On the Uniform Distribution of Strings

Sébastien Rebecchi* and Jean-Michel Jolion

Université de Lyon, F-69361 Lyon
INSA Lyon, F-69621 Villeurbanne
CNRS, LIRIS, UMR 5205
{sebastien.rebecchi, jean-michel.jolion}@liris.cnrs.fr

Abstract. In this paper, we propose the definition of a measure for sets of strings of length not greater than a given number. This measure leads to an instanciation of the uniform distribution definition in sets of such limited-size strings, for which we provide a linear time complexity generative algorithm.
Some ideas could rather easily be extended to other ordered structure types.

Keywords: string, uniform distribution, measure

1 Introduction

For many years, several research teams have tried to blend the two main approaches of pattern recognition, namely the statistical and structural ones [3,2]. This choice is justified by the desire for being able at the same time to benefit from the undeniable advantages of the two approaches, while being detached of their respective drawbacks.

The statistical pattern recognition is based on a coding of the data in the form of numerical vectors, often unable to accurately reproduce the complexity of the data. However this choice is justified by the broad pallet of statistical algorithms published in the literature and recognized as powerful for the classification of numerical data[1].

In the structural pattern recognition paradigm, the coding part is rich because of being based on data structures of great expressivity (graphs, strings, trees...), allowing in particular to represent in an adequate way any kind of intra/inter patterns relations (reflexivity, sequentiality, hierarchy...). However, the tools related to the classification of structures are too restrictive (isomorphism, edit distance[4,6]...) and not robust enough for some applications specific to pattern recognition. Another limitation comes from the lack of a structural formalism for the processing of sets of data, in the sense that the tools classically used are generally based on only unary or binary operators. Finally, association between, on the one hand, the size of the data structures used, and, on the other hand, the complexity of the relative algorithms, tends to reject this approach for the processing of large volumes of data.

To be able to reconcile these two approaches, a necessary condition is to define a statistical characterization of spaces of structures. We propose to contribute to this vision, by translating the concept of distribution. We especially concentrate our attention to strings for which we propose the definition of a uniform distribution.

The uniform process is specified in a set S, with respect to a measure function μ, by the distribution for which the probability of a subset E of S, measurable by μ, is proportional to its measure: $\mathrm{P}(E) = \alpha \times \mu(E)$. More precisely, α is just the

* A part of this work was done when the author visited the *Gruppo di Ricerca sulle Macchine Intelligenti per il riconoscimento di Video, Immagini e Audio*, *Università degli Studi di Salerno*, Italy. The visit was supported by a grant from the *Région Rhône-Alpes*, France.

Sébastien Rebecchi, Jean-Michel Jolion: On the Uniform Distribution of Strings, pp. 116–125.
Proceedings of PSC 2008, Jan Holub and Jan Žďárek (Eds.), ISBN 978-80-01-04145-1

inverse of total measure $\mu(S)$ of S, that makes the uniform process be considered as a mere normalization one, passing from a measure to a distribution while respecting the *relative* measure of E in S.

Two well-known examples of such a specification are given in the discrete and continous one-dimension numerical cases, where the measure functions are, respectively, the cardinality ($\mu(E) = |E|$), and the Lebesgue measure (roughly speaking, $\mu(E) = \sum_{I \in \text{MaxLen}(E)} l(I)$, with $\text{MaxLen}(E)$ the set of maximal-length intervals that are subset of E, and $l(I)$ the length of the interval I). With respect to these measures, the uniform property demands, in the discrete case, that all elements have the same probability, and, in the continous one, that all intervals of the same length have the same probability.

As for the string case, we have to define a consistent measure for sets of such structures, *i.e.* a one that would take care of the inherent structural and combinatorial nature of this type of elements. Before going into details in section 3, we introduce some necessary notations and definitions.

