

Algorithmic Rate-Distortion Theory

Nikolai Vereshchagin*

Paul Vitányi†

December 7, 2005

Abstract

We propose and develop rate-distortion theory in the Kolmogorov complexity setting. This gives the ultimate limits of lossy compression of individual data objects, taking all effective regularities of the data into account.

1 Introduction

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. A similar absolute approach is needed for lossy compression, that is, a rate-distortion theory giving the ultimate effective limits for individual finite data objects. We give natural definitions of the rate-distortion functions of individual data (independent of a random source producing those data). We analyze the possible shapes of the rate-distortion graphs for all data and all computable distortions. The classic Shannon rate-distortion curve corresponds approximately to the individual curves of typical (random) data from the postulated random source, while the nonrandom data have completely 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). Finally, we formulate a new problem related to the practice of lossy compression. Is it the case that a lossily compressed representation that realizes least distortion of the source object at a given rate also captures an appropriate part of the “essense” or “meaning” of that object? We are deliberately vague here until we introduce the appropriate formalism. Clearly, this question cannot be well posed in the Shannon setting, where we deal with expected distortion in communication, but explicitly not with “meaning.” We show that in our setting this question is answered in the affirmative for every distortion measure that satisfies a certain parsimony-of-covering property.

2 Preliminaries

Compared to the classical information theory setting 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

*Department of Mathematical Logic and Theory of Algorithms, Faculty of Mechanics and Mathematics, Moscow State University, Leninskie Gory, Moscow, Russia 119992. Email: ver@ccme.ru.

†CWI, Kruislaan 413, 1098 SJ Amsterdam, The Netherlands. Email: Paul.Vitanyi@cwi.nl.

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 [11, 12, 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* :

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

Here, like in [16, 9, 13], 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 [6], and the textbook [8]. In our treatment it is not essential which version of 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 $K(y) \leq r$ and $d(x, y) \leq a$. 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* :

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

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* :

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

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

2.1 Related work

In Shannon’s paper [12], 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 $\lim_{n \rightarrow \infty} r^n(a)/n$ exists and determines its non-constructive description in terms of $\mathcal{A}, \mathcal{B}, d, a, X_i$. In the next paragraph this limit will be denoted by $R(a)$. More general distortion measures and random variables, that were studied later, are treated in [1, 2].

The papers [16, 9, 13], 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 $\lim_{n \rightarrow \infty} r_x(a)/n$ 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 (ergodic stationary) source, then with high probability its rate distortion function is close to the function $nR(a)$. The reason is as follows. If, for example, we flip a fair coin n times, then the resulting sequence has maximal Kolmogorov complexity (within a small additive constant) with overwhelming probability. As we show in Section 5, for every distortion measure (of $O(\log n)$ Kolmogorov complexity), the individual rate-distortion functions of all these x 's will be close to the Shannon rate-distortion function (trivially achieving the “almost surely” part), and since almost all probability is concentrated on these x 's they also coincide with the expectation up to negligible error. Our results will show that for nonrandom x (containing structure and regularity), there are many different shapes of r_x , and all of them are vastly different from that of R .

Ziv [17] considers also a kind of the 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 [14], we treated the special case of *list decoding distortion* (related to model selection in statistics): $X = \{0, 1\}^n$, and 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 is such a special case of distortion 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.

2.2 Results

Given \mathcal{X}, \mathcal{Y} and the distortion measure d , satisfying certain mild properties, we determine all possible shapes of the profiles P_x for the different source words $x \in \mathcal{X}$, within a small margin of error. (Equivalently, we establish all possible shapes of the graph of r_x up to that margin.) 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), the new approach establishes that different x 's can lead to vastly different profiles. This analysis may illuminate the dichotomy between Shannon's rate-distortion theory and practical lossy compression where one tries to exploit syntactical and other regularities in an application domain. A successful lossy compressor succeeds in capturing a good distortion measure and lossily compresses all individual data appropriately; many of them below the rate-distortion curve of an assumed random variable (since the latter gives the point-wise average). 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 $d(x, y) = a/n$ if y differs from x in a bit positions. Other examples we have analyzed are *list decoding distortion*, and *Euclidean distortion*.

A second issue, mentioned already in the introduction, has not been addressed before in either theory or practice of lossy compression (as far as the authors know). This is the question whether a destination word with least distortion at a given rate also represents the “meaning” of the source word in an appropriate manner. It is difficult (if possible at all) to formulate this question in Shannon's theory, but here it is related to the Kolmogorov complexity notion of “randomness deficiency” of source word x with respect to destination word y (the precise

definition of randomness deficiency is given in Section 4). We define the notion of a “property” of a source word x , and “simple properties” represented by a destination word y . We show that if the randomness deficiency is about 0, then all simple properties represented by the destination word are possessed by x . We show that a destination word y with $K(y) \leq r$, that witnesses the distortion-rate graph of $d_x(r)$ ($d(x, y) = d_x(r)$), minimizes also the randomness deficiency of x with respect to y among all y 's of complexity at most r .

Finally, we show that given a random variable X of low complexity, the Shannon distortion-rate function $d(r)$ pointwise equals to the average of $d_x(r \pm \varepsilon)$ for a small ε , the expectation taken over the probabilities of the outcomes of X .

3 Rate-Distortion Graph

Under certain mild restrictions on the considered distortion measures, the shape of the rate-distortion function will follow a fixed pattern, associated with the particular distortion measure involved, and, moreover, every function that follows that pattern is the rate distortion function of some source word $x \in \mathcal{X}$ within a negligible tolerance for that distortion measure. Before formulating this claim we need to list these restrictions and introduce some notation.

Although the results of this section 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} .

The distortion measure d is a function mapping $\mathcal{X} \times \mathcal{Y}$ to the non-negative rationals. (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 . We assume that there is a $y \in \mathcal{Y}$ with distortion $d(x, y) \leq d_{\max}$ for all source words $x \in \mathcal{X}$.

A *ball of radius a in \mathcal{X}* is a set of the form $B_y(a) = \{x \in \mathcal{X} : d(x, y) \leq a\}$. The destination word y is called the *center* of the ball. Note that such a ball, being a subset of \mathcal{X} , may have more than one center. 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 $d(x, y) = 0$.) Note that $B(d_{\max}) = |\mathcal{X}|$.

Let α denote the *covering coefficient* related to $\mathcal{X}, \mathcal{Y}, d$, defined as the minimal number that satisfies the following conditions: (i) For every a the cardinality of every ball of radius a is at least $B(a)/\alpha$; (ii) For every $a > 0$ there is $c < a$ with $B(c) \geq B(a)/\alpha$; and (iii) For all $0 \leq a < a'$, every ball of radius a' in \mathcal{X} can be covered by at most $\alpha B(a')/B(a)$ balls of radius a .

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 .

