9-1997

We Must Choose the Simplest Physical Theory: Levin-Li-Vitanyi Theorem and its Potential Physical Applications

Dirk Fox

Martin Schmidt

Misha Kosheleva

Vladik Kreinovich
The University of Texas at El Paso, vladik@utep.edu

Luc Longpre
The University of Texas at El Paso, longpre@utep.edu

See next page for additional authors

Comments:
Technical Report: UTEP-CS-97-21
In: Gary J. Erickson, Joshua T. Rychert, and C. Ray Smith (eds.), *Maximum Entropy and Bayesian Methods*, Kluwer, Dordrecht, 1998, pp. 238-251.

WE MUST CHOOSE THE SIMPLEST PHYSICAL THEORY: LEVIN-LI-VITÁNYI THEOREM AND ITS POTENTIAL PHYSICAL APPLICATIONS

D. FOX, M. SCHMIDT, M. KOSHELEV
V. KREINOVICH, L. LONGPRÉ
Department of Computer Science
University of Texas at El Paso
El Paso, TX 79968, USA[‡]

AND

J. KUHN
National Solar Observatory/Sacramento Park,
Sunspot, NM 88349-0062, USA[§]

Abstract. If several physical theories are consistent with the same experimental data, which theory should we choose? Physicists often choose the *simplest* theory; this principle (explicitly formulated by Occam) is one of the basic principles of physical reasoning. However, until recently, this principle was mainly a *heuristic* because it uses the *informal* notion of simplicity.

With the explicit notion of simplicity coming from the Algorithmic Information theory, it is possible not only to *formalize* this principle in a way that is consistent with its traditional usage in physics, but also to *prove* this principle, or, to be more precise, *deduce* it from the fundamentals of mathematical statistics as the choice corresponding to the least informative prior measure. Potential physical applications of this formalization (due to Li and Vitányi) are presented.

In particular, we show that, on the *qualitative* level, most fundamental ideas of physics can be re-formulated as natural steps towards choosing a theory that is the simplest in the above *precise* sense (although on the intuitive level, it may seem that, e.g., *classical* physics is easier than quantum physics): in particular, we show that such ideas as Big Bang cosmology,atomism, uncertainty principle, Special Relativity, quark confinement, quantization, symmetry, supersymmetry, etc. can all be justified by this (Bayesian justified) preference for formalized simplicity.

[‡]Emails: {dfox,mschmidt,mkosh,vladik}@cs.utep.edu.
[§]E-mail: jkuhn@sunspot.noao.edu

Key words: Kolmogorov complexity, Algorithmic Information theory, Occam razor, Bayesian statistics, fundamental physics

1. The problem

The problem: what theory should we choose? If several physical theories are consistent with the same experimental data, which theory should we choose?

How is a theory chosen now. Physicists often choose the *simplest* theory.

This principle (explicitly formulated by Occam) is one of the most basic principles of physical reasoning.

The idea of choosing the simplest theory was actively advocated by Einstein. The most famous promoter of this idea was A. Einstein. For example, in his lecture "On the method of theoretical physics", he said that "It is the grand object of all theory to make ... irreducible elements as simple and as few in number as possible" ([3], p. 272).

This principle used to be heuristic. The principle of choosing the simplest physical theory was, until recently, only *heuristic*, because until recently, there was no well-accepted definition of simplicity.

Simplicity can now be defined formally. Now, with the advent of Algorithmic Information Theory (see, e.g., [16]), there are formal (and well-accepted) definitions of simplicity. The first such definition was proposed, in mid-60s, practically simultaneously, by A. Kolmogorov, R. Solomonoff, and G. Chaitin; this notion is called *Kolmogorov complexity*. Crudely speaking, Kolmogorov complexity of a text is the shortest length of a program (in some universal language) that generates this text. There are also several modifications of this definition. For different problems, different formalizations turn out to be more adequate.

This formalization leads to the following two (natural and related) questions:

- First, which of the definitions of simplicity is more adequate for choosing a physical theory?
- Second, now that the idea of choosing the simplest theory becomes formal, the question is: is this principle mathematical or physical? In other words:
  - does this formal principle *follow* from the traditional principles of mathematical *statistics* (that are used to process physical data), or
  - is this principle *independent* from the general principles of mathematical statistics, and somehow reflecting the specific structure of *our* physical world (as opposed to other mathematically possible "worlds")?

