Information Theory over Multisets

Cosmin Bonchiş[1], Cornel Izbaşa[1], Gabriel Ciobanu[2]

[1] Research Institute "e-Austria" Timişoara, Romania
{cosmin, cornel}@ieat.ro
[2] "A.I. Cuza" University, Faculty of Computer Science and
Romanian Academy, Institute of Computer Science
gabriel@info.uaic.ro

"The words, the sad words,
Sometimes surround the time
As a pipe, the water which flows within."
Nichita Stănescu

Summary. Starting from Shannon theory of information, we present the case of producing information in the form of multisets, and encoding information using multisets. We compute the entropy of a multiset information source by constructing an equientropic string source (with interdependent symbols), and we compare this with a string information source with independent symbols. We then study the encoder and channel part of the system, obtaining some results about multiset encoding length and channel capacity.

1 Motivation

The attempt to study information sources which produce multisets instead of strings, and ways to encode information on multisets rather than strings, originates in observing new computational models like membrane systems which employ multisets [5]. Membrane systems have been studied extensively and there are plenty of results regarding their computing power, language hierarchies and complexity. However, while any researcher working with membrane systems (called also P systems) would agree that P systems process information, and that living cells and organisms do this too, we are unaware of any attempt to precisely describe natural ways to encode information on multisets or to study sources of information which produce multisets instead of strings. One could argue that, while some of the information in a living organism is encoded in a sequential manner, like in DNA for example, there might be important molecular information sources which involve multisets (of molecules) in a non-trivial way.

A simple question: given a P system with, say, 2 objects a and 3 objects b from a known vocabulary V (suppose there are no evolution rules), how much

information is present in that system? Also, many examples of P systems perform various computational tasks. Authors of such systems encode the input (usually numbers) in various ways, some by superimposing a string-like structure on the membrane system [1], some by using the natural encoding or the unary numeral system, that is, the natural number n is represented with n objects, for example, a^n. However, just imagine a gland which uses the bloodstream to send molecules to some tissue which, in turn, sends back some other molecules. There is for sure an energy and information exchange. How to describe it? Another, more general way to pose that question is: what are the natural ways to encode numbers, and more generally, information on multisets, and how to measure the encoded information?

If membrane systems, living cells and any other (abstract or concrete) multiset processing machines are understood as information processing machines, then we believe that such questions should be investigated. According to our knowledge, this is the first attempt of such an investigation. We start from the idea that a study of multiset information theory might produce interesting, useful results at least in systems biology; if we understand the *natural* ways to encode information on multisets, there is a chance that *Nature* might be using similar mechanisms.

Another way in which this investigation seems interesting to us is that there is more challenge in efficiently encoding information on multisets, because they constitute a poorer encoding media compared to strings. Encoding information on strings or even richer, more organized and complex structures are obviously possible and have been studied. Removing the symbol order, or their position in the representation as strings can lead to multisets carrying a certain penalty, which deserves a precise description. Order or position do *not* represent essential aspects for information encoding; symbol multiplicity, a native quality of multisets, is *enough* for many valid purposes. We focus mainly on such "natural" approaches to information encoding over multisets, and present some advantages they have over approaches that superimpose a string structure on the multiset. Then we encode information using multisets in a similar way as it is done using strings.

There is also a connection between this work and the theory of numeral systems. The study of number encodings using multisets can be seen as a study of a class of purely non-positional numeral systems.

2 Entropy of an Information Source

Shannon's information theory represents one of the great intellectual achievements of the twentieth century. Information theory has had an important and significant influence on probability theory and ergodic theory, and Shannon's mathematics is a considerable and profound contribution to pure mathematics.

Shannon's important contribution comes from the invention of the source-encoder-channel-decoder-destination model, and from the elegant and general solution of the fundamental problems which he was able to pose in terms of this model. Shannon has provided significant demonstration of the power of coding with delay

in a communication system, the separation of the source and channel coding problems, and he has established the fundamental natural limits on communication. As time goes on, the information theoretic concepts introduced by Shannon become more relevant to day-to-day more complex process of communication.

