On Lossy Compression (Extended Abstract)

Nikolai Vereshchagin* Paul Vitányi†

Abstract

Our proposal resolves the gap left by Shannon's theory of lossy compression (rate distortion theory), and, indeed, exposes the inadequacies of that theory by showing that globally meaningful objects can, and commonly do, have rate distortion characteristics that are radically different from those supplied by Shannon's approach. The theory in this paper may lead to novel schemes for lossy compression, and it establishes ultimate limits that future such schemes can be judged against.

1 Introduction

The Apple iPOD webpage [6] states: "iTunes makes it easy to quickly transfer your favorite songs and albums. Just pop a CD into your Mac or PC and click the Import button. You can import music in a variety of formats, such as MP3 or AAC, and at whatever quality level youd prefer. You can even choose the new Apple Lossless encoder. Music encoded with that option offers sound quality indistinguishable from the original CDs at about half the file size of the original." Formats MP3 or AAC are *lossy compressed* music files. With lossy compression, one compresses a lot, but one cannot recover the original file faithfully from the compressed version. The encoded version represents the compression program's reinterpretation of the original. This sort of compression cannot be used for anything that needs to be reproduced exactly, including software applications, databases and presidential inauguration speeches. For the latter, we use lossless compressors like gzip, bzip2, or PPMZ. From the quote above we understand that Apple losslessly compresses music by about 50%. But the massive popularity of iPOD and its clones, mega storage devices, comes from the fact that lossy compressed versions of the music pieces are stored at a fraction of their original sizes on CD. Downloading music from the web, using Napster, Kazaa, iTunes, constitutes a considerable segment of all network traffic, and is made possible only by MP3 lossy compression of that music. The same holds for JPEG lossy compression of (web) images and MPEG lossy compression of (web) videos. In fact, a good deal of all network traffic consists of lossy compressed objects. With lossy compression we can lower the bit rate to the desired level, at the cost of decreasing the fidelity of the encoding.

These are just a couple of examples, the top of the ice berg, of the ubiquitous use of lossy compression. The theoretical underpinning of lossy compression is commonly understood to be Shannon's rate-distortion theory. There we are interested in the *rate distortion function* that gives the relation between the cost in number of bits per encoded object (the rate) and the resulting lack of fidelity (the distortion) with respect to the source object. This function gives the expected distortion for every rate, the expectation taken over the probability characteristic of a random source. The problem is that this single function does not represent the vastly

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

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

different rate distortion characteristics of different objects, but only shows the characteristics of an average. In the worst case, this function does not correspond to the distortion of any of the individual data strings. In the best case it corresponds with many objects in a high-probability typical set of outcomes of a random source. For example, such outcomes can be music consisting of randomly generated notes, or pictures consisting of randomly generated pixels. However, commonly we have a piece of music with complex global structure, or a picture or video with a complex global structure, and the rate distortion function does not represent their characteristics. Lossy compression of such complex objects requires the preservation of the global meaning, and Shannon's rate distortion theory is inadequate for this use. As a result, the practice of lossy compression is forced to use ad-hoc considerations. What is needed is a theory of lossy compression of individual objects. This is precisely what we aim at in this paper: Our proposal resolves the gap left by Shannon's theory, and, indeed, exposes the inadequacies of that theory by showing that globally meaningful objects have rate distortion characteristics that are radically different from those supplied by Shannon's approach. In a practical sense we may "approximate" Kolmogorov complexity with real-world compressors. This procedure was successful in our application of the "normalized compression distance" for parameter-free clustering, classification, and data-mining in [11, 2, 7], which has led to a new class of applied methods. Apart from being applicable in approximate form, the theory in this paper may lead to novel schemes for lossy compression, and it establishes ultimate limits that future such schemes can be judged against.

2 Rate Distortion

Let X be a given source of messages. Suppose we want to communicate source data $x \in X$ using r bits. The best way to look at what happens for the individual outcomes, irrespective of the source X, is to use Kolmogorov complexity [8], textbook [10], or similar measures of the information in an individual object. If the Kolmogorov complexity $K(x)$ of the data is greater than r, or if x is not a finite object, then we can only transmit a lossy encoding $y \in Y$ of x with $K(y) \leq r$. Here Y is a given set of finite objects called *models*. Assume also that we are given a *distortion* function $d : X \times Y \to \mathbb{R}^+ \cup \{+\infty\}$, that measures the fidelity of the coded version against the source data. Different notions of fidelity will result in different distortion functions, for example music requires a different distortion measure than do images. Here we consider classical mathematical examples:

Hamming distortion. The data space X and the model space Y are both equal to the set $\{0,1\}^n$ of all binary strings of length n. The distortion function $d(x,y)$ is equal to the fraction of bits where y differs from x.

