Complexity Distortion Theory

Daby M. Sow, *Member, IEEE,* and Alexandros Eleftheriadis, *Senior Member, IEEE*

***Abstract*—Complexity distortion theory (CDT) is a mathematical framework providing a unifying perspective on media representation. The key component of this theory is the substitution of the decoder in Shannon's classical communication model with a universal Turing machine. Using this model, the mathematical framework for examining the efficiency of coding schemes is the algorithmic or Kolmogorov complexity. CDT extends this framework to include distortion by defining the complexity distortion function. We show that despite their different natures, CDT and rate distortion theory (RDT) predict asymptotically the same results, under stationary and ergodic assumptions. This closes the circle of representation models, from probabilistic models of information proposed by Shannon in information and rate distortion theories, to deterministic algorithmic models, proposed by Kolmogorov in Kolmogorov complexity theory and its extension to lossy source coding, CDT.**

***Index Terms*—Kolmogorov complexity, Markov types, rate distortion function, universal coding.**

I. Introduction

CURRENT methodologies for media representation have their roots in systems designed and conceived several decades ago. The theoretical foundation for addressing such representation problems was established in 1948, when Shannon introduced information and rate distortion theories. These theories ignore the meaning of the message considered "irrelevant" [13]. They are based on the measure theoretical concept of probability that was proposed by Kolmogorov in 1929.

In this setting, source observations are produced by a source $\{X_n\}, -\infty < n < \infty$, in a probability space $(A, \mathcal{A}, \mu)$, with source alphabet A, σ-algebra $\mathcal{A}$, and probability measure μ. Focusing on source coding, let $\hat{A}$ be the reproduction alphabet with σ-algebra $\hat{\mathcal{A}}$. A^n and $\hat{A}^n$ denote, respectively, the nth-fold Cartesian product of A and $\hat{A}$.

The main goal in lossy universal compression is to represent efficiently a source observation $x_1^n = (x_1, x_2, \ldots, x_n) \in A^n$, with a reproduction sequence $y_1^n \in \hat{A}^n$ without assuming the knowledge of μ at both ends of the communication system.

In general, efficiency is measured in terms of distortion and rate. To simplify the discussion, we assume that distortion is measured with a family of single-letter distortion measures

$$d_n\left(x_1^n, y_1^n\right) = \frac{1}{n}\sum_{i=1}^{n} d(x_i, y_i), \;\; x_1^n \in A^n, \;\; y_1^n \in \hat{A}^n \tag{1}$$

where $d: A \times \hat{A} \to [0; +\infty)$.

Manuscript received June 2, 2000; revised November 12, 2002. The material in this paper was presented in part at the IEEE International Symposium on Information Theory, Ulm, Germany, June/July 1997.

D. M. Sow is with the IBM T. J. Watson Research Center, Hawthorne, NY 10532 USA (email: sowdaby@us.ibm.com).

A. Eleftheriadis is with the Department of Electrical Engineering, Columbia University, NewYork, NY 10027 USA (email: eleft@ee.columbia.edu).

Communicated by T. Fuja, Associate Editor At Large.

Digital Object Identifier 10.1109/TIT.2002.808135

To formalize the notion of rate, a variable length code C_n is defined as

1) an encoding function $\Phi_n: A^n \to B = \{0,1\}^*$, where $\{0,1\}^*$ is the set of all binary sequences;
2) a decoding function $\Psi_n: B \to \hat{A}^n$.

Let $l: B \to \mathcal{N}$ be the length function that associates to each string its length, $\mathcal{N}$ being the set of natural numbers. The rate of the code C_n is defined as the expectation of the length of $\Phi_n(X_1^n)$ divided by n. Let the expected value of a random variable f with respect to a measure μ be defined by $E_\mu[f] = \int f(x)d\mu(x)$, then the rate of the code C_n is

$$R = \frac{E_\mu\left[l(\Phi_n(X_1^n))\right]}{n}. \tag{2}$$

For a fixed distortion level D, C_n is called D-semifaithful or D-admissible if $d_n(x_1^n, \Psi_n\left(\Phi_n(x_1^n)\right)) \leq D$ for all $x_1^n \in A^n$.

One of the many important contributions that Shannon made to this field is the definition of the rate distortion function (RDF) $R(D)$. He proved that it is a lower bound for rates of D-admissible codes

$$\liminf_{n\to\infty} \frac{E_\mu[l(\Phi_n(X_1^n))]}{n} \geq R(D). \tag{3}$$

He also proved the existence of codes achieving this lower bound.[1] These proofs revolve around the probabilistic concept of typicality [2].

