* looks like a simplicity of the description: if we can describe an object by using a shorter text, then its order is higher. If the only way of describing an object is to enumerate all its pixels (all its nodes for a melody, all its vertices for a polygon, etc.), then this object does not have much order in it. Intuitively, it seems reasonable that an object with some order in it should (all else being equal) look prettier than an object with less order. How can we formalize this notion of "order"?

By a *description*, we mean a *complete* description, i.e., a description which is detailed enough so that, given this description, we can uniquely reconstruct the object. In other words, the description must serve as a *program* for a computational device which, given this description, reconstructs the object. In these terms, the length of the description is the length $l(p)$ of this program p. So, the smaller $l(p)$, the more order is there in the object.

Summarizing our discussion of complexity and order, we can conclude that the beauty $B(s)$ of an object s depends on the *time* $t(p)$ of the program p which generates s, and on the *length* $l(p)$ of this program, i.e., that $B(x) = f(t(p), l(p))$ for some function $f(t, l)$. The only thing we know about the function $f(t, l)$ is that it should monotonically decrease with the increase of each of the variables t and l.

In these term, the question of formalizing beauty can be reformulated in more mathematically-sounding terms: Which function $f(t, l)$ should we choose?

Which function $f(t, l)$ should we choose? It is well known in computer science that there is a *trade-off* between the program time and the program length. A short program usually uses only a few ideas of speeding up computations, and thus, takes a reasonable amount of time to run. If we want to speed up the computations, we must add some complicated ideas and modify the algorithm. As a result, to make the program faster, we must usually make it longer. Vice versa, we can often shorten the program by eliminating some of the time-saving parts and thus, by making its running time longer.

This trade-off is not only true for programs written in the same programming language, the same trade-off is true if we compare programs written on programming languages of different level. For example, we can write a program in machine code (or in assembler language, which is close to the machine code).

- In a machine-code program, we have to spell out all necessary steps, so this program will be reasonably long. On the other hand, in a machine code program, every instruction will be immediately implemented, so running this program does not take too long.
- Alternatively, we can write our program in a high level programming language (e.g., in C++). In this case, the program is usually shorter, because we do not need to spell out all the details, it is sufficient to describe the construction that we want to use (like a loop or calling a function). However, when we run this short program, we first need to translate it into the machine code (i.e., *compile* it), and this compiling takes extra time.

Thus, we can get a shorter program which runs longer, or we can have a longer program which runs faster.

Our definition of the formalized beauty depended on the program p. It is reasonable to require that the "beauty" of an object x should not depend on which level we write this program p. Let us formalize this requirement.

By going to a different level of programming, we can cut a lot of bits from the length of the program. Let us describe this cut step-by-step and analyze what happens if we cut exactly one bit.

We are interested not in the abstract notion of beauty, but in the much more specific notion of the beauty of a state. If we cut a bit from the program that generates the state s, we get a new program p' which is exactly one bit shorter ($l(p') = l(p) - 1$). To generate the desired state s, since we do not know whether the deleted bit was 0 or 1, we can try both possible values of this bit (i.e., run two programs $p'0$ and $p'1$) and find out which of the two states is better. Thus, if we delete a bit, then instead of running the original program p once, we run *two* programs $p'0$ and $p'1$. Hence, crudely speaking, when we decrease the length of the program by 1, we thus get a double increase in the running time: $t(p') = 2t(p)$.

From this viewpoint, the fact that the beauty should not depend on the level means, in particular, that the values of $B(s)$ computed as $f(t(p), l(p))$ should stay the same if we replace the original program p by a one-bit-shorter program p'. In other words, we should have $f(t(p'), l(p')) = f(t(p), l(p))$. Since we know that $l(p') = l(p) - 1$ and $t(p') = 2t(p)$, we thus conclude that $f(2t(p), l(p) - 1) = f(t(p), l(p))$ for every program p. In other words, the desired function $f(t, l)$ must satisfy, for every two integers t and l, the following equation:

$$f(2t, l - 1) = f(t, l). \tag{1}$$

Functions which satisfy this equation can be explicitly described:

5.3. Main result

Definition 10. *We say that a function $f(t, l)$ is invariant if it satisfies the equation (1) for all positive integers t and l.*

Theorem 5. *A function $f(t, l)$ is invariant if and only if $f(t, l) = F(t \cdot 2^l)$ for some function $F(z)$ of one variable.*

Proof. From the equation (1), we conclude that $f(t, l) = f(2t, l - 1)$. Applying the same equation (1) to the right-hand side of the new equality, we conclude that $f(2t, l - 1) = f(2^2 t, l - 2)$, and thus, that $f(t, l) = f(2^2 t, l - 2)$. Similarly, we can prove that

$$f(t, l) = f(2^2 t, l - 2) = f(2^3 t, l - 3) = \ldots = f(2^k t, l - k)$$

for an arbitrary k. In particular, for $k = l$, we conclude that $f(k, l) = f(2^l t, 0)$. Thus, the Theorem is true, for the function $F(z) = f(z, 0)$. Q.E.D.