The results in this paper will usually be precise up to an additive term of order $O(s)$ where

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

We say that $\varepsilon = O(\delta)$ where ε, δ are functions of $\mathcal{X}, \mathcal{Y}, d$ if $|\varepsilon| \leq c\delta + C$, 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} .

For the Hamming distortion example, all the assumptions are satisfied, and s is of order $O(\log n)$ (see Lemma 1 below). For the list decoding distortion, again all the assumptions are satisfied, and s is of order $O(\log n)$. This is a good accuracy, as the the values in question are proportional to $n = \log |\mathcal{X}|$.

3.1 The common properties of rate distortion profile

We first establish three easy properties, Theorem 1, that are satisfied by the rate distortion profile P_x of every source word $x \in \mathcal{X}$. Item (i) is self-explanatory. Item (ii) states that the profile P_x of x contains a pair which is close to the pair $\langle 0, d_{\max} \rangle$ (every word can be easily described with maximal distortion). The set P_x also contains a pair that is close to $\langle K(x), 0 \rangle$ (each word can be exactly described in about $K(x)$ bits), and no pairs $\langle r, 0 \rangle$ with r significantly less than $K(x)$. Item (iii) states that we can decrease the distortion from a to a' at the expense of increasing the rate by at most $\log B(a)/B(a') + O(s)$. That is, if P_x has a pair $\langle r, a \rangle$, then it contains all the pairs $\langle r + \log(B(a)/B(a')) + O(s), a' \rangle$ for $a' < a$. Note that moving in the other direction is impossible in general: in some cases we cannot decrease the rate at the expense of increasing the distortion by any reasonable amount. This is a corollary of Theorem 2 below (see Fig 1).

THEOREM 1. *For some $\varepsilon = O(s)$ and all $\mathcal{X}, \mathcal{Y}, d$ and $x \in \mathcal{X}$:*

(i) *The set P_x is upward closed: if $\langle r, a \rangle \in P_x$, then P_x also contains all $\langle r', a' \rangle$ with $r' \geq r$ and $a' \geq a$.*

(ii) *The set P_x contains the pairs $\langle \varepsilon, d_{\max} \rangle$ and $\langle K(x) + \varepsilon, 0 \rangle$, and does not contain the pair $\langle K(x) - \varepsilon, 0 \rangle$.*

(iii) *If a pair $\langle r, a \rangle \in P_x$ and $0 \leq a' < a$, then also $\langle r + \log(B(a)/B(a')) + \varepsilon, a' \rangle \in P_x$.*

Proof. Item (i) is obvious.

To prove Item (ii), consider the first destination word y_0 with distortion at most d_{\max} with respect to every source word. Its complexity is at most $K(\mathcal{Y}) + O(1) \leq K(d) + O(1)$. By Item (i), this proves that the pair $\langle \varepsilon, d_{\max} \rangle$ is in P_x . Towards the other extreme, we have assumed that for every source word x there is a code word y such that $d(x, y) = 0$. The complexity of the first such y is at most $K(x) + O(s)$, which proves that the pair $\langle K(x) + \varepsilon, 0 \rangle$ is in P_x . On the other hand, every y having distortion 0 with x completely identifies x (given the graph of d). Therefore, $\langle r, 0 \rangle \in P_x$ implies that the witness y satisfies $K(y) \leq r + O(s)$.

To prove Item (iii), let y witness $r_x(a) = K(y)$. Find a covering of $B_y(a)$ by at most $\alpha B(a)/B(a')$ balls of radius a' . Let B' be a ball in the covering containing x . Its center y' can be specified by y and the index i of B' among the covering balls, given the following extra information: the graph of d , and the values of a and a' . Without loss of generality we can assume that both a, a' belong to D (otherwise we can decrease them to the closest values in D). We need also $O(\log \log i) = O(\log \log |\mathcal{X}|)$ extra bits to separate the description of y from the binary representation of i . All the extra information and separator bits are included in $O(s)$. Altogether, $K(y') \leq K(y) + \log B(a)/B(a') + O(s)$ bits. \square

COROLLARY 1.

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

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

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

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 2 below will establish that

$d_x(r)$ can decrease a lot for r increasing only a little (see Fig 1). Corollary 1 shows that the rate distortion function is confined within the following bounds:

$$K(x) - \log B(a) - O(s) \leq r_x(a) \leq \log |\mathcal{X}| - \log B(a) + O(s).$$

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). It can also be argued as follows: Let y witness $r_x(a)$. Then we have the following two part description of x : y , and the index of x in $B_y(a)$. Since the complexity of this description cannot be less than $K(x)$, we obtain

$$K(x) \leq K(y) + \log B(a) + O(s) = r_x(d) + \log B(a) + O(s).$$

If x is a random element of \mathcal{X} , that is, the complexity of x equals $\log |\mathcal{X}| + O(s)$, then the lower and upper bounds for $r_x(a)$ coincide and we can conclude that $r_x(a) = \log |\mathcal{X}| - \log B(a) + O(s)$. If x is not such a random element, then there are many possible behaviors of $r_x(a)$, and we will show in the next section that they are all realizable.

3.2 Every function is realized by some data

Assume that we are given a non-increasing function $r(a) : D \rightarrow \mathbb{N}$ satisfying the constraints in Corollary 1. The graph of r is the set of pairs $\langle a, r(a) \rangle$ ($a \in D$) ordered lexicographically. Is there a source word $x \in \mathcal{X}$ whose distortion function $r_x(a)$ is close to $r(a)$? The next theorem answers this question in the affirmative:

THEOREM 2. *Let $r : D \rightarrow \mathbb{N}$ 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 $k \pm \varepsilon$ such that*

$$|r(a) - r_x(a)| \leq \varepsilon, \tag{5}$$

where $\varepsilon = O(\sqrt{\log |\mathcal{X}|} \log(2\alpha) + K(d) + K(r))$ and $K(r)$ stands for the complexity of the graph of r .

Proof. First note that the complexity of the potential witness x is fixed by r : If $r_x(0) = k \pm \varepsilon$, then by Equation (3) in Corollary 1, we have $K(x) = k \pm \varepsilon$.

Next, we claim that there is a sequence of elements $a_0 = d_{\max} > a_1 > \dots > a_N = 0$ in D , where $N = O(\sqrt{\log |\mathcal{X}|})$, such that every $a \in D \setminus \{a_0, \dots, a_N\}$ belongs to a segment $[a_{i+1}; a_i]$ with $\log B(a_i) - \log B(a_{i+1}) \leq \sqrt{\log |\mathcal{X}|}$. Indeed, chop the interval $[0; \log B(d_{\max})]$ into subintervals of length $\sqrt{\log |\mathcal{X}|}$, and for each subinterval $[b; c]$ consider the set $B^{-1}[b; c] = \{a : b \leq B(a) \leq c\}$. Include the least element and the greatest element of the intersection $D \cap B^{-1}[b; c]$ in the sequence of the a_i 's (if the intersection is empty, do not include anything).

