Kolmogorov Complexity with Error

Lance Fortnow[1], Troy Lee[2], and Nikolai Vereshchagin[3,⋆]

[1] University of Chicago, 1100 E. 58th Street, Chicago, IL, USA 60637.
fortnow@cs.uchicago.edu,
http://people.cs.uchicago.edu/~fortnow

[2] CWI and University of Amsterdam, 413 Kruislaan, 1098 SJ Amsterdam, The Netherlands.
tlee@cwi.nl,
http://www.cwi.nl/~tlee

[3] Moscow State University, Leninskie Gory, Moscow, Russia 119992.
ver@mccme.ru,
http://lpcs.math.msu.su/~ver

Abstract. We introduce the study of Kolmogorov complexity with error. For a metric d, we define $C_a(x)$ to be the length of a shortest program p which prints a string y such that $d(x, y) \leq a$. We also study a conditional version of this measure $C_{a,b}(x|y)$ where the task is, given a string y' such that $d(y, y') \leq b$, print a string x' such that $d(x, x') \leq a$. This definition admits both a uniform measure, where the *same* program should work given any y' such that $d(y, y') \leq b$, and a nonuniform measure, where we take the length of a program for the worst case y'. We study the relation of these measures in the case where d is Hamming distance, and show an example where the uniform measure is exponentially larger than the nonuniform one. We also show an example where symmetry of information does not hold for complexity with error under either notion of conditional complexity.

1 Introduction

Kolmogorov complexity measures the information content of a string typically by looking at the size of a smallest program generating that string. Suppose we received that string over a noisy or corrupted channel. Such a channel could change random bits of a string, possibly increasing its Kolmogorov complexity without adding any real information.

Alternatively, suppose that we do not have much memory and are willing to sacrifice fidelity to the original data in order to save on compressed size. What is the cheapest approximation to a string within our level of tolerance to distortion? Such compression where some, less important we hope, information about the original data is lost is known as lossy compression.

Intuitively, these scenarios are in some sense complementary to one another: we expect that if we lossy compress a string received over a corrupted channel

⋆ Supported in part by the RFBR grants 03-01-00475, 358.20003.1. Work done while visiting CWI.

with our level of tolerance equal to the number of expected errors, then the cheapest string within the level of tolerance will be the one with the high complexity noise removed. Ideally we would get back our original string. For certain compression schemes and models of noise this intuition can be made precise [8].

In this paper we explore a variation of Kolmogorov complexity designed to help us measure information in these settings. We define the Kolmogorov complexity of a string x with error a as the length of a smallest program generating a string x' that differs from x in at most a bits. We give tight bounds (up to logarithmic factors) on the maximum complexity of such strings and also look at time-bounded variations.

We also look at conditional Kolmogorov complexity with errors. Traditional conditional Kolmogorov complexity looks at the smallest program that converts a string y to a string x. In our context both x and y could be corrupted. We want the smallest program that converts a string close to y to a string close to x. We consider two variations of this definition, a uniform version where we have a single program that that converts any y' close to y to a string x' close to x and a nonuniform version where the program can depend on y'. We show examples giving a large separation between the uniform and nonuniform definitions.

Finally we consider symmetry of information for Kolmogorov complexity with error. Traditionally the complexity of the concatenation of strings x, y is roughly equal to the sum of the complexity of x and the complexity of y given x. We show that for any values of d and a the complexity of xy with error d is at most the sum of the complexity of x with error a and the complexity of converting a string y with $d-a$ error given x with a bits of error. We show the other direction fails in a strong sense—we do not get equality for any a.

2 Preliminaries

We use $|x|$ to denote the length of a string x, and $\|A\|$ to denote the cardinality of a set A. All logarithms are base 2.

We use $d_H(x, y)$ to denote the Hamming distance between two binary strings x, y, that is the number of bits on which they differ. For $x \in \{0,1\}^n$ we let $B_n(x, R)$ denote the set of n-bit strings within Hamming distance R from x, and $V(n, R) = \sum_{i=0}^{R} \binom{n}{i}$ denote the volume of a Hamming ball of radius R over n-bit strings. For $0 < \lambda \leq 1/2$ the binary entropy of λ is $H(\lambda) = -\lambda \log \lambda - (1-\lambda)\log(1-\lambda)$. The binary entropy is useful in the following approximation of $V(n, R)$ which we will use on several occasions (a proof can be found in [1]).