2.1 Short Review of Shannon Information Theory

We use the notions defined in the classical paper [6] where Shannon has formulated a general model of a communication system which is tractable to a mathematical treatment.

Definition 1. *The quantity H is a reasonable measure of choice or information.*

Consider an information source modeled by a discrete Markov process. For each possible state i of the source there is a set of probabilities $p_i(j)$ associated to the transitions to state j. Each state transition produces a symbol corresponding to the destination state, e.g., if there is a transition from state i to state j, the symbol x_j is produced. Each symbol x_i has an initial probability $p_{i\in\overline{1..n}}$ corresponding to the transition probability from the initial state to each state i.

We can also view this as a random variable X with x_i as events with probabilities p_i, $X = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \\ p_1 & p_2 & \cdots & p_n \end{pmatrix}$.

There is an entropy H_i for each state. The entropy of the source is defined as the average of these H_i weighted in accordance with the probability of occurrence of the states:

$$H(X) = \sum_i P_i H_i = -\sum_{i,j} P_i p_i(j) \log p_i(j) \tag{1}$$

Suppose there are two symbols x_i, x_j and $p(i,j)$ is the probability of the successive occurrence of x_i and then x_j. The entropy of the joint event is

$$H(i,j) = -\sum_{i,j} p(i,j) \log p(i,j)$$

The probability of symbol x_j to appear after the symbol x_i is the conditional probability $p_i(j)$.

String Entropy

Consider an information source which produces sequences of symbols selected from a set of n independent symbols x_i with probabilities p_i. The entropy formula for such a source is given in [6]:

$$H(X) = \sum_{i=1}^{n} p_i \log_b \frac{1}{p_i}$$

2.2 Multiset Entropy

We consider a discrete information source modeled by a discrete-time first-order Markov process (or Markov chain) which produces multiset messages (as opposed to string messages). A message is a multiset of symbols. To compute the entropy of such a source, we construct an equientropic source which produces strings with mutually dependent symbols. Each string produced by this equientropic source is an exponent of a multiset produced by the multiset source, because a multiset is a string equivalence class.

The entropy of such a source is computed by Shannon's formula 1, where P_i is the probability of state i, and $p_i(j)$ is the transition probability from state i to state j. To compute the probability of the state i we must first observe what is specific for the multisets. The corresponding state trees are presented in the next figures.

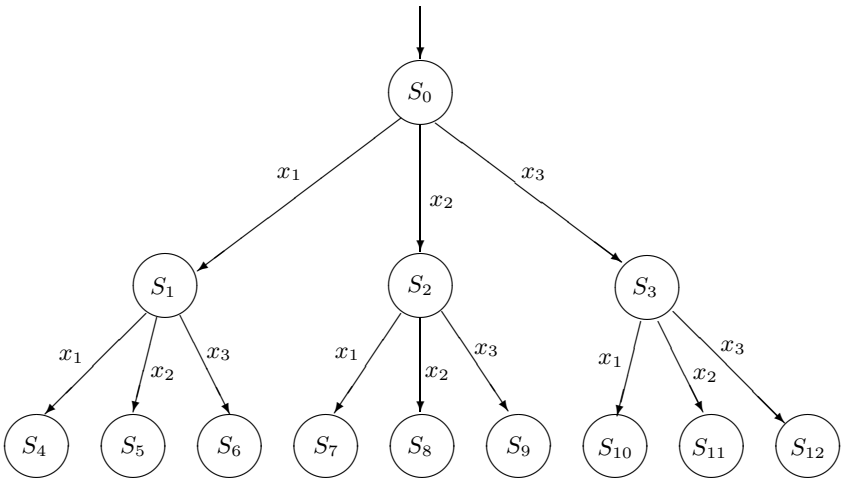

Fig. 1. String source states tree

We take the P_i for the first level of the tree, and because $P_0 = 1$ we get:

$$P_i = P_0 p_0(i) = p_i \tag{2}$$

To compute the transition probability $p_i(j)$ we know that for multisets $p(i, j) = 0$ for $i > j$.

