The Annals of Probability
1989, Vol. 17, No. 3, 840–865

KOLMOGOROV'S CONTRIBUTIONS TO INFORMATION THEORY AND ALGORITHMIC COMPLEXITY

BY THOMAS M. COVER, PETER GACS[1] AND ROBERT M. GRAY

Stanford University, Boston University and Stanford University

1. Introduction. Kolmogorov's sustained interest in randomness and complexity led to his early contributions to information theory through seminars and papers in the 1950s, and culminated in the crucial idea of algorithmic (or descriptional) complexity in the 1960s.

Briefly, information theory says a random object $X \sim p(x)$ has complexity (entropy) $H = -\sum p(x)\log p(x)$, with the attendant interpretation that H bits are sufficient to describe X on the average. Algorithmic complexity says an object x has a complexity $K(x)$ equal to the length of the shortest (binary) program that describes x. It is a beautiful fact that these ideas are much the same. In fact, it is roughly true that $EK(X) \approx H$. Moreover, if we let $P_U(x) = \Pr\{U \text{ prints } x\}$ be the probability that a given computer U prints x when given a random program, it can be shown that $\log(1/P_U(x)) \approx K(x)$ for all x, thus establishing a vital link between the "universal" probability measure P_U and the "universal" complexity K. More on this later.

The relationship of these ideas to probability theory was summed up in Kolmogorov's 1983 paper which was based on his 1970 talk in Nice. Perhaps only the founder of modern probability theory would have the audacity to place it in the following unusual perspective ([K462][2]):

> Information theory must precede probability theory, and not be based on it. By the very essence of this discipline, the foundations of information theory have a finite combinatorial character.
>
> The applications of probability theory can be put on a uniform basis. It is always a matter of consequences of hypotheses about the impossibility of reducing in one way or another the complexity of the description of the objects in question. Naturally, this approach to the matter does not prevent the development of probability theory as a branch of mathematics being a special case of general measure theory.
>
> The concepts of information theory as applied to infinite sequences give rise to very interesting investigations, which, without being indispensable as a basis of probability theory,

Received November 1988.

[1] On leave at IBM Almaden Research Center.

[2] Reference citations preceded by K refer to the list of Kolmogorov's publications on pages 945–964.

> can acquire a certain value in the investigation of the algorithmic side of mathematics as a whole.

We will discuss Kolmogorov's contributions to Shannon information theory followed by a development of his ideas in algorithmic complexity. It becomes clear in hindsight that an interest in complexity was a dominant theme in Kolmogorov's thinking. In particular, he was interested in finding the determinism in random events and in defining the structure of discrete objects, whether it be the law of large numbers or a notion of intrinsic complexity. Even his work on turbulence theory can be seen in this light as an attempt to find deterministic order in chaotic processes.

We will provide the background theory and the consequences of his ideas to information theory and algorithmic complexity so that the reader may draw some conclusions about Kolmogorov's above-stated point of view.

2. Kolmogorov's contributions to probabilistic information theory. The unusual flow of ideas from engineering system design to and from mathematics is due in large part to Kolmogorov's pivotal role as a teacher and contributor. He has had a profound impact on the mathematical development of Shannon's information theory both directly through his own technical contributions and indirectly through his influence on a generation of information theorists. His impact has been amplified by his applications of information theoretic ideas to other fields, most notably ergodic theory, approximation theory and complexity theory.

In this section we describe the principal contributions of Kolmogorov and his school to Shannon theory and we briefly refer to some of his use of information theoretic ideas and techniques in related fields.

Kolmogorov first became interested in information theory in 1955 [5]. In June of the following year, he presented a paper with I. M. Gel'fand and A. M. Yaglom at the Third All-Union Mathematical Congress in Moscow [K276] which laid the foundation for the development of mathematical information theory by extending Shannon's notions of communication systems, entropy and information measures from discrete time processes with discrete or real alphabets to general alphabets and both discrete and continuous time. This paper reported and elaborated on results in several then recent papers of Kolmogorov and his colleagues, most notably [K267], [15] and [K272]. Several of the engineering-oriented results were published in English in [K264]. In addition to these papers, Kolmogorov had a strong influence through his work and personal contact on others who continued the early mathematical generalizations of Shannon's ideas and results. Notable among these are Dobrushin's development of asymptotic properties of general information measures and of conditional information measures and his formulation of source and channel coding theorems with multiple fidelity criteria [11] and Pinsker's development of the properties of information, information rate and information stability of random variables and processes,

especially Gaussian processes. (See Pinsker's book [35] for a good survey of much of the early work of Kolmogorov and his colleagues and students, together with many references. It is notable that Pinsker thanks Kolmogorov in the Author's Preface for drawing his attention to the questions treated in the book.)

We attempt here to sketch some of the basic Shannon ideas and their extension by Kolmogorov and his colleagues. Readers interested in more details and the subsequent development of mathematical information theory are referred to the excellent 1966 survey by Kotz [19]. This book is perhaps the best available English-language guide to the Eastern European information theory literature for the period 1956–1966.

Shannon information theory. In order to describe Kolmogorov's contributions, we begin with several ideas from Shannon's original work on information theory, the "mathematical theory of communication" [38]. Suppose that X is a discrete random variable or random vector described by a probability mass function (pmf) $p_X(x) = \Pr(X = x)$, where x takes values in an "alphabet" A_X. Shannon defined the *entropy* of X by

$$H(X) = -\sum_{x \in A_X} p_X(x) \ln p_X(x). \tag{2.1}$$

Similarly, given two discrete random objects X and Y, one can define the various entropies $H(X, Y)$, $H(X)$ and $H(Y)$ as well as conditional entropies such as

$$H(X|Y) = H(X, Y) - H(Y) \tag{2.2}$$

and the *mutual information*

$$\begin{aligned} I(X; Y) &= H(X) + H(Y) - H(X, Y) \\ &= \sum_{xy} p_{XY}(x, y) \ln \frac{p_{XY}(x, y)}{p_X(x) p_Y(y)}. \end{aligned} \tag{2.3}$$

Shannon also defined similar quantities for real random variables with probability density functions (distributions absolutely continuous with respect to Lebesgue measure). For example, if random variables X, Y are described by a probability density function $F_{XY}(x, y)$, then the mutual information can be defined by

$$I(X; Y) = \int f_{XY}(x, y) \ln \frac{f_{XY}(x, y)}{f_X(x) f_Y(y)} \, dx \, dy. \tag{2.4}$$

The notion of entropy did not extend naturally to continuous outcomes, although the similar notion of differential entropy $h(X) = -\int f_X(x) \ln f_X(x) \, dx$ is useful in formulas, for example, $I(X; Y) = h(X) + h(Y) - h(X, Y)$ in analogy with the discrete case. We shall see that Kolmogorov's work led to generalizations of these ideas from discrete and real random variables to general random objects taking values in abstract spaces.

A great deal of literature has been devoted to interpreting such information and entropy measures, finding relations among them and finding axioms that

characterize them. For example, it is easy to show that $I(X;Y)$ is symmetric, is nonnegative with equality to 0 only if X and Y are independent and is reduced by coding or mapping in the sense that

$$I(f(X);Y) \le I(X;Y). \tag{2.5}$$

Shannon's main focus, and the principal use of these quantities, however, has been in the development of coding theorems quantifying the optimal achievable performance in idealized communication systems. A simple discrete communications system model can be thought of as a sequence

$$U \to X \to Y \to V$$

with the following components:

the *source* or *input signal* X is a random variable described by a pmf p_U;
the deterministic mapping of U into X is called an *encoder*;
the stochastic mapping of X into Y is called a *channel* or *noisy channel* and the occurrence of a channel output Y, given the channel input X, is described by a conditional pmf $p_{Y|X}(y|x) = \Pr(Y = y|X = x)$;
the deterministic mapping of Y into the *reproduction* or *output signal* V is called the decoder.

In practical communication systems one does not communicate a single random variable or a single sample of a random process, one communicates a sample function of a random process. All of the above definitions can easily be extended to vectors of N coordinates, for example, we can consider an n-dimensional system

$$U^n \to X^n \to Y^n \to V^n,$$

where $X^n = X_1, X_2, \ldots, X_n$. Infinite sequences or processes require limiting operations.

Shannon showed that the optimal achievable performance when communicating a memoryless discrete source ($p_{U^n} = \prod_i p_{U_i}$) with entropy $H(U)$ over a discrete memoryless noisy channel ($p_{Y^n|X^n} = \prod_i p_{Y_i|X_i}$) is completely described by the quantity

$$C = \sup_{p_X} I(X;Y),$$

called the *capacity* of the channel, in a sense made precise by the following theorem.

THEOREM 2.1. *If $H(U) < C$, then there exists a sequence of deterministic mappings f_n: $A_U^n \to A_X^n$ (encoder) and g_n: $A_Y^n \to A_V^n$ (decoder) such that*

$$\lim_{n\to\infty} \Pr(g_n(Y^n) \ne U^n) = 0.$$

No such sequence exists if $H(U) > C$.