In practice, stationary and ergodic assumptions are made on the source probability measure to design coding algorithms with performances close to the theoretical limits defined by Shannon. Stationarity assumes that the statistics of the source do not change with time and ergodicity justifies the approximation of the source distribution from time averages obtained from a single infinite source observation.

Unfortunately, in many practical situations, models are not inherently probabilistic [17] and source observations are finite. For example, in image representation, source observations usually contain a significant amount of spatial regularities that escape all probabilistic models. Furthermore, the finiteness of the observed data (e.g., an image) does not allow us to fit probabilistic models for the representation simply because there is not enough physical evidence to estimate probabilities from time averages. In these situations, the stationary and ergodic assumptions are difficult to justify and it becomes difficult to attach a physical meaning to the measure theoretical concept of probability [6].

Interestingly, Kolmogorov introduced in [8] the notion of Kolmogorov complexity to measure the amount of randomness in individual objects. He found the need to measure it from

[1]See [1] for precise definitions and statements.

lengths of descriptions of objects on a universal Turing machine (UTM). The Kolmogorov complexity of a sequence $x_1^n \in A^n$ is defined as the length of the shortest program written for a UTM, that halts and outputs x_1^n. There exist many variants of the Kolmogorov complexity in the literature. Throughout this paper, we will use the prefix Kolmogorov complexity introduced by Levin and Gacs. We denote it by $C(\cdot)$.

In this paper, we follow Kolmogorov's approach to measure information. We put computational constraints on the decoding function Ψ_n by replacing the decoder in Shannon's communication system with a UTM. We investigate the effect of these constraints on the limits of compression. Since its introduction, Kolmogorov complexity has grown substantially. It has concentrated, however, on data compression, i.e., lossless representation. We extend it to include distortion for lossy representations. This step allows the application of the complexity framework to audio-visual information representation, where the introduction of (ideally nonperceptible) distortion is a key step for allowing nontrivial compression.

The structure of the paper is as follows. In Section II, we introduce the main entity in complexity distortion theory (CDT), namely, the complexity distortion function (CDF) and state the conditions for the equivalence between the CDF and the RDF. General equivalences have been proposed in the lossless and lossy cases in [9], [11], [19], and [20]. In this paper, we simplify these results significantly and extend them with pragmatic considerations for the coding of finite objects. A formal proof of the equivalence is proposed in Section III. We end this paper with important remarks on CDT in Section IV.

II. Complexity Distortion Theory (CDT)

A key component of Kolmogorov complexity theory is the introduction of a computational model at the decoding end of the communication system. In contrast, RDT makes no assumption on the structure of the decoding function Ψ_n. Assuming that Ψ_n is computable,[2] it can be implemented by a Turing machine (TM). By replacing the decoder in Shannon's communication system with a UTM, we can emulate any decoding function by programming the UTM. If we allow distortion in the representation, it is very natural to measure the coding performances with an extension of the Kolmogorov complexity to the semifaithful case and introduce a new mathematical entity, the CDF.

To define the CDF, we contrast it with the RDF and consider the following standard construction sometimes used to introduce the RDF. Fix $y_1^n \in \hat{A}^n$. The set

$$B_D^n = \{x_1^n \in A^n \colon d_n(x_1^n, y_1^n) \le D\}$$

is called a (D, n)-ball with center y_1^n or simply a D-ball if n is understood. Remark that since A and $\hat{A}$ could differ, some D-balls could be empty. To avoid this, we define $D_{\min}$ as a real number equal to the infimum amount of distortion that can be obtained using any coding/decoding scheme Φ_n, Ψ_n. Obviously, if $\hat{A} \supseteq A$, $D_{\min} = 0$. But in practice it is more common to have $\hat{A} \subset A$. From now on, we always assume that such a real number $D_{\min}$ exists and that $D \ge D_{\min}$. If $A = \hat{A}$, and if $D = 0$, the code is noiseless or faithful. Let $G(S)$ be a union of D-balls covering $S \subseteq A^n$. $G(S)$ is called a D-cover of S. Let $N(D, S)$ denote the minimum number of D-balls needed to cover S. By definition, the operational RDF $R^o(D)$ is given by

$$R^o(D) = \lim_{\epsilon \to 0} \lim_{n \to \infty} R_n(D, \epsilon) \tag{4}$$

where

$$R_n(D, \epsilon) = \min_{S: \mu(S) \ge 1-\epsilon} \frac{1}{n} \log_2 N(D, S). \tag{5}$$

