On Algorithmic Rate-Distortion Function

Nikolai Vereshchagin
Dept. Math. Logic & Theor. Algor.,
Moscow State Univ., Russia.
Email: ver@mccme.ru

Paul Vitanyi
CWI, Amsterdam, the Netherlands.
Email: paulv@cwi.nl

Abstract— **We develop rate-distortion theory in the Kolmogorov complexity setting. This is a theory of lossy compression of individual data objects, using the computable regularities of the data.**

I. Introduction

The classical theory of lossy compression was initiated by Shannon in [14], where a rate distortion function is associated to every random variable X and every distortion measure. A similar approach for lossy compression of individual objects (and not random variables) was proposed by Yang and Shen in [19], where a rate-distortion function r_x was assigned to each individual object x. The function r_x maps every real a to the minimal possible Kolmogorov complexity of y with distortion at most a with x. (Kolmogorov complexity is the accepted absolute measure of the information content of an individual finite object. It gives the ultimate limit on the number of bits resulting from lossless compression of the object—more precisely, the number of bits from which effective lossless decompression of the object is possible.)

We call Yang and Shen's approach algorithmic, as Kolmogorov complexity is an algorithmic notion. The papers [19], [12], [16] relate the classical and algorithmic approach. They prove that if the object x is obtained by random sampling of a sequence of i.i.d. random variables $X_1, \dots, X_n$ then with high probability its rate distortion function r_x is close to the Shannon's rate distortion function of the random variable X_i, and some generalizations of this statement to arbitrary ergodic random sources.

In the present paper we describe all possible shapes of Yang and Shen's rate-distortion curve r_x (with some accuracy, and under some conditions on the distortion measure, Theorems 2 and 3). Our description implies that nonrandom data can have a variety of different curves. It is easy to see that one is generally interested in the behavior of lossy compression on complex structured nonrandom data, like pictures, movies, music, while the typical unstructured random data like noise (represented by the Shannon curve) is discarded (we are not likely to want to store it).

Second, we formulate a new theoretical problem related to the practice of lossy compression. It has been implicitly addressed before in proposals for denoising of signals by lossy compression [11], [13], [3], [4]. Is it the case that a lossily compressed representation that realizes least compressed size at a given distortion with respect to the source object also captures an appropriate part of the "essence" or "meaning" of that object? Clearly, this question cannot be well posed in the Shannon setting. We show that in algorithmic setting this question is answered in the affirmative for every distortion measure. More specifically, we prove that if y witnesses the rate-distortion curve at point a (that is, y has minimal Kolmogorov complexity among all y's with distortion at most a with respect to x) then the randomness deficiency of x in the set of all x' with distortion at most a with respect to y is small, Theorem 5. We are deliberately vague here until we introduce the appropriate formalism in Section V.

II. Preliminaries

Compared to the classical information theory setting, in the algorithmic rate-distortion theory we dispense with sequences of random variables, and we also generalize the distortion measures from single-letter distortion measures to full generality. We start from some set $\mathcal{X}$. Its elements will be called *source words*. Suppose we want to communicate source words x from $\mathcal{X}$ using a code of at most r bits for each such word. (We call r the *rate*.) If 2^r is smaller than $|\mathcal{X}|$, then this is clearly impossible. However, for every x we can use a representation y that in some sense is close to x. For example, assume that we want to communicate a real number $x \in [0;1]$. Using r bits we are able to communicate a representation y that is a rational number at distance $\leq 2^{-r}$ from x.

Assume that the representations are chosen from a set $\mathcal{Y}$, possibly different from $\mathcal{X}$. We call its elements *destination words*. Assume furthermore that we are given a function d from $\mathcal{X} \times \mathcal{Y}$ to the reals, called the *distortion measure*. It measures the lack of fidelity, which we call *distortion*, of the destination word y against the source word x. (In our example, this is the Euclidean distance between x and y.)