Let N be the number of all symbols (with repetition allowed). Then the most probable number of symbols x_j is $p_j N$. For $i \leq j$, in order to obtain j after i, we observe that the symbols $x_{i>j}$ cannot be produced. Therefore, the probability to obtain j after i is given by the number of favorable cases over all possible cases

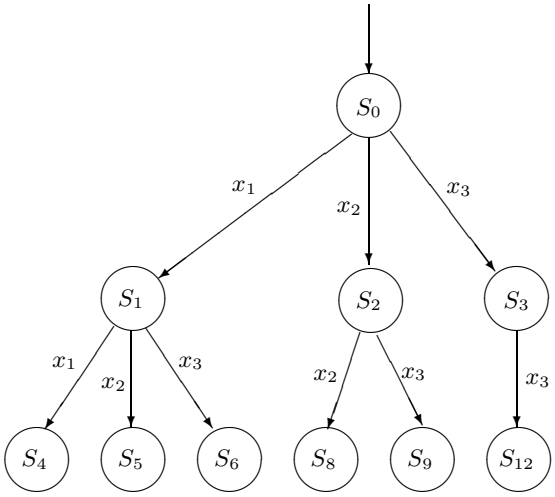

Fig. 2. Multiset source states tree

$$p_i(j) = \frac{p_j N}{N - \sum_{j=1}^{i-1} p_j N} = \frac{p_j}{\sum_{j=i}^{n} p_j}$$

namely

$$p_i(j) = \begin{cases} 0, & i > j \\ \frac{p_j}{\sum_{j=i}^{n} p_j}, & i \leq j \end{cases} \tag{3}$$

Theorem 1. *The entropy formula of a multiset generating information source is:*

$$H(X) = -\sum_{i=1}^{n} p_i \sum_{j=i}^{n} \frac{p_j}{\sum_{k=i}^{n} p_k} \log\left(\frac{p_j}{\sum_{k=i}^{n} p_k}\right). \tag{4}$$

Proof. From 1, 2, and 3 we infer

$$\begin{aligned} H(X) &= -\sum_{i,j,i\leq j} p_i \frac{p_j}{\sum_{k=i}^{n} p_k} \log\left(\frac{p_j}{\sum_{k=i}^{n} p_k}\right) \\ &= -\sum_{i=1}^{n} p_i \sum_{j=i}^{n} \frac{p_j}{\sum_{k=i}^{n} p_k} \log\left(\frac{p_j}{\sum_{k=i}^{n} p_k}\right). \end{aligned}$$

Proposition 1. *When the events are equiprobable, i.e.,* $p_i = \frac{1}{n}$*, then*

$$H(X) = \frac{\log n!}{n}.$$

Proof. We substitute $\frac{1}{n}$ for p_i in equation (4), and get

$$\begin{aligned} H(X) &= -\sum_{i=1}^{n} \frac{1}{n} \sum_{j=i}^{n} \left(\frac{\frac{1}{n}}{\sum_{j=i}^{n} \frac{1}{n}} \log \left(\frac{\frac{1}{n}}{\sum_{j=i}^{n} \frac{1}{n}} \right) \right) \\ &= -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=i}^{n} \frac{1}{n-i+1} \log \frac{1}{n-i+1} \\ &= -\frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{n-i+1} \log \frac{1}{n-i+1} \right) \sum_{j=i}^{n} 1 \\ &= \frac{1}{n} \sum_{i=1}^{n} \log(n-i+1) \\ &= \frac{1}{n} \sum_{i=1}^{n} \log i = \frac{\log n!}{n}. \end{aligned}$$

String Source Entropy vs. Multiset Source Entropy

Theorem 2. *The entropy of a multiset-producing source* is lower than or equal to *the entropy of an equiprobable string-producing source:*

$$H_{multiset} \leq H_{string(x_i - equiprobable)}$$

Proof. We know that $\sum_{i=1}^{n} p_i = 1 \Rightarrow \sum_{k=i}^{n} p_k \leq 1 \Rightarrow$

$$\frac{p_j}{\sum_{k=i}^{n} p_k} \geq p_j \tag{5}$$

Gibbs inequality suppose that $P = \{p_1, p_2, \ldots, p_n\}$ is a probability distribution. Then for any other probability distribution $Q = \{q_1, q_2, \ldots, q_n\}$ the following inequality holds