To prove the theorem, it suffices to find an x such that (5) holds for all a_i . To additionally show that the inequality also holds for the remaining a 's, let $[a_{i+1}; a_i]$ be the subsegment containing a . Since both functions $r(a), r_x(a)$ are non-increasing, we have

$$\begin{aligned} r(a) &\in [r(a_i), r(a_{i+1})], \\ r_x(a) &\in [r_x(a_i), r_x(a_{i+1})] \subset [r(a_i) - \varepsilon, r(a_{i+1}) + \varepsilon]. \end{aligned}$$

By the conditions of the theorem, and the property of the sequence of a_i 's, the length of the segment $[r(a_i), r(a_{i+1})]$ is at most

$$r(a_{i+1}) - r(a_i) \leq \log(B(a_i)/B(a_{i+1})) \leq \sqrt{\log |\mathcal{X}|}.$$

Hence $|r(a) - r_x(a)| \leq \sqrt{\log |\mathcal{X}|} + \varepsilon$ and we are done.

To find the desired x , we run the following non-halting algorithm that takes as input the graphs of d and r .

Algorithm: *Outline:* Enumerate all the balls in \mathcal{X} of radiuses a_i and complexities less than $r(a_i) - \varepsilon$, for $0 \leq i \leq N$, respectively. Call such balls *forbidden*, since the desired source word x cannot belong to any such ball. Maintain a variable G containing \mathcal{X} minus the union of all forbidden balls discovered so far.

Construct, in parallel, balls B_0, \dots, B_N of radiuses a_0, \dots, a_N , respectively, as described below. Call them *candidate balls* since their intersection contains the desired word x . These are balls ensuring the inequality $r_x(a_i) \leq r(a_i) + \varepsilon$. Every candidate ball is changed from time to time to maintain the following invariant: for all $i \leq N$ the cardinality of the intersection $B_0 \cap \dots \cap B_i \cap G$ is at least

$$B(a_i)2^{-i-1}\alpha^{-i}. \quad (6)$$

Intialize: Find balls B_0, \dots, B_N of radiuses a_0, \dots, a_N such that $|B_0 \cap \dots \cap B_i| \geq 2^{-i-1}\alpha^{-i}$. We will amply fulfill the requirement by producing balls with a much larger intersection—without the factor of 2^{-i-1} . Let B_0 be the ball of the radius d_{\max} centered at the first element in \mathcal{Y} with distortion at most d_{\max} for every $x \in \mathcal{X}$. The next balls are constructed inductively. Assume that B_0, \dots, B_i are already defined, and the cardinality of their joint intersection is at least $B(a_i)\alpha^{-i}$. To find B_{i+1} , cover B_i by at most $\alpha B(a_i)/B(a_{i+1})$ balls of radius a_{i+1} (this cover exists by the definition of α). This covering also covers the set $B_0 \cap \dots \cap B_i$, which has at least $B(a_i)\alpha^{-i}$ elements. Thus the intersection of at least one covering balls with this set has at least

$$B(a_i)\alpha^{-i}/(\alpha B(a_i)/B(a_{i+1})) = B(a_{i+1})\alpha^{-i-1}$$

elements. Let B_{i+1} be the first such ball.

Enumerate all forbidden balls and maintain invariant: Enumerating forbidden balls we update G . Whenever the invariant (6) becomes false, we change some candidate balls to restore the invariant. Let us prove first that for $i = 0$ the invariant never becomes false. In other words the cardinality of G never gets smaller than half of the cardinality of $B_0 = \mathcal{X}$. Indeed, for fixed i the total cardinality of all the balls of radius a_i and complexity less than $r(a_i) - \varepsilon$ does not exceed $2^{r(a_i)-\varepsilon}B(a_i)$. Since the function $r(a) + \log B(a)$ is monotonic non-decreasing, the total number of elements in all forbidden balls is at most

$$\sum_{i=0}^N 2^{r(a_i)-\varepsilon}B(a_i) \leq (N+1)2^{r(d_{\max})-\varepsilon}B(d_{\max}) = (N+1)2^{-\varepsilon}B(d_{\max}) \ll B(d_{\max}) = |\mathcal{X}|,$$

where we note that $r(d_{\max}) = 0$ by definition, and the last inequality holds provided ε is chosen appropriately.

Now assume that the invariant has become false for some $i > 0$. Let i be the least such index. Since the invariant is true for $i - 1$, the cardinality of the joint intersection G' of all the balls B_1, \dots, B_{i-1} and G is at least $B(a_{i-1})2^{-i}\alpha^{-i+1}$. We update B_i, \dots, B_N as follows. To define the new B_i find a covering of B_{i-1} by at most $\alpha B(a_{i-1})/B(a_i)$ balls of radius a_i . The cardinality of $G' \cap B$ for at least one covering ball B is at least

$$|G'|/(\alpha B(a_{i-1})/B(a_i)) \geq B(a_i)2^{-i}\alpha^{-i}.$$

Let $B_i = B$ be for the first such ball. Note that $B(a_i)2^{-i}\alpha^{-i}$ exceeds twice the threshold required by the invariant. We will use this in the sequel: after each change of any candidate ball B_j the required threshold for j is exceeded at least two times. Using the same procedure find B_{i+1}, \dots, B_N . **End of Algorithm**

Although the algorithm does not halt, at some (unknown) moment the last forbidden ball is enumerated. After this moment the candidate balls are not changed. Take as x any object

in the intersection of G and all the candidate balls. The intersection is not empty, since its cardinality is positive by the invariant. By construction x avoids all the forbidden balls, thus $r_x(a)$ satisfies the required lower bound.

