Counting dependent and independent strings

Marius Zimand [*]

Department of Computer and Information Sciences, Towson University, Baltimore, MD, USA

Abstract. We derive quantitative results regarding sets of n-bit strings that have different dependency or independency properties. Let $C(x)$ be the Kolmogorov complexity of the string x. A string y has α dependency with a string x if $C(y) - C(y \mid x) \geq \alpha$. A set of strings $\{x_1, \ldots, x_t\}$ is pairwise α-independent if for all $i \neq j$, $C(x_i) - C(x_i \mid x_j) \leq \alpha$. A tuple of strings $(x_1, \ldots, x_t)$ is mutually α-independent if $C(x_{\pi(1)} \ldots x_{\pi(t)}) \geq C(x_1) + \ldots + C(x_t) - \alpha$, for every permutation π of $[t]$. We show that:

- For every n-bit string x with complexity $C(x) \geq \alpha + 7 \log n$, the set of n-bit strings that have α dependency with x has size at least $(1/\text{poly}(n))2^{n-\alpha}$. In case α is computable from n and $C(x) \geq \alpha + 12 \log n$, the size of same set is at least $(1/C)2^{n-\alpha} - \text{poly}(n)2^\alpha$, for some positive constant C.

- There exists a set of n-bit strings A of size $\text{poly}(n)2^\alpha$ such that any n-bit string has α-dependency with some string in A.

- If the set of n-bit strings $\{x_1, \ldots, x_t\}$ is pairwise α-independent, then $t \leq \text{poly}(n)2^\alpha$. This bound is tight within a $\text{poly}(n)$ factor, because, for every n, there exists a set of n-bit strings $\{x_1, \ldots, x_t\}$ that is pairwise α-dependent with $t = (1/\text{poly}(n)) \cdot 2^\alpha$ (for all $\alpha \geq 5 \log n$).

- If the tuple of n-bit strings $(x_1, \ldots, x_t)$ is mutually α-independent, then $t \leq \text{poly}(n)2^\alpha$ (for all $\alpha \geq 7 \log n + 6$).

1 Introduction

A fact common to many mathematical settings is that in a sufficiently large set some relationship emerges among its elements. Generically, these are called Ramsey-type results. We list just a few examples: any $n+1$ vectors in an n-dimensional vector space must be dependent; for every k and sufficiently large n, any subset of $[n]$ of constant density must have k elements in arithmetic progression; any set of 5 points in the plane must contain 4 points that form a convex polygon. All these results show that in a sufficiently large set, some attribute of one element is determined by the other elements.

We present in this paper a manifestation of this phenomenon in the very general framework of algorithmic information theory. We show that in a sufficiently large set some form of algorithmical dependency among its elements must exist. Informally speaking, $\text{poly}(n) \cdot 2^\alpha$ binary strings of length n must share at least α bits of information. For one interpretation of "share", we also show that this bound is tight within a $\text{poly}(n)$ factor.

Central to our investigation are the notions of information in a string and the derived notion of dependency between strings. The information in a string x is captured by its Kolmogorov complexity $C(x)$. A string y has α-dependency with string x if $C(y) - C(y \mid x) \geq \alpha$. The expression $C(y) - C(y \mid x)$, denoted usually more concisely as $I(x : y)$, represents *the quantity of information in x about y* and is a key concept in information theory. It

---

[*] http://triton.towson.edu/~mzimand.

is known that $I(x : y) = I(y : x) \pm O(\log n)$ (Symmetry of Information Theorem [20]), where n is the length of the longer between the strings x and y, and therefore $I(x : y)$ is also called the mutual information of x and y. For any n-bit string x and positive integer α, we are interested in estimating the size of the set $A_{x,\alpha}$ of n-bit strings y such that $C(y) - C(y \mid x) \geq \alpha$. One can see by a standard counting argument that $|A_{x,\alpha}| \leq 2^{n-\alpha+c}$ for some constant c. Regarding a lower bound for $|A_{x,\alpha}|$, it is easy to see that if $C(x) \preceq \alpha$, then $A_{x,\alpha}$ is empty (intuitively, in order for x to have α bits of information about y, it needs to have α bits of information to start with, regardless of y). The lower bound that we establish holds for any string having Kolmogorov complexity $\succeq \alpha$.[1] For such strings x, we show that $|A_{x,\alpha}| \geq (1/\mathrm{poly}(n))2^{n-\alpha}$. A related set is $B_{x,\alpha}$ consisting of the n-bit strings y with the property $C(y \mid n) - C(y \mid x) \geq \alpha$. This is the set of n-bit strings about which x has α bits of information besides the length. Note that $B_{x,\alpha} \subseteq A_{x,\alpha}$. The same observations regarding an upper bound for $|B_{x,\alpha}|$ and the emptiness of $B_{x,\alpha}$ in case $C(x) \preceq \alpha$ remain valid. For x with $C(x) \succeq \alpha$ and α computable from n, we show the lower bound $|B_{x,\alpha}| \geq (1/C) \cdot 2^{n-\alpha} - \mathrm{poly}(n) \cdot 2^{\alpha}$, for some positive constant C.

We turn to the Ramsey-type results announced above. A set of n-bit strings $\{x_1, \ldots, x_t\}$ is pairwise α-independent if for all $i \neq j$, $C(x_i) - C(x_i \mid x_j) \leq \alpha$. Intuitively, this means that any two strings in the set have in common at most α bits of information. For the notion of mutual independence we propose the following definition (but other variants are conceivable). The tuple of n-bit strings $(x_1, \ldots, x_t) \in (\{0,1\}^n)^t$ is mutually α-independent if $C(x_{\pi(1)} \ldots x_{\pi(t)}) \geq C(x_1) + \ldots + C(x_t) - \alpha$, for every permutation π of $[t]$. Intuitively this means that $x_1, \ldots, x_t$ share at most α bits of information. We show that if $\{x_1, \ldots, x_t\}$ is pairwise α-independent or if $(x_1, \ldots, x_t)$ is mutually α-independent then $t \leq \mathrm{poly}(n)2^{\alpha}$. The bound in the pairwise independent case is tight within a polynomial factor.

