Clustering with Respect to the Information Distance

Andrei Romashchenko

Clustering with Respect to the Information Distance

Andrei Romashchenko

June 29, 2022

Abstract

We discuss the notion of a dense cluster with respect to the information distance and prove that all such clusters have an extractable core that represents the mutual information shared by the objects in the cluster.

1 Introduction

In the seminal paper [1], Bennett et al. introduced the notion of *information distance* based on Kolmogorov complexity. Loosely speaking, the distance between two individual finite objects is defined as the length of the shortest program that can translate these objects to each other. A surprising result proven in [1] claims that the length of such a program (that performs the translation of the given objects to each other, in both directions) is substantially equal to the maximum of the lengths of two separate programs translating the objects to each other. The optimal lengths of the programs depend, of course, on the choice of the programming language. Such a choice may look arbitrary. However, the framework of Kolmogorov complexity (see [2]) provides us with an *optimal* programming language where the required programs have the minimum (up to an additive constant) possible length.

The notion of information distance attracted the attention of theoretical computer scientists and also inspired many experimental works, where various practical approximations of the information distance were computed for real-world data. This technique was typically used to reveal clusters and classify data of some specific type (texts, music recordings, genetic codes, and so on). In such experiments, the revealed clusters (groups of objects with small pairwise information distances) consist of the data having something in common: texts written in the same language, music pieces of the same genre, genetic information of closely related species, and so on (see, e.g, [3, 4, 5]). In the studied practical examples, it can be usually observed that the revealed dense clusters have some core (e.g., the common language vocabulary, specific characteristics of a particular music genre, the genetic information of the common predecessor of several biological species), even if providing an explicit description of such a core is not immediate.

In this paper, we study a similar phenomenon in a purely theoretical setting. We show that the *only* reason why a large set of objects may form a dense cluster with respect to the information distance is that these objects share some *common information* in the sense of Gács and Körner [6]. In other words, there must exist a *core* that is simple conditional on each object of the cluster and that represents in some sense the mutual information shared by all these objects. Such results were used (more or less explicitly) as technical tools in [7, 8, 9]. We believe that the observed phenomenon is interesting in its own right, and we want to draw more attention to this issue. In this paper, we propose a self-contained explanation of the mentioned result using a simplified definition of a cluster proposed by Alexander Shen (personal communication, June 10, 2021), see Definition 1 below. The proofs of the main results follow the arguments suggested in [7, 8].

Notation and standard properties of Kolmogorov complexity In what follows we use the standard notation $\mathrm{C}(x)$ for the plain Kolmogorov complexity of a string x and $\mathrm{C}(x|y)$ for the plain Kolmogorov complexity of a string x conditional on a string y (see, e.g., [10] or [11]). In this paper we do not discuss prefix-free complexity, monotone complexity, or any other subtler versions of algorithmic complexity, so we may use for $\mathrm{C}(x)$ and $\mathrm{C}(x|y)$ the term *Kolmogorov complexity* without risk of ambiguity.

As usual, we assume to be fixed an *encoding* of tuples, i.e., a computable bijection between binary strings and tuples (finite ordered lists) of strings. Thus, every tuple of strings is associated with its *code*, which is an individual binary string. Keeping this in mind, we assume that $\mathrm{C}(x, y)$ denotes Kolmogorov complexity of the code of the pair $\langle x, y\rangle$; $\mathrm{C}(x, y, z)$ denotes Kolmogorov complexity of the code of the triple $\langle x, y, z\rangle$, and so on.

Observe that any two encodings of this type (computable bijections between tuples and individual strings) are equivalent: there are translation algorithms converting a code of a tuple in one encoding system into the code of the same tuple in the other encoding systems, and the other way around. This means that the choice of the encoding is quite arbitrary and affects the value of Kolmogorov complexity by at most an additive constant, see [11, Section 2.1].