5.4. Discussions

Our result justifies Birkhoff's formula. Let us show that this result justifies Birkhoff's formula. Namely, we will show that the search for the "most beautiful" state is equivalent to looking for a state for which Birkhoff's ratio takes the largest possible value for appropriately defined quantities O and C.

Indeed, since we assumed that the function $f(t, l)$ is monotonically decreasing in both variables, we can conclude that the function $F(z)$ is monotonically decreasing too. So, looking for the "most beautiful" state means looking for the state generated by a program p for which the product $t(p) \cdot 2^{l(p)}$ takes the smallest possible value, or, equivalently, for which the inverse value $2^{-l(p)}/t(p)$ takes the largest possible value. We have already mentioned that the running

time $t(p)$ is a natural formalization of Birkhoff's complexity C, and that Birkhoff's "order" O is a monotonically decreasing function of the program length $l(p)$. Thus, *looking for the most beautiful state means looking for a state for which the ratio O/C takes the largest possible value, where $C = t(p)$ and $O = 2^{-l(p)}$.* So, we indeed get a justification for Birkhoff's formula.

Our result makes perfect sense from the pragmatic viewpoint. For each possible state s, we can define its "beauty" $b(s)$ as the smallest possible value of the product $t(p) \cdot 2^{l(p)}$ for all possible programs p which generate this state. Then, finding the most beautiful state means finding the state s which if consistent with all the observations and for which thus defined quantity $b(s)$ takes the smallest possible value.

This notion $b(s)$ is known in the theory of Kolmogorov complexity: namely, it was introduced by Leonid A. Levin[17] as one of the possible modifications of Kolmogorov complexity which takes into consideration not only the length $l(p)$ of the program, but its running time $t(p)$ as well. Levin has proven that if we are looking for an optimal (asymptotically fastest) universal algorithm for solving different search problems (see, e.g., Ref. 13), then this optimal algorithm should check all possible states in the increasing order of their Levin's complexity $b(s)$ (see[1,17,18]).

Thus, our formalization of beauty makes perfect pragmatic sense: if we want to find the best state s as fast as possible, we must first look among the prettiest states (i.e., among the states with the smallest possible value of $b(s)$), then among the next prettiest states, etc.

This theoretical idea is not yet a practically working tool. *Theoretically*, Birkhoff's idea seems to work well. However, *in practice*, there is a big obstacle to applying this idea, because Kolmogorov complexity is not algorithmically computable.[18]

Levin's modification of Kolmogorov complexity is actually computable but computing it requires too long time, so for all practical purposes, it is not computable at all. What can we do to make this criterion practically useful?

How to transform this theoretical idea into a practically working tool? A possible approach towards making this notion practical is to take into consideration the fact that the Kolmogorov complexity is not computable because it is based on considering *all* possible algorithms. If we limit the class of algorithms, we get a computable version of Kolmogorov complexity. This idea was used, e.g., in Ref. 14, where a similarly modified version of Kolmogorov complexity was used to successfully predict the time required for a human to remember a geometric pattern. How can we come up with reasonable computable analogues of complexity and order (symmetry)?

Complexity of a computer object (string, image, etc.) can be measured by the ability of compressing programs to compress them. Thus, to get a computable estimate for complexity, we can use an advanced compression algorithm (e.g., an algorithm that underlies the widely used zip compression), and measure the complexity by the length of the compressed object: *if the compressed text is short, the object was easy; if the compressed text still takes many bits, the compressed object was complex.*

To measure the *order* (= *symmetry*, see, e.g., Ref. 22) of an object, we can, similarly, use compression procedures, but this time, only procedures which use symmetry to compress. The most widely known symmetry-motivated compression techniques is the *wavelet* compression (see, e.g., Ref. 19). In view of this, we use the length of the wavelet compression as an indication of the order: *an image with a short wavelet compression has high order, while an image whose wavelet compression has approximately the same length as the original image has low order.*

We are planning to experimentally check that these definitions indeed lead to a reasonable characterization of beauty.

ACKNOWLEDGMENTS

This work was supported in part by NASA under cooperative agreement NCC5-209, by NSF grant No. DUE-9750858, by the United Space Alliance, grant No. NAS 9-20000 (P.O. 297A001153), and by the Future Aerospace Science and Technology Program (FAST) Center for Structural Integrity of Aerospace Systems, effort sponsored by the Air Force Office of Scientific Research, Air Force Materiel Command, USAF, under grant number F49620-95-1-0518.

