Towards an Algorithmic Statistics

(Extended Abstract)

Peter Gács*, John Tromp, and Paul Vitányi**

Abstract. While Kolmogorov complexity is the accepted absolute measure of information content of an individual finite object, a similarly absolute notion is needed for the relation between an individual data sample and an individual model summarizing the information in the data, for example, a finite set where the data sample typically came from. The statistical theory based on such relations between individual objects can be called algorithmic statistics, in contrast to ordinary statistical theory that deals with relations between probabilistic ensembles. We develop a new algorithmic theory of typical statistic, sufficient statistic, and minimal sufficient statistic.

1 Introduction

We take statistical theory to ideally consider the following problem: Given a data sample and a family of models (hypotheses) one wants to select the model that produced the data. But a priori it is possible that the data is atypical for the model that actually produced it, or that the true model is not present in the considered model class. Therefore we have to relax our requirements. If selection of a "true" model cannot be guarantied by any method, then as next best choice "modeling the data" as well as possible, irrespective of truth and falsehood of the resulting model, may be more appropriate. Thus, we change 'true" to "as well as possible." The latter we take to mean that the model expresses all significant regularities present in the data.

Probabilistic Statistics: In ordinary statistical theory one proceeds as follows, see for example [3]: Suppose two random variables X, Y have a joint probability mass function $p(x, y)$ and marginal probability mass functions $p(x)$ and $p(y)$. Then the (probabilistic) *mutual information* $I(X; Y)$ is the relative entropy between the joint distribution and the product distribution $p(x)p(y)$:

$$I(X;Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}. \tag{1}$$

Every function $T(D)$ of a data sample D—like the sample mean or the sample variance—is called a *statistic* of D. Assume we have a probabilistic ensemble of

* Address: Computer Science Department, Boston University, Boston MA 02215, U.S.A. Email: gacs@bu.edu. The paper was partly written during this author's visit at CWI.

** Address: CWI, Kruislaan 413, 1098 SJ Amsterdam, The Netherlands. Email: {tromp, paulv}@cwi.nl

models, say a family of probability mass functions $\{f_\theta\}$ indexed by θ, together with a distribution over θ. A statistic $T(D)$ is called *sufficient* if the probabilistic mutual information

$$I(\theta; D) = I(\theta; T(D)) \tag{2}$$

for all distributions of θ. Hence, the mutual information between parameter and data sample is invariant under taking sufficient statistics and vice versa. That is to say, a statistic $T(D)$ is called sufficient for θ if it contains all the information in D about θ. For example, consider n tosses of a coin with unknown bias θ with outcome $D = d_1 d_2 \ldots d_n$ where $d_i \in \{0,1\}$ $(1 \leq i \leq n)$. Given n, the number of outcomes "1" is a sufficient statistic for θ: the statistic $T(D) = \sum_{i=1}^n d_i$. Given T, every sequence with $T(D)$ "1"s are equally likely independent of parameter θ: Given k, if D is an outcome of n coin tosses and $T(D) = k$ then $\Pr(D \mid T(D) = k) = \binom{n}{k}^{-1}$ and $\Pr(D \mid T(D) \neq k) = 0$. This can be shown to imply (2) and therefore T is a sufficient statistic for θ. According to Fisher [4]: "The statistic chosen should summarise the whole of the relevant information supplied by the sample. This may be called the Criterion of Sufficiency ... In the case of the normal curve of distribution it is evident that the second moment is a sufficient statistic for estimating the standard deviation." Note that one cannot improve on sufficiency: for every (possibly randomized) function T we have

$$I(\theta; D) \geq I(\theta; T(D)), \tag{3}$$

that is, mutual information cannot be increased by processing the data sample in any way. All these notions and laws are probabilistic: they hold in an average sense. Our program is to develop a sharper theory, which we call *algorithmic* statistics to distinguish it from the standard *probabilistic* statistics, where the notions and laws hold in the individual sense.

Algorithmic Statistics: In algorithmic statistics, one wants to select an individual model (described by, say, a finite set) for which the data is individually typical. To express the notion "individually typical" one requires Kolmogorov complexity—standard probability theory cannot express this. The basic idea is as follows: In a two-part description, we first describe such a model, a finite set, and then indicate the data within the finite set by its index in a natural ordering of the set. The optimal models make the two-part description as concise as the shortest one-part description of the data. Moreover, for such optimal two-part descriptions it can be shown that the data will be "individually typical" for the model concerned. A description of such a model is an algorithmic sufficient statistic since it summarizes all relevant properties of the data. Among the algorithmic sufficient statistics a simplest one (the algorithmic minimal sufficient statistic) is best in accordance with Ockham's razor principle since it summarizes the relevant properties of the data as concisely as possible. In probabilistic data or data subject to noise this involves separating regularities (structure) in the data from random effects.

Background and Related Work: At a Tallinn conference in 1973, A.N. Kolmogorov formulated this task rigorously in terms of Kolmogorov complexity (according to [14, 2]). This approach can also be viewed as a two-part code separating the *structure* of a string from meaningless *random* features. Cover [2, 3] interpreted this approach as (sufficient) statistic. Related aspects of "randomness deficiency" (formally defined later in (11)) were formulated in [9, 10] and studied in [14, 17]. Algorithmic mutual information, and the associated non-increase law, were studied in [11, 12]. Despite its evident epistimological prominence in the theory of hypothesis selection and prediction, only some scattered aspects of the subject have been studied before, for example as related to the "Kolmogorov structure function" [14, 2], and "absolutely non-stochastic objects" [14, 17, 15, 18], notions also defined or suggested by Kolmogorov at the mentioned meeting. For the relation with inductive reasoning according to minimum description length principle see [16]. The entire approach is based on Kolmogorov complexity [8] (also known as algorithmic information theory). For a general introduction to Kolmogorov complexity, its mathematical theory, and application to induction see [7].