The classical Kolmogorov–Levin theorem (the *chain rule*), [12, 13], establishes the relation between Kolmogorov complexity of a pair and conditional Kolmogorov complexity,

$$\mathrm{C}(x, y) = \mathrm{C}(x) + \mathrm{C}(y|x) + O(\log(\mathrm{C}(x) + \mathrm{C}(y))).$$

The mutual information of two strings and the conditional mutual information are defined as $\mathrm{I}(x : y) \stackrel{\text{def}}{=} \mathrm{C}(y) - \mathrm{C}(y|x)$ and $\mathrm{I}(x : y|z) \stackrel{\text{def}}{=} \mathrm{C}(y|z) - \mathrm{C}(y|x, z)$ respectively. From the Kolmogorov–Levin theorem it follows that mutual information is symmetric up to a logarithmic additive term:

$$\mathrm{I}(x : y) = \mathrm{C}(x) + \mathrm{C}(y) - \mathrm{C}(x, y) + O(\log n) = \mathrm{I}(y : x) + O(\log n)$$

and

$$\mathrm{I}(x:y|z) = \mathrm{C}(x,z)+\mathrm{C}(y,z)-\mathrm{C}(x,y,z)-\mathrm{C}(z)+O(\log m) = \mathrm{I}(y:x|z)+O(\log m),$$

where $n = \mathrm{C}(x) + \mathrm{C}(y)$ and $m = \mathrm{C}(x) + \mathrm{C}(y) + \mathrm{C}(z)$.

2 Clusters with respect to the information distance

The *information distance between* x and y can be defined as

$$\mathrm{dist}(x,y) = \max\{\mathrm{C}(x|y), \mathrm{C}(y|x)\}.$$

(It is not a distance in the proper sense since the triangle inequality is true only up to an additive logarithmic term.) Information distance measures the amount of information needed to obtain one of the strings given another one. Speaking informally, this value can be understood as a measure of "similarity" (or rather "non-similarity") between strings: if x, y, z are three strings of the same length, and $\mathrm{dist}(x,y)$ is much less than $\mathrm{dist}(x,z)$, we can say that x is more "similar" to y than to z.

We can consider clusters defined in the sense of this information distance, i.e., large sets of strings with small diameters. Roughly speaking, a cluster is a set of strings of cardinality at least 2^m and diameter at most m. As is usual in the theory of Kolmogorov complexity, we should admit a minor imprecision (say, logarithmic in m) of the parameters. The formal definition of a cluster involves two parameters, the diameter and the logsize (logarithm of the cardinality):

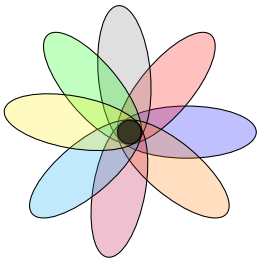
Definition 1. We say that a set of strings S is an (m,ℓ)-*cluster* if for all $x_1, x_2 \in S$ we have $\mathrm{dist}(x_1,x_2) \le m$ and $\#S \ge 2^\ell$. An equivalent wording: we can say that this S is a cluster with parameters (m,ℓ). The minimal suitable value of m is called the cluster's *diameter* and the maximal suitable integer number ℓ is called the cluster's *logsize*.

The "density" of a cluster can be measured by the difference between the diameter and the logsize: the closer they are to each other, the denser is the cluster. We usually deal with clusters where this difference is bounded by $O(\log m)$.

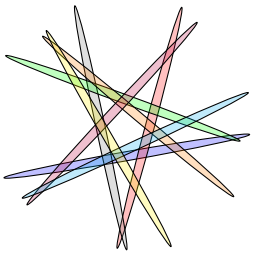
First of all, do the dense clusters exist? The answer to this question is yes, we can find (m,ℓ)-clusters with only a logarithmic gap between m and ℓ. Indeed, for every string z we may consider a "canonical" cluster or a *daisy* that consists of strings x such that $\mathrm{C}(z|x) \approx 0$ and $\mathrm{C}(x|z) \lesssim m$. To make these approximate equality and inequality more specific, we fix a parameter d and define a daisy with imprecision d as follows.

Definition 2. An (m,d)-*daisy with a core* z is the set of all x such that

$$\mathrm{C}(z|x) \le d \text{ and } C(x|z) \le m + d.$$



(a) A bunch of intersecting sets having a common core.