2 Preliminary notations and definitions

Definition 1 (Alphabet). *An alphabet is a non-empty finite set whose elements are called letters.*

In the rest of this paper, A denotes an alphabet, and λ a special object, called *empty letter*, that does not belong to A . Moreover, we denote by $|A|$ the size of A, *i.e.* its cardinal.

Definition 2 (String). *A string over A is a finite-length sequence of letters of A.*

Let X be a string over A, and $n \in \mathbb{N}$. We use the following notation:

- $|X|$ the size of X, *i.e.* its length
- X_i the letter at position i in X, $i \in \{1, \ldots, |X|\}$
- Λ the empty string over A, *i.e.* of length 0
- A^n the set of strings over A of length n
- $A^{\leqslant n}$ the set of strings over A of length not greater than n
- A^* the set of strings over A

Notice that:

- $A^0 = \{\Lambda\}$
- $A^{\leqslant n} = \bigcup_{i=0}^{n} A^i$
- $A^* = \bigcup_{i \in \mathbb{N}} A^i$

Definition 3 (Concatenation). *The concatenation over A is the binary operation $. : A^* \times (A^* \cup A \cup \{\lambda\}) \longrightarrow A^*$, such that: $\forall (X, Y \in A^*, a \in A)$:*

- $X.\lambda = X$
- $[|X.a| = |X| + 1] \wedge [\forall i \in \{1, \ldots, |X|\}, (X.a)_i = X_i] \wedge [(X.a)_{|X|+1} = a]$
- $[|X.Y| = |X| + |Y|] \wedge [\forall i \in \{1, \ldots, |X|\}, (X.Y)_i = X_i] \wedge [\forall i \in \{1, \ldots, |Y|\}, (X.Y)_{|X|+i} = Y_i]$

Remark. $\Lambda.X = X.\Lambda = X$ follows from the last point of this definition.

Thanks to definition 3, we can "promote" a letter of $A \cup \{\lambda\}$ as a string of $A^{\leqslant 1}$, simply by concatenating it to Λ.

Definition 4 (Promotion). *The promotion over A is the bijection:*

$$\begin{aligned} A \cup \{\lambda\} &\to A^{\leqslant 1} \\ a &\to \Lambda.a \end{aligned}$$

Moreover, we say that a is the *promoted* of $\Lambda.a$. Notice that the empty letter λ is promoted to the empty string Λ, and that a non-empty letter is promoted to a non-empty string of length 1.

Definition 5 (σ-algebra). *Let S be a set. A σ-algebra σ over S is a set of subsets of S, such that:*

- *σ contains the empty set: $\emptyset \in \sigma$*
- *σ is closed under complementation: $E \in \sigma =: (S \setminus E) \in \sigma$*
- *σ is closed under countably infinite union:*

$$[\forall n \in \mathbb{N}, E_n \in \sigma] =: \left(\bigcup_{n \in \mathbb{N}} E_n \right) \in \sigma$$

Remark. If S is countable, then the power set (set of all subsets) of S is the only σ-algebra over S containing all singletons $\{x\}$, $x \in S$.

Definition 6 (Measure). *Let S be a set, and σ a σ-algebra over S. A measure μ over σ is a function $\sigma \to \mathbb{R}^+ \cup \{\infty\}$, such that:*

- *The empty set has a null measure: $\mu(\emptyset) = 0$*
- *μ is additive under disjoint countably infinite union:*

$$[\forall (i, j \in \mathbb{N} | i < j), E_i, E_j \in \sigma, E_i \cap E_j = \emptyset]$$

$$=:$$

$$\mu \left(\bigcup_{n \in \mathbb{N}} E_n \right) = \sum_{n \in \mathbb{N}} \mu(E_n)$$

In the rest of this paper, we simplify the notation $\mu(\{x\})$ by $\mu(x)$, for all singletons $\{x\}$, $x \in S$.

