mathematics

Article

Retrieving a Context Tree from EEG Data

Aline Duarte [1], Ricardo Fraiman [2], Antonio Galves [1,*] and Guilherme Ost [3] and Claudia D. Vargas [4]

[1] Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo 05508-090, Brazil; alineduarte@usp.br

[2] Centro de Matemática, Universidad de la República, Uruguay and Instituto Pasteur de Montevideo, Montevideo 11400, Uruguay; rfraiman@cmat.edu.uy

[3] Instituto de Matemática, Universidade Federal do Rio de Janeiro, Rio de Janeiro 21941-909, Brazil; guilhermeost@im.ufrj.br

[4] Instituto de Biofísica, Universidade Federal do Rio de Janeiro, Rio de Janeiro 21941-902, Brazil; cdvargas@biof.ufrj.br

* Correspondence: galves@usp.br; Tel.: +551-11-30911712

Received: 28 March 2019; Accepted: 5 May 2019; Published: 14 May 2019

Abstract: It has been repeatedly conjectured that the brain retrieves statistical regularities from stimuli. Here, we present a new statistical approach allowing to address this conjecture. This approach is based on a new class of stochastic processes, namely, sequences of random objects driven by chains with memory of variable length.

Keywords: stochastic chains with memory of variable length; sequences of random objects driven by context tree models; stochastic modeling of EEG data

1. Introduction

Consider the following experimental situation. A listener is exposed to a sequence of auditory stimuli, generated by a stochastic chain, while electroencephalographic (EEG) signals are recorded from his scalp. Starting from Von Helmholtz [1], a classical conjecture in neurobiology claims that the listener's brain automatically identifies statistical regularities in the sequence of stimuli (see, for instance, [2,3]). If this is the case, then a signature of the stochastic chain generating the stimuli should somehow be encoded in the brain activity. The question is whether this signature can be identified in the EEG data recorded during the experiment. The goal of this paper is to discuss a new probabilistic framework in which this conjecture can be formally addressed.

To model the relationship between the random chain of auditory stimuli and the corresponding EEG data, we introduce a new class of stochastic processes. A process in this class has two components. The first one is a stochastic chain taking values in the set of auditory units. The second one is a sequence of functions corresponding to the sequence of EEG chunks recorded during the exposure of the successive auditory stimuli.

We use a stochastic chain with memory of variable length to model the dependence from the past characterizing the sequence of auditory stimuli. Stochastic chains with memory of variable length were introduced by Rissanen [4], as a universal system for data compression. In his seminal paper, Rissanen observed that in many real life stochastic chains the dependence from the past has not a fixed length. Instead, it changes at each step as a function of the past itself. He called a *context* the smallest final string of past symbols containing all the information required to predict the next symbol. The set of all contexts defines a partition of the past and can be represented by a rooted and labeled oriented tree. For this reason, many authors call stochastic chains with memory of variable length *context tree models*. We adopt this terminology here. A non-exhaustive list of articles on context tree models, with applications in biology and linguistics, includes [5–13].

An interesting point about stochastic chains with memory of variable length with finite context trees is that they are dense in the $\bar{d}$-topology in the class of chains of infinite order with continuous and non-null transition probabilities and summable continuity rates. This result follows easily from Fernández and Galves [14] and Duarte et al. [15]. We refer the reader to these articles for definitions and more details.

Besides modeling the chain of auditory units, we must also model the relationship between the chain of stimuli and the sequence of EEG chunks. To that end, we assume that at each time step a new EEG chunk is chosen according to a probability measure (defined on suitable class of functions) which depends only on the context assigned to the sequence of auditory units generated up to that time. In particular, this implies that to describe the new class of stochastic chains introduced in this paper, we also need to consider a family of probability measures on the set of functions corresponding to the EEG chunks, indexed by the contexts of the context tree characterizing the chain of auditory stimuli.

In this probabilistic framework, the neurobiological question can now be rigorously addressed as follows. Is it possible to retrieve the context tree characterizing the chain of stimuli from the corresponding EEG data? This is a problem of statistical model selection in the class of stochastic processes we have just informally described.

This article is organized as follows. In Section 2, we provide an informal overview of our approach. In Section 3, we introduce the notation, recall what is a *context tree model* and introduce the new class of *sequences of random objects driven by context tree models*. A statistical procedure to select, given the data, a member on the class of sequences of random objects driven by context tree models is presented in Section 4. The theoretical result supporting the proposed method, namely Theorem 1, is given in the same section. In Section 5, we conduct a brief simulation study to illustrate the statistical selection procedure presented in Section 4. The proof of Theorem 1 is given in Section 6.

2. Informal Presentation of Our Approach

Volunteers are exposed to sequences of auditory stimuli generated by a context tree models while EEG signals are recorded. The auditory units used as stimuli are *strong beats*, *weak beats* or *silent units*, represented by symbols $2, 1$ and 0, respectively.

The way the sequence of auditory units was generated can be informally described as follows. Start with the deterministic sequence

$$2\,1\,1\,2\,1\,1\,2\,1\,1\,2\,1\,1\,2\,\ldots.$$

Then, replace each weak beat (symbol 1) by a silent unit (symbol 0) with probability ϵ in an independent way.

An example of a sequence produced by this procedure acting on the basic sequence would be

$$2\,1\,1\,2\,0\,1\,2\,1\,1\,2\,0\,0\,2\,\ldots.$$

In the sequel, this stochastic chain is denoted by the symbols $(X_0, X_1, X_2, \ldots)$.

The stochastic chain just described can be generated step by step by an algorithm using only information from the past. We impose to the algorithm the condition that it uses, at each step, the shortest string of past symbols necessary to generate the next symbol.

This algorithm can be described as follows. To generate X_n, given the past $X_{n-1}, X_{n-2}, \ldots$, we first look to the last symbol X_{n-1}.

- If $X_{n-1} = 2$, then

$$X_n = \begin{cases} 1, & \text{with probability } 1-\epsilon, \\ 0, & \text{with probability } \epsilon. \end{cases}$$

- If $X_{n-1} = 1$ or $X_{n-1} = 0$, then we need to go back one more step,
  - if $X_{n-2} = 2$, then
$$X_n = \begin{cases} 1, & \text{with probability } 1 - \epsilon, \\ 0, & \text{with probability } \epsilon; \end{cases}$$
  - if $X_{n-2} = 1$ or $X_{n-2} = 0$, then $X_n = 2$ with probability 1.

