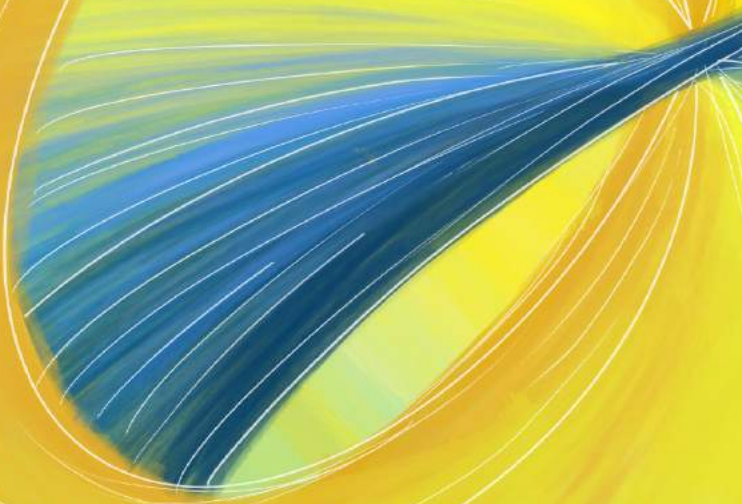
Statistical model selection for stochastic systems with applications to Bioinformatics, Linguistics and Neurobiology



Antonio Galves
Florencia Leonardi
Guilherme Ost

impa

33º Colóquio Brasileiro de Matemática

Statistical model selection for stochastic systems with applications to Bioinformatics, Linguistics and Neurobiology

Statistical model selection for stochastic systems with applications to Bioinformatics, Linguistics and Neurobiology

Primeira impressão, julho de 2021

Publicado no Brasil / Published in Brazil.

ISBN 978-65-89124-29-0

MSC (2020) Primary: 62M09, Secondary: 62M20, 60G17, 60J05, 62M05, 60J20

Coordenação Geral Carolina Araujo

Produção Books in Bytes **Capa** Izabella Freitas & Jack Salvador

Realização da Editora do IMPA

IMPA
Estrada Dona Castorina, 110
Jardim Botânico
22460-320 Rio de Janeiro RJ

www.impa.br
editora@impa.br

Contents

1 Introduction

These lecture notes present new results on statistical model selection for stochastic systems. The majority of the results are original and first appeared in several recent papers co-authored by us. They all share a common feature; they propose a new conceptual framework to assign appropriate models to specific samples of scientific data, displaying non-trivial interactions in time and space.

The papers at the origin of these notes found their primary motivation in problems and data coming from linguistics or biology. Despite their original specific inspiration, we believe that the models and statistical procedures presented here can be applied to a large variety of data sets produced from different scientific disciplines, representing time evolutions with structural time and space dependencies. This belief justifies the existence of the present book.

The models and statistical procedures discussed here are attractive from an applied point of view, but not only. They are also interesting from a purely theoretical point of view, as mathematical objects. All the results presented here have been rigorously proved, and these proofs are presented in the book. However, this rigour should not scare applied researchers. These notes are written so that the models, statistical procedures and results are presented intuitively. Proofs appear only in a separate section at the end of the chapters. They are there to be read by those interested in the technical details related to the theoretical properties of

models and procedures. Those who are only interested in applications can skip the proofs.

Let us now summarise the goal content of this book. Let us start by discussing the meaning of the title: statistical model selection for stochastic systems.

What is statistical model selection?
Statistical model selection is a domain of Statistics. It refers to a crucial issue, namely, how to assign models to samples of experimental data.

What is a model?
A model is a description of a procedure that can generate samples with the same statistical features displayed by the sample of experimental data we are analysing. By procedure, we mean, for instance, a computational algorithm able to generate a piece of data.

For example, suppose the sample is a string of symbols. In that case, a possible model is a computational algorithm, producing sequentially the symbols, one by one, by taking into account, at each step, the last symbols already generated.

A naive model could, for instance, assume that each next symbol is produced independently of the string of past symbols. Or it could assume that each next symbol depends only on the last symbol already generated. This class of models was introduced by the Russian mathematician Andrey Andreyvich Markov in 1913 to model the occurrence of consonants and vowels in Pushkin's poem Eugene Oneguin (Markov 2006).

We could generalise Markov's original assumption and assume that each next symbol depends on the last k symbols, where k is a fixed integer greater or equal 1. More recently, in 1983, the Finish information computer scientist Jorma Rissanen observed that typically strings of symbols produced by scientific experiments have a dependence from the past, which is not fixed but has a length which is a function of the past itself. This leads Rissanen to introduce what was later called the class of *chains with memory of variable length*.

What statistics has to do with this?
The intrinsic randomness of typical samples of scientific data makes it unavoidable to use statistical criteria to select a model. In other words, we do not look for a procedure that generates a sample identical to the original sample of scientific data; instead, we look for a procedure that can generate samples with the same statistical features as those displayed by the sample of experimental data.

What is the motivation of this quest for models?
In 1867, the physicist von Helmholtz observed that the human brain does statistical model selection all the time, by making hypotheses and assigning models to sequences of stimuli to be able to make predictions about what will occur in the near future. This neurobiological ability was called *unconscious inference* by von Helmholtz.

Assigning models to stimuli to make predictions about the future is crucial to make good choices in real life, in all kinds of situations, from driving a car without touching or being touched by other vehicles to simply surviving in a hostile environment.

Less dramatically, other examples of the need for statistical model selection in real life include making reliable predictions about the stock market's time evolution, the weather, the options of a set of voters, etc.

In computer science, assigning models to strings of bits is necessary to compress data. Medical diagnostic imaging is essentially a matter of statistical model selection. More generally, in all branches of science, assigning models to data samples is necessary to understand the structure and typical features of samples of scientific data.

What are the classes of models that we consider in this lecture notes?
In this book, we consider mainly two classes of stochastic systems. First of all, the class of stochastic chains with memory of variable length, introduced by Jorma Rissanen in his 1983 seminal paper: A universal system for data compression (Rissanen 1983). The paper's title refers to the fact that models in this class are dense in the class of chains with memory of infinite order. From an applied point of view, these models are attractive because they are flexible enough to recover essential patterns in the processes. At the same time, they can be economical in the number of degrees of freedom, giving a good balance between the goodness of fit and the complexity of the final model. The second model considered in the book is interacting systems of point processes with memory of variable length and, in particular, systems of interacting chains with memory of infinite length. They extend Rissanen's' ideas to systems with space-time interactions, which are required to deal with medical imagery, multiunit records of neuronal activity and samples representing systems with many components interacting in time and space. From a mathematical point of view, this class of systems extends in a non-trivial way the class of interacting Markov systems introduced by Spitzer (Spitzer 1970). .

Is this a course in Probability Theory or is this a course in Statistics?

In this book, we introduce probabilistic models, which are interesting mathematical objects by themselves. We also discuss how these mathematical models can be used to model sets of scientific data.

To apply the models to data analysis is necessary to study rigorously the properties of the algorithms used to select the model which best fits the data. This requires proving theorems that are mathematically challenging and technically difficult.

Besides discussing the rigorous mathematical framework required to make statistical analysis with these models, we also face the challenge of analysing real scientific data, with samples and scientific questions coming from linguistics, proteomics and neurobiology.

Is this course related to Data Science?

The answer is clearly: yes! Data Science's goal is to assign models to huge sets of data to make predictions, putting together data with the same type of features. It turns out that identifying essential features in the data is rarely a task that can be solved by naive "visual inspection". Accurate predictions require identifying a model able to generate samples with the same statistical features as those displayed by the original data set.

A naive point of view that considers that Data science requires only computational power will only produce superficial and non-interesting results. To be successful, data science requires developing new classes of stochastic systems and new statistical selection procedures. This is precisely the goal of this book.

By the way, one of the articles that we discuss in the book, Galves, Galves, et al. (2012), received in 2020 the Johannes Kepler award discerned for the first time by the SBMAC, the Brazilian Society for Applied and Computational Mathematics. The award's name comes from the fact that Johannes Kepler can be considered the first data scientist in history.

So the answer is yes. This book is related to Data Science. We hope that it will be useful for young researchers interested in the stochastic modelling of very large samples of complex data.

Acknowledgements

Each chapter of this book starts by mentioning the original papers co-authored by us, where the models, procedures and results discussed here were first presented. So it is just fair to conclude this introduction by naming and thanking all of our co-authors. They are Ludmila Brochini, Aline Duarte, Charlotte Galves, Jesus Enrique Garcia, Pierre Hodara, Aurélien Garivier, Eva Löcherbach, Nancy Lopes Garcia, Christophe Pouzat and Patricia Reynaud-Bouret.

This work is part of USP project *Mathematics, computation, language and the brain* and FAPESP projects *Research, Innovation and Dissemination Center for Neuromathematics* (grant 2013/07699-0), *Model selection in high dimensions: theoretical properties and applications* (grant 2019/17734-3) and *Stochastic Modeling of Interacting Systems* (grant 2017/10555-0). AG is partially supported by CNPq fellowship (grant 309501/2011-3). FL is partially supported by CNPq fellowship (grant 311763/2020-0). GO is partially supported by FAPERJ fellowships (grant E-26/211.343/2019 and E-26/201.397/2021).

2 Stochastic chains with memory of variable length

In this chapter we introduce the main definitions concerning stochastic chains with memory of variable length. We also describe the main algorithms in the literature to estimate the parameters and the structure of the context tree associated to the model. The material in this chapter is based mainly on the articles Galves, Galves, et al. (2012), Galves and Leonardi (2008), Garivier and Leonardi (2011), and Leonardi (2010).

2.1 Model definition

The idea behind the notion of stochastic chains with memory of variable length is that the probabilistic definition of each symbol only depends on a finite part of the past and the length of this relevant portion is a function of the past itself. The minimal relevant part of each past is called *context*. The set of all contexts satisfies the suffix property which means that no context is a proper suffix of another context. This property allows to represent the set of all contexts as a rooted labeled tree. With this representation the process is described by the tree of all contexts and an

associated family of probability measures, indexed by the tree of contexts. Given a context, its associated probability measure gives the probability of the next symbol for any past having this context as a suffix. In the sequel we put these ideas in formal terms.

2.1.1 Irreducible trees

We write $\mathbb{N}$ to denote the set of natural numbers $\{0, 1, 2, \ldots\}$. The set of integers $\{\ldots, -1, 0, 1, \ldots\}$ is denoted by $\mathbb{Z}$. The set of strictly negative and positive integers are denoted by $\mathbb{Z}_-$ and $\mathbb{Z}_+$, respectively.

Let A be a finite alphabet. We denote by $|A|$ the cardinal of the set A. For integers $m, n \in \mathbb{Z}$ with $m \leqslant n$, we will use the shorthand notation $w_{m:n}$ to denote the string $(w_m, \ldots, w_n)$ of symbols in the alphabet A. The length of this string will be denoted by $\ell(w_{m:n}) = n - m + 1$. If $m > n$ we let $w_{m:n}$ denote the empty string λ. For any $j \in \mathbb{N}$, we let A^j denote the set of strings in A having length j, in particular $A^0 = \{\lambda\}$. We also let $A^\star = \cup_{j \geqslant 0} A^j$ denote the set of all finite strings on A and we denote by A^∞ the set of all left-infinite sequences $w_{-\infty:n}$ with symbols in A.

We say that a sequence $s_{j:k}$ is a *suffix* of a sequence $w_{m:n}$ if $\ell(s_{j:k}) \leqslant \ell(w_{m:n})$ and $s_{k-i} = w_{n-i}$ for all $i = 0, \ldots, k - j$. This will be denoted as $s_{j:k} \preceq w_{m:n}$. If $\ell(s_{j:k}) < \ell(w_{m:n})$ then we say that s is a proper suffix of w and denote this relation by $s \prec w$. Given a sequence w, the maximal proper suffix of w (obtained bu removing the leftmost symbol) will be denoted by $\mathrm{suf}(w)$.

Definition 2.1. A subset $\tau \subset A^\star \cup A^\infty$ is a *tree* if it satisfies the *suffix property*, what means that no $w \in \tau$ is a proper suffix of another $s \in \tau$. If in addition, a tree τ satisfies the *irreducibility property*, which states that no string belonging to τ can be replaced by a proper suffix without violating the suffix property, then it is called *irreducible tree*.

It is easy to see that the set τ can be identified with the set of leaves of a rooted tree with a finite set of labeled branches. Elements of τ will be denoted either as w or as $w_{-k:-1}$ if we want to stress the number of symbols in the string.