Lemma 1. *Suppose that $0 < \lambda \leq 1/2$ and λn is an integer. Then*

$$\frac{2^{nH(\lambda)}}{\sqrt{8n\lambda(1-\lambda)}} \leq V(n, \lambda n) \leq 2^{nH(\lambda)}.$$

3 Defining Kolmogorov Complexity with Error

We consider several possible ways of defining Kolmogorov complexity with error. In this section we present these alternatives in order to evaluate their relative merits in the coming sections. First, we review the standard definition of Kolmogorov complexity. More details can be found in [6].

For a Turing machine T, the Kolmogorov complexity $C_T(x|y)$ of x given y is the length of a shortest program p such that $T(p, y) = x$. The theory of Kolmogorov complexity begins from the following invariance theorem: there is a universal machine U such that for any other Turing machine T, there exists a constant c_T such that $C_U(x|y) \leq C_T(x|y) + c_T$, for all x, y. We now fix such a U and drop the subscript. Now we define also the unconditional Kolmogorov complexity $C(x) = C(x|\text{empty string})$.

Definition 1. *Let $d : (\{0,1\}^n)^2 \to R$ be a metric, and $a \in R$. The complexity of x with error a, denoted $C_a(x)$ is $C_a(x) = \min_{x'}\{C(x') : d(x', x) \leq a\}$.*

We will also consider a time bounded version of this definition, $C_a^t(x) = \min_{x'}\{C^t(x'|\text{empty string}) : d(x, x') \leq a\}$, where $C^t(x|y)$ is the length of a shortest program p such that $U(p, y)$ prints x in less than $t(|x| + |y|)$ time steps. Here we assume that the machine U is universal in the following sense: for any other Turing machine T, there exists a constant c_T and a polynomial q such that $C_U^{q(|x|,|y|,t)}(x|y) \leq C_T^t(x|y) + c_T$, for all x, y, t.

A relative version of Kolmogorov complexity with error is defined by Impagliazzo, Shaltiel and Wigderson [4]. That is, they use the definition $C_\delta(x) = \min\{C(y) : d_H(x, y) \leq \delta|x|\}$. We prefer using absolute distance here as it behaves better with respect to concatenations of strings—using relative distance has the disadvantage of severe nonmonotonicity over prefixes. Take, for example, $x \in \{0,1\}^n$ satisfying $C(x) \geq n$. Let $y = 0^{2n}$. Then $C_{1/3}(x) \geq n - \log V(n, n/3)$ while $C_{1/3}(xy) \leq \log n + O(1)$. Using absolute error we have that $C_a(xy) \geq C_a(x) - O(\log n)$, that is it only suffers from logarithmic dips as with standard definition.

Defining conditional complexity with error is somewhat more subtle. We introduce both uniform and nonuniform versions of conditional complexity with error.

Definition 2. *For a Turing machine T, the uniform conditional complexity, denoted $(C_{a,b}^u)_T(x|y)$, is the length of a shortest program p such that,* for any y' *satisfying $d(y, y') \leq b$ it holds that $T(p, y')$ outputs a string whose distance from x is less than a.*

The invariance theorem remains true: there is a universal machine U such that for any other Turing machine T, there exists a constant c_T such that $(C_{a,b}^u)_U(x|y) \leq (C_{a,b}^u)_T(x|y) + c_T$, for all x, y, a, b. We fix such a U and drop the subscript.

Definition 3. *Nonuniform conditional complexity, which we denote $C_{a,b}(x|y)$ is defined as $C_{a,b}(x|y) = \max_{y'} \min_{x'}\{C(x'|y') : d(x', x) \leq a \text{ and } d(y', y) \leq b\}$.*

In section 6 we study the difference between these two measures.

4 Strings of Maximal Complexity

One of the most famous applications of Kolmogorov complexity is the incompressibility method (see [6], Chapter 6). To prove there exists an object with a certain property, we consider an object with maximal Kolmogorov complexity and show that it could be compressed if it did not possess this property.

This method relies on a simple fact about strings of maximal complexity: for every length n, there is a string x of complexity at least n. This follows from simple counting. It is also easy to see that, up to an additive constant, every string has complexity at most its length. What is the behavior of maximal complexity strings in the error case? In this paper we restrict ourselves to the Hamming distance case.