The theorem states that under the given conditions the source can be communicated over the channel arbitrarily reliably by using sufficiently long

"block codes" if the entropy is strictly less than the channel capacity. The theorem further states that such communication is not possible if the inequality is reversed (the case of equality has no general solution).

In the special case where the channel has equal input and output alphabets and is *noiseless* in the sense that $p_{Y_i|X_i}(y|x) = 1$ if $x = y$ and 0 otherwise, then $C = \log\|A_X\|$, the logarithm of the cardinality of the alphabet. In this case the theorem is called a *noiseless coding theorem* or, more correctly, an *almost noiseless source coding theorem* since it states that a source with one discrete alphabet can be coded into another alphabet in an almost invertible way provided the entropy of the source is smaller than the cardinality of the second alphabet.

An immediate question is whether similar results exist for more general sources and channels for suitable definitions of entropy and capacity. The general notions of entropy and information developed by Kolmogorov and his colleagues combined with the extensions of sources to general random processes with abstract alphabets and channels to regular conditional probabilities on sequence spaces provided the tools to formulate such general results and for subsequent solutions to many special cases. Before describing these ideas, however, we first consider another form of Shannon coding theorem.

An immediate (and practically important) issue not treated in Theorem 2.1 is how much performance need be lost if $H(U) > C$; that is, it is not enough to know that near zero probability of error is impossible, it is important to know how small the probability of error can be made under the given conditions. Here Shannon first developed [38] and then elaborated [39] the idea of source coding with a fidelity criterion. He introduced the idea of a distortion measure $d(u, v)$ measuring the distortion or cost of reproducing a source symbol u as a reproduction symbol v and defined the performance or fidelity of a communication system to be the expected distortion $E(d(U, V))$. Common examples of distortion measures are the Hamming distance $d(u, v) = 0$ if $u = v$ and 1 if $u \neq v$ (which yields the probability of error when the expectation is taken) and the squared error $d(u, v) = (u - v)^2$. Shannon emphasized distortion measures that were additive when looking at sequences, that is,

$$d_N(u^N, v^N) = \sum_{k=1}^{N} d(u_k, v_k).$$

Shannon conjectured in [38] and proved in [39] that the optimal performance in the sense of minimizing the average distortion when communicating a memoryless source over a memoryless channel with capacity C was determined by the *rate-distortion function* of the source and distortion measure defined by

$$R(D) = \min_{p_{UV} \in \mathscr{W}} I(U; V), \tag{2.6}$$

where $\mathscr{W}$ is the set of all joint pmf's p_{UV} which are consistent with the input distribution, i.e.,

$$\sum_v p_{UV}(u, v) = p_U(u)$$

and for which the following constraint is met:

$$\sum_{u,v} d(u,v)p_{UV}(u,v) \leq D,$$

i.e., the average distortion is less than some given quantity. Shannon called such a constraint a *fidelity criterion*. Specifically, Shannon proved the following theorem.

THEOREM 2.2. *If* $R(D) < C$, *then there exists a sequence of codes* f_n: $A_U^n \to A_X^n$ (*encoder*) *and* g_n: $A_Y^n \to A_V^n$ (*decoder*) *such that*

$$\lim_{n\to\infty} (1/n)E\big(d_n(U^n, g_n(Y^n))\big) = D.$$

If $R(D) > C$, *then any sequence of codes will yield asymptotic average distortion strictly greater than* D.

Observe that $R(D)$ plays a role similar to that of $H(U)$, a fact which led Kolmogorov to consider $R(D)$ explicitly as a useful generalization of entropy which he called *epsilon entropy*.

These remarkable results showed that in the simple cases considered, two information theoretic quantities depending separately on the source and channel categorize the potential performance. Finding the D for which $R(D) = C$ gives an unbeatable lower bound to the average distortion which can be approximately achieved by deterministic block codes if long enough codes are used.

When Kolmogorov became interested in Shannon's results in 1955, the notions of entropy and information and the coding theorems for discrete memoryless sources and channels had been established in Shannon's original paper, but the ideas of rate-distortion theory had only been sketched and Shannon's 1959 source coding paper had not appeared. Kolmogorov proceeded both to popularize Shannon's ideas among Soviet mathematicians and to extend the basic tools and results to more general systems. Dobrushin [12] points out that it was a difficult task for Kolmogorov to overcome the suspicions of many mathematicians of the time that information theory, although fashionable, was mathematically superficial. Kolmogorov understood the real values underlying Shannon's work and correctly saw its eventual importance to both mathematics and engineering. Kolmogorov sought, by his rigorous development of Shannon's ideas, to make the theory applicable in the most general possible setting and to develop information measures for abstract spaces. Such generalizations are important not only for mathematical reasons; Kolmogorov realized that generality and rigor would also prove important for eventual applications. Discrete memoryless sources and channels are not always adequate models for real-world signal sources and communication and storage media. Metric spaces of functions, vectors and sequences as well as random fields naturally arise as models of source and channel outcomes; real sources and channels usually possess memory. Other advantages of general and rigorous definitions are the potential for proving useful new properties, for gaining insight into their behavior and for finding formulas for computing such measures for specific processes.

Information and entropy. The first and most fundamental contribution of Kolmogorov and his colleagues to information theory was a new and general definition of average mutual information, together with a formula for its computation. The definition can be developed in a variety of ways, but it is most natural here to consider the definition for a pair of random variables X, Y described by a probability distribution P_{XY} on a measurable space $(A_X \times A_Y, \mathscr{B}_X \times \mathscr{B}_Y)$, where A_X and A_Y are abstract spaces with σ-fields $\mathscr{B}_X$ and $\mathscr{B}_Y$, respectively, and where $(\mathscr{B}_X \times \mathscr{B}_Y)$ is the product σ-field. The mutual information between X and Y is then defined as

$$I(X;Y) = \sup \sum_{k,l} P_{XY}(F_k \times G_l) \ln \frac{P_{XY}(F_k \times G_l)}{P_X(F_k)P_Y(G_l)}, \tag{2.7}$$

where the supremum is over all finite measurable partitions $\{F_k\}$ and $\{G_l\}$ of A_X and A_Y, respectively. Dobrushin later proved that the supremum could be restricted to partitions measurable with respect to a generating field [11]. This definition works for arbitrary random objects, including infinite sequences and sample waveforms. It has the intuitive interpretation that the mutual information between two continuous random objects is the largest mutual information that can be obtained by quantizing the continuous objects into a finite number of levels. Given this definition, the entropy of a general random variable is defined by

$$H(X) = I(X;X).$$

These general definitions easily reduce to the Shannon definitions for the case of discrete random variables. They permit easy generalization of several of the basic properties of mutual information (including those previously mentioned) by taking limits of the corresponding discrete results. The mutual information between two random processes or the *mutual information rate* was defined by

$$\bar{I}(X;Y) = \lim_{N \to \infty} (1/N) I(X^N; Y^N)$$

when the limit exists. The *entropy rate* was defined by $\bar{H}(X) = \bar{I}(X;X)$.

Entropy and ergodic theory. There is an obvious similarity between the definitions of mutual information and mutual information rate on the one hand and Kolmogorov and Sinai's extension of Shannon's entropy to dynamical systems, the so-called *Kolmogorov–Sinai invariant* or *metric entropy* of a dynamical system [K280, K284, 41], on the other. This similarity merits elaboration and can be used to provide an information theoretic interpretation of the Kolmogorov–Ornstein isomorphism theorem of ergodic theory [K280, K284, 31].