$$-\sum_{i=1}^{n} p_i \log p_i \leq -\sum_{i=1}^{n} p_i \log q_i \tag{6}$$

Then

$$H_m(X) = -\sum_{i=1}^{n} p_i \sum_{j=i}^{n} \frac{p_j}{\sum_{k=i}^{n} p_k} \log\left(\frac{p_j}{\sum_{k=i}^{n} p_k}\right) \overset{(5)}{\leq}$$

$$\leq -\sum_{i=1}^{n} p_i \sum_{j=i}^{n} \frac{p_j}{\sum_{k=i}^{n} p_k} \log p_j = -\sum_{i=1}^{n} \frac{p_i}{\sum_{k=i}^{n} p_k} \sum_{j=i}^{n} p_j \log p_j \overset{(6)}{\leq}$$

$$\overset{(6)}{\leq} -\sum_{i=1}^{n} \frac{p_i}{\sum_{k=i}^{n} p_k} \sum_{j=i}^{n} p_j \log q_j$$

with $Q_i = \{q_j | \text{where, } j = \overline{i,n} \text{ and } \sum_{j=i}^{n} q_j = 1\}$, and $q_j = \dfrac{1}{n-i+1}$:

$$-\sum_{i=1}^{n} \frac{p_i}{\sum_{k=i}^{n} p_k} \sum_{j=i}^{n} p_j \log q_j = -\sum_{i=1}^{n} \frac{p_i}{\sum_{k=i}^{n} p_k} \sum_{j=i}^{n} p_j \log \frac{1}{n-i+1}$$

$$= -\sum_{i=1}^{n} \frac{p_i}{\sum_{k=i}^{n} p_k} \left(\log \frac{1}{n-i+1}\right) \sum_{j=i}^{n} p_j = \sum_{i=1}^{n} p_i \log(n-i+1) \leq$$

$$\leq \sum_{i=1}^{n} p_i \log n = \log n = H_{string}(X)_{x_i - equiprobable}$$

Corollary 1. *When X is equiprobable, $H_m \leq H_s$.*

Proof. For $p_i = \dfrac{1}{n}$ we have

$$H_{multiset} = \frac{1}{n} \sum_{i=1}^{n} \log i = \frac{\log n!}{n} \leq \log n = H_{string}$$

Maximum Entropy for a Multiset Source

For a multiset source, equiprobable events do not generate the maximum entropy. This is obtained by maximizing expression 4, which seems difficult in the general case, but we give an example for the simplest case - with two events (a binary multiset source):

$$X = \begin{pmatrix} x_1 & x_2 \\ p_1 & p_2 \end{pmatrix}$$

The multiset entropy for these events is: $H_{multiset}(X) = -p_1(p_1 \log p_1 + p_2 \log p_2)$. Let $p = p_1 \Rightarrow H_{multiset}(X) = -p[p \log p + (1-p)\log(1-p)]$. Since this function has only one maximum in $[0,1]$, we need to solve:

$$H'_{multiset}(X) = 2p[\log(1-p) - \log p] - \log(1-p) = 0.$$

A numerical solution is $p \approx 0.703506$. The maximizing probability distribution is $X \approx \begin{pmatrix} x_1 & x_2 \\ 0.703506 & 0.296494 \end{pmatrix}$ and the maximum entropy is

$$H_{multiset}(X) \approx 0.427636 < H_{string}(X_{equiprobable}) = \log 2 \approx 0.6931472.$$

3 Multiset Encoding and Channel Capacity

After exploring the characteristics of a multiset generating information source, we move to the channel part of the communication system. Properties of previously developed multiset encodings are analyzed in [2, 3]. A formula for the capacity of multiset communication channel is derived based on the Shannon's general formula. Please note that one can have a multiset information source and a usual sequence-based encoder and channel. All the following combinations are possible:

Source/Encoder	Sequential	Multiset
Sequential	[6]	this paper
Multiset	this paper	this paper

Table 1. Source/Encoder types

3.1 String Encoding

We shortly review the results concerning the string encoding.