The algorithm described above is characterized by two elements. The first one is a partition of the set of all possible sequences of past units. This partition is represented by the set

$$\tau = \{00, 10, 20, 2, 01, 11, 21, 2\}.$$

In partition τ, the string 00 represents the set of all strings ending by the ordered pair $(0, 0)$; 10 represents the set of all strings ending by the ordered pair $(1, 0)$, $\ldots$; and finally the symbol 2 represents the set of all strings ending by 2. Following Rissanen [4], let us call *context* any element of this partition.

For instance, if

$$\ldots, X_{n-3} = 1, X_{n-2} = 2, X_{n-1} = 0, X_n = 1.$$

the context associated to this past sequence is 01.

The partition τ of the past as described above can be represented by a rooted and labeled *tree* (see Figure 1) where each element of the partition is described as a leaf of the tree.

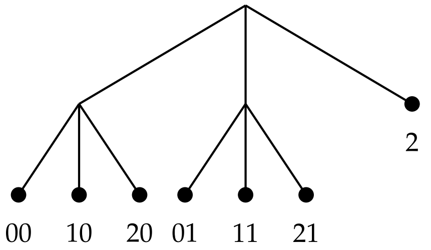

Figure 1. Graphical representation of the context tree τ.

In the construction described above, for each sequence of past symbols, the algorithm first identifies the corresponding context w in the partition τ. Once the context w is identified, the algorithm chooses a next symbol $a \in \{0, 1, 2\}$ using the transition probability $p(a|w)$. In other terms, each context w in τ defines a probability measure on $\{0, 1, 2\}$. The family of transition probabilities indexed by elements of the partition is the second element characterizing the algorithm.

The families of transition probabilities associated to τ are shown in Table 1.

Table 1. Transition probabilities associated to the context tree τ.

Context w	$\mathbf{p(0\|w)}$	$\mathbf{p(1\|w)}$	$\mathbf{p(2\|w)}$
2	ϵ	$1-\epsilon$	0
21	ϵ	$1-\epsilon$	0
20	ϵ	$1-\epsilon$	0
11	0	0	1
10	0	0	1
01	0	0	1
00	0	0	1

Using the notion of context tree, the neurobiological conjecture can now be rephrased as follows. Is the brain able to identify the context tree generating the sample of auditory stimuli? From an experimental point of view, the question is whether it is possible to retrieve the tree presented in Figure 1 from the corresponding EEG data. To deal with this question we introduce a new statistical model selection procedure described below.

Let Y_n be the chunk of EEG data recorded while the volunteer is exposed to the auditory stimulus X_n. Observe that Y_n is a continuous function taking values in $\mathbb{R}^d$, where $d \geq 1$ is the number of electrodes. Its domain is the time interval of length, say T, during which the acoustic stimulus X_n is presented.

The statistical procedure introduced in the paper can be informally described as follows. Given a sample $(X_0, Y_0), ..., (X_n, Y_n)$ of auditory stimuli and associated EEG chunks and for a suitable initial integer $k \geq 1$, do the following.

1. For each string $\mathbf{u} = u_1, u_2, ..., u_k$ of symbols in $\{0, 1, 2\}$, identify all occurrences in the sequence $X_0, X_1, ..., X_n$ of the string $a\mathbf{u}$, obtained by concatenating the symbol $a \in \{0, 1, 2\}$ and the string $\mathbf{u}$.
2. For each $a \in \{0, 1, 2\}$, define the subsample of all EEG chunks $Y_m = Y_m^{(a\mathbf{u})}$ such that $X_{m-k} = a, X_{m-k+1} = u_1, ..., X_m = u_k$ (see Figure 2).
3. For any pair $a, b \in \{0, 1, 2\}$, test the null hypothesis that the law of the EEG chunks $Y^{(a\mathbf{u})}$ and $Y^{(b\mathbf{u})}$ collected at Step 2 are equal.
   (a) If the null hypothesis is not rejected for any pair of final symbols a and b, we conclude that the occurrence of a or b before the string $\mathbf{u}$ do not affect the law of EEG chunks. Then, we start again the procedure with the one step shorter sequence $\mathbf{u} = u_2, ..., u_k$.
   (b) If the null hypothesis is rejected for at least one pair of final symbols a and b, we conclude that the law of EEG chunks depend on the entire string $a\mathbf{u}$ and we stop the pruning procedure.
4. We keep pruning the sequence $u_1, ..., u_k$ until the null-hypothesis is reject for the first time.
5. Call $\hat{\tau}_n$ the tree constituted by the strings which remained after the pruning procedure.

The question is whether $\hat{\tau}_n$ coincides with the context tree τ generating the sequence of auditory stimuli.

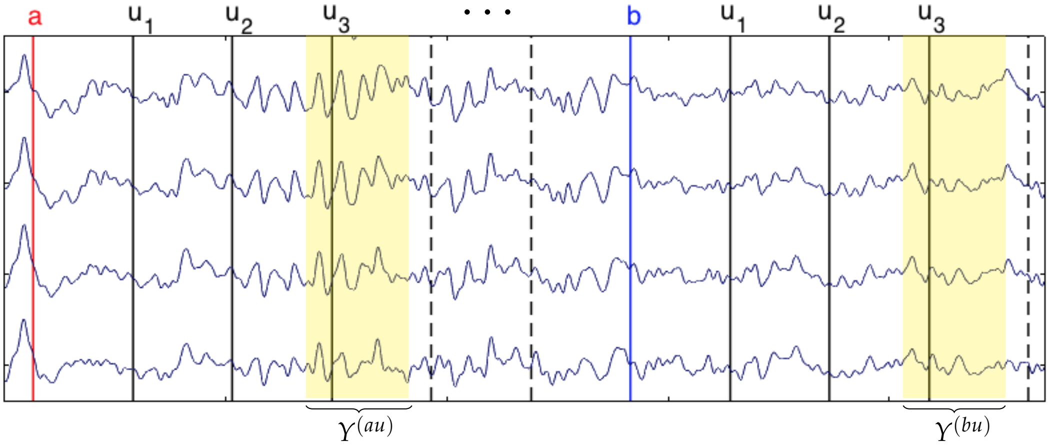

Figure 2. EEG signals recorded from four electrodes. The sequence of stimuli is indicated in the top horizontal line. Vertical lines indicate the beginning of the successive auditory units. The distance between two successive vertical lines is $T > 0$. Solid vertical lines indicate the successive occurrence times of the string $\mathbf{u}$. The first yellow strip corresponds to the chunk $Y_n^{(a\mathbf{u})}$ associated to the string $a\mathbf{u}$. The second yellow strip corresponds to the chunk $Y_n^{(b\mathbf{u})}$ associated to the string $b\mathbf{u}$.