We also show that there exists a set B of size $\mathrm{poly}(n)2^{\alpha}$ that "α-covers" the entire set of n-bit strings, in the sense that for each n-bit string y there exists a string x in B that has α bits of information about y (i.e., y is in $A_{x,\alpha}$).

The main technical novelty of this paper is the technique used to lower bound the size of $B_{x,\alpha} = \{y \in \{0,1\}^n \mid C(y \mid n) - C(y \mid x) \geq \alpha\}$, which should be contrasted with a known and simple approach. This "normal" and simple approach is best illustrated when x is random. In this case, the prefix $x(1 : \alpha)$ of x of length α is also random and, therefore, if we take z to be an $(n - \alpha)$ long string that is random conditioned by $x(1 : \alpha)$, then $C(zx(1 : \alpha)) = n - O(\log n)$, $C(zx(1 : \alpha) \mid x(1 : \alpha)) = n - \alpha - O(\log n)$, and thus, $zx(1 : \alpha) \in B_{x,\alpha+O(\log n)}$. There are approximately $2^{n-\alpha}$ strings z as above, and this leads to a lower bound of $2^{n-\alpha}$ for $|B_{x,\alpha+O(\log n)}|$, which implies a lower bound of $(1/\mathrm{poly}(n))2^{n-\alpha}$ for $|B_{x,\alpha}|$. This method is so basic and natural that it looks hard to beat. However, using properties of Kolmogorov complexity extractors, we derive a better lower bound for $|B_{x,\alpha}|$ that does not have the slack of $1/\mathrm{poly}(n)$, in case α is computable from n (even if α is not computable from n, the new method gives a tighter estimation than the above "normal" method). A Kolmogorov complexity extractor is a function that starting with several strings that have Kolmogorov complexity relatively small compared to their lengths, computes a string that has Kolmogorov complexity almost close to its length. A related notion, namely multi-source randomness extractors, has been studied extensively

---

[1] We use notation $\mathrm{poly}(n)$ for $n^{O(1)}$ and $\approx$, $\preceq$ and $\succeq$ to denote that the respective equality or inequality holds with an error of at most $O(\log n)$.

in computational complexity (see[3,1,2,12,11]). Hitchcock, Pavan and Vinodchandran [8] have shown that Kolmogorov complexity extractors are equivalent to a type of functions that are close to being multisource randomness extractors. Fortnow, Hitchcock, Pavan, Vinodchandran and Wang [7] have constructed a polynomial-time Kolmogorov complexity extractor based on the multi-source randomness constractor of Barak, Impagliazzo and Wigderson [1]. The author has constructed Kolmogorov complexity extractors for other settings, such as extracting from infinite binary sequences [18,16] or from binary strings that have a bounded degree of dependence [16,19,17]. The latter type of Kolmogorov complexity extractors is relevant for this paper. Here we modify slightly an extractor E from [17], which, on inputs two n-bit strings x and y that have Kolmogorov complexity at least s and dependency at most α, constructs an m-bit string z with $m \approx s$ and Kolmogorov complexity equal to $m - \alpha - O(1)$ even conditioned by any one of the input strings. Let us call a pair of strings x and y with the above properties as *good-for-extraction*. We fix $x \in \{0,1\}^n$ with $C(x) \geq s$. Let z be the most popular image of the function E restricted to $\{x\} \times \{0,1\}^n$. Because it is distinguishable from all other strings, given x, z can be described with only $O(1)$ bits (we only need a description of the function E and of the input length). Choosing m just slightly larger than α we arrange that $C(z \mid x) < m - \alpha - O(1)$. This implies that all the preimages of z under E restricted as above are *bad-for-extraction*. Since the size of $E^{-1}(z) \cap (\{x\} \times \{0,1\}^n)$ is at least 2^{n-m}, we see that at least 2^{n-m} pairs (x,y) are bad-for-extraction. A pair of strings (x,y) is bad-for-extraction if either y has Kolmogorov complexity below s (and it is easy to find an upper bound on the number of such strings), or if $y \in B_{x,\alpha}$. This allows us to find the lower bound for the size of $B_{x,\alpha}$.

2 Preliminaries

We work over the binary alphabet $\{0,1\}$; $\mathbb{N}$ is the set of natural numbers. A string x is an element of $\{0,1\}^*$; $|x|$ denotes its length; $\{0,1\}^n$ denotes the set of strings of length n; $|A|$ denotes the cardinality of a finite set A; for $n \in \mathbb{N}$, $[n]$ denotes the set $\{1,2,\ldots,n\}$. We recall the basics of (plain) Kolmogorov complexity (for an extensive coverage, the reader should consult one of the monographs by Calude [4], Li and Vitányi [10], or Downey and Hirschfeldt [6]; for a good and concise introduction, see Shen's lecture notes [13]). Let M be a standard Turing machine. For any string x, define the *(plain) Kolmogorov complexity* of x with respect to M, as

$$C_M(x) = \min\{|p| \mid M(p) = x\}.$$

There is a universal Turing machine U such that for every machine M there is a constant c such that for all x,

$$C_U(x) \leq C_M(x) + c. \tag{1}$$

We fix such a universal machine U and dropping the subscript, we let $C(x)$ denote the Kolmogorov complexity of x with respect to U. We also use the concept of conditional Kolmogorov complexity. Here the underlying machine is a Turing machine that in addition to the read/work tape which in the initial state contains the input p, has a second tape containing initially a string y, which is called the conditioning information. Given such a

machine M, we define the Kolmogorov complexity of x conditioned by y with respect to M as

$$C_M(x \mid y) = \min\{|p| \mid M(p, y) = x\}.$$

Similarly to the above, there exist universal machines of this type and they satisfy the relation similar to Equation 1, but for conditional complexity. We fix such a universal machine U, and dropping the subscript U, we let $C(x \mid y)$ denote the Kolmogorov complexity of x conditioned by y with respect to U.