What we are planning to do. In Section 2, we describe the result by Li and Vitányi (which is, in its turn, based on Levin's theorem) that the basic principles of mathematical statistics do indeed lead to a formalization of the principle of choosing the simplest theory, and we explicitly describe the formalization of complexity that this theory corresponds to. Thus, we get answers to both questions.

The main ideas of our exposition are from Li and Vitányi [16]. However, since our main goal is formalizing this principle with respect to *physical theories*,

our exposition will be specifically oriented towards the physics readers and therefore, our formulation and our exposition will be somewhat different from [16]. In particular, we will give physical motivations for all definitions and requirements that are used in the proof.

In Section 3, we show that the resulting formalization is consistent with the traditional use of this principle in theoretical physics. Hopes and speculations are described in the final Section 4.

2. Justification of the principle of choosing the simplest physical theory

Basic statistical approach: brief reminder. According to the fundamentals of mathematical statistics, in order to choose a theory based on the experimental data, we must do the following:

- find the prior probabilities $P_0(H_i)$ of different hypotheses H_i;
- use the experimental data E to compute the posterior probabilities $P(H_i)$ of different hypotheses; to compute these probabilities, we can use the Bayes rule
$$P(H_i) = \frac{P(E|H_i) \cdot P_0(H_i)}{\sum P(E|H_j) \cdot P_0(H_j)},$$
where $P(E|H)$ is the probability of observing the data E in case the hypothesis H is true;
- and, finally, select a hypothesis whose probability is the largest $P(H_i) \to \max$.

The posterior probability $P(H_i)$ of a hypothesis H_i is also called its *likelihood*, and correspondingly, the above basic statistical approach is also called the *maximum likelihood* (ML) approach.

From the computational viewpoint, it is often convenient to use *negative logarithms* $-\log(P(H_i))$ of the probabilities: e.g., for the most widely used Gaussian distribution, the probabilities $\exp(-c \cdot x^2)$ require calling an exponential function and are, therefore, much more difficult to compute than their negative logarithms $-c \cdot x^2$. Since negative logarithm is a decreasing function, in terms of negative logarithms, the criterion $P(H_i) \to \max$ takes the form $-\log(P(H_i)) \to \min$.

At first glance, this approach may sound somewhat *subjective* because the resulting choice of a hypothesis depends on the original (rather subjective) choice of prior probabilities. However, it is known that in the long run (as we get more and more experimental data), this dependence disappears. To be more precise, let us assume that the actual theory is H_a. Then:

- If the prior probability of H_a is 0 (i.e., $P_0(H_a) = 0$), then no matter how many experimental data we get, if we apply the above Bayes formula, we always end up with $P(H_a) = 0$.
- However, if we make sure that the prior probability of the correct hypothesis is positive ($P_0(H_0) > 0$), then, as the number of experimental data increases, the probability $P(H_a)$ of this hypothesis will increase, while the probabilities of other hypotheses will decrease and eventually, we will get $P(H_a) \gg P(H_i)$ for all $H_i \neq H_a$ and therefore, eventually, the hypothesis H_a will be chosen.

Since we do not know which hypothesis is correct, we must require that an prior probability of each hypothesis is positive.

This mathematically sounding result is actually very intuitively clear: if we do not know which of the hypotheses is true, we must not throw away any of them, otherwise, we may throw away the only theory that is consistent with the observations.

How to apply this general approach to fundamental physics (first approximation). One of the major goals of physics is to predict the results of different experiments. Whatever *mathematical* language a theory or hypothesis is formulated in, eventually, we are interested in what *observation* results this theory predicts. These results can also be described in different terms, but since most measurements and practically all data processing is done by using computers, we can use the fact that inside the computers, everything is represented by 0's and 1's. Each 0 or 1 is called a *binary unit*, or, for short, a *bit*. Therefore, for simplicity of analysis (and without losing any generality), we can assume that all possible experimental results form a (potentially infinite) binary sequence $X = x_1 x_2 \ldots x_n \ldots$

Our goal is to predict either the results of *all* the experiments, or at least the results of *some* of them. So, our goal is to predict either the entire (potentially infinite) sequence X, or a finite portion of this sequence. Therefore, to apply the standard statistical methodology, we need a prior measure on the set of all such sequences.