Example 2.2. Suppose $A = \{0, 1\}$. Consider the following sets of sequences with

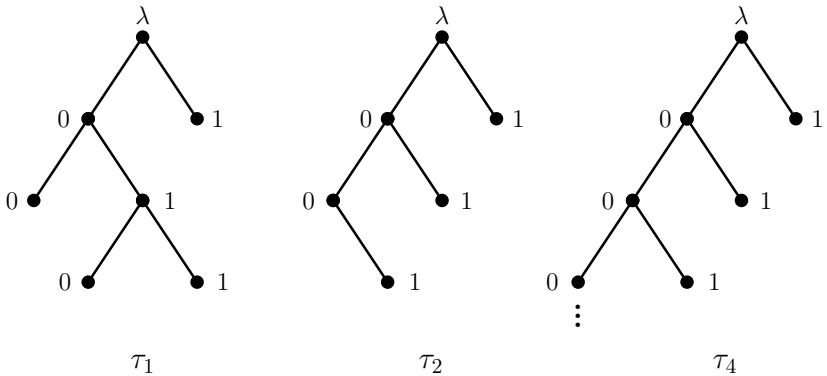


Figure 2.1: Examples of the tree representation of the sets τ_1, τ_2 and τ_4 that satisfy the suffix property in Definition 2.1. The sequences in the set (read from left to right) are read in the tree bottom-up (from leaves to root). The set τ_2 is not irreducible, because we can substitute the sequence 100 by the sequence 00 without violating the suffix property. The set τ_4 is infinite then we represent a truncated version with sequences of length up to three.

symbols in A.

$$\begin{aligned}
\tau_1 &= \{00, 010, 110, 1\} \\
\tau_2 &= \{100, 10, 1\} \\
\tau_3 &= \{000, 00, 100, 10, 1\} \\
\tau_4 &= \{10_{-k:-1} : k = 0, \dots\} \cup \{0_{-\infty:-1}\}\,.
\end{aligned}$$

Here, $10_{-k:-1}$ represents the sequence obtained by concatenating a 1 with k 0's. Similarly, the sequence $0_{-\infty:-1}$ is a semi-infinite sequence with all 0's. It can be seen that τ_1 and τ_4 correspond to irreducible trees over A, satisfying all the conditions in Definition 2.1. On the other hand, τ_2 does not satisfy the irreducibility property and τ_3 does not satisfy the suffix property. As τ_1, τ_2 and τ_4 satisfy the suffix property, they can be represented graphically as an (inverted) tree where each sequence is represented by a leaf in the tree, see Figure 2.1.

In the set of all trees over the alphabet A we can define a partial ordering.

Definition 2.3. We will say that $\tau \preceq \tau'$ if for every $v \in \tau'$ there exists $w \in \tau$ such that $w \preceq v$. As usual, whenever $\tau \preceq \tau'$ with $\tau \neq \tau'$ we will write $\tau \prec \tau'$.

For example, considering the trees defined in Example 2.2 we can say that $\tau_2 \prec \tau_4$ but $\tau_2 \not\prec \tau_1$.

The height $\ell(\tau)$ of the tree τ is the maximal length of a sequence in τ, that is

$$\ell(\tau) = \max\{\ell(w) \colon w \in \tau\}.$$

In Example 2.2 we have $\ell(\tau_1) = \ell(\tau_2) = 3$ and $\ell(\tau_4) = \infty$. Given a tree τ and $K \in \mathbb{N}$ we will denote by $\tau|_K$ the tree τ *truncated* at level K, that is

$$\tau|_K = \{w \in \tau \colon \ell(w) \leqslant K\} \cup \{w \in A^K \colon w \prec u \text{ for some } u \in \tau\}.$$

For example,

$$\tau_4|_3 = \{000, 100, 10, 1\}.$$

2.1.2 Context tree of a stationary ergodic process

Let $\{X_i \colon i \in \mathbb{Z}\}$ be a stationary and ergodic process with law, or measure, $\mathbb{P}$ assuming values in the alphabet A[1]. If w is a finite sequence, we denote by $p(w)$ the probability of observing w at any time, that is

$$p(w) = \mathbb{P}\big(X_{1:\ell(w)} = w\big).$$

If $s \in A^\star$ is such that $p(s) > 0$ we write

$$p(a|s) = \mathbb{P}\big(X_0 = a \mid X_{-\ell(s):-1} = s\big), \tag{2.1}$$

for the transition probabilities of the process, with the convention that if $s = \lambda$ then $p(a|s) = p(a)$.

A process as above is said to have kernel of transition probabilities given by $p(\cdot|\cdot) \colon A \times A^\star \to [0, 1]$ as defined in (2.1).

Definition 2.4. We say that the string $s \in A^\star$ is a *context* for a process with law $\mathbb{P}$ if it satisfies

1. $p(s) > 0$ or $s = \lambda$.
2. For all $a \in A$ and all $w \in A^\star$ such that $s \prec w$

$$p(a|w) = p(a|s). \tag{2.2}$$

[1] In this book we assume the reader is familiar with the definition and properties of stationary ergodic processes over finite alphabets. For an introduction of this subject we recommend the book Shields (1996).

3. No proper suffix of s satisfies 2.

An *infinite context* is a left-infinite sequence $s_{-\infty:-1}$ such that its finite suffixes $s_{-k:-1}, k = 1, 2, \ldots$ have positive probability but none of them nor λ is a context for the process.

By this definition, the set of contexts of a process with measure $\mathbb{P}$ is an irreducible tree, as defined in Definition 2.1, and will be denoted simply by τ. We leave this proof as an exercise to the reader, see Exercise 2.4.

Example 2.5. Consider the stationary Markov chain of order 3 over the alphabet $A = \{0, 1\}$ defined by the transition probabilities

w	$p(0\|w)$	$p(1\|w)$
$ab1$	0.2	0.8
$a00$	0.5	0.5
010	0.3	0.7
110	0.7	0.3

where $a, b \in A$ are arbitrary. This is an example of what is called in the statistics literature a *Variable Length Markov Chain* (VLMC). By Definition 2.4, the only contexts of this process are the strings 1, 00, 010 and 110. The context tree of this process is the tree τ_1 represented in Figure 2.1.

Example 2.6. Suppose the ergodic process $\{X_i : i \in \mathbb{Z}\}$ takes values in $A = \{0, 1\}$, and in order to decide the probability distribution of the next symbol based on the past realization, we only need to know the distance to the last occurrence of a 1[2]. That is, for any $k \geqslant 0$ and any $v, w \in A^\star$ assume that

$$p(1|v10_{-k:-1}) = p(1|w10_{-k:-1}),$$

and that these conditional probabilities are well defined. According to Definition 2.4, the strings $10_{-k:-1}, k \geqslant 0$, as well as the semi-infinite sequence $0_{-\infty:-1}$ are contexts of this process.

Therefore, the context tree of this process is τ_4 shown partially in Figure 2.1.

[2]One can show that this process is obtained, for example, when $X_i = Y_i Z_i$, where $\{Y_i : i \in \mathbb{Z}\}$ is a Markov chain of order 1 over A and $\{Z_i : i \in \mathbb{Z}\}$ is an i.i.d (Bernoulli) sequence with values in A.

2.2 Inference on model parameters

In the sequel we will assume we have a finite sample $x_1, \ldots, x_n$ of elements in A generated by a stochastic process with law $\mathbb{P}$ and context tree τ^*. In this context, the inference problem is related to the estimation of the transition probabilities p that completely determine the law of the process, and the model selection problem is related to finding a procedure based on $x_{1:n}$ to select the context tree τ^*.

Let d be an integer such that $d < n$. For any finite string w with $\ell(w) \leqslant d$ and any symbol $a \in A$ we denote by $N_n(w, a)$ the number of occurrences of the string wa in the sample $x_{(d-\ell(w)+1):n}$, that is

$$N_n(w, a) = \sum_{t=d+1}^{n} \mathbf{1}\{x_{(t-\ell(w)):(t-1)} = w, x_t = a\}. \tag{2.3}$$

We also define the counter $N_n(w)$ as the sum of $N_n(w, a)$ over all $a \in A$, that is

$$N_n(w) = \sum_{a \in A} N_n(w, a). \tag{2.4}$$

Observe that in general we can have $N_n(wa) \neq N_n(w, a)$ for some sequence w and symbol $a \in A$, because $N_n(wa)$ considers the sub sample $x_{(d-\ell(w)+1):(n-1)}$, while $N_n(w, a)$ takes into account the sub sample $x_{(d-\ell(w)+1):n}$[3].

Given any tree τ with $\ell(\tau) \leqslant d$ and any family of probability distributions over the finite set A indexed by the contexts in the tree $q = \{q(\cdot|w) \colon w \in \tau\}$, the likelihood function $L_\tau(q; x_{1:n})$ conditioned on $x_{1:d}$ is given by

$$L_\tau(q; x_{1:n}) = \prod_{w \in \tau} \prod_{a \in A} q(a|w)^{N_n(w,a)}, \tag{2.5}$$

with the convention that $0^0 = 1$[4]. Then, it is not difficult to prove that the maxi-

[3] It can be shown that all the results in this section apply straightforward to the real number of occurrences of any sequence wa in the sample $x_{1:n}$, by substituting d by $\ell(w)$ in (2.3). The fact of using the same value d independent of the length of the sequence facilitates the proofs of the consistency results of the context tree estimators defined in the following sections, where we need to handle a set of sequences with different lengths.

[4] Informally speaking, the likelihood function is defined as the probability of the sample for a given set of parameters in the model. As the sample is assumed to be fixed, the likelihood function is a function of the parameters. Then, the set of parameters maximizing this function are called maximum likelihood estimators. For more details on the definition of the likelihood function for some stochastic processes see Guttorp (1995).

mum likelihood estimators of the transition probabilities are given by

$$\hat{p}_n(a|w) = \frac{N_n(w,a)}{N_n(w)} \tag{2.6}$$

if $N_n(w) > 0$ (and any distribution on A if $N_n(w) = 0$), see Exercise 2.6. By plugging-in these estimators in (2.5) we obtain the maximum likelihood of $x_{1:n}$ conditioned on $x_{1:d}$ for the tree τ, given by

$$\hat{P}_\tau(x_{1:n}) = \prod_{w\in\tau}\prod_{a\in A} \hat{p}_n(a|w)^{N(w,a)}\,. \tag{2.7}$$

2.2.1 Concentration inequalities

We know, by the ergodicity of the chain and the Ergodic Theorem[5], that for any finite sequence $w \in A^\star$ and all $a \in A$, the empirical transition probability $\hat{p}_n(a|w)$ converges to $p(a|w)$ almost surely as $n \to \infty$, see Exercise 2.7. But in order to prove the consistency of the context tree estimators defined in the following sections, we need finer results on the rate of convergence of the empirical transition probabilities.

Suppose $\{X_i : i \in \mathbb{Z}\}$ is a stochastic chain with memory of variable length with kernel p and context tree τ^*. From now on we assume the process satisfies the following condition.

Assumption 2.7. The process $\{X_t : t \in \mathbb{Z}\}$ satisfies the following loss of memory condition:

$$\alpha := 1 + \sum_{k=1}^{\infty} \alpha_k \;<\; \infty, \tag{2.8}$$

where

$$\alpha_k : \;= \sup_{a\in A}\sup_{u\in A^\infty} |\mathbb{P}(X_k = a \mid X_{-\infty:0} = u) - p(a)| \;<\; \infty\,.$$

This assumption is sufficient to prove concentration of the empirical counter $N_n(w,a)$ around its mean $(n-d)p(wa)$. This result implies a concentration bound for the transition probability $\hat{p}_n(a|w)$ around its limiting value $p(a|w)$, as shown originally in Galves, Maume-Deschamps, and Schmitt (2008) for processes with bounded context tree and extended in Galves and Leonardi (2008) for

[5] See for instance Shields (1996), Theorems I.3.1 and I.3.2

processes with unbounded context trees, i.e. processes with infinite memory. Assumption 2.7 can be proved under different conditions on the parameters of the model, as shown in Galves and Leonardi (ibid., Lemma 3.4), see also Garivier and Leonardi (2011) for slightly different conditions.

We present below, in Theorem 2.8 and Corollary 2.9, the concentration inequalities proved in Galves and Leonardi (2008).