Results: We develop the outlines of a new general mathematical theory of algorithmic statistics, in this initial approach restricted to models that are finite sets. A set S is "optimal" if the best two-part description consisting of a description of S and a straightforward description of x as an element of S by an index of size $\log |S|$, is as concise as the shortest one-part description of x. Descriptions of such optimal sets are algorithmic sufficient statistics, and the shortest description among them is an algorithmic minimal sufficient statistic. The mode of description plays a major role in this. We distinguish between "explicit" descriptions and "implicit" descriptions—that are introduced in this paper as a proper restriction on recursive enumeration based description mode. We establish new precise range constraints of cardinality and complexity imposed by implicit (and hence explicit) descriptions for typical and optimal sets, and exhibit for the first time concrete algorithmic minimal (or near-minimal) sufficient statistics for both description modes. There exist maximally complex objects for which no finite set of less complexity is an explicit sufficient statistic—such objects are absolutely non-stochastic. This improves a result of Shen [14] to the best possible.

Application: In all practicable inference methods, one must use background information to determine the appropriate model class first—establishing what meaning the data can have—and only then obtain the best model in that class by optimizing its parameters. For example in the "probably approximately correct (PAC)" learning criterion one learns a concept in a given concept class (like a class of Boolean formulas over n variables); in the "minimum description length (MDL)" induction, [1], one first determines the model class (like Bernoulli processes). Note that MDL has been shown to be a certain generalization of the (Kolmogorov) minimum sufficient statistic in [16].

To develop the onset of a theory of algorithmic statistics we have used the mathematically convenient model class consisting of the finite sets. An illustration of background information is Example 3. An example of selecting a model parameter on the basis of compression properties is the precision at which we represent the other parameters: too high precision causes accidental noise to be modeled as well, too low precision may cause models that should be distinct to be confused. In general, the performance of a model for a given data sample depends critically on what we may call the "degree of discretization" or the "granularity" of the model: the choice of precision of the parameters, the number of nodes in the hidden layer of a neural network, and so on. The granularity is often determined ad hoc. In [5], in two quite different experimental settings the MDL predicted best model granularity values are shown to coincide with the best values found experimentally.

2 Kolmogorov Complexity

We assume familiarity with the elementary theory of Kolmogorov complexity. For introduction, details, and proofs, see [7]. We write *string* to mean a finite binary string. Other finite objects can be encoded into strings in natural ways. The set of strings is denoted by $\{0,1\}^*$. The *length* of a string x is denoted by $l(x)$, distinguishing it from the *cardinality* $|S|$ of a finite set S. The (prefix) Kolmogorov complexity, or algorithmic entropy, $K(x)$ of a string x is the length of a shortest binary program to compute x on a universal computer (such as a universal Turing machine). Intuitively, $K(x)$ represents the minimal amount of information required to generate x by any effective process, [8]. We denote the *shortest program* for x by x^*; then $K(x) = l(x^*)$. (Actually, x^* is the first shortest program for x in an appropriate standard enumeration of all programs for x such as the halting order.) The conditional Kolmogorov complexity $K(x \mid y)$ of x relative to y is defined similarly as the length of a shortest program to compute x if y is furnished as an auxiliary input to the computation.

From now on, we will denote by $\stackrel{+}{<}$ an inequality to within an additive constant, and by $\stackrel{+}{=}$ the situation when both $\stackrel{+}{<}$ and $\stackrel{+}{>}$ hold. We will also use $\stackrel{*}{<}$ to denote an inequality to within an multiplicative constant factor, and $\stackrel{*}{=}$ to denote the situation when both $\stackrel{*}{<}$ and $\stackrel{*}{>}$ hold.

We will use the "Additivity of Complexity" (Theorem 3.9.1 of [7]) property (by definition $K(x,y) = K(\langle x,y\rangle)$):

$$K(x,y) \stackrel{+}{=} K(x) + K(y \mid x^*) \stackrel{+}{=} K(y) + K(x \mid y^*). \tag{4}$$

The conditional version needs to be treated carefully. It is

$$K(x,y \mid z) \stackrel{+}{=} K(x \mid z) + K(y \mid x, K(x \mid z), z). \tag{5}$$

Note that a naive version

$$K(x,y \mid z) \stackrel{+}{=} K(x \mid z) + K(y \mid x^*, z)$$

is incorrect: taking $z = x$, $y = K(x)$, the left-hand side equals $K(x^* \mid x)$, and the right-hand side equals $K(x \mid x) + K(K(x) \mid x^*, x) \stackrel{+}{=} 0$.

We derive a (to our knowledge) new "directed triangle inequality" that is needed below.