In the Shannon theory [14], [15], [1], [2], we are given a random variable X with values in $\mathcal{X}$. Thus every source word appears with a given probability. The goal is, for a given rate r and distortion a, to find an encoding function E, with a range of cardinality at most 2^r, such that the expected distortion between a source word $x \in \mathcal{X}$ and the corresponding destination word $y = E(x)$ is at most a. The set P of all pairs $\langle r, a \rangle$ for which this is possible is called the *rate-distortion profile of the random variable* X. For every distortion a, we consider the minimum rate r such that the pair $\langle r, a \rangle$ is an element of the profile of X. This way we obtain the *rate-*

distortion function of the random variable X:

$$r(a) = \min\{r : \langle r, a \rangle \in P\}. \tag{1}$$

Here, like in [19], [12], [16], we are interested in what happens for individual source words, irrespective of the probability distribution on $\mathcal{X}$ induced by a random variable. To this end we use Kolmogorov complexity $K(y)$ and conditional Kolmogorov complexity $K(x|y)$, as defined in [9], and the textbook [10]. In our treatment it is not essential which version of Kolmogorov complexity we use, the plain one or the prefix one. We assume that the set of destination words $\mathcal{Y}$ consists of finite objects and thus $K(y)$ is defined for all $y \in \mathcal{Y}$. For every $x \in \mathcal{X}$ we want to identify the set of pairs $\langle r, a \rangle$ such that there is $y \in \mathcal{Y}$ with $K(y) \le r$ and $d(x, y) \le a$. The set P_x of all such pairs will be called the *rate-distortion profile of the source word x*. For every distortion a consider the minimum rate r such that the pair $\langle r, a \rangle$ belongs to the profile of x. This way we obtain the *rate-distortion function of the source word x* [19]:

$$r_x(a) = \min\{K(y) : d(x, y) \le a\}.$$

The quantity $r_x(a)$ for Hamming distortion was independently defined in the paper [8] under the notation $C_a(x)$ (Hamming distortion is defined in the example following Theorem 3).

It is often more intuitive to consider, for every rate r, the minimum distortion a such that the pair $\langle r, a \rangle$ belongs to the profile of x. This way we obtain the *distortion-rate function of the individual word x*:

$$d_x(r) = \min\{d(x, y) : K(y) \le r\}.$$

It is straightforward from the definitions that $d_x(r)$ is a sort of "inverse" from $r_x(a)$.

III. Related work

In Shannon's paper [15], it is assumed that $\mathcal{X} = \mathcal{A}^n$, $\mathcal{Y} = \mathcal{B}^n$ are the sets of strings of certain length n over finite alphabets $\mathcal{A}, \mathcal{B}$. (Here we ignore the generalizations to infinite or continuous sets.) The distortion measure has the form $d^n(x, y) = \sum_{i=1}^n d(x_i, y_i)/n$ where d maps pairs of letter from $\mathcal{A} \times \mathcal{B}$ to the reals (the single-letter distortion measure). The random variable X is taken as $X = X_1, \dots, X_n$ where the X_i's are independent random variables identically distributed over $\mathcal{A}$. For every n we obtain the rate-distortion function $r^n(a)$. Shannon shows that the limit $\lim_{n\to\infty} r^n(a)/n$ exists (we denote it by $R(a)$) and determines its non-constructive description in terms of $\mathcal{A}, \mathcal{B}, d, a, X_i$:

Theorem 1 (Shannon): $R(a) = \min\{I(X_i : Y) \mid \mathbf{E}\, d(X_i, Y) \le a\}$ where $I(X_i : Y) = H(Y) - H(Y|X_i)$ denotes the common information in X_i and Y and $\mathbf{E}\, d(X_i, Y)$ stands for the expected value of $d(X_i, Y)$.

More general distortion measures and random variables, that were studied later, are treated in [1], [2].

The papers [19], [12], [16], using the same i.i.d. assumptions on $\mathcal{X}, \mathcal{Y}, X, d$, establish the value of the rate-distortion functions $r_x(a)$ for specific x's in A^n and compare them with $R(a)$. It is shown that the limit $\lim_{n\to\infty} r_x(a)/n$ is equal to $R(a)$ almost surely (i.e. with probability 1), and that the limit of the expectation of $r_x(a)/n$ is also equal to $R(a)$. These results show that if x is obtained from a random source, then with high probability its rate-distortion function is close to the function $nR(a)$. Our results will show that for individual data x (containing structure and regularity), there are many different shapes of $r_x(a)$, and all of them are very different from that of $nR(a)$.