Theorem 2.8. *Assume the process $\{X_t : t \in \mathbb{Z}\}$ satisfies Assumption 2.7. For any finite sequence w with $\ell(w) \leqslant d$, any symbol $a \in A$ and any $t > 0$ the following inequality holds*

$$\mathbb{P}\big(\,|N_n(w,a) - (n-d)\,p(wa)| \,>\, t\,\big) \,\leqslant\, e^{\frac{1}{e}} \exp\Big[\frac{-t^2}{4e\alpha(n-d)(d\,+\,1)}\Big].$$

As a direct consequence of Theorem 2.8 we obtain the following corollary.

Corollary 2.9. *Assume the process $\{X_t : t \in \mathbb{Z}\}$ satisfies Assumption 2.7. For any finite sequence w with $p(w) > 0$ and $\ell(w) \leqslant d$, any symbol $a \in A$, any $t > 0$ and any $n > d$ the following inequality holds*

$$\mathbb{P}\big(|\hat{p}_n(a|w) - p(a|w)| > t\big) \,\leqslant\, e^{\frac{1}{e}}\,(|A|+1)\,\exp\Big[-\,\frac{(n-d)t^2\,p(w)^2}{16e\alpha|A|^2(d\,+\,1)}\Big]. \quad (2.9)$$

In order to prove the consistency of the context tree estimators that we will introduce in the following section, we need to control the empirical transition probabilities associated to different context tree models. When the candidate context trees are bounded with an *a priori* upper bound, then Corollary 2.9 is sufficient to obtain these consistency results. But for unbounded candidate context trees, Corollary 2.9 is not optimal because of the $p(w)$ factor in the exponent of the right-hand side of (2.9), that in general decreases exponentially fast with the length of the sequence. In this case, we can use finer results based on martingale techniques that can be applied when the sequence w has a suffix in the context tree of the process. We state this result in the following theorem.

Theorem 2.10. *For any sequence w having a suffix in the context tree of the process with $\ell(w) \leqslant d$, any $a \in A$ and any $t > 1$ we have that*

$$\mathbb{P}\big(N_n(w)\max_{a\in A}|\hat{p}_n(a|w) - p(a|w)|^2 > t\big) \,\leqslant\, e|A|\log(n)t\exp(-t).$$

Theorem 2.10 was first obtained in Garivier and Leonardi (2011), considering the *Kullback–Leibler* divergence between the empirical and theoretical transition probabilities. This version of the theorem was considered in Leonardi, Carvalho, and Frondana (2021) for Markov random fields on graphs.

The proofs of all the theoretical results in this section are postponed to Section 2.4.

2.3 Model selection

In this section we present the main approaches in the literature for model selection, that is the estimation of the context tree τ^* of the process, based on the finite sample $x_{1:n}$. We begin by describing in Section 2.3.1 the algorithm Context introduced by Rissanen (1983). Then, in Section 2.3.2 we present the Bayesian Information Criterion defined in Csiszár and Talata (2006b) and finally, in Section 2.3.3 we describe the *Smallest Maximizer Criterion* introduced in Galves, Galves, et al. (2012).

2.3.1 The algorithm Context

The algorithm Context was introduced by Rissanen (1983) in its original work. This algorithm computes, for each node of a given tree, a discrepancy measure between the transition probability associated to this context and the corresponding transition probabilities of the nodes obtained by concatenating a single symbol to the context. Beginning with the largest leaves of a candidate tree, if the discrepancy measure is greater than a given threshold, the contexts are maintained in the tree; otherwise, they are pruned. The procedure continues until no more pruning of the tree can be performed.

The discrepancy measure used by the algorithm Context is the *Küllback–Leibler divergence*, defined for two probability measures p and q on A by

$$D(p;q) = \sum_{a\in A} p(a)\log\frac{p(a)}{q(a)} \tag{2.10}$$

where, by convention, $p(a)\log\frac{p(a)}{q(a)} = 0$ if $p(a) = 0$ and $p(a)\log\frac{p(a)}{q(a)} = +\infty$ if $p(a) > q(a) = 0$.

Denote by $\mathcal{V}_n$ the set of all sequences $w \in A^\star$ that appear at least once in the sample, that is

$$\mathcal{V}_n = \{w \in A^\star \colon N_n(w) \geqslant 1\}\,. \tag{2.11}$$

Observe that the set of maximal sequences (with respect to the suffix order $\prec$) in $\mathcal{V}_n$ is a tree as defined in Definition 2.1. This tree is not necessarily irreducible. We leave these proofs to the reader, see Exercise 2.8. Moreover, there is a one-to-one correspondence between the set $\mathcal{V}_n$ and the nodes of this tree (considering also the internal nodes). Therefore, even if $\mathcal{V}_n$ is not a tree as defined in Definition 2.1, by an abuse of notation we will refer sometimes to $\mathcal{V}_n$ as a tree.

For a given sequences $w \in \mathcal{V}_n$ let

$$\Delta_n(w) = \sum_{b\colon bw \in \mathcal{V}_n} N_n(bw)\, D\left(\hat{p}_n(\cdot|bw);\, \hat{p}_n(\cdot|w)\right) . \tag{2.12}$$

Remark 2.11. The discrepancy measure $\Delta_n(w)$ defined in Equation (2.12) is the one originally proposed by Rissanen (ibid.), but other possibilities have been proposed in the literature, see for instance Bühlmann and Wyner (1999) and Galves, Maume-Deschamps, and Schmitt (2008).

We will denote the threshold used in the algorithm Context on samples of length n by δ_n, where $(\delta_n)_{n\in\mathbb{N}}$ is a sequence of positive real numbers such that $\delta_n \to +\infty$ and $\delta_n/n \to 0$ when $n \to +\infty$. This asymptotic property on δ_n is necessary to obtain the consistency of the algorithm Context when the sample size increases, see Theorem 2.12 later in this section.

The algorithm Context works as follows. First we construct the maximal possible tree with depth at most d and such that each node w in the tree belongs to $\mathcal{V}_n$, that is $N(w) \geqslant 1$. This first tree is not necessarily irreducible. Then, we assign to each node in this tree an indicator function, denoted by $C_w(x_{1:n}) \in \{0, 1\}$, beginning by assigning a value of 0 to the terminal nodes and going up until reaching the root, in a recursive form. The recursion is defined as follows. For a given node w, and assuming that the function $C_v(x_{1:n})$ was defined for every node v in the tree such that $w \prec v$, we compute $C_w(x_{1:n})$ as

$$C_w(x_{1:n}) = \max\left\{\mathbf{1}\{\Delta_n(w) > \delta_n\},\, \max_{b\in A} C_{bw}(x_{1:n})\right\} . \tag{2.13}$$

This definition implies that if a given node in the tree has a significantly bigger discrepancy measure, then this node must be an internal node in the context tree of the process and it must be maintained in the final tree, as well as all the nodes in the way from this node to the root. For this reason all these nodes will receive value 1 as its associated indicator function. Then, the final estimated tree by the algorithm Context is the tree of sequences with associated indicator function 0 and such that all the nodes in the way to the root has associated value 1. Formally,

the context tree estimator $\hat{\tau}_{\mathrm{C}}(x_{1:n})$ given by the algorithm Context is the set of sequences given by

$$\hat{\tau}_{\mathrm{C}}(x_{1:n}) = \{w \in \mathcal{V}_n : C_w(x_{1:n}) = 0 \text{ and } C_u(x_{1:n}) = 1 \text{ for all } u \prec w\}. \quad (2.14)$$

The algorithm Context is consistent for any threshold sequence δ_n such that $\frac{\delta_n}{n} \to 0$ as $n \to \infty$ and $\frac{\delta_n}{\log\log n} > c_0$ for all n, where c_0 is a constant depending on the process. As the constant c_0 is unknown, the most common approach is to use a threshold sequence $\delta_n = c \log n$, with $c > 0$. The consistency of the algorithm Context is stated in the following theorem.

Theorem 2.12. *Assume the process generating the sample $x_{1:n}$ has context tree τ^* and satisfies Assumption 2.7. Let $K, d \in \mathbb{N}$ be such that*

$$\min_{\substack{s \prec w \in \tau^* \\ \ell(s) \leqslant K}} \max_{\substack{u \in A^r \\ r \leqslant d - \ell(w)}} \max_{a \in A} \{|p(a|us) - p(a|\mathrm{suf}(us))|\} \geqslant \epsilon > 0 \quad (2.15)$$

and let δ_n such that $\delta_n / n \to \infty$ when $n \to \infty$. Then, there exist explicit constants $c_i > 0$, $i = 1, \ldots, 4$ depending only on the process and $n_0 \in \mathbb{N}$ such that for any $n \geqslant n_0$ it holds that

$$\mathbb{P}(\hat{\tau}_C(x_{1:n})|_K \neq \tau^*|_K) \leqslant c_1 \exp\Big[-\frac{c_2(n-d)}{d+1}\Big] + c_3 \delta_n \log(n) n^2 \exp\big[-c_4 \delta_n\big]. \quad (2.16)$$

The constants in (2.16) are given by

$$c_1 = 3e^{\frac{1}{e}} (|A| + 1)|A|^K \qquad c_2 = \frac{\epsilon^2 q_{\min}^2}{256 e \alpha |A|^2}$$

$$c_3 = e p_{\min} \qquad c_4 = \frac{p_{\min}}{|A|},$$

where $p_{\min} > 0$ and $q_{\min} > 0$ depend on the process.

Remark 2.13. It can be seen that for any $K \in \mathbb{N}$ there is always a value of d such that (2.15) holds. This hypothesis can be avoided by letting d increase with the sample size n and by controlling the upper bounds in (2.16). Extensions of Theorem 2.12 can also be obtained by allowing K to be a function of the sample size n. In this case, the rate at which K increases must be controlled together with the rate at which ϵ, $p_{\min}$ and $q_{\min}$ decrease with the sample size. This leads to a rather technical condition, see for instance Talata and Duncan (2009).

From Theorem 2.12 we can obtain the consistency of the algorithm Context for an appropriate choice of δ_n.

Corollary 2.14. *Under the same hypotheses of Theorem 2.12, take* $\delta_n \geqslant \frac{4|A|}{p_{min}} \log(n)$ *for all* n. *Then for* $d = o(n)$ *we have that*

$$\mathbb{P}(\hat{\tau}_C(x_{1:n})|_K \neq \tau^*|_K) \to 0\,,$$

as $n \to \infty$.

Observe that Corollary 2.14 implies in particular that when τ^* is finite, if we take $K \geqslant \ell(\tau^*)$ we have that $\hat{\tau}_C(x_{1:n}) = \tau^*$ with probability converging to one when $n \to \infty$. In general we take $d = O(\log(n))$ and in the finite case we can take $K = d$. In the infinite case we need stronger assumptions to enable $K \to \infty$.

Remark 2.15. The proof of Theorem 2.12 and Corollary 2.14 are based on the exponential inequalities presented in Section 2.2, specifically Corollary 2.9 and Theorem 2.10. But other approaches exist in the literature to prove the consistency of the algorithm Context, see for instance Duarte, Galves, and Garcia (2006) and the review article Galves and Löcherbach (2008) for details.

2.3.2 Penalized maximum likelihood

The penalized maximum likelihood approach is based on the regularization of the maximum likelihood function defined in (2.7) with an appropriate penalizing term. The estimated tree is then the tree maximizing this regularized function. Although the set of all possible irreducible trees over the alphabet A grows exponentially fast with the depth of the trees d, Csiszár and Talata (2006b) showed that the estimation can be done in linear time in the size of the sample, by an efficient recursion over the nodes of the tree, in a similar way as the algorithm Context works. By using this construction, it can be shown that the estimated tree with penalized maximum likelihood is always smaller than the tree obtained with the algorithm Context, as shown in Garivier and Leonardi (2011).

The most well known method of penalized maximum likelihood for stochastic chains with memory of variable length is the Bayesian Information Criterion. The BIC is a classic approach in statistics originally proposed for general model selection by Schwarz (1978). In the area of stochastic chains with memory of variable length it was first considered in the work Csiszár and Talata (2006b), where the authors also proposed an efficient algorithm to compute the estimator of the context tree.

It is worth noting that on the definition of a penalized maximum likelihood criterion, the likelihood function (2.7) does not need to be computed on all possible trees. Rather, it is only necessary to compute the likelihood function for all trees whose associated contexts are observed in the sample. This is made precise in the following definition.

Definition 2.16. We say the irreducible tree τ is *admissible* for the sample $x_{1:n}$ if $\ell(\tau) \leqslant d$, $N_n(w) > 0$ for any $w \in \tau$ and for any $j = d \ldots, n-1$ there exists a sequence $w \in \tau$ such that $w \preceq x_{1:j}$.