To finish the proof it remains to show that the complexity of every candidate ball B_i (after the stabilization moment) does not exceed $r(a_i) + \varepsilon$ for an appropriate $\varepsilon = O(s)$. Fix $i \leq N$. Consider the description of the final B_i consisting of i , the graphs of d, r , and the total number M of changes to intermediate versions of B_i . The final ball B_i can be algorithmically found from this description by running the Algorithm. Thus it remains to upper bound $\log M$ by something close to $r(a_i)$. Let us prove that the candidate ball B_i is changed at most $2^{r(a_i)+i}$ times. Distinguish two possible cases when B_i is changed: (1) the invariant has become false for an index strictly less than i , and (2) the invariant has become false for i and remained true for all smaller indexes. By induction, the number of changes of the first kind can be upper bounded by $2^{r(a_{i-1})+i-1} \leq 2^{r(a_i)+i-1}$. To upper bound the number of changes of the second kind divide them again in two categories: (2a) after the last change of B_i at least one forbidden ball of radius greater than a_i has been enumerated, (2b) after the last change of B_i no forbidden ball of radius greater than a_i has been enumerated. The number of changes of type (2a) is at most the number of forbidden balls of radiuses $a_j \geq a_i$. By monotonicity of $r(a)$, this is at most $(N+1)2^{r(a_i)-\varepsilon} \ll 2^{r(a_i)}$. Finally, for every change of type (2b), between the last change of B_i and the current one no candidate balls with indexes less than i have been changed and no forbidden balls with radiuses $a_j \geq a_i$ have been enumerated. Thus the cardinality of G has decreased by at least $B(a_i)2^{-i-1}\alpha^{-i}$ due to enumerating forbidden balls with radiuses $a_j < a_i$ (recall that after the last change of B_i the threshold was exceeded at least two times). The total cardinality of forbidden balls of these radiuses does not exceed $N2^{r(a_i)-\varepsilon}B(a_i)$ (we use the monotonicity of $r(a) + \log B(a)$). The number of changes of types (2b) is less than the ratio of this number to the threshold $B(a_i)2^{-i-1}\alpha^{-i}$. Hence it is less than $N2^{r(a_i)-\varepsilon}2^{i+1}\alpha^i$. One can choose $\varepsilon = O(N \log(2\alpha)) = O(\sqrt{|\mathcal{X}|} \log(2\alpha))$ so that this is much less than $2^{r(a_i)}$. The theorem is proved. \square

3.3 Example: Hamming distortion

Recall that in the case of Hamming distortion, $\mathcal{X} = \mathcal{Y}$ is the set $\{0, 1\}^n$ of all binary strings of length n . The distortion measure $d(x, y)$ is equal to the fraction of bits where y differs from x . Let us estimate the term $\log B(a)$ in Corollary 1 and Theorem 2. For all $a \leq 1/2$ 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 $a \in [1/2; 1]$ the function $B(a)$ is almost constant: $n-1 \leq \log B(a) \leq n$.

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. (Although one would think that everything is known about covering Hamming balls, as far as the authors were able to ascertain this is a new combinatorial result.)

LEMMA 1. *For all $a \leq a' \leq 1/2$ every Hamming ball of radius a' can be covered by at most $\alpha B(a')/B(a)$, where α is a polynomial of n , Hamming balls of radius a .*

Proof. The lemma implies that the set of all strings of length n can be covered by at most

$$N = \text{poly}(n)2^n/B(a)$$

balls of radius a . We will first prove this corollary, and then use the same method to prove the full lemma.

Fix a string x . The probability that x is *not* covered by a randomly selected ball of radius a is equal to $1 - B(a)2^{-n}$. Thus the probability that no ball out of N randomly selected balls of radius a covers x is $(1 - B(a)2^{-n})^N < e^{-N \cdot B(a)2^{-n}}$.

For $N = n2^n/B(a)$, the exponent in the right hand side of the latter inequality is at most $-n$ and the probability that x is not covered is less than e^{-n} . This probability remains exponentially small even after multiplying by 2^n , the number of different x 's. Hence, with probability close to 1, N random balls cover all the strings of length n .

Let us proceed to the proof of the lemma. Fix a ball with center y and radius a' . All the strings in the ball that are at Hamming distance at most a from y can be covered by one ball of radius a with center y . Thus it suffices, for every a'' of the form i/n such that $a < a'' \leq a'$, to cover by $\text{poly}(n)B(a')/B(a)$ balls of radius a all the strings at distance a'' from y .

Fix a'' and let S denote the set of all strings at distance exactly a'' from y . Let f be the solution to the equation $a + f(1 - 2a) = a''$ rounded to the closest rational of the form i/n . As $a < a'' \leq a' \leq \frac{1}{2}$ this equation has the unique solution and it lies in the interval $[0; 1]$. Consider a ball B of radius a with a random center z at distance f from y . As in the first argument, it suffices to show that

$$\text{Prob}[x \in B] \geq \frac{B(a)}{\text{poly}(n)B(a')}$$

for all $x \in S$.

Fix any string z at distance f from y . We claim that the ball of radius a with center z covers $\frac{B(a)}{\text{poly}(n)}$ strings in S . W.l.o.g. assume that the string y consists of only zeros and string z consists of fn ones and $(1 - f)n$ zeros. Flip a set of fan ones and a set of $(1 - f)an$ zeros in z . The total number of flipped bits is equal to an , therefore, the resulting string is at distance a from z . The number of ones in the resulting string is $fn - fan + (1 - f)an = a''n$, therefore it belongs to S . Different choices of flipped bits result in different strings in S . The number of ways to choose flipped bits is equal to $\binom{fn}{fan} \binom{(1-f)n}{(1-f)an}$. By Stirling's formula, the second factor is $2^{(1-f)nH(a) - O(\log n)}$ (we use that $a < \frac{1}{2}$ and that $H(a)$ increases on $[0; \frac{1}{2}]$). The first factor can be estimated as $\binom{fn}{fan} \geq 2^{fnH(a) - O(\log n)}$. Therefore, the number of ways to choose flipped bits is at least

$$2^{fnH(a) + (1-f)nH(a) - O(\log n)} = 2^{nH(a) - O(\log n)} \geq \frac{B(a)}{\text{poly}(n)}.$$

By symmetry, the probability that a random ball B covers a fixed string $x \in S$ does not depend on x . We have shown that a random ball B covers $\frac{B(a)}{\text{poly}(n)}$ strings in S . Hence with probability

$$\frac{B(a)}{\text{poly}(n)|S|} \geq \frac{B(a)}{\text{poly}(n)B(a')}$$

a random ball B covers a fixed string in S . The lemma is proved. (A more accurate calculation shows that the lemma holds with $\alpha = O(n^4)$.) \square

COROLLARY 2. *For every x of length n the rate distortion function r_x of x satisfies the inequalities:*

$$r_x(1/2) = O(\log n), \quad r_x(0) = K(x) + O(\log n) \tag{7}$$

$$0 \leq r_x(a) - r_x(a') \leq n(H(a) - H(a')) + O(\log n) \tag{8}$$

for all $0 \leq a < a' \leq \frac{1}{2}$. On the other hand, let r be a function mapping the set $\{0, 1/n, 2/n, \dots, 1/2\}$ to the naturals satisfying the condition (8) without $O(\log n)$ term and such that $r_x(1/2) = 0$ and $r_x(0) = k$. Then there is a string x of length n and complexity $k \pm O(\log n)$ such that $r_x(a) = r(a) + O(\sqrt{n} \log n)$ for all $a \leq 1/2$.

For example, we can apply the second part of Corollary 2 to the function $r(a)$ shown on Fig. 1. The rate distortion graph of the string x existing by Corollary 2 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 the segment $[1/6; 1/3]$.