Ziv [20] considers also a rate-distortion function for individual data. The rate-distortion function is assigned to every infinite sequence ω of letters of a finite alphabet $\mathcal{A}$ (and not to a finite object, as in the present paper). The source words x are prefixes of ω and the encoding function is computed by a finite state transducer. Kolmogorov complexity is not involved.

In [17], we treated the special case of *list decoding distortion* (related to model selection in statistics): $\mathcal{X} = \{0, 1\}^n$, and $\mathcal{Y}$ is the set of all finite subsets of $\{0, 1\}^n$; the distortion function $d(x, y)$ is equal to $\lceil \log |y| \rceil$ if y contains x, and is equal to infinity otherwise (we need $\lceil \log |y| \rceil$ of extra information to identify x given y). This type of code was pioneered by [5], [6], [18]. The associated distortion is such a special case that the proofs and techniques do not generalize. Nonetheless, and surprisingly so to the authors, the results generalize by more powerful techniques to somewhat weaker versions, yielding a completely general algorithmic rate-distortion theory.

IV. Possible shapes of the rate-distortion curve

Given $\mathcal{X}, \mathcal{Y}$ and the distortion measure d, satisfying certain mild properties, we determine all possible shapes of the graph of r_x for the different source words $x \in \mathcal{X}$, within a small margin of error. In contrast to the Shannon case where one obtains a single profile P and rate-distortion function $r(a)$ (for every $\mathcal{X}, \mathcal{Y}, X$ and distortion measure d), we establish that different x's can lead to different profiles.

Although some of our results can be naturally generalized to infinite or even uncountable sets $\mathcal{X}$, like the segment $[0; 1]$, for simplicity we will assume further that $\mathcal{X}$ is a finite subset of a fixed set of finite objects $\mathcal{U}$. We assume that a computable bijection σ between $\mathcal{U}$ and the set of all binary strings is fixed. The Kolmogorov complexity $K(x)$ for $x \in \mathcal{U}$ is defined as $K(\sigma(x))$. The bijection σ induces a well order on $\mathcal{U}$ and hence on $\mathcal{X}$. We make the same assumptions about $\mathcal{Y}$, the fixed set of finite objects $\mathcal{Y}$ is included to is denoted by $\mathcal{V}$. We will assume that the distortion measure d takes only non-negative rational values. (This is not an essential restriction for this paper, since every real can be approximated by a rational and all bounds in the paper hold to limited precision only.) Let D denote the range of d and $d_{\max}$ the maximal element of D.

A *ball of radius a in $\mathcal{X}$* is a set of the form $B_y(a) = \{x \in \mathcal{X} : d(x, y) \le a\}$. The destination word y is called the *center* of the ball. Let $B(a)$ stand for maximal cardinality of a ball of radius a in $\mathcal{X}$. We assume that $B(0) = 1$, and for every $x \in \mathcal{X}$ there is $y \in \mathcal{Y}$ with $B_y(0) = \{x\}$. (This is equivalent to the statement that, for every $x \in \mathcal{X}$, there is a $y \in \mathcal{Y}$ with distortion $d(x, y) = 0$.) The *graph* of distortion measure d is the set of triples $\langle x, y, d(x, y) \rangle$ ordered lexicographically. Note that this list identifies also $\mathcal{X}, \mathcal{Y}$ and D. Let $K(d)$ stand for the Kolmogorov complexity of the graph of d. Let α denote the *covering coefficient* related to $\mathcal{X}, \mathcal{Y}, d$, defined as the minimal number that satisfies the following condition: for all $0 \le a < a'$, every ball of radius a' in $\mathcal{X}$ can be covered by at most $\alpha B(a')/B(a)$ balls of radius a. (In the examples considered in this paper the covering coefficient is of order $\log^{O(1)} |\mathcal{X}|$.)

The following Theorems 2 and 3 describe all possible shapes of the rate-distortion functions r_x of individual $x \in \mathcal{X}$.