An important technical issue must be clarified at this point, namely, how to test the equality of the laws of two subsamples of EEG chunks. This is done using the projective method informally explained below.

Suppose we have two samples of random functions, each sample composed by independent realizations of some fixed law. To test whether the two samples are generated by the same law, we choose at random a "direction" and project each function in the samples in this direction. This produces two new samples of real numbers. Now, we test whether the samples of the projected real numbers have the same law. Under suitable conditions, a theorem by Cuesta-Albertos et al. [16] ensures that for almost all directions if the test does not reject the null hypothesis that the projected samples have the same law, then the original samples also have the same law.

The arguments informally sketched in this section are formally developed in the subsequent sections.

3. Notation and Definitions

Let A be a finite alphabet. Given two integers $m, n \in \mathbb{Z}$ with $m \leq n$, the string $(u_m, \ldots, u_n)$ of symbols in A is often denoted by u_m^n; its length is $\ell(u_m^n) = n - m + 1$. The empty string is denoted by $\emptyset$ and its length is $\ell(\emptyset) = 0$. Fixing two strings u and v of elements of A, we denote by uv the string in $A^{\ell(u)+\ell(v)}$ obtained by the concatenation of u and v. By definition $u\emptyset = \emptyset u = u$ for any string $u \in A^{\ell(u)}$. The string u is said to be a *suffix* of v if there exists a string s satisfying $v = su$. This relation is denoted by $u \preceq v$. When $v \neq u$, we say that u is a *proper suffix* of v and write $u \prec v$. Hereafter, the set of all finite strings of symbols in A is denoted by $A^* := \bigcup_{k=1}^{\infty} A^k$. For any finite string $w = w_{-k}^{-1}$ with $k \geq 2$, we write $\mathrm{suf}(w)$ to denote the one-step shorter string w_{-k+1}^{-1}.

Definition 1. *A finite subset τ of A^* is a context tree if it satisfies the following conditions:*

1. *Suffix Property. For no $w \in \tau$ we have $u \in \tau$ with $u \prec w$.*
2. *Irreducibility. No string belonging to τ can be replaced by a proper suffix without violating the suffix property.*

The set τ can be identified with the set of leaves of a rooted tree with a finite set of labeled branches. The elements of τ are always denoted by $w, u, v, \ldots$.

The height of the context tree τ is defined as $\ell(\tau) = \max\{\ell(w) : w \in \tau\}$. In the present paper, we only consider context trees with finite height.

Definition 2. *Let τ and τ' be two context trees. We say that τ is smaller than τ' and write $\tau \preceq \tau'$, if for every $w' \in \tau'$ there exists $w \in \tau$ such that $w \preceq w'$.*

Given a context tree τ, let $p = \{p(\cdot \mid w) : w \in \tau\}$ be a family of probability measures on A indexed by the elements of τ. The pair (τ, p) is called a *probabilistic context tree* on A. Each element of τ is called a *context*. For any string $x_{-n}^{-1} \in A^n$ with $n \geq \ell(\tau)$, we write $c_\tau(x_{-n}^{-1})$ to denote the only context in τ which is a suffix of x_{-n}^{-1}.

Definition 3. *A probabilistic context tree (τ, p) with height $\ell(\tau) = k$ is irreducible if for any $a_{-k}^{-1} \in A^k$ and $b \in A$ there exist a positive integer $n = n(a_{-k}^{-1}, b)$ and symbols $a_0, a_1, \ldots, a_n = b \in A$ such that*

$$p(a_0|c_\tau(a_{-k}^{-1})) > 0, p(a_1|c_\tau(a_0a_{-k}^{-1})) > 0, \ldots, p(a_n|c_\tau(a_{n-1}, \ldots, a_0a_{-k}^{-1})) > 0.$$

Definition 4. *Let (τ, p) be a probabilistic context tree on A. A stochastic chain $(X_n)_{n\in\mathbb{N}}$ taking values in A is called a context tree model compatible with (τ, p) if*

1. *for any $n \geq \ell(\tau)$ and any finite string $x_{-n}^{-1} \in A^n$ such that $P(X_0^{n-1} = x_{-n}^{-1}) > 0$, it holds that*

$$P\Big(X_n = a \mid X_0^{n-1} = x_{-n}^{-1}\Big) = p\big(a \mid c_\tau(x_{-n}^{-1})\big) \text{ for all } a \in A, \tag{1}$$

where $c_\tau(x_{-n}^{-1})$ is the only context in τ which is a suffix of x_{-n}^{-1}.

2. *For any* $1 \leq j < \ell(c_\tau(x_{-n}^{-1}))$, *there exists* $a \in A$ *such that*

$$P\Big(X_n = a \mid X_0^{n-1} = x_{-n}^{-1}\Big) \neq P\Big(X_n = a \mid X_{n-j}^{n-1} = x_{-j}^{-1}\Big).$$

With this notation, we can now introduce the class of random objects driven by a context tree model.

Definition 5. *Let A be a finite alphabet, (τ, p) a probabilistic context tree on A, $(F, \mathcal{F})$ a measurable space and $(Q^w : w \in \tau)$ a family of probability measures on $(F, \mathcal{F})$. The bivariate stochastic chain $(X_n, Y_n)_{n\in\mathbb{N}}$ taking values in $A \times F$ is a sequence of random objects driven by a context tree model compatible with (τ, p) and $(Q^w : w \in \tau)$ if the following conditions are satisfied:*

1. $(X_n)_{n\in\mathbb{N}}$ *is a context tree model compatible with* (τ, p).
2. *The random elements $Y_0, Y_1, \ldots$ are $\mathcal{F}$-measurable. Moreover, for any integers $\ell(\tau) \leq m \leq n$, any string $x_{m-\ell(\tau)+1}^n \in A^{n-m+\ell(\tau)}$ and any sequence $J_m, \ldots, J_n$ of $\mathcal{F}$-measurable sets,*

$$P(Y_m \in J_m, \ldots, Y_n \in J_n | X_{m-\ell(\tau)+1}^n = x_{m-\ell(\tau)+1}^n) = \prod_{k=m}^{n} Q^{c_\tau(x_{k-\ell(\tau)+1}^k)}(J_k),$$

where $c_\tau(x_{k-\ell(\tau)+1}^k)$ is the context in τ assigned to the string of symbols $x_{k-\ell(\tau)+1}^k$.