Again by a counting argument, we see that for every n there is an x of length n with $C_a(x) \geq \log 2^n/V(n,a) = n - \log V(n,a)$. Upper bounding the complexity of strings in the error case requires a bit more work, and has a close connection with the construction of covering codes. A covering code $\mathcal{C}$ of radius a is a set of strings such that for every $x \in \{0,1\}^n$ there is an element $y \in \mathcal{C}$ such that $d_H(x,y) \leq a$. Thus an upper bound on the maximum complexity strings will be given by the existence of covering codes of small size. The following Lemma is well known in the covering code literature, (see [1] or [5]).

Lemma 2. *For any n and integer $R \leq n$, there exists a set $\mathcal{C} \subseteq \{0,1\}^n$ with the following properties:*

1. $\|\mathcal{C}\| \leq n2^n/V(n,R)$
2. *for every $x \in \{0,1\}^n$, there exists $c \in \mathcal{C}$ with $d_H(x,c) \leq R$*
3. *The set $\mathcal{C}$ can be computed in time* $\mathrm{poly}(2^n)$

Proof: For the first two items we argue by the probabilistic method. Fix a point $x \in \{0,1\}^n$. We uniformly at random choose k elements $x_1, \ldots, x_k$ of $\{0,1\}^n$. The probability P_x that x is not contained in $\cup_{i=1}^k B(x_i, R)$ is precisely

$$P_x = (1 - V(n,R)/2^n)^k \leq e^{-kV(n,R)/2^n}.$$

For the inequality we have used the fact that $e^z \geq 1 + z$ for any z. Taking k to be $n2^n/V(n,R)$ makes this probability strictly less than 2^{-n}. Thus the probability of the union of the events P_x over $x \in \{0,1\}^n$ is, by the union bound, less than 1 and there exists a set of $n2^n/V(n,R)$ centers which cover $\{0,1\}^n$. This gives items 1 and 2.

For item 3 we now derandomize this argument using the method of conditional probabilities. The argument is standard as found in [7], and omitted here. $\square$

To achieve part 3 of Lemma 2 one could alternatively apply a general theorem that the greedy algorithm always finds a covering of a set X of size at most a $\ln \|X\|$ multiplicative factor larger than the optimal covering (see Corollary 37.5 in [2]). This would give the slightly worse bound of $O(n^2 2^n/V(n,R))$.

Theorem 1. *For every n, a and $x \in \{0,1\}^n$, $C_a(x) \leq n - \log V(n,a) + O(\log n)$.*

Proof: Use the lexicographically first covering code of radius a whose existence is given by Lemma 2. $\square$

One nice property of covering codes is that they behave very well under concatenation. Let $\mathcal{C}_1$ be a covering code of $\{0,1\}^{n_1}$ of radius R_1 and $\mathcal{C}_2$ be a covering code of $\{0,1\}^{n_2}$ of radius R_2. Now let $\mathcal{C} = \{cc' : c \in \mathcal{C}_1, c' \in \mathcal{C}_2\}$ be the set of all ordered concatenations of codewords from $\mathcal{C}_1$ with codewords from $\mathcal{C}_2$. Then $\mathcal{C}$ is a covering code over $\{0,1\}^{n_1+n_2}$ of radius $R_1 + R_2$.

We can use this idea in combination with item 3 of Lemma 2 to efficiently construct near-optimal covering codes. This construction has already been used for a complexity-theoretic application in [3].

Theorem 2. *There is a polynomial time bound $p(n)$ such that $C_a^{p(n)}(x) \leq n - \log V(n,a) + O(n \log\log n / \log n)$ for every $x \in \{0,1\}^n$ and every a.*

Proof: We construct a covering code over $\{0,1\}^n$ with radius a such that the ith element of the covering can be generated in time polynomial in n. Let $\ell = \log n$ and divide n into n/ℓ blocks of length ℓ. Let $r = (a/n)\ell$. Now by item 3 of Lemma 2 we can in time polynomial in n construct a covering code over $\{0,1\}^\ell$ of radius r and of cardinality $\ell 2^\ell / V(\ell, r)$. Call this covering $\mathcal{C}_\ell$. Our covering code $\mathcal{C}$ over $\{0,1\}^n$ will be the set of codewords $\{c_1 c_2 \cdots c_{n/\ell} : c_i \in \mathcal{C}_\ell\}$. The size of this code will be:

$$\begin{aligned} \|\mathcal{C}\| &\leq (2^{\ell - \log V(\ell,r) + \log \ell})^{n/\ell} = (2^{\ell - \ell H(a/n) + O(\log \ell)})^{n/\ell} \\ &= 2^{n - nH(a/n) + O(n \log \ell / \ell)} = 2^{n - \log V(n,a) + O(n \log \ell/\ell)}. \end{aligned} \tag{1}$$

The second and last inequalities hold by Lemma 1.