To describe this measure, it is sufficient to describe, for each finite sequence $x = x_1 \ldots x_n$, the probability $p(x)$ that the (possible infinite) sequence X of measurement results will start with x. These probabilities must satisfy the two natural requirements:

- Since every sequence starting with x must continue either with 0, or with 1, we get the *additivity* condition for probabilities: $p(x) = p(x0) + p(x1)$.
- Also, the total probability must be equal to 1, i.e., $p(0) + p(1) = 1$.

The second requirement can be formulated as a particular case of the additivity requirement, if we allow an empty string $x = \Lambda$ and assume that

$$p(\Lambda) = 1 \tag{1}.$$

Comment. In the following text, we will see that we need to make a (minor) modification to these formulas; that is why we called this description the *first approximation*.

The prior probabilities must be computable (in some reasonable sense). The above-described statistical methodology requires us to *compute* the posterior probabilities based on the prior ones. Bayes' formula is explicit, but in order to use it, we must assume that the prior probabilities $P_0(H)$ are, in some reasonable sense, computable. What does this "computable" mean?

Crudely speaking, a prior probability of an event means the probability of this event that is computed based on some *a priori* ideas, i.e., probabilities computed based on some prior "theory".

If an event is simple enough, then, in this prior "theory", we can simply compute its probability *directly*. To be more precise, we can compute this probability with an arbitrary *accuracy*, i.e., there exists an algorithm that, given a positive integer k, computes a rational number r_k that is 10^{-k}-close to the desired probability p (i.e., for which $|r_k - p| \leq 10^{-k}$). Real numbers p for which such an algorithmic approximation is possible are called *computable*. If we have an algorithm that, given x and k, returns a 10^{-k}-approximation to $p(x)$, then $p(x)$ is called a *computable function*.

Often, however, the desired event can have many different reasons; it is *possible* to compute the probabilities related to *each reason* but it is *difficult* to compute the *total probability*. This effect is well known to experimental physicists who estimate the reliability of their experimental results (i.e., the probability that the observed results are caused not by the analyzed phenomena, but rather by noise). There are well-known examples when experimental claims, that were initially thought to be correct, later on turned out to be false because of some additional factors: e.g., Weber's experimental discovery of gravitational waves turned out to be an observation of a different phenomenon [17].

If we take this feature of experimental physics into consideration, we will arrive at the conclusion that for each finite sequence x, the desired probability $p(x)$ is not necessarily a computable real number, but that it is the *limit* $p(x) = \lim p_N(x)$, where $p_N(x)$ is the probability that only takes first N factors into consideration. The sequence $\{p_N(x)\}$ is:

- *non-decreasing* in the sense that $p_1(x) \leq p_2(x) \leq \ldots \leq p_N(x) \leq p_{N+1}(x) \leq \ldots$, and
- *computable* in the sense that there is an algorithm that, for every x, N, and k, returns a 10^{-k}-approximation to $p_N(x)$.

Functions that can be represented as limits of non-decreasing computable sequences are called *enumerable* (crudely speaking, this name comes from the fact that after we *enumerate* all factors, we get the desired probabilities). Thus, we can re-formulate the requirement on the function $p(x)$ by saying that p should be an *enumerable* function.

How to distinguish a reasonable prior probability measure from other possible probability measures. When choosing prior probabilities $p_0(x)$, the main requirement is not to assign probability 0 to an event E that could end up with a non-zero probability if we use a different prior probability measure $p(x)$. In particular, if, as the event E, we take the statement that X must start with a given finite sequence x, this requirement means that we must avoid the situations when $p_0(x)/p(x) = 0$, i.e., we must require that $p_0(x)/p(x) > 0$ for all words x.

Events E can be more complicated than that; for example, we may have events that are defined as *limits* of the above simple events. To cover all possible more complicated events, it is reasonable to require that not only all the ratios $p_0(x)/p(x)$ be positive, but that also the *limits* $p_0(x^{(N)})/p(x^{(N)})$ of such ratios, corresponding to different sequences $x^{(N)}$, should also be always positive.