Theorem 1. *For all x, y, z,*

$$K(x \mid y^*) \stackrel{+}{<} K(x, z \mid y^*) \stackrel{+}{<} K(z \mid y^*) + K(x \mid z^*).$$

Proof. Using (4), an evident inequality introducing an auxiliary object z, and twice (4) again:

$$\begin{aligned} K(x, z \mid y^*) &\stackrel{+}{=} K(x, y, z) - K(y) \stackrel{+}{<} K(z) + K(x \mid z^*) + K(y \mid z^*) - K(y) \\ &\stackrel{+}{=} K(y, z) - K(y) + K(x \mid z^*) \stackrel{+}{=} K(x \mid z^*) + K(z \mid y^*). \end{aligned}$$

□

This theorem has bizarre consequences. Denote $k = K(y)$ and substitute $k = z$ and $K(k) = x$ to find the following counterintuitive corollary:

Corollary 1. $K(K(k) \mid y, k) \stackrel{+}{=} K(K(k) \mid y^*) \stackrel{+}{<} K(K(k) \mid k^*) + K(k \mid y, k) \stackrel{+}{=} 0$. *We can iterate this: given y and $K(y)$ we can determine $K(K(K(y)))$ in $O(1)$ bits. So $K(K(K(k))) \mid y, k) \stackrel{+}{=} 0$ and so on.*

If we want to find an appropriate model fitting the data, then we are concerned with the information in the data about such models. To define the algorithmic mutual information between two individual objects x and y with no probabilities involved, rewrite (1) as

$$\sum_x \sum_y p(x, y)[-\log p(x) - \log p(y) + \log p(x, y)],$$

and note that $-\log p(s)$ is the length of the prefix-free Shannon-Fano code for s. Consider $-\log p(x) - \log p(y) + \log p(x, y)$ over the individual x, y, and replace the Shannon-Fano code by the "shortest effective description" code. [1] The *information in y about x* is defined as

$$I(y : x) = K(x) - K(x \mid y^*) \stackrel{+}{=} K(x) + K(y) - K(x, y), \tag{6}$$

where the second equality is a consequence of (4) and states the celebrated result that the information between two individual objects is symmetrical, $I(x : y) \stackrel{+}{=} I(y : x)$, and therefore we talk about *mutual information*.[2] In the full paper [6] we show that the expectation of the algorithmic mutual information $I(x : y)$ is close the the probabilistic mutual information $I(x; y)$—which corroborates that

[1] The Shannon-Fano code has optimal expected code length equal to the entropy with respect to the distribution of the source [3]. However, the prefix-free code of shortest effective description, that achieves code word length $K(s)$ for source word s, has both about expected optimal code word length and individual optimal effective code word length, [7].

[2] The notation of the algorithmic (individual) notion $I(x : y)$ distinguishes it from the probabilistic (average) notion $I(x; y)$. We deviate slightly from [7] where $I(y : x)$ is defined as $K(x) - K(x \mid y)$.

the algorithmic notion is a sharpening of the probabilistic notion to individual objects.

The mutual information between a pair of strings x and y cannot be increased by processing x and y separately by some deterministic computations, and furthermore, randomized computation can increase the mutual information only with negligible probability, [11, 12]. Since the first reference gives no proofs and the second reference is not easily accessible, in the full version of this paper [6] we use the triangle inequality of Theorem 1 to give new simple proofs of this information non-increase.

3 Algorithmic Model Development

In this initial investigation, we use for mathematical convenience the *model class* consisting of the family of finite sets of finite binary strings, that is, the set of subsets of $\{0,1\}^*$.

3.1 Finite Set Representations

Although all finite sets are recursive there are different ways to represent or specify the set. We only consider ways that have in common a method of recursively enumerating the elements of the finite set one by one, and which differ in knowledge of its size. For example, we can specify a set of natural numbers by giving an explicit table or a decision procedure for membership and a bound on the largest element, or by giving a recursive enumeration of the elements together with the number of elements, or by giving a recursive enumeration of the elements together with a bound on the running time. We call a representation of a finite set S *explicit* if the size $|S|$ of the finite set can be computed from it. A representation of S is *implicit* if the size $|S|$ can be computed from it only up to a factor of 2.

Example 1. In Section 3.4, we will introduce the set S^k of strings whose elements have complexity $\leq k$. It will be shown that this set can be represented implicitly by a program of size $K(k)$, but can be represented explicitly only by a program of size k.

Such representations are useful in two-stage encodings where one stage of the code consists of an index in S of length $\stackrel{+}{=} \log |S|$. In the implicit case we know, within an additive constant, how long an index of an element in the set is. In general S^* denotes the shortest binary program from which S can be computed and whether this is an implicit or explicit description will be clear from the context.

The worst case, a recursively enumerable representation where nothing is known about the size of the finite set, would lead to indices of unknown length. We do not consider this case. We may use the notation

$$S_{\text{impl}}, S_{\text{expl}}$$

for some implicit and some explicit representation of S. When a result applies to both implicit and explicit representations, or when it is clear from the context which representation is meant, we will omit the subscript.

3.2 Optimal Models and Sufficient Statistics

In the following we will distinguish between "models" that are finite sets, and the "shortest programs" to compute those models that are finite strings. Such a shortest program is in the proper sense a statistics of the data sample as defined before. In a way this distinction between "model" and "statistics" is artificial, but for now we prefer clarity and unambiguousness in the discussion.