THEOREM 2: For all $\mathcal{X}, \mathcal{Y}, d$ and $x \in \mathcal{X}$ we have

$$r_x(d_{\max}) \le \varepsilon, \tag{2}$$

$$r_x(0) = K(x) \pm \varepsilon, \tag{3}$$

$$0 \le r_x(a') - r_x(a) \le \log(B(a)/B(a')) + \varepsilon \text{ for all } a' < a, \tag{4}$$

where $\varepsilon = O(\log \alpha + K(d) + \log |D| + \log\log |\mathcal{X}|)$.

We say that $\varepsilon = O(\delta)$ where ε, δ are functions of $\mathcal{X}, \mathcal{Y}, d$ if $|\varepsilon| \le c\delta + C$, with c an absolute constant, and C depends on the choice of the optimal description method in the definition of Kolmogorov complexity, and on the choice of computable bijections between the set of binary strings and the universes $\mathcal{U}, \mathcal{V}$.

Property (4) implies that $r_x(a)$ is a rather smooth function provided $\log B(a)$ is so. The similar property doesn't hold for the "inverse" $d_x(r)$. Theorem 3 will establish that $d_x(r)$ can decrease a lot for r increasing only a little (see Fig 1). Theorem 2 shows that the rate-distortion function is confined within the following bounds:

$$K(x) - \log B(a) - \varepsilon \le r_x(a) \le \log |\mathcal{X}| - \log B(a) + \varepsilon.$$

The right-hand bound is obtained by letting $a = d_{\max}$ in Equation (4). The left-hand bound can be derived by letting $a' = 0$ in (4), or can be shown also by a simple direct argument. If x is a random element of $\mathcal{X}$, that is, $K(x) = \log |\mathcal{X}| \pm \varepsilon$, then the lower and upper bounds for $r_x(a)$ are close to each other and we can conclude that $r_x(a) = \log |\mathcal{X}| - \log B(a) \pm \varepsilon$. If x is not such a random element, then there are many possible behaviors of $r_x(a)$, and the next theorem shows that they are all realizable.

THEOREM 3: Let $r : D \to \mathbb{N}$ satisfy (2) by having $r(d_{\max}) = 0$, satisfy (3) by having $r(0) = k$, and satisfy (4) with $\varepsilon = 0$. Then there is a source word $x \in \mathcal{X}$ of complexity $k \pm \delta$ such that

$$|r(a) - r_x(a)| \le \delta, \tag{5}$$

where $\delta = O(\sqrt{\log |\mathcal{X}|} \log(2\alpha) + K(d) + K(r))$ and $K(r)$ stands for the complexity of the graph of r, which is the set of pairs $\langle a, r(a) \rangle$ $(a \in D)$ ordered lexicographically.

The proof of this theorem is similar to the proof of its special case for list decoding distortion in [17]. However, there is an essential difference: in the case of list decoding distortion we can let x be equal to the first x satisfying the inequality $r_x(a) \ge r(a) - \delta$ for all $a \in D$. In the general case this does not work any more: we construct x together with balls ensuring the inequalities $r_x(a) \le r(a) + \delta$ for all $a \in D$.

We will illustrate the variability of the shapes by the instructive example of *Hamming distortion:* The set of source words $\mathcal{X}$ and the set of destination words $\mathcal{Y}$ are both equal to the set $\{0,1\}^n$ of all binary strings of length n. The distortion function d is defined by $d(x,y) = t/n$ if y differs from x in t bit positions. For all $a \le \frac{1}{2}$ the term $\log B(a)$ differs by at most $O(\log n)$ from $nH(a)$, where $H(a) = a \log 1/a + (1-a) \log 1/(1-a)$ is the Shannon entropy function. For $a \in [\frac{1}{2}; 1]$ the function $B(a)$ is almost constant: $n - 1 \le \log B(a) \le n$. The terms $\log\log |\mathcal{X}|$, $\log |D|$, $K(d)$ are all of order $O(\log n)$. As to the term $\log \alpha$, it also is of the same order, as the following lemma shows.

LEMMA 1: For all $a \le a' \le \frac{1}{2}$ every Hamming ball of radius a' can be covered by at most $\alpha B(a')/B(a)$, where $\alpha = \text{poly}(n)$, Hamming balls of radius a. (We denote by poly(n) a polynomial of n.)