In this proof we assumed that $\log n$, $n/\log n$, and $a \log n/n$ are all integer. The general case follows with simple modifications. $\square$

5 Dependence of Complexity on the Number of Allowed Errors

Both the uniform and the non-uniform conditional complexities $C^u_{a,b}$ and $C_{a,b}$ are decreasing functions in a and increasing in b. Indeed, if b decreases and a increases then the number of y''s decreases and the number of x''s increases, thus the problem to transform every y' to some x' becomes easier. What is the maximal possible rate of this decrease/increase? For the uniform complexity, we have no non-trivial bounds. For the non-uniform complexity, we have the following

Theorem 3. *For all x, y of length n and all $a \leq a'$, $b' \leq b$ it holds*

$$C_{a,b}(x|y) \leq C_{a',b'}(x|y) + \log(V(n,a)/V(n,a')) + \log(V(n,b')/V(n,b)) + O(\log n).$$

Proof: Let y' be a string at distance b from y. We need to find a short program mapping it to a string at distance a from x. To this end we need the following lemma from [9].

Lemma 3. *For all $d \le d' \le n$ having the form i/n, every Hamming ball of radius d' in the set of binary strings of length n can be covered by at most $O(n^4 V(n,d')/V(n,d))$ Hamming balls of radius d.*

Apply the lemma to $d' = b$, $d = b'$ and to the ball of radius b centered at y'. Let $B_1, \dots, B_N$, where $N = O(n^4 V(n,b)/V(n,b'))$, be the covering balls. Let B_i be a ball containing the string y and let y'' be its center. There is a program, call it p, of length at most $C_{a',b'}(x|y)$ mapping y'' to a string at distance a' from x. Again apply the lemma to $d = a$, $d' = a'$ and to the ball of radius d' centered at x'. Let $C_1, \dots, C_M$, where $M = O(n^4 V(n,a')/V(n,a))$, be the covering balls. Let C_j be a ball containing the string x and let x'' be its center. Thus x'' is at distance a from x and can be found from y', p, i, j. This implies that $K(x''|y') \le |p| + \log N + \log M + O(\log n)$ (extra $O(\log n)$ bits are needed to separate p, i and j). □

In the above proof, it is essential that we allow the program mapping y' to a string close to x depend on y'. Indeed, the program is basically the triple (p, i, j) where both i and j depend on y'. Thus the proof is not valid for the uniform conditional complexity. And we do not know whether the statement itself is true for the uniform complexity.

By using Theorem 2 one can prove a similar inequality for time bounded complexity with the $O(\log n)$ error term replaced by $O(n \log\log n/\log n)$.

6 Uniform vs. Nonuniform Conditional Complexity

In this section we show an example where the uniform version of conditional complexity can be exponentially larger than the nonuniform one. Our example will be for $C_{0,b}(x|x)$. This example is the standard setting of error correction: given some x' such that $d_H(x, x') \le b$, we want to recover x exactly. An obvious upper bound on the nonuniform complexity $C_{0,b}(x|x)$ is $\log V(n,b) + O(1)$—as we can tailor our program for each x' we can simply say the index of x in the ball of radius b around x'.

In the uniform case the same program must work for every x' in the ball of radius b around x and the problem is not so easy. The following upper bound was pointed out to us by a referee.

Proposition 1. $C^u_{0,b}(x|x) \le \log V(n, 2b) + O(1)$.

Proof: Let $\mathcal{C} \subseteq \{0,1\}^n$ be a set with the properties:

1. For every $x, y \in \mathcal{C} : B_n(x,b) \cap B_n(y,b) = \varnothing$.
2. For every $y \in \{0,1\}^n\ \exists x \in \mathcal{C} : d_H(x,y) \le 2b$.

We can greedily construct such a set as if there is some string y with no string $x \in \mathcal{C}$ of distance less than $2b$, then $B_n(y, b)$ is disjoint from all balls of radius b around elements of $\mathcal{C}$ and so we can add y to $\mathcal{C}$.