We would like to thank all the researchers who discussed different parts of the presented material with us. Especially, we want to thank P. Benioff, N. Cartwright, P. Cheeseman, P. Cohen, S. A. Cook, Ya. Eliashberg, A. M. Finkelstein, B. van Fraassen, I. Frenkel, M. Gelfond, A. A. Grib, Yu. Gurevich, J. Halpern, R. B. Kearfott, L. Levin, V. Lifschitz, J. McCarthy, A. Meyer, V. Moncrief, B. Z. Moroz, J. Moschovakis, A. Neumaier, R. Parikh, H. Przymusinska, T. Przymusinski, V. Sazonov, A. Semenov, G. N. Solopchenko, P. Suppes, A. Troelstra, and V. Vovk.

REFERENCES

1. M. Beltran, G. Castillo, and V. Kreinovich, "Algorithms that still produce a solution (maybe not optimal) even when interrupted: Shary's idea justified", *Reliable Computing*, **4**, No. 1, pp. 39–53, 1998.
2. G. D. Birkhoff, "A mathematical approach to aesthetics", *Scietia*, **50**, Vol. 50, pp. 133–146, September 1931 (reprinted in Ref. 7, **3**, pp. 320–333).
3. G. D. Birkhoff, "A mathematical theory of aesthetics", *Rice Institute Pamphlet*, **19**, pp. 189–342, July 1932 (reprinted in Ref. 7, **3**, pp. 382–535).
4. G. D. Birkhoff, *Aesthetic measure*, Harvard University Press, Cambridge, MA, 1933.
5. G. D. Birkhoff, "Three public lectures on scientific subjects", *Rice Institute Pamphlet*, **28**, pp. 1–76, January 1941 (reprinted in Ref. 7, **3**, pp. 755–777).
6. G. D. Birkhoff, "Mathematics of aesthetics", In: J. R. Newman (ed.), *The World of Mathematics*, Simon and Schuster, N.Y., **4**, pp. 2185–2208, 1956.
7. G. B. Birkhoff, *Collected Mathematical Papers*, Dover, N.Y., 1960.
8. G. Birkhoff, "Mathematics and psychology", *SIAM Review*, **11**, No. 4, pp. 429–467, 1969.
9. E. Bishop, *Foundations of Constructive Analysis*, McGraw-Hill, 1967.
10. E. Bishop, D. S. Bridges, *Constructive Analysis*, Springer, N.Y., 1985.
11. D. S. Bridges, *Constructive Functional Analysis*, Pitman, London, 1979.
12. A. M. Finkelstein and V. Kreinovich. "Impossibility of hardly possible events: physical consequences," *Abstracts of the 8th International Congress on Logic, Methodology and Philosophy of Science*, **5**, Pt. 2, pp. 23–25, Moscow, 1987.
13. M. R. Garey and D. S. Johnson, *Computers and intractability: a guide to the theory of NP-completeness*, Freeman, San Francisco, 1979.
14. R. M. Granovskaya, I. Ya. Bereznaya, and A. N. Grigorieva, *Perception of form and forms of perception*, Erlbaum, Hillsdale, NJ, 1987.
15. V. Kozlenko, V. Kreinovich, and G. N. Solopchenko, *A method for solving ill-defined problems*, Technical Report No. 1067, Leningrad Center of Scientific and Technical Information, Leningrad, 1984 (in Russian).
16. V. Kreinovich and L. Longpré, "Guaranteed predictions based on probabilistic knowledge: why and how", *International Conference on Interval Methods and their Application in Global Optimization (INTERVAL'98), April 20–23, Nanjing, China, Abstracts*, pp. 61–63, 1998.
17. L. A. Levin, "Universal sequential search problems", *Problems of Information Transmission*, **9**, No. 3, pp. 265–266, 1973.
18. M. Li and P. Vitányi, *An introduction to Kolmogorov complexity and its applications*, Springer-Verlag, N.Y., 1997.
19. Y. Meyer, *Wavelets algorithms and applications*, SIAM, Philadelphia, 1993.
20. M. Nakamura, R. Mines, and V. Kreinovich. "Guaranteed intervals for Kolmogorov's theorem (and their possible relation to neural networks)", *Interval Computations*, No. 3, pp. 183–199, 1993.
21. A. N. Tikhonov and V. Y. Arsenin. *Solutions of ill-posed problems,* V. H. Winston & Sons, Wasnington, DC, 1977.
22. H. Weyl, "Symmetry", In: J. R. Newman (ed.), *The World of Mathematics*, **1**, pp. 671–724, Simon and Schuster, N.Y., 1956.