There exists a constant c_U such that for all strings x, $C(x) \leq |x| + c_U$. Strings $x_1, x_2, \ldots, x_k$ can be encoded in a self-delimiting way (i.e., an encoding from which each string can be retrieved) using $|x_1| + |x_2| + \ldots + |x_k| + 2\log|x_2| + \ldots + 2\log|x_k| + O(k)$ bits. For example, x_1 and x_2 can be encoded as $\overline{(bin(|x_2|)}01x_1x_2$, where $bin(n)$ is the binary encoding of the natural number n and, for a string $u = u_1 \ldots u_m$, $\overline{u}$ is the string $u_1 u_1 \ldots u_m u_m$ (i.e., the string u with its bits doubled).

Given a string x and its Kolmogorov complexity $C(x)$, one can effectively enumerate all descriptions y of x of length $C(x)$, i.e., the set $\{y \in \{0,1\}^{C(x)} \mid U(y) = x\}$. We denote x^* the first string in this enumeration. Note that $C(x) - O(1) \leq C(x^*) \leq |x^*| + O(1) = C(x) + O(1)$.

The Symmetry of Information Theorem [20] states that for any two strings x and y,

(a) $C(xy) \leq C(y) + C(x \mid y) + 2\log C(y) + O(1)$.

(b) $C(xy) \geq C(x) + C(y \mid x) - 2\log C(xy) - 4\log\log C(xy) - O(1)$.

(c) If $|x| = |y| = n$, $C(y) - C(y \mid x) \geq C(x) - C(x \mid y) - 5\log n$

Since the theorem is usually stated in a slightly different form and since we use the constants specified above, we present in the appendix the proof (which follows the standard method).

As discussed in the Introduction, our main focus is on sets of strings having certain dependency or independency properties. For convenience, we restate here the main definitions.

Definition 1. *The string y has α-dependency (where $\alpha \in \mathbb{N}$) with the string x if $C(y) - C(y \mid x) \geq \alpha$ or if x coincides with y.*

We have included the case "x coincides with y" to make a string dependent with itself even in case it has low Kolmogorov complexity.

Definition 2. *The strings $x_1, \ldots, x_t$ are pairwise α-independent if for all $i \neq j$, $C(x_i) - C(x_i \mid x_j) \leq \alpha$.*

Definition 3. *The tuple of strings $(x_1, \ldots, x_t)$ is mutually α-independent (where $\alpha \in \mathbb{N}$) if $C(x_{\pi(1)} x_{\pi(2)} \ldots x_{\pi(t)}) \geq C(x_1) + C(x_2) + \ldots + C(x_t) - \alpha$, for every permutation π of $[t]$.*

3 Strings dependent with a given string

Given a string $x \in \{0,1\}^n$, and $\alpha \in \mathbb{N}$, how many strings have dependency with x at least α? That is we are interested in estimating the size of the set

$$A_{x,\alpha} = \{y \in \{0,1\}^n \mid C(y) - C(y \mid x) \geq \alpha\}.$$

4

This is the set of strings about which, roughly speaking, x has at least α bits of information. A related set is

$$B_{x,\alpha} = \{y \in \{0,1\}^n \mid C(y \mid n) - C(y \mid x) \geq \alpha\},$$

consisting of the n-bit strings about which x provides α bits of information besides the length n. Clearly, $B_{x,\alpha} \subseteq A_{x,\alpha}$, and thus an upper bound for $|A_{x,\alpha}|$ also holds for $|B_{x,\alpha}|$, and a lower bound for $|B_{x,\alpha}|$ also holds for $|A_{x,\alpha}|$.

We show that for some polynomial p and for some constant C, for all x and α except some special values,

$$(1/p(n)) \cdot 2^{n-\alpha} \leq |A_{x,\alpha}| \leq C2^{n-\alpha},$$

and, in case $\alpha(n)$ is computable from n,

$$(1/C) \cdot 2^{n-\alpha} - p(n)2^\alpha \leq |B_{x,\alpha}| \leq C2^{n-\alpha},$$

The upper bounds for the sizes of $A_{x,\alpha}$ and $B_{x,\alpha}$ can be readily derived. Observe that the set $A_{x,\alpha}$ is included in $\{y \in \{0,1\}^n \mid C(y \mid x) < n-\alpha+c\}$ for some constant c, and therefore

$$|A_{x,\alpha}| \leq C \cdot 2^{n-\alpha},$$

for $C = 2^c$.

We move to finding a lower bound for the size of $A_{x,\alpha}$. A first observation is that for $A_{x,\alpha}$ to be non-empty, it is needed that $C(x) \succeq \alpha$. Indeed, it is immediate to observe that for any strings x and y of length n,

$$C(y) \leq C(x) + C(y \mid x) + 2\log C(x) + O(1) \leq C(x) + C(y \mid x) + 2\log n + O(1),$$

and thus, if $C(y) - C(y \mid x) \geq \alpha$, then $C(x) \geq \alpha - 2\log n - O(1)$. Intuitively, if the information in x is close to α, not too many strings can be α-dependent with it.

We provide a lower bound for $|A_{x,\alpha}|$, for every string x with $C(x) \geq \alpha + 7\log n$. The proof uses the basic "normal" approach presented in the Introduction. To simplify the discussion, suppose $C(x) = \alpha$. Then if we take a string z of length $n - \alpha$ that is random conditioned by x^*, it holds that $C(x^*z) \approx n$ and $C(x^*z \mid x^*) \approx n-\alpha$. Thus, $C(x^*z) - C(x^*z \mid x^*) \succeq \alpha$. Note that there are approximately $2^{n-\alpha}$ such strings x^*z. Since x^* can be obtained from x and $C(x)$, we can replace x^* by x in the conditioning at a small price. We obtain approximately $2^{n-\alpha}$ strings in $A_{x,\alpha}$.