Now for a given x, let x^* be the closest element of $\mathcal{C}$ to x, with ties broken by lexicographical order. Let $z = x \oplus x^*$. By the properties of $\mathcal{C}$ this string has Hamming weight at most $2b$ and so can be described with $\log V(n, 2b)$ bits. Given input x' with $d_H(x, x') \leq b$, our program does the following: computes the closest element of $\mathcal{C}$ to $x' \oplus z$, call it w, and then outputs $w \oplus z = w \oplus x^* \oplus x$. Thus for correctness we need to show that $w = x^*$ or in other words that $d_H(x' \oplus z, x^*) \leq b$. Notice that $d_H(\alpha \oplus \beta, \beta) = d_H(\alpha, 0)$, thus

$$d_H(x' \oplus z, x^*) = d_H(x' \oplus x \oplus x^*, x^*) = d_H(x' \oplus x, 0) = d_H(x, x') \leq b.$$

□

We now turn to the separation between the uniform and nonuniform measures. The intuition behind the proof is the following: say we have some computable family S of Hamming balls of radius b, and let x be the center of one of these balls. Given any x' such that $d_H(x, x') \leq b$, there may be other centers of the family S which are also less than distance b from x'. Say there are k of them. Then x has a nonuniform description of size about $\log k$ by giving the index of x in the k balls which are of distance less than b from x'.

In the uniform case, on the other hand, our program can no longer be tailored for a particular x', it must work for any x' such that $d_H(x, x') \leq b$. That is, intuitively, the program must be able to distinguish the ball of x from any other ball intersecting the ball of x. To create a large difference between the nonuniform and uniform conditional complexity measures, therefore, we wish to construct a large family of Hamming balls, every two of which intersect, yet that no single point is contained in the intersection of too many balls. Moreover, we can show the stronger statement that $C_{0,b}(x|x)$ is even much smaller than $C^u_{a,b}(x|x)$, for a non-negligible a. For this, we further want that the contractions of any two balls to radius a are disjoint. The next lemma shows the existence of such a family.

Lemma 4. *For every length m of strings and a, b, and N satisfying the inequalities*

$$N^2 V(m, 2a) \leq 2^{m-1}, \quad N^2 V(m, m-2b) \leq 2^{m-1}, \quad NV(m, b) \geq m2^{m+1} \qquad (2)$$

there are strings $x_1, \ldots, x_N$ such that the balls of radius a centered at $x_1, \ldots, x_N$ are pairwise disjoint, and the balls of radius b centered at $x_1, \ldots, x_N$ are pairwise intersecting but no string belongs to more than $NV(m, b)2^{1-m}$ of them.

Proof: The proof is by probabilistic arguments. Take N independent random strings $x_1, \ldots, x_N$. We will prove that with high probability they satisfy the statement.

First we estimate the probability that there are two intersecting balls of radius a. The probability that two fixed balls intersect is equal to $V(m, 2a)/2^m$.

The number of pairs of balls is less than $N^2/2$, and by union bound, there are two intersecting balls of radius a with probability at most $N^2V(m,2a)/2^{m+1} \le 1/4$ (use the first inequality in (2)).

Let us estimate now the probability that there are two disjoint balls of radius b. If the balls of radius b centered at x_j and x_i are disjoint then x_j is at distance at most $m-2b$ from the string $\bar{x}_i$, that is obtained from x_i by flipping all bits. Therefore the probability that for a fixed pair (i,j) the balls are disjoint is at most $V(m,m-2b)/2^m$. By the second inequality in (2), there are two disjoint balls with probability at most $1/4$.

It remains to estimate the probability that there is a string that belongs to more than $NV(m,b)2^{1-m}$ balls of radius b. Fix x. For every i the probability that x lands in B_i, the ball of radius b centered at x_i, is equal to $p = |B_i|/2^m = V(m,b)/2^m$. So the average number of i with $x \in B_i$ is $pN = NV(m,b)/2^m$. By Chernoff inequality the probability that the number of i such that x lands in B_i exceeds twice the average is at most

$$\exp(-pN/2) = \exp(-NV(m,b)/2^{m+1}) \le \exp(-m) \ll 2^{-m}$$

(use the third inequality in (2)). Thus even after multiplying it by 2^m the number of different x's we get a number close to 0. $\square$

Using this lemma we find x with exponential gap between $C_{0,b}(x|x)$ and $C^u_{0,b}(x|x)$ and even between $C_{0,b}(x|x)$ and $C^u_{a,b}(x|x)$ for a,b linear in the length n of x.