Thus in the Hamming distortion case the following corollary of Theorems 2 and 3 describes all possible shapes of rate-distortion function.

COROLLARY 1: For every x of length n the rate-distortion function r_x of x satisfies the inequalities:

$$r_x(\frac{1}{2}) = O(\log n), \quad r_x(0) = K(x) + O(\log n) \tag{6}$$

$$0 \le r_x(a) - r_x(a') \le n(H(a) - H(a')) + O(\log n) \tag{7}$$

for all $0 \le a < a' \le \frac{1}{2}$. On the other hand, let r be a function mapping the set $\{0, 1/n, 2/n, \dots, \frac{1}{2}\}$ to the naturals satisfying the condition (7) without $O(\log n)$ term and such that $r_x(\frac{1}{2}) = 0$ and $r_x(0) = k$. Then there is a string x of length n and complexity $k \pm O(\log n)$ such that $r_x(a) = r(a) + O(\sqrt{n} \log n + K(r))$ for all $a \le \frac{1}{2}$.
(The bound (7) was announced, without a complete proof, in the paper [7].)

For example, we can apply Corollary 1 to the function $r(a)$ shown in Fig. 1. The rate-distortion graph of the string x existing by Theorem 1 is in the strip of size $O(\sqrt{n} \log n)$ of the graph of $r(a)$. Therefore $r_x(a)$ is almost constant on

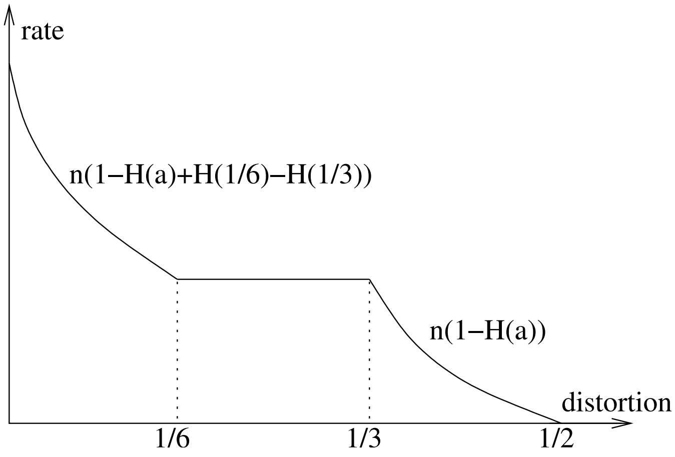

Fig. 1. A possible shape of the rate-distortion function for Hamming distortion

the segment $[\frac{1}{6};\frac{1}{3}]$. Allowing the distortion to increase on this interval, all the way from $\frac{1}{6}$ to $\frac{1}{3}$, so allowing $n/6$ incorrect extra bits, we still cannot decrease the rate. This means that the distortion-rate function $d_x(r)$ of x drops from $\frac{1}{3}$ to $\frac{1}{6}$ near the point $r = n(1 - H(\frac{1}{3}))$, exhibiting a very non-smooth behavior.

Other examples we have analyzed are *list decoding distortion*, and *Euclidean distortion*. In the former the accuracy $\varepsilon = O(\log n)$ in Theorem 2 and $\delta = O(\sqrt{n}\log n + K(r))$ in Theorem 3 (as shown in [17], the accuracy in Theorem 3 can be improved to $O(\log n + K(r))$ in this example). In Euclidean distortion we let $\mathcal{X} = \mathcal{Y}$ be the set of rational numbers in the segment $[0,1]$ having n binary digits and we let $d(x,y)$ be equal the 0 if $x = y$ and to $n + 1 + \lceil \log |x - y| \rceil$ otherwise. The accuracy is again of order $O(\log n)$ in Theorem 2 and $O(\sqrt{n}\log n + K(r))$ in Theorem 3.