Consider a string x of length n and prefix complexity $K(x) = k$. We identify the *structure* or *regularities* in x that are to be summarized with a set S of which x is a *random* or *typical* member: given S (or rather, an (implicit or explicit) shortest program S^* for S), x cannot be described much shorter than by its maximal length index in S. Formally this is expressed by $K(x \mid S^*) \stackrel{+}{>} \log|S|$. More formally, we fix some constant

$$\beta \geq 0,$$

and require $K(x \mid S^*) \geq \log|S| - \beta$. We will not indicate the dependence on β explicitly, but the constants in all our inequalities ($\stackrel{+}{<}$) will be allowed to be functions of this β. This definition requires a finite S. In fact, since $K(x \mid S^*) \stackrel{+}{<} K(x)$, it limits the size of S to $O(2^k)$ and a set S (rather, the shortest program S^* from which it can be computed) is a *typical statistic* for x iff

$$K(x \mid S^*) \stackrel{\pm}{=} \log|S|. \tag{7}$$

Depending on whether S^* is an implicit or explicit program, our definition splits into implicit and explicit typicality.

Example 2. Consider the set S of binary strings of length n whose every odd position is 0. Let x be element of this set in which the subsequence of bits in even positions is an incompressible string. Then S is explicitly as well as implicitly typical for x. The set $\{x\}$ also has both these properties.

Remark 1. It is not clear whether explicit typicality implies implicit typicality. Section 4 will show some examples which are implicitly very non-typical but explicitly at least nearly typical.

There are two natural measures of suitability of such a statistic. We might prefer either the simplest set, or the largest set, as corresponding to the most likely structure 'explaining' x. The singleton set $\{x\}$, while certainly a typical statistic for x, would indeed be considered a poor explanation. Both measures relate to the optimality of a two-stage description of x using S:

$$K(x) \leq K(x, S) \stackrel{\pm}{=} K(S) + K(x \mid S^*) \stackrel{+}{<} K(S) + \log|S|, \tag{8}$$

where we rewrite $K(x,S)$ by (4). Here, S can be understood as either S_{impl} or S_{expl}. Call a set S (containing x) for which

$$K(x) \stackrel{\pm}{=} K(S) + \log |S|, \tag{9}$$

optimal. (More precisely, we should require $K(x) \geq K(S) + \log |S| - \beta$.) Depending on whether $K(S)$ is understood as $K(S_{\text{impl}})$ or $K(S_{\text{expl}})$, our definition splits into implicit and explicit optimality. The shortest program for an optimal set is a *algorithmic sufficient statistic* for x [3]. Furthermore, among optimal sets, there is a direct trade-off between complexity and logsize, which together sum to $\stackrel{\pm}{=} k$. Equality (9) is the algorithmic equivalent dealing with the relation between the individual sufficient statistic and the individual data sample, in contrast to the probabilistic notion (2).

Example 3. The following restricted model family illustrates the difference between the algorithmic individual notion of sufficient statistics and the probabilistic averaging one. Following the discussion in section 1, this example also illustrates the idea that the semantics of the model class should be obtained by a restriction on the family of allowable models, after which the (minimal) sufficient statistics identifies the most appropriate model in the allowable family and thus optimizes the parameters in the selected model class. In the algorithmic setting we use all subsets of $\{0,1\}^n$ as models and the shortest programs computing them from a given data sample as the statistics. Suppose we have background information constraining the family of models to the $n+1$ finite sets $S_k = \{x \in \{0,1\}^n : x = x_1 \dots x_n \& \sum_{i=1}^n x_i = k\}$ $(0 \leq k \leq n)$. Then, in the probabilistic sense for every data sample $x = x_1 \dots x_n$ there is only one single sufficient statistics: for $\sum_i x_i = k$ this is $T(x) = k$ with the corresponding model S_k. In the algorithmic setting the situation is more subtle. (In the following example we use the complexities conditional n.) For $x = x_1 \dots x_n$ with $\sum_i x_i = \frac{n}{2}$ taking $S_{\frac{n}{2}}$ as model yields $|S_{\frac{n}{2}}| = \binom{n}{\frac{n}{2}}$, and therefore $\log |S_{\frac{n}{2}}| \stackrel{\pm}{=} n - \frac{1}{2}\log n$. The sum of $K(S_{\frac{n}{2}}|n) \stackrel{\pm}{=} 0$ and the logarithmic term gives $\stackrel{\pm}{=} n - \frac{1}{2}\log n$ for the right-hand side of (9). But taking $x = 1010\dots10$ yields $K(x|n) \stackrel{\pm}{=} 0$ for the left-hand side. Thus, there is no algorithmic sufficient statistics for the latter x in this model class, while every x of length n has a probabilistic sufficient statistics in the model class. In fact, the restricted model class has algorithmic sufficient statistics for data samples x of length n that have maximal complexity with respect to the frequency of "1"s, the other data samples have no algorithmic sufficient statistics in this model class.

Example 4. It can be shown that the set S of Example 2 is also optimal, and so is $\{x\}$. Typical sets form a much wider class than optimal ones: $\{x,y\}$ is still typical for x but with most y, it will be too complex to be optimal for x.

For a perhaps less artificial example, consider complexities conditional to the length n of strings. Let y be a random string of length n, let S_y be the set of strings of length n which have 0's exactly where y has, and let x be a random element of S_y. Then x is a string random with respect to the distribution in which 1's are chosen independently with probability 0.25, so its complexity is much less than n. The set S_y is typical with respect to x but is too complex to be optimal, since its (explicit or implicit) complexity conditional to n is n.

It follows that (programs for) optimal sets are typical statistics. Equality (9) expresses the conditions on the algorithmic individual relation between the data and the sufficient statistic. Later we demonstrate that this relation implies that the probabilistic optimality of mutual information (1) holds for the algorithmic version in the expected sense.