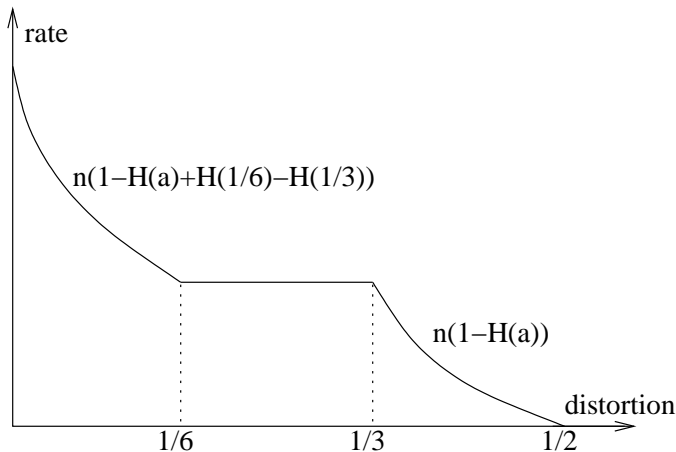


Figure 1: A possible shape of the rate distortion function.

Allowing the distortion to increase on this interval, all the way from $1/6$ to $1/3$, 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 $1/3$ to $1/6$ near the point $r = n(1 - H(1/3))$, exhibiting a very non-smooth behavior.

3.4 Example: List decoding distortion

The value s is again of order $O(\log n)$, and hence the accuracy ε is of order $O(\log n)$ in Corollary 1, and $\varepsilon = O(\sqrt{n} \log n + K(r))$ in Theorem 2. However, we can achieve better accuracy in Theorem 2, as shown in [14]:

THEOREM 3. *Let $r_x(a)$ stand for the minimal complexity of a set of cardinality at most 2^a containing x of length n . Then*

$$\begin{aligned} r_x(0) &= K(x) + O(1), & r_x(n) &= O(\log n), \\ 0 \leq r_x(a) - r_x(a') &\leq a' - a + O(\log n) \end{aligned}$$

for all $0 \leq a \leq a' \leq n$. On the other hand, let $r : \{0, 1, 2, \dots, n\} \rightarrow \mathbb{N}$ be a non-increasing function such that $r(n) = 0$, $r(0) = k$ and the function $r(a) + a$ is monotonic non-decreasing. Then there is a string x of length n and complexity $k + O(\log n)$ such that

$$r_x(a) = r(a) + O(\log n + K(r)).$$

List decoding distortion is the distortion proposed by A.N. Kolmogorov, [7], in the context of model selection and a non-probabilistic approach to statistics. It is representative of a family of distortions of which the rate-distortion graphs coincide up to logarithmic terms: *Shannon-Fano distortion* where \mathcal{Y} is the set of computable probability mass functions p with distortion $\log 1/p(x)$, and minimizing distortion means finding the distribution that maximizes probability (maximum likelihood estimation), [14]; *total recursive function distortion* where \mathcal{Y} is the set of total recursive functions f with distortion d for $f(d) = x$, [14]; and *communication complexity distortion*, [3].

3.5 Example: Euclidean distortion

Let $\mathcal{X} = \mathcal{Y}$ be the set of rational numbers in the segment $[0, 1]$ having n binary digits. Let $d(x, y)$ be equal to 0 if $x = y$ and to $n + 1 + \lceil \log |x - y| \rceil$ otherwise. Given any y with

$d(x, y) \leq a$ we can find about $n - a$ first bits of the binary expansion of x and vice versa. Thus in this example $r_x(a)$ differs by at most $O(1)$ from the Kolmogorov complexity of the prefix of length $n - a$ of the binary expansion of x .

The value s is again of order $O(\log n)$ and hence we have the accuracy $\varepsilon = O(\log n)$ in Corollary 1 and $\varepsilon = O(\sqrt{n} \log n + K(r))$ in Theorem 2. In [10], improving an earlier treatment in a draft of this paper, Salnikov strengthened this result and extended it to the case of all reals in $[0; 1]$:

THEOREM 4. *Let $r_x(a)$ stand for the minimal complexity of a rational number y at distance at most a from a real $x \in [0; 1]$. For all x we have*

$$\begin{aligned} r_x(1/2) &= O(1), \\ 0 \leq r_x(a) - r_x(a') &\leq \log a' - \log a + O(\log \log(a'/a)) \end{aligned}$$

for all $0 < a \leq a' \leq \frac{1}{2}$. On the other hand, let $r : \mathbb{Q} \rightarrow \mathbb{N}$ be a given non-increasing function such that $r(1/2) = 0$ and the function $r(a) + \log a$ is monotonic non-decreasing. Then there is a real $x \in [0; 1]$ such that

$$r_x(a) = r(a) + O(\sqrt{\log 1/a})$$

for all $0 < a \leq 1/2$. The constant in $O(\sqrt{\log 1/a})$ does not depend on r .

4 Minimizing rate and randomness deficiency

Assume that a destination word y witnesses a point $\langle r, d_x(r) \rangle$ on the distortion-rate graph of x ($d(x, y) = d_x(r)$ and $K(y) \leq r$). We will demonstrate that y can be considered as a “best-fit” description of x of the quality that is as good, as is possible for any destination word z with $K(z) \leq r$. We will measure the quality of destination words, as fitting descriptions of x , by the randomness deficiency of x in the ball $B_y(d(x, y))$. 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 $A \subseteq \mathcal{X}$ in the conditional of $K(x|A)$ is given as the list of elements of A (in the fixed order of \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 very 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 such x 's satisfying the inequality cannot be larger. Thus, elements of small deficiency form a majority of A , which has a simple description conditional to A (there is a program of size about $\log \beta$ enumerating all elements with deficiency at least β).

(3) Elements with small deficiency belong to *all* simply described majorities $B \subseteq A$ with, say, $|B| \geq (1 - 2^{-\beta})|A|$. Assume to the contrary, that $x \in A \setminus B$ and x has small randomness deficiency in A . Assume furthermore that $A \setminus B$ is enumerated by a program p that is simple conditional to A , say $K(p|A) \leq \gamma$. Then the randomness deficiency of x in A is large for large β and small γ : $\delta(x|A) > \beta - \gamma$, contradicting the assumption. Indeed, $K(x|A) \leq \log |A \setminus B| + K(p|A) \leq \log |A \setminus B| + \gamma < \log |A| - \beta + \gamma$. We omitted logarithmic terms.

Roughly speaking, items (2) and (3) mean, that an element of A has small deficiency in A if and only if it belongs to the intersection of all simply described large subsets of A (where a

description of a subset of A is a program that enumerates its complement given the list of A as input).