Kolmogorov distortion (= list decoding distortion). $X = \{0,1\}^n$, and Y is the set of all finite subsets of $\{0,1\}^n$; the distortion function $d(x,y)$ is equal to the cardinality of y if y contains x and is equal to infinity otherwise. The idea is as follows: the smaller cardinality $|y|$ is, the less auxiliary information we need to identify x given y.

Shannon-Fano distortion. This is a very similar to the previous example. Again $X = \{0,1\}^n$, but this time Y is the set of all probability distributions over $\{0,1\}^n$ that take only rational values; the distortion function $d(x,y)$ is defined as the inverse of probability of x with respect to y: $d(x,y) = 1/y(x)$. (In the case of Kolmogorov distortion we consider only uniform distributions on subsets of X.)

Euclidean distortion. X is the set of reals in the segment $[0,1]$, and Y is the set of all rational numbers in this segment; $d(x,y) = |x-y|$. Given any approximation of x with precision d we can find about $\lfloor \log 1/d \rfloor$ first bits of the binary expansion of x and vice verse. Hence $r_x(d)$ differs by at most $O(1)$ from the Kolmogorov complexity of the prefix of length

$\lfloor \log 1/d \rfloor$ of the binary expansion of x.

Given $X, Y, d(\cdot,\cdot)$ and a particular $x \in X$ the *rate-distortion function* r_x is defined as the minimum number of bits we need to transmit a code word y (so that y can be effectively reconstructed from the transmission) to obtain a distortion of at most d:

$$r_x(d) = \min\{K(y) \mid d(x,y) \leq d\}$$

This is an analog for individual data x of the famous *rate-distortion function* of Shannon expressing the least average rate at which outcomes from a random source X can be transmitted with distortion at most d (see Optional Appendix). We can also consider the "inverse" function

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

This is called the *distortion-rate function*. Using general codes and distortion measures we obtain a theory of *lossy compression*: Given a model family (code word set) and a particular distortion measure or "loss" measure, for given data we obtain the relation between the number of bits used for the model or code word and the least attending distortion or "loss" of the information in the individual data. Our goal is to investigate possible shapes of the graph of $r_x(i)$, as a function of i.

2.1 Related work

Shannon's approach [13, 14, 1] to rate distortion determines, nonconstructively, for every distortion measure, a single rate distortion function related to the (assumed) stationary and ergodic random source generating the data. This is largely irrelevant for most large data files, which in general cannot be modeled by such sources. Recall, that similarly Shannon's theory of communication [13] concerns nonconstructive efficient average coding of outcomes of a random variable, and hence does not help much for the problem of lossless compression of individual data strings that are not typical sequences of outcomes of an i.i.d. sequence of such random variables. Thus, already Ziv (having co-designed the breakthrough Lempel-Ziv lossless compression method) in [19] proposed a notion of a distortion-rate function for lossy compression of individual infinite sequences which is shown to be a lower bound on the distortion that can be achieved by a finite-state encoder operating at a fixed output rate. In another direction, in [18, 5, 15] it is shown that the least Kolmogorov complexity string within a given distortion d of the data string, divided by the latter's length, for the length growing unboundedly, equals Shannon's rate-distortion function almost surely. All these previous approaches deal with stationary ergodic sources, and give, for every distortion measure, a *single rate distortion characteristic*, and then show that this is appropriate for a high-probability typical subset of the data, or asymptotically for growing data (which have the property that the average over long enough data approximates the ensemble average over the high-probability typical subset). This characteristic always equals Shannon's rate distortion characteristic (up to tolerance due to the technicalities). Thus, they do not really contribute anything going beyond Shannon's original work, and do not say anything about the rate distortion characteristic of individual data that differ from the average. But data carrying meaningful information, the data we are interested in and want to store, display, and manipulate, by their very nature must differ from the average.

