On mutual information of a binary word and its entropy

Andrej Muchnik[*]

We consider two kinds of the entropy (and the conditional entropy) of binary words, namely, the simple entropy (introduced by Kolmogorov) denoted by KS, and the prefix entropy (introduced by Levin) denoted by KP. When a statement is true for both kinds of entropy, we write K instead of KS and KP. We suppose that there are fixed isomorphisms between the set of naturals and the set of binary words, the set of pairs of binary words, and so on. All logarithms are binary. The length of a word x is denoted by $\ell(x)$.

As is known, both entropies under consideration are enumerable from above and close by the value:

$$KS(x|y) \leq KP(x|y) + O(1)\,,$$
$$KP(x|y) \leq KS(x|y) + O(\log KS(x|y))\,.$$

It is no wonder that KS and KP have many indentical properties. But in some cases they behave differently. An example of a recursion-theoretical property that substantially distinguishes KS and KP was published in [1]. In this paper we study quantitative properties, with respect to which KS and KP behave in an opposite way. We are interested in the following questions: does a word help to find its entropy and does the entropy of a word help to find this word itself? More exactly, we want to describe the behaviour of functions $I(x : K(x)) = K(K(x)) - K(K(x)|x)$ and $I(K(x) : x) = K(x) - K(x|K(x))$ (we write IS instead of I, when $K = KS$, and IP, when $K = KP$). By Kolmogorov–Levin theorem about the symmetry of mutual information $K(x) - K(x|y) \approx K(y) - K(y|x) \approx K(x) + K(y) - K(\langle x, y\rangle)$. If $y = K(x)$,

[*] Institute of New Technologies, 10, Nizhnyaya Radishevskaya, Moscow, 109004, Russia. E-mail: muchnik@lpcs.math.msu.ru, fax: (095)9156963. The work was partially supported by the Russian Foundation for Basic Research grants 01-01-00505 and 02-01-10904.

then (as Gács showed)

$$K(\langle x, y\rangle) = K(\langle x, K(x)\rangle) = K(x) + O(1)$$

(a minimal code of a word x gives x itself, and code's length gives $K(x)$). But mutual information is symmetrical only up to the term $O\big(\log K(x) + \log K(y)\big)$ (for KS it was proved by Kolmogorov and Levin, and for KP by Gács). Since $K(K(x)) \leq O(\log K(x))$, the logarithmic accuracy is not sufficient.

Our main results that proved in Theorems 1–4 are represented in the following table.

	KS	KP
$I(K(x):x)$	Mutual information is bounded	Mutual information tends to ∞
$I(x:K(x))$	Mutual information is not bounded, but does not tend to ∞	Mutual information tends to ∞

We see that for KP knowing a word helps to find its entropy and knowing the entropy helps to find the word. For KS knowing the entropy does not help to find the word itself and knowing a word sometimes helps to find its entropy and sometimes does not.

Theorem 1. *The function $IS(KS(x):x)$ is bounded.*

Proof. Let us consider a large enough natural number D. Assume that $KS(x|KS(x)) < KS(x) - D$. Let y be a minimal code of x when $KS(x)$ is known. Denote by n a special binary representation of the number $N = KS(x) - \ell(y)$, in which each digit is repeated two times. Given the word $n01y$ one can effectively find x. Indeed, the first from the left occurence of 01 separate y from n; given n one finds $KS(x) - \ell(y)$; knowing y, one finds $KS(x)$; given y and $KS(x)$ one finds x. Therefore $KS(x) \leq \ell(n01y) + C \leq 2\log N + 4 + \ell(y) + C$, where C depends on the choice of an optimal programming language only. That is, $N = KS(x) - \ell(y) \leq 2\log N + C + 4$. Since $N > D$, for large D we get a contradiction. Thus for large enough D we have $KS(x) \leq KS(x|KS(x)) + D$. □

Perhaps, the reader is surprised why the same reasoning cannot be applied to the function $IP(KP(x):x)$. In this case y is a minimal *prefix* code of x

when $KP(x)$ is known, $n01$ is a *prefix* code of N. It seems that $n01y$ is a prefix code of x, and we get a similar result. But actually y is a prefix code of x for a *fixed* condition only. The set of all codes that correspond to the function $KP(\cdot|\cdot)$ for all conditions is not a prefix set.