We now define “properties.” Properties can be defined intentionally and extensionally. For example, a binary string has equal number of occurrences of 0’s as of 1’s is an intentional definition of a property. The subset of binary strings that satisfies this intentional definition is an extensional definition of the property. It is convenient here to use the extensional definition to define properties of source words. Thus, a *property* is a set $S \subset \mathcal{X}$. We say that a set $A \subseteq \mathcal{X}$ *represents a property* B if $|B \cap A| \geq (1 - 2^{-\beta})|A|$ for an agreed upon large positive β . Therefore, Item (3) above states that if x has small randomness deficiency in A , then *all* properties that are represented by A , that are simple to describe (relative to A), are properties of x . In our treatment below, such A ’s are constituted by distortion balls centered on a destination word y , and we identify the property-representation faculties of the ball with its center y .

Let us give an example in the context of “denoising.” Assume that we are given a source word x of length n bits, and the hypothesis is “ x is obtained from an unknown string y of low Kolmogorov complexity, $K(y) \leq r$, by adding noise, at most a bits.” Possibly, the complexity of x is higher than that of y , but this increase is caused by noise we are not interested in. We would like to denoise x to retrieve the original string y . We may even be satisfied to obtain a string y' of low complexity from which x could be obtained by adding an appropriate amount of noise. For example, let $x = uv$ where u has length $n/2$ and is obtained from $00 \dots 0$ by flipping εn bits at random, and v is a string of length $n/2$ obtained from $00 \dots 0$ by flipping $n/2 - \varepsilon n$ bits at random. As a candidate for y consider the string $y_1 = 00 \dots 0$. Its Hamming distortion with respect to x is $1/2$. Intuitively, the hypothesis above with $y = y_1$ is very unlikely, since the noise is then concentrated mostly in the second half of x . The hypothesis appears much more plausible for $y_2 = 00 \dots 011 \dots 1$ (half zeros, half ones) and $a \approx 2\varepsilon$.

Assume that, to recover a y satisfying our hypothesis, we apply algorithmic rate distortion theory: Look for a string y' of low complexity ($K(y') \leq r$) that is as close to x , as it possible for strings of this complexity. The question we address is whether the hypothesis is as likely for this y' as it is for the original y ? The answer will be as follows: The string y_2 has less distortion with respect to x than the string y_1 , which will be shown below to imply that x has smaller randomness deficiency in the ball $B_{y_2}(2\varepsilon)$ than in the ball $B_{y_1}(\frac{1}{2})$. Hence we know for sure that x belongs to all majority subsets in the ball centered on y_2 , while x does not belong to some majority subsets in the ball centered on y_1 . We interpret this (as above) that y_2 represents only properties of x , while y_1 represents many properties not possessed by x . Thus, y_2 is a better explanation for x at rate r than is y_1 .

We claim that if y witnesses $d_x(r) = a$, then the randomness deficiency of x in the ball $B_y(a)$ is as small as is possible for balls of complexity at most r (minus a small value). To formulate this rigorously, let the *deficiency-rate* function $\beta_x(r)$ be defined by

$$\beta_x(r) = \min\{\delta(x|B_z(b)) : z \in \mathcal{Y}, b \in D, K(z) \leq r, x \in B_z(b)\}.$$

From now on, we use $\delta(x|z, b)$ as a shortcut for $\delta(x|B_z(b))$, and $\delta(x|z)$ denotes $\delta(x|z, d(x, z))$.

THEOREM 5. *For some $\varepsilon = O(s)$ for all $\mathcal{X}, \mathcal{Y}, d, x$ and $r \leq K(x)$ the following holds. If a destination word $y \in \mathcal{Y}$ witnesses $d_x(r + \varepsilon) = a$, that is, $d(x, y) = a$ and $K(y) \leq r + \varepsilon$, then*

$$\delta(x|y) \leq \beta_x(r) + \varepsilon.$$

Moreover, for every $y \in \mathcal{Y}$ we have

$$\delta(x|y) \leq \beta_x(r) + \varepsilon + \log B(d(x, y))/B(d_x(r + \varepsilon)) + K(y) - r. \quad (9)$$

Note that Equation (9) gives the bound

$$\delta(x|y) \leq \beta_x(r) + \varepsilon + \log B(d_x(r))/B(d_x(r + \varepsilon))$$

for every y witnessing $d_x(r)$ (and not $d_x(r + \varepsilon)$). This bound is useful when the term

$$\log B(d_x(r))/B(d_x(r + \varepsilon))$$

is small. Unfortunately this is not always the case, since $d_x(r)$ can be much greater than $d_x(r + \varepsilon)$. Theorem 2 shows that this is indeed possible (see Fig 1).

Note that for the destination words y witnessing the rate-distortion curve, instead of the distortion-rate curve, Theorem 5 is not necessarily true. For example let, for Hamming distortion, x be the string of n zeros. The string $y = 0101 \dots 01$ witnesses $r_x(n/2)$ close to 0. The randomness deficiency of x in the ball $B_y(n/2)$ is about $n - \varepsilon \approx n$. However, $\beta_x(r_x(0))$ is close to 0, with $r_x(0)$ also close to 0, as witnessed by the string x itself.

Proof of Theorem 5. The proof is based on relating the randomness deficiency of x in a set A with the *optimality deficiency of x in A* defined by

$$\log |A| + K(A) - K(x).$$

This is the number of extra bits incurred by the two-part code for x using A compared to the most optimal one-part code of x using $K(x)$ bits. The randomness deficiency is always less than the optimality deficiency, and the difference between them is equal to

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

We ignore in the proof additive terms of order $O(s)$. The last equality is true by the Symmetry of Information (see [8]). Since $K(A|x)$ is nonnegative, we obtain the inequality

$$\delta(x|A) \leq \log |A| + K(A) - K(x), \tag{10}$$

which explains why a ball $A \ni x$ with minimal radius, and hence cardinality (among balls of complexity at most r), can minimize randomness deficiency. To prove the theorem by this argument we need to show that this way to obtain balls of small deficiency is optimal. Formally, this translates to the inequality

$$\log B(d_x(r + \varepsilon)) + r - K(x) \leq \beta_x(r) + \varepsilon \tag{11}$$

for appropriate $\varepsilon = O(s)$. The combination of inequalities (11), (10) with $A = B_y(d(x, y))$, and the inequality $K(A) \leq K(y)$, we obtain inequality (9).