$R_n(D, \epsilon)$ is a measure of the ratio between the number of bits needed to index the minimum number of balls required to cover S, S being a subset of A^n of measure greater than $1 - \epsilon$. The intuition behind this statement is that each element of S can be represented "D-semifaithfully" by the index of the D-ball containing the element. Gray, Neuhoff, and Ornstein have shown in [5] that this definition of the operational RDF is equivalent to the definition of the RDF [1]. From now on, we drop the subscript on it and represent it with $R(D)$.

The CDF is introduced in a similar manner. In a D-ball centered around x_1^n this time, let $\mathcal{Q}_D(x_1^n)$ be the sequence in $\hat{A}^n$ with the smallest Kolmogorov complexity. If we have many sequences inside the ball with small complexity, we pick the closest one to x_1^n, according to d_n. If we still have more than one candidate, we list them in a lexicographic order and arbitrarily pick the sequence with the smallest index in the list

$$\mathcal{Q}_D(x_1^n) = \arg \min_{y_1^n \in \hat{A}^n : d_n(x_1^n, y_1^n) \le D} C(y_1^n). \tag{6}$$

Definition 1: The CDF is defined as

$$C_D(x_1^n) = C\left(\mathcal{Q}_D(x_1^n)\right). \tag{7}$$

There are two striking differences between the RDF and the CDF.

1) The CDF is a deterministic quantity that does not depend on probabilities. In contrast, the RDF relies on the probability measure associated to the source.

2) The CDF describes the best element inside a D-ball and does not ignore the information content. The RDF does not describe the content of any elements of a D-ball. It simply indexes one of these balls following the general information-theoretic principle that the actual meaning of messages is irrelevant to the communication problem [13].

Despite these fundamental differences between the RDF and the CDF, these two lossy measures of information are equivalent for a very wide class of information sources. In the lossless setting, Zvonkin and Levin were the first one to propose equivalences between Kolmogorov complexity and the entropy in [21]. The following theorem is formulated without a proof in [21].

Theorem 1: For a stationary ergodic source $(A, \mathcal{A}, \mu)$, with a probability measure μ

$$\lim_{n \to \infty} \frac{C(x_1^n)}{n} = H(\mu), \qquad x_1^n \in A^n \mu\text{-a.s.} \tag{8}$$

where $H(\mu)$ denotes the entropy rate of the source.

[2] See [3] for a precise definition of computable or recursive functions.

We extend this result to the lossy case.

Theorem 2: For a stationary ergodic source $(A, \mathcal{A}, \mu)$, with a probability measure μ

$$\lim_{n\to\infty} \frac{C_D(x_1^n)}{n} = R(D), \qquad x_1^n \in A^n \mu\text{-a.s.} \tag{9}$$

In both of these theorems, the ergodic decomposition of stationary processes allows us to drop the ergodicity assumption if we allow the introduction of an expectation term on the Kolmogorov complexity or on the CDF.

The importance of these results cannot be overstated. It clearly shows that despite introducing a mechanical structure at the decoding end of Shannon's communication system, we do not lose much in terms of performance.

III. Proof of Equivalence

In this section, we prove Theorem 2. Since Theorem 1 is a restriction of Theorem 2 to the lossless case where $D_{\min} = D = 0$, we do not prove it.

Proof: To prove Theorem 2, we proceed in two steps. In the first step, we construct a code with rate equal to the RDF. This code can be decoded by a UTM. We then invoke source coding results in the second step, to show that no code, decodable by a UTM or not, can have a rate below the RDF. To accomplish the first step, we prove the following lemma.