One can also consider notions of *near*-typical and *near*-optimal that arise from replacing the β above by some slow growing functions, such as $O(\log l(x))$ or $O(\log k)$ as in [14, 15].

3.3 Properties of Sufficient Statistics

We start with a sequence of lemmas that will be used in the later theorems. Several of these lemmas have two versions: for implicit and for explicit sets. In these cases, S will denote S_{impl} or S_{expl} respectively.

Below it is shown that the mutual information between every typical set and the datum is not much less than $K(K(x))$, the complexity of the complexity $K(x)$ of the datum x. For optimal sets it is at least that, and for algorithmic minimal statistic it is equal to that. The number of elements of a typical set is determined by the following:

Lemma 1. *Let $k = K(x)$. If a set S is (implicitly or explicitly) typical for x then $I(x : S) \stackrel{+}{=} k - \log |S|$.*

Proof. By definition $I(x : S) \stackrel{+}{=} K(x) - K(x \mid S^*)$ and by typicality $K(x \mid S^*) \stackrel{+}{=} \log |S|$. □

Typicality, optimality, and minimal optimality successively restrict the range of the cardinality (and complexity) of a corresponding model for a datum x. The above lemma states that for (implicitly or explicitly) typical S the cardinality $|S| = \Theta(2^{k-I(x:S)})$. The next lemma asserts that for implicitly typical S the value $I(x : S)$ can fall below $K(k)$ by no more than an additive logarithmic term.

Lemma 2. *Let $k = K(x)$. If a set S is (implicitly or explicitly) typical for x then $I(x : S) \stackrel{+}{>} K(k) - K(I(x : S))$ and $\log |S| \stackrel{+}{<} k - K(k) + K(I(x : S))$. (Here, S is understood as S_{impl} or S_{expl} respectively.)*

Proof. Writing $k = K(x)$, since

$$k \stackrel{+}{=} K(k, x) \stackrel{+}{=} K(k) + K(x \mid k^*) \tag{10}$$

by (4), we have $I(x : S) \stackrel{+}{=} K(x) - K(x \mid S^*) \stackrel{+}{=} K(k) - [K(x \mid S^*) - K(x \mid k^*)]$. Hence, it suffices to show $K(x \mid S^*) - K(x \mid k^*) \stackrel{+}{<} K(I(x : S))$. Now, from an implicit description S^* we can find $\stackrel{+}{=} \log|S| \stackrel{+}{=} k - I(x:S)$ and to recover k we only require an extra $K(I(x:S))$ bits apart from S^*. Therefore, $K(k \mid S^*) \stackrel{+}{<} K(I(x:S))$. This reduces what we have to show to $K(x \mid S^*) \stackrel{+}{<} K(x \mid k^*) + K(k \mid S^*)$ which is asserted by Theorem 1.

□

The term $I(x:S)$ is at least $K(k) - 2\log K(k)$ where $k = K(x)$. For x of length n with $k \stackrel{+}{>} n$ and $K(k) \stackrel{+}{>} l(k) \stackrel{+}{>} \log n$, this yields $I(x:S) \stackrel{+}{>} \log n - 2\log\log n$.

If we further restrict typical sets to optimal sets then the possible number of elements in S is slightly restricted. First we show that implicit optimality of a set with respect to a datum is equivalent to typicality with respect to the datum combined with effective constructability (determination) from the datum.

Lemma 3. *A set S is (implicitly or explicitly) optimal for x iff it is typical and $K(S \mid x^*) \stackrel{+}{=} 0$.*

Proof. A set S is optimal iff (8) holds with equalities. Rewriting $K(x,S) \stackrel{+}{=} K(x) + K(S \mid x^*)$ the first inequality becomes an equality iff $K(S \mid x^*) \stackrel{+}{=} 0$, and the second inequality becomes an equality iff $K(x \mid S^*) \stackrel{+}{=} \log|S|$ (that is, S is a typical set). □

Lemma 4. *Let $k = K(x)$. If a set S is (implicitly or explicitly) optimal for x, then $I(x:S) \stackrel{+}{=} K(S) \stackrel{+}{>} K(k)$ and $\log|S| \stackrel{+}{<} k - K(k)$.*

Proof. If S is optimal for x, then $k = K(x) \stackrel{+}{=} K(S) + K(x \mid S^*) \stackrel{+}{=} K(S) + \log|S|$. From S^* we can find both $K(S) \stackrel{+}{=} l(S^*)$ and $|S|$ and hence k, that is, $K(k) \stackrel{+}{<} K(S)$. We have $I(x:S) \stackrel{+}{=} K(S) - K(S \mid x^*) \stackrel{+}{=} K(S)$ by (4), Lemma 3, respectively. This proves the first property. Substitution of $I(x:S) \stackrel{+}{>} K(k)$ in the expression of Lemma 1 proves the second property. □

3.4 A Concrete Implicit Minimal Sufficient Statistic

A simplest implicitly optimal set (that is, of least complexity) is an implicit algorithmic minimal sufficient statistic. We demonstrate that $S^k = \{y : K(y) \leq k\}$, the set of all strings of complexity at most k, is such a set. First we establish the cardinality of S^k:

Lemma 5. $\log|S^k| \stackrel{+}{=} k - K(k)$.