V. A THEORETICAL SUPPORT OF DENOISING VIA COMPRESSION

Consider the following idealized procedure of denoising via compression. Given the data $x \in \mathcal{X}$ to denoise and $a \in D$ (the amount of noise to remove) do the following:
1. Find a destination word y with distortion at most a with respect to x having the minimum Kolmogorov complexity, which is equal to $r_x(a)$. (Note that such y is very hard to find. First, Kolmogorov complexity is not computable. Second, there are exponentially many $y \in Y$ with distortion at most a with respect to x. Thus we consider a very idealized version of denoising via compression.)
2. If $K(y) + \log |B_y(a)|$ is close to $K(x)$ then output y. (In this case the ball $B_y(a)$ is called *an algorithmic sufficient statistic* of x.) Otherwise the procedure fails.

Note that $K(y) + \log |B_y(a)|$ cannot be less than $K(x)$. Indeed, consider the code for x consisting of the minimum length description of y and the index i of x in the ball $B_y(a)$. As the total length of this two-part code cannot be greater than the Kolmogorov complexity of x, we obtain the inequality $K(y) + \log |B_y(a)| \geq K(x)$. (We ignore here additive terms of order $O(K(d) + \log |D| + \log\log |X|)$.)

If we have no idea how much noise is there in the data, we try all $a \in D$ to find maximum a such that the procedure succeeds. The corresponding y is called *a minimal sufficient statistic* of x, as $K(y)$ is minimal.

The following two questions arise: (1) Assume that the procedure succeeds on inputs x and a, that is, $B_y(a)$ is a sufficient statistic of x. Why do we think that all the information present in x but not in y is a noise? (2) Assume that for some (unknown) destination word z the source word x is chosen at random in the ball $B_z(a)$. Is it true that in this case the procedure succeeds with high probability on inputs x and a? In other words, is it true that with high probability $r_x(a) + \log B(a) - K(x)$ is small?

The first question is easy to answer. If $B_y(a)$ is an algorithmic sufficient statistic of x then the complexity of the index i of x in the ball $B_y(a)$ conditional to $B_y(a)$ is close to the binary length l of i. Indeed,

$$K(x) \leq K(B_y(a)) + K(i|B_y(a)) \leq K(y) + l$$
$$\leq K(y) + \log |B_y(a)| \approx K(x),$$

hence all the inequalities here are equalities. That is, in the two-part representation $\langle y, i \rangle$ of x the second part has no regularities and can be considered as a random noise.

The second question is answered in affirmative by the following theorem.

THEOREM 4: Let x is chosen at random in the ball $B_z(a)$ (all elements in the ball are equiprobable). Then the probability of the event

$$r_x(a) + \log B(a) - K(x) > \delta$$

is less than $2^{-\delta+\varepsilon}$. Here $\varepsilon = O(K(d) + \log\log |\mathcal{X}| + \log |D| + \log \gamma)$, where γ is the maximal ratio $|B_u(a)|/|B_v(a)|$ over all a, u, v. Note that in all our examples γ is bounded by a constant.

This theorem easily follows from its algorithmic version, which uses the notion of randomness deficiency. The *randomness deficiency of x in a set $A \subset \mathcal{X}$* containing x is defined as

$$\delta(x|A) = \log |A| - K(x|A),$$

where A in the conditional of $K(x|A)$ is given as the list of elements of A (in the fixed order on $\mathcal{X}$). The following properties of randomness deficiency explain its meaning:
(1) Randomness deficiency is almost non-negative, that is, $\delta(x|A) \geq C$ for some constant C and all $x \in A$. Indeed, every element x of A can be described by its $\log |A|$-bit index in A conditional to A. Thus $K(x|A) \leq \log |A| + O(1)$.
(2) For all A, the randomness deficiency of almost all elements of A is small: the number of $x \in A$ with $\delta(x|A) > \beta$ is less than $|A|2^{-\beta}$. Indeed, $\delta(x|A) > \beta$ implies $K(x|A) < \log |A| - \beta$. Since there are at most $2^{\log |A| - \beta}$ programs of less than $\log |A| - \beta$ bits, the number of x's satisfying the inequality cannot be larger.
(3) Every element with small deficiency in A possesses every property possessed by majority of elements in A (we identify a property of elements of A with a subset of A consisting of all elements having the property). More specifically, assume that B is a subset of A with $|B| \geq (1 - 2^{-\beta})|A|$ and $K(B|A) \leq \gamma$. Then the randomness deficiency of all $x \notin B$ in A satisfies $\delta(x|A) > \beta - \gamma - O(\log\log |A|)$, which is large if β is large and γ is small.