(b) A bunch of pairwise intersecting sets with no common core.

Figure 1: Geometric representation of clusters.

Observe that the definition of a daisy involves two integer parameters, but they do not play the same role as the parameters in the general definition of an (m, ℓ)-cluster.

Every (m, d)-*daisy* with $d = O(\log m)$ is a cluster, i.e., it satisfies Definition 1 with a logarithmic gap between the diameter and the logsize. Indeed, from the definition of an (m, d)-daisy it follows that for all x_1, x_2 in this set we have

$$\mathrm{C}(x_1 \,|\, x_2) \le \mathrm{C}(z \,|\, x_2) + \mathrm{C}(x_1 \,|\, z) + O(\log m).$$

A similar bound applies to $\mathrm{C}(x_2 \,|\, x_1)$. Therefore,

$$\mathrm{dist}(x_1, x_2) \le m + O(d + \log m) = m + O(\log m).$$

The cardinality of this set is at least 2^m since it contains all pairs $\langle z, w \rangle$ where w is an m-bit string. Thus, this set is an $(m + O(\log m), m)$-cluster.

A daisy is by definition a cluster with an explicitly given core. A daisy of strings intuitively resembles a flower with a core and many petals. We can draw it on the plane as a family of 2^m pairwise intersecting sets (so that the mutual information of every two strings corresponds to the size of the intersection between two sets) such that all these sets have a common part (the core), as shown in Fig. 1a, and the size of each "petal" (outside the core) is at most m. However, a naive parallelism between the mutual information and intersections of sets can be deceiving. Indeed, Gács and Körner showed that the mutual information of two strings may not correspond to any material object (see [6]). Moreover, even if the mutual information of each pair of objects can be "materialized," it seems possible that pairwise intersecting objects do not share any common core, as in the example in Fig. 1b.

So a natural question arises: do there exist clusters substantially different from a daisy? Rather surprisingly, it turns out that there are no other clusters besides the daisies and their subsets. We can say informally that all clusters (in the sense of Definition 1) resemble Fig. 1a and not Fig. 1b. This is the main result of this paper.

Main Result (informal version)**.** *Every cluster in the sense of information distance is a sufficiently large subset of some daisy.*

Observe that every large enough subset of a cluster is still a cluster (of high enough density). Thus, this theorem may be interpreted as a description of all dense enough clusters as sufficiently large parts of daisies.

Now we proceed with a more formal statement.

Theorem 1 (**main result**, the formal version)**.** *Let S be a set of strings such that $\mathrm{C}(x|x') \le m$ for every two strings $x, x' \in S$. Assume that $\log \#S \ge m - d$ for some d. Then there exists a string z such that*

$$\mathrm{C}(z|x) \le O(d + \log m) \text{ and } \mathrm{C}(x|z) \le m + O(d + \log m)$$

for all $x \in S$.

Theorem 1 can be naturally rephrased in terms of clusters and daisies: it claims that for every $(m, m-d)$-cluster S there exists a string z such that S is included in an $(m + O(d + \log m), O(d + \log m))$-daisy with the core z. Notice that the found core z possibly does not belong to S. In fact, a cluster may even not contain any element close to its core.

Proof. We start the proof with the following lemma.

Lemma 1 (Many paths $x - y - z$ imply one shorter path $x - z$)**.** *Assume that for given strings x, z and for given numbers u, v, w there are at least 2^u strings y such that*

$$\mathrm{C}(y|x) < v \quad \text{and} \quad \mathrm{C}(z|y) < w.$$

Then $\mathrm{C}(z|x) \le v + w - u + O(\log(v + w))$.

Proof of Lemma 1. Given x, v, w, we can enumerate all z for which there exists y with the required properties. There are at most 2^{v+w} paths of length 2 and at least 2^u of these paths should lead to such a z. So there are at most 2^{v+w-u} different z, and this implies the bound for $\mathrm{C}(z|x)$. □

The previous lemma can be used to merge clusters, as the following remark shows.