In mathematical terms, this requirement can be formulated as follows: we have a set S of all possible values of the ratio $p_0(x)/p(x)$, and we require that if a

sequence of real numbers from this set S has a limit, then this limit should be positive. For an arbitrary set S of real numbers, the smallest possible limit of sequences from S is known to be equal to the *infimum* $\inf(S)$ of this set S (i.e., to its *greatest lower bound* (g.l.b.)). Thus, the requirement that *all* the limits are positive is equivalent to requiring that *the smallest* of these limits is positive, i.e., that $\inf(p_0(x)/p(x)) > 0$. This, in turn, means that there exists a positive number $c > 0$ for which, for all x, $p_0(x) \geq c \cdot p(x)$.

Thus, we can say that an enumerable probability measure $p_0(x)$ is an *ideal prior measure* if for every other enumerable probability measure $p(x)$, there exists a constant $c > 0$ for which $p_0(x) \leq c \cdot p(x)$ for all x.

Modification is needed. The above idea of an "ideal" prior probability measure was proposed, in the early 60s, by R. Solomonoff, one of the authors of Algorithmic Information Theory. However, Solomonoff himself showed that, in effect, the seemingly reasonable formalization of this idea (the one we have just described) does not work, in the sense that the above-defined "ideal" prior measure cannot exist (for detailed and precise history, see [16]).

In view of this impossibility, we must *modify* our definitions. Such a modification has actually been proposed. The modification, as well see in a second, is very natural from the computer science viewpoint (and from the common sense viewpoint as well).

Final definitions. The prior probability $p(x)$ can be obtained, e.g., if we poll several experts and take, as $p(x)$, the fraction $N(x)/N$ between the total number $N(x)$ of experts who believe that x will occur in the measurement, and the total number N of experts. If every expert is *a priori* definite in his beliefs, i.e., if he believes, for every word x, either that x will occur, or that x will not occur, then the resulting values $p(x)$ indeed form a probability measure: e.g., among all the experts who believe in $x = x_1 \ldots x_n$, some believe in $x0 = x_1 \ldots x_n 0$ and some in $x1 = x_1 \ldots x_n 1$ (but everyone does believe in one of these two), and therefore, $N(x) = N(x0) + N(x1)$ and $p(x) = p(x0) + p(x1)$.

In reality, however, experts may be *undecided* about some measurement results. So, among $N(x)$ experts who initially believe in x, some will believe in $x0$, some in $x1$, while some will be undecided about what will follow x. In databases and in other areas of computer science, the case when we do not know a certain value is called a *wild card* and is denoted by $*$. Thus, we can express the situation in which an expert has no opinion about the continuation of x by $x*$, and conclude that $N(x) = N(x0) + N(x1) + N(x*)$ and therefore, $p(x) = p(x0) + p(x1) + p(x*)$.

We are interested only in the probabilities $p(x)$ of *definite* sequences (i.e., sequences that contain only 0's and 1's but not the wild card $*$). For these values, the above complicated *additivity* requirement turns into an *inequality*:

$$p(x) \geq p(x0) + p(x1). \tag{2}$$

It is known that other methods of soliciting the degree of belief (e.g., methods that take into consideration not only how voted for and who against, but also to what extent each expert believes in x) also lead to numerical measures that satisfy the inequality (2).

Thus, instead of probability measures, we must consider functions p that maps finite binary strings into real numbers and that satisfy the properties (1) and (2). In Algorithmic Information theory, such functions are called *semi-measures*.

Main results. Using the same arguments as above, we conclude that it is reasonable to require:

- that the function $p(x)$ is *enumerable*, and
- that the *ideal* prior semi-measure $p_0(x)$ must be such that for any other enumerable semi-measure $p(x)$, there exists a constant $c > 0$ for which for all x, $p_0(x) \geq c \cdot p(x)$.

For thus modified definition (with probability measure replaced by semi-measure), it is already possible to prove that the ideal prior semi-measure exists (and that it is, in some reasonable sense, unique; this theorem was first proven by L. Levin; see [16] for a detailed history). The resulting ideal prior semi-measure is usually denoted by $\mathbf{M}(x)$.

Thus, according to the statistical methodology, we must choose, for each n, the sequence $x = x_1 \ldots x_n$ for which $\mathbf{M}(x) \to \max$, or, equivalently, for which $KM(x) \to \min$, where we denoted $KM(x) = -\log_2(\mathbf{M}(x))$.

It turns out that this negative binary logarithm $KM(x) = -\log_2(\mathbf{M}(x))$ is closely related to one of the versions of *Kolmogorov complexity*: namely, to the *monotone complexity* $Km(x)$ that is defined (crudely speaking) as the length of the shortest program that computes a sequence x or a sequence that starts with x on a universal *monotone* Turing machine (i.e., on a computer with a one-way read-only input tape, a one-way write-only output tape (and some work tapes).