Theorem 2. *The function $IP(KP(x):x)$ tends to infinity.*

Proof. As is known, the enumerable from below function $\mu(x|y)=2^{-KP(x|y)}$ for every y specifies a semimeasure on the domain of x and it is the greatest one among such semimeasures up to a multiplicative constant. And respectively $\mu(x)=2^{-KP(x)}$. By μ^t denote the result of a step t in the enumeration from below of the function μ. For every t, the domain where $\mu^t\neq 0$ is finite.

Let us define a computable function $\nu_k^t(x|n)$ of natural arguments x, n, t and $k>0$ with rational values. For $t=0$, the function ν is identically equal to zero. While t is increasing, the function is not decreasing, and for all t, n, k it holds $\sum\limits_x \nu_k^t(x|n)\leq 2^{-k}$.

Put $\nu_k^{t+1}(x|n) = 2^k\cdot\min\{\mu^t(x),2^{-n}\}$ if $\sum\limits_x \min\{\mu^t(x),2^{-n}\}\leq 2^{-2k}$, and $\nu_k^{t+1}(x|n)=\nu_k^t(x|n)$ otherwise.

The function $\nu(x|n) = \sum\limits_k \max\limits_t \nu_k^t(x|n)$ is enumerable from below and $\sum\limits_x \nu(x|n)\leq 1$. Therefore $\nu(x|n)\leq O(\mu(x|n))$.

For every k, there exists x_0 such that $\sum\limits_{x\geq x_0}\mu(x)\leq 2^{-2k-1}$, and there exists n_0 such that for all $n\geq n_0$ it holds $\sum\limits_{x<x_0} 2^{-n}\leq 2^{-2k-1}$. Hence for $n\geq n_0$ we get $\sum\limits_x \min\{\mu^t(x),2^{-n}\}\leq 2^{-2k}$, and therefore $\nu(x|n)\geq 2^k\cdot\min\{\mu(x),2^{-n}\}$. For all large enough x the value of $\mu(x)$ is not greater than 2^{-n_0}; it implies that for $n=-\log\mu(x)=KP(x)$, we have $\nu(x|n)\geq 2^k\cdot\mu(x)$. Since $\mu(x|n)\geq 2^{-C}\cdot\nu(x|n)\geq 2^{k-C}\cdot\mu(x)$, we obtain $KP(x|KP(x))\leq KP(x)-k+C$, where C depends on the choice of an optimal programming language only. □

Theorem 3. *The function $IP(x:KP(x))$ tends to infinity.*

Proof. The proof of this theorem is similar to the proof of the previous one.

Let us define a computable function $\nu_k^t(n|x)$ of natural arguments n, x, t and $k>0$ with rational values. For $t=0$, the function ν is identically equal to zero. While t is increasing, the function is not decreasing, and for all t, n, k it holds $\sum\limits_n \nu_k^t(n|x)\leq 2^{-k}$.

To define $\nu_k^{t+1}(n|x)$ we use an auxiliary function

$$\alpha_k(n|x) = \begin{cases} 2^k \cdot \mu^t(n) & \text{if } \mu^t(x) = 2^{-n}, \\ \nu_k^t(n|x) & \text{otherwise.} \end{cases}$$

If $\sum\limits_n \alpha_k(n|x) \leq 2^{-k}$, put $\nu_k^{t+1}(n|x) = \alpha_k(n|x)$, and $\nu_k^{t+1}(n|x) = \nu_k^t(n|x)$ in the converse case.