Theorem 1. *For every natural number n, for every natural number α and for every $x \in \{0,1\}^n$ such that $C(x) \geq \alpha + 7\log n$,*

$$|A_{x,\alpha}| \geq \frac{1}{2n^7} 2^{n-\alpha},$$

provided n is large enough.

Proof. Let $k = C(x)$ and let $\beta = \alpha + 7\log n$. Let x^* be the smallest description of x as described in the Preliminaries. Let x^*_β be the prefix of x^* of length β. Since x^* is described

5

by x_β^* and by its suffix of length $k - \beta$, $C(x^*) \leq C(x_\beta^*) + (k - \beta) + 2\log C(x_\beta^*) + O(1)$ and, thus

$$C(x_\beta^*) \geq C(x^*) - (k - \beta) - 2\log C(x_\beta^*) - O(1)$$
$$\geq (k - O(1)) - (k - \beta) - 2\log C(x_\beta^*) - O(1)$$
$$\geq \beta - 2\log \beta - O(1).$$

The set $B = \{z \in \{0,1\}^{n-\beta} \mid C(z \mid x_\beta^*) \geq n - \beta - 1\}$ has size at least $(1/2) \cdot 2^{n-\beta}$ (using a standard counting argument). Consider a string $y \in \{0,1\}^n$ of the from $y = x_\beta^* z$ with $z \in B$. There are at least $(1/2) \cdot 2^{n-\beta}$ such strings.

By symmetry of information,

$$C(y) = C(x_\beta^* z) \geq C(x_\beta^*) + C(z \mid x_\beta^*) - (2\log n + 4\log\log n + O(1))$$
$$\geq (\beta - 2\log\beta) + (n - \beta - 1) - (2\log n + 4\log\log n + O(1))$$
$$\geq n - (4\log n + 4\log\log n + O(1)) \geq n - 5\log n.$$

On the other hand, $C(y \mid x_\beta^*) = C(x_\beta^* z \mid x_\beta^*) \leq C(z) + O(1) \leq (n - \beta) + O(1)$. Note that

$$C(y \mid x) \leq C(y \mid x_\beta^*) + 2\log n + 4\log\log n + O(1),$$

because one can effectively construct x_β^* from x, k and β. Therefore,

$$C(y \mid x) \leq (n - \beta) + 2\log n + 4\log\log n + O(1),$$

and thus

$$C(y) - C(y \mid x) \geq \beta - (6\log n + 8\log\log n + O(1)) \geq \beta - 7\log n.$$

So, $y \in A_{x,\beta-7\log n} = A_{x,\alpha}$. Since this holds for all the strings y mentioned above, it follows that $|A_{x,\alpha}| \geq (1/2)2^{n-\beta} = (1/(2n^7)) \cdot 2^{n-\alpha}$. ∎

The lower bound for $|B_{x,\alpha}|$ is obtained using a technique based on Kolmogorov complexity extractors, as explained in the Introduction. We use the following theorem which can be obtained by a simple modification of a result from [17].

Theorem 2. *For any computable functions $s(n), m(n)$ and $\alpha(n)$ with $n \geq s(n) \geq \alpha(n) + 7\log n$ and $m(n) \leq s(n) - 7\log n$, there exists a computable ensemble of functions $E : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^{m(n)}$ such that for all x and y in $\{0,1\}^n$*

- *if $C(x) \geq s(n), C(y \mid n) \geq s(n)$ and $C(y \mid n) - C(y \mid x) \leq \alpha(n)$*
- *then $C(E(x,y) \mid x) \geq m(n) - \alpha(n) - O(1)$.*

Theorem 3. *Let $\alpha(n)$ be a computable function. For every sufficiently large natural number n, for every $x \in \{0,1\}^n$ such that $C(x) \geq \alpha(n) + 8\log n$,*

$$|B_{x,\alpha(n)}| \geq \frac{1}{C} \cdot 2^{n-\alpha(n)} - n^8 2^{\alpha(n)},$$

for some positive constant C.

Proof. Let $m = \alpha(n) + c$ and $s = \alpha(n) + 8 \log n$, where c is a constant that will be specified later. Consider $E : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^m$ the Kolmogorov extractor given by Theorem 2 for these parameters. Let $z \in \{0,1\}^m$ be the string that has the largest number of E preimages in the set $\{x\} \times \{0,1\}^n$. Note that, for some constant c_1, $C(z \mid x) \leq c_1$, because, given x, z can be constructed from a table of E, which at its turn can be constructed from n which is given because it is the length of x. On the other hand, if $y \in \{0,1\}^n$ is a string with $C(y \mid n) \geq s$ and $C(y \mid n) - C(y \mid x) \leq \alpha(n)$, then Theorem 2 guarantees that, for some constant c_2, $C(E(x,y) \mid x) \geq m - \alpha(n) - c_2 = c - c_2 > c_1$, for an appropriate c. Therefore all the strings y such that $E(x,y) = z$ are bad for extraction, i.e., they belong to

$$\{y \in \{0,1\}^n \mid C(y \mid n) < s\} \cup \{y \in \{0,1\}^n \mid C(y \mid n) \geq s \text{ and } C(y \mid n) - C(y \mid x) \geq \alpha\}.$$

Since there are at least 2^{n-m} such strings y and the first set above has less than 2^s elements, it follows that

$$|\{y \in \{0,1\}^n \mid C(y \mid n) - C(y \mid x) \geq \alpha(n)\}| \geq 2^{n-m} - 2^s = \frac{1}{2^c} \cdot 2^{n - \alpha(n)} - n^8 2^{\alpha(n)}.$$