Lemma 1:

$$\limsup_{n\to\infty} \frac{C_D(x_1^n)}{n} \le R(D), \qquad x_1^n \in A^n \mu\text{-a.s.} \tag{10}$$

Proof: We use the concept of Markov types as proposed by Shields in [14] to design universal codes that can be decoded by a UTM. Let $k \in \mathcal{N}$. Following [14], the Markov k-type is defined by sliding a window of length $k+1$ along x_1^n and counting frequencies. These frequencies are then used to define empirical transition probabilities. Let $\bar{x}$ be an infinite periodic sequence defined by

$$\bar{x}_{tn+i} = x_i, \qquad i = 1, 2, \ldots, n, \; t = 0, 1, 2, \ldots. \tag{11}$$

For each integer $0 \le k < n$ and for each $a_1^k \in A^k$, define

$$\hat{p}_k\left(a_1^k\right) = \lim_{L\to\infty} \frac{1}{L} \left|\left\{i : \bar{x}_{i+1}^{i+k} = a_1^k, \; 0 \le i < L\right\}\right|. \tag{12}$$

where the notation $|A|$ denotes the cardinal of the set A. The periodicity of $\bar{x}$ implies that

$$\hat{p}_k\left(a_1^k\right) = \frac{1}{n} \left|\left\{i : \bar{x}_{i+1}^{i+k} = a_1^k, \; 0 \le i < n\right\}\right| \tag{13}$$

and

$$\hat{p}_{k-1}\left(a_1^{k-1}\right) = \sum_{a_k \in A_0} \hat{p}_k(a_1^k), \qquad a_1^{k-1} \in A^{k-1}. \tag{14}$$

The Markov k-type of x_1^n is the Markov measure $\hat{\mu}^{(k)}$ with state space A^k, stationary probabilities $\hat{p}_k\left(a_1^k\right)$ and transition probabilities $p(\cdot \mid \cdot)$ defined by

$$p\left(a_{k+1} \mid A_1^k\right) = \frac{\hat{p}_{k+1}\left(a_1^{k+1}\right)}{\hat{p}_k\left(a_1^k\right)}. \tag{15}$$

The type class $T_k\left(x_1^n\right)$ is defined as the set of all sequences of length n with Markov type equal to the Markov type of x_1^n.

To encode x_1^n efficiently, recall that a D-cover of a set $S \subset A^n$ is a collection G of D-balls covering S. The sequence x_1^n can be represented by the following two-part code proposed in [14].

1) The index of the Markov k-type of x_1^n is transmitted.
2) Choose a D-cover $G_k\left(T_k(x_1^n)\right)$ with the least cardinal among all D-covers of $T_k\left(x_1^n\right)$. The second part of the code is an address of an element of $G_k\left(T_k(x_1^n)\right)$ that contains x_1^n.

The first part of the code requires only a number of bits $m = o(n)$ to represent the index of the Markov k-type of x_1^n, If $k = \left\lfloor \frac{1}{2} \log_{|A|} n \right\rfloor$ [14], $\lfloor \alpha \rfloor$ denoting the greatest integer $\le \alpha$. The second part of the code requires at most $\log_2 \left|G_k\left(T_k(x_1^n)\right)\right|$ bits. It is shown in [14] that the rate of this two-part code converges almost surely to the RDF of the source and we have the fol- lowing result:

$$\lim_{n\to\infty} \frac{l\left(\Phi_n(x_1^n)\right)}{n} = R(D) \text{ a.s.} \tag{16}$$

Φ_n being the encoding function corresponding to the code construction presented above. See [14] for a proof of this result. This code is prefix free [14] and is a valid program for a prefix-free TM. Its decoding operations are clearly computable. To see this, note that the first part of the representation encodes the type class $T_k\left(x_1^n\right)$. With this description of $T_k\left(x_1^n\right)$, the decoder can generate all sequences belonging to $T_k\left(x_1^n\right)$. A program computing this Markov-type class requires shift, addition, and division operations which are all Turing computable. Since the composition of computable functions is also computable (see [3, Theorem 2.1, p. 36]), there is a TM able to decode this part of the representation. From the second part of the description, the machine has to index the D-ball that contains the sequence x_1^n. The final representation of x_1^n is obtained by identifying the right reproduction sequence y_1^n associated with the indexed D-ball. Consequently, using (16) and Definition 1 of the CDF, we have

$$\limsup_{n\to\infty} \frac{C_D(x_1^n)}{n} \le \limsup_{n\to\infty} \frac{l\left(\Phi_n(x_1^n)\right)}{n} = R(D), \qquad x_1^n \in A^n \mu\text{-a.s.} \tag{17}$$

proving Lemma 1. □

In the second part of the proof of Theorem 2, we prove the following lemma.