Proof. The lower bound is easiest. Denote by k^* of length $K(k)$ a shortest program for k. Every string s of length $k - K(k) - c$ can be described in a self-delimiting manner by prefixing it with k^*c^*, hence $K(s) \stackrel{+}{<} k - c + 2\log c$. For a large enough constant c, we have $K(s) \leq k$ and hence there are $\Omega(2^{k-K(k)})$ strings that are in S^k.

For the upper bound: by (10) all $x \in S^k$ satisfy $K(x \mid k^*) \stackrel{+}{<} k - K(k)$ and there can only be $O(2^{k-K(k)})$ of them. □

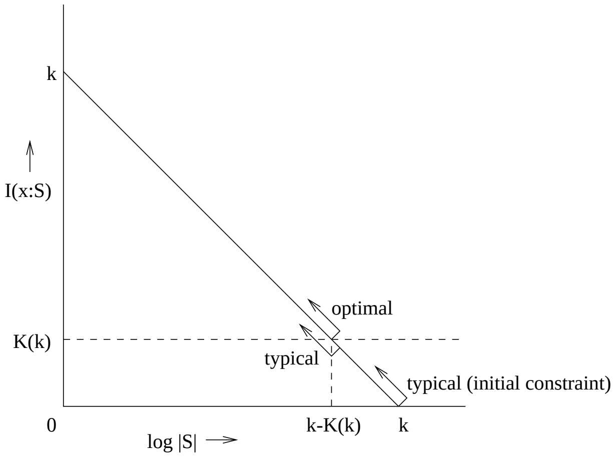

Fig. 1. Range of typical statistics on the straight line $I(x:S) \stackrel{+}{=} K(x) - \log |S|$.

From the definition of S^k it follows that it is defined by k alone, and it is the same set that is optimal for all objects of the same complexity k.

Theorem 2. *The set S^k is implicitly optimal for every x with $K(x) = k$. Also, we have $K(S^k) \stackrel{+}{=} K(k)$.*

Proof. From k^* we can compute both k and $k - l(k^*) = k - K(k)$ and recursively enumerate S^k. Since also $\log |S^k| \stackrel{+}{=} k - K(k)$ (Lemma 5), the string k^* plus a fixed program is an implicit description of S^k so that $K(k) \stackrel{+}{>} K(S^k)$. Hence, $K(x) \stackrel{+}{>} K(S^k) + \log |S^k|$ and since $K(x)$ is the shortest description by definition equality ($\stackrel{+}{=}$) holds. That is, S^k is optimal for x. By Lemma 4 $K(S^k) \stackrel{+}{>} K(k)$ which together with the reverse inequality above yields $K(S^k) \stackrel{+}{=} K(k)$ which shows the theorem. □

Again using Lemma 4 shows that the optimal set S^k has least complexity among all optimal sets for x, and therefore:

Corollary 2. *The set S^k is an implicit algorithmic minimal sufficient statistic for every x with $K(x) = k$.*

All algorithmic minimal sufficient statistics S for x have $K(S) \stackrel{+}{=} K(k)$, and therefore there are $O(2^{K(k)})$ of them. At least one such a statistic (S^k) is associated with every one of the $O(2^k)$ strings x of complexity k. Thus, while the idea of the algorithmic minimal sufficient statistic is intuitively appealing, its unrestricted use doesn't seem to uncover most relevant aspects of reality. The only relevant structure in the data with respect to a algorithmic minimal sufficient statistic is the Kolmogorov complexity. To give an example, an initial segment of $3.1415\ldots$ of length n of complexity $\log n + O(1)$ shares the same algorithmic sufficient statistic with many (most?) binary strings of length $\log n + O(1)$.

3.5 A Concrete Explicit Minimal Sufficient Statistic

Let us now consider representations of finite sets that are explicit in the sense that we can compute the cardinality of the set from the representation. For example, the description program enumerates all the elements of the set and halts. Then a set like $S^k = \{y : K(y) \leq k\}$ has complexity $\stackrel{+}{=} k$ [15]: Given the program we can find an element not in S^k, which element by definition has complexity $> k$. Given S^k we can find this element and hence S^k has complexity $\stackrel{+}{>} k$. Let

$$N^k = |S^k|,$$

then by Lemma 5 $\log N^k \stackrel{+}{=} k - K(k)$. We can list S^k given k^* and N^k which shows $K(S^k) \stackrel{+}{<} k$.

One way of implementing explicit finite representations is to provide an explicit generation time for the enumeration process. If we can generate S^k in time t recursively using k, then the previous argument shows that the complexity of every number $t' \geq t$ satisfies $K(t', k) \geq k$ so that $K(t') \stackrel{+}{>} K(t' \mid k^*) \stackrel{+}{>} k - K(k)$ by (4). This means that t is a huge time which as a function of k rises faster than every computable function. This argument also shows that explicit enumerative descriptions of sets S containing x by an enumerative process p plus a limit on the computation time t may take only $l(p) + K(t)$ bits (with $K(t) \leq \log t + 2 \log\log t$) but $\log t$ unfortunately becomes noncomputably large!

In other cases the generation time is simply recursive in the input: $S_n = \{y : l(y) \leq n\}$ so that $K(S_n) \stackrel{+}{=} K(n) \leq \log n + 2 \log\log n$. That is, this typical sufficient statistic for a random string x with $K(x) \stackrel{+}{=} n + K(n)$ has complexity $K(n)$ both for implicit and explicit descriptions: differences in complexity arise only for nonrandom strings (but not too nonrandom, for $K(x) \stackrel{+}{=} 0$ these differences vanish again).