This concludes the proof. ∎

The proof of Theorem 1 actually shows more: The lower bound applies even to a subset of $A_{x,\alpha}$ containing only strings with high Kolmogorov complexity. More precisely, if we denote $A_{x,\alpha,s} = \{y \in \{0,1\}^n \mid C(y) \geq s \text{ and } C(y) - C(y \mid x) \geq \alpha\}$, then $|A_{x,\alpha,n-5 \log n}| \geq \frac{1}{2n^7} 2^{n-\alpha}$. Note that there is an interesting "zone" for the parameter s that is not covered by this result. Specifically, it would be interesting to lower bound the size of $A_{x,\alpha,n}$. This question remains open. Nevertheless, the technique from Theorem 3 can be used to tackle the variant in which access to the set $R = \{u \in \{0,1\}^n \mid C(u) \geq |u|\}$ is granted for free. Thus, let $A^R_{x,\alpha,n} = \{y \in \{0,1\}^n \mid C^R(y) \geq n \text{ and } C^R(y) - C^R(y \mid x) \geq \alpha\}$.

Proposition 1. *For the same setting of parameters as in Theorem 3, $|A^R_{x,\alpha,n}| \geq \frac{1}{C} \cdot 2^{n-\alpha(n)}$, for some positive constant C.*

Proof. Omitted from this extended abstract. ∎

4 Pairwise independent strings

We show that if the n-bit strings $x_1, \ldots, x_t$ are pairwise α-independent, then $t \leq \text{poly}(n)2^\alpha$. This upper bound is relatively tight, since there are sets with $(1/\text{poly}(n)) \cdot 2^\alpha$ n-bit strings that are pairwise α-independent.

Theorem 4. *For every sufficiently large n and for every natural number α, the following holds. If $x_1, \ldots, x_t$ are n-bit strings that are α-independent, then $t < 2n^3 \cdot 2^\alpha$.*

Proof. There are less than $2^{\alpha+3 \log n}$ strings with Kolmogorov complexity less than $\alpha + 3 \log n$. We discard such strings from $x_1, \ldots, x_t$ and assume that $x_1, \ldots, x_{t'}$ are the strings that are left. Since $t < 2^{\alpha+3 \log n} + t'$, we need to show that $t' \leq n^3 2^\alpha$.

For $1 \leq i \leq t'$, let $k_i = C(x_i)$ and let x_i^* be the shortest description of x_i as described in the Preliminaries. Let $\beta = \alpha + 3 \log n$ (we assume that $\alpha \leq n - 3 \log n$, as otherwise the

statement is trivial). We show that the prefixes of length β of the strings $x_1, \ldots, x_{t'}$ are all distinct, from which we conclude that $t' \leq 2^\beta = n^3 \cdot 2^\alpha$.

Suppose that there are two strings in the set that have equal prefixes of length β. W.l.o.g. we can assume that they are x_1 and x_2. Then

$$C(x_1^* \mid x_2^*) \leq (k_1 - \beta) + \log \beta + 2 \log \log \beta + O(1),$$

because, given x_2^*, x_1^* can be constructed from β and the suffix of length $k_1 - \beta$ of x_1^*. Note that

$$C(x_1^* \mid x_2) \leq C(x_1^* \mid x_2^*) + \log k_2 + 2 \log \log k_2 + O(1),$$

because x_2^* can be constructed from x_2 and k_2. Also note that $C(x_1 \mid x_2) \leq C(x_1^* \mid x_2) + O(1)$. Thus,

$$C(x_1 \mid x_2) \leq C(x_1^* \mid x_2^*) + \log k_2 + 2 \log \log k_2 + O(1).$$

Therefore,

$$
\begin{aligned}
C(x_1) - C(x_1 \mid x_2) &\geq k_1 - (C(x_1^* \mid x_2^*) + \log k_2 + 2 \log \log k_2 + O(1)) \\
&\geq k_1 - (k_1 - \beta) - \log \beta - 2 \log \log \beta - \log k_2 - 2 \log \log k_2 - O(1) \\
&\geq \beta - 3 \log n = \alpha,
\end{aligned}
$$

which is a contradiction. ∎

The next result shows that the upper bound in Theorem 4 is relatively tight. It relies on the well-known Turán's Theorem in Graph Theory [14], in the form due to Caro (unpublished) and Wei [15] (see [9, page 248]): Let G be a graph with n vertices and let d_i be the degree of the i-th vertex. Then G contains an independent set of size at least $\sum \frac{1}{d_i+1}$.

Theorem 5. *For every natural number n and for every natural number α satisfying $5 \log n \leq \alpha \leq n$, there exists a constant C and $t = \frac{1}{Cn^5} \cdot 2^\alpha$ n-bit strings $x_1, \ldots, x_t$ that are pairwise α-independent.*

Proof. Let $\beta = \alpha - 5 \log n$. Consider the graph $G = (V, E)$, where $V = \{0,1\}^n$ and $(u,v) \in E$ iff $C(u) - C(u \mid v) \geq \beta$ and $C(v) - C(v \mid u) \geq \beta$. Note that for every $u \in \{0,1\}^n$, the degree of u is bounded by $|A_{u,\beta}| \leq 2^{n-\beta+c}$, for some constant c. Therefore, by Turán's theorem, the graph G contains an independent set I of size at least $2^n \cdot \frac{1}{2^{n-\beta+c}+1} \geq 2^{\beta-c-1} = \frac{1}{Cn^5} \cdot 2^\alpha$. For any two elements u, v in I, we have either $C(u) - C(u \mid v) < \beta$ or $C(v) - C(v \mid u) < \beta$. In the second case, by symmetry of information, $C(u) - C(u \mid v) < \beta + 5 \log n = \alpha$. It follows that the strings in I are pairwise α-independent. ∎

5 Mutually independent strings

In this section we show that the size of a mutually α-independent tuple of n-bit strings is bounded by $\mathrm{poly}(n)2^\alpha$.