Definition 6. *A sequence of random objects driven by a context tree model compatible with (τ, p) and $(Q^w : w \in \tau)$ is identifiable if for any context $w \in \tau$ there exists a context $u \in \tau$ such that suf(w)=suf(u) and $Q^w \neq Q^u$.*

The process (X_n) is called the *stimulus chain* and (Y_n) is called the *response chain*.

The experimental situation described in Section 2 can now be formally presented as follows.

- The stimulus chain (X_n) is a context tree model taking values in an alphabet having as elements symbols indicating the different types of auditory units appearing in the sequence of stimuli. We call (τ, p) its probabilistic context tree.
- Each element Y_n of the response chain (Y_n) represents the EEG chunk recorded while the volunteer is exposed to the auditory stimulus X_n. Thus, $Y_n = (Y_n(t), t \in [0, T])$ is a function taking values in $\mathbb{R}^d$, where T is the time distance between the onsets of two consecutive auditory stimuli and d the number of electrodes used in the analysis. The sample space F is the Hilbert space $L^2([0, T], \mathbb{R}^d)$ of $\mathbb{R}^d$-valued functions on $[0, T]$ having square integrable components. The Hilbert space F is endowed with its usual Borel σ-algebra $\mathcal{F}$.
- Finally, $(Q^w, w \in \tau)$ is a family of probability measures on $L^2([0, T], \mathbb{R}^d)$ describing the laws of the EEG chunks.

From now on, the pair $(F, \mathcal{F})$ always denotes the Hilbert space $L^2([0, T], \mathbb{R}^d)$ endowed with its usual Borel σ-algebra.

4. Statistical Selection for Sequences of Random Objects Driven by Context Tree Models

Let $(X_0, Y_0), \ldots, (X_n, Y_n)$, with $X_k \in A$ and $Y_k \in F$ for $0 \leq k \leq n$, be a sample produced by a sequence of random objects driven by a context tree model compatible with $(\bar{\tau}, \bar{p})$ and $(\bar{Q}^w : w \in \bar{\tau})$. Before introducing the statistical selection procedure, we need two more definitions.

Definition 7. *Let τ be a context tree and fix a finite string $s \in A^*$. We define the branch in τ induced by s as the set $B_\tau(s) = \{w \in \tau : w \succ s\}$. The set $B_\tau(s)$ is called a terminal branch if for all $w \in B_\tau(s)$ it holds that $w = as$ for some $a \in A$.*

Given a sample $X_0, \ldots, X_n$ of symbols in A and a finite string $u \in A^*$, the number of occurrences of u in the sample $X_0, \ldots, X_n$ is defined as

$$N_n(u) = \sum_{m=l(u)-1}^{n} 1\{X_{m-\ell(u)+1}^{m} = u\}.$$

Definition 8. *Given integers $n > L \geq 1$, an admissible context tree of maximal height L for the sample $X_0, \ldots, X_n$ of symbols in A, is any context tree τ satisfying*

1. *$w \in \tau$ if and only if $\ell(w) \leq L$ and $N_n(w) \geq 1$.*
2. *Any string $u \in A^*$ with $N_n(u) \geq 1$ is a suffix of some $w \in \tau$ or has a suffix $w \in \tau$.*

For any pair of integers $1 \leq L < n$ and any string $u \in A^*$ with $\ell(u) \leq L$, call $I_n(u)$ the set of indexes belonging to $\{\ell(u) - 1, \ldots, n\}$ in which the string u appears in sample $X_0, \ldots, X_n$, that is

$$I_n(u) = \{\ell(u) - 1 \leq m \leq n : X_{m-\ell(u)+1}^{m} = u\}.$$

Observe that by definition $|I_n(u)| = N_n(u)$. If $I_n(u) = \{m_1, \ldots, m_{N_n(u)}\}$, we set $Y_k^{(u)} = Y_{m_k}$ for each $1 \leq k \leq N_n(u)$. Thus, $Y_1^{(u)}, \ldots, Y_{N_n(u)}^{(u)}$ is the *subsample of* $Y_0, \ldots, Y_n$ *induced by the string* u.

Given $u \in A^*$ such that $N_n(u) \geq 1$ and $h \in F$, we define the empirical distribution associated to the projection of the sample $Y_1^{(u)}, \ldots, Y_{N_n(u)}^{(u)}$ onto the direction h as

$$\hat{Q}_n^{u,h}(t) = \frac{1}{N_n(u)} \sum_{m=1}^{N_n(u)} 1_{(-\infty,t]}(\langle Y_m^{(u)}, h\rangle),\ t \in \mathbb{R},$$

where for any pair of functions $f, h \in F$,

$$\langle f, h\rangle = \sum_{i=1}^{d} \int_0^T f_i(t)h_i(t)dt.$$

For a given pair $u, v \in A^*$, with $\max\{\ell(u), \ell(v)\} \leq L$ and $h \in F$, the Kolmogorov–Smirnov distance between the empirical distributions $\hat{Q}_n^{u,h}$ and $\hat{Q}_n^{v,h}$ is defined by

$$\mathrm{KS}(\hat{Q}_n^{u,h}, \hat{Q}_n^{v,h}) = \sup_{t \in \mathbb{R}} |\hat{Q}_n^{u,h}(t) - \hat{Q}_n^{v,h}(t)|.$$

Finally, we define for any pair $u, v \in A^*$ such that $\max\{\ell(u), \ell(v)\} \leq L$ and $h \in F$,

$$D_n^h((Y_1^{(u)}, \ldots, Y_{N_n(u)}^{(u)}), (Y_1^{(v)}, \ldots, Y_{N_n(v)}^{(v)})) = \sqrt{\frac{N_n(u)N_n(v)}{N_n(u) + N_n(v)}}\mathrm{KS}(\hat{Q}_n^{u,h}, \hat{Q}_n^{v,h}).$$

Our selection procedure can now be described as follows. Fix an integer $1 \leq L < n$ and let $\mathcal{T}_n$ be the largest admissible context tree of maximal height L for the sample $X_0, \ldots, X_n$. The largest means that if τ is any other admissible context tree of maximal height L for the sample X_1^n, then $\tau \preceq \mathcal{T}_n$.