Definition 7 (Uniform distribution). *Let S be a set, σ a σ-algebra over S, and μ a measure over σ. A distribution is uniform w.r.t. μ iff: $\forall E \in \sigma$:*

$$\mathrm{P}(E) = \mu(E) \times \mu(S)^{-1}$$

In the rest of this paper, we simplify the notation $\mathrm{P}(\{x\})$ by $\mathrm{P}(x)$, for all singletons $\{x\}$, $x \in S$.

3 String uniform distribution

3.1 The measure

Let μ_A be a measure over the power set of $A \cup \{\lambda\}$, $n \in \mathbb{N}$, and σ^n the power set of $A^{\leqslant n}$. Our wish is to define a measure μ^n over σ^n.

We wish μ^n to respect two necessary properties relatives to the connection between, on the one hand, the combinatorial and structural nature of the string type, and, on the other hand, the set $A^{\leqslant n}$ for which is defined this measure:

1. μ^n has an additive effect under combination
2. μ^n has an multiplicative effect under concatenation

The first property would be related with the combinatorial nature of a string, considering it within the set $A^{\leqslant n}$: one can attempt to describe a string X of $A^{\leqslant n}$ with a n-tuple of letters of $A \cup \{\lambda\}$, simply by preserving the respective order of the letters of X, and padding with the appropriate number of λ. But in this case, one must face the problem that the number of possible λ-paddings is equal to the number of combinations of size $|X|$ from a set of cardinal n. This caracteristic must hold in the definition of μ^n: if we see a string X of $A^{\leqslant n}$ as being the canonical representation of a set of n-tuples, then, according to the additive property of a measure (*cf.* definition 6), the measure $\mu^n(X)$ should be obtained by the sum of the measures of all tuples associated to X, with respect to a measure defined over the power set of the set of n-tuples of letters of $A \cup \{\lambda\}$. If we denote by μ_t^n this intermediate n-tuple measure, then we have:

$$\mu^n(X) = \sum_{T \in \text{tuples}_n(X)} \mu_t^n(T)$$

where $\text{tuples}^n(X)$ stands for the set of n-tuples of $A \cup \{\lambda\}$ associated to X.

As for the second property, it would be related with the structural sequential nature of a string. Keeping in mind the viewpoint introduced above, according to which a string is the canonical representation of a set of tuples, we wish to consider a letter of X as its expression in a certain dimension of each of these tuples. We wish not to regard the particular position of this letter in a tuple, because it could introduce a specific relative importance of any particular position in a string, which is not our purpose in this general study, as we wish to preserve the one and only this one induced by μ_A. Then, each tuple composed of the same letters should have the same measure, and, according to this dimensional point of view, we have:

$$\mu_t^n(T) = \prod_{i=1}^{|T|} \mu_A(T_i)$$

The reasonning above drives us to the following definition:

Definition 8 (String measure). $\forall X \in A^{\leqslant n}$*:*

$$\mu^n(X) = \mathrm{C}_n^{|X|} \times \left(\prod_{i=1}^{|X|} \mu_A(X_i) \right) \times \mu_A(\lambda)^{n-|X|}$$

where $\mathrm{C}_n^{|X|}$ stands for the number of combinations of size $|X|$ from a set of cardinal n.

The measure of non-singleton sets simply follows from the property of a measure (*cf.* definition 6): sum of the measures of all the singletons that are subset of it (hence 0 for the empty set). Therefore, it would be straightforward to prove that μ^n *is* a measure over the power set of $A^{\leqslant n}$.

Finally, notice that the above formula follows the recursive rule below, that is going to be useful in 3.2:

Proposition 9 (String measure recursion). $n > 0 =: \forall (X \in A^{\leqslant n-1}, a \in A)$*:*

$$\mu^n(X.a) = \mu^n(X) \times \mathrm{C}_n^{|X|+1} \,/\, \mathrm{C}_n^{|X|} \times \mu_A(a)/\mu_A(\lambda)$$

Proof.