The function $\nu(n|x) = \sum\limits_k \max\limits_t \nu_k^t(n|x)$ is enumerable from below and $\sum\limits_n \nu(n|x) \leq 1$. Therefore $\nu(n|x) \leq O(\mu(n|x))$.

For every k, there exists n_0 such that $\sum\limits_{n \geq n_0} 2^k \mu(n) \leq 2^{-k}$, and there exists x_0 such that for all $x \geq x_0$ it holds $\mu(x) \leq 2^{-n_0}$. Hence for $x \geq x_0$ we get $\nu(n|x) = 0$ if $n < n_0$, and the inequality $\sum\limits_n \alpha_k(n|x) \leq 2^{-k}$ from the definition of ν holds for all t. Thus for $n = -\log \mu(x) = KP(x)$ we have $\nu(n|x) \geq 2^k \cdot \mu(n)$. Since $\mu(n|x) \geq 2^{-C} \cdot \nu(n|x) \geq 2^{k-C} \cdot \mu(n)$, we obtain $KP(KP(x)|x) \leq KP(x) - k + C$, where C depends on the choice of an optimal programming language only. □

Theorem 4. *The function $IS(x : KS(x))$ is not bounded, but does not tend to infinity.*

Proof. The unboundedness of the function $IS(x : KS(x))$ is almost obvious. Let x be a random word of length n (that is $KS(x) = n + O(1)$). Then $KS(KS(x)|x) = O(1)$, and on the other hand $KS(KS(x)) \to \infty$.

Now our aim is to prove that the function $IS(x : KS(x))$ is bounded on an infinite set. For any word x of length n, the simple entropy of x is not greater than $n + O(1)$, therefore $KS(KS(x)) \leq \log n + O(1)$. We shall show that

$$\forall n\, \exists x \quad \ell(x) = n \wedge KS(KS(x)|x) \geq \log n - O(1)\,, \tag{$*$}$$

from where the statement of the theorem follows immediately.

We say that a computable function $f(x,t)$ (where x is a binary word, t and values of f are naturals) is a *semi-enumerator* if $f(x,t)$ is monotonically non-increasing in t and $\lim\limits_{t\to\infty} f(x,t) \geq KS(x)$. If $\forall x\ \lim\limits_{t\to\infty} f(x,t) = KS(x)$, then f is called an *enumerator*. Values of f are called *hypotheses* about the entropy of x.

Let f_0 be an enumerator, C be a large enough natural number. Let us transform them into a semi-enumerator f_1. For every x it passes to a new

```
0. For all natural $a$  $n > 5$ in parallel do the following.
 1. Sequentially for each $x$ of length $n$:
   2. Put $t = 0$.
   3. If $f(x,t) \leq n/2 + a$,
       then stop the process for this $n$.
     4. Look for a word $y$ of length $f(x,t) - a$
          such that $g(y,a)$ has not defined yet;
        if the word has been found,
         then put $g(y,a) = x$,
         else stop the process for this $n$.
     5. Wait for $t' > t$ such that $f(x,t') < f(x,t)$.
     6. If $f(x,t') < f(x,t) - a - 1$,
         then go to the next $x$ from the Item 1,
         else put $t = t'$ and go to the Item 3.
```

Figure 1: The algorithm that enumerates the graph of the function g.

hypothesis $f_1(x,t)$ only if within the time t this hypothesis was introduced by the enumerator f_0 and it was discovered that $KS(f_1(x,t)|x) < \lfloor \log \ell(x) \rfloor - C$. The number of hypotheses of f_1 for every x is less than $\ell(x)/2^C$. The simple entropy of a program for computing f_1 is less than $\log C + O(1)$. Assume that the statement $(*)$ is false, then $\exists n \, \forall x \; [\ell(x) = n \;\Rightarrow\; \lim\limits_{t\to\infty} f(x,t) = KS(x)]$. Now it is sufficient to prove the following lemma.