Remark. (Merging two clusters). If S and S' are two clusters of diameter m that have at least 2^{m-d} common elements, then their union is a cluster of diameter at most $m + d + O(\log m)$. Indeed, if x and x' are elements from S and S' respectively, then there are at least 2^{m-d} paths $x - x'' - x'$ such that $x'' \in S \cap S'$. Therefore, from Lemma 1 it follows that $\mathrm{dist}(x, x') \le 2m - (m-d) + O(\log m) = m + d + O(\log m)$.

More specifically, we will need the following version of cluster merging:

Lemma 2 (Merging a cluster with a daisy)**.** *Assume that an $(m+d_1, m-d_2)$-cluster S has at least 2^{m-d_3} common elements with an (m, d_4)-daisy S' with a core z. Then S is contained in the daisy S'' with the same core z with the parameters*

$$\Big(m + O\big(\sum_i d_i + \log m\big), O\big(\sum_i d_i + \log m\big)\Big). \tag{1}$$

Proof. For every $x \in S$ there are at least 2^{m-d_3} chains $z - x' - x$ such that $x' \in S \cap S'$. From Lemma 1 it follows that

$$\mathrm{C}(z|x) \le m + d_1 + d_4 - (m - d_3) + O(\log m) = O\big(\sum_i d_i + \log m\big)$$

and

$$\mathrm{C}(x|z) \le m + d_4 + m + d_1 - (m - d_3) + O(\log m) = m + O\big(\sum_i d_i + \log m\big).$$

Therefore, x belongs to the daisy S'' with parameters (1) and the base z. □

To prove the theorem, it is enough (thanks to Lemma 2) to find a daisy with parameters

$$\big(m + O\big(d + \log m\big), O\big(d + \log m\big)\big)$$

that has a large (of cardinality at least $2^{m-O\left(d+\log m\right)}$) intersection with the given cluster S. We do it as follows. The property of being a cluster with given parameters is enumerable. So we can run a process enumerating all $(m, m-d)$-clusters. We do not restrict the length or complexity of strings in the clusters, so the enumeration will be infinite. As any other cluster with the same parameters, our cluster S will be enumerated at some stage of this process.

Let us fix some threshold d' (that will be slightly greater than d, see below). We say that two clusters S_1, S_2 have a *large* intersection if $\#(S_1 \cap S_2) > 2^{m-d'}$. To make the enumeration procedure defined above more economic, we will drop some clusters from this enumeration. We will keep only the ones that do not have large intersections with one of the clusters enumerated (and not dropped) earlier. We call the clusters that are not dropped *referential clusters*. We assign to the referential clusters their ordinal numbers in the order they appear in the enumeration. (Observe again that there can be infinitely many referential clusters.) We will see that either S itself or some cluster that has a large intersection with S will become a referential cluster, and we plan to take its ordinal number in the enumeration as the core z of the daisy that we are looking for.

First of all, we argue that every referential cluster is a part of a daisy with parameters

$$\big(m + O(d + \log m), O(d + \log m)\big).$$

Let S_i be the referential cluster with ordinal number i. We start with the observation that every element $x \in S_i$ can be determined if we know i and the ordinal number of x in some standard ordering of S_i. To organize the process

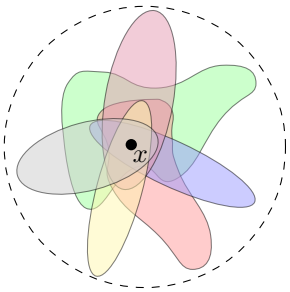


Figure 2: An element x (shown as a dot) is covered by a family of sets (shown in different colors) all of which lie in a neighborhood of this element (shown as an area with a dashed borderline).

of enumeration of the referential clusters we also need to know the number m, which requires $O(\log m)$ bits. Therefore, $\mathrm{C}(x|i) \le m + O(d + \log m)$ for all $x \in S_i$. We want to show that S_i is a part of a daisy with a core i. To this end, we show that $\mathrm{C}(i|x) \approx 0$ for all $x \in S_i$. We use the following lemma saying that (under some conditions on the parameters) the multiplicity of the family S_i is small, i.e., every x is covered by only a small number of S_i. Then, to reconstruct i given x, we need only the ordinal number of S_i in the list of referential clusters containing x (and also the number m, as before).