The close relationship between $Km(x)$ and $KM(x)$ is described, e.g., by the following two properties:

- $KM(x) \leq Km(x) \leq KM(x) + Km(KM(x)) + O(1)$;
- for almost all infinite sequences $X = x_1 \ldots x_n \ldots$, the difference between $Km(x_1 \ldots x_n)$ and $KM(x_1 \ldots x_n)$ grows slower than any unbounded computable function.

Due to this close relationship, the negative logarithm $KM(x)$ of a universal semi-measure is also considered one of the versions of Kolmogorov complexity.

Conclusion. *From the basis statistical methodology, we can deduce that, if there are several physical theories consistent with the known experimental data, we must always choose the simplest theory x, i.e., a theory for which $KM(x) \to \min$, where $KM(x)$ is the above formalization of complexity.*

Historical comment. This derivation was first proposed by M. Li and P. Vit'anyi; see [14–16,26,27].

Why is this particular formalization of Kolmogorov complexity the most adequate for choosing a theory? Physical explanation. Let us give a physical explanation of why not the *original* Kolmogorov complexity, but its *monotone* version turned out to be more adequate for choosing physical theories. Indeed, the main difference between $Km(x)$ and the original Kolmogorov complexity $C(x)$ is that:

– in the definition of original Kolmogorov complexity, we only consider programs that produce exactly x, while
– in the definition of $Km(x)$, we also consider programs whose output contains x followed by something else.

This difference has a simple physical interpretation: In physics, often, when we want to predict a certain value of the field at some specific moment of time, a natural way is:

– to find a *general* solution of the corresponding system of partial differential equation, and then
– to extract the desired values from the general solution (this "general" solution is, frequently, only applicable to a certain area, e.g., a vacuum electromagnetic field solution only holds outside the bodies).

The resulting algorithm consists of two parts:

– solving the system of equations, and
– extracting the desired values from the general solution.

From the viewpoint of the original Kolmogorov complexity, we have to count the complexity of *both* parts when we estimate the complexity of the solution. The complexity of the first part is indeed indicative of the complexity of the physical world, while the complexity of the second part (extraction) has nothing to do with the physical world and is caused solely by the fact that we are not currently interested in *all* the solution, only in the *part* of it. Therefore, if we are interested in characterizing the physical world itself (and not in our own goals), we should neglect the complexity of the second (extraction) part. This is exactly what $Km(x)$ is doing.

This idea has been successfully used in many applications. In the above derivation, we did not use many specifics of physics as opposed to data processing in general. It will, therefore, not be a surprise to a reader to know that this idea (of choosing the hypothesis for which some version of Kolmogorov complexity takes the smallest possible value) has been originally proposed and successfully used in data processing. Namely, it was first described, under the name of the *minimum description length principle*, by J. J. Rissanen in [20]; for a more recent modifications and updates, see, e.g., [16,21,22].

This idea has been applied to various areas such as handwriting recognition [9], surface reconstruction in computer vision [19], economic forecasts [11], cognitive psychology [2], biology, etc. [16]. These successes make us believe that this formalized principle will be useful in fundamental physics as well.

3. Qualitative physical examples

3.1. MAIN IDEA BEHIND THE EXAMPLES

Our intention. In this section, we will show that, on the *qualitative* level, most fundamental ideas of physics can be re-formulated as natural steps towards choosing a theory that is the simplest in the above precise sense.

Is this intention consistent with our intuition? At first glance, this claim may seem counter-intuitive, because the advance of physics usually means going to more and more complicated theories. For example:

- a simple pre-physical theory that everything in this world is determined by the good and evil spirits and is, therefore, completely unpredictable, a theory that uses no mathematics at all, seems to be very simple,
- while Newtonian mechanics, a theory that uses complicated differential equations, seems to be much more complicated.

This seeming contradiction with the intuition only comes from the fact that the word "simple" is very ambiguous. If we use the word "simple" in the precise way indicated in the previous section, then Newtonian mechanics is no longer simpler that the belief in a completely unpredictable world. Indeed, according to this definition, a complexity of a theory is, crudely speaking, the smallest length of the program that is able to predict (according to this theory) the results of all possible observations and measurements.