Lemma 2:

$$\liminf_{n\to\infty} \frac{C_D(x_1^n)}{n} \ge R(D), \qquad x_1^n \in A^n \mu\text{-a.s.} \tag{18}$$

Proof: Recall [7, Theorem 2]. This theorem states that for any semifaithful code C_n with encoding function Φ_n

$$\liminf_{n\to\infty} \frac{l\left(\Phi_n(x_1^n)\right)}{n} \ge R(D), \qquad x_1^n \in A^n \mu\text{-a.s.} \tag{19}$$

Note that in (19), there is no computational restriction on the decoding process and the decoding function Ψ_n. Any

D-admissible encoding function that represents x_1^n at a rate equal to $\frac{C_D(x_1^n)}{n}$ will verify (19). Hence, we conclude that

$$\liminf_{n\to\infty} \frac{C_D\left(x_1^n\right)}{n} \geq R(D), \qquad x_1^n \in A^n \ \mu\text{-a.s.} \tag{20}$$

we have proved Lemma 2. □

From Lemma 1 and Lemma 2, Theorem 2 follows. ■

We end this section with two important remarks on the proof.

Remark 1: Any universal coding scheme could have been used to prove Lemma 1. We used the code construction proposed in [14] because it highlights the separation between the model and the data. The model is represented by the first part of the code, the index of the Markov k-type. This model is clearly probabilistic. The data part of the representation is the index of the D-ball that contains the source sequence x_1^n. In the representation of x_1^n, the size of the model is asymptotically negligible when compared to the size of the data. In practice, when objects are finite, the size of the model could be a significant part of the code.

Remark 2: The simplicity of the proof of Theorem 2 is due to the power of source coding theorems. In fact, while Lemma 1 is derived from a well-known universal coding scheme, Lemma 2 is a restriction of a theorem proved by Kieffer (in [7, Theorem 2]) to computable decoding functions. These well-known information-theoretic results shield us from the complexity of this equivalence.

IV. Concluding Remarks

In this paper, we have extended the notion of Kolmogorov complexity to lossy descriptions of information by defining the CDF. We have shown that this function is almost surely equal to the RDF for infinite observations of stationary-ergodic sources. Using types, we highlight the separation principle between model and data. The model part can be identified as the routine used to describe the source probability distribution. The data part is essentially an index of a D-ball containing the source observation. We end this paper with three key points on the scope of CDT.

First, restricting the decoding function to be computable does not reduce the performances of the system. In fact the Church–Turing thesis [10] guarantees that any coding technique belongs to the set of computable functions, from traditional entropy approaches to modern approaches like fractal and model-based coding. These modern techniques do not rely on statistical models of the source. The result is a unification of all coding algorithms under the same mathematical framework.

The second point addresses the computability of the Kolmogorov complexity. The Kolmogorov complexity is not computable. While there exists no TM able to compute $K\left(x_1^n\right)$ for any sequence x_1^n, it is possible to approximate it, as shown in [10]. Such approximations are obtained by adding computational resource bounds on the decoding UTM [16]. Similarly, we are not aware of any systematic way to estimate source distributions from finite observations [12]. Jeffrey [6] takes this observation even further and asserts that the concept of probability has no physical meaning. These observations did not prevent the field of source coding to blossom with the development of efficient coding algorithms. In statistics, the minimum description length principle is often used to select efficient models from general observations. This principle by-passes the concept of probabilities considered artificial and has been successfully applied to develop many practical statistical learning techniques. Similarly, we believe that it could be beneficial to develop programmatic representations based on Kolmogorov's approach for the design of practical media representation systems. In practice, programmatic representation techniques are already starting to gain momentum in audio representation with the MPEG-4 Structured Audio standard [4]. It is shown in [4] that such programmatic representations of sound outperform today's probabilistic audio representation schemes.

The last point that we would like to emphasize is that Shannon's communication system does not bound the amount of computational resources available at the decoding end. In practice, the computational power of the decoding device is always bounded. With arbitrary computational resource bounds on the decoding UTM, the equivalence may not hold. Putting such computational resource bounds on the decoder transforms it into a finite-state machine (FSM). In fact, any real computer has only a finite amount of computational power. The key here is that a TM may have infinite tapes but at each time of computation, only a finite amount of memory and time has been used. The amount of computational resources is unbounded but finite at any time. Indeed, as mentioned in [18], "we can even go so far as to compute a finite bound on the maximum amount of memory constructible out of the material composing the known universe and be tempted to claim that, for all practical purposes, FSMs serve as models of effective procedures. Unfortunately, the finiteness of an FSM is not a mathematically useful concept. The finiteness constraints can often get in the way of a concise, understandable description of an algorithm. We often, for example, write programs in which we assume that all our intermediate results will fit in their respective variables. Even though the assumption may not always be strictly justified, by making it, we greatly reduce the amount of detail we have to handle and this is certainly desirable."