Lemma 3 (Multiplicity bound)**.** *Assume that $d' > 2d + 1$. Then every string x can be covered by at most 2^{d+1} referential $(m, m-d)$-clusters.*

Proof of Lemma 3. Assume that some string x is covered by N referential clusters. Each cluster S_i has size at least 2^{m-d}, and the intersection of every two clusters is at most $2^{m-d'}$ (otherwise the second cluster could not be selected as a referential one). Note also that all elements of all clusters that contain x have conditional complexity at most m conditional on x. Therefore, the union of all these clusters has a size at most 2^{m+1} (all clusters covering x are included in a rather small neighborhood of x, see Fig. 2).

We use the following probabilistic claim:

Claim. *Let ε be the inverse of a positive integer number*[1]*. If there are N events of probability greater than ε, and all pairwise intersections have probability less than $\varepsilon^2/2$, then $N < 2/\varepsilon$.*

(The bounds in the claim are pretty tight: $1/\varepsilon$ events of probability ε could be disjoint, and any number of independent events of probability ε have intersection ε^2.)

Proof. Assume that we have $N = 2/\varepsilon$ events (we decrease N if needed). By the principle of inclusion and exclusion, the probability of the union of these events

[1] A similar claim is true for all real numbers $\varepsilon > 0$. The assumption that $1/\varepsilon$ is an integer number slightly simplifies the calculations since we can ignore rounding.

is strictly greater than

$$N \cdot \varepsilon - \frac{N^2}{2} \cdot \frac{\varepsilon^2}{2} = 2 - \frac{4}{2\varepsilon^2} \cdot \frac{\varepsilon^2}{2} = 1,$$

a contradiction. □

We consider the union of all $(m, m-d)$-clusters covering x as the probability space with equiprobable points, where each cluster is an event. To apply this claim, we note that each cluster has a probability of at least $\varepsilon := 2^{-d}$, and the intersections are of probability at most $2^{-d'}$. Since $d' > 2d+1$, we can apply the Claim and obtain the bound 2^{d+1} for the number of clusters covering x. Therefore, we have

$$\mathrm{C}(x|i) \le m + O(\log m) \quad \text{and} \quad \mathrm{C}(i|x) \le d + O(\log m)$$

for every element x of every referential cluster S_i. Thus, each referential cluster is a part of an $\big(m + O(d + \log m), O(d + \log m)\big)$-daisy. □

Now we bind together all parts of the argument. Assume that we have an $(m-d, m)$-cluster S. We let $d' = 2d+2$ (so the condition of Lemma 3 is true) and start the process of enumeration of referential clusters using d' as the "large intersection" threshold. The construction guarantees that S is one of the referential clusters or at least it has a large intersection with some referential cluster S_i. Every referential cluster S_i is a part of an $\big(m + O(d + \log m), O(d + \log m)\big)$-daisy. Therefore, we can apply Lemma 2 and conclude that S is a part of a slightly bigger $\big(m + O(d + \log m), O(d + \log m)\big)$-daisy, and the theorem is proven. □

3 Clusters and the mutual information of a triple

In this section, we discuss an application of Theorem 1 that motivated the definition of "bunches" proposed in [7] (similar to the definition of clusters discussed in the previous section).

Theorem 2 ([7, 8])**.** *For every triple of strings x, y, z there exists a string w such that*

$$\mathrm{C}(w) = \mathrm{I}(x : y : z) + O(\varepsilon + \log \mathrm{C}(x, y, z))$$

and

$$\max\{\mathrm{C}(w|x), \mathrm{C}(w|y), \mathrm{C}(w|z)\} = O(\varepsilon + \log \mathrm{C}(x, y, z)),$$

where $\varepsilon := \max\big\{\mathrm{I}(x : y|z), \mathrm{I}(x : z|y), \mathrm{I}(y : z|x)\big\}$ and

$$\mathrm{I}(x : y : z) := \mathrm{C}(x) + \mathrm{C}(y) + \mathrm{C}(z) - \mathrm{C}(x, y) - \mathrm{C}(x, z) - \mathrm{C}(y, z) + \mathrm{C}(x, y, z).$$