Note that the converse of (10) is not necessarily true: the randomness deficiency can be much less than the optimality deficiency. Here is an example for $\mathcal{X} = \{0, 1\}^n$. Let x be a random string ($K(x)$ is close to n). Let A be the set of all strings of length n except a string x' that is random and independent of x . Then $\delta(x|A) = n - n = 0$, but the optimality deficiency is close to $n + n - n = n$. This set A can be improved by adding the string x' . The resulting set B has much smaller complexity and almost the same cardinality. It turns out that this is rather general situation: for every ball $A \ni x$ with large complexity $K(A|x)$ there is another ball $B \ni x$ of much smaller complexity and the same radius. We need this statement to prove Equation (11). Let us state it rigorously:

LEMMA 2. *For every ball $A \ni x$ there is a ball $B \ni x$ of the same radius b , as A , with $K(B) \leq K(A) - K(A|x)$. (We ignore here additive terms of order $O(s)$.)*

Proof. Indeed, let N be the number of balls of radius b and complexity at most $K(A)$ covering x . We claim that $\log N \geq K(A|x)$ (in the proof we ignore additive terms of order $O(s)$). Indeed, given $x, K(A), b$ we can generate all the balls of radius b and complexity at most $K(A)$ covering x . Thus we can describe A by its index among generated balls and $K(A|x) \leq \log N$. Applying Theorem 6 below to the family of balls of radius b and $r = K(A)$, $k = \lfloor \log N \rfloor$, we conclude that there is a ball $B \ni x$ of radius b with $K(B) \leq K(A) - k \leq K(A) - K(A|x)$. \square

Apply Lemma 2 to a ball $A = B_z(b) \ni x$ minimizing $\delta(x|z, b)$ subject to $K(z) \leq r$. By Lemma 2 there is a ball $B \ni x$ of radius b with $K(B) \leq K(A) - K(A|x)$ and hence

$$\begin{aligned} K(B) + \log |B| - K(x) &\leq K(A) - K(A|x) + \log |B| - K(x) \\ &= K(A) - K(A|x) + \log |A| - K(x) = \delta(x|A) = \beta_x(r). \end{aligned} \tag{12}$$

(we ignore additive terms of order $O(s)$). To prove Equation (11) we need to show that the optimality deficiency has the following property:

LEMMA 3. *If $d_x(i) = a$ and $i \leq K(x)$ then $i + \log B(a)$ is less than $K(u) + \log B(b) + O(s)$ for every ball $B_u(b) \ni x$ with $K(u) \leq i$. (Here the inequality $K(u) \leq i$ is understood literally, without hidden $O(s)$ terms.)*

Assuming that the lemma is true (we prove it later), we prove Equation (11) as follows: With A and B as in (12), let u be the first code word in \mathcal{Y} with $B = B_u(b)$ (note that a ball might have many centers). Then

$$K(u) \leq K(B) \leq K(A) - K(A|x) \leq K(A) \leq K(z).$$

Recall that this inequalities hold to within an additive $O(s)$ term. We have required that $K(z) \leq r$. Therefore, for appropriate $\varepsilon = O(s)$ these inequalities imply that $K(u) \leq r + \varepsilon$. Thus we can apply Lemma 3 for $i = r + \varepsilon$, use (12), and conclude that

$$r + \log B(a) - K(x) \leq K(u) + \log B(b) - K(x) = K(B) + \log |B| - K(x) \leq \beta_x(r).$$

Proof of Lemma 3. Let u, b satisfy its conditions. Since a is the minimal possible radius of a ball containing x with center of complexity at most i , we can conclude that $b \geq a$. By definition of α there is $c < a$ with $B(a)/B(c) \leq \alpha$. (If $a = 0$, then the statement is true, since $K(u) + \log B(b) \geq K(x) \geq i$.) By Item (iii) in Theorem 1, there is a destination word v such that $d(x, v) \leq c$ and

$$K(v) \leq K(u) + \log(B(b)/B(c)) = K(u) + \log(B(b)/B(a)),$$

where we ignore $O(s)$ terms. Since $d(v, x) < a$ we know that

$$i < K(v) \leq K(u) + \log(B(b)/B(a))$$

and we are done. \square

\square

In the proof of Lemma 2 we have used Theorem 6 below. We exhibit it separately, since the theorem is interesting in its own right. Previously an analog of this theorem was known for exact descriptions: If a string x has at least 2^k descriptions of length at most r (with respect to an optimal description method), then $K(x) \leq r - k + O(\log k + \log r)$ (Ex. 4.3.8 in [8]). It was also known in the case of the list decoding distortion [14]: if a binary string belongs to at least 2^k sets A , of a fixed cardinality 2^c and complexity $K(A) \leq r$, then x belongs to a set

B of cardinality 2^c and complexity $K(B) \leq r - k + O(\log r + \log k + \log c)$. The difference is that in the first case there is no restriction on description syntax, and in the second case the descriptions are restricted to be finite sets of fixed cardinality. In the next theorem we restrict the description format further to allow an arbitrary set family (for example, Hamming balls).

THEOREM 6. *Let \mathcal{A} be a family of subsets of \mathcal{X} . If $x \in \mathcal{X}$ is covered by at least 2^k sets $A \in \mathcal{A}$ with $K(A) \leq r$, then x belongs to a set A in \mathcal{A} with $K(A) \leq r - k + \varepsilon$ where $\varepsilon = O(\log k + \log r + \log \log |\mathcal{X}| + K(\mathcal{A}) + K(\mathcal{X}))$.*

Proof. The statement of the theorem easily follows from its combinatorial version: Consider a game between two players, called Producer (P) and Consumer (C). The game consists of alternating moves by the players, each making 2^r moves, starting with P's move. A move of P consists in producing a subset of \mathcal{X} . A move of C consists in marking some sets previously produced by P (the number of marked sets can be 0). There are two versions of the game: the on-line version and the off-line one. In the on-line game, C wins if, following every one of his moves, every $x \in \mathcal{X}$ that is covered at least 2^k times by P's sets belongs to a marked set. In the off-line game C wins if this condition holds after his last move. Consumer can easily win (in both games), if he marks every set produced by P. However, we are interested in minimizing the total number of marked sets.

LEMMA 4. *In the off-line game, Consumer has a winning strategy that marks at most $2^{r-k} \log |\mathcal{X}|$ sets. In the on-line game, Consumer has a winning strategy that marks at most $r 2^{r-k} \log |\mathcal{X}|$ sets.*

Proof. Off-line case: We show that there is a selection of $2^{r-k} \log |\mathcal{X}|$ sets produced by P, that cover all $x \in \mathcal{X}$ that are covered by at least 2^k sets produced by P. Choose at random $2^{r-k} \log |\mathcal{X}|$ of P's sets. Let $x \in \mathcal{X}$ be covered by at least 2^k sets produced by P. Then, the probability that x is not covered by the chosen sets is at most

$$(1 - 2^{k-r})^{2^{r-k} \log |\mathcal{X}|} \leq e^{-\log |\mathcal{X}|} \ll 1/|\mathcal{X}|.$$

Multiplying this upper bound by $|\mathcal{X}|$ we get less than 1. Therefore, there is a selection of $2^{r-k} \log |\mathcal{X}|$ sets produced by P that covers all $x \in \mathcal{X}$ with multiplicity 2^k or more.