From a pure compression point of view, the FSM model is sufficient, under stationary and ergodic assumptions. In this case, traditional compression algorithms (like Shannon–Fano, Huffman, arithmetic, and Lempel–Ziv coding) yield asymptotically optimal representations. This fact is clearly shown in [13], where encoders and decoders are modeled by transducers which are FSMs with a finite amount of memory. But when descriptions of content at higher semantic levels are desired, the content and the clarity of the descriptions become important factors that force us to extend the FSM model to the TM model with infinite memory, as we commonly do with real computers which are physically FSMs with a large but finite amount of memory.

ACKNOWLEDGMENT

The authors thank Prof. Dimitris Anastassiou and Prof. Predrag Jelenkovic for their helpful suggestions. They also would like to thank the anonymous reviewers of this manuscript for their constructive comments.

REFERENCES

[1] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Englewood Cliffs, NJ: Prentice-Hall, 1971.
[2] T. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
[3] M. Davis, *Computability and Unsolvability*. New York: Dover, 1982.
[4] E. Scheirer, "Structured audio, Kolmogorov complexity, and generalized audio coding," *IEEE Trans. Speech Audio Processing*, vol. 9, pp. 914–931, Aug. 2001.
[5] R. M. Gray, D. L. Neuhoff, and D. S. Ornstein, "Non-block source coding with a fidelity criterion," *Ann. Probab.*, vol. 3, pp. 478–491, 1975.
[6] R. C. Jeffrey, "Mises redux," in *Basic Problems in Methodology and Linguistics*, R. Butts and J. Hintikka, Eds. Dordrecht, The Netherlands: Reidel, 1977, pp. 213–222.
[7] J. Kieffer, "Sample converses in source coding theory," *IEEE Trans. Inform. Theory*, vol. 37, pp. 263–268, Mar. 1991.
[8] A. N. Kolmogorov, "Three approaches to the quantitative definition of information," *Probl. Inform. Transm.*, vol. 1, no. 1, pp. 1–7, 1965.
[9] S. K. Leung-Yan-Cheong and T. M. Cover, "Some equivalence between Shannon entropy and Kolmogorov complexity," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 331–339, May 1978.
[10] M. Li and P. Vitanyi, *An Introduction to Kolmogorov Complexity and its Apllications*, 2nd ed. New York: Springer-Verlag, 1997.
[11] J. Muramatsu and F. Kanaya, "Distortion-complexity and rate-distortion function," *IEICE Trans. Fundamentals*, vol. E77-A, pp. 1224–1229, 1994.
[12] J. Rissanen, *Stochastic Complexity in Statistical Inquiry*. Singapore: World Scientific, 1989.
[13] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, 1948.
[14] P. C. Shields, "Universal almost sure data compression using Markov types," *Probl. Contr. Inform. Theory*, vol. 19, pp. 269–277, 1990.
[15] D. Sow and A. Eleftheriadis, "Complexity distortion theory," in *Proc. IEEE Int. Symp. Information Theory*, Ulm, Germany, June/July 1997, p. 188.
[16] ——, "Approximation of the resource bounded complexity distortion function," in *Proc. 2000 IEEE Int. Symp. Information Theory*, Sorrento, Italy, 2000, p. 206.
[17] P. Vitanyi and M. Li, "Minimum description length induction, Bayesianism and Kolmogorov complexity," *IEEE Trans. Inform. Theory*, vol. 46, pp. 446–464, Mar. 2000.
[18] W. Wulf, M. Shaw, and P. N. Hilfinger, *Fundamental Structures of Computer Science*. Readidng, MA: Addison-Wesley, 1981.
[19] E. H. Yang, "The proof of Levin's conjecture," *Chinese Sci. Bull.*, vol. 34, pp. 1761–1765, 1989.
[20] E. H. Yang and S. Shen, "Distortion program-size complexity with respect to a fidelity criterion and rate distortion function," *IEEE Trans. Inform. Theory*, vol. 39, pp. 288–292, Jan. 1993.
[21] A. K. Zvonkin and L. A. Levin, "The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms," *Russian Math. Surv.*, vol. 25, no. 6, pp. 83–124, 1970.