In particular, if the three values of conditional mutual information $\mathrm{I}(x : y|z)$, $\mathrm{I}(x : z|y)$, $\mathrm{I}(y : z|x)$ are negligibly small (say, logarithmic in $\mathrm{C}(x, y, z)$) as shown in Fig. 3, then the mutual information shared by x, y, z can be materialized in the sense of *common information* by Gács and Körner, [6].

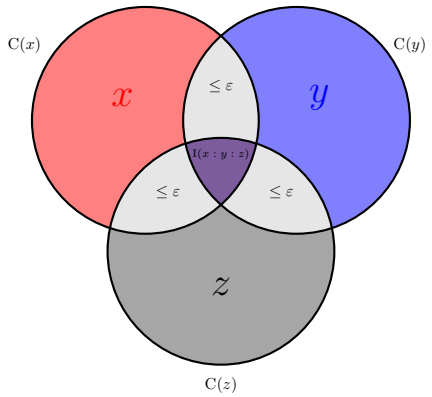


Figure 3: A Venn-like diagram representing information quantities for a triplet (x, y, z). The area of each circle represents the value of Kolmogorov complexity of x, y, and z respectively. The areas of the unions of any two circles represent Kolmogorov complexity of pairs, and the area of the union of all three circles represents Kolmogorov complexity of the triple. Accordingly, the intersections of every two circles represent the mutual information of pairs, and the intersection of all three circles represents the value $\mathrm{I}(x : y : z)$. The areas shown in light gray represent the values of the three conditional mutual information $\mathrm{I}(x : y|z)$, $\mathrm{I}(x : z|y)$, $\mathrm{I}(y : z|x)$. Theorem 2 claims that if the three values of conditional mutual information are negligibly small, then the mutual information of the triple $\mathrm{I}(x : y : z)$ (which is in this case ε-close to each of the values $\mathrm{I}(x : y)$, $\mathrm{I}(x : z)$, and $\mathrm{I}(y : z)$) can be materialized.

Proof. For each triple of strings (x, y, z) we call by its *complexity profile* the tuple of seven complexity quantities

$$\big(\mathrm{C}(x), \mathrm{C}(y), \mathrm{C}(z), \mathrm{C}(x, y), \mathrm{C}(x, z), \mathrm{C}(y, z), \mathrm{C}(x, y, z)\big).$$

For every precision parameter (an integer number) δ we define the set $\mathcal{C}_\delta$ of *δ-clones of z conditional on (x, y)* as the set of all z' such that the complexity profile of (x, y, z') differs in each component from the complexity profile of (x, y, z) by at most δ.

A simple counting argument implies (see, e.g., [14]) that there exists a constant D (independent of x, y, z) such that for all x, y, z the set of δ-clones of z conditional on (x, y) with $\delta = D \log \mathrm{C}(x, y, z)$ consists of $2^{\mathrm{C}(z|x,y)-O(\delta)}$ strings z'. In what follows we fix such a δ.

We claim that $\mathcal{C}_\delta$ is a cluster. To prove this fact we need the following inequality, where z' and z'' are two arbitrary elements from $\mathcal{C}_\delta$:

$$\begin{aligned}\mathrm{I}(x : y) \quad \leq \quad & \mathrm{I}(x : y|z') + \mathrm{I}(x : y|z'') + \mathrm{I}(z' : z'') \\ & + \mathrm{I}(x : y|z) + \mathrm{I}(x : z|y) + \mathrm{I}(y : z|x) + O(\log \mathrm{C}(x, y, z)).\end{aligned}$$

This inequality[2] is valid for all strings x, y, z, z', z'', see [17]. Since z' and z'' are clones of z conditional on x, y, the inequality rewrites to

$$\mathrm{I}(x:y) \quad \leq \quad \mathrm{I}(z':z'') + O(\varepsilon + \log \mathrm{C}(x,y,z)).$$