$$\begin{aligned}
\mu^n(X.a) = &\ \mathrm{C}_n^{|X.a|} \times \left(\prod_{i=1}^{|X.a|} \mu_A((X.a)_i)\right) \times \mu_A(\lambda)^{n-|X.a|} \\
=^* &\ \mathrm{C}_n^{|X|+1} \times \left(\prod_{i=1}^{|X|} \mu_A(X_i)\right) \times \mu_A(a) \times \mu_A(\lambda)^{n-|X|-1} \\
= &\ \left[\mathrm{C}_n^{|X|} \times \left(\prod_{i=1}^{|X|} \mu_A(X_i)\right) \times \mu_A(\lambda)^{n-|X|}\right] \times \mathrm{C}_n^{|X|+1} \,/\, \mathrm{C}_n^{|X|} \times \mu_A(a) \times \mu_A(\lambda)^{-1} \\
= &\ \mu^n(X) \times \mathrm{C}_n^{|X|+1} \,/\, \mathrm{C}_n^{|X|} \times \mu_A(a)/\mu_A(\lambda)
\end{aligned}$$

$*$ Definition 3 □

We impose $n > 0$ because the set A^{-1} is undefined ($-1 \notin \mathbb{N}$).

3.2 The distribution

Following the general requirement of the uniform specification (*cf.* definition 7) with respect to the measure μ^n, the probability of a string is given by the following formula:

Definition 10 (Uniform string distribution). $\forall X \in A^{\leqslant n}$*:*

$$\mathrm{P}^n(X) = \mu^n(X) \times \mu^n(A^{\leqslant n})^{-1}$$

It would be much complex to compute the total measure $\mu^n(A^{\leqslant n})$ of $A^{\leqslant n}$ by a naive recursive exponential time demanding enumeration of all strings of this set (remind that $\mu^n(A^{\leqslant n}) = \sum_{Y \in A^{\leqslant n}} \mu^n(Y)$). Fortunately, we can simplify it analytically:

Proposition 11 (Total string measure).

$$\mu^n(A^{\leqslant n}) = \mu_A(A \cup \{\lambda\})^n$$

The following formula is going to be helpful to prove the proposition:

Lemma 12. $\forall i \in \{0, \ldots, n-1\}$*:*

$$\mu^n(A^{i+1}) = \mu^n(A^i) \times \mathrm{C}_n^{i+1} \,/\, \mathrm{C}_n^{i} \times \mu_A(A)/\mu_A(\lambda)$$

Then, the proof of the proposition follows in a simple way:

Proof (Proposition 11).

$$\text{Lemma 12} \wedge \left[\mu^n(A^0) = \mu^n(\Lambda) =^* \mu_A(\lambda)^n\right] =: \forall i \in \{0, \ldots, n\}:$$

$$\begin{aligned}\mu^n(A^i) &= \mu_A(\lambda)^n \times \textstyle\prod_{j=1}^{i} \left(\mathrm{C}_n^j \,/\, \mathrm{C}_n^{j-1} \times \mu_A(A)/\mu_A(\lambda)\right) \\ &= \mu_A(\lambda)^n \times \mu_A(A)^i/\mu_A(\lambda)^i \times \textstyle\prod_{j=1}^{i} \mathrm{C}_n^j \,/\, \mathrm{C}_n^{j-1} \\ &= \mu_A(\lambda)^{n-i} \times \mu_A(A)^i \times \mathrm{C}_n^i \,/\, \mathrm{C}_n^0 \\ &= \mu_A(\lambda)^{n-i} \times \mu_A(A)^i \times \mathrm{C}_n^i\end{aligned}$$

Then, we have:

$$\begin{aligned}\mu^n(A^{\leqslant n}) = &\ \textstyle\sum_{i=0}^{n} \mu^n(A^i) \\ = &\ \textstyle\sum_{i=0}^{n} \left(\mu_A(\lambda)^{n-i} \times \mu_A(A)^i \times \mathrm{C}_n^i\right) \\ =^{**} &\ \left(\mu_A(\lambda) + \mu_A(A)\right)^n \\ = &\ \mu_A(A \cup \{\lambda\})^n\end{aligned}$$