On-line case: Consumer simultaneously uses r strategies denoted by $j = 1, 2, \dots, r$. Strategy number j works as follows. Divide the sequence of Producer's sets into 2^{r-j} segments of 2^j sets. After receiving each segment, cover all the $x \in \mathcal{X}$ of multiplicity $\geq 2^k/r$ in that segment by marked sets. From the off-line case we know that it suffices to use $2^{r-k} r \log |\mathcal{X}|$ marked sets. Since there are 2^{r-j} segments (for fixed j), the total number of marked sets C needs to use is $2^{j-k} 2^{r-j} r \log |\mathcal{X}| = 2^{r-k} r \log |\mathcal{X}|$ (for fixed j). Summing over all j , this comes to $2^{r-k} r^2 \log |\mathcal{X}|$ marked sets.

We claim that after every move $t = 1, \dots, 2^r$ of C, each $x \in \mathcal{X}$ of multiplicity 2^k belongs to a marked set. Assume to the contrary, that there is an x that has multiplicity 2^k following step t of C, and x belongs to no set marked on step t or earlier. Let $t = 2^{j_1} + 2^{j_2} + \dots$ where $j_1 > j_2 > \dots$ be the binary expansion of t . The element x has multiplicity less than $2^k/r$ in the first segment of 2^{j_1} P's sets, less than $2^k/r$ in the next segment of 2^{j_2} sets, and so on. Thus its total multiplicity among t first sets is less than $r 2^k/r = 2^k$. The contradiction proves the claim. \square

Let us finish the proof of the theorem. Given \mathcal{A}, \mathcal{X} and k, r , enumerate the sets in \mathcal{A} of complexity at most r . Using the on-line strategy of Lemma 4, mark at most $2^{r-k} r^2 \log |\mathcal{X}|$ of the generated sets that cover all the strings of multiplicity 2^k , i.e., that are covered 2^k times by the generated sets. The complexity of each marked set is at most the logarithm

$r - k + 2 \log r + \log \log |\mathcal{X}|$ of the number of marked sets plus the amount of information needed to run the Algorithm. The latter is $O(\log r + \log k + K(\mathcal{X}) + K(\mathcal{A}))$. \square

5 Relation with Shannon's notion

Recall the initial paragraphs of Section 2.1. We generalize the approach from i.i.d. random variables to more general random variables, provided they are computable by a short program. Let X be a random variable in \mathcal{X} . Let $f(x)$ denote the probability of $X = x$, which is assumed to be rational for all $x \in \mathcal{X}$. Let $K(X)$ denote the Kolmogorov complexity of the set of pairs $\langle x, f(x) \rangle$ ordered lexicographically. We define the distortion-rate function as the “inverse” of (1) in this setting:

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

We relate $d(r)$ with the expected value of $d_x(r)$, the expectation taken over the distribution $f(x)$. Roughly speaking we prove that $d(r)$ is equal to the expected value of $d_x(r)$ for all r .

THEOREM 7. *Assume the discussion above. Then,*

$$\mathbf{E}d_x(r + \varepsilon) \leq d(r) \leq \mathbf{E}d_x(r - 1),$$

with $\varepsilon = O(K(d) + K(r) + K(X))$, the expectations taken over $f(x)$.

Proof. Left Inequality: By definition,

$$d(r) = \min_{E: \mathcal{X} \rightarrow \mathcal{Y}, |E(\mathcal{X})| \leq 2^r} \sum_{x \in \mathcal{X}} f(x) d(x, E(x)). \quad (13)$$

Given r and the graphs of d and $f(x)$, we can compute an optimal E of (13). Then, $K(E(x)) \leq r + O(K(d) + K(r) + K(X))$, and we have $d_x(r + O(K(d) + K(r) + K(X))) \leq d(x, E(x))$. Therefore, $\mathbf{E}d_x(r + O(K(d) + K(r) + K(X))) \leq d(r)$, the expectation taken over $f(x)$.

Right Inequality: By definition, $\mathbf{E}d_x(r) = \sum_{x \in \mathcal{X}} f(x) d_x(r)$. Define $E(x)$ to be a y with $K(y) < r$ witnessing $d_x(r)$. Then the range of E has at most $2^r - 1$ strings. For E defined in this way, we have

$$\sum_{x \in \mathcal{X}} f(x) d(x, E(x)) = \mathbf{E}d_x(r - 1),$$

and, since in this case $E(\mathcal{X})$ is restricted to a limited selection of subsets of cardinality $< 2^r$ of \mathcal{Y} , we have $d(r) \leq \mathbf{E}d_x(r - 1)$. \square

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] H. Buhrman, H. Klauck, N.K. Vereshchagin, and P.M.B. Vitányi. Individual communication complexity. In *Proc. 21th Symp. Theoret. Aspects of Comput. Sci.*, Lecture Notes in Computer Science, Vol. 2996, Springer-Verlag, Berlin, 2004, 19–30.
- [4] P. Elias, List decoding for noisy channels. *Wescon Convention Record*, Part 2, Institute for Radio Engineers (now IEEE), 1957, 94–104.

- [5] P. Elias, Error-correcting codes for List decoding, *IEEE Trans. Inform. Th.*, 37:1(1991), 5–12.
- [6] A.N. Kolmogorov, Three approaches to the quantitative definition of information, *Problems Inform. Transmission* 1:1 (1965) 1–7.
- [7] A.N. Kolmogorov. Complexity of Algorithms and Objective Definition of Randomness. A talk at Moscow Math. Soc. meeting 4/16/1974. An abstract available in *Uspekhi Mat. Nauk* 29:4(1974),155; English translation in [14].
- [8] M. Li and P.M.B. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer-Verlag, 1997. 2nd Edition.
- [9] J. Muramatsu, F. Kanaya, Distortion-complexity and rate-distortion function, *IEICE Trans. Fundamentals*, E77-A:8(1994), 1224–1229.
- [10] S. Salnikov. Kolmogorov complexity of initial segments of binary sequences. Manuscript, 2004.
- [11] C.E. Shannon. The mathematical theory of communication. *Bell System Tech. J.*, 27:379–423, 623–656, 1948.
- [12] C.E. Shannon. Coding theorems for a discrete source with a fidelity criterion. In *IRE National Convention Record, Part 4*, pages 142–163, 1959.
- [13] D.M. Sow, A. Eleftheriadis, Complexity distortion theory, *IEEE Trans. Inform. Th.*, 49:3(2003), 604–608.
- [14] N.K. Vereshchagin and P.M.B. Vitanyi, Kolmogorov’s Structure functions and model selection, *IEEE Trans. Inform. Theory*, 50:12(2004), 3265– 3290.
- [15] J.M. Wozencraft, List decoding. *Quarterly Progress Report*, Research Laboratory for Electronics, MIT, Vol. 58(1958), 90–95.
- [16] 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.
- [17] J. Ziv, Distortion-rate theory for individual sequences, *IEEE Trans. Inform. Th.*, 26:2(1980), 137–143.