Then, the set of candidate trees considered in the penalized maximum likelihood criterion, denoted by $\mathcal{T}_n$, will be the set of all admissible trees. The penalized maximum likelihood criterion for the sequence $x_{1:n}$ is then defined by

$$\hat{\tau}_{\text{PML}}(x_{1:n}) = \arg\max_{\tau \in \mathcal{T}_n} \left\{ \log \hat{P}_\tau(x_{1:n}) - |\tau| \text{pen}(n) \right\}, \tag{2.17}$$

where $|\tau|$ is the size of the tree τ, i.e. the number of sequences in τ and $\text{pen}(n)$ is some positive function such that $\text{pen}(n) \to +\infty$ and $\text{pen}(n)/n \to 0$ when $n \to \infty$. As we will show below, the penalty function is closely related to the threshold δ_n in the algorithm Context.

The BIC as introduced by Csiszár and Talata (2006b) is obtained by using the penalty function $\text{pen}(n) = c(|A| - 1)\log(n)$, for a given constant $c > 0$. It may first appear practically impossible to compute $\hat{\tau}_{\text{PML}}(x_{1:n})$, because the maximization in (2.17) must be performed over the set of all admissible trees $\mathcal{T}_n$. Fortunately, Csiszár and Talata (ibid.) showed how to adapt the Context Tree Maximizing (CTM) method introduced in Willems, Shtarkov, and Tjalkens (1995) to obtain a simple and efficient algorithm computing $\hat{\tau}_{\text{PML}}(x_{1:n})$. We describe this algorithm in the sequel.

For any sequence $w \in \mathcal{V}_n$, see (2.11), we define

$$\hat{P}_w(x_{1:n}) = \prod_{a \in A} \hat{p}_n(a|w)^{N_n(w,a)}.$$

Then we have that the maximum likelihood function, introduced in (2.7), is given by

$$\hat{P}_\tau(x_{1:n}) = \prod_{w \in \tau} \hat{P}_w(x_{1:n}).$$

This equality allows us to decompose the penalized maximum likelihood function in the argument of (2.17) as a sum over the sequences in the tree τ; that is

$$\log \hat{P}_\tau(x_{1:n}) - |\tau| \text{pen}(n) \;=\; \sum_{w \in \tau} \left[\log \hat{P}_w(x_{1:n}) - \text{pen}(n) \right]. \tag{2.18}$$

This is the key to obtain the efficient algorithm based on the CTM method of Willems, Shtarkov, and Tjalkens (ibid.).

We consider, as in Section 2.3.1, the set of nodes given by (2.11). We first define, for any $w \in \mathcal{V}_n$ with $\ell(w) \geqslant d$ the value

$$V_w(x_{1:n}) = e^{-\mathrm{pen}(n)} \hat{P}_w(x_{1:n})$$

and the indicator function

$$\mathcal{X}_w(x_{1:n}) = 0\,.$$

We then assign to any $w \in \mathcal{V}_n$ recursively from bottom to top of the tree, the value

$$V_w(x_{1:n}) = \max\Big\{e^{-\mathrm{pen}(n)} \hat{P}_w(x_{1:n}),\ \prod_{b\in A:\, bw\in\mathcal{V}_n} V_{bw}(x_{1:n})\Big\} \tag{2.19}$$

and the indicator

$$\mathcal{X}_w(x_{1:n}) = \mathbf{1}\Big\{\prod_{b\in A:\, bw\in\mathcal{V}_n} V_{bw}(x_{1:n}) \ >\ e^{-\mathrm{pen}(n)} \hat{P}_w(x_{1:n})\Big\}\,. \tag{2.20}$$

Now, for any finite string w, with $\ell(w) \leqslant d$ and for any tree $\tau \in \mathcal{T}_n$, we define the irreducible tree τ_w as the set of branches in τ which have w as a suffix, that is

$$\tau_w = \{u \in \tau : w \preceq u\}.$$

Let $\mathcal{T}_w$ be the set of all trees defined in this way, that is

$$\mathcal{T}_w = \{\tau_w : \tau \in \mathcal{T}_n\}.$$

If w is a sequence such that $\mathcal{X}_w(x_{1:n}) = 1$ we define the maximizing tree assigned to the sequence w as the tree $\tau_w^{\mathrm{M}} \in \mathcal{T}_w$ given by

$$\tau_w^{\mathrm{M}} = \{u \in \mathcal{V}_n : \mathcal{X}_u(x_{1:n}) = 0 \text{ and } \mathcal{X}_v(x_{1:n}) = 1 \text{ for all } w \preceq v \prec u\}. \tag{2.21}$$

On the other hand, if $\mathcal{X}_w(x_{1:n}) = 0$ we define $\tau_w^{\mathrm{M}} = \{w\}$.

The following lemma, proven in Csiszár and Talata (2006b), is the key for the efficient computation of the penalized maximum likelihood context tree estimator given in (2.17).

Lemma 2.17. *For any finite string w, with $\ell(w) \leqslant d$, we have*

$$V_w(x_{1:n}) = \max_{\tau\in\mathcal{T}_w} \prod_{u\in\tau} e^{-\mathrm{pen}(n)} \hat{P}_u(x_{1:n}) = \prod_{u\in\tau_w^{M}} e^{-\mathrm{pen}(n)} \hat{P}_u(x_{1:n}). \tag{2.22}$$

The second equality in (2.22) implies, in particular, that the context tree estimator defined in (2.17) is given by $\hat{\tau}_{\text{PML}}(x_{1:n}) = \tau_\lambda^{\text{M}}$, that is

$$\hat{\tau}_{\text{PML}}(x_{1:n}) = \{u \in \mathcal{V}_n \colon \mathcal{X}_u(x_{1:n}) = 0 \text{ and } \mathcal{X}_v(x_{1:n}) = 1 \text{ for all } v \prec u\} \,. \quad (2.23)$$

Observe that this definition of the context tree estimator with penalized maximum likelihood is very similar to that given for the algorithm Context in (2.14). This similarity was exploited in Garivier and Leonardi (2011) by showing that if the threshold function δ_n in the algorithm Context is smaller than the penalizyng function $\text{pen}(n)$ in the definition of the penalized maximum likelihood estimator we have that $\hat{\tau}_{\text{PML}}(x_{1:n})$ is *smaller* than $\hat{\tau}_{\text{C}}(x_{1:n})$. This statement is made precise in the following proposition.

Proposition 2.18. *For any $n \geqslant 1$ and all sequences $x_{1:n}$, if $\delta_n \leqslant \text{pen}(n)$ we have that*

$$\hat{\tau}_{PML}(x_{1:n}) \preceq \hat{\tau}_C(x_{1:n}) \,,$$

with the partial order given in Definition 2.3.

As in the case of the algorithm Context, we can prove that the penalized maximum likelihood criterion is consistent for a convenient penalizing function $\text{pen}(n)$.

Theorem 2.19. *Assume the process generating the sample $x_{1:n}$ has context tree τ^* and satisfies Assumption 2.7. Let $K, d \in \mathbb{N}$ be such that*

$$\min_{\substack{s \prec w \in \tau^* \\ \ell(s) \leqslant K}} \max_{\substack{u \in A^r \\ r \leqslant d - \ell(w)}} \max_{a \in A} \{|p(a|us) - p(a|s)|\} \geqslant \epsilon > 0 \quad (2.24)$$

and let $\text{pen}(n)$ such that $\text{pen}(n)/n \to \infty$ when $n \to \infty$. Then, there exist explicit constants $c_i > 0$, $i = 1, \ldots, 4$ depending only on the process and $n_0 \in \mathbb{N}$ such that for any $n \geqslant n_0$ it holds that

$$\begin{aligned} \mathbb{P}(\hat{\tau}_{PML}(x_{1:n})|_K \neq \tau^*|_K) \leqslant\ & c_1 \exp\Big[-\frac{c_2(n-d)}{d+1}\Big] \\ & + c_3 \text{pen}(n) \log(n) n^2 \exp\big[-c_4 \text{pen}(n)\big] \,. \end{aligned} \quad (2.25)$$

The constants in (2.16) are given by

$$\begin{aligned} c_1 &= 3e^{\frac{1}{e}} (|A| + 1)|A|^K & c_2 &= \frac{\epsilon^2 q_{\min}^2}{256 e \alpha |A|^2} \\ c_3 &= e p_{\min} & c_4 &= \frac{p_{\min}}{|A|} \,, \end{aligned}$$

where $p_{\min} > 0$ and $q_{\min} > 0$ depend on the process.

As in the case of the algorithm Context, we obtain as a consequence of Theorem 2.19 the consistency of the penalized maximum likelihood criterion for an appropriate penalizing function pen(n).

Corollary 2.20. *Under the same hypotheses of Theorem 2.19, take* $\text{pen}(n) \geqslant \frac{4|A|}{p_{min}} \log(n)$ *for all* n. *Then for* $d = o(n)$ *we have that*

$$\mathbb{P}(\hat{\tau}_{PML}(x_{1:n})|_K \neq \tau^*|_K) \to 0\,,$$

as $n \to \infty$.

Remark 2.21. Corollary 2.20 establishes the consistency of the penalized maximum likelihood criterion for a penalizing term $\text{pen}(n) = o(n)$ and such that $\text{pen}(n) \geqslant c_0 \log(n)$ for a specified constant c_0. This result follows from the concentration inequalities presented in Section 2.2. It is worth noting that consistency results also exist in the literature for penalizing terms of the form $\text{pen}(n) = c \log(n)$ for any constant $c > 0$, see for instance Csiszár and Talata (2006b) and Garivier (2006).

2.3.3 Smallest Maximizer Criterion

The Smallest Maximizer Criterion (henceforth SMC) is a constant free procedure that selects a context tree model, given the sample $x_{1:n}$. Informally speaking this criterion can be described as follows. First of all, using the penalized maximum likelihood approach described in Section 2.3.2 with a penalty $\text{pen}(n) = c \log(n)$, we identify the set of *"champion trees"*, which are the context trees maximizing the penalized likelihood for each possible constant c in the penalization term. It turns out that the set of context trees identified in this way is totally ordered with respect to the natural ordering among rooted trees. The sample likelihood increases when we go through the ordered sequence of champion trees: the bigger the tree, the bigger the likelihood of the sample. The noticeable fact is that there is a change of regime in the way the sample likelihood increases from a champion tree to the next one. The function mapping the successive champion trees to their corresponding log-likelihood values starts with a very steep slope which becomes almost flat when it crosses a certain tree. The tree corresponding to this change of regime is the estimated tree with the SMC.

As before, let $\mathcal{T}_n$ be the set of all admissible context trees for the sample $x_{1:n}$. Let $\text{df}\colon \mathcal{T}_n \to \mathbb{N}$ be a function that assigns to each tree $\tau \in \mathcal{T}_n$ the number of degrees of freedom of the model corresponding to the context tree τ. The definition of $\text{df}(\tau)$ depends on the class of models considered. Without any restriction

$\mathrm{df}(\tau) = (|A|-1)|\tau|$. However, in many scientific data sets, we know beforehand that some transitions are not allowed by the problem's nature. That is the case of the linguistic data set we consider in our case study presented in Chapter 3. In general, we can define an incidence function $\chi\colon A^\star \times A \to \{0,1\}$ which indicates in a consistent way which are the possible transitions, by using the convention that $\chi(w,a) = 0$ means that the transition from w to a is not allowed. By consistent we mean that if $\chi(w,a) = 0$ for some $w \in A^\star$ and $a \in A$ then $\chi(uw,a) = 0$ for all $u \in A^\star$. In this case,

$$\mathrm{df}(\tau;\chi) = \sum_{w\in\tau}\sum_{a\in A}\chi(w,a).$$

In order to construct our constant-free selection procedure, we consider the penalized maximum likelihood estimator defined in (2.17) with $\mathrm{pen}(n) = c\log n$ and $|\tau|$ replaced by $\mathrm{df}(\tau)$. That is, for each constant $c \in [0,+\infty]$ define the estimated tree with constant c by

$$\hat{\tau}_{\mathrm{SMC}}(c) \;=\; \arg\max_{\tau\in\mathcal{T}_n}\{\log \hat{P}_\tau(x_{1:n}) - c\cdot\mathrm{df}(\tau)\cdot\log n\}. \tag{2.26}$$

Then define the map

$$c \in [0,+\infty) \mapsto \hat{\tau}_{\mathrm{SMC}}(c) \in \mathcal{T}_n\,,$$

and denote by C_n its image

$$C_n = \{\tau_n^c = \hat{\tau}_{\mathrm{SMC}}(c)\colon c\in[0,+\infty)\}. \tag{2.27}$$

The trees belonging to C_n are called *champion trees*. Observe that the *champion trees* are the ones which maximize the likelihood of the sample for each available number of degrees of freedom. The set $\mathcal{T}_n$ of all admissible context trees is not totally ordered with respect to the ordering introduced in Definition 2.3. But its subset C_n containing only the champion trees is totally ordered, as can be concluded from the following result.