It turns out that some strings cannot thus be explicitly represented parsimonously with low-complexity models (so that one necessarily has bad high complexity models like S^k above). For explicit representations, there are *absolutely non-stochastic* strings that don't have efficient two-part representations with $K(x) \stackrel{+}{=} K(S) + \log |S|$ $(x \in S)$ with $K(S)$ significantly less than $K(x)$, Section 4.

Again, consider the special set $S^k = \{y : K(y) \leq k\}$. As we have seen earlier, S^k itself cannot be explicitly optimal for x since $K(S^k) \stackrel{+}{=} k$ and $\log N^k \stackrel{+}{=} k - K(k)$, and therefore $K(S^k) + \log N^k \stackrel{+}{=} 2k - K(k)$ which considerably exceeds k. However, it turns out that a closely related set ($S^k_{m_x}$ below) is explicitly near-optimal. Let I^k_y denote the index of y in the standard enumeration of S^k, where all indexes are padded to the same length $\stackrel{\pm}{=} k - K(k)$ with 0's in front. For $K(x) = k$, let m_x denote the longest joint prefix of I_x^k and N^k, and let

$$I_x^k = m_x 0 i_x, \quad N^k = m_x 1 n_x,$$
$$S_{m_x}^k = \{y \in S^k : m_x 0 \text{ a prefix of } I_y^k\}$$

Theorem 3. *The set $S_{m_x}^k$ is an explicit algorithmic minimal near-sufficient statistic for x among subsets of S^k in the following sense:*

$$|K(S_{m_x}^k) - K(k) - l(m_x)| \stackrel{+}{<} K(l(m_x)),$$
$$\log |S_{m_x}^k| \stackrel{\pm}{=} k - K(k) - l(m_x).$$

Hence $K(S_{m_x}^k) + \log |S_{m_x}^k| \stackrel{\pm}{=} k \pm K(l(m_x))$. Note, $K(l(m_x)) \stackrel{+}{<} \log k + 2 \log \log k$.

The proof is given in the full paper [6]. We have not completely succeeded in giving a concrete algorithmic exlicit minimal sufficient statistic. However, we show [6] that $S_{m_x}^k$ is *almost always* minimal sufficient—also for the nonstochastic objects of Section 4.

4 Non-Stochastic Objects

Every data sample consisting of a finite string x has an sufficient statistics in the form of the singleton set $\{x\}$. Such a sufficient statistics is not very enlightening since it simply replicates the data and has equal complexity with x. Thus, one is interested in the minimal sufficient statistics that represents the regularity, (the meaningful) information, in the data and leaves out the accidental features. This raises the question whether every x has a minimal sufficient statistics that is significantly less complex than x itself. At a Tallinn conference in 1973 Kolmogorov (according to [14, 2]) raised the question whether there are objects x that have no minimal sufficient statistics that have relatively small complexity. In other words, he inquired into the existence of objects that are not in general position (random with respect to) every finite set of small enough complexity, that is, "absolutely non-random" objects. Clearly, such objects x have neither minimal nor maximal complexity: if they have minimal complexity then the singleton set $\{x\}$ is a minimal sufficient statistics of small complexity, and if $x \in \{0,1\}^n$ is completely incompressible (that is, it is individually random and has no meaningful information), then the uninformative universe $\{0,1\}^n$ is the minimal sufficient statistics of small complexity. To analyze the question better we need a technical notion.

Define the *randomness deficiency* of an object x with respect to a finite set S containing it as the amount by which the complexity of x as an element of S falls short of the maximal possible complexity of an element in S when S is known explicitly (say, as a list):

$$\delta_S(x) = \log |S| - K(x \mid S). \tag{11}$$

The meaning of this function is clear: most elements of S have complexity near $\log |S|$, so this difference measures the amount of compressibility in x compared to the generic, typical, random elements of S. This is a generalization of the sufficiency notion in that it measures the discrepancy with typicality and hence sufficiency: if a set S is a sufficient statistic for x then $\delta_S(x) \stackrel{\pm}{=} 0$.

Kolmogorov Structure Function: We first consider the relation between the minimal unavoidable randomness deficiency of x with respect to a set S containing it, when the complexity of S is upper bounded by α. Such functional relations are known as *Kolmogorov structure functions.* He did not specify what is meant by $K(S)$ but it was noticed immediately, as the paper [15] points out, that the behavior of $h_x(\alpha)$ is rather trivial if $K(S)$ is taken to be the complexity of a program that lists S without necessarily halting. Section 3.4 elaborates this point. So, this section refers to explicit descriptions only. For technical reasons, we introduce the following variant of randomness deficiency (11):

$$\delta_S^*(x) = \log |S| - K(x \mid S, K(S)).$$

The function $\beta_x(\alpha)$ measuring the minimal unavoidable randomness deficiency of x with respect to every finite set S of complexity $K(S) < \alpha$. Formally, we define

$$\beta_x(\alpha) = \min_S \{ \delta_S(x) : K(S) < \alpha \},$$

and its variant β_x^* defined in terms of δ_S^*. Note that $\beta_x(K(x)) \stackrel{\pm}{=} \beta_x^*(K(x)) \stackrel{\pm}{=} 0$.