For $u \in \{0,1\}^n$, we define $D_\alpha(u) = \{x \in \{0,1\}^n \mid u \in A_{x,\alpha}\} = \{x \in \{0,1\}^n \mid C(u) - C(u \mid x) \geq \alpha\}$ and $d_\alpha(u) = |D_\alpha(u)|$.

Lemma 1. *For every natural number n sufficiently large, for every natural number α, and for every $u \in \{0,1\}^n$, with $C(u) \geq \alpha + 12\log n$,*

$$\frac{1}{2n^{12}}2^{n-\alpha} \leq d_\alpha(u) \leq n^5 \cdot 2^{n-\alpha}.$$

Proof. For every $x \in A_{u,\alpha+5\log n}$,

$$C(x) - C(x \mid u) \geq \alpha + 5\log n$$

which by symmetry of information implies

$$C(u) - C(u \mid x) \geq \alpha + 5\log n - 5\log n = \alpha,$$

and therefore, $u \in A_{x,\alpha}$. Thus

$$d_\alpha(u) \geq |A_{u,\alpha+5\log n}| \geq \frac{1}{2n^7}2^{n-\alpha-5\log n} = \frac{1}{2n^{12}}2^{n-\alpha}.$$

For every $u \in \{0,1\}^n$,

$$
\begin{aligned}
x \in D_{u,\alpha} &\Rightarrow u \in A_{x,\alpha}\\
&\Rightarrow C(u) - C(u \mid x) \geq \alpha\\
&\Rightarrow C(x) - C(x \mid u) \geq \alpha - 5\log n\\
&\Rightarrow C(x \mid u) \leq n - \alpha + 5\log n.
\end{aligned}
$$

Thus, $d_\alpha(u) \leq |\{x \in \{0,1\}^n \mid C(x \mid u) \leq n - \alpha + 5\log n\}| \leq n^5 \cdot 2^{n-\alpha}$. ∎

Since for any string x and natural number α, $|A_{x,\alpha}| \leq 2^{n-\alpha-c}$, for some constant c, it follows that we need at least $T = 2^{\alpha-c}$ strings $x_1, \ldots, x_T$ to "α-cover" the set of n-bit strings, in the sense that for each n-bit string y, there exists x_i, $i \in [T]$ such that y is α-dependent with x_i. The next theorem shows that $\text{poly}(n)2^\alpha$ strings are enough to α-cover the set of n-bit strings.

Theorem 6. *For every natural number n sufficiently large, for every natural number α, there exists a set $B \subseteq \{0,1\}^n$ of size $\text{poly}(n)2^\alpha$ such that each string in $\{0,1\}^n$ is α-dependent with some string in B, i.e., $\{0,1\}^n = \bigcup_{x \in B} A_{x,\alpha}$. More precisely the size of B is bounded by $(2n^{13} + n^{12}) \cdot 2^\alpha$.*

Proof. (a) We choose $T = 2n^{13}2^\alpha$ strings $x_1, \ldots, x_T$, uniformly at random in $\{0,1\}^n$. The probability that a fix u with $C(u) \geq \alpha + 12\log n$ does not belong to any of the sets $A_{x_i,\alpha}$, for $i \in [T]$, is at most $(1 - \frac{1}{2n^{12}2^\alpha})^T < e^{-n}$ (by Lemma 1). By the union bound, the probability that there exists $u \in \{0,1\}^n$ with $C(u) \geq \alpha + 12\log n$, that does not belong to any of the sets $A_{x_i,\alpha}$, for $i \in [T]$, is bounded by $2^n \cdot e^{-n} < 1$. Therefore there are strings $x_1, \ldots, x_T$ in $\{0,1\}^n$ such that $\bigcup A_{x_i,\alpha}$ contains all the strings $u \in \{0,1\}^n$ having $C(u) \geq \alpha + 12\log n$. By adding to $x_1, \ldots, x_T$, the strings that have Kolmogorov complexity $< \alpha + 12\log n$, we obtain the set B that α-covers the entire $\{0,1\}^n$. ∎

To estimate the size of a mutually α-independent tuple of strings, we need the following lemma.

9

Lemma 2. *Let $\alpha, \beta \in \mathbb{N}$ and let the tuple of n-bit strings $(x_1, x_2, \ldots, x_k)$ satisfy $C(x_1 \ldots x_k) \geq C(x_1) + \ldots + C(x_k) - \beta$. Then there exists a constant d such that*

$$|A_{x_1,\alpha} \cap \ldots \cap A_{x_k,\alpha}| \leq dn^{7k+5}k^3 2^{n-k\alpha+\beta}.$$

Proof. Let $u \in \{0,1\}^n$ be a string in $A_{x_1,\alpha} \cap \ldots \cap A_{x_k,\alpha}$. Then $C(u) - C(u \mid x_i) \geq \alpha$, for all $i \in [k]$. Therefore, by symmetry of information, $C(x_i) - C(x_i \mid u) \geq \alpha - 5 \log n$, for all $i \in [k]$. It follows that for every $i \in [k]$, there exists a string p_i of length $|p_i| \leq C(x_i) - \alpha + 5 \log n$ such that, given u, is a descriptor of x_i (i.e., $U(p_i, u) = x_i$). The strings $p_1, \ldots, p_k$ describe the string $x_1 x_2 \ldots x_k$, given u, and therefore

$$\begin{aligned} C(x_1 x_2 \ldots x_k \mid u) &\leq |p_1| + \ldots + |p_k| + 2 \log |p_1| + \ldots + 2 \log |p_k| + O(1) \\ &\leq C(x_1) + \ldots + C(x_k) - k\alpha + 5k \log n + 2 \log |p_1| + \ldots + 2 \log |p_k| + O(1) \\ &\leq C(x_1) + \ldots + C(x_k) - k\alpha + 7k \log n + O(1) \\ &\leq C(x_1 \ldots x_k) + \beta - k\alpha + 7k \log n + O(1). \end{aligned}$$