Suppose that $\{U_n\}$ is a discrete alphabet stationary source with entropy rate $\bar{H}(U)$. Consider the corresponding abstract dynamical system consisting of the probability space of sequences drawn from this process and the shift transformation T defined by $Tu = (\ldots, u_0, u_1, u_2, \ldots)$ if $u = (\ldots, u_{-1}, u_0, u_1, \ldots)$. Then the Kolmogorov–Sinai invariant of the dynamical system is the same as the entropy rate of the corresponding random process. A generalization of Shannon's

almost noiseless coding theorem (based on the Shannon–McMillan–Breiman theorem) states that a stationary and ergodic discrete alphabet source can be coded using block codes in an almost invertible manner into an arbitrary discrete alphabet provided that the cardinality of the alphabet is strictly greater than the entropy rate of the source. Here "almost invertible" means that the source can be recovered with probability arbitrarily close to 1. The Kolmogorov–Sinai theorem can be viewed as a limiting perfectly invertible version of the almost noiseless coding theorem with an extra property, as shall be seen next. To describe the theorem, we first need an alternative notion of coding which was not present in Shannon's works but was developed in ergodic theory: A *stationary code* is a measurable mapping $\bar{f}\colon A^\infty \to B^\infty$ of one infinite sequence space onto another with the property that if T_A and T_B are the shift operators on these spaces, then $\bar{f}(T_A x) = T_B \bar{f}(x)$; that is, shifting the input sequence yields the corresponding shifted output sequences. The corresponding mapping $f\colon A^\infty \to B$ of sequences into a single symbol defined by $f(u) = (\bar{f}(u))_0$, the time zero coordinate of the output sequence, is called a *sliding block code*, in contrast to a Shannon block code, since it produces an output sequence by "sliding" or shifting the input sequence; that is, $\bar{f}(u) = \{f(T^n u);\ n = \ldots, -1, 0, 1, \ldots\}$. Two discrete alphabet random processes $\{U_n\}$ and $\{X_n\}$ are *isomorphic* if one can be coded into the other using an invertible stationary code, that is, there are stationary codes $\bar{f}\colon A_U^\infty \to A_X^\infty$ and $\bar{g}\colon A_X^\infty \to A_U^\infty$ such that the process $\{f(T_{A_U}^n \overline{U})\}$ has the same distribution as $\{X_n\}$, where $\overline{U} = (\ldots, U_{-1}, U_0, U_1, \ldots)$, and

$$\Pr\bigl(U_0 \neq g(\bar{f}(\overline{U}))\bigr) = 0.$$

THEOREM 2.3 (Kolmogorov and Ornstein). *A necessary condition for two processes to be isomorphic is that they have the same entropy rate. If both processes are B-processes (i.e., are stationary codings of memoryless processes), then this condition is also sufficient.*

The Kolmogorov–Ornstein theorem can thus be interpreted as a limiting version of the almost noiseless coding theorem where the coding is perfectly invertible *and* where the encoded process has exactly some prespecified distribution, the distribution of the second given process. As block codes do not really make sense in the limit of infinite length, the stationary code structure is required to get the limit code. The "negative" portion of the coding theorem was due to Kolmogorov and introduced Shannon entropy into ergodic theory. The "positive" coding theorem was proved by Ornstein by embedding block codes into stationary codes (using the Rohlin–Kakutani theorem) and constructing suitable limits. The stationary codes of ergodic theory were subsequently introduced to information theory and used to obtain Shannon-style coding theorems for nonblock codes (see, e.g., [16] and [17]).

The general definitions of average mutual information between random variables and processes and of metric entropy for dynamical systems were developed at approximately the same time and formed a powerful bridge between information theory and ergodic theory that has proved fruitful to both fields.

Evaluation of information. Another major contribution was the derivation of a formula for computing the mutual information, which reduced to the known result for real-valued random variables with distributions absolutely continuous with respect to Lebesgue measure. Gel'fand, Kolmogorov and Yaglom [K267] showed that (1) if the distribution P_{XY} is not absolutely continuous with respect to the product distribution $P_X \times P_Y$ of its marginal distributions, then the mutual information $I(X; Y)$ is infinite, and (2) if the absolute continuity holds, then the Radon-Nikodym derivative

$$a(x, y) = \frac{dP_{XY}}{d(P_X \times P_Y)}$$

is well defined and

$$\begin{aligned} I(X; Y) &= \int_{A_X \times A_Y} \log a(x, y)\, dP_{XY}(x, y) \\ &= \int_{A_X \times A_Y} a(x, y) \log a(x, y)\, d(P_X \times P_Y)(x, y). \end{aligned} \tag{2.8}$$

The quantity $\log a(x, y)$ is called the *information density* and is the focus of the general Shannon–McMillan–Breiman theorems (information stability theorems). Generalizations and elaborations of this result were developed by Gel'fand and Yaglom [15] and independently by Perez [34]. Complete treatments of the results may be found in Dobrushin [11] and Pinsker [35]. It is interesting to note that both Pinsker and Kolmogorov [K264] attribute the basic result to Gel'fand and Yaglom, although the result first appears in the report to the Third All-Union Math Conference [K276].

Communication systems. The Shannon communication model was generalized to the case of abstract alphabets by requiring that $U \to X \to Y$ and $X \to Y \to V$ be Markov chains and requiring that the channel be described by a regular conditional probability measure $P_{Y|X}(F|X = x)$. [$U \to X \to Y$ is a Markov chain if $\Pr(Y \in F|X, U) = \Pr(Y \in F|X)$ with probability 1.] Combined with the general notions of average mutual information, the appropriate channel capacity and rate-distortion functions could then be defined in a natural way for general random vectors and processes.

Epsilon entropy. As previously mentioned, Kolmogorov also realized that the ideas sketched by Shannon regarding source coding with a fidelity criterion provide an appropriate generalization for the entropy concept to continuous random variables. Although the general notion of mutual information provides a definition for entropy, in general the entropy of a continuous variable is infinite and this sheds no light on information content or coding. Kolmogorov extended Shannon's general notion of a rate-distortion function to abstract probability spaces: Given a source distribution P_U, let $\mathscr{W}$ denote the family of all joint distributions P_{UV} having P_U as a marginal distribution [$P_{UV}(U \in F) = P_U(F)$] and having some additional constraints. This includes Shannon's principal

example of $E(d(U, V)) \leq D$, but it can also incorporate multiple average distortion constraints or maximum distortion constraints. Define the quantity

$$H_{\mathscr{W}}(U) = \inf_{P_{UV} \in \mathscr{W}} I(U; V), \tag{2.9}$$

where the general notion of average mutual information is used. If the simple average distortion constraint is used, then this is exactly Shannon's rate-distortion function. Kolmogorov focused on two examples: (1) the Shannon example with a squared-error distortion measure $d(u, v) = (u - v)^2$, and (2) a distortion measure $d(u, v) = 0$ or ∞ accordingly as $|u - v| < \varepsilon$ or not. The second example effectively replaces Shannon's average distortion constraint by a maximum distortion constraint. Kolmogorov renamed these examples *ϵ-entropy* because they form a natural extension of entropy: They should give the actual entropy of an approximation to the original random variable that is within a specified average or maximum distortion. For example, if one required a maximum error of 0, then the ε-entropy reduces to the ordinary entropy. A general converse coding theorem follows easily from the construction: If one wishes to communicate the given source over the given channel subject to the given fidelity criterion, then necessarily

$$H_{\mathscr{W}}(U) < C. \tag{2.10}$$

Thus Kolmogorov and his colleagues both generalized the formulation of entropy and the rate-distortion function and obtained a new general converse coding theorem in the process.

Kolmogorov's work initiated the development of approximate expressions for ε-entropy in the limit of small ε as well as exact evaluations for the finite-dimensional Gaussian case. The average distortion constraint results were subsequently extended to nonsquared error distortion measures by Linkov [27], and the maximum error form of ε-entropy was developed and applied to the theory of approximations of functions by Kolmogorov and Tihomirov [K285] and Vitushkin [45].

General coding theorem formulation. The open problem (which led to many years of work by many researchers) was (and is) to find general conditions on sources, codes and channels under which the inequality of (2.10) is also sufficient for the existence of codes meeting the given constraint. Alternatively, how can one define $H_{\mathscr{W}}(U)$ and C for general sources and channels so that the appropriate extensions of Theorems 2.1 and 2.2 hold.

It should be noted that Kolmogorov and his colleagues also pointed out that, as in Shannon's original paper, additional constraints could be added to the channel capacity definition, for example, one might wish to constrain the average power of channel signals.

In the general case of sources and channels with memory, one must first evaluate the mutual information, capacity and rate-distortion function for vectors of n samples [e.g., $I(X^n; Y^n)$, the corresponding capacity C^n, and the rate-distortion function $H^n_{\mathscr{W}}(U)$] and then consider the limiting behavior as

$n \to \infty$. One obtains necessary conditions for all n and hence also in the limit, but now the problem is to find conditions under which

$$\liminf_{n\to\infty} \frac{C^n}{H^n_{\mathscr{W}}(U)} > 1 \tag{2.11}$$

is sufficient for the existence of good codes.

Key to such proofs are ergodic theorems for information densities, results called *information stability theorems* in the Soviet literature. These results have their origins in Shannon's theorem on the entropy of ergodic processes and are often called Shannon–McMillan–Breiman theorems or asymptotic equipartition theorems. It is interesting to note that this remains an active area of research and that the most general known results are both recent and strongly dependent on the general formulation of Kolmogorov [1, 30].

Continuous time. The general formulation was extended to continuous time processes by using sampling arguments to form discrete time processes and taking suprema over all possible choices of sampling. Again this permitted the general information measures to inherit many of its properties from the simple discrete case.

Many of these ideas were elaborated on by Dobrushin [11], who explicitly considered multiple distortion measures, and the information theory literature is replete with proofs for many examples of sources, channels and code structures.

Information rates of Gaussian processes. Kolmogorov's work led to the study of the behavior of asymptotic information rates $\lim_{n\to\infty} n^{-1} I(X^n; Y^n)$ and ε-entropies $\lim_{n\to\infty} n^{-1} H^n_{\mathscr{W}}$ for continuous and discrete time Gaussian processes. Many alternative definitions of information rate were proposed and conditions under which they are equivalent were developed. This line of research was developed in detail in the book by Pinsker [35] and the references cited therein.