Lemma. *Let a number c be large enough and a semi-enumerator f be defined by a program of simple entropy $d > 0$. If $\forall x \; [\ell(x) = n \;\Rightarrow\; \lim\limits_{t\to\infty} f(x,t) = KS(x)]$, then on a certain x of length n the number of hypotheses of f is greater than $n/(cd)$.*

This lemma strengthens two results from the paper [2].

Proof. Without loss of generality it can be assumed that $f(x,0) = \ell(x) + O(1) > \ell(x)$. Let us construct an auxiliary partial computable function $g(y,a)$ (where y is a binary word, a and values of g are naturals). On Figure 1 we present an algorithm that enumerates the graph of the function g (in the sequel we refer to the items of this algorithm).

It is obvious that $\forall a, x \;\; KS(x) \leq \min\{\ell(y) \mid g(y,a) = x\} + O(d + \log a)$. Suppose b is a large enough number (it depends on the choice of an optimal programming language). Let us fix $a = bd$. In Item 5 we can always find $t' > t$ such that $f(x, t') < f(x, t)$, because $a > O(d + \log a)$. For all n, m there is not more than one y of length m such that $g(y, a) = x$, $\ell(x) = n$ and $KS(x) \geq m - 1$ (such y will be called an *exception*). For $\ell(y) = m$, a value $g(y, a)$ could be defined while considering of numbers $n < 2m$ in Item 0 only. The cardinality of $\{x \mid KS(x) < m - 1\}$ is less than $2^m/2$, the number of exceptions of length m is less than $2m$; therefore if in Item 4 we did not managed to find y, then it would hold $2^m/2 < 2m$, that contradicts the condition $n > 5$ in Item 0.

For $n < 4a$, the statement of the lemma is true if $c \geq 4b$. Suppose $n \geq 4a$.

If for every x of length n the number of hypotheses of f is not greater than $n/(cd)$, then for some t the condition in Item 6 will necessarily be true. Indeed, while $f(x, t)$ increases from $n + O(1)$ to $n/2 + a$ it makes less than $n/(cd)$ jumps, hence, one of them is greater than $(n/2 - a) : (n/(cd)) \geq cd/4 > a + 1$ (if $c > 4b + 4$). Thus for every x of length n the length of some hypothesis of f is strictly less than n. But the cardinality of $\{x \mid KS(x) < n\}$ is strictly less than 2^n, and the number of words of length n is equal to 2^n. A contradiction. □

This completes the proof of the theorem. □

In Theorems 2 and 3 there was no effective bound on the growth rate of the functions $IP(KP(x) : x)$ $IP(x : KP(x))$. This fact is not surprising: as is known, even for the functions $KP(x)$ $KP(KP(x))$ there are no unlimited partial computable lower bounds. Nonetheless, the two following theorems show that for each of the functions $IP(KP(x) : x)$ $IP(x : KP(x))$ there exists an infinite (and not too thin) set such that the function grows on it fast enough.

Theorem 5. *For every natural k, there exists a binary word z of length not greater than k such that $IP(KP(z) : z) \geq \log k - O(1)$.*

Proof. Assume that the number k is great enough (for small k, the statement is trivial).

Let us define a computable function $\nu^t(n|x)$ of natural arguments n, x, and t with rational values. For $t = 0$, the function ν is identically equal to zero. While t is increasing, the function is not decreasing, and for all t, n it holds $\sum\limits_n \nu^t(n|x) \leq 1$.

Put $\nu^{t+1}(x|n) = (n/4)\cdot 2^{-n}$ if $\mu^t(x) \geq 2^{-n}$ and $\bigl|\{x|\mu^t(x) \geq 2^{-n}\}\bigr| \leq 2^{n+2}/n$, in the other case $\nu^{t+1}(x|n) = \nu^t(x|n)$.

The function $\nu(x|n) = \max_t \nu^t(x|n)$ is enumerable from below, $\sum_x \nu(x|n) \leq 1$, therefore $\nu(n|x) \leq O(\mu(n|x))$.