For any string $u \in A^*$ such that $B_{\mathcal{T}_n}(u)$ is a terminal branch, we test the null hypothesis

$$H_0^{(u)}\colon \mathcal{L}\big(Y_1^{(au)}, \ldots, Y_{N_n(au)}^{(au)}\big) = \mathcal{L}\big(Y_1^{(bu)}, \ldots, Y_{N_n(bu)}^{(bu)}\big),\ \forall\, au, bu \in B_{\mathcal{T}_n}(u) \tag{2}$$

using the test statistic

$$\Delta_n(u) = \Delta_n^W(u) = \max_{a,b \in A} D_n^W\big((Y_1^{(au)}, \ldots, Y_{N_n(au)}^{(au)}), (Y_1^{(bu)}, \ldots, Y_{N_n(bu)}^{(bu)})\big), \tag{3}$$

where $W = ((W_1(t), \ldots, W_d(t)) : t \in [0, T])$ is a realization of a d-dimensional Brownian motion in $[0, T]$.

We reject the null hypothesis $H_0^{(u)}$ when $\Delta_n(u) > c$, where $c > 0$ is a suitable threshold. When the null hypothesis $H_0^{(u)}$ is not rejected, we prune the branch $B_{\mathcal{T}_n}(u)$ in $\mathcal{T}_n$ and set as a new candidate context tree

$$\mathcal{T}_n = (\mathcal{T}_n \setminus B_{\mathcal{T}_n}(u)) \cup \{u\}.$$

On the other hand, if the null hypothesis $H_0^{(u)}$ is rejected, we keep $B_{\mathcal{T}_n}(u)$ in $\mathcal{T}_n$ and stop testing $H_0^{(s)}$ for strings $s \in A^*$ such that $s \preceq u$.

In each pruning step, take a string $s \in A^*$ that induces a terminal branch in $\mathcal{T}_n$ and has not been tested yet. This pruning procedure is repeated until no more pruning is performed. We denote by $\hat{\tau}_n$ the final context tree obtained by this procedure. The formal description of the above pruning procedure is provided in Algorithm 1 as pseudocode.

Algorithm 1 Pseudocode describing the pruning procedure used to select the tree $\hat{\tau}_n$.

```
Input: A sample $(X_0, Y_0), \ldots, (X_n, Y_n)$ with $X_k \in A$ and $Y_k \in F$ for $0 \leq k \leq n$, a positive threshold $c$
    and a positive integer $L$.
Output: A tree $\hat{\tau}_n$
$\tau \leftarrow \mathcal{T}_n$
Flag$(s) \leftarrow$ "not visited" for all string $s$ such that $s \preceq w \in \mathcal{T}_n$
for k in L to 1 do
    while $\exists s \in \tau$: $\ell(s) = k$, Flag$(s)$ = "not visited" and $B_\tau(s)$ is a terminal branch do
        Choose a $s$ such that $\ell(s) = k$, Flag$(s)$ = "not visited" and $B_\tau(s)$ is a terminal branch
        Compute the test statistic $\Delta_n(s)$ to test $H_0^{(s)}$
        if $\Delta_n(s) > c$ then
            Flag$(u) \leftarrow$ "visited" $\forall u \preceq s$
        else
            $\tau \leftarrow (\tau \setminus B_\tau(s)) \cup \{s\}$
        end if
    end while
end for
Return $\hat{\tau}_n = \tau$.
```

To state the consistency theorem, we need the following definitions.

Definition 9. *A probability measure P defined on $(F, \mathcal{F})$ satisfies Carleman condition if all the absolute moments $m_k = \int ||h||^k P(dh)$, $k \geq 1$, are finite and*

$$\sum_{k \geq 1} m_k^{-1/k} = +\infty.$$

Definition 10. *Let P be a probability measure on $(F, \mathcal{F})$. We say that P is* continuous *if P^h is continuous for any $h \in F$, where P^h is defined by*

$$P^h((-\infty, t]) = P(x \in F : \langle x, h \rangle \leq t), \; t \in \mathbb{R}.$$

Let V be a finite set of indexes and $(P_i : i \in V)$ be a family of probability measures on $(F, \mathcal{F})$. We say that $(P_i : i \in V)$ is continuous if for all $i \in V$, the probability measure P_i is continuous.

In what follows, let $c_\alpha = \sqrt{(1/2)\ln(2/\alpha)}$, where $\alpha \in (0, 1)$. We say that $\alpha_n \to 0$ *slowly enough* as $n \to \infty$ if

$$\frac{\sqrt{n}}{c_{\alpha_n}} \to \infty \text{ as } n \to \infty.$$

Theorem 1. *Let $(X_0, Y_0), \ldots, (X_n, Y_n)$ be a sample produced by a identifiable sequence of random objects driven by a context tree model compatible with $(\bar{\tau}, \bar{p})$ and $(\bar{Q}_w : w \in \bar{\tau})$, and let $\hat{\tau}_n$ be the context tree selected from the sample by Algorithm 1 with $L \geq \ell(\bar{\tau})$ and threshold $c_{\alpha_n} = \sqrt{(1/2)\ln(2/\alpha_n)}$, where $\alpha_n \in (0,1)$. If $(\bar{\tau}, \bar{p})$ is irreducible and $(\bar{Q}_w : w \in \bar{\tau})$ is continuous and satisfies Carleman condition, then for $\alpha_n \to 0$ slowly enough as $n \to \infty$,*

$$\lim_{n\to\infty} P(\hat{\tau}_n \neq \bar{\tau}) = 0.$$

The proof of Theorem 1 is presented in Section 6.