Encoding Length

We have a set of symbols X to be encoded, and an alphabet A. We consider the uniform encoding. Considering the length l of the encoding, then $X = \{x_i = a_1 a_2 \dots a_l \mid a_j \in A\}$.

If $p_i = P(x_i) = \frac{1}{n}$, then we have

$$H(X) = \sum_{i=1}^{n} \frac{1}{n} \log_b(n) = \log_b(n) \leq l$$

It follows that $n \leq b^l$. For $n \in \mathbb{N}$, $n - b^x = 0$ implies $x_0 = \log_b n$ and so $l = \lceil x_0 \rceil = \lceil \log_b n \rceil$.

Channel Capacity

Definition 2. *[6] The* capacity C *of a discrete channel is given by*

$$C = \lim_{T \to \infty} \frac{\log N(T)}{T}$$

where $N(T)$ is the number of allowed signals of duration T.

Theorem 3. *[6] Let $b_{ij}^{(s)}$ be the duration of the s^{th} symbol which is allowable in state i and leads to state j. Then the channel capacity C is equal to* $\log W$ *where W is the largest real root of the determinant equation:*

$$\left| \sum_s W^{-b_{ij}^{(s)}} - \delta_{ij} \right| = 0$$

where $\delta_{ij} = 1$ if $i = j$, and zero otherwise.

3.2 Multiset Encoding

We present some results related to the multiset encoding.

Encoding Length

We consider a set X of N symbols, an alphabet A, and the length of encoding l, therefore:

$X = \{x_i = a_1^{n_1} a_2^{n_2} \dots a_b^{n_b} \mid \sum_{j=1}^{b} n_j = l,\ a_j \in A\}$.

Proposition 2. *Non-uniform encodings over multisets are shorter than uniform encodings over multisets.*

Proof. Over multisets we have

1. for an uniform encoding: $N \leq N(b,l) = \left\langle {b \atop l} \right\rangle = \binom{b+l-1}{l} = \frac{(b+l-1)!}{l!(b-1)!} = \frac{\prod_{i=1}^{b-1}(l+i)}{(b-1)!}$. If x_0 is the real root of $n - \frac{\prod_{i=1}^{b-1}(x+i)}{(b-1)!} = 0$ then $l = \lceil x_0 \rceil$.
2. for non-uniform encoding: $N \leq N(b+1, l-1) = \left\langle {b+1 \atop l-1} \right\rangle = \binom{b+l-1}{l-1}$ $= \frac{(b+l-1)!}{(l-1)!b!} = \frac{\prod_{i=0}^{b-1}(l+i)}{b!} = \frac{l}{b}\frac{\prod_{i=1}^{b-1}(l+i)}{(b-1)!} = \frac{l}{b}N(b,l)$. Let $x_0^{'}$ be the real root of $n - \frac{\prod_{i=0}^{b-1}(x+i)}{b!} = 0$. Then $l' = \left\lceil x_0^{'} \right\rceil$.

From $n - N(b, x_0) = 0$ and $n - \frac{x_0^{'}}{b}N(b, x_0^{'}) = 0$ we get $N(b, x_0) = \frac{x_0^{'}}{b}N(b, x_0^{'})$. In order to prove $l > l' \iff x_0 > x_0'$, let suppose that $x_0 \leq x_0^{'}$. We have $x_0^{'} > b$ (for sufficiently large numbers), and this implies that $N(b, x_0) \leq N(b, x_0^{'}) < \frac{x_0^{'}}{b}N(b, x_0^{'})$. Since this is false, it follows that $x_0 > x_0'$ implies $l \geq l'$.

Channel Capacity

We consider that a sequence of multisets is transmitted along the channel. The capacity of such a channel is computed for base 4, then some properties of it for any base are presented.