Lemma 2.22. *Let $0 < c_1 < c_2$ be arbitrary positive constants. Then*

$$\hat{\tau}_{SMC}(c_1) \;\succeq\; \hat{\tau}_{SMC}(c_2).$$

Assume τ^* is finite. By the consistency of the penalized maximum likelihood estimator in Corollary 2.20 we have that if the sample size n is big enough, then the tree τ^*, which, by assumption, generated the sample, belongs to C_n with high

probability. In fact, stronger results in the literature show that τ^* belongs to C_n with probability one eventually for large n, see Remark 2.21. It turns out that when $\tau^* \in C_n$ it has a remarkable property: it is an inflection point for the penalized maximum likelihood function. This makes it possible to identify τ^* in the set C_n. This is the basis for the SMC approach and the content of the next theorems. As in Csiszár and Talata (2006b) we assume $d = o(\log(n))$.

Theorem 2.23. *Assume the process generating the sample $x_{1:n}$ has finite context tree τ^*. Then, C_n is totally ordered with respect to the order $\prec$ and eventually almost surely $\tau^* \in C_n$ as $n \to \infty$.*

The next theorem is the basis for the smallest maximizer criterion. It shows that there is a change of regime in the gain of likelihood at τ^*.

Theorem 2.24. *Assume the process generating the sample $x_{1:n}$ has context tree τ^*. Then, the following results hold eventually almost surely as $n \to \infty$.*

1. *For any $\tau \in C_n$, with $\tau \prec \tau^*$, there exists a constant $c(\tau^*, \tau) > 0$ such that*

$$\log \hat{P}_{\tau^*}(x_{1:n}) - \log \hat{P}_{\tau}(x_{1:n}) \geqslant c(\tau^*, \tau)\, n. \tag{2.28}$$

2. *For any $\tau \prec \tau' \in C_n$, with $\tau^* \preceq \tau$, there exists a constant $c(\tau, \tau') \geqslant 0$ such that*

$$\log \hat{P}_{\tau'}(x_{1:n}) - \log \hat{P}_{\tau}(x_{1:n}) \leqslant c(\tau, \tau') \log n. \tag{2.29}$$

Theorem 2.23 and Theorem 2.24 lead to the following *Smallest Maximizer Criterion*: select the smallest tree $\hat{\tau}_{\text{SMC}}$ in the set of champion trees C_n such that

$$\log \hat{P}_{\tau}(x_{1:n}) - \log \hat{P}_{\hat{\tau}}(x_{1:n}) \;\leqslant\; r_n \tag{2.30}$$

with r_n a function satisfying $r_n/n \to 0$ and $r_n/\log n \to \infty$ as $n \to \infty$. The next theorem states the consistency of this criterion.

Theorem 2.25. *Let $x_{1:n}$ be a stochastic chain with memory of variable length with context tree τ^*, with τ^* finite. Then*

$$\mathbb{P}(\hat{\tau}_{SMC} \neq \tau^*) \;\to\; 0$$

for any sequence $\{r_n\}_{n\in\mathbb{N}}$ satisfying $r_n/n \to 0$ and $r_n/\log n \to \infty$ as $n \to \infty$.

In order to implement the SMC we first need an algorithm to compute the set of champion trees C_n. This can be done efficiently using the Context Tree Maximizing method presented in Section 2.3.2, for different values of the penalizing constant. In fact, for any given tree $\tau \in C_n$ associated to a given constant c, it is possible to compute the minimum $c' > c$ leading to a strictly smaller tree $\tau' \prec \tau$ by a direct exploration of the leaves in τ. Then beginning with the constant $c = 0$ and the maximal possible tree $\mathcal{V}_n$ defined in (2.11), we can iterate this procedure until we arrive to the root tree composed uniquely by the empty string λ. The sequence of trees obtained in this way corresponds to the set of champion trees C_n.

Once the set of champion trees C_n has been obtained, the next step is to identify a tree $\hat{\tau}$ belonging to C_n exhibiting a change of regime as specified by Theorem 2.24. One possibility would be to take a given threshold r_n and compare the differences in log-likelihood between two subsequent trees in C_n as in (2.30). That is, denoting by $\tau_0 \prec \tau_1 \prec \cdots \prec \tau_k$ the different trees in C_n, we could define

$$\hat{\tau}_{\text{SMC}} = \min\{\tau_i : \log \hat{P}_{\tau_{i+1}}(x_{1:n}) - \log \hat{P}_{\hat{\tau}_i}(x_{1:n}) < r_n\},$$

where the minimum is taken with respect to the order $\prec$. But this procedure would require the specification of the threshold r_n and it is not clear how this number should be selected. Moreover, as we only have one sample $x_{1:n}$ this would not take into account the variability in the selection principle. For this reason, Galves, Galves, et al. (2012) proposed a Bootstrap procedure[6] to detect the tree in C_n exhibiting the change of regime. This procedure is specified below.

Bootstrap Procedure:

1. For two different sample sizes $n_1 < n_2 < n$ obtain B independent bootstrap resamples of $x_{1:n}$. Denote these resamples by $\mathbf{x}^{*,(b,j)} = \{x_i^{*,(b,j)}, i = 1, \ldots, n_j\}$ where $b = 1, \ldots, B$ and $j = 1, 2$.

2. For $j = 1, 2$ and for all $\tau_n \in C_n$ and its successor $\tau_n' \in C_n$ in the $\prec$ order, compute the average

$$\Delta^{(\tau_n, \tau_n')}(n_j) = \frac{1}{B} \sum_{b=1}^{B} \frac{\log \hat{P}_{\tau_n}(\mathbf{x}^{*,(b,j)}) - \log \hat{P}_{\tau_n'}(\mathbf{x}^{*,(b,j)})}{n_j^{0.9}}.$$

[6]The Bootstrap is a classical resampling method in Statistics where the observed sample is used to produce resamples and to assess the variability of a statistical procedure. We refer the reader to the seminal book Efron and Tibshirani (1993) for an extensive presentation of this method.

3. Apply a one-sided t-test for comparing the two means $\mathbb{E}(\Delta^{(\tau_n,\tau'_n)}(n_1))$ and $\mathbb{E}(\Delta^{(\tau,\tau'_n)}(n_2))$.

4. Select the tree $\hat{\tau}_{\text{SMC}}$ as the smallest champion tree τ_n such that the test rejects the equality of the means in favor of the alternative hipothesis that

$$\mathbb{E}(\Delta^{(\tau_n,\tau'_n)}(n_1)) < \mathbb{E}(\Delta^{(\tau_n,\tau'_n)}(n_2)) .$$

In Step 1 above, any bootstrap resampling method for stochastic chains with memory of variable length can be used. In the specific application of the SMC to linguistic data described in Chapter 3, we use a remarkable feature of the data set, that is, the fact that one of the symbols is a renewal point. This makes it possible to sample randomly with replacement independent strings between two successive such symbols, see Galves, Galves, et al. (ibid.) for details.

2.4 Proofs of this chapter

Before presenting the proof of Theorem 2.8, for completeness we state a proposition obtained by Dedecker and Doukhan (2003, Proposition 4) that is essential to obtain such result.

Proposition 2.26. *Let $\{X_i : i \in \mathbb{N}\}$ be a sequence of centered and square integrable random variables, and $\mathcal{M}_i = \sigma(X_j, 0 \leqslant j \leqslant i)$. Define $S_n = X_1 + \cdots + X_n$ and*

$$b_{n,i} = \max_{i \leqslant l \leqslant n} \left\| X_i \sum_{k=i}^{l} \mathbb{E}(X_k | \mathcal{M}_i) \right\|_{p/2} .$$

Then, for any $p \geqslant 2$, the following inequality holds:

$$\|S_n\|_p \leqslant \left(2p \sum_{i=1}^{n} b_{i,n} \right)^{1/2} .$$

We present now the proofs of the main results in this chapter.

Proof of Theorem 2.8. Let w be a finite sequence and a any symbol in A. Define the random variables

$$U_t = \mathbf{1}\{x_{(t-\ell(w)):t} = wa\} - p(wa),$$

for $t = d+1, \ldots, n$. Observe that $(U_t)_{t \geqslant d+1}$ is a sequence of centered and square integrable random variables. Then, using Proposition 2.26 we have that, for any $r \geqslant 2$

$$\begin{aligned}
&\| N_n(w,a) - (n-d)\, p(wa) \|_r \\
&\quad \leqslant \Big(2r \sum_{t=d+1}^{n} \max_{t \leqslant l \leqslant n} \| \, U_t \sum_{k=t}^{l} \mathbb{E}(U_k \mid U_{d+1}, \ldots, U_t) \, \|_{r/2} \Big)^{\frac{1}{2}} \\
&\quad \leqslant \Big(2r \sum_{t=d+1}^{n} \sum_{k=t}^{n} \sup_{u \in A^t} |\mathbb{P}(X_{(k-\ell(w)):k} = wa \mid X_{1:t} = u) - p(wa)| \Big)^{\frac{1}{2}} \\
&\quad \leqslant \Big(2r\, (n-d)\, (d+1)\, \alpha \Big)^{\frac{1}{2}},
\end{aligned}$$

where in the last inequality we used Assumption 2.7 and the fact that we can use the bounds

$$\sup_{u \in A^t} |\mathbb{P}(X_{(k-\ell(w)):k} = wa \mid X_{1:t} = u) - p(wa)| \ \leqslant \ 1$$

for $k = t, \ldots, t + \ell(w)$ and

$$\sup_{u \in A^t} |\mathbb{P}(X_{(k-\ell(w)):k} = wa \mid X_{1:t} = u) - p(wa)| \ \leqslant \ \ell(wa)\alpha_{k-\ell(w)-t}$$

for $k > t + \ell(w)$, see Galves and Leonardi (2008, Lemma 3.4) for details. Let $B = 2(d+1)(n-d)\alpha$. Then, as in Dedecker and Prieur (2005), by using Markov's inequality we obtain that, for any $t > 0$,

$$\begin{aligned}
\mathbb{P}(|N_n(w,a) - (n-d)\, p(wa)| > t) \ &\leqslant \ \min\Big(1, \frac{\mathbb{E}(|N_n(w,a) - (n-d)\, p(wa)|^r)}{t^r}\Big) \\
&\leqslant \ \min\Big(1, \Big[\frac{rB}{t^2}\Big]^{\frac{r}{2}}\Big).
\end{aligned}$$

The function $r \to (cr)^{\frac{r}{2}}$ has a unique minimum at $r_0 = \frac{1}{ec}$ and is increasing on the interval $[r_0, +\infty)$. Then, comparing the value of r_0 with 2 we have that

$$\mathbb{P}(|N_n(w,a) - (n-d)\, p(wa)| > t) \ \leqslant \ g\Big(\frac{t^2}{2eB}\Big),$$

where g is the function from $\mathbb{R}_+$ to $\mathbb{R}_+$ defined by

$$g(y) = \mathbf{1}_{y \leqslant e^{-1}} + (ey)^{-1}\mathbf{1}_{e^{-1} < y \leqslant 1} + e^{-y}\mathbf{1}_{y>1}\,.$$

Observe that $g(y) \leqslant \exp(-y + e^{-1})$ for any positive y. Then we conclude that

$$\mathbb{P}(|N_n(w,a) - (n-d)p(wa)| > t) \;\leqslant\; e^{\frac{1}{e}} \exp\Big[\frac{-t^2}{4e(d+1)(n-d)\alpha}\Big]. \qquad \square$$

Proof of Corollary 2.9. First observe that

$$p(a|w) = \frac{(n-d)p(wa)}{(n-d)p(w)}\,.$$

Then, summing and subtracting the term $\frac{N_n(w,a)}{(n-d)p(w)}$ we obtain that

$$\Big|\frac{N_n(w,a)}{N_n(w)} - \frac{(n-d)p(wa)}{(n-d)p(w)}\Big| \;\leqslant\; \frac{N_n(w,a)}{N_n(w)(n-d)p(w)}\,\big|(n-d)p(w) - N_n(w)\big| + \frac{1}{(n-d)p(w)}\,\big|N_n(w,a) - (n-d)p(wa)\big|.$$