- By the intuitive understanding of a completely unpredictable world, the only way for a program to *predict* all the measured values is to actually store them. (If we would be able to store *some* of these values and predict the others, the theory would *not* be completely unpredictable.) Thus, within this theory, we must store *all* the values to predict them. The complexity of this theory is (in the formal sense described in the previous section) the largest possible.
- On the other hand, in Newtonian mechanics, e.g., for particles, we only need to know the positions and velocities of all the particles in the initial moment of time t_0; then we will be able, by integrating the equations of motion, to predict the values at all other moments of time. Thus:
  - instead of storing the values of the particles' coordinates and velocities at *all* possible moments of time,
  - it is sufficient to store these values only for a *single* moment of time.

Thus, when we go from a lawless world to a world described by Newtonian mechanics, we get a drastic *decrease* in the (formally understood) complexity.

Comment. The idea that progress in physics actually leads to theories that are *simpler* in some reasonable sense was emphasized and advocated by Einstein; e.g., he wrote that as "our experience grows larger and larger ... the simpler the logical structure ⟨of the physical theory⟩ becomes – that is to say, the smaller the number of logically independent conceptual elements which are found necessary to support the structure" ([3], p. 273).

How to minimize the complexity: a general idea. One of the major goals of a physical theory is to describe what is happening everywhere in the Universe. In other words, we must describe the values of all possible physical quantities in different points of space at different moments of time.

In the above-mentioned hypothetical completely unpredictable and lawless world, to describe all these values, we would actually need to describe the value of each quantity at each point of space-time. (Since all these values are unlimited and unrelated, if we miss one of these values, we will not be able to reconstruct it

from the other stored ones, and therefore, if we want to be able to reconstruct the state of the Universe, we must actually store *all* the values.) In other words, the Kolmogorov complexity of this description is equal to the total length of all the values stored. How can we decrease the complexity of this description?

In order to answer this question, let us estimate how many bits we need to store the complete information about the lawless world. By using b bits, we can store $V = 2^b$ possible values. Therefore, if a certain physical quantity has V possible values, we must use $\approx \log_2(V)$ bits to store its value. Therefore, if we have T moments of time, S points in space, if we have the total number Q of measurable quantities, and if we have V possible values of each quantity, we need to store $T \cdot S \cdot Q$ different values, each of which requires $\log_2(V)$ bits. Therefore, totally, we need $B = T \cdot S \cdot Q \cdot \log_2(V)$ bits to store all this information. (If one of these numbers is infinite, then, of course, we must store infinitely many bits.)

From the purely *mathematical* viewpoint, to decrease the complexity B of this description, we must, therefore, do one (or several) of the following things:

- decrease the number T of moments of time;
- decrease the number S of possible points in space;
- decrease the number Q of possible quantities;
- decrease the number V of possible values of each quantity, and
- introduce *dependency* between:
  - values of different quantities,
  - values at different moments of time, and/or
  - in different spatial points.

3.2. HOW TO MINIMIZE COMPLEXITY: EXAMPLES FROM FUNDAMENTAL PHYSICS

Let us show that, on the qualitative level, the above possibilities are exactly what major fundamental physical ideas have been achieving.

Restricting T: Big Bang. Restricting the "number of different moments of time" means, in effect, restricting the lifetime of the Universe. Thus, we naturally arrive at the idea of a Universe that has the beginning.

Restricting S: atomism. Similarly, restricting the "number of different spatial points" means that, according to our theory, most of the spatial points inside the Universe have no objects in them, and the matter is concentrated in a few places while others are filled with vacuum. Thus, we get the idea of *atomism*.

Restricting Q: uncertainty principle. Since the total number of *possible* physical quantities is given, the only way to restrict the total number of *measured* quantities is to impose restriction according to which measuring *one* of the properties restricts or even prohibits the measurement of the other quantities.

Restricting V: Special Relativity Theory, quark confinement, quantization. Similarly to the above two methods of decreasing the total number S of spatial points, there are two ways to decrease the total number of possible values V of a physical quantity:

– First, we can *bound* the possible values of this quantity.
  - The best known example of such bounding is the bounding of possible velocities which is the basis of *special relativity*.
  - Another example, also well known but much less fundamental, is the restriction on the relative location of two quarks known as *quark confinement* (see, e.g., [28]).
– Second, we can assume that not *all* values within the bound are possible, but only *some* of these values. This is the original fundamental idea of *quantization*: in quantum mechanics, many quantities can also take values from a certain discrete set (spectrum).