Multiset channel capacity in base 4

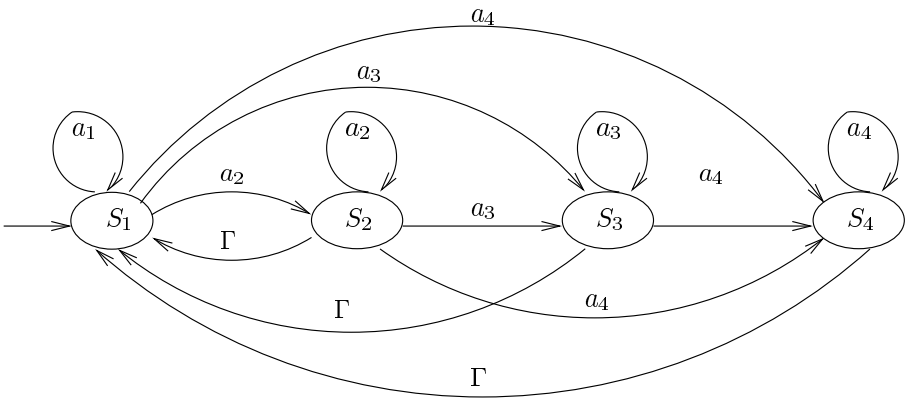

Fig. 3. Multiset channel capacity

In Figure 3 we have a graph $G(V, E)$ with 4 vertices $V = \{S_1, S_2, S_3, S_4\}$ and $E = \{(i, j) \mid i, j = \overline{1..4},\ i \leq j\} \cup \{(i, j) \mid i = 4,\ j = \overline{1..3}\}$

In Theorem 3 we get $b_{ij}^{(a_k)} = t_k$ because we consider that the duration to produce a_k is the same for each $(i, j) \in E$. The determinant equation is

$$\begin{vmatrix} W^{-t_1}-1 & W^{-t_2} & W^{-t_3} & W^{-t_4} \\ W^{-t_\Gamma} & W^{-t_2}-1 & W^{-t_3} & W^{-t_4} \\ W^{-t_\Gamma} & 0 & W^{-t_3}-1 & W^{-t_4} \\ W^{-t_\Gamma} & 0 & 0 & W^{-t_4}-1 \end{vmatrix} = 0$$

If we consider $t_\Gamma = t_k = t$, then the equation becomes

$1 - \dfrac{4}{W^t} + \dfrac{3}{W^{2t}} - \dfrac{1}{W^{3t}} = 0$, and $W_{real} = \sqrt[t]{\dfrac{4 + \sqrt[3]{\frac{47-3\sqrt{93}}{2}} + \sqrt[3]{\frac{47+3\sqrt{93}}{2}}}{3}} \approx \sqrt[t]{3.147899}$.
Therefore $C = \log_4 \sqrt[t]{3.147899}$ for $t = 1$, and so $C \approx 0.827194$.

Multiset channel capacity in base b

The determinant equation is

$$\begin{vmatrix} W^{-t_1}-1 & W^{-t_2} & W^{-t_3} & \cdots & W^{-t_b} \\ W^{-t_\Gamma} & W^{-t_2}-1 & W^{-t_3} & \cdots & W^{-t_b} \\ W^{-t_\Gamma} & 0 & W^{-t_3}-1 & \cdots & W^{-t_b} \\ \vdots & \vdots & \vdots & & \vdots \\ W^{-t_\Gamma} & \cdots & 0 & W^{-t_{b-1}}-1 & W^{-t_b} \\ W^{-t_\Gamma} & 0 & 0 & \cdots & W^{-t_b}-1 \end{vmatrix} = 0$$

Proposition 3. *If $t_\Gamma = t_k = t$, then the determinant equation becomes*

$$(1-\frac{1}{W^t})^{b-1} - \frac{1}{W^t} = 0. \tag{7}$$

The capacity C is given by $C = \log_b W$, where W is the largest real root of the equation (7). Considering $x = W^{-t}$, then we have

$$W = \frac{1}{\sqrt[t]{x}} \Rightarrow C = -\frac{1}{t}\log_b x. \tag{8}$$

Since we need the largest real root W then we should find the smallest positive root x of the equation

$$(1-x)^{b-1} - x = 0 \tag{9}$$

Let $f_b(x) = (1-x)^{b-1} - x$.

Lemma 1. *For all b there is a unique $x_b \in (0,1)$ such that $f_b(x_b) = 0$.*

Proof. We have $f_b'(x) = -(b-1)(1-x)^{b-2} - 1$.