Kolmogorov's contribution to Shannon or probabilistic information theory far exceeds that suggested by a count of his publications in the area. His technical contributions and his expansion of the audience for the techniques and ideas of information theory has been of enormous benefit to the field. In addition, as we shall next argue, his work in probabilistic information theory left him in position to develop the key idea of algorithmic information theory.

3. Algorithmic information theory.

3.1. *Kolmogorov's early work on complexity.* In his 1965 article in *Problems of Information Transmission* [K320], Kolmogorov presented a fundamental notion of intrinsic complexity. (Further work was presented in 1969. Related notions were introduced independently by Solomonoff in 1964 and Chaitin in 1966–1969.) In this work he defined the notion of the complexity of a finite object x with respect to a certain fixed universal computer. This complexity is the length of the shortest binary program that causes the computer to print x.

He argued that this definition does not depend too much on the choice of the computer. This follows from the fact that if there is a rival computer to be used one can always mimic its operation on the fixed universal computer by a short pre-program which specifies the transition table of the rival computer. Thus any short description for the rival computer becomes a short description for the fixed computer. The penalty in description length is an additive constant necessary to describe the workings of the first computer. This leads to a natural notion of algorithmic information theory, provides a notion of program complexity to go with existing ideas in computational complexity and points toward the notion of universal probability measures. Before going into these ramifications we would like to discuss how it is that Kolmogorov might have been led to his notion of algorithmic complexity.

As a young mathematician, Kolmogorov attended Stepanov's seminar on trigonometric series, and wrote his first paper [K2] at age 19 on the order of magnitude of Fourier coefficients. He found an example of a very complex Fourier series which diverged almost everywhere. This was likely the beginning of his thinking about the complexity of functions. In 1955 and 1956, Kolmogorov introduced the concept of the ε-entropy of a set of functions in a metric space. Here he asked how the number of functions in an ε-net or in an ε-covering must grow with ε, thus characterizing the size or complexity of the set. In particular he was able to prove Vitushkin's results on the impossibility of representing an r times differentiable function of n variables by composition of l times differentiable functions of m variables. By a neat counting argument it was possible to show that such a composition is possible if $n/r > m/l$. The amount of information necessary to describe the constituent functions well enough to characterize the desired function requires these conditions.

This then led Kolmogorov to Hilbert's thirteenth problem where he was able to prove that every continuous function of any number of variables can be represented as a composition of continuous functions of three variables. He later improved this result with Arnold by showing that a continuous function of any number of variables can be represented as a composition of continuous functions of a single variable with the addition operation as the only function of several variables.

Kolmogorov also found, through his work on rate-distortion theory, that $2^{nR(\varepsilon)}$ sequences ε-cover a collection of random sequences with high probability, where $R(\varepsilon)$ is defined in (2.6). The emphasis here is on the high probability. One can ignore the "atypical" sequences.

During this period in the 1950s, Kolmogorov also considered a recursive function theoretic approach to complexity. He had a seminar in the 1950s on the subject and published a paper with Uspenskii proposing a machine model (now named after them) general enough to convince him of the validity of Church's thesis. The storage structure of this machine is a graph built from pointers which the finite state control can change locally at each step. This model, along with a variant of it (called the storage modification machine), developed independently by Schoenhage, is widely used in computational complexity theory. Kolmogorov's seminar also investigated Shannon's estimates for the complexity of Boolean

functions. Kolmogorov's question on the computational complexity of multiplication resulted in Karatsuba's fast multiplication algorithm.

Clearly, Kolmogorov had a life-long interest in complexity. Hilbert's problem and ε-entropy had to do with the complexity of functions, while the rate-distortion work dealt with the complexity of random sequences. In both cases a descriptive complexity point of view led to counting arguments characterizing the solution. It is no surprise then, given Kolmogorov's interest in recursive function theory, that he was led to the universal notion of the intrinsic complexity of sequences.

We now describe that work.

3.2. *Algorithmic complexity.* The definition of the algorithmic complexity

$$K(x) = \min_{U(p)=x} l(p)$$

of a string x as the length of the shortest program for a universal computer U to output x had immediate impact. Various versions of this notion were discovered independently at approximately the same time. However, the clarity, brevity and accuracy of Kolmogorov's formulation and his view of its role in information theory and probability theory allowed advanced work to follow immediately from his contribution. The other discoverers were Solomonoff and, to some extent, Chaitin. Solomonoff [42] formalized the notion of Occam's razor for inductive inference. He attributed to every object x an intrinsic probability that is essentially

$$\sum_{p:\, U(p)=x} 2^{-l(p)},$$

where the latter can be interpreted as the probability that a computer U will print x when fed a random program p. In the attempted proof of the machine invariance of this notion he defined the length of shortest description as well. (Solomonoff's paper is loaded with intuitively interesting and significant ideas. But its use of mathematics is not rigorous. Several of the statements are unclear, and the right concepts are hidden among less useful alternatives.) In 1966, Chaitin, then an undergraduate at City College of New York, published a paper which proposed the smallest number of states of a Turing machine outputting the string x as the measure of the complexity of x.

3.3. *Kolmogorov's introduction of complexity.* In [K320], Kolmogorov motivates the definition of algorithmic complexity by reviewing two earlier ideas: (1) that $\log|M|$ is the number of bits needed to describe the elements of a set M, and (2) that $H(X)$ is the number of bits needed to describe a random variable X.

In the new approach, to capture the amount of information required to specify an object x, Kolmogorov used the notion of the theory of algorithms. This theory gives a precise mathematical definition of the notion of a computable function. For this purpose, a simple symbolic computing device, the Turing machine is introduced.

Turing machines. A Turing machine is a device which has a so-called control unit with a finite number of possible states, and a potentially infinite memory arranged in the form of a tape. The tape is divided into squares. The machine has a finite tape alphabet, and each tape square contains one symbol of this alphabet written into it. The machine works in discrete time. At each time step, the control unit is in one of the possible states and is scanning one of the tape squares. In one computation step, a certain elementary action is performed. Elementary actions are the change of the control state, the move of the scanner left or right and the change of the symbol written in the tape square currently scanned. The step to be performed depends on the current control state and the current scanned symbol. This transition function can be written in a finite table which can be considered to be the description of the "hardware" of the machine.

Let us confine ourselves to Turing machines whose tape alphabet includes (among others) the symbols 0, 1 and "blanks." To a machine M, we can define a partial function $M(s)$ that assigns finite binary strings to finite binary strings wherever defined, as follows. We write the string s on an otherwise blank tape and place the scanner on the first symbol of s. We run the machine M until it comes to a state called the "halting state." If at this time, the tape contains a binary string surrounded by blanks, then this is $M(s)$; otherwise $M(s)$ is undefined. A function $f(s)$ is called *computable* if there is a Turing machine M such that $f(s) = M(s)$. Work in the decades following Turing's definitions confirmed that the work of all imaginable computers can be simulated by appropriate Turing machines. In particular, all possible binary functions $f(s)$ computable in any intuitive sense are computable in the above formal sense. (This statement is known as the Church–Turing thesis.)

Turing has shown that it is enough to use a single Turing machine for the definition of computable functions since there are Turing machines that are *universal*: They can simulate any other Turing machine. More specifically, he showed the existence of a Turing machine U such that for every other Turing machine M there is a binary string p_M such that for all binary strings s we have $M(s) = U(p_M s)$. Here, $p_M s$ is the concatenation of the string p_M and the string s. The string p_M can be considered to be the description of the machine M for the universal machine U, or the "program" of the function $M(s)$ on the universal machine U.

The invariance theorem. Kolmogorov considered an arbitrary computable partial function φ and defined $K_\varphi(x)$ as the minimum program length $l(p)$ over all binary programs p such that $\varphi(p) = x$. The function φ acts as the decoder, or interpreter, of the description p. The complexity of x is the length of the shortest description, with respect to the interpreter φ.

The main theorem of Kolmogorov's paper [K320], which we will call the invariance theorem, says that this notion of complexity can be made fairly independent of the choice of the interpreter, that is, there are "asymptotically optimal" interpreters $U(p)$ with the property that for any other computable partial function $\varphi(p)$ we have the inequality

$$K_U(x) \le K_\varphi(x) + c_\varphi, \tag{3.1}$$

where the constant c_φ does not depend on x. This theorem is proved by simply choosing U as a Turing machine universal in the sense defined above. Indeed, the universality of U implies that there is a binary string q_φ with the property that for all p we have $\varphi(p) = U(q_\varphi p)$.