2.2 Results

We depart from the previous approaches in our aim to analyze the rate-distortion characteristics for every individual data string, for each given distortion measure. This characteristic is a property of the data string, and is independent of the source producing the data. For every

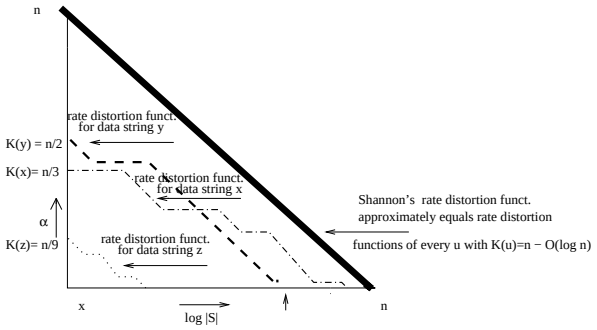

Figure 1: We give an example situation for Kolmogorov distortion of data strings of n bits (this is easier to draw than the other distortion measures). Here the code words are finite sets, and the distortion measure is the number of bits to index the elements of the finite set involved. It is essentially list coding, and the rate distortion function for a given data string gives the minimum number of bits to represent a list code word for the data string under consideration, as a function of the distortion bound (maximal allowed log-cardinality). In particular we can see whether slightly more distortion gives a gain in code length decrease. We depict Shannon's rate distortion curve (straight line) for a uniform random source; the rate distortion graphs for *existing* individual data strings x of complexity $K(x) = n/3$, data strings y of complexity $K(y) = n/2$, and data strings z of complexity $K(z) = n/9$. All $(1 - 1/n)2^n$ data strings u of complexity $n - \log n \leq K(u) \leq n + O(1)$ will have rate distortion curves that approximately coincide with Shannon's curve. There are data strings realizing all curves decreasing at a slope of at least -1 in the triangle below the diagonal, [16]. However, since all individual rate-distortion graphs of data strings of about maximal complexity (say $\geq n - \log n$) about coincide with Shannon's rate distortion curve, also the point-wise expectation of the individual rate distortion curves coincides with Shannon's up to logarithmic precision. But these data strings are precisely the ones that are *random noise*, and have no meaning other than being outcomes of random fair coin flips. Every data string that has a meaning we may be interested in, music, picture, text, has regularities expressing that meaning and hence will have low Kolmogorov complexity and a completely different rate distortion graph from Shannon's.

distortion measure: (i) For every distortion measure we give simple conditions on the shape of the function graph satisfied by the rate distortion functions of all data strings. (ii) Every function satisfying the conditions of Item (i) is realized by some data string. (iii) Functions in the family can be very different indeed. (iv) Shannon's rate distortion function for a random source with as outcomes the considered data strings, is point-wise approximated by the expectation of the rate distortion functions of the individual data strings, up to precision depending on the complexity of the random source. Result (iv) is for space reasons omitted in this abstract but can be accessed in the preliminary version http://arxiv.org/abs/cs.IT/0411014.

Technically, we give upper- and lower bounds, and shape, of the rate distortion graph of given data x, for general distortion measures in terms of "distortion balls" (Theorem 5), establishing Item (i). It is shown that for every function, satisfying the constraints of Theorem 1 on shape for a given distortion measure, there are data realizing that shape (Theorem 6), establishing Item (iii). Together, this establishes Item (iii), since the conditions in Item (i) allow for vastly different functions. We show how to apply this general theory to particular distortion measures. The particular case of Hamming distortion is worked out in detail (Theorems 1 and 2). The techniques we use are algorithmic and apply Kolmogorov complexity. We obtain a new result about covering of Hamming balls that is of independent interest, Lemma 1.

It turns out that in another context (model selection), and under another name, we analysed Kolmogorov distortion in [16]. The distortion rate function $d_x(r)$ was called the *Kolmogorov structure function* [9] and was denoted by $h_x(\alpha)$. We cite a result of [16] in Section 3.2 below. (Kolmogorov distortion is also related to list decoding, introduced by Elias [3] and Wozencraft [17], where the decoder can output a list of codewords as answer provided it contains the code word for the correct message. For a more recent survey see [4].) Following an earlier draft of the current paper, the case of Euclidean distortion was (further) analyzed in [12]. We cite the result of [12] in Section 3.3 below. In the cases of Euclidean distortion and Kolmogorov distortion the general Theorem 6 is weaker than known results [16, 12] (the price of generalization).

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 data x within a negligible tolerance for that distortion measure. Before formulating this claim precisely in the general case we consider the particular case of Hamming distortion. For the Hamming distortion we obtain new results: Theorems 1 and 2 below.

3.1 Hamming distortion

Let $X = Y$ be the set $\{0,1\}^n$ of all binary strings of length n. Let the distortion function $d(x,y)$ be equal to the fraction of bits where y differs from x: the Hamming distortion.