The randomness deficiency measures our disbelief that x can be obtained by random sampling in A (where all elements of A are equiprobable). By property (2) with high probability the randomness deficiency of an element randomly chosen in A is small. On the other hand, if $\delta(x|A)$ is small, then there is no way to refute the hypothesis that x was obtained by random sampling in A: every such refutation is based on a simply described property possessed by a majority of A but not by x. Here it is important that we consider only simply described properties, as otherwise we can refute the hypothesis by exhibiting the property $B = A \setminus \{x\}$.

THEOREM 5: Let x belong to a ball $B_z(a)$. Then

$$r_x(a) + \log B(a) - K(x) \leq \delta(x|B_z(a)) + \varepsilon,$$

where ε is of the same order that in Theorem 4.

An easy calculation shows that

$$\delta(x|A) \leq K(A) + \log|A| - K(x)$$

for all $A \ni x$. Thus if x is chosen at random in a ball $B_z(a)$ and y has minimal Kolmogorov complexity among all destination words with distortion a with x then $\delta(x|B_y(a))$ is only a little larger than $\delta(x|B_z(a))$. Indeed,

$$\begin{aligned}\delta(x|B_y(a)) &\leq K(B_y(a)) + \log|B_y(a)| - K(x)\\ &\leq \delta(x|B_z(a)) + \varepsilon.\end{aligned}$$

This gives an additional support to denoising via compression: the hypothesis "x is chosen at random in $B_y(a)$" is almost as plausible as the hypothesis "x is chosen at random in $B_z(a)$".

VI. AN ALGORITHMIC ANALOG OF THEOREM 1

The proofs of the results of the previous section are based on the following theorem.

THEOREM 6: For every ball $B_y(a) \ni x$ there is a ball $B_z(a) \ni x$ with $K(z) \leq I(x:y) + O(\log K(y)) + \varepsilon$ where $I(x:y) = K(y) - K(y|x)$ stands for the information in x about y. Hence $r_x(a) = \min\{I(x:y) \mid d(x,y) \leq a\} \pm \varepsilon + O(\log \max_{y \in \mathcal{Y}} K(y)$. Here $\varepsilon = O(K(d) + \log\log|\mathcal{X}| + \log|D|)$.

It is worth to remark that this theorem is very similar the above Shannon's Theorem 1. However the proofs of Theorem 1 and 6 are quite different. The proof of Theorem 6 is based on the following Theorem 7, which is interesting in its own right. Let $\mathcal{A}$ be a family of finite subsets of $\mathcal{X}$ (for instance, distortion balls). A set in $\mathcal{A}$ is called a *model of x* if it contains x and $\mathcal{A}(x)$ denotes the set of all models of x. The Kolmogorov complexity $K(A)$ of A is defined as Kolmogorov complexity of the list of elements of A in the fixed order on $\mathcal{X}$.

THEOREM 7: If $|\mathcal{A}(x)| \geq 2^k$ then x has a model in $\mathcal{A}$ with $K(A) \leq \log|\mathcal{A}| - k + \varepsilon$ where $\varepsilon = O(\log\log|\mathcal{X}| + K(\mathcal{X}) + KE(\mathcal{A}) + \log\log|\mathcal{A}|)$.

Here $KE(\mathcal{A})$ stands for the Kolmogorov complexity of "enumerating $\mathcal{A}$", defined as follows. Fix a computable bijection $A \mapsto [A]$ between the set of all finite subsets of $\mathcal{A}$ and the naturals. Then $KE(\mathcal{A})$ is the minimal size of a non-halting program that prints all elements in the set $\{[A] \mid A \in \mathcal{A}\}$ in some order (it is essential that we do not know the moment when the last element in the set is printed out).

Previously an analog of this theorem was known in the case when $\mathcal{A}$ is the class of all sets of fixed cardinality 2^c and of complexity not exceeding a certain level r. For $c = 0$ this is an exercise (Ex. 4.3.8 in [10]): if a string x has at least 2^k descriptions of length at most r (p is called a description of x if $U(p) = x$ where U is an optimal description method), then $K(x) \leq r - k + O(\log k + \log r)$. The paper [17] proves this for all c: if a binary string belongs to at least 2^k sets A of cardinality 2^c and complexity $K(A) \leq r$, then x belongs to a set B of cardinality 2^c and complexity $K(B) \leq r - k + O(\log r + \log k + \log c)$.