$*$ Definition 8
$**$ Binomial theorem □

It now remains to prove the lemma:

Proof (Lemma 12). If $n = 0$, then $\{0, \ldots, n-1\} = \emptyset$, and thereby the lemma is true by vacuity. Else ($n > 0$), we have:

$$\begin{aligned}\mu^n(A^{i+1}) = &\ \textstyle\sum_{Y \in A^{i+1}} \mu^n(Y) \\ =^* &\ \textstyle\sum_{X \in A^i} \sum_{a \in A} \mu^n(X.a) \\ =^{**} &\ \textstyle\sum_{X \in A^i} \sum_{a \in A} \left(\mu^n(X) \times \mathrm{C}_n^{|X|+1} \,/\, \mathrm{C}_n^{|X|} \times \mu_A(a)/\mu_A(\lambda)\right) \\ = &\ \textstyle\sum_{X \in A^i} \sum_{a \in A} \left(\mu^n(X) \times \mathrm{C}_n^{i+1} \,/\, \mathrm{C}_n^{i} \times \mu_A(a)/\mu_A(\lambda)\right) \\ = &\ \textstyle\sum_{X \in A^i} \left(\mu^n(X) \times \mathrm{C}_n^{i+1} \,/\, \mathrm{C}_n^{i} \times \sum_{a \in A} \mu_A(a)/\mu_A(\lambda)\right) \\ = &\ \textstyle\sum_{X \in A^i} \left(\mu^n(X) \times \mathrm{C}_n^{i+1} \,/\, \mathrm{C}_n^{i} \times \mu_A(A)/\mu_A(\lambda)\right) \\ = &\ \left(\textstyle\sum_{X \in A^i} \mu^n(X)\right) \times \mathrm{C}_n^{i+1} \,/\, \mathrm{C}_n^{i} \times \mu_A(A)/\mu_A(\lambda) \\ = &\ \mu^n(A^i) \times \mathrm{C}_n^{i+1} \,/\, \mathrm{C}_n^{i} \times \mu_A(A)/\mu_A(\lambda)\end{aligned}$$

$*$ Definition 3
$**$ Proposition 9 □

According to definition 8 and proposition 11, we can compute the probability of a string in $O(n)$:

Definition 13 (Uniform string distribution). $\forall X \in A^{\leqslant n}$*:*

$$\mathrm{P}^n(X) = \mathrm{C}_n^{|X|} \times \left(\prod_{i=1}^{|X|} \mu_A(X_i)\right) \times \mu_A(\lambda)^{n-|X|} \times \mu_A(A \cup \{\lambda\})^{-n}$$

3.3 Preservation

An interesting property of our string uniform distribution is the preservation under concatenation: the concatenation of two uniform strings remains a uniform string:

Proposition 14 (Uniform string preservation). $\forall (i \in \{0, \ldots, n\}, Y \in A^{\leqslant i}, Z \in A^{\leqslant n-i})$*, if* Y *is uniformly distributed w.r.t.* μ^i*, and* Z *uniformly distributed w.r.t.* μ^{n-i}*, then* $X = Y.Z$ *is uniformly distributed w.r.t.* μ^n*.*

Proof. First, remind that we have: $\forall i, k \in \{0, \ldots, n\}$:

$$\mathrm{C}_n^k = \sum_{j=0}^{k} \left(\mathrm{C}_i^j \times \mathrm{C}_{n-i}^{k-j} \right)$$

Then, we have: $\forall X \in A^{\leqslant n}$:

$$\begin{aligned} \mathrm{C}_n^{|X|} =& \sum_{j=0}^{|X|} \left(\mathrm{C}_i^j \times \mathrm{C}_{n-i}^{|X|-j} \right) \\ =^* & \sum_{Y \in A^{\leqslant i}, Z \in A^{\leqslant n-i} | Y.Z = X} \left(\mathrm{C}_i^{|Y|} \times \mathrm{C}_{n-i}^{|Z|} \right) \end{aligned}$$