3.1.1 Bounds

A *ball of radius d in X* is any set of the form $\{x \in X \mid d(x,y) \le d\}$. The string y is called the *center* of the ball. Let $B(n,d)$ stand for the cardinality of a ball of radius d (it does not depend on the center). For this Hamming distortion measure, the rate distortion graphs of all data strings, satisfy the following conditions:

THEOREM 1. *For every x of length n the function $r_x(d)$ is monotonic non-increasing and satisfies the inequalities:*

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

$$r_x(d) + \log B(n,d) \le r_x(d') + \log B(n,d') + O(\log n) \tag{2}$$

for all $0 \le d \le d' \le \frac{1}{2}$.

The term $\log B(n,d)$ in (2) can be replaced by $nH(d)$, where $H(d) = d\log 1/d + (1-d)\log 1/(1-d)$ is the Shannon entropy function. Indeed, $\log B(n,d)$ and $nH(d)$ differ by at most $O(\log n)$ for all $d \le 1/2$.

The inequalities (1) and (2) imply that $r_x(d) + nH(d)$ lies between $K(x)$ and n (up to the error term $O(\log n)$):

$$K(x) - O(\log n) \le r_x(d) + nH(d) \le n + O(\log n).$$

Indeed, the left inequality is obtained by letting $d = 0$ in (2), the right one by letting $d' = 1/2$ in (2). If x is a random string of length n, that is, $K(x) = n + O(\log n)$, then the right hand side and the left hand side of this inequality coincide. Hence $r_x(d) = n - nH(d) + O(\log n)$ for all $0 \le d \le \frac{1}{2}$. If $K(x)$ is much less than n then these bounds leave much freedom for $r_x(d)$ (we will prove that this freedom indeed can be used).

Proof of Theorem 1. Immediately from the definition it follows that r_x is a non-increasing function.

The first equality in (1) is obvious as the only string at distance 0 from x is x itself. The second equality in (1) is true, as every string is at distance at most $\frac{1}{2}$ from either the string $00\ldots0$ or $11\ldots1$.

To prove (2) we need a combinatorial lemma about the number of small Hamming balls sufficient to cover a larger Hamming ball. To the best of the authors' knowledge, this covering result is new.

LEMMA 1. *For all $d \le d' \le 1/2$ every Hamming ball of radius d' can be covered by at most $\alpha B(n,d')/B(n,d)$, where $\alpha = O(n^4)$, Hamming balls of radius d.*

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

$$N = c \cdot n^4 \cdot 2^n / B(n,d)$$

balls of radius d. 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 random ball of radius d is equal to $1 - B(n,d)2^{-n}$. Thus the probability that no ball in a random family of N balls of radius d covers x is $(1 - B(n,d)2^{-n})^N < e^{-N \cdot B(n,d) 2^{-n}}$.

For $c \ge 1$, the exponent in the right hand side of the latter inequality is at most $-n^4$ and the probability that x is not covered is less than e^{-n^4}. 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. As an aside, these arguments show that there is a family of $n2^n \ln 2 / B(n,d)$ balls of radius d covering all the strings of length n.

Let us proceed to the proof of the lemma. Fix a ball with center y and radius d'. All the strings in the ball that are at Hamming distance at most d from y can be covered by one ball of radius d with center y. Thus it suffices to cover by $O(n^3 B(n,d')/B(n,d))$ balls of radius d all the strings at distance d'' from y for every d'' of the form i/n such that $d < d'' \le d'$.

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

$$\mathrm{Prob}[x \in B] \ge \Omega\Big(\frac{B(n,d)}{n^2 B(n,d')}\Big)$$

for all $x \in S$.

Fix any string z at distance f from y. We claim that the ball of radius d with center z covers $\Omega\Big(\frac{B(n,d)}{n^2}\Big)$ strings in S. W.l.o.g. assume that the string y consists of only zeros and z of fn ones and $(1-f)n$ zeros. Flip a set of $\lfloor fdn \rfloor$ ones and a set of $\lceil (1-f)dn \rceil$ zeros in z. The total number of flipped bits is equal to dn, therefore, the resulting string is at distance d from z. The number of ones in the resulting string is $fn - \lfloor fdn \rfloor + \lceil (1-f)dn \rceil = d''n$, [1] therefore it belongs to S. Different choices of flipped bits result in different strings in S. The number

[1] Formally, we need f to satisfy the equation $fn - \lfloor fdn \rfloor + \lceil (1-f)dn \rceil = d''n$, and not the equation $d + f(1-2d) = d''$. The existence of a solution of the form i/n in the segment $[0,1]$ can be proved as follows: for $f = 0$ its left hand side is equal to dn, which is less than the right hand side. For $f = 1$ the left hand side is equal to $n - dn \ge n/2$, which is greater than or equal to the right hand side. As f is increased by $1/n$, the left hand side is increased by at most 1. Hence increasing f from 0 to 1 by step $1/n$ we will find an appropriate solution.