Therefore we have that

$$\begin{aligned}\mathbb{P}\big(|\hat{p}_n(a|w) - p(a|w)| > t\,,\, N_n(w) \geqslant 1\big) &\leqslant \mathbb{P}\Big(\big|(n-d)p(w) - N_n(w)\big| > \frac{t(n-d)p(w)}{2}\Big) \\ &\quad + \mathbb{P}\Big(\big|N_n(w,a) - (n-d)p(wa)\big| > \frac{t(n-d)p(w)}{2}\Big)\,.\end{aligned}$$

We can write $N_n(w) = \sum_{b\in A} N_n(w,b)$ and $p(w) = \sum_{b\in A} p(wb)$, then the right hand side of the last inequality can be bounded above by the sum

$$\sum_{b\in A} \mathbb{P}\big(\,|N_n(w,b) - (n-d)p(wb)| \;>\; \frac{t(n-d)p(w)}{2|A|}\big) \;+\; \mathbb{P}\big(\,|N_n(w,a) - (n-d)p(wa)| \;>\; \frac{t(n-d)p(w)}{2}\big).$$

Using Theorem 2.8 we can bound above this expression by

$$e^{\frac{1}{e}}\,(|A|+1)\,\exp\Big[-\,\frac{(n-d)t^2p(w)^2}{16e\alpha|A|^2(d+1)}\Big]\,.$$

This finishes the proof of the corollary. $\square$

Proof of Theorem 2.10. First observe that

$$\begin{aligned}\mathbb{P}\big(N_n(w)\sup_{a\in A}|\hat{p}_n(a|w)-p(a|w)|^2 \;>\; t\big)\\ &\leqslant \sum_{a\in A}\mathbb{P}\big(N_n(w)\,|\hat{p}_n(a|w)-p(a|w)|^2 \;>\; t\big)\\ &= \sum_{a\in A}\mathbb{P}\big(\,|N_n(w,a)-N_n(w)p(a|w)| \;>\; \sqrt{tN_n(w)}\,\big)\,.\end{aligned} \tag{2.31}$$

Now we will fix $a \in A$ and bound above each term in the right-hand side separately. For simplifying the notation we write $\hat{p} = \hat{p}_n(a|w)$, $p = p(a|w)$, $O_n = N_n(w,a)$ and $N_n = N(w)$. Observe that $\hat{p} = O_n/N_n$. For $\lambda > 0$ define $\phi(\lambda) = \log(1-p+e^{\lambda}p)$. Let $W_d^{\lambda} = 1$ and for $n > d$ define

$$W_n^{\lambda} = e^{\lambda O_n - N_n\phi(\lambda)}\,.$$

Observe that W_n^{λ} is a martingale with respect to $\mathcal{F}_{n-1} = \sigma(X_1^{n-1})$ with $\mathbb{E}[\,W_n^{\lambda}\,] = 1$ for all $n \geqslant d$. In fact, we have that

$$O_n - O_{n-1} = \begin{cases}1\,, & \text{if } x_n = a,\ x_{n-\ell(w)}^{n-1} = w\,;\\ 0\,, & \text{c.c}\ .\end{cases}$$

and similarly

$$N_n - N_{n-1} = \begin{cases}1\,, & \text{if } x_{n-\ell(w)}^{n-1} = w\,;\\ 0\,, & \text{c.c}\ .\end{cases}$$

Observe that if $x_{n-\ell(w)}^{n-1} = w$ and w has a suffix in the context tree of the process then

$$\begin{aligned}\mathbb{E}\left[\,e^{\lambda(O_n-O_{n-1})}\,|\,\mathcal{F}_{n-1}\,\right] &= \mathbb{E}\left[\,e^{\lambda\mathbf{1}\{x_n=a\}}\,|\,\mathcal{F}_{n-1}\,\right]\\ &= e^{\phi(\lambda)}\\ &= e^{(N_n-N_{n-1})\phi(\lambda)}\,.\end{aligned} \tag{2.32}$$

On the other hand, if $x_{n-\ell(w)}^{n-1} \neq w$ the equality trivially holds. Then rearranging the terms in (2.32) we conclude that

$$\mathbb{E}\left[e^{\lambda O_n - N_n\phi(\lambda)}\,|\,\mathcal{F}_{n-1}\right] = e^{\lambda O_{n-1} - N_{n-1}\phi(\lambda)}$$

and W_n^λ is a martingale with respect to $\mathcal{F}_{n-1}$. Now divide the interval $\{1,\dots,n\}$ of possible values for N_n into "slices" $\{\theta_{k-1}+1,\dots,\theta_k\}$ of geometrically increasing size, and treat the slices independently. Assume $t > 1$ and take $\eta = 1/(t-1)$, $\theta_0 = 0$ and for $k \geqslant 1$, $\theta_k = \left\lfloor (1+\eta)^k \right\rfloor$. Let m be the first integer such that $\theta_m \geqslant n$, that is

$$m = \left\lceil \frac{\log n}{\log(1+\eta)} \right\rceil .$$

Define the events $B_k = \{\theta_{k-1} < N_n \leqslant \theta_k\} \cap \left\{N_n\,|\hat{p}-p|^2 > t\right\}$. We have

$$\mathbb{P}\left(N_n\,|\hat{p}-p|^2 \;>\; t\right) \;\leqslant\; \mathbb{P}\left(\bigcup_{k=1}^{m} B_k\right) \;\leqslant\; \sum_{k=1}^{m} \mathbb{P}\left(B_k\right). \tag{2.33}$$

Without loss of generality we can assume that $\hat{p} \geqslant p$ (the case $\hat{p} \leqslant p$ holds by symmetry). Observe that $|x-p|^2$ is a continuous increasing function for $x \in [p;1]$, with $0 \leqslant |x-p|^2 \leqslant |1-p|^2$. Let x be such that $|x-p|^2 = t/(1+\eta)^k$, that is we take

$$x \;=\; \sqrt{\frac{t}{(1+\eta)^k}} + p\,.$$

Observe that $x \in [p,1]$ unless $t/(1+\eta)^k > |1-p|^2$. But in this case we have that if $N_n \leqslant (1+\eta)^k$ then

$$t \;>\; (1+\eta)^k |1-p|^2 \;\geqslant\; N_n |\hat{p}_n - p|^2$$

so $\mathbb{P}(B_k) = 0$. So we may assume that such an x always exists over the non-empty events B_k. Moreover, on B_k we have that $|\hat{p}_n - p|^2 \geqslant t/N_n \geqslant t/(1+\eta)^k$ then we must have $\hat{p}_n \geqslant x$. Now take $\lambda = \log(x(1-p)) - \log(p(1-x))$. It can be verified that $\lambda x - \phi(\lambda) = d(x;p)$ with

$$d(x;p) = x \log \frac{x}{p} + (1-x) \log \frac{1-x}{1-p}\,.$$

Moreover, by Pinsker's Inequality (see Cesa-Bianchi and Lugosi 2006, Section A.2) we also have that $d(x;p) \geqslant |x-p|^2$. Then on B_k we have that

$$\lambda \hat{p}_n - \phi(\lambda) \;\geqslant\; \lambda x - \phi(\lambda) \;\geqslant\; |x-p|^2 \;=\; \frac{t}{(1+\eta)^k} \;\geqslant\; \frac{t}{(1+\eta)N_n}$$

therefore

$$\begin{aligned} B_k &\subset \left\{\lambda\,\hat{p}_n - \phi(\lambda) > \frac{t}{(1+\eta)N_n}\right\} \\ &\subset \left\{W_n^\lambda > e^{t/(1+\eta)}\right\}. \end{aligned}$$

As $\mathbb{E}\left[W_n^\lambda\right] = 1$, by Markov's inequality we have that

$$\begin{aligned} \mathbb{P}\,(B_k) &\leqslant \mathbb{P}\left(W_n^\lambda > e^{t/(1+\eta)}\right) \\ &\leqslant e^{-t/(1+\eta)}. \end{aligned} \tag{2.34}$$

Finally, by (2.33) we have that

$$\mathbb{P}\left(N_n|\hat{p}-p|^2 > t\right) \leqslant m\,e^{-t/(1+\eta)}.$$

But as $\eta = 1/(t-1)$, $m = \left\lceil \frac{\log n}{\log(1+\eta)} \right\rceil$ and $\log(1+1/(t-1)) \geqslant 1/t$ we obtain

$$\mathbb{P}\left(N_n\,|\hat{p}-p|^2 > t\right) \leqslant e\log(n)\,t\,e^{-t}\,.$$

Finally, by (2.31) we obtain that

$$\mathbb{P}\big(N_n(w)\sup_{a\in A}|\hat{p}_n(a|w)-p(a|w)|^2 > t\big) \leqslant e|A|\log(n)\,te^{-t}\,. \qquad \square$$

Before presenting the poof of Theorem 2.12 we state and prove two basic lemmas that are based on the results in Theorem 2.8 and Corollary 2.9.

Lemma 2.27. *Assume the process $\{X_t : t \in \mathbb{Z}\}$ satisfies Assumption 2.7. Then for any $w \in A^\star$ with $\ell(w) \leqslant d$ and any $t > 0$ such that $t < (n-d)\,p(w)$ we have*

$$\mathbb{P}(\,N_n(w) \leqslant t\,) \leqslant |A|\,e^{\frac{1}{e}}\,\exp\Big[-\frac{[(n-d)\,p(w)-t]^2}{4e\alpha|A|^2(n-d)(d+1)}\Big]. \tag{2.35}$$

Proof. Using that $N_n(w) = \sum_{a\in A} N_n(w,a)$, $p(w) = \sum_{a\in A} p(wa)$ and $t-(n-d)\,p(w) < 0$ we have that

$$\begin{aligned} \mathbb{P}(N_n(w) \leqslant t) &= \mathbb{P}\Big(\sum_{a\in A}[N_n(w,a)-(n-d)\,p(wa)] \leqslant t-(n-d)\,p(w)\Big) \\ &\leqslant \sum_{a\in A}\mathbb{P}\Big(|N_n(w,a)-(n-d)\,p(wa)| \geqslant \frac{(n-d)\,p(w)-t}{|A|}\Big). \end{aligned}$$

Using Theorem 2.8 we can bound above the right hand side of the last inequality by

$$|A|\, e^{\frac{1}{e}}\ \exp\Big[-\frac{[(n-d)\,p(w)-t]^2}{4e\alpha|A|^2(n-d)(d+1)}\Big].$$

This finishes the proof of Lemma 2.27. □

Lemma 2.28. *Assume the process $\{X_t : t \in \mathbb{Z}\}$ satisfies Assumption 2.7. Let $u, w \in A^\star$ with $\max(\ell(u), \ell(w)) \leqslant d$ and $b \in A$ such that $p(b|u) - p(b|w) > 0$. Then, for any $t < [p(b|u) - p(b|w)]^2/5$ we have that*

$$\mathbb{P}\big(D(\hat{p}_n(\cdot|u);\, \hat{p}_n(\cdot|w)) \leqslant t\big) \leqslant\ 2e^{\frac{1}{e}}\ (|A|+1)\ \exp\Big[-\frac{(n-d)t\,\min(p(w)^2,\, p(u)^2)}{32e\alpha|A|^2(d+1)}\Big].$$

Proof. By Pinsker's inequality (see Cesa-Bianchi and Lugosi 2006, Section A.2), we have that

$$\begin{aligned} D(\hat{p}_n(\cdot|u);\, \hat{p}_n(\cdot|w)) \ &\geqslant\ \frac{1}{2}\Big[\sum_{a\in A} |\hat{p}_n(a|u) - \hat{p}_n(a|w)|\Big]^2 \\ &\geqslant\ \frac{1}{2}\big[\hat{p}_n(b|u) - \hat{p}_n(b|w)\big]^2. \end{aligned} \tag{2.36}$$

Now, set $\nu = 2[p(b|u) - p(b|w)]^2/9 > t$ and define the event

$$C_n = \{\, |\hat{p}_n(b|u) - p(b|u)| \leqslant \sqrt{\nu/2}\,\} \cap \{\, |\hat{p}_n(b|w) - p(b|w)| \leqslant \sqrt{\nu/2}\,\}. \tag{2.37}$$

Observe that

$$\{\, D(\hat{p}_n(\cdot|u);\, \hat{p}_n(\cdot|w)) \leqslant t\,\} \cap C_n = \emptyset.$$