Moreover, we have: $\forall (Y \in A^{\leqslant i}, Z \in A^{\leqslant n-i} | Y.Z = X)$:

$$\begin{aligned} \mathrm{P}^i(Y) \times \mathrm{P}^{n-i}(Z) =^{**} & \left[\mathrm{C}_i^{|Y|} \times \left(\prod_{j=1}^{|Y|} \mu_A(Y_j) \right) \times \mu_A(\lambda)^{i-|Y|} \times \mu_A(A \cup \{\lambda\})^{-i} \right] \times \\ & \left[\mathrm{C}_{n-i}^{|Z|} \times \left(\prod_{j=1}^{|Z|} \mu_A(Z_j) \right) \times \mu_A(\lambda)^{(n-i)-|Z|} \times \mu_A(A \cup \{\lambda\})^{-(n-i)} \right] \\ =^* & \ \mathrm{C}_i^{|Y|} \times \mathrm{C}_{n-i}^{|Z|} \times \left(\prod_{j=1}^{|Y.Z|} \mu_A((Y.Z)_j) \right) \times \mu_A(\lambda)^{n-|Y.Z|} \times \\ & \ \mu_A(A \cup \{\lambda\})^{-n} \\ = & \ \mathrm{C}_i^{|Y|} \times \mathrm{C}_{n-i}^{|Z|} \times \left(\prod_{j=1}^{|X|} \mu_A(X_j) \right) \times \mu_A(\lambda)^{n-|X|} \times \\ & \ \mu_A(A \cup \{\lambda\})^{-n} \end{aligned}$$

Finally, we have:

$$\begin{aligned} \mathrm{P}^n(X) =^{**} & \ \mathrm{C}_n^{|X|} \times \left(\prod_{j=1}^{|X|} \mu_A(X_j) \right) \times \mu_A(\lambda)^{n-|X|} \times \mu_A(A \cup \{\lambda\})^{-n} \\ = & \ \sum_{Y \in A^{\leqslant i}, Z \in A^{\leqslant n-i} | Y.Z = X} \left(\mathrm{C}_i^{|Y|} \times \mathrm{C}_{n-i}^{|Z|} \right) \times \left(\prod_{j=1}^{|X|} \mu_A(X_j) \right) \times \mu_A(\lambda)^{n-|X|} \times \\ & \ \mu_A(A \cup \{\lambda\})^{-n} \\ = & \ \sum_{Y \in A^{\leqslant i}, Z \in A^{\leqslant n-i} | Y.Z = X} \left[\mathrm{C}_i^{|Y|} \times \mathrm{C}_{n-i}^{|Z|} \times \left(\prod_{j=1}^{|X|} \mu_A(X_j) \right) \times \mu_A(\lambda)^{n-|X|} \times \right. \\ & \left. \mu_A(A \cup \{\lambda\})^{-n} \right] \\ = & \ \sum_{Y \in A^{\leqslant i}, Z \in A^{\leqslant n-i} | Y.Z = X} \left[\mathrm{P}^i(Y) \times \mathrm{P}^{n-i}(Z) \right] \end{aligned}$$

We deduce that, for all $i \in \{0, \ldots, n\}$, the probability of a uniform string X w.r.t. μ^n is equal to the sum of the probabilities of all the possible concatenations of a uniform string Y w.r.t. μ^i, with a uniform string Z w.r.t. μ^{n-i}, such that $X = Y.Z$. This is exactly the meaning of the proposition.