of ways to choose flipped bits is equal to $\binom{fn}{\lfloor fdn\rfloor}\binom{(1-f)n}{\lceil(1-f)dn\rceil}$. By Stirling formula the second binomial coefficient is $\Omega(2^{(1-f)nH(d)-\log n/2})$ (we use that $d<\frac{1}{2}$ and that $H(d)$ increases on $[0;\frac{1}{2}]$). The first binomial coefficient can be estimated as

$$\binom{fn}{\lfloor fdn\rfloor} \geq \binom{fn}{\lceil fdn\rceil}/n = \Omega(2^{fnH(d)-3\log n/2}).$$

Therefore, the number of ways to choose flipped bits is at least

$$\Omega(2^{fnH(d)-3\log n/2+(1-f)nH(d)-\log n/2}) = \Omega(2^{nH(d)-2\log n}) = \Omega\Big(\frac{B(n,d)}{n^2}\Big).$$

By symmetry reasons 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 $\Omega\Big(\frac{B(n,d)}{n^2}\Big)$ strings in S. Hence with probability

$$\Omega\Big(\frac{B(n,d)}{n^2|S|}\Big) = \Omega\Big(\frac{B(n,d)}{n^2B(n,d')}\Big)$$

a random ball B covers a fixed string in S. The lemma is proved. □

Let us continue the proof of the theorem. By definition of the function r_x there is a ball of radius d' and complexity $r_x(d')$ containing x. Given d,n,d' and the center y' of this ball we can find a cover of it by at most $N=\alpha B(n,d')/B(n,d)$ balls of radius d. Consider the first generated ball among the covering balls that contains x. That ball can be found given d,n,d',y' and its index among the covering balls. Hence its complexity is at most $K(y')+\log N+O(\log\log N+K(n,d,d'))$. Thus we have $r_x(d)\leq r_x(d')+\log N+O(\log\log N+K(n,d,d'))$. The theorem is proved. □

3.1.2 Every function is realized by some data

Assume now that we are given a non-increasing function $r(d):\mathbb{Q}\to\mathbb{N}$ satisfying the constraints in Theorem 1. Is there a string x of length n whose distortion function $r_x(d)$ is close to $r(d)$? The next theorem answers this question in positive: every graph satisfying the conditions is realized by certain data strings (up to the stated precision):

THEOREM 2. *Let $r:\{0,1/n,2/n,\dots,1/2]\to\mathbb{N}$ be a non-increasing function such that the function $r(d)+\log B(n,d)$ is monotonic non-decreasing and $r(1/2)=0$. Then there is a string x of length n such that*

$$|r(d)-r_x(d)|\leq\varepsilon=O(\sqrt{n}\log n+K(r)) \tag{3}$$

for all d. Here $K(r)$ stands for the Kolmogorov complexity of the graph of r.

Proof. Let $d_0=1/2>d_1>\dots>d_N=0$, where $N=O(\sqrt{n})$, be points of the form i/n that divide the segment $[0;\frac{1}{2}]$ into subsegments each of length at most $1/\sqrt{n}$. Let us prove first that it suffices to find a string x such that (3) holds only for $d=d_0,\dots,d_N$. Indeed, to show that the inequality holds also for remaining d let $[d_{i+1};d_i]$ be the subsegment containing d. As both functions $r(d),r_x(d)$ are non-increasing, we have

$$\begin{aligned} r(d)&\in[r(d_i),r(d_{i+1})]\\ r_x(d)&\in[r_x(d_i),r_x(d_{i+1})]\subset[r(d_i)-\varepsilon,r(d_{i+1})+\varepsilon]\end{aligned}$$

Let us upper bound the length of the segment $[r(d_i),r(d_{i+1})]$. As the function $r(d)+\log B(n,d)$ is non-decreasing we have

$$r(d_{i+1})-r(d_i)\leq\log B(n,d_i)-\log B(n,d_{i+1})\leq(nd_i-nd_{i+1})\log n\leq\sqrt{n}\log n.$$

Hence $|r(d) - r_x(d)| \le \varepsilon + \sqrt{n}\log n$ and we are done.

To find x we run the following non-halting algorithm that takes as input n and the graph of r.