Optimal Non-Stochastic Objects: We are now able to formally express the notion of non-stochastic ojects using the Kolmogorov structure functions $\beta_x(\alpha), \beta_x^*(\alpha)$. For every given $k < n$, Shen constructed in [14] a binary string x of length n with $K(x) \leq k$ and $\beta_x(k - O(1)) > n - 2k - O(\log k)$.

Here, we improve on this result, replacing $n - 2k - O(\log k)$ with $n - k$ and using β^* to avoid logarithmic terms. This is the best possible, since by choosing $S = \{0,1\}^n$ we find $\log |S| - K(x \mid S, K(S)) \stackrel{\pm}{=} n - k$, and hence $\beta_x^*(c) \stackrel{+}{<} n - k$ for some constant c, which implies $\beta_x^*(\alpha) \leq \beta_x(c) \stackrel{+}{<} n - k$ for every $\alpha > c$. The proof is relegated to the full version of this paper [6].

Theorem 4. *For any given $k < n$, there are constants c_1, c_2 and a binary string x of length n with $K(x \mid n) \leq k$ such that for all $\alpha < k - c_1$ we have*

$$\beta_x^*(\alpha \mid n) > n - k - c_2.$$

Let x be one of the non-stochastic objects of which the existence is established by Theorem 4. Substituting $k \stackrel{\pm}{=} K(x|n)$ we can contemplate the set $S = \{x\}$ with complexity $K(S|n) \stackrel{\pm}{=} k$ and x has randomness deficiency $\stackrel{\pm}{=} 0$ with respect to S. This yields $0 \stackrel{\pm}{=} \beta_x^*(K(x|n)) \stackrel{+}{>} n - K(x|n)$. Since it generally holds that $K(x|n) \stackrel{+}{<} n$, it follows that $K(x|n) \stackrel{\pm}{=} n$. That is, these non-stochastic objects have complexity $K(x|n) \stackrel{\pm}{=} n$ and are *not random, typical, or in general position* with respect to every set S containing them with complexity $K(S|n) \stackrel{\pm}{\neq} n$, but they are random, typical, or in general position only for sets S with complexity $K(S|n) \stackrel{+}{>} n$ like $S = \{x\}$ with $K(S|n) \stackrel{\pm}{=} n$. That is, every explicit sufficient statistic S for x has complexity $K(S|n) \stackrel{\pm}{=} n$, and $\{x\}$ is such a statistic.

References

1. A.R. Barron, J. Rissanen, and B. Yu, The minimum description length principle in coding and modeling, *IEEE Trans. Inform. Theory*, IT-44:6(1998), 2743–2760.
2. T.M. Cover, Kolmogorov complexity, data compression, and inference, pp. 23–33 in: *The Impact of Processing Techniques on Communications*, J.K. Skwirzynski, Ed., Martinus Nijhoff Publishers, 1985.
3. T.M. Cover and J.A. Thomas, *Elements of Information Theory*, Wiley, New York, 1991.
4. R. A. Fisher, On the mathematical foundations of theoretical statistics, *Philosophical Transactions of the Royal Society of London, Ser. A*, 222(1922), 309-368.
5. Q. Gao, M. Li and P.M.B. Vitányi, Applying MDL to learn best model granularity, *Artificial Intelligence*, To appear. http://xxx.lanl.gov/abs/physics/0005062
6. P. Gács, J. Tromp, P. Vitányi, Algorithmic statistics, Submitted. http://xxx.lanl.gov/abs/math.PR/0006233
7. M. Li and P.M.B. Vitányi, *An Introduction to Kolmogorov Complexity and its Applications*, Springer-Verlag, New York, 2nd Edition, 1997.
8. A.N. Kolmogorov, Three approaches to the quantitative definition of information, *Problems Inform. Transmission* 1:1 (1965) 1-7.
9. A.N. Kolmogorov, On logical foundations of probability theory, Pp. 1–5 in: *Probability Theory and Mathematical Statistics*, Lect. Notes Math., Vol. 1021, K. Itô and Yu.V. Prokhorov, Eds., Springer-Verlag, Heidelberg, 1983.
10. A.N. Kolmogorov and V.A. Uspensky, Algorithms and Randomness, *SIAM Theory Probab. Appl.*, 32:3(1988), 389–412.
11. L.A. Levin, Laws of information conservation (nongrowth) and aspects of the foundation of probability theory, *Problems Inform. Transmission* 10:3(1974), 206–210.
12. L.A.Levin Randomness conservation inequalities: information and independence in mathematical theories, *Information and Control* 61 (1984) 15-37.
13. P. Martin-Löf, The definition of random sequences, *Inform. Contr.*, 9(1966), 602-619.
14. A.Kh. Shen, The concept of (α, β)-stochasticity in the Kolmogorov sense, and its properties, *Soviet Math. Dokl.*, 28:1(1983), 295-299.
15. A.Kh. Shen, Discussion on Kolmogorov complexity and statistical analysis, *The Computer Journal*, 42:4(1999), 340–342.
16. P.M.B. Vitányi and M. Li, Minimum Description Length Induction, Bayesianism, and Kolmogorov Complexity, *IEEE Trans. Inform. Theory*, IT-46:2(2000), 446–464.
17. V.V. V'yugin, On the defect of randomness of a finite object with respect to measures with given complexity bounds, *SIAM Theory Probab. Appl.*, 32:3(1987), 508–512.
18. V.V. V'yugin, Algorithmic complexity and stochastic properties of finite binary sequences, *The Computer Journal*, 42:4(1999), 294–317.