Dependency between different values: general idea. Finally, we can assume some *dependency* between the values of different quantities, and between the values of the same quantity at different moments of time, and in different spatial points.

Dependency between the values of different quantities: symmetry, supersymetry, Dirac's large numbers. Dependency between the values of different components of a multi-component physical field (e.g., a vector, a spinor, or a tensor field) corresponds, crudely speaking, to the fundamental notion of *symmetry* of a physical theory (see, e.g., [7,8,10,12,18,25]).

> The relationship between the Kolmogorov-complexity based principle and symmetries is not surprising to us, because in [13], we used *Kolmogorov complexity* (and related notion of randomness) to *explain* why *symmetries* are a universal language of physics.

Specifically, dependency between components of fields of different types corresponds to the idea of *supersymmetry*.

If we do not have a complete relations between different fields, we can at least get *some* relation, e.g., in terms of equality of some numerical characteristics corresponding to different fields. Such equalities have been discovered since Dirac, in 1937, first observed that the ratio between the electromagnetic and gravitational forces between, say, two protons ($\approx 10^{40}$), is almost equal to the ratio of the Universe's size and the time during which light passes through a proton. This equality is not very easy to explain within specific theories of modern physics (see detailed discussion in [17]), but it is quite in line with our general principle of choosing the simplest theory.

4. Speculations and hopes

In this paper, we have shown two things:

– First, we have described, for a physics reader, how the principle of choosing the simplest physical theory, the idea that is usually formulated in *informal, qualitative* terms, can be reformulated in *precise, quantitative* terms; this formalization has been successfully used in various areas of data processing.
– Second, we have shown that this formalization explains, on the *qualitative* level, many fundamental principles of physics.

The success of formalizing the principle itself makes us hope that it may be possible to formalize these *explanations* as well, and thus, come out with a new universal physical theory in which the choice of the simplest hypothesis (or, equivalently, algorithmic information theory) is the sole axiom.

This possibility is in good accordance with the vision of Einstein who said ([3], p. 274) that "Our experience hitherto justifies us in believing that nature is the realization of the simplest conceivable mathematical ideas."

This hope of ours, that *physics* may be eventually reduced to *algorithmic* information theory, is also in clear agreement with the physically motivated ideas of J. A. Wheeler of logic as pre-geometry and pre-physics (see, e.g., [17] and references therein).

Acknowledgment

This work was supported in part by NASA under cooperative agreement NCCW-0089. It was also partially supported by NSF under grants No. DUE-9750858 and EEC-9322370, 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.

The authors are thankful to all participants of the MaxEnt'97 Workshop, especially to Peter Cheeseman, Anton Garrett, Steve Gull, and John Skilling, for important discussions.

References

1. L. Brink and M. Henneaux, *Principles of string theory*, Plenum Press, N.Y., 1988.
2. N, Chater, "Reconciling simplicity and likelihood principles in perceptual organization", *Psychological Reviews*, 1996, Vol. 103, pp. 566–581.
3. A. Einstein, "On the method of theoretical physics", The Herbert Spencer Lecture delivered at Oxford on June 10, 1933. Reprinted in: A. Einstein, *Ideas and opinions*, Crown Publishers, N.Y., 1954, pp. 270–276.
4. A. Finkelstein, O. Kosheleva, and V. Kreinovich, "Astrogeometry, error estimation, and other applications of set-valued analysis", *ACM SIGNUM Newsletter*, 1996, Vol. 31, No. 4, pp. 3–25.
5. A. Finkelstein, O. Kosheleva, and V. Kreinovich, "Astrogeometry: towards mathematical foundations", *International Journal of Theoretical Physics*, 1997, Vol. 36, No. 4, pp. 1009–1020.
6. A. Finkelstein, O. Kosheleva, and V. Kreinovich, "Astrogeometry: geometry explains shapes of celestial bodies", *Geombinatorics*, 1997, Vol. VI, No. 4, pp. 125–139.
7. A. M. Finkelstein and V. Kreinovich, "Derivation of Einstein's, Brans-Dicke and other equations from group considerations," *On Relativity Theory. Proceedings of the Sir Arthur Eddington Centenary Symposium, Nagpur India 1984*, Vol. 2, Y. Choque-Bruhat and T. M. Karade (eds), World Scientific, Singapore, 1985, pp. 138–146.
8. A. M. Finkelstein, V. Kreinovich, and R. R. Zapatrin, "Fundamental physical equations uniquely determined by their symmetry groups," *Lecture Notes in Mathematics*, Springer-Verlag, Berlin-Heidelberg-N.Y., Vol. 1214, 1986, pp. 159–170.
9. Q. Gao and M. Li, "An application of minimum description length principle to online recognition of handprinted alphanumerals", In: *Proc. 11th Int's Join Conferences on Artificial Intelligence IJCAI*, Morgan Kaufmann, San Mateo, CA, 1989, pp. 843–848.