Algorithm: Enumerate all the balls in X of radiuses d_i and complexities less than $r(d_i)-\varepsilon$, respectively. Call such balls *forbidden*, as the object x cannot belong to any such ball. Let G denote X minus the union of all forbidden balls discovered so far.

Construct, in parallel, balls $B_0, \dots, B_N$ of radiuses $d_0, \dots, d_N$, respectively, as described further. Call them *candidate balls*. These are balls ensuring the inequality $r_x(d_i) \le r(d_i) + \varepsilon$. Every candidate ball is changed time to time so that the following invariant is true: for all $i \le N$ the measure of the intersection $B_0 \cap \dots \cap B_i \cap G$ is at least

$$B(n, d_i)2^{-i-1}\alpha^{-i},$$

where $\alpha = O(n^4)$ is the constant from Lemma 1.

First perform the initialization step to find the initial candidate balls. Let say B_i be the ball of radius d_i with the center in $00\dots0$.

Enumerating forbidden balls we update G. Once the invariant 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 $B(n, 1/2)$. Indeed, for fixed i the total cardinality of all the balls of radius d_i and complexity less than $r(d_i)-\varepsilon$ does not exceed $2^{r(d_i)-\varepsilon}B(n, d_i)$. As the function $r(d)+\log B(n, d)$ is monotonic non-decreasing, the total number of elements on all forbidden balls is at most

$$\sum_{i=0}^{N} 2^{r(d_i)-\varepsilon}B(n, d_i) \le (N+1)2^{r(1/2)-\varepsilon}B(n, 1/2) = (N+1)2^{-\varepsilon}B(n, 1/2) \ll B(n, 1/2).$$

Assume that the invariant has become false for some $i > 0$. Let i be the least such index. As the invariant is true for $i-1$, the measure of the intersection G' of all the balls $B_1, \dots, B_{i-1}$ and G is at least $B(n, d_{i-1})2^{-i}\alpha^{-i+1}$. 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 $\alpha B(n, d_{i-1})/B(n, d_i)$ balls of radius d_i (it exists by Lemma 1). The cardinality of $G' \cap B$ for at least one covering ball is at least

$$|G'|/(\alpha B(n, d_{i-1})/B(n, d_i)) \ge B(n, d_i)2^{-i}\alpha^{-i}.$$

Let B_i be equal to any such ball. Note that $B(n, d_i)2^{-i}\alpha^{-i}$ 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 $B_{i+1}, \dots, B_N$.

The algorithm is described. Although it 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, as its cardinality is positive by the invariant. By construction x avoids all the forbidden balls, thus $r_x(d)$ 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 $r(d_i) + \varepsilon$. Fix $i \le N$. Consider the description of B_i consisting of n, i, the graph of r, and the total number M of changes of B_i. The 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 $r(d_i)$. Let us prove that the candidate ball B_i is changed at most $2^{r(d_i)+i}$ times. Distinguish two possible cases when B_i is changed: (1) the invariant has become false for an index strictly less than i, (2) the invariant has become false for i and remains true for all smaller indexes. Arguing by induction, the number of changes of

the first kind can be upper bounded by $2^{r(d_{i-1})+i-1} \le 2^{r(d_i)+i-1}$. 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 d_i has been enumerated, (2b) after the last change of B_i no forbidden ball of radius greater than d_i have been enumerated. The number of changes of type (2a) is at most the number of forbidden balls of radiuses $d_j \ge d_i$. By monotonicity of $r(d)$, this is at most $(N+1)2^{r(d_i)-\varepsilon} \ll 2^{r(d_i)}$. 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 $d_j \ge d_i$ have been enumerated. Thus the cardinality of G has decreased by at least $B(n,d_i)2^{-i-1}\alpha^{-i}$ due to enumerating forbidden balls with radiuses $d_j < d_i$ (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 $N2^{r(d_i)-\varepsilon}B(n,d_i)$ (we use the monotonicity of $r(d)+\log B(n,d)$). The number of changes of types (2b) is less than the ratio of this number to the threshold $B(n,d_i)2^{-i-1}\alpha^{-i}$. Hence it is less than $N2^{r(d_i)-\varepsilon}2^{i+1}\alpha^i \ll 2^{r(d_i)}$. The theorem is proved. □

3.2 Kolmogorov distortion

The case of Kolmogorov distortion was analyzed in [16]. In this case we can achieve better accuracy, as the following theorem shows. Let $r_x(d)$ stand for the minimal complexity of a set of cardinality d containing x.