It follows that

$$\begin{array}{rcl} \mathrm{C}(z''|z') & = & \mathrm{C}(z'') - \mathrm{I}(z':z'') \\ & \leq & \mathrm{C}(z) - \mathrm{I}(x:y) + O(\varepsilon + \log \mathrm{C}(x,y,z)) \\ & = & \mathrm{C}(z|x,y) + O(\varepsilon + \log \mathrm{C}(x,y,z)). \end{array}$$

Therefore, $\mathcal{C}_\delta$ is a cluster with the parameters

$$\big(\mathrm{C}(z|x,y) + O(\delta + \log \mathrm{C}(x,y,z)), \mathrm{C}(z|x,y) - O(\delta + \log \mathrm{C}(x,y,z))\big).$$

Theorem 1 implies that there exists a string w (the core of the cluster) such that $\mathrm{C}(w|\hat{z}) = O(\varepsilon + \log \mathrm{C}(x,y,z))$ and $\mathrm{C}(\hat{z}|w) = \mathrm{C}(z|x,y) + O(\varepsilon + \log \mathrm{C}(x,y,z))$ for all $\hat{z} \in \mathcal{C}_\delta$ (including the original string z). It is easy to compute Kolmogorov complexity of w:

$$\mathrm{C}(w) = \mathrm{I}(z : \langle x, y \rangle) + O(\varepsilon + \log \mathrm{C}(x,y,z)) = \mathrm{I}(x:y:z) + O(\varepsilon + \log \mathrm{C}(x,y,z)).$$

It remains to observe that due to Lemma 1

$$\mathrm{C}(w|x) = O(\varepsilon + \log \mathrm{C}(x,y,z)) \text{ and } \mathrm{C}(w|y) = O(\varepsilon + \log \mathrm{C}(x,y,z))$$

(it is enough to count the number of chains $x - \hat{z} - w$ and $y - \hat{z} - w$ with $\hat{z} \in \mathcal{C}_\delta$). □

4 Discussion

The questions addressed in this paper seem to be related to the density properties studied in [1, Section IX], where the authors estimated the rate of growth of the number of elements in balls of radius r in the metric spaces induced by the information distance. We should stress, however, that a ball (the set of strings x' at the distance at most r from a given center x) is not a cluster in the sense of Definition 1.

The main result of this theorem is formulated and proven with a "logarithmic precision," which is quite typical for the theory of Kolmogorov complexity. In many applications the logarithmic precision is enough. At the same time, it seems that the residue terms in Theorem 1 can be made somewhat tighter. In this vein, an anonymous referee of the *Theoretical Computer Science* journal suggested the following stronger version of Theorem 1:

[2]This inequality is a *non–Shannon type* one, i.e., it cannot be represented as a linear combination of several instances of inequalities representing non negativity of conditional Kolmogorov complexity, or mutual information, or conditional mutual information. The very first example of a non–Shannon type linear inequality for Shannon's entropy was proven by Zhang and Yeung in [15]. It is known, see [16], that the same linear inequalities are true for Shannon's entropy and Kolmogorov complexity. The inequality used in our proof is a little generalization of the inequality discovered by Zhang and Yeung, see [17] for details.

Theorem $1'$**.** *Let S be a set of strings such that $C(x|x') \le m$ for every two strings $x, x' \in S$. Assume that $\log \#S \ge m - d$ for some d. Then there exists a string z such that $C(z|x,m) < O(d)$ and $C(x|z,m) < m + O(d)$.*

(Compared with Theorem 1, the conclusion of this statement contains no additive terms $O(\log m)$; on the other hand, the number m is included into the conditions of two expressions with Kolmogorov complexity.) This version of the theorem can be proven by an argument very similar to the proof of Theorem 1 presented in Section 2; we only need to relativize to m the expressions with Kolmogorov complexity that appear in the proof. However, if we want to rephrase Theorem $1'$ in terms of clusters and daisies (cf. the paragraph after Theorem 1 on p. 5), we would need to revise the definition of a daisy. This observation suggests that the definitions of clusters and daisies might need to be refined.