5. Simulation Study

In this section, we illustrate the performance of Algorithm 1 by applying it in a toy example. We consider the context tree model compatible with $(\bar{\tau}, \bar{p})$ described in Section 2 with $\epsilon = 0.2$. For each $w \in \bar{\tau}$, we assume $\bar{Q}^w$ is the law of a diffusion process with drift coefficient $f_w = (f_w(t))_{0\leq t\leq 1}$ and constant diffusion coefficient. For simplicity, all diffusion coefficients are assumed to be 1. For each context $w \in \bar{\tau}$, we assume $f_w = Kg_w$, where K is a positive constant and g_w is the density of a Gaussian random variable with mean μ_w and standard deviation σ_w, restricted to the interval $[0,1]$. In the simulation, we take $K = 5$. The shapes of the functions f_w and corresponding values of μ_w and σ_w are shown in Figure 3. One can check that the assumptions of Theorem 1 are satisfied by this toy example.

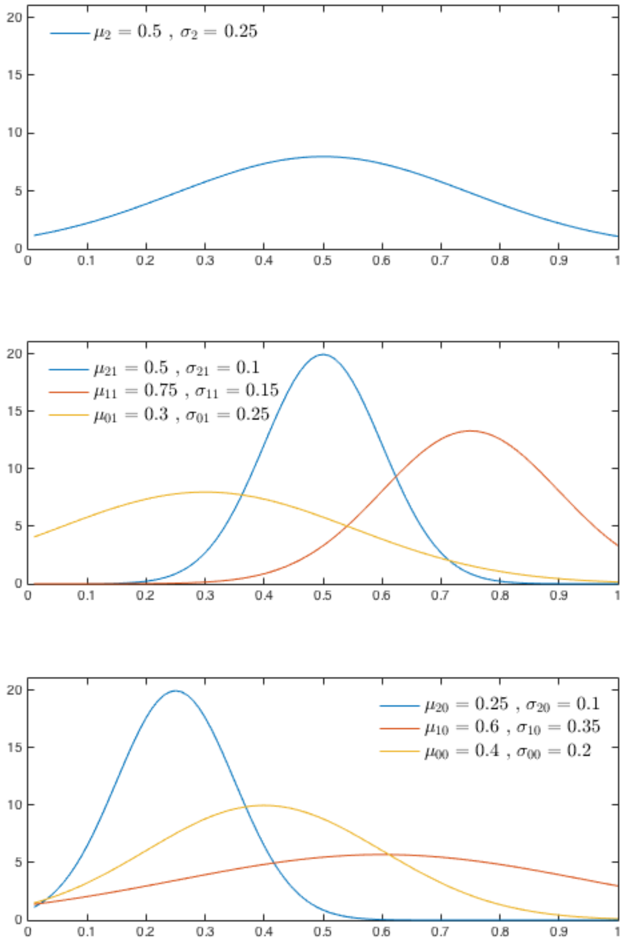

Figure 3. Functions g_w and the corresponding values of μ_w and σ_w for $w \in \tau = \{2, 11, 21, 01, 00, 10, 20\}$: (**top**) function g_2; (**middle**) functions g_{21}, g_{11} and g_{01}; and (**bottom**) functions g_{20}, g_{10} and g_{00}.

To numerically implement Algorithm 1, we assume that all trajectories of the diffusion processes are observed on equally spaces point $0 = t_0 < t_1 < \ldots < t_{100} = 1$, where $t_i = \frac{i}{100}$ for each $1 \leq i \leq 100$. For each sample size $n = 100, 120, 140, \ldots, 1000$, we estimate the fraction of times Algorithm 1, with $\alpha_n = 1/n$ and $L = 4$, correctly identifies the context tree $\bar{\tau}$ based on 100 random samples of the model with size n. The results are reported in Figure 4.

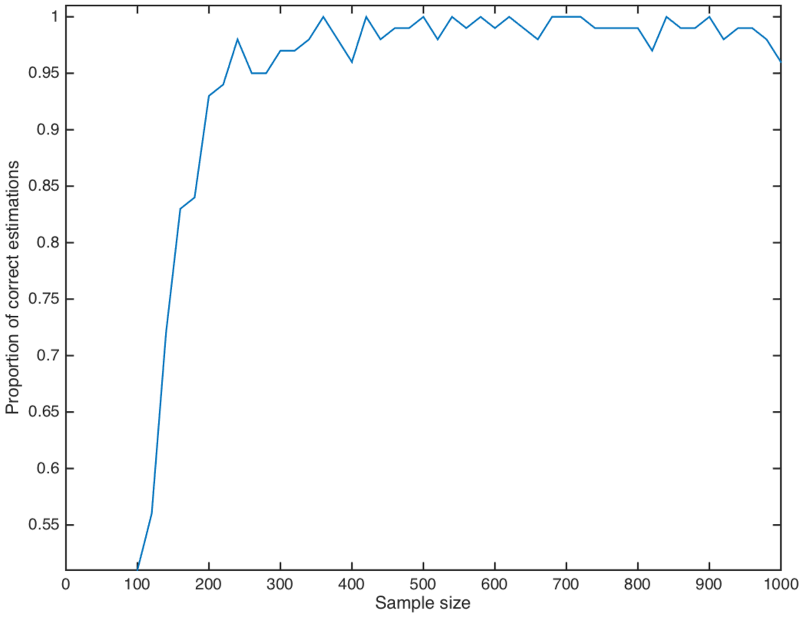

Figure 4. Proportion of correct identification of the context tree $\bar{\tau} = \{2, 01, 11, 21, 20, 10, 00\}$ by applying Algorithm 1 to simulated data with sample sizes $n = 100, 120, 140, \ldots, 1000$. For sample sizes larger than 200, the proportion of correct identification is at least 95%.

6. Proof of Theorem 1

The proof of Theorem 1 is a direct consequence of Propositions 1 and 2 presented below.