THEOREM 3. *For all strings x of length n we have*

$$r_x(1) = K(x) + O(1), \qquad r_x(2^n) = O(\log n),$$
$$r_x(d) + \log d \le r_x(d') + \log d' + O(\log n)$$

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

$$|r(d) - r_x(d)| = O(\log n + K(r))$$

for all $d \ge 1$.

The theorem implies that $r_x(d) + \log d$ is between $K(x)$ and n (up to a logarithmic error term):

$$K(x) - O(\log n) \le r_x(d) + \log d \le n + O(\log n).$$

If x is random, that is, $K(x) = n + O(\log n)$, then we obtain $r_x(d) = n - \log d + O(\log n)$ for all $1 \le d \le 2^n$.

3.3 Euclidean distortion

The case of Euclidean distortion was investigated in [12]. Therefore, we can be short. Let $r_x(d)$ stand for the minimal complexity of a rational number y at distance at most d from a real $x \in [0;1]$.

THEOREM 4. *For all x we have*

$$r_x(1/2) = O(1),$$
$$r_x(d) + \log d \le r_x(d') + \log d' + O(\log\log(d'/d))$$

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

$$|r(d) - r_x(d)| = O(\sqrt{\log 1/d})$$

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

4 General theory

To generalize Theorems 1 and 2 to arbitrary spaces and distortion measures we need some assumptions on X, Y, d of algorithmic and combinatorial nature. Below we list all of them. Basically, we just list the properties of X, Y, d used in the proofs of Theorems 1 and 2. We will assume that a measure μ on X is given. It will be used in the assumptions and in the theorems.

(a) Call any subset of X that is a finite Boolean combination of balls of rational radiuses a *simple set* (balls are defined just as in the Hamming distortion case). Every simple set can be finitely described (a ball is described by its center and radius). We require that the measure of every simple subset of X is a rational number that can be computed given the subset by a certain algorithm A_1 (its complexity will be used in the theorems).

(b) There exists an algorithm A_2 that given a simple set decides whether the set is empty.

(c) Let $B(d)$ stand for the maximal measure of a ball in X of radius d (we assume that the maximum is attained). We assume that $B(d)$ is rational for all rational d and can be computed given d by a certain algorithm A_3. We assume also that the set X is itself a ball of a rational radius $d_{\max}$ with a center y_0.

(d) We assume that there is an integer α that satisfies the following analog of Lemma 1:

> For all $d \le d'$ such that $B(d) > 0$, every ball of radius d' in X can be covered by at most $\alpha \frac{B(d')}{B(d)}$ balls of radius d.

It follows from previous items that such a covering can be found given the initial ball, α and algorithms A_1, A_2, A_3. Obviously $\alpha \ge 1$. The smaller α is, the more precisely we can describe the possible shapes of $r_x(d)$ in terms of the measure μ.

Here is the analog of Theorem 1 for general distortions:

THEOREM 5. *For all* $x \in X$ *the function* r_x *is monotonic non-increasing and satisfies the following inequalities*

$$r_x(d_{max}) \le K(y_0),$$
$$r_x(d) + \log B(d) \le r_x(d') + \log B(d') + O(\log \alpha + \log\log(B(d')/B(d)) + K(A_1, A_2, A_3))$$

for all rational $d \le d'$.

The proof is essentially the same as for Theorem 1 and therefore we omit it.

Theorem 1 and the first parts of Theorems 3 and 4 are special cases of Theorem 5. To deduce them we let μ to be the uniform measure. In the Hamming distortion case we can set $\alpha = O(n^4)$ and both terms $\log\log(B(d')/B(d))$, $K(A_1, A_2, A_3)$ are of order $O(\log n)$. In the Kolmogorov distortion case we can set $\alpha = 2$ and again both terms $\log\log(B(d')/B(d))$, $K(A_1, A_2, A_3)$ are of order $O(\log n)$. For Euclidean distortion we can also set $\alpha = 2$, and both $\log\log(B(d')/B(d)) = \log\log(d'/d)$ and the quantity $K(A_1, A_2, A_3)$ is constant.