It follows that Martians, humans and computers will all approximately agree on the intrinsic complexity of n bits of *War and Peace*, the Mona Lisa and a Bernoulli sequence with parameter p. (Experimental evidence on language compression and image compression leads one to believe that these quantities are approximately given by $n/3$, $n/10$ and $n(-p \log p - (1-p)\log(1-p))$, respectively. Shiryaev's paper in the present issue refers to the research initiated by Kolmogorov on the estimation of complexity in works of art. Another approach to these problems is suggested in [3].) Solomonoff's independent proof of the invariance theorem appeared in 1964.

The work [K320] concludes with a statement concerning the algorithmic notion of randomness:

> If a finite set M containing a very large number N of members admits determination by means of a program of length negligibly small in comparison with $\log_2 N$ then almost all members of M have complexity $K(x)$ close to $\log_2 N$. The elements $x \in M$ of this complexity are also considered as "random" elements of the set M.

For the case when M is the set of all binary strings of length n it is easy to see that the relation $K(x) \leq n + O(1)$ holds for all x. On the other hand, the number of strings x with complexity less than $n - k$ is at most 2^{n-k}. Indeed, there are at most 2^{n-k} descriptions of length less than $n - k$.

Algorithmic properties of complexity. The complexity function $K(x)$ is not computable. Indeed, if $K(x)$ were computable, then we could define a string of high complexity with a short program. [The program would make use of the algorithm to compute $K(x)$.] The uncomputability theorem in the form stated by Kolmogorov is cited in [48]. (This property is intimately connected to the so-called Berry–Richardson paradox, asking for the smallest number undefinable in 100 words.) Such ideas also appear in Chaitin's 1966 paper [6]. The uncomputability of $K(x)$ confines its role somewhat to that of a theoretical clarifying tool. In particular, the definition of random strings of length n as those with complexity close to n is not operational.

In some sense, $K(x)$ possesses one-half of the property of computability. Only *nonrandomness* can ever be established. Let us make this more exact using a notion introduced in [48]. A real function $f(x)$ is called *enumerable* (or semicomputable) if one can recursively enumerate the set of pairs (x, r) such that r is a rational number and $f(x) > r$. In other words, one can compute arbitrarily close lower bounds to $f(x)$. A real function f is *computable* if both f and $-f$ are enumerable, that is, if one can compute arbitrarily close lower and upper bounds to f. The complexity function $K(x)$ is not computable but has the property that $-K$ is enumerable.

The theorems of Kolmogorov and Barzdin' in [48] show that there are no nontrivial enumerable lower bounds to $K(x)$. The proof of this result also shows that if the axioms of a formal theory have complexity k, then the theory can have only a small number of theorems of the form "$K(x) > m$" for various strings x and numbers m. Indeed, all such m are bounded by $m \leq k + O(1)$. In this way, an abundant source of undecidable mathematical statements was opened. See also Chaitin's development of these ideas in [8, 9].

The above considerations also raised interest in explicitly defined infinite sequences whose initial segments are complex. Of course, such sequences cannot be computable. The best known noncomputable sequence is the following. Let T be a universal computer with integer inputs. Let $\chi_n = 1$ if T halts with input n, and 0 otherwise. The sequence χ is not computable, but not random either. Barzdin' showed in 1968 [2] that the complexity $K(\chi^n|n)$ of an initial segment of χ is bounded above by $\log n + O(1)$ and that the upper bound is achieved. This complexity must be near n for random sequences.

Conditional complexity and conditional entropy. Some properties of statistical entropy have meaningful nontrivial analogs in algorithmic information theory. Let $K(x|y)$ be the length of the shortest program computing x on a given universal computer when y is given, and let $K(x, y)$ be the complexity of the pair x, y. Kolmogorov stated in 1968 the relation

$$K(x, y) = K(x) + K(y|x) + O(\log K(x))$$

proved by him and Levin and remarked that the logarithmic correction term is necessary. This relation is much less obvious than the analogous information theoretic identity $H(X, Y) = H(X) + H(Y|X)$. Using a slightly modified version $K^{(p)}(x)$ of complexity (where no program can be the prefix of another), one can arrive at a form of this relation that is exact to within an additive constant:

$$K^{(p)}(x, y) = K^{(p)}(x) + K^{(p)}\big(y|x, K^{(p)}(x)\big) + O(1). \tag{3.2}$$

This relation, proved independently by Levin in [13] and Chaitin in [10], cannot be improved by omitting $K^{(p)}(x)$ from the condition. Chaitin noted that the quantity

$$\overline{H}(y|x) = K^{(p)}\big(y|x, K^{(p)}(x)\big)$$

obeys identities completely analogous to those of conditional entropy.

In [22] and [23], Levin extended the information notion

$$I(\alpha; \beta) = K^{(p)}(\alpha) + K^{(p)}(\beta) - K^{(p)}(\alpha, \beta)$$

to infinite sequences α, β and proved an inequality analogous to the data processing inequality for random variables. Let the random variable Y be obtained from β by a probabilistic algorithm (an algorithm permitted to use the independent tosses of an unbiased coin). Then

$$E2^{I(\alpha, Y)} \leq 2^{I(\alpha, \beta)}, \tag{3.3}$$

where E denotes the expectation. This theorem implies a remarkable strengthening of Gödel's incompleteness theorem. Let u be a sequence describing in some standardized way all our current knowledge that can have a bearing on the infinite sequence χ (defined above) describing the halting problem. Assume that the information $I(\chi; u)$ in u about χ is some finite number c. Let U be an infinite random sequence obtained by the application of some probabilistic algorithm to u. The above inequality and the Markov inequality imply that the probability that $I(\chi; U)$ becomes greater than $c + k$ is less than 2^{-k}. In other words, no algorithm, even one using "creative" randomization, will ever increase our information about the halting problem (or, for that matter, about the solvability of Diophantine equations) by more than a few bits, even over infinite time.

The problem of randomness. In his 1987 paper [K475] written with Uspenskii, Kolmogorov gives what can now be considered the standard introduction to the paradox of randomness. Suppose that persons A and B give us a sequence of 20 digits each, saying that they were obtained from independent coin flips. If A gives the string 10101110011010001111 and B gives a string of twenty 0's, then we would believe A and would not believe B despite the fact that both strings have the same probability of occurrence in a series of coin flips. The string given by A seems random and the string given by B does not. This problem is at the root of mathematical statistics, that is, the study of testing probability models against experimental results. Laplace was aware of this [20]:

> In the game of heads and tails, if head comes up a hundred times in a row then this appears to us extraordinary, because after dividing the nearly infinite number of combinations that can arise in a hundred throws into regular sequences, or those in which we observe a rule that is easy to grasp, and into irregular sequences, the latter are incomparably more numerous.

The quote shows that Laplace's idea of a solution for the paradox above seems somewhat similar to Kolmogorov's: The nonrandom strings are the ones with some regularity in them, and since the number of all those strings is small, the occurrence of such a string is extraordinary. But a convincingly general formal notion of irregularity, for which it could be proven that the number of regular strings is small, was not found until 1965. von Mises attempted, beginning with his work in 1919 (see, e.g., the edition [46]), to define the notion of a random infinite sequence. von Mises's *Kollektiv* is a sequence in which the relative frequencies converge for all subsequences selected by certain selection rules. von Mises's ideas were elaborated upon by a number of others, especially Ville and Church. Kolmogorov's criticism (see the 1956 work [K268]) pointed out that the infinite sequences used by von Mises just shifted the problem of untestability from one area to another one.

In his 1963 paper on tables of random numbers, Kolmogorov came to the view that meaningful finite versions of von Mises's definition are possible, and

explicitly revoked his earlier criticism. Despite the interesting open technical problems still left by that paper, its randomness definition seems less appropriate today than the one based on description complexity.

Martin-Löf 's tests. In 1966 in [29] the young Swedish mathematician Martin-Löf,who worked briefly with Kolmogorov, gave a formal theory of finite and infinite random sequences that has great convincing power of its own and is in harmony with Kolmogorov's suggestions based on complexity. Just as later work on algorithmic complexity can be considered "second-order" corrections to Kolmogorov's original concept, all later work on the theory of randomness can be considered as extensions and second-order corrections to Martin-Löf's randomness tests. We now discuss these tests for coin-tossing sequences of length n, using the more concise development given by Zvonkin and Levin in 1970. A *test* is an enumerable (lower semicomputable) function $F(x)$ of finite binary strings x with the property that for all k, n, the number of binary strings of length n with $F(x) > k$ is at most 2^{n-k}. In computing lower bounds to $F(x)$ we find out gradually the degree to which x violates our ideas of randomness. But the total probability of those x rejected at level k is not allowed to be more than 2^{-k}. We can call $F(x)$ the *randomness defect* of x, and it can be interpreted as the logarithm of the significance level at which, after the testing, we reject the hypothesis that x is a sample from a Bernoulli process with parameter $\vartheta = 1/2$.

Here is an example of a test. For a binary string x of length n let $f(x)$ be the number of 1's in x. Chebyshev's inequality says that for all positive λ, the probability of $2n^{-1/2}|f(x) - n/2| > \lambda$ is at most $1/\lambda$. Therefore the function

$$F(x) = \log\bigl(2n^{-1/2}|f(x) - n/2|\bigr)$$

is a test.