Theorem 4. *Fix rational constants α,β,γ satisfying $\gamma \ge 1$ and*

$$0<\alpha<1/4<\beta<1/2, \quad 2H(\beta)>1+H(2\alpha), \quad 2H(\beta)>1+H(1-2\beta) \quad (3)$$

Notice that if β is close to 1/2 and α is close to 0 then these inequalities are satisfied. Then for all sufficiently large m there is a string x of length $n=\gamma m$ with $C_{0,\beta m}(x|x) = O(\log m)$ while $C^u_{\alpha m,\beta m}(x|x) \ge m(1-H(\beta)) - O(\log m)$.

Proof: Given m let $a=\alpha m$, $b=\beta m$ and $N = m2^{m+1}/V(m,b)$. Let us verify that for large enough m the inequalities (2) in the condition of Lemma 4 are fulfilled. Taking the logarithm of the first inequality (2) and ignoring all terms of order $O(\log m)$ we obtain

$$2(m-mH(\beta)) + mH(2\alpha) < m$$

This is true by the second inequality in (3). Here we used that, ignoring logarithmic terms, $\log V(m,b) = mH(\beta)$ and $\log V(m,2a) = mH(2\alpha)$ as both $\beta, 2\alpha$ are less than 1/2. Taking the logarithm of the second inequality (2) we obtain

$$2(m-mH(\beta)) + mH(1-2\beta) < m.$$

This is implied by the third inequality in (3). Finally, the last inequality (2) holds by the choice of N.

Find the first sequence $x_1, \ldots, x_N$ satisfying the lemma. This sequence has complexity at most $C(m) = O(\log m)$. Append 0^{n-m} to all strings $x_1, \ldots, x_N$. Obviously the resulting sequence also satisfies the lemma. For each string x_i we have $C_{0,b}(x_i|x_i) = O(\log m)$, as given any x' at distance at most b from x_i we can specify x_i by specifying its index among centers of the balls in the family containing x' in $\log(NV(m,b)2^{1-m}) = \log 4m$ bits and specifying the family itself in $O(\log m)$ bits.

It remains to show that there is x_i with $C^u_{a,b}(x_i|x_i) \geq \log N$. Assume the contrary and choose for every x_i a program p_i of length less than $\log N$ such that $U(p, x')$ is at distance a from x_i for every x' at distance at most b from x_i. As N is strictly greater than the number of strings of length less than $\log N$, by the Pigeon Hole Principle there are different x_i, x_j with $p_i = p_j$. However the balls of radius b with centers x_i, x_j intersect and there is x' at distance at most b both from x_i, x_j. Hence $U(p, x')$ is at distance at most a both from x_i, x_j, a contradiction. $\square$

Again, at the expense of replacing $O(\log m)$ by $O(m \log\log m / \log m)$ we can prove an analog of Theorem 4 for time bounded complexity. We defer the proof to the final version.

Theorem 5. *There is a polynomial p such that for all sufficiently large m there is a string x of length $n = \gamma m$ with $C^{p(n)}_{0,\beta m}(x|x) = O(m \log\log m / \log m)$ while $C^u_{\alpha m,\beta m}(x|x) \geq m(1-H(\beta)) - O(m \log\log m / \log m)$. (Note that C^u has no time bound; this makes the statement stronger.)*

7 Symmetry of Information

The principle of symmetry of information, independently proven by Kolmogorov and Levin [10], is one of the most beautiful and useful theorems in Kolmogorov complexity. It states $C(xy) = C(x) + C(y|x) + O(\log n)$ for any $x, y \in \{0,1\}^n$. The direction $C(xy) \leq C(x) + C(y|x) + O(\log n)$ is easy to see—given a program for x, and a program for y given x, and a way to tell these programs apart, we can print xy. The other direction of the inequality requires a clever proof.

Looking at symmetry of information in the error case, the easy direction is again easy: The inequality $C_d(xy) \leq C_a(x) + C_{d-a,a}(y|x) + O(\log n)$ holds for any a — let p be a program of length $C_a(x)$ which prints a string x^* within Hamming distance a of x. Let q be a shortest program which, given x^*, prints a string y^* within Hamming distance $d - a$ of y. By definition, $C_{d-a,a}(y|x) = \max_{x'} \min_{y'} C(y'|x') \geq \min_{y'} C(y'|x^*) = |q|$. Now given p and q and a way to tell them apart, we can print the string xy within d errors.