$*$ Definition 3
$**$ Definition 13 □

This general reasoning leads to the following specific corollary, that is going to be useful in 3.4:

Corollary 15 (Uniform string preservation). *$\forall (X \in A^{\leqslant n}, a \in A \cup \{\lambda\})$, if X is uniformly distributed w.r.t. μ^n, and a uniformly distributed w.r.t. μ_A, then $X.a$ is uniformly distributed w.r.t. μ^{n+1}.*

The proof of this corollary follows from the following lemma:

Lemma 16. *$\forall a \in A \cup \{\lambda\}$, a is uniformly distributed w.r.t. μ_A iff $\Lambda.a$ is uniformly distributed w.r.t. μ^1.*

Proof (Corollary 15). According to lemma 16, $\Lambda.a$ is uniformly distributed w.r.t. μ^1. Then, according to proposition 14, $X.(\Lambda.a) =^* X.a$ is uniformly distributed w.r.t. μ^{n+1}.

$*$ Definition 3 □

Proof (Lemma 16). Let us denote by $\mathrm{P}_A(a)$ the probability of a acording to a uniform distribution w.r.t. μ_A. Then, we have:

$$\mathrm{P}_A(a) =^* \mu_A(a) \times \mu_A(A \cup \{\lambda\})^{-1}$$

If $a = \lambda$, then we have: $\mathrm{P}_A(a) = \mathrm{C}_1^0 \times 1 \times \mu_A(\lambda)^{1-0} \times \mu_A(A \cup \{\lambda\})^{-1}$.
Else ($a \in A$), we have: $\mathrm{P}_A(a) = \mathrm{C}_1^1 \times \mu_A(a) \times \mu_A(\lambda)^{1-1} \times \mu_A(A \cup \{\lambda\})^{-1}$.
So in all cases, we have:

$$\begin{aligned}\mathrm{P}_A(a) &=^{**} \mathrm{C}_1^{|\Lambda.a|} \times \left(\prod_{i=1}^{|\Lambda.a|} \mu_A((\Lambda.a)_i)\right) \times \mu_A(\lambda)^{1-|\Lambda.a|} \times \mu_A(A \cup \{\lambda\})^{-1} \\ &=^{***} \mathrm{P}^1(\Lambda.a)\end{aligned}$$

This proves the lemma, as the promotion is a bijection (*cf.* definition 4).

$*$ Definition 7
$**$ Definition 3
$***$ Definition 13 □

3.4 Generation

We wish to generate strings according to our uniform distribution. First, notice that, according to definition 13, we have: $\forall X \in A^{\leqslant n}$:

$$\begin{aligned}\mathrm{P}^n(X) &= \mathrm{C}_n^{|X|} \times \left(\prod_{i=1}^{|X|} \mu_A(X_i)\right) \times \mu_A(\lambda)^{n-|X|} \times \mu_A(A \cup \{\lambda\})^{-n} \\ &= \mathrm{C}_n^{|X|} \times \prod_{i=1}^{|X|} (\mu_A(X_i) \times \mu_A(A \cup \{\lambda\})^{-1}) \times \prod_{i=1}^{n-|X|} (\mu_A(\lambda) \times \mu_A(A \cup \{\lambda\})^{-1})\end{aligned}$$

This equation tells us that it is sufficient to, first, generate a n-tuple by the concatenation of n elements of $A \cup \{\lambda\}$ generated according to a uniform distribution w.r.t. μ_A, and, then, remove all the λ in T, to finaly obtain a string X generated according to a uniform distribution w.r.t. μ^n: according to this procedure, T would have a probability $\prod_{i=1}^{|X|} (\mu_A(X_i) \times \mu_A(A \cup \{\lambda\})^{-1}) \times \prod_{i=1}^{n-|X|} (\mu_A(\lambda) \times \mu_A(A \cup \{\lambda\})^{-1})$, and therefore X a probability $\mathrm{P}^n(X)$, as we have seen in 3.1 that X is the canonical representation of $\mathrm{C}_n^{|X|}$ such tuples T.

But this sufficiency can obviously also be retrieved in a slightly different form from a recursive use of the uniformity conversation of a string when concatenated with a uniform string w.r.t. μ^1, or its promoted uniform letter w.r.t. μ_A, as exhibit by corollary 15.