10. *Group theory in physics: proceedings of the international symposium held in honor of Prof. Marcos Moshinsky, Cocoyoc, Morelos, Mexico, 1991*, American Institute of Physics, N.Y., 1992.
11. H. A. Keuzenkamp and M. McAleer, "Simplicity, scientific inference, and econometric modelling", *The Economic Journal*, 1995, Vol. 105, pp. 1–21.
12. V. Kreinovich. "Derivation of the Schroedinger equations from scale invariance," *Theoretical and Mathematical Physics*, 1976, Vol. 8, No. 3, pp. 282–285.
13. V. Kreinovich and L. Longpré, "Unreasonable effectiveness of symmetry in physics", *International Journal of Theoretical Physics*, 1996, Vol. 35, No. 7, pp. 1549–1555.
14. M. Li and P. M. B. Vitányi, "Inductive reasoning and Kolmogorov complexity", *J. Comput. System. Sci.*, 1992, Vol. 44, No. 2, pp. 343–384.
15. M. Li and P. M. B. Vitányi, "Computational machine learning in theory and practice", In: J. van Leeuwen (ed.), *Computer Science Today, Recent Trends and Developments*, Springer Lecture Notes in Computer Science, Springer-Verlag, Berlin-Heidelberg-N.Y., 1995, Vol. 1000, pp. 518–535.
16. M. Li and P. M. B. Vitányi, *An Introduction to Kolmogorov Complexity and its Applications*, Springer-Verlag, N.Y., 1997.
17. Ch. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation*, W. H. Freeman and Co., San Francisco, 1973.
18. P. J. Olver, *Equivalence, invariants, and symmetry*, Cambridge University Press, Cambridge, N.Y., 1995.
19. E. P. D. Pednault, "Some experiments in applying inductive inference principles to surface reconstruction", In: *Proc. 11th International Joint Conferences on Artificial Intelligence IJCAI*, Morgan Kaufmann, San Mateo, CA, 1989, pp. 1603–1609.
20. J. J. Rissanen, "Modeling by the shortest data description", *Automatica*, 1978, Vol. 14, pp. 465–471.
21. J. J. Rissanen, *Stochastic complexity and statistical inquiry*, World Scientific, Singapore, 1989.
22. J. J. Rissanen, "Fisher information and stochastic complexity", *IEEE Transactions on Information Theory*, 1996, Vol. IT–42, No. 1, pp. 40–47.
23. I. Rosenthal-Schneider, "Presuppositions and anticipations", In: P. A. Schlipp (ed.), *Albert Einstein: philosopher-scientist*, Tudor Publ., N.Y., 1951.
24. W. Siegel, *Introduction to string field theory*, World Scientific, Singapore, 1988.
25. *Symmetries in physics: proceedings of the international symposium held in honor of Prof. Marcos Moshinsky, Cocoyoc, Morelos, Mexico, 1991*, Springer-Verlag, Berlin, N.Y., 1992.
26. P. M. B. Vitányi and M. Li, "Ideal MDL and its relation to Bayesianism", In: D. Dowe, K. Korb, and J. Oliver (eds.), *Proc. ISIS: Information, Statistics, and Induction in Science Conference*, World Scientific, Singapore, 1996, pp. 282–291.
27. P. M. B. Vitányi and M. Li, *Minimum description length induction, Bayesianism, and Kolmogorov complexity*, Manuscript, CWI, Amsterdam, 1996.
28. F. J. Yndurain, *Quantum chromodynamics: an introduction to the theory of quarks and gluons*, Springer-Verlag, N.Y., 1983.