So,

$$C(x_1 \ldots x_k) - C(x_1 \ldots x_k \mid u) \geq -(\beta - k\alpha + 7k \log n + O(1)).$$

By symmetry of information,

$$C(u) - C(u \mid x_1 \ldots x_k) \geq C(x_1 \ldots x_k) - C(x_1 \ldots x_k \mid u) - 2 \log C(u) - 2 \log C(x_1 \ldots x_k u) \\ - 4 \log \log C(x_1 \ldots x_k u) - O(1).$$

It follows that

$$C(u) - C(u \mid x_1 \ldots x_k) \geq -(\beta - k\alpha + 7k \log n) - 5 \log n - 3 \log k$$

and thus

$$\begin{aligned} C(u \mid x_1 \ldots x_k) &\leq C(u) + \beta - k\alpha + (7k+5) \log n + 3 \log k \\ &\leq n + \beta - k\alpha + (7k+5) \log n + 3 \log k + O(1). \end{aligned}$$

Therefore,

$$A_{x_1,\alpha} \cap \ldots \cap A_{x_k,\alpha} \subseteq \{u \in \{0,1\}^n \mid C(u \mid x_1 \ldots x_k) \leq n + \beta - k\alpha + (7k+5) \log n + 3 \log k + O(1)\}.$$

The conclusion follows. ∎

Finally, we prove the upper bound for the size of a mutually α-independent tuple of n-bit strings.

Theorem 7. *For every sufficiently large natural number n the following holds. Let α be an integer such that $\alpha > 7 \log n + 6$. Let $(x_1, \ldots, x_t)$ be a mutually α-independent tuple of n-bit strings. Then $t \leq \text{poly}(n)2^\alpha$.*

Proof. By Theorem 6, there exists a set B of size at most $\text{poly}(n)2^{\alpha+5 \log n}$ such that every n-bit string x is in $A_{y,\alpha+5 \log n}$, for some $y \in B$. We view $\{x_1, \ldots, x_t\}$ as a multiset. Let y be the string in B that achieves the largest size of multiset $A_{y,\alpha+5 \log n} \cap \{x_1, \ldots, x_t\}$ (we take every common element with the multiplicity in $\{x_1, \ldots, x_t\}$). Let k be the size of the above intersection. Clearly, $k \geq t/|B|$. We will show that $k = \text{poly}(n)$, and, therefore, $t \leq k \cdot |B| = \text{poly}(n) \cdot 2^\alpha$.

Without loss of generality suppose $A_{y,\alpha+5\log n} \cap \{x_1, \ldots, x_t\} = \{x_1, \ldots, x_k\}$ (as multi-sets). Since, for every $i \in [k]$, $C(x_i) - C(x_i \mid y) \geq \alpha + 5\log n$, by symmetry of information, it follows that $C(y) - C(y \mid x_i) \geq \alpha$. Thus $y \in A_{x_1,\alpha} \cap \ldots \cap A_{x_k,\alpha}$. In particular, $A_{x_1,\alpha} \cap \ldots \cap A_{x_k,\alpha}$ is not empty. We want to use Lemma 2 but before we need to estimate the difference between $C(x_1 \ldots x_k)$ and $C(x_1) + \ldots + C(x_k)$.

Claim. $C(x_1 \ldots x_k) \geq C(x_1) + \ldots + C(x_k) - \beta$, where $\beta = \alpha + 4\log(nt/2)$.

Proof of claim. Suppose $C(x_1 \ldots x_k) < C(x_1) + \ldots + C(x_k) - \beta$. Note that

$$
\begin{aligned}
C(x_1 \ldots x_t) &\leq C(x_1 \ldots x_k) + C(x_{k+1} \ldots x_t) + 2\log C(x_1 \ldots x_k) + O(1) \\
&\leq C(x_1) + \ldots C(x_k) + C(x_{k+1} \ldots x_t) - \beta + 2\log kn + O(1).
\end{aligned}
$$

Since $C(x_1 \ldots x_t) \geq C(x_1) + \ldots + C(x_t) - \alpha$, it follows that

$$
C(x_{k+1}) + \ldots + C(x_t) - \alpha \leq C(x_{k+1} \ldots x_t) - \beta + 2\log kn + O(1).
$$

On the other hand,

$$
C(x_{k+1} \ldots x_t) \leq C(x_{k+1}) + \ldots + C(x_t) + 2\log(t-k)n + O(1).
$$

It follows that
$$
\beta - \alpha \leq 2\log kn + 2\log(t-k)n + O(1).
$$

However, from the definition of β,

$$
\beta - \alpha = 4\log(nt/2) > 2\log kn + 2\log(t-k)n + O(1).
$$

The contradiction proves the claim. ∎

Now, by Lemma 2,

$$
\begin{aligned}
|A_{x_1,\alpha} \cap \ldots \cap A_{x_k,\alpha}| &\leq dn^{7k+5}k^3 2^{n-k\alpha+\beta} \\
&= dn^{7k+5}k^3 2^{n-(k-1)\alpha+4\log t+4\log(n/2)} \\
&\leq dn^{7k+5}k^3 2^{5n-(k-1)\alpha+4\log(n/2)},
\end{aligned}
$$

where in the last line we used the fact that $t \leq 2^n$.

It can be checked that if $\alpha > 7\log n + 6$ and $k \geq n$, then the above upper bound is less than 1, which is a contradiction. It follows that $k < n$. ∎

6 Final remarks

This paper provides tight bounds (within a polynomial factor) for the size of $A_{x,\alpha}$ (the set of n-bit strings that have α-dependency with x) and for the size of sets of n-bit strings that are pairwise α-independent.