A simple $O(n)$-complex pseudo-code implementation of such a procedure is given by algorithm 1. Instead of passing by an intermediate tuple initialisation, filling, to a final canonization to string, the algorithm works directly under the string representation, thanks to a sequence of concatenations of a uniform letter to a string initialised to the empty one, for the same result, as mentioned above.

```
Input: A positive integer n
Output: A string X generated according to a uniform distribution w.r.t. μ^n
begin
    D ⟵ uniform distribution w.r.t. μ_A: ∀a ∈ A ∪ {λ}, P(a) = μ_A(a)/μ_A(A ∪ {λ});
    X ⟵ Λ;
    for i ⟵ 1 to n do
        l ⟵ random choice according to D;
        X ⟵ X.l;
    end
    return X;
end
```

Algorithm 1: Uniform string generation

4 Relation with a previous work

The present work subsumes and generalizes the part concerning the uniform distribution of the one exposed in [5] (in french): one can retrieve the special definition proposed in the later paper by setting $\mu_A(\lambda) = 0$, and $\mu_A(a) = \mu_A(b), \forall a, b \in A$ (this last equality can reasonably be supposed in numerous applications). One can also consider the restriction of μ^n over the restriction of σ^n over A^n, by ignoring the null λ influence, that leads to $\mathrm{P}^n(A^{\leqslant n-1}) = 0$ if $n > 0$. This would also define a uniform distribution w.r.t. the considered restriction of μ^n.

5 Discussion and further work

In fact, this work can be summarized as an analytical justification of this proposition: one can generate a uniform string over the alphabet A by concatenating uniform letters of $A \cup \{\lambda\}$. The simplicity of this property is obsiously due to the particular measure defined for sets of strings, but we have seen that this definition does not only please our wish for simplicity, but also fulfill some relevant arguments concerning the nature of the type of elements taken into consideration.

This simplicity could be rather easily extended to more complex structure types, for which we can define an order of the primitives. Let us consider, for instance, the (ordered) bounded arity and depth tree over A: if a is the maximum number of children that can have a node (arity), and d the maximum depth that can a tree, we could recursively define such a tree by a couple composed of the root and a string of children of length not greater than a, the children alphabet being the set of non-empty trees of arity a and maximum depth $(d-1)$, with the empty tree as associated empty letter. Then, we could take some ideas to the work above to instanciate a uniform distribution of such trees. A part of the associated tree measure would be defined thanks to the definition of the string one, with a necessary supplementary condition (and potential difficulty) induced by the hierarchical structural nature of trees.

To conclude, we can say that this work is a prelude and it would be interesting to study the possible properties that could arise when combining, in a manner to be specified yet, independent executions of the same uniform string distribution. This would tend to a string distributed according to a gaussian string distribution, as told by the central limit theorem. A necessary condition is to define the concept of random variable in a measurable vector space for strings. It could be sufficient to order the strings to keep the real line as the state set of such random variables, and thus take advantage of the classical (numerical) probabilistic statistical theory, but we could also develop a specific different point of view.

References

1. A. K. Jain, R. P. W. Duin, and J. Mao: *Statistical pattern recognition: a review.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(1) 2000, pp. 4–37.
2. J.-M. Jolion: *The deviation of a set of strings.* Pattern Analysis and Applications, 6(3) 2003, pp. 224–231.
3. T. Kohonen: *Median strings.* Pattern Recognition Letters, 3(5) 1985, pp. 309–313.
4. V. I. Levenshtein: *Binary codes capable of correcting deletions, insertions, and reversals.* Soviet Physics Doklady, 10(8) 1966, pp. 707–710.
5. S. Rebecchi and J.-M. Jolion: *Lois uniformes et normales de chanes discrtes*, in RFIA, Amiens, France, January 2008, pp. 471–480.
6. R. A. Wagner and M. J. Fischer: *The string-to-string correction problem.* Journal of the ACM, 21(1) 1974, pp. 168–173.