ACKNOWLEDGMENT

The work of N. Vereshchagin was done in part while visiting CWI and was supported in part by the grant 02-01-22001 from Russian Federation Basic Research Fund. The work of P. Vitanyi was supported by the EU fifth framework project QAIP, IST–1999–11234, the NoE QUIPROCONE IST–1999–29064, the ESF QiT Programme, and the EU Fourth Framework BRA NeuroCOLT II Working Group EP 27150.

REFERENCES

[1] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*, Prentice-Hall, Englewood Cliffs, NJ, 1971.

[2] T. Berger, J.D. Gibson, Lossy source coding, *IEEE Trans. Inform. Th.*, 44:6(1998), 2693–2723.

[3] S.C. Chang, B. Yu, M. Vetterli, Image denoising via lossy compression and wavelet thresholding, *Proc. Intn'l Conf. Image Process. (ICIP'97)*, 1997, 604-607 in Volume 1.

[4] D. Donoho, The Kolmogorov sampler, *Annals of Statistics*, submitted.

[5] P. Elias, List decoding for noisy channels. *Wescon Convention Record,* Part 2, Institute for Radio Engineers (now IEEE), 1957, 94–104.

[6] P. Elias, Error-correcting codes for List decoding, *IEEE Trans. Inform. Th.*, 37:1(1991), 5–12.

[7] L. Fortnow, T. Lee, N. Vereshchagin, Kolmogorov Complexity with Error, *Proc. Symposium Theoretical Aspects of Comput. Science 2006,* Lecture Notes in Computer Science, vol. 3884 (2006) 137–148

[8] R. Impagliazzo, R. Shaltiel, and A. Wigderson. Extractors and pseudo-random generators with optimal seed length. In *Proceedings of the 32nd ACM Symposium on the Theory of Computing*, pages 1–10. ACM, 2000.

[9] A.N. Kolmogorov, Three approaches to the quantitative definition of information, *Problems Inform. Transmission* 1:1 (1965) 1–7.

[10] M. Li and P.M.B. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer-Verlag, 1997. 2nd Edition.

[11] B.K. Natarajan, Filtering random noise from deterministic signals via data compression, *IEEE Trans. on Signal Processing*, 43:11(1995), 2595-2605.

[12] J. Muramatsu, F. Kanaya, Distortion-complexity and rate-distortion function, *IEICE Trans. Fundamentals*, E77-A:8(1994), 1224–1229.

[13] N. Saito, Simultaneous noise suppression and signal compression using a library of orthonormal bases and the minimum description length criterion, Pp. 299–324 in *Wavelets in Geophysics*, E. Foufoula-Georgiou, P. Kumar, Eds., Academic Press, 1994.

[14] C.E. Shannon. The mathematical theory of communication. *Bell System Tech. J.*, 27:379–423, 623–656, 1948.

[15] C.E. Shannon. Coding theorems for a discrete source with a fidelity criterion. In *IRE National Convention Record, Part 4*, pages 142–163, 1959.

[16] D.M. Sow, A. Eleftheriadis, Complexity distortion theory, *IEEE Trans. Inform. Th.*, 49:3(2003), 604–608.

[17] N.K. Vereshchagin and P.M.B. Vitányi, Kolmogorov's Structure functions and model selection, *IEEE Trans. Inform. Theory*, 50:12(2004), 3265- 3290.

[18] J.M. Wozencraft, List decoding. *Quarterly Progress Report*, Research Laboratory for Electronics, MIT, Vol. 58(1958), 90–95.

[19] E.-H. Yang, S.-Y. Shen, Distortion program-size complexity with respect to a fidelity criterion and rate-distortion function, *IEEE Trans. Inform. Th.*, 39:1(1993), 288–292.

[20] J. Ziv, Distortion-rate theory for individual sequences, *IEEE Trans. Inform. Th.*, 26:2(1980), 137–143.