And here is the analog of Theorem 2 for general distortions:

THEOREM 6. *Let a finite set $D \subset \mathbb{Q}$ whose maximal element is equal to d_{max} and a non-increasing function $r : D \to \mathbb{N}$ be given. Assume that the function $r(d)+\log B(d)$ is monotonic non-decreasing and $r(d_{max}) = 0$. Then there is a string x such that*

$$|r(d) - r_x(d)| \le \varepsilon \tag{4}$$

for all $d \in D$. Here $\varepsilon = O(\sqrt{\log(B(d_{max})/B(d_{min}))}\log(2\alpha) + K(A_1, A_2, A_3, r, d_{max}, y_0))$ where d_{min} stands for the minimal element in D

Proof. Let $\delta = \sqrt{\log(B(d_{\max})/B(d_{\min}))}$. We claim that there are points $d_0 = d_{\max} > d_1 > \cdots > d_N = d_{\min}$ in D, where $N = O(\delta)$, such that every $d \in D \setminus \{d_0, \ldots, d_N\}$ belongs to a segment $[d_{i+1}; d_i]$ with $\log B(d_i) - \log B(d_{i+1}) \le \delta$. Indeed, chop the segment $[\log B(d_{\min}); \log B(d_{\max})]$ into subsegments of length δ and for each subsegment $[a; b]$ include the leftmost and the rightmost points in $D \cap B^{-1}([a; b])$ in the sequence of d_i (if this set is empty, do not include anything). As in the proof of Theorem 2 it is enough to find x such that (4) holds for all d_i.

To find such an x we run an algorithm analogous to that in the proof of Theorem 2. To run it we need to know $r, d_{\max}, y_0, \alpha$ and we need algorithms A_1, A_2, A_3. Again we keep the same invariant as in the proof of Theorem 2. To perform the initialization step we let B_0 to be equal to the ball of the radius $d_{\max}$ centered at y_0, and then find $B_1, \ldots, B_N$ inductively using the property (d) and algorithms A_1, A_2, A_3.

Then we enumerate forbidden ball. Again all the balls with radiuses d_i and centers of complexity at most $r(d_i) - \varepsilon$ are forbidden. By assumption on r the total measure of all forbidden balls is at most $(N + 1)2^{-\varepsilon}B(d_{\max}) \ll B(d_{\max})$. This implies that the invariant is true for $i = 0$ at any step of the algorithm. Once the invariant becomes false we restore it just by the same procedure using algorithms A_1, A_2, A_3. The upper bound for the number of changes of B_i is proved just as in Theorem 2. All we need are the inequalities $(N + 1)2^{r(d_i)-\varepsilon} \ll 2^{r(d_i)}$ and $N2^{r(d_i)-\varepsilon}2^{i+1}\alpha^i \ll 2^{r(d_i)}$, which are true by the choice of ε. □

Theorem 2 is a direct corollary of Theorem 6 for $D = \{0, 1/n, 2/n, \ldots, 1/2\}$. Indeed, in this case the term $\log(B(d_{\max})/B(d_{\min}))$ is of order $O(n)$ and $K(A_1, A_2, A_3, d_{\max}, y_0)$ is of order $\log n$.

In the case of Kolmogorov distortion we can apply Theorem 6 for $D = \{1, 2, 3, \ldots, 2^n\}$. Then $\log(B(d_{\max})/B(d_{\min}))$ is again of order $O(n)$ thus the accuracy of the general Theorem 6 is $O(\sqrt{n}+K(r))$, which is worse than the accuracy $O(\log n+K(r))$ we achieved for this specific distortion by special-purpose arguments in the second part of Theorem 3.

The second part Theorem 4 is not derivable from Theorem 6 with any non-trivial accuracy. Indeed, to deduce Theorem 4 we need to let D be infinite. We can only deduce the following weaker statement: Let $r : \mathbb{Q} \to \mathbb{N}$ be a given computable non-increasing function such that $r(\frac{1}{2}) = 0$ and the function $r(d) + \log d$ is monotonic non-decreasing. Then for every N there is a real $x \in [0; 1]$ such that $|r(d) - r_x(d)| = O(\sqrt{N} + K(r))$ for all $2^{-N} \le d \le \frac{1}{2}$.

References

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

[2] R. Cilibrasi, P.M.B. Vitanyi, Clustering by compression, *IEEE Trans. Information Theory*, 51:4(2005)

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

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

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

[6] iPOD + iTUNES webpage at http://www.apple.com/ipod/

[7] E. Keogh, S. Lonardi, and C.A. Rtanamahatana, Toward parameter-free data mining, In: *Proc. 10th ACM SIGKDD Intn'l Conf. Knowledge Discovery and Data Mining*, Seattle, Washington, USA, August 22—25, 2004, 206–215.

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

[9] 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 [16].

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

[11] M. Li, X. Chen, X. Li, B. Ma, P.M.B. Vitanyi, The similarity metric, *IEEE Trans. Inform. Th.*, 50:12(2004), 3250- 3264.

[12] S. Salnikov. Kolmogorov complexity of initial segments of binary sequences. Manuscript, 2004.

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

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

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

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

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

[18] 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.

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