In analogy with the invariance theorem above, Martin-Löf proved the existence of a *universal test* $d(x)$ that dominates all other statistical tests, that is, a test with the property that for each other test $F(x)$ there is a constant c_F such that for all x we have $F(x) \le d(x) + c_F$. The computation of $d(x)$ is equivalent to testing that the string satisfies all laws of probability theory whether these laws are explicitly known or not. [Unfortunately, $d(x)$ is not computable.] Strings x for which the value $d(x)$ is small can therefore be considered random.

The correspondence between the complexity definition of randomness and its definition via tests is shown by the following surprisingly exact theorem of Martin-Löf:

$$d(x) = l(x) - K(x|l(x)) + O(1). \tag{3.4}$$

This equation characterizes randomness in terms of complexity. It says that the "log significance level" $d(x)$ and the algorithmic complexity $K(x|l(x))$ sum to a constant.

Randomness with respect to a set: An algorithmic "sufficient statistic." Unlike most other developers of algorithmic information theory, Kolmogorov wanted to use the theory to eliminate the need for a direct interpretation of

probabilities. Therefore instead of randomness with respect to a probability distribution (the choice of Martin-Löf and others), he preferred the notion of randomness of an element x with respect to a set S containing it. To approximate the notion of a random sequence of n tosses of a biased coin, in this approach one chooses the set S as the set of all n-strings with a certain given frequency of 1's. Note that sets of this form provide a sufficient partition with respect to the Bernoulli process with unknown parameter θ, and that x is uniformly distributed over the set S. By analogy to the relation (3.4), Kolmogorov introduced (see [K462]) the quantity

$$d(x|S) = \log|S| - K(x|S)$$

as the defect of randomness of a string x with respect to the set S. There are ways to move from each of the two approaches (randomness with respect to a set and randomness with respect to a probability distribution) to the other, as pointed out, for example, in [47].

At a Tallin conference in 1973, Kolmogorov proposed a variant of the function

$$\delta_x(k) = \min_S \{d(x|S): K(S) < k, x \in S\},$$

considering it an interesting characteristic of the object x. If x is some sequence of experimental results, then the set S can be considered to be the extraction of those features in x that point to nonrandom regularities. At the point $k^*(x)$ where the decreasing function $\delta_x(k)$ becomes 0 (or less than some constant agreed on in advance), we can say that it is useless to explain x in greater detail than by giving the set S^* such that $d(x|S^*) = \delta_x(k^*)$. Indeed, the added explanation would be as large as the number of extra bits it accounts for. The set S^* expresses all the relevant structure in the sequence x, the remaining details of x being conditionally maximally random. For example, S^* would describe the Mona Lisa up to brush strokes, and k^*, the length of description of S^*, is the "structural complexity of x."

The set S^* plays the role of a *universal minimal sufficient statistic* for x. In mathematical statistics, a function T: $\mathcal{X} \to \mathcal{T}$ is said to be a *sufficient statistic* relative to the family of densities $\{f_\vartheta(x)\}$ if $\vartheta \to T(X) \to X$ forms a Markov chain, that is, if $f_\vartheta(x, T(x)) = g_\vartheta(T(x))h(x|T(x))$, for some choice of densities $g_\vartheta(\cdot)$, and $h(\cdot|\cdot)$. The statistic T is *minimal sufficient* if $\vartheta \to T \to T' \to X$ for every other sufficient statistic T'. It is in this sense that X is "as random as it can be" given $T(X)$. Now, the set S^* defined above by Kolmogorov is such that x is conditionally maximally random given S^*, that is, $K(x|S^*) \approx \log|S^*|$. Note that the definition of S^* stands on its own as an algorithmic definition of structure, without any probabilistic interpretation.

For $k < k^*(x)$, the function $\delta_x(k)$ represents the trade-off between the size of the explanation and its value. If the object x is random with respect to a probability distribution which can be defined by a short program then $k^*(x)$ is small. Kolmogorov asked whether for any decreasing function $f(k)$ there are objects x for which $\delta_x(k)$ is close to $f(k)$, especially, whether there are "absolutely nonrandom" strings x, for which $k^*(x)$ is close to $K(x)$. Positive answers to these questions were given in [40] and [47].

Kolmogorov and Uspenskii write in 1987: "The question, of course, remains whether such (absolutely non-random) strings exist in the real world." The quantity $k^*(x)$ provides a lower bound on the information $I(x, \chi)$ of x on the infinite sequence χ describing the halting problem. Levin's previously mentioned algorithmic data processing inequality implies that the strings x having information about the halting problem are very exotic and unlikely to occur in nature. Therefore, strings x with a $k^*(x)$ of any significant size are unlikely to occur in nature.

3.4. *Further development of algorithmic information theory.* We now highlight the ideas resulting from algorithmic information theory. The consequences of the theory demonstrate its naturalness as well as its relationship to probability theory.

Semimeasures. Let $\mathbf{N} = \{0, 1, 2, \dots\}$. The nonnegative real function $\nu(x)$ is called a *discrete semimeasure* over $\mathbf{N}$ if $\sum_x \nu(x) \leq 1$. It will be called a *probability measure* in case of equality.

Let $\mathbf{B}$ be a (finite or infinite) alphabet, and $\mathbf{B}^*$ the set of finite strings over $\mathbf{B}$. A nonnegative function μ on $\mathbf{B}^*$ is called a *semimeasure* if $\mu(\Lambda) \leq 1$ for the empty string Λ and

$$\sum_{b \in \mathbf{B}} \mu(xb) \leq \mu(x) \tag{3.5}$$

holds for all x, where xb denotes the concatenation of the string x and the symbol b. It is a *probability measure* if equality holds in both of the above relations. In terms of classical measure theory, $\mu(x)$ is really the measure of the set of all infinite sequences beginning with x, that is, the probability that x is a prefix of the sample sequence.

Monotonic and prefix complexity. For aesthetically appealing results, it is useful to modify Kolmogorov's original definition slightly. We define two such variants, introduced by Levin in [21] and [22] (see also [13]). Similar or related notions were defined in [37] and [10]. The difference between the newer quantities and $K(x)$ can be bounded by $O(\log l(x) + \log K(x))$.

The so-called *prefix complexity* $K^{(p)}(x)$ is the one to be used for the discrete space $\mathbf{N}$. For strings x, y let $x \subset y$ denote that x is a prefix of y. Let us consider computing machines T using binary strings as inputs, and outputting an element of $\mathbf{N}$. Such a machine is *self-delimiting* if $p \subset p'$ implies $T(p) = T(p')$ whenever $T(p)$ is defined. Let us denote $K^{(p)}(x) = K_U(x)$ with a self-delimiting machine U optimal in the sense of the relation (3.1).

The so-called *monotonic complexity* $K^{(m)}(x)$ is the one preferred in Kolmogorov's last article [K475], and the one more suitable for the "continuous" space $\mathbf{B}^*$. The interpreter for this notion is a relation E rather than a function. It is recursively enumerable, that is, there is a computable function f such that E consists of the values $f(1), f(2), \dots$. We say that the binary string p is a description of string x if $(p, x) \in E$. We require that E *preserves comparability*,

in the sense that if $(p_1, x_1), (p_2, x_2) \in E$ and p_1 is a prefix of p_2, then x_1 is a prefix of x_2 or x_2 is a prefix of x_1. We define $K_E^{(m)}(x) = \min\{l(p): (p, x) \in E\}$. We denote $K^{(m)}(x) = K_U^{(m)}(x)$ with an interpreter U optimal in the sense of (3.1).

A priori probability. Solomonoff's article in 1964 defines a notion of "a priori probability," to be used for inductive inference in a very general situation. This notion was closely related to the probability $\mathbf{M}(x)$ that the optimal monotonic machine U produces an extension of x when presented with an infinite coin-tossing random binary string as input. The function $\mathbf{M}(x)$ is a semimeasure. It is not necessarily a measure since the machine U may never terminate computation on some inputs. Zvonkin and Levin showed in 1970, that $\mathbf{M}$ is *maximal*, to within a multiplicative constant, among the enumerable semimeasures. Thus, for all enumerable semimeasures μ there is a positive constant c_μ with

$$\mathbf{M}(x) \geq c_\mu \mu(x) \tag{3.6}$$

for all x. The measure $\mathbf{M}(x)$ is also called the *universal semimeasure*. It "contains" all other enumerable semimeasures. (Solomonoff conjectured a certain property of "optimality" for his notions of a priori probability, but this conjecture was mathematically vague in the sense that he didn't define the class of measures with respect to which the a priori measure should be optimal.

The papers [48] and [21] also show that the quantity $\mathbf{H}(x) = -\log \mathbf{M}(x)$ is close to the complexity $K^{(m)}(x)$. In particular, the difference can be bounded by $O(\log \mathbf{H}(x))$. (A lower bound on the difference is given in [14].)