The size of a mutually α-independent tuple of n-bit strings is at most $\text{poly}(n)2^\alpha$. We do not know how tight this bound is and leave this issue as an interesting open problem.

We have recently learned about the paper [5], which obtains similar results regarding the size of sets of pairwise and k-independence strings, for a notion of independence that is suitable for strings with large Kolmogorov complexity.

References

1. Barak, B., Impagliazzo, R., Wigderson, A.: Extracting randomness using few independent sources. In: Proceedings of the 36th ACM Symposium on Theory of Computing. pp. 384–393 (2004)
2. Barak, B., Kindler, G., Shaltiel, R., Sudakov, B., Wigderson, A.: Simulating independence: new constructions of condensers, ramsey graphs, dispersers, and extractors. In: Proceedings of the 37th ACM Symposium on Theory of Computing. pp. 1–10 (2005)
3. Bourgain, J.: More on the sum-product phenomenon in prime fields and its applications. International Journal of Number Theory 1, 1–32 (2005)
4. Calude, C.: Information and Randomness: An Algorithmic Perspective. Springer-Verlag (2002), 2nd edition, 1st edition in 1994
5. Chang, C., Lyuu, Y., Ti, Y., Shen, A.: Sets of k-independent sets. International Journal of Foundations of Computer Science (2009), to appear.
6. Downey, R., Hirschfeldt, D.: Algorithmic randomness and complexity. Springer Verlag (2010)
7. Fortnow, L., Hitchcock, J., Pavan, A., Vinodchandran, N., Wang, F.: Extracting Kolmogorov complexity with applications to dimension zero-one laws. In: Proceedings of the 33rd International Colloquium on Automata, Languages, and Programming. pp. 335–345. Springer-Verlag *Lecture Notes in Computer Science #4051*, Berlin (2006)
8. Hitchcock, J., Pavan, A., Vinodchandran, N.: Kolmogorov complexity in randomness extraction. Electronic Colloquium on Computational Complexity (ECCC) (09-071) (2009)
9. Jukna, S.: Extremal Combinatorics. Springer Verlag (2001)
10. Li, M., Vitanyi, P.: An introduction to Kolmogorov complexity and its applications. Springer-Verlag (2008), 3rd edition. 1st edition in 1993.
11. Rao, A.: Extractors for a constant number of polynomially small min-entropy independent sources. In: Proceedings of the 38th ACM Symposium on Theory of Computing. pp. 497–506 (2006)
12. Raz, R.: Extractors with weak random seeds. In: Gabow, H.N., Fagin, R. (eds.) STOC. pp. 11–20. ACM (2005)
13. Shen, A.: Algorithmic information theory and Kolmogorov complexity. Tech. Rep. 2000-034, Uppsala Universitet (December 2000)
14. Turán, P.: On an extremal problem in graph theory. Math.Fiz.Lapok 48, 436–452 (1941), in Hungarian
15. Wei, V.: A lower bound on the stability number of a simple graph. Tech. Rep. 81-11217-9, Bell Laboratories (1981)
16. Zimand, M.: Extracting the Kolmogorov complexity of strings and sequences from sources with limited independence. In: Proceedings 26th STACS, Freiburg, Germany (February 26–29 2009)
17. Zimand, M.: Impossibility of independence amplification in Kolmogorov complexity theory. In: MFCS (2010)
18. Zimand, M.: Two sources are better than one for increasing the Kolmogorov complexity of infinite sequences. In: Hirsch, E.A., Razborov, A.A., Semenov, A.L., Slissenko, A. (eds.) CSR. Lecture Notes in Computer Science, vol. 5010, pp. 326–338. Springer (2008)
19. Zimand, M.: On generating independent random strings. In: Ambos-Spies, K., Löwe, B., Merkle, W. (eds.) CiE. Lecture Notes in Computer Science, vol. 5635, pp. 499–508. Springer (2009)
20. Zvonkin, A., Levin, L.: 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)

A

Symmetry of Information Theorem

Theorem 8. *For any two strings x and y,*

(a) $C(xy) \leq C(y) + C(x \mid y) + 2\log C(y) + O(1)$.

(b) $C(xy) \geq C(x) + C(y \mid x) - 2\log C(xy) - 4\log\log C(xy) - O(1)$.

(c) *If $|x| = |y| = n$, $C(y) - C(y \mid x) \geq C(x) - C(x \mid y) - 5\log n$*

Proof (sketch): (a) is easy and (c) follows immediately from (a) and (b). We prove (b). Let $C(xy) = t$, $A = \{(u,v) \mid C(uv) \leq t\}$, $A_u = \{v \mid C(uv) \leq t\}$. Note that $|A| < 2^{t+1}$. Let $e = \lfloor \log |A_x| \rfloor$. Let $B = \{u \mid |A_u| \geq 2^e\}$. Note that $x \in B$ and $|B| < |A|/2^e < 2^{t-e+1}$.

FACT: x can be described by: t, rank in B (which is written on exactly $t - e + 1$ bits so that e can be also reconstructed), $O(1)$ bits. So $C(x) \leq (t - e + 1) + \log t + 2\log\log t + O(1)$.

FACT: y, given x, can be described by: t, rank in A_x, $O(1)$ bits. So, $C(y \mid x) \leq e + \log t + 2\log\log t + O(1)$.

Combining the last two: $C(x) \leq t - (C(y \mid x) - \log t - 2\log\log t - O(1)) + \log t + 2\log\log t + (1) = C(xy) - C(y \mid x) + 2\log t + 4\log\log t + O(1) = C(xy) - C(y \mid x) + 2\log C(xy) + 4\log\log C(xy) + O(1)$.