For the converse direction we would like to have the statement

$$\begin{array}{l}\text{For every } d, x, y \text{ there exists } a \leq d \text{ such that} \\ C_d(xy) \geq C_a(x) + C_{d-a,a}(y|x) - O(\log n).\end{array} \qquad (*)$$

We do not expect this statement to hold for every a, as the shortest program for xy will have a particular pattern of errors which might have to be respected in the programs for x and y given x. We now show, however, that even the formulation $(*)$ is too much to ask.

Theorem 6. *For every n and all $d \le n/4$ there exist $x, y \in \{0,1\}^n$ such that for all $a \le d$ the difference*

$$\Delta(a) = (C_a(y) + C_{d-a,a}(x|y)) - C_d(xy)$$

is more than both

$$\log V(n,d) - \log V(n,a), \qquad \log V(n,d+a) - \log V(n,d-a) - \log V(n,a),$$

up to an additive error term of the order $O(\log n)$.

Since $C^u_{d-a,a}(x|y) \ge C_{d-a,a}(x|y)$, Theorem 6 holds for uniform conditional complexity as well.

Before proving the theorem let us show that in the case, say, $d = n/4$ it implies that for some positive ε we have $\Delta(a) \ge \varepsilon n$ for all a. Let $\alpha < 1/4$ be the solution to the equation

$$H(1/4) = H(1/4 + \alpha) - H(1/4 - \alpha).$$

Note that the function in the right hand side increases from 0 to 1 as α increases from 0 to 1/4. Thus this equation has a unique solution.

Corollary 1. *Let $d = n/4$ and let x, y be the strings existing by Theorem 6. Then we have $\Delta(a) \ge n(H(1/4) - H(\alpha)) - O(\log n)$ for all a.*

The proof is simply a calculation and is omitted. Now the proof of Theorem 6.

Proof: Coverings will again play an important role in the proof. Let $\mathcal{C}$ be the lexicographically first minimal size covering of radius d. Choose y of length n with $C(y) \ge n$, and let x be the lexicographically least element of the covering within distance d of y. Notice that $C_d(xy) \le n - \log V(n,d)$, as the string xx is within distance d of xy, and can be described by giving a shortest program for x and a constant many more bits saying "repeat". (In the whole proof we neglect additive terms of order $O(\log n)$). Let us prove first that $C(x) = n - \log V(n,d)$ and $C(y|x) = \log V(n,d_1) = \log V(n,d)$, where d_1 stands for the Hamming distance between x and y. Indeed,

$$\begin{aligned} n \le C(y) \le C(x) + C(y|x) &\le n - \log V(n,d) + C(y|x) \\ &\le n - \log V(n,d) + \log V(n,d_1) \le n. \end{aligned}$$

Thus all inequalities here are equalities, hence $C(x) = n - \log V(n,d)$ and $C(y|x) = \log V(n,d_1) = \log V(n,d)$.

Let us prove now the first lower bound for $\Delta(a)$. As y has maximal complexity, for any $0 \le a \le d$ we have $C_a(y) \ge n - \log V(n,a)$. Summing the inequalities

$$\begin{aligned} -C_d(xy) &\ge -n + \log V(n,d), \\ C_a(y) &\ge n - \log V(n,a), \\ C_{d-a,a}(x|y) &\ge 0, \end{aligned}$$

we obtain the lower bound $\Delta(a) \geq \log V(n,d) - \log V(n,a)$. To prove the second lower bound of the theorem, we need to show that

$$C_{d-a,a}(x|y) \geq \log V(n, d+a) - \log V(n, d-a) - \log V(n,d). \tag{4}$$

To prove that $C_{d-a,a}(x|y)$ exceeds a certain value v we need to find a y' at distance at most a from y such that $C(x'|y') \geq v$ for all x' at distance at most $d-a$ from x. Let y' be obtained from y by changing a random set of a bits on which x and y agree. This means that $C(y'|y,x) \geq \log V(n-d_1, a)$. It suffices to show that

$$C(x|y') \geq \log V(n, d+a) - \log V(n,d).$$

Indeed, then for all x' at distance at most $d-a$ from x we will have

$$C(x'|y') + \log V(n, d-a) \geq C(x|y')$$

(knowing x' we can specify x by its index in the ball of radius $d-a$ centered at x'). Summing these inequalities will yield (4).