Over a discrete space $\mathbf{N}$, the relation between semimeasures and complexity is much simpler. Let $\mathbf{m}(x)$ denote the probability that x is produced on the fixed optimal prefix machine with coin-tossing binary input. Just as in the continuous case, this enumerable semimeasure is maximal to within a multiplicative constant, and is called the a priori probability. Obviously, $-\log \mathbf{m}(x) < K^{(p)}(x)$ since the optimal program is one of the possible binary programs giving x. Levin showed (see [22] and [13])

$$K^{(p)}(x) = -\log \mathbf{m}(x) + O(1). \tag{3.7}$$

This nontrivial result was also obtained independently by Chaitin in [10]. It is fitting to call it the coding theorem. The code that assigns small descriptions to objects of high a priori probability is related to the variable-length codes of Shannon [38]. The technical difficulty is the noncomputability of the function $\mathbf{m}(x)$.

Characterizing randomness by complexity. Let μ be a computable probability distribution over $\mathbf{B}^*$. The following theorems by Levin from 1973 are analogous to the upper bound theorem and lower bound theorem of $K(x)$. First: There is a constant k_μ such that we have for all x,

$$-\log \mathbf{M}(x) = \mathbf{H}(x) \leq K^{(m)}(x) \leq -\log \mu(x) + k_\mu.$$

Thus, for a computable measure, each sample has a minimal description bounded

by the logarithm of its probability plus a constant. Second: Let V_k be the set of those ω for which there is an n with $k < -\log \mu(\omega^n) - K^{(m)}(x)$. Then $\mu(V_k) \leq 2^{-k}$. In other words, the above estimate is close for most samples. These results suggest that the quantities

$$\begin{aligned} d(x, \mu) &= -\log \mu(x) - \mathbf{H}(x), \\ d(\omega, \mu) &= \sup_n d(\omega^n, \mu) \end{aligned} \tag{3.8}$$

can measure the defect of randomness in the finite string x and the infinite string ω, respectively. Here, ω^n is the prefix of length n of ω. Levin showed that indeed, it is asymptotically equal to the universal randomness defect introduced by Martin-Löf (see the discussion of Martin-Löf's tests). This results in the following beautiful characterizations of random sequences: The sequence ω is random iff for all n, the logarithm $\mathbf{H}(\omega^n)$ of a priori probability is within an additive constant $d(\omega, \mu)$ of its upper bound $-\log \mu(\omega^n)$. [Also, iff the same is true of the complexity $K^{(m)}(\omega^n)$ in place of $\mathbf{H}(\omega^n)$.]

For a discrete computable probability distribution ν there is no sharp distinction between random and nonrandom. The corresponding quantity measuring the defect of randomness of an outcome x is $-\log \nu(x) - K^{(p)}(x)$.

Another interpretation of the defect of randomness

$$-\log \mu(x) - \mathbf{H}(x) = \log \frac{\mathbf{M}(x)}{\mu(x)} \tag{3.9}$$

is that it is the likelihood ratio of the hypothesis μ and the fixed alternative hypothesis $\mathbf{M}(x)$. The string x is random with respect to μ if and only if its a priori probability is close to its μ-probability.

Testing for randomness by betting. Suppose that a casino claims that the distribution of the outcomes ω is the computable measure μ. Then, given any function $f(\omega)$ with $\int f(\omega)\, d\mu < 1$, the casino should accept 1 unit for an obligation to pay $f(\omega)$ on outcome ω. A sequential payoff function leading to such a global payoff function is a function $t(x)$ on the space B^* of finite strings with $t(\Lambda) < 1$ and

$$\sum_{b \in \mathbf{B}} t(xb)\mu(xb) \leq t(x)\mu(x). \tag{3.10}$$

This inequality says that the function $t(\omega^n)$ is a *submartingale*.

We could call the outcome ω nonrandom if it allows us to win against the casino by choosing an appropriate payoff function. Thus ω is nonrandom if there is a computable (or at least enumerable) submartingale $t(x)$ such that $t(\omega^n)$ grows unboundedly. It turns out that this characterization of random infinite strings is equivalent to the one using the quantity $d(\omega, \mu)$ in (3.8) above. Indeed, comparison with the inequality (3.5) shows that $t(x)$ is a submartingale iff $t(x)\mu(x)$ is a semimeasure. It follows immediately that $\mathbf{M}(x)/\mu(x)$ is maximal, within a multiplicative constant, among all enumerable submartingales. It can be called a *universal payoff function*. Now, the logarithm of this martingale is just the randomness defect $d(x, \mu)$ found in (3.8).

The characterization of randomness by martingales was the earliest attempt to go beyond the limited tests provided by von Mises's selection schemes, and was proposed by Ville in [44]. It was shown in Schnorr's 1971 book [36] that gambling tests based on enumerable martingales are equivalent to Martin-Löf's universal tests.

Information. Description complexity was recommended by Kolmogorov as the right definition of individual information. It is therefore interesting to clarify its relation to the entropy of a probability distribution, its statistical analog. For a discrete random variable X with a computable mass function $f(x)$, let $H(X)$ denote its entropy $-\sum_x f(x)\log f(x)$. It follows from the coding theorem (3.7), the "Kraft inequality" $\sum_x 2^{-K^{(p)}(x)} \leq 1$ and Shannon's theorems in [38] for variable-length codes that

$$H(X) = \sum_x f(x)K^{(p)}(x) + c_f,$$

where the constant c_f depends on the length of program needed to define the distribution $f(x)$. In other words, statistical entropy is equal, within an additive constant, to the expected value of the complexity, or individual entropy. This theorem becomes interesting, for example, in the case when f is the distribution of a string of n independent, identically distributed random variables. In this case, the entropy is proportional to n, but c_f is $O(\log n)$.

Inductive inference. Kolmogorov's introduction of complexity was motivated by information theory and the problem of randomness. Solomonoff introduced algorithmic complexity independently of Kolmogorov, but for a different reason: inductive inference.

Solomonoff defined a notion close to our a priori probability $\mathbf{M}(x)$ of a finite sequence defined above in order to use it for inductive inference. Given the initial segment x, he suggests estimating the conditional probability that the next segment will be y by the expression

$$\frac{\mathbf{M}(xy)}{\mathbf{M}(x)}. \tag{3.11}$$

This formula can be considered a technical formalization of Occam's razor: "Entities should not be multiplied beyond necessity." This principle is generally interpreted as the prescription: "Find the simplest theory accounting for x and then infer y according to it." We encounter difficulties in formalizing this principle for probabilistic theories μ since a trade-off occurs between the complexity of μ and the defect of randomness in x with respect to μ.

Formula (3.11) solves the trade-off in the case of large x and simple μ. Let μ be an arbitrary computable measure. This case includes all computable sequences as well as many Bernoulli sequences. If the length of y is fixed and the length of x grows to infinity, then we have

$$\frac{\mathbf{M}(xy)/\mathbf{M}(x)}{\mu(xy)/\mu(x)} \to 1$$

with μ-probability 1, that is, the conditional a priori probability is almost always asymptotically equal to the conditional probability. A similar statement is proved in [43].

The quantity in (3.11) is uncomputable and thus impractical. Still, it is very helpful in discussions of the theoretical possibilities of inductive inference.

Lower bounds on computational complexity. Random finite strings are not compressible by any simple algorithm. This idea is the basis of an important class of mathematical results making use of description complexity. The results themselves concern the (time, memory, etc.) complexity of computations, and description complexity enters the proofs only where the technical advantage of its use is overwhelming. For the desired computation, an incompressible input is chosen. The proof shows that if the computation uses too little time or memory, there is a way to compress the input, contrary to the assumption. Typical examples can be found in [32, 33, 28, 24]. The survey [25] and the book [26] provide an overview of the subject.

The philosophers' stone. It follows from the result of Barzdin' discussed above that an explicit definition of a random string uses at least two quantifiers. Such a definition was given in [48], but the simplest is given by Chaitin in [10]. Let U be an optimal prefix computer. Chaitin defines Ω to be the probability that U halts on a random (fair coin tossing) input string. Chaitin proved that the complexity $K^{(p)}(\Omega^n)$ of the initial segments of Ω grows like $n + O(1)$, and Schnorr proved that such a complexity growth is equivalent to randomness. Levin's algorithmic data processing inequality (3.3) suggests that we will never know more than a handful of bits of Ω. An oracle answering questions on the halting problem χ would be of invaluable help in answering any kind of mathematical question. The metaphysical fascination of Ω comes from the fact that Ω is computationally equivalent to χ, but contains the same information in maximally compressed form.

Let us summarize. First, we will never learn more than a handful of those tantalizingly information-packed bits of Ω. Second, even if we had an oracle supplying all bits of Ω, we would not make any practical use of them, since the time to decompress Ω to χ grows nonrecursively. On the other hand, with enough patience, from the first n bits of Ω we can recover the proofs (or refutations) of all provable (or refutable) n-bit assertions. These issues are explored in [4] and in several articles by Chaitin and account for much of the popular interest in Kolmogorov's complexity.