To see this note that by (2.36), if (2.37) holds then

$$D(\hat{p}_n(\cdot|u);\, \hat{p}_n(\cdot|w)) \ \geqslant\ \frac{1}{2}\Big[\,(p(b|u) - \sqrt{\nu/2}) - (p(b|w) + \sqrt{\nu/2})\Big]^2 = \ \nu\ >\ t\,.$$

Therefore, using the bounds in Corollary 2.9 we obtain for any $t < \nu$ that

$$\begin{aligned} \mathbb{P}\big(D(\hat{p}_n(\cdot|u);\, \hat{p}_n(\cdot|w)) \leqslant t\big) \ &\leqslant\ \mathbb{P}\big(|\hat{p}_n(b|u) - p(b|u)| > \sqrt{\nu/2}\,\big) \\ &\quad +\ \mathbb{P}\big(|\hat{p}_n(b|w) - p(b|w)| > \sqrt{\nu/2}\,\big) \\ &\leqslant\ 2e^{\frac{1}{e}}\ (|A|+1)\ \exp\Big[-\frac{(n-d)t\,\min(p(w)^2,\, p(u)^2)}{32e\alpha|A|^2(d+1)}\Big] \end{aligned}$$

and this concludes the proof of Lemma 2.28. □

Proof of Theorem 2.12. First observe that

$$\mathbb{P}(\hat{\tau}_{\mathrm{C}}(x_{1:n})|_K \neq \tau^*|_K) \;\leqslant\; \mathbb{P}(O_n) + \mathbb{P}(U_n)\,,$$

where

$$O_n \;=\; \{\,\tau^*|_K \prec \hat{\tau}_{\mathrm{C}}(x_{1:n})|_K\,\}$$

is the *overestimation* event and

$$U_n \;=\; \{\,\tau^*|_K \not\preceq \hat{\tau}_{\mathrm{C}}(x_{1:n})|_K\,\}$$

is the *underestimation* event (more precisely, non-overestimation event). We will divide the proof bounding separately these two events. In the case of O_n observe that

$$O_n \;\subset \bigcup_{\substack{s\in\tau^* \\ \ell(s)\leqslant K}} \bigcup_{u\in A^\star} \{\Delta_n(us) > \delta_n\}$$

and

$$\begin{aligned}\Delta_n(us) \;&=\; \sum_{b\in A} N_n(bus)\, D\left(\hat{p}_n(\cdot|bus);\, \hat{p}_n(\cdot|us)\right) \\ &=\; \sum_{b\in A} N_n(bus) \sum_{a\in A}\Big[\hat{p}_n(a|bus)\log\hat{p}_n(a|bus) - \hat{p}_n(a|bus)\log\hat{p}_n(a|us)\Big]\,.\end{aligned}$$

For any sequence $us \in A^\star$ we have that $\hat{p}_n(\cdot|us)$ are the maximum likelihood estimators of the transition probabilities $p(\cdot|us)$, therefore we have that

$$\begin{aligned}\sum_{a\in A} N_n(us,a)\log\hat{p}_n(a|us) \;&\geqslant\; \sum_{a\in A} N_n(us,a)\log p(a|us) \\ &=\; \sum_{b\in A}\sum_{a\in A} N_n(bus,a)\log p(a|bus)\end{aligned}$$

where the equality in the second line follows because us has a suffix in the context tree τ^*. Then

$$\begin{aligned}\Delta_n(us) \;&\leqslant\; \sum_{b\in A} N_n(bus) \sum_{a\in A}\Big[\hat{p}_n(a|bus)\log\hat{p}_n(a|bus) - \hat{p}_n(a|bus)\log p(a|bus)\Big] \\ &=\; \sum_{b\in A} N_n(bus)\, D\left(\hat{p}_n(\cdot|bus);\, p(\cdot|bus)\right)\,.\end{aligned} \tag{2.38}$$

Now we use a well-known inequality between two probability distributions p and q over A, that states that

$$D(p;q) \leqslant \sum_{a\in A} \frac{(p(a)-q(a))^2}{q(a)} \tag{2.39}$$

for a proof see for example Csiszár and Talata (2006b, Lemma 6.3). Hence, from (2.38) and (2.39) we obtain that

$$\mathbb{P}\left(\Delta_n(us) > \delta_n\right) \leqslant \mathbb{P}\left(\sum_{b\in A} N_n(bus) \sum_{a\in A} \frac{[\hat{p}_n(a|bus) - p(a|bus)]^2}{p(a|s)} > \delta_n\right),$$

as $p(a|bus) = p(a|s)$ for all $a, b \in A$. Taking

$$p_{\min} = \min\{p(a|s)\colon a \in A, s \in \tau^*, \ell(s) \leqslant K\}$$

we have that

$$\begin{aligned}\mathbb{P}(O_n) &\leqslant \sum_{\substack{s\in\tau^*\\ \ell(s)\leqslant K}} \sum_{u\in A^\star} \mathbb{P}\left(\Delta_n(us) > \delta_n\right)\\ &\leqslant \sum_{\substack{s\in\tau^*\\ \ell(s)\leqslant K}} \sum_{u\in A^\star} \sum_{a\in A} \mathbb{P}\Big(N_n(us) \max_{a\in A} |\hat{p}_n(a|us) - p(a|us)|^2 > \frac{p_{\min}}{|A|}\delta_n\Big).\end{aligned}$$

Now it is enough to observe that the sum over s and u has at most $\frac{n(n-1)}{2}$ terms corresponding to those sequences us such that $N_n(us) > 0$. Each term can be bounded above using Theorem 2.10 by

$$\begin{aligned}&\mathbb{P}\Big(N_n(us) \max_{a\in A} |\hat{p}_n(a|us) - p(a|us)|^2 > \frac{p_{\min}}{|A|}\delta_n\Big)\\ &\qquad\leqslant ep_{\min}\delta_n \log(n) \exp\Big\{-\frac{p_{\min}\delta_n}{|A|}\Big\}\end{aligned}$$

then

$$\mathbb{P}(O_n) \leqslant ep_{\min}\delta_n n^2 \log(n) \exp\Big\{-\frac{p_{\min}\delta_n}{|A|}\Big\}. \tag{2.40}$$

In the case of U_n we have that

$$U_n \subset \bigcup_{\substack{s\prec w\in\tau^*\\ \ell(s)\leqslant K}} \bigcap_{u\in A^\star} \{\Delta_n(us) < \delta_n\}$$

and the union is over a finite set of sequences. Denote by S the set of sequences s such that $s \prec w \in \tau^*$ and $\ell(s) \leqslant K$. By hypothesis, we can take some $bu \in A^r$ with $r \leqslant d - \ell(w)$ and $p(bus) > 0$ such that

$$\max_{a \in A} |p(a|bus) - p(a|us)| \geqslant \epsilon \,.$$

Denote by S' the (finite) set of sequences of the form bus, with $s \in S$ satisfying the above inequality. Then

$$\mathbb{P}(U_n) \leqslant \sum_{u \in S'} \mathbb{P}(\Delta_n(us) < \delta_n) \,. \tag{2.41}$$

Now, for any fixed $bus \in S'$ define the events A_n and B_n by

$$\begin{aligned} A_n &= \{\, \Delta_n(us) < \delta_n \} \\ B_n &= \left\{\, D(\hat{p}_n(\cdot|bus)\,;\, \hat{p}_n(\cdot|us)) > \epsilon^2/8 \right\} . \end{aligned}$$

Then we can bound above the probability in (2.41) by

$$\mathbb{P}\big(\, A_n \cap B_n \,\big) + \mathbb{P}\big(\, B_n^c \,\big) \,.$$

To bound the first term note that by Lemma 2.27, if n is sufficiently large so that

$$\frac{\delta_n}{(n-d)} < \frac{\epsilon^2 p(bus)}{16}$$

then we obtain

$$\begin{aligned} \mathbb{P}\big(\, A_n \cap B_n \,\big) &\leqslant \mathbb{P}\Big(\, N_n(bus) \leqslant \frac{8\delta_n}{\epsilon^2} \,\Big) \\ &\leqslant |A|\, e^{\frac{1}{e}} \, \exp\Big[- \frac{(n-d)\, p(bus)^2}{16 e \alpha |A|^2 (d+1)} \Big] . \end{aligned}$$

On the other hand, by Lemma 2.28 we have that

$$\mathbb{P}\,\big(\, B_n^c \,\big) \leqslant 2 e^{\frac{1}{e}} \, (|A| + 1) \, \exp\Big[- \frac{(n-d)\epsilon^2 \, p(bus)^2}{256 e \alpha |A|^2 (d+1)} \Big] .$$

We conclude the proof of Theorem 2.12 by observing that S' has the same cardinality as S, the set of sequences $s \prec w \in \tau^*$ with $\ell(s) \leqslant K$, therefore we obtain

$$\mathbb{P}(U_n) \leqslant 3 e^{\frac{1}{e}} \, (|A| + 1) |A|^K \, \exp\Big[\frac{-(n-d)\epsilon^2 q_{\min}^2}{256 e \alpha |A|^2 (d+1)} \Big] ,$$

with

$$q_{\min} = \min_{bus \in S'} \{\, p(bus) \,\} \;>\; 0\,. \qquad \square$$

Proof of Corollary 2.14. If $\delta_n \geqslant \frac{4|A|}{p_{min}} \log(n)$ for all n we have that

$$\mathbb{P}(\hat{\tau}_{\mathrm{C}}(x_{1:n})|_K \neq \tau^*|_K) \;\to\; 0$$

when $n \to \infty$ and this completes the proof. $\square$

Proof of Lemma 2.17. This proof is given in Csiszár and Talata (2006b, Lemma 4.4), and follows by induction on the length of the sequence w. If $\ell(w) = d$ the statement is obvious. Supposing the assertion holds for all strings of length d, we have for any w with $\ell(w) = d - 1$ that

$$\begin{aligned}\prod_{a\colon aw \in \mathcal{V}_n} V_{aw}(x_{1:n}) \;&=\; \prod_{a\colon aw \in \mathcal{V}_n} \Big(\max_{\tau_a \in \mathcal{T}_{aw}} \prod_{s \in \tau_a} e^{-\mathrm{pen}(n)}\, \hat{P}_s(x_{1:n}) \Big) \\ &=\; \max_{\tau_w \in \mathcal{T}_w\colon \ell(\tau_w) \geqslant 1} \prod_{s \in \tau_w} e^{-\mathrm{pen}(n)}\, \hat{P}_s(x_{1:n})\,.\end{aligned}$$

The second inequality follows since any family of trees $\tau_a \in \mathcal{T}_{aw}$, $a \in A$, uniquely corresponds to a tree $\tau_w \in \mathcal{T}_w$ via $\tau_w = \cup_{a \in A} \tau_{aw}$. Then by the definition of $V_w(x_{1:n})$ in (2.19) and the equality above we have that

$$\begin{aligned} V_w(x_{1:n}) \;&=\; \max\Big\{ e^{-\mathrm{pen}(n)}\, \hat{P}_w(x_{1:n}),\; \prod_{a \in A\colon aw \in \mathcal{V}_n} V_{aw}(x_{1:n}) \Big\} \\ &=\; \max\Big\{ e^{-\mathrm{pen}(n)}\, \hat{P}_w(x_{1:n}),\; \max_{\tau_w \in \mathcal{T}_w\colon \ell(\tau_w) \geqslant 1} \prod_{s \in \tau_w} e^{-\mathrm{pen}(n)}\, \hat{P}_s(x_{1:n}) \Big\} \\ &=\; \max_{\tau_w \in \mathcal{T}_w} \prod_{s \in \tau_w} e^{-\mathrm{pen}(n)}\, \hat{P}_s(x_{1:n}) \end{aligned}$$

and the first equality in the lemma follows. The last equality also follows from the last identity, by the induction hypothesis and (2.19)-(2.21). $\square$

Proof of Proposition 2.18. We must prove that a leaf in $\hat{\tau}_{\mathrm{PML}}(x_{1:n})$ is always a leaf or an internal node in $\hat{\tau}_{\mathrm{C}}(x_{1:n})$. By the characterization of $\hat{\tau}_{\mathrm{C}}(x_{1:n})$ and $\hat{\tau}_{\mathrm{PML}}(x_{1:n})$ given by equations (2.14) and (2.23), respectively, this is equivalent to prove that $\mathcal{X}_w(x_{1:n}) \leqslant C_w(x_{1:n})$ for all $w \in \mathcal{V}_n$ with $\ell(w) < d$. In fact, assume that $\mathcal{X}_w(x_{1:n}) = 1$ implies $C_w(x_{1:n}) = 1$, and take $w \in \hat{\tau}_{\mathrm{PML}}(x_{1:n})$; then, either

$\ell(w) = d$ and $w \in \hat{\tau}_{\mathrm{C}}(x_{1:n})$, or it holds that for all $u \prec w$, $\mathcal{X}_u(x_{1:n}) = 1$, which implies that $C_u(x_{1:n}) = 1$ for all $u \prec w$ and thus there exists some $v \succeq w$ such that $v \in \hat{\tau}_{\mathrm{C}}(x_{1:n})$.