- b is odd $\Rightarrow f_b'(x) = 0$ has the real root $x = 1 + \frac{1}{\sqrt[k-1]{k}} > 1$ and so $f_b'(x) < 0$ for all $x \in (-\infty, 1]$;
- b is even $\Rightarrow f_b'(x) \leq 0$ for all $x \in \mathbb{R}$.

Therefore $f_b(x)$ is decreasing for $x \in (0,1)$, $f_b(0) = 1$ and $f_b(1) = -1$. Then there exists a unique $x_b \in (0,1)$ such that $f_b(x_b) = 0$.

Lemma 2. *The smallest positive root of Equation (9) is decreasing with respect to b. More exactly, for all b we have $x_b \geq x_{b+1}$, where x_b is the smallest positive root of $f_b(x) = 0$.*

Proof. $f_{b+1}(x) - f_b(x) = (1-x)^b - x - ((1-x)^{b-1} - x) = -x(1-x)^{b-1}$. Then $f_{b+1}(x) - f_b(x) \leq 0$ for all $x \in (0,1)$. Since $f_{b+1}(x_b) \leq 0$ and $f_{b+1}(0) = 1$, then we have $x_{b+1} \in (0, x_b)$ according to Lemma 1.

Theorem 4. *Channel capacity is an increasing function with respect to b.*

Proof. This follows by Lemma 2 and Equation (8).

Remark 1. When $n = 2$, the capacity is $C = \frac{1}{t}$.

Proof. From $1 - \frac{2}{W^t} = 0$ we get $C = \log_2 \sqrt[t]{2} = \frac{1}{t}$.

4 Conclusion

Based on Shannon's classical work, we present a multiset entropy formula of an information source. We also present some relationships between this entropy and the string entropy. For a binary multiset source, we compute an approximate maximal value for the entropy. Using the determinant capacity formula, we compute the multiset channel capacity in base 4, and we describe some properties of the multiset channel capacity in base b. As future work we plan to further explore the properties of multiset based communication systems, and to develop some methods for computing the maximal multiset entropy in the general case.

A poetic vision of communication

Nichita Stănescu (1933-1983) was a Romanian poet proposed for Nobel Prize for literature. Here is his view of words and communication, first in Romanian and then in English (translation is ours).

> "Cuvintele / nu au loc decât în centrul lucrurilor, / numai înconjurate de lucruri. // Numele lucrurilor / nu e niciodata afară.
> Şi totuşi / cuvintele, tristele, / înconjoară câteodata timpul / ca o ţeavă, apa care curge prin ea. // ... ca şi cum ar fi lucruri..., / oho, ca şi cum ar fi lucruri..."
>
> "Words / do not belong but are in the center of things, / only surrounded by things. // The names of the things are never outside.
> But still / the words, the sad words, / sometimes surround the time / as a pipe, the water which flows within.// ... as they would be things..., / oh, as they would be things."
>
> *Nichita Stănescu*
>
> "For Nichita Stănescu, the pipe of words is for time what the communication channel is for a message; time flows through the pipe made of words as a message passes as a fluid through that which we call a communication channel."
>
> *Solomon Marcus [4]*

References

1. A. Atanasiu: Arithmetic with Membranes. *Pre-Proceedings of the Workshop on Multiset Processing*, Curtea de Argeş, 2000, 1–17.
2. C. Bonchiş, G. Ciobanu, C. Izbaşa: Encodings and Arithmetic Operations in Membrane Computing. *Theory and Applications of Models of Computation*, Lecture Notes in Computer Science vol. 3959, Springer, Berlin, 2006, 618–627.
3. C. Bonchiş, G. Ciobanu, C. Izbaşa: Number Encodings and Arithmetics over Multisets. *SYNASC'06: 8th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing*. Timişoara, IEEE Computer Society, 2006, 354–361.

4. S. Marcus: *Intâlnirea Extremelor*. Paralela 45, Bucharest, 2005, 166.
5. Gh. Păun: *Membrane Computing. An Introduction*. Springer, Berlin, 2002.
6. C.E. Shannon: A Mathematical Theory of Communication. *Bell System Technical Journal*, 27 (1948), 379–423, and 623–656.