REFERENCES

[1] BARRON, A. R. (1985). The strong ergodic theorem for densities: Generalized Shannon–Mcmillan–Breiman theorem. *Ann. Probab.* **13** 1292–1303.

[2] BARZDIN', YA. M. (1968). The complexity of programs to determine whether natural numbers not greater than n belong to a recursively enumerable set. *Soviet Math. Dokl.* **9** 1251–1254.

[3] BENNETT, C. H. (1988). Logical depth and physical complexity. In *Universal Turing Machine: A Half-Century Survey* (R. Herken, ed.) 227–257. Oxford Univ. Press, Oxford.

[4] BENNETT, C. H. and GARDNER, M. (1979). The random number omega bids fair to hold the mysteries of the universe. *Scientific American* **241**(5) 20–34.
[5] BOGOLYUBOV, N. N., GNEDENKO, B. V. and SOBOLEV, S. L. (1983). Andrei Nikolaevich Kolmogorov (on his eightieth birthday). *Russian Math. Surveys* **38**(4) 9–27.
[6] CHAITIN, G. J. (1966). On the length of programs for computing binary sequences. *J. Assoc. Comput. Mach.* **13** 547–569.
[7] CHAITIN, G. J. (1969). On the length of programs for computing binary sequences: Statistical considerations. *J. Assoc. Comput. Mach.* **16** 145–159.
[8] CHAITIN, G. J. (1974). Information-theoretic limitations of formal systems. *J. Assoc. Comput. Mach.* **21** 403–424.
[9] CHAITIN, G. J. (1975). Randomness and mathematical proof. *Scientific American* **232**(5) 47–52.
[10] CHAITIN, G. J. (1975). A theory of program-size formally identical to information theory. *J. Assoc. Comput. Mach.* **22** 329–340.
[11] DOBRUSHIN, R. L. (1959). General formulation of Shannon's basic theorems of the theory of information. *Uspekhi Mat. Nauk* **14**(6) 3–104. (In Russian.)
[12] DOBRUSHIN, R. L. (1988). Information theory. In *Information Theory and the Theory of Algorithms* 254–257. Nauka, Moscow. (Third volume of *Selected Works* of A. N. Kolmogorov; in Russian.)
[13] GÁCS, P. (1974). On the symmetry of algorithmic information. *Soviet Math. Dokl.* **15** 1477–1480.
[14] GÁCS, P. (1983). On the relation between descriptional complexity and algorithmic probability. *Theoret. Comput. Sci.* **2** 71–93.
[15] GEL'FAND, I. M. and YAGLOM, A. M. (1957). On the calculation of the quantity of information about a random function contained in another such function. *Uspekhi Mat. Nauk* **12**(1) 3–52. (In Russian; English translation in *Amer. Math. Soc. Transl.* **12** 199–246, 1959.)
[16] GRAY, R. M., ORNSTEIN, D. S. and DOBRUSHIN, R. L. (1980). Block synchronization, sliding-block coding, unvulnerable sources, and zero error codes for discrete noisy channels. *Ann. Probab.* **8** 639–674.
[17] KIEFFER, J. C. (1989). *Basic and Advanced Information Theory*. To appear.
[18] KOLMOGOROV, A. N. (1968). Logical basis for information theory and probability theory. *IEEE Trans. Inform. Theory* **IT-14** 662–664.
[19] KOTZ, S. (1966). *Recent Results in Information Theory*. Methuen, London. (First published in *J. Appl. Probab.*, 1966.)
[20] LAPLACE, P. S. (1951). *A Philosophical Essay on Probabilities* 16–17. Dover, New York.
[21] LEVIN, L. A. (1973). On the notion of a random sequence. *Soviet Math. Dokl.* **14** 1413–1416.
[22] LEVIN, L. A. (1974). Laws of information conservation (nongrowth) and aspects of the foundations of probability theory. *Problems Inform. Transmission* **10**(3) 206–210.
[23] LEVIN, L. A. (1984). Randomness conservation inequalities: Information and independence in mathematical theories. *Inform. and Control* **61** 15–37.
[24] LI, M. and VITÁNYI, P. M. B. (1988). Tape versus queue and stacks: The lower bounds. *Inform. and Computation* **78** 56–85.
[25] LI, M. and VITÁNYI, P. M. B. (1989). Kolmogorov complexity and its applications. In *Handbook of Theoretical Computer Science* (J. van Leeuwen, ed.). North-Holland, Amsterdam. To appear.
[26] LI, M. and VITÁNYI, P. M. B. (1990). *Introduction to Kolmogorov Complexity and Its Applications*. Addison-Wesley, Reading, Mass. To appear.
[27] LINKOV, U. N. (1965). Evaluation of ε-entropy of random variables for small ε. *Problems Inform. Transmission* **1**(2) 12–18.
[28] MAAS, W. (1985). Combinatorial lower bound arguments for deterministic and nondeterministic Turing machines. *Trans. Amer. Math. Soc.* **292** 675–693.
[29] MARTIN-LÖF, P. (1966). The definition of random sequences. *Inform. and Control* **9** 602–619.
[30] OREY, S. (1985). On the Shannon–Perez–Moy theorem. In *Particle Systems, Random Media and Large Deviations* (R. Durrett, ed.). *Contemp. Math.* **41** 319–327. Amer. Math. Soc., Providence, R.I.
[31] ORNSTEIN, D. S. (1970). Bernoulli shifts with the same entropy are isomorphic. *Adv. in Math.* **4** 337–352.

[32] PAUL, W. (1979). Kolmogorov's complexity and lower bounds. In *Fundamentals of Computation Theory* (L. Budach, ed.) 325–334. Akademie, Berlin.

[33] PAUL, W., SEIFERAS, J. and SIMON, J. (1981). An information-theoretical approach to time-bounds for on-line computation. *J. Comput. System Sci.* **23** 108–126.

[34] PEREZ, A. (1957). Sur la théorie de l'information dans le cas d'un alphabet abstrait. In *Trans. First Prague Conference on Information Theory, Statistical Decision Functions and Random Processes* 209–244. Academia, Prague.

[35] PINSKER, M. S. (1960). *Information and Stability of Random Variables and Processes*. Izd. Akad. Nauk, Moscow. (English translation, 1964.)

[36] SCHNORR, C. P. (1971). *Zufälligkeit und Wahrscheinlichkeit. Eine algorithmische Behandlung der Wahrscheinlichkeitstheorie. Lecture Notes in Math.* **218**. Springer, New York.

[37] SCHNORR, C. P. (1973). Process complexity and effective random tests. *J. Comput. System Sci.* **7** 376–388.

[38] SHANNON, C. E. (1948). A mathematical theory of communication. *Bell Systems Tech. J.* **27** 379–423. (Published as a book with the same title with a tutorial by W. Weaver, Univ. Illinois Press, Urbana, 1949.)

[39] SHANNON, C. E. (1960). Coding theorems for a discrete source with a fidelity criterion. In *Information and Decision Processes* (R. E. Machol, ed.) 93–126. McGraw-Hill, New York.

[40] SHEN, A. KH. (1983). The concept of Kolmogorov (α, β)-stochasticity and its properties. *Soviet Math. Dokl.* **28** 295–299.

[41] SINAI, YA. G. (1959). On the concept of entropy for a dynamic system. *Dokl. Akad. Nauk SSSR* **124** 768–771. (In Russian.)

[42] SOLOMONOFF, R. J. (1964). A formal theory of inductive inference. I, II. *Inform. and Control* **7** 1–22, 224–254.

[43] SOLOMONOFF, R. J. (1978). Complexity-based induction systems: Comparisons and convergence theorems. *IEEE Trans. Inform. Theory* **IT-24** 422–432.

[44] VILLE, J. (1939). *Etude critique de la notion de collectif*. Gauthier-Villars, Paris.

[45] VITUSHKIN, A. G. (1961). *The Theory of Transmission and Processing of Information*. Pergamon, New York. (Original Russian ed., Fizmatgiz, Moscow, 1959.)

[46] VON MISES, R. (1964). *Mathematical Theory of Probability and Statistics*. Academic, New York.

[47] V'YUGIN, V. V. (1987). On the defect of randomness of a finite object with respect to measures with given complexity bounds. *Theory Probab. Appl.* **32** 508–512.

[48] ZVONKIN, A. K. and LEVIN, L. A. (1970). The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms. *Russian Math. Surveys* **25**(6) 83–124.

THOMAS M. COVER AND ROBERT M. GRAY
DEPARTMENTS OF STATISTICS
AND ELECTRICAL ENGINEERING
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305

PETER GACS
IBM ALMADEN RESEARCH CENTER
650 HARRY ROAD
SAN JOSE, CALIFORNIA 95120-6099