Proposition 1. *Let $(X_0, Y_0), \ldots, (X_n, Y_n)$ be a sample produced by a sequence of random objects driven by a context tree model compatible with $(\bar{\tau}, \bar{p})$ and $(\bar{Q}_w : w \in \bar{\tau})$ satisfying the assumptions of Theorem 1. Let $\alpha \in (0, 1)$ and set $c_\alpha = \sqrt{1/2 \ln(2/\alpha)}$. For any integer $L \geq \ell(\bar{\tau})$, context $w \in \bar{\tau}$, direction $h \in F \setminus \{0\}$, and strings $u, v \in \cup_{k=1}^{L-\ell(w)} A^k$ such that $w \preceq u$ and $w \preceq v$, it holds that*

$$\lim_{n\to\infty} P(D_n^h((Y_1^{(u)}, \ldots, Y_{N_n(u)}^{(u)}), (Y_1^{(v)}, \ldots, Y_{N_n(v)}^{(v)})) > c_\alpha) = \alpha.$$

In particular, for any $\alpha_n \to 0$ as $n \to \infty$, we have

$$\lim_{n\to\infty} P(D_n^h((Y_1^{(u)}, \ldots, Y_{N_n(u)}^{(u)}), (Y_1^{(v)}, \ldots, Y_{N_n(v)}^{(v)})) > c_{\alpha_n}) = 0.$$

Proof. The irreducibility of $(\bar{\tau}, \bar{p})$ implies that P-a.s. both $N_n(u)$ and $N_n(v)$ tend to $+\infty$ as n diverges. Thus, Theorem 3.1(a) of [16] implies that the law of $D_n^h((Y_1^{(u)}, \ldots, Y_{N_n(u)}^{(u)}), (Y_1^{(v)}, \ldots, Y_{N_n(v)}^{(v)}))$ is independent of the strings u and v, and also of the direction $h \in F \setminus \{0\}$. It also implies that $D_n^h((Y_1^{(u)}, \ldots, Y_{N_n(u)}^{(u)}), (Y_1^{(v)}, \ldots, Y_{N_n(v)}^{(v)}))$ converges in distribution to $K = \sup_{t\in[0,1]} |B(t)|$ as $n \to \infty$, where $B = (B(t) : t \in [0, 1])$ is a Brownian Bridge. Since $P(K > c_\alpha) = \alpha$, the first part of the result follows.

By the first part of the proof, for any fixed $\alpha \in (0,1)$, we have that for all n large enough,

$$P(D_n^h((Y_1^{(u)},\ldots,Y_{N_n(u)}^{(u)}),(Y_1^{(v)},\ldots,Y_{N_n(v)}^{(v)})) > c_\alpha)) \leq 2\alpha.$$

Thus, given $\epsilon > 0$, take $\alpha \in (0,1)$ such that $2\alpha < \epsilon$ to deduce that for all n large enough,

$$P(D_n^h((Y_1^{(u)},\ldots,Y_{N_n(u)}^{(u)}),(Y_1^{(v)},\ldots,Y_{N_n(v)}^{(v)})) > c_\alpha)) < \epsilon.$$

Since $c_{\alpha_n} \to \infty$ as $n \to \infty$, we have that for all n sufficiently large $c_{\alpha_n} > c_\alpha$ so that the result follows from the previous inequality. □

Proposition 2 reads as follows.

Proposition 2. *Let $(X_0,Y_0),\ldots,(X_n,Y_n)$ be a sample produced by a identifiable sequence of random objects driven by a context tree model compatible with $(\bar\tau,\bar p)$ and $(\bar Q_w : w \in \bar\tau)$, and let $\hat\tau_n$ satisfying the assumptions of Theorem 1. Let $\alpha \in (0,1)$ and define $c_\alpha = \sqrt{1/2\ln(2/\alpha)}$. For any string $s \in A^*$ such that $B_{\bar\tau}(s)$ is a terminal branch there exists a pair $w,w' \in B_{\bar\tau}(s)$ such that for almost all realization of a Brownian motion $W = (W(t) : t \in [0,T])$ on $[0,T]$,*

$$\lim_{n\to\infty} P(D_n^W((Y_1^{(w)},\ldots,Y_{N_n(w)}^{(w)}),(Y_1^{(w')},\ldots,Y_{N_n(w')}^{(w')})) \leq c_{\alpha_n}) = 0,$$

whenever $\alpha_n \to 0$ slowly enough as $n \to \infty$.

Proof. Since the sequence of random objects $(X_0,Y_0),(X_1,Y_1),\ldots$ is identifiable and $B_{\bar\tau}(s)$ is a terminal branch, there exists a pair $w,w' \in B_{\bar\tau}(\mathrm{suf}(w))$ whose associated distributions $\bar Q^w$ and $\bar Q^w$ on F are different, and both $\bar Q^w$ and $\bar Q^{w'}$ satisfy the Carleman condition. For each $n \geq 1$, define

$$N_n := \sqrt{\frac{N_n(w)N_n(w')}{N_n(w)+N_n(w')}},$$

if $\min\{N_n(w),N_n(w')\} \geq 1$. Otherwise, we set $N_n = 0$. The irreducibility of $(\bar\tau,\bar p)$ implies that $n^{-1/2}N_n \to C$ as $n \to \infty$ P-a.s., where C is a positive constant depending on w and w'.

Now, Theorem 3.1(b) of [16] implies that, for almost all realization of a Brownian motion W on F,

$$\liminf_{n\to\infty} \mathrm{KS}(\hat Q_n^{W,w},\hat Q_n^{W,w'}) > 0 \; P\text{-a.s.} \tag{4}$$

Since $D^W((Y_1^{(w)},\ldots,Y_{N_n(w)}^{(w)}),(Y_1^{(w')},\ldots,Y_{N_n(w')}^{(w')})/c_{\alpha_n} = \frac{\sqrt{n}}{c_{\alpha n}}\frac{N_n}{\sqrt{n}}\mathrm{KS}(\hat Q_n^{h,w},\hat Q_n^{h,w'})$ and $\alpha_n \to 0$ slowly enough, the result follows. □

Proof of Theorem 1. Let $C_{\bar\tau}$ be the set of contexts belonging to a terminal branch of $\bar\tau$. Define also the following events

$$U_n = \bigcup_{w\in C_{\bar\tau}} \{\Delta_n^W(\mathrm{suf}(w)) \leq c_{\alpha_n}\} \text{ and } O_n = \bigcup_{w\in\bar\tau} \bigcup_{\substack{s\succ w:\\ \ell(s)\leq L}} \{\Delta_n^W(s) > c_{\alpha_n}\}.$$

It follows from the definition of Algorithm 1 that

$$P(\hat\tau_n \neq \bar\tau) = P(U_n) + P(O_n).$$

Thus, it is enough to prove that for any $\epsilon > 0$ there exists $n_0 = n_0(\epsilon)$ such that $P(U_n) \leq \epsilon/2$ and $P(O_n) \leq \epsilon/2$ for all $n \geq n_0$.