We use symmetry of information in the nonerror case to turn the task of lower bounding $C(x|y')$ into the task of lower bounding $C(y'|x)$ and $C(x)$. This works as follows: by symmetry of information,

$$C(xy') = C(x) + C(y'|x) = C(y') + C(x|y').$$

As $C(y')$ is at most n, using the second part of the equality we have $C(x|y') \geq C(x) + C(y'|x) - n$. Recall that $C(x) = n - \log V(n,d)$. Thus to complete the proof we need to show the inequality $C(y'|x) \geq \log V(n, d+a)$, that is, y' is a random point in the Hamming ball of radius $d+a$ with the center at x. To this end we first note that $\log V(n, d+a) = \log V(n, d_1 + a)$ (up to a $O(\log n)$ error term). Indeed, as $a + d \leq n/2$ we have $\log V(n, d+a) = \log \binom{n}{d+a}$ and $\log V(n,d) = \log \binom{n}{d}$. The same holds with d_1 in place of d. Now we will show that $\log V(n,d) - \log V(n, d_1) = O(\log n)$ implies that $\log V(n, d+a) - \log V(n, d_1 + a) = O(\log n)$. It is easy to see that $\binom{n}{d+1}/\binom{n}{d_1+1} \leq \binom{n}{d}/\binom{n}{d_1}$ provided $d_1 \leq d$. Using the induction we obtain $\binom{n}{d+a}/\binom{n}{d_1+a} \leq \binom{n}{d}/\binom{n}{d_1}$.

Thus we have

$$\log V(n, d+a) - \log V(n, d_1 + a) = \log \left(\binom{n}{d+a} / \binom{n}{d_1 + a} \right)$$
$$\leq \log \left(\binom{n}{d} / \binom{n}{d_1} \right) = \log V(n,d) - \log V(n, d_1) = O(\log n).$$

Again we use (the conditional form of) symmetry of information:

$$C(y'y|x) = C(y|x) + C(y'|y,x) = C(y'|x) + C(y|y',x).$$

The string y differs from y' on a bits out of the $d_1 + a$ bits on which y' and x differ. Thus $C(y|y',x) \leq \log \binom{d_1+a}{a}$. Now using the second part of the equality we have

$$\begin{aligned} C(y'|x) &= C(y|x) + C(y'|y,x) - C(y|y',x) \\ &\geq \log V(n,d_1) + \log V(n-d_1,a) - \binom{d_1+a}{a}. \end{aligned}$$

We have used that $\log V(n-d_1,a) = \log \binom{n-d_1}{a}$, as $a \leq (n-d_1)/2$. Hence,

$$C(y'|x) \geq \log \binom{n}{d_1} + \log \binom{n-d_1}{a} - \log \binom{d_1+a}{a} = \log V(n, d+a).$$

$\square$

Again, at the expense of replacing $O(\log n)$ by $O(n \log\log n/\log n)$ we can prove an analog of Theorem 6 for time bounded complexity.

Acknowledgment

We thank Harry Buhrman for several useful discussions and the anonymous referees for valuable remarks and suggestions.

References

1. G. Cohen, I. Honkala, S. Litsyn, and A. Lobstein. *Covering Codes*. North-Holland, Amsterdam, 1997.
2. T. Cormen, C. Leiserson, and R. Rivest. *Introduction to Algorithms*. MIT Press, 1990.
3. E. Dantsin, A. Goerdt, E. Hirsch, and U. Schöning. Deterministic algorithms for k-SAT based on covering codes and local search. In *Proceedings of the 27th International Colloquium On Automata, Languages and Programming*, Lecture Notes in Computer Science, pages 236–247. Springer-Verlag, 2000.
4. R. Impagliazzo, R. Shaltiel, and A. Wigderson. Extractors and pseudo-random generators with optimal seed length. In *Proceedings of the 32nd ACM Symposium on the Theory of Computing*, pages 1–10. ACM, 2000.
5. M. Krivelevich, B. Sudakov, and V. Vu. Covering codes with improved density. *IEEE Transactions on Information Theory*, 49:1812–1815, 2003.
6. M. Li and P. Vitányi. *An Introduction to Kolmogorov Complexity and its Applications*. Springer-Verlag, New York, second edition, 1997.
7. R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, 1997.
8. B. Natarajan. Filtering random noise from deterministic signals via data compression. *IEEE transactions on signal processing*, 43(11):2595–2605, 1995.
9. N. Vereschagin and P. Vitányi. Algorithmic rate-distortion theory. http://arxiv.org/abs/cs.IT/0411014, 2004.
10. A. Zvonkin and L. Levin. The complexity of finite objects and the algorithmic concepts of information and randomness. *Russian Mathematical Surveys*, 25:83–124, 1970.