Let us show that there is $n \in [k/2, k]$ such that for all t it holds $\bigl|\{x|\mu^t(x) \geq 2^{-n}\}\bigr| \leq 2^{n+2}/n$. If this was not true, then

$$\forall n \in [k/2, k] \quad \bigl|\{x|\mu(x) \geq 2^{-n}\}\bigr| \cdot 2^{-n} > 4/n\,.$$

Summing, we obtain

$$\sum_{n=k/2}^{k} \bigl|\{x|\mu(x) \geq 2^{-n}\}\bigr| \cdot 2^{-n} > \sum_{n=k/2}^{k} 4/n > 2\,.$$

On the other hand,

$$2 \geq \sum_x 2\mu(x) \geq \sum_{n=k/2}^{k} \bigl|\{x|\mu(x) \geq 2^{-n}\}\bigr| \cdot 2^{-n}\,,$$

a contradiction.

Let us fix this n.

Consider a random word y of length k. Its simple, and hence prefix, entropy is greater than $k - O(1)$. Let y_i be the beginning of y of length i. It is clear that in the sequence $KP(y_k), KP(y_{k-1}), \ldots$ the difference between neighbouring numbers is less than a constant. If j is the least natural number such that $KP(y_{k-j}) \leq n$, then $KP(y_{k-j}) \geq n - C$, where C depends on the choice of an optimal programming language only. Put $z = y_{k-j}$, then $\mu(z) \geq 2^{-n}$ and $\mu^t(z) \geq 2^{-n}$ for t greater than some t_0. For $t > t_0$, we have $\nu^t(z|n) = (n/4)\cdot 2^{-n}$ (by the choice of n). Hence $\mu(z|n) \geq 2^{-O(1)} \cdot (n/4) \cdot 2^{-n}$. Taking a logarithm, we get $KP(z|n) \leq n - \log n + O(1)$. Since $|KP(z) - n| \leq C$, we obtain $KP(z|KP(z)) \leq KP(z|n) + O(1)$. Therefore $KP(z) - KP(z|KP(z)) \geq (n - C) - (n - \log n + O(1)) \geq \log k - O(1)$. This completes the proof. □

Theorem 6. *For every natural k, there exists a binary word z of length not greater than k such that $IP(z : KP(z)) \geq \log k - O(\log\log k)$.*

Proof. Assume that the number k is great enough (for small k, the statement is trivial).

Consider a natural number $n \leq k$ such that the binary representation of it is a random word of length $\lfloor \log k \rfloor$, that is $KS(n) = \lfloor \log k \rfloor + O(1)$. Let z be a random word of length n, that is $KS(z) = n + O(1)$. It is clear that $KS(n) - O(1) \leq KP(n) \leq KP(KP(z)) + KP(KP(z) - n) + O(1) \leq KP(KP(z)) + KP(O(\log n)) + O(1) \leq KP(KP(z)) + O(\log \log n)$. Hence $KP(KP(z)) \geq \log k - O(\log \log k)$. On the other hand, $KP(KP(z)|z) \leq KP(n|z) + KP(KP(z) - n|z) + O(1) \leq O(1) + KP(O(\log n)|z) \leq O(\log \log k)$. Therefore, $KP(KP(z)) - KP(KP(z)|z) \geq \log k - O(\log \log k)$. The theorem is proved. □

It was useful and helpful for the author to be acquainted with the paper [2] and a comment of an anonymous referee of that paper about its connections with the work [3]. The author is grateful to Alexey Chernov for his help in preparing the text.

References

[1] An. A. Muchnik, S. Ye. Positselsky. Kolmogorov entropy in the context of computability theory. *Theoretical Computer Science*, 2002, v. 271, pp. 15–35.

[2] R. Beigel, H. Buhrman, P. Fejer, L. Fortnow, L. Longpré, F. Stephan, L. Torenvliet. Enumerators of the Kolmogorov Function. Manuscript, 2002.

[3] P. Gács. On the symmetry of algorithmic information. *Sov. Math. Dokl.*, 1974, 15, 1477–1480