By the union bound, we see that

$$P(U_n) \leq \sum_{w \in \bar{\tau}} P(\Delta_n^W(\text{suf}(w)) \leq c_{\alpha_n}). \tag{5}$$

The sequence of random objects $(X_0, Y_0), (X_1, Y_1), \ldots$ is identifiable. Thus by observing that for each $w \in C_{\bar{\tau}}$, $B_{\bar{\tau}}(\text{suf}(w))$ is a terminal branch, we have that there exists $w' \in B_{\bar{\tau}}(\text{suf}(w))$ such that the associated distributions $\bar{Q}^w$ and $\bar{Q}^{w'}$ on F are different, and both $\bar{Q}^w$ and $\bar{Q}^{w'}$ satisfies Carleman condition. Since

$$\{\Delta_n^W(\text{suf}(w)) \leq c_{\alpha_n}\} \subset \{D_n^W((Y_1^{(w)}, \ldots, Y_{N_n(w)}^{(w)}), (Y_1^{(w')}, \ldots, Y_{N_n(w')}^{(w')}) \leq c_{\alpha_n}\},$$

and $\bar{\tau}$ is finite, Proposition 2 implies that $P(U_n) \to 0$ as $n \to \infty$, if $\alpha_n \to 0$ slowly enough. As a consequence, for any $\epsilon > 0$ there exists $n_0 = n_0(\epsilon)$ such that $P(U_n) \leq \epsilon/2$ for all $n \geq n_0$.

Using again the union bound, we have

$$P(O_n) \leq \sum_{w \in \bar{\tau}} \sum_{\substack{s \succ w: \\ \ell(s) \leq L}} P(\Delta_n^W(s) > c_{\alpha_n}). \tag{6}$$

By observing that $\bar{\tau}$ is finite, the alphabet A is finite and

$$\{\Delta_n^W(s) > c_{\alpha_n}\} = \bigcup_{a,b \in A} \{D_n^W((Y_1^{(as)}, \ldots, Y_{N_n(as)}^{(as)}), (Y_1^{(bs)}, \ldots, Y_{N_n(bs)}^{(bs)}) > c_{\alpha_n}\},$$

we deduce from Proposition 1 and the inequality in Equation (6) that, for any $\epsilon > 0$, we have $P(O_n) \leq \epsilon/2$ for all n large enough. This concludes the proof of the theorem. □

Author Contributions: All authors contributed equally to this work.

Funding: This work is part of USP project *Mathematics, computation, language and the brain*, FAPESP project *Research, Innovation and Dissemination Center for Neuromathematics* (grant 2013/07699-0), and project *Plasticity in the brain after a brachial plexus lesion* (FAPERJ grants E26/010002902/2014, E26/010002474/2016). A. Galves and C.D. Vargas are partially supported by CNPq fellowships (grants 311 719/2016-3 and 309560/2017-9, respectively). C.D. Vargas is also partially supported by a FAPERJ fellowship (CNE 202.785/2018). A. Duarte was fully and successively supported by CNPq and FAPESP fellowships (grants 201696/2015-0 and 2016/17791-9). G. Ost was fully and successively supported by CNPq and FAPESP fellowships (grants 201572/2015-0 and 2016/17789-4).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Von Helmholtz, H. *Handbuch der Physiologischen Optik*; Translated by The Optical Society of America in 1924 from the third germand edition, 1910, Treatise on physiological optics, Volume III; Leopold Voss: Leipzig, Germany, 1867; Volume 3.
2. Garrido, M.I.; Sahani, M.; Dolan, R.J. Outlier responses reflect sensitivity to statistical structure in the human brain. *PLOS Comput. Biol.* **2013**, *9*. [CrossRef] [PubMed]
3. Wacongne, C.; Changeux, J.; Dehaene, S. A Neuronal Model of Predictive Coding Accounting for the Mismatch Negativity. *J. Neurosci.* **2012**, *32*, 3665–3678. [CrossRef] [PubMed]
4. Rissanen, J. A Universal Data Compression System. *IEEE Trans. Inf. Theory* **1983**, *29*, 656–664. [CrossRef]
5. Bühlmann, P.; Wyner, A.J. Variable length Markov chains. *Ann. Stat.* **1999**, *27*, 480–513. [CrossRef]
6. Csiszár, I.; Talata, Z. Context tree estimation for not necessarily finite memory processes, Via BIC and MDL. *IEEE Trans. Inf. Theory* **2006**, *52*, 1007–1016. [CrossRef]
7. Leonardi, F.G. A generalization of the PST algorithm: modeling the sparse nature of protein sequences. *Bioinformatics* **2006**, *22*, 1302–1307. [CrossRef] [PubMed]
8. Galves, A.; Löcherbach, E. Stochastic chains with memory of variable length. *TICSP Ser.* **2008**, *38*, 117–133.

9. Garivier, A.; Leonardi, F. Context tree selection: A unifying view. *Stoch. Processes Appl.* **2011**, *121*, 2488–2506. [CrossRef]
10. Gallo, S. Chains with unbounded variable length memory: perfect simulation and a visible regeneration scheme. *Adv. Appl. Probab.* **2011**, *43*, 735–759. [CrossRef]
11. Galves, A.; Galves, C.; García, J.E.; Garcia, N.L.; Leonardi, F. Context tree selection and linguistic rhythm retrieval from written texts. *Ann. Appl. Stat.* **2012**, *6*, 186–209. [CrossRef]
12. Galves, A.; Garivier, A.; Gassiat, E. Joint Estimation of Intersecting Context Tree Models. *Scand. J. Stat.* **2013**, *40*, 344–362. [CrossRef]
13. Belloni, A.; Oliveira, R.I. Approximate group context tree. *Ann. Stat.* **2017**, *45*, 355–385. [CrossRef]
14. Fernández, R.; Galves, A. Markov approximations of chains of infinite order. *Bull. Braz. Math. Soc.* **2002**, *33*, 295–306. [CrossRef]
15. Duarte, D.; Galves, A.; Garcia, N.L. Markov approximation and consistent estimation of unbounded probabilistic suffix trees. *Bull. Braz. Math. Soc.* **2006**, *37*, 581–592. [CrossRef]
16. Cuesta-Albertos, J.A.; Fraiman, R.; Ransford, T. Random projections and goodness-of-fit tests in infinite-dimensional spaces. *Bull. Braz. Math. Soc. New Ser.* **2006**, *37*, 477–501. [CrossRef]