Let us mention also that slightly different variants of the definition of a cluster may be helpful in some applications, see [9]. An interesting variant of the definition was proposed by S. Epstein in [18], where the principal parameter was not the *maximum* but the *average* distance between elements of a cluster. Epstein argued that the density of a cluster is connected with the mutual information between this cluster and the *halting sequence* (characterizing the stopping Turing machine in the universal enumeration). Thus, the formulation of the most natural and practical definition of a dense cluster (in the sense of information distance) remains an open question.

Acknowledgments. The author is grateful to Alexander Shen and Marius Zimand for fruitful discussions, especially for the elegant form of the probabilistic claim (see the *Claim* on p. 7) suggested by Alexander Shen. The author also thanks the anonymous referees of the Theoretical Computer Science journal for the careful review of the paper and valuable comments and suggestions.

References

[1] Charles H Bennett, Péter Gács, Ming Li, Paul MB Vitányi, and Wojciech H Zurek. Information distance. *IEEE Transactions on information theory*, 44(4):1407–1423, 1998.

[2] Andrei N Kolmogorov. Three approaches to the quantitative definition of information. *Problems of information transmission*, 1(1):1–7, 1965.

[3] Ming Li, Xin Chen, Xin Li, Bin Ma, and Paul MB Vitányi. The similarity metric. *IEEE transactions on Information Theory*, 50(12):3250–3264, 2004.

[4] Rudi Cilibrasi, Paul Vitányi, and Ronald de Wolf. Algorithmic clustering of music based on string compression. *Computer Music Journal*, 28(4):49–67, 2004.

[5] Rudi Cilibrasi and Paul MB Vitányi. Clustering by compression. *IEEE Transactions on Information theory*, 51(4):1523–1545, 2005.

[6] Peter Gács and János Körner. Common information is far less than mutual information. *Problems of Control and Information Theory*, 2(2):149–162, 1973.

[7] Andrei Romashchenko. Extracting the mutual information for a triple of binary strings. In *18th IEEE Annual Conference on Computational Complexity, 2003. Proceedings.*, pages 221–229. IEEE, 2003.

[8] Andrei E Romashchenko. A criterion for extractability of mutual information for a triple of strings. *Problems of Information Transmission*, 39(1):148–157, 2003.

[9] An A Muchnik and Andrei E Romashchenko. Stability of properties of kolmogorov complexity under relativization. *Problems of information transmission*, 46(1):38–61, 2010.

[10] Ming Li and Paul Vitányi. *An introduction to Kolmogorov complexity and its applications*. Springer, 3 edition, 2008.

[11] Alexander Shen, Vladimir A Uspensky, and Nikolay Vereshchagin. *Kolmogorov complexity and algorithmic randomness*, volume 220. American Mathematical Society, 2022.

[12] Andrei Kolmogorov. Logical basis for information theory and probability theory. *IEEE Transactions on Information Theory*, 14(5):662–664, 1968.

[13] Alexander K Zvonkin and Leonid A Levin. The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms. *Russian Mathematical Surveys*, 25(6):83–124, 1970.

[14] AE Romashchenko. Pairs of words with nonmaterializable mutual information. *Problems of Information Transmission*, 36(1):3–20, 2000.

[15] Zhen Zhang and Raymond W Yeung. On characterization of entropy function via information inequalities. *IEEE Transactions on Information Theory*, 44(4):1440–1452, 1998.

[16] Daniel Hammer, Andrei Romashchenko, Alexander Shen, and Nikolai Vereshchagin. Inequalities for Shannon entropy and Kolmogorov complexity. *Journal of Computer and System Sciences*, 60(2):442–464, 2000.

[17] Konstantin Makarychev, Yury Makarychev, Andrei Romashchenko, and Nikolai Vereshchagin. A new class of non-Shannon-type inequalities for entropies. *Communications in Information and Systems*, 2(2):147–166, 2002.

[18] Samuel Epstein. On the conditional complexity of sets of strings. *arXiv preprint arXiv:1907.01018*, 2019.