Assume there exists $w \in \mathcal{V}_n$, $\ell(w) < d$, such that $\mathcal{X}_w(x_{1:n}) = 1$ and $C_w(x_{1:n}) = 0$. Note that by (2.13), $C_w(x_{1:n}) = 0$ implies $C_{uw}(x_{1:n}) = 0$ for all $uw \in \mathcal{V}_n$, $\ell(uw) \leqslant d$; hence, w can be chosen such that $\mathcal{X}_{bw}(x_{1:n}) = 0$ for any $bw \in \mathcal{V}_n$, $b \in A$. In this case we have, by the definitions (2.19) and (2.20) that

$$\begin{aligned} e^{-\mathrm{pen}(n)}\, \hat{P}_w(x_{1:n}) \;&<\; \prod_{b\colon bw \in \mathcal{V}_n} V_{bw}(x_{1:n}) \\ &= \prod_{b\colon bw \in \mathcal{V}_n} e^{-\mathrm{pen}(n)}\, \hat{P}_{bw}(x_{1:n})\,. \end{aligned} \tag{2.42}$$

The equality in the second line of the last expression follows by the fact that $\mathcal{X}_{bw}(x_{1:n}) = 0$ for any $bw \in \mathcal{V}_n$, $b \in A$; therefore we must have $V_{bw}(x_{1:n}) = e^{-\mathrm{pen}(n)}\, \hat{P}_{bw}(x_{1:n})$ for any $bw \in \mathcal{V}_n$, $b \in A$.

Now, observe that for any $a \in A$, $N_n(w,a) = \sum_{b\colon bw \in \mathcal{V}_n} N_n(bw,a)$ and $|\{b\colon bw \in \mathcal{V}_n\}| \geqslant 2$. If not, $N_n(w,a)$ would be equal to $N_n(cw,a)$ for some $c \in A$ and for all $a \in A$, implying that $\hat{P}_{cw}(x_{1:n}) = \hat{P}_w(x_{1:n})$; hence

$$\prod_{b\colon bw \in \mathcal{V}_n} V_{bw}(x_{1:n}) = V_{cw}(x_{1:n}) = e^{-\mathrm{pen}(n)}\, \hat{P}_{cw}(x_{1:n}) = e^{-\mathrm{pen}(n)}\, \hat{P}_w(x_{1:n})$$

and thus, by definition, $\mathcal{X}_w(x_{1:n}) = 0$. Using these facts, and taking logarithm on both sides of (2.42), we obtain

$$\begin{aligned} \big(\big|\{b\colon bw \in \mathcal{V}_n\}\big| - 1\big)\, \mathrm{pen}(n) \;&<; \sum_{b\colon bw \in \mathcal{V}_n} \sum_{a \in A} N_n(bw,a) \log \frac{\hat{p}_n(a|bw)}{\hat{p}_n(a|w)} \\ &= \sum_{b\colon bw \in \mathcal{V}_n} N_n(bw)\, D\left(\hat{p}_n(\cdot|bw);\, \hat{p}_n(\cdot|w)\right) \\ &= \Delta_n(w)\,. \end{aligned}$$

Therefore, if $\delta_n \leqslant \mathrm{pen}(n)$ we have $\delta_n < \Delta_n(w)$ which contradicts the fact that $C_w(x_{1:n}) = 0$. This concludes the proof of Proposition 2.18. □

Proof of Theorem 2.19. As in the proof of Theorem 2.12, observe that

$$\mathbb{P}(\hat{\tau}_{\mathrm{PML}}(x_{1:n})|_K \neq \tau^*|_K) \;\leqslant\; \mathbb{P}(O_n') + \mathbb{P}(U_n')\,,$$

where

$$O'_n \;=\; \{\, \tau^*|_K \prec \hat{\tau}_{\mathrm{PML}}(x_{1:n})|_K \,\}$$

is the *overestimation* event and

$$U'_n \;=\; \{\, \tau^*|_K \not\preceq \hat{\tau}_{\mathrm{PML}}(x_{1:n})|_K \,\}$$

is the underestimation (more precisely, non-overestimation) event. By Proposition 2.18, for a given $\mathrm{pen}(n)$ we can take $\delta_n = \mathrm{pen}(n)$ and we have that

$$\hat{\tau}_{\mathrm{PML}}(x_{1:n}) \preceq \hat{\tau}_{\mathrm{C}}(x_{1:n})$$

and thus $O'_n \subset O_n$. Then

$$\mathbb{P}(O'_n) \;\leqslant\; e p_{\min} \mathrm{pen}(n) n^2 \log(n) \exp\Big\{-\frac{p_{\min}\mathrm{pen}(n)}{|A|}\Big\}$$

as obtained in (2.40). The proof for bounding the probability of U'_n follows almost identical to the analogue in the proof of Theorem 2.12. We first observe that

$$U'_n \;\subset \bigcup_{\substack{s \prec w \in \tau^* \\ \ell(s) \leqslant K}} \{\mathcal{X}_s(x_{1:n}) = 0\}\,.$$

By hypothesis, for any $s \prec w \in \tau^*$ with $\ell(s) \leqslant K$ there exists $r \leqslant d - \ell(s)$ and $u \in A^r$ such that

$$\max_{a \in A} |p(a|us) - p(a|s)| \;\geqslant\; \epsilon \;>\; 0\,.$$

Denote by S' the (finite) set of sequences us satisfying the above inequality. Now let $s \prec w \in \tau^*$ with $\ell(s) \leqslant K$ and $us \in S'$. Observe that

$$\mathbb{P}\left(\,\mathcal{X}_s(x_{1:n}) = 0\,\right) \;=\; \mathbb{P}\Bigg(\prod_{b \in A\colon bs \in \mathcal{V}_n} V_{bs}(x_{1:n}) \;\leqslant\; e^{-\mathrm{pen}(n)}\,\hat{P}_s(x_{1:n})\Bigg)\,.$$

If $u = (u_1 \ldots u_r)$, denote by $A_i = A \setminus \{u_i\}$ and let τ be the tree given by

$$\tau = \cup_{i=2}^{r+1} \cup_{b \in A_i} \{b u_i^r s\} \cup \{us\}\,.$$

By Lemma 2.17, for any $bs \in \mathcal{V}_n$ we have that

$$V_{bs}(x_{1:n}) = \max_{\tau' \in \mathcal{T}_n} \prod_{v \in \tau'_{bs}} e^{-\mathrm{pen}(n)}\,\hat{P}_v(x_{1:n})\,.$$

Therefore

$$\begin{aligned}\mathbb{P}\Big(\prod_{b\in A:\, bs\in\mathcal{V}_n} V_{bs}(x_{1:n}) \;\leqslant\; e^{-\mathrm{pen}(n)}\,\hat{P}_s(x_{1:n})\Big)\\ \leqslant\; \mathbb{P}\left(\prod_{v\in\tau} e^{-\mathrm{pen}(n)}\,\hat{P}_v(x_{1:n}) \;\leqslant\; e^{-\mathrm{pen}(n)}\,\hat{P}_s(x_{1:n})\right)\end{aligned} \tag{2.43}$$

by noticing that

$$\begin{aligned}\prod_{b\in A:\, bs\in\mathcal{V}_n}\max_{\tau'\in\mathcal{T}_n}\prod_{v\in\tau'_{bs}} e^{-\mathrm{pen}(n)}\,\hat{P}_v(x_{1:n}) \;&\geqslant\; \prod_{b\in A:\, bs\in\mathcal{V}_n}\prod_{v\in\tau_{bs}} e^{-\mathrm{pen}(n)}\,\hat{P}_v(x_{1:n})\\ &\geqslant\; \prod_{v\in\tau} e^{-\mathrm{pen}(n)}\,\hat{P}_v(x_{1:n})\,.\end{aligned}$$

Applying logarithm and using that $N_n(s,a)=\sum_{v\in\tau}N_n(v,a)$ for any $a\in A$ we can write the probability in (2.43) by

$$\begin{aligned}\mathbb{P}\Big(\sum_{v\in\tau} N_n(v)D(\hat{p}_n(\cdot|v)\,;\,\hat{p}_n(\cdot|s)) \;\leqslant\; (|\tau|-1)\mathrm{pen}(n)\Big)\\ \leqslant\; \mathbb{P}\Big(N_n(us)D(\hat{p}_n(\cdot|us)\,;\,\hat{p}_n(\cdot|s)) \;\leqslant\; (|\tau|-1)\mathrm{pen}(n)\Big)\,.\end{aligned} \tag{2.44}$$

Define the events A_n and B_n by

$$\begin{aligned}A_n \;&=\; \{\,N_n(us)D(\hat{p}_n(\cdot|us)\,;\,\hat{p}_n(\cdot|s)) \;\leqslant\; (|\tau|-1)\mathrm{pen}(n)\,\}\\ B_n \;&=\; \{\,D(\hat{p}_n(\cdot|us)\,;\,\hat{p}_n(\cdot|s)) > \epsilon^2/8\,\}\,.\end{aligned}$$

Then we can bound above the probability in (2.44) by $\mathbb{P}\big(A_n\cap B_n\big)+\mathbb{P}\big(B_n^c\big)$. To bound the first term note that by Lemma 2.27, if n satisfies

$$\frac{\mathrm{pen}(n)}{n-d} \;<\; \frac{\epsilon^2 p(us)}{16(|\tau|-1)}$$

then, using the bound $|\tau|-1\leqslant|A|r\leqslant|A|d$ we obtain

$$\begin{aligned}\mathbb{P}\big(A_n\cap B_n\big) \;&\leqslant\; \mathbb{P}\Big(N_n(us)\;\leqslant\;\frac{8(|\tau|-1)\mathrm{pen}(n)}{\epsilon^2}\Big)\\ &\leqslant\; |A|e^{\frac{1}{e}}\exp\Big(-\frac{(n-d)\,p(us)^2}{16e\alpha|A|^2(d+1)}\Big)\,.\end{aligned}$$

On the other hand, by Lemma 2.28 we have

$$\mathbb{P}\left(B_n^c\right) \leqslant 2e^{\frac{1}{e}}(|A|+1)\exp\Big[-\frac{(n-d)\epsilon^2 p(us)^2}{256e\alpha|A|^2 1(d+1)}\Big].$$

We conclude the proof of Theorem 2.19 by observing that we only have a finite number of sequences $s \prec w \in \tau^*$ with $\ell(s) \leqslant K$, therefore we obtain

$$\mathbb{P}(U_n) \leqslant 3e^{\frac{1}{e}}(|A|+1)|A|^K \exp\Big[-\frac{(n-d)\epsilon^2 q_{\min}^2}{256e\alpha|A|^2 1(d+1)}\Big]$$

with

$$q_{\min} = \min_{us\in S'}\{\, p(us)\,\} > 0\,. \qquad \square$$

Proof of Corollary 2.20. If $\mathrm{pen}(n) \geqslant \frac{4|A|}{p_{min}}\log(n)$ for all n we have that

$$\mathbb{P}(\hat{\tau}_{\mathrm{PML}}(x_{1:n})|_K \neq \tau^*|_K) \to 0$$

when $n \to \infty$ and this completes the proof. $\square$

Proof of Lemma 2.22. Denote by $\tau^1 = \hat{\tau}_{\mathrm{SMC}}(c_1)$ and $\tau^2 = \hat{\tau}_{\mathrm{SMC}}(c_2)$. Suppose that it is not true that $\tau^1 \succeq \tau^2$. Then there exists a sequence $w \in \tau^1$ and $w' \in \tau^2$ such that w is a proper suffix of w'. This implies that $\tau^2_w \neq \emptyset$. Since τ^2 is irreducible we have that $|\tau^2_w| \geqslant 2$. Then, using the definition of maximizing tree we obtain

$$\begin{aligned}
\log \hat{P}_w(x_{1:n}) &\geqslant \sum_{w'\in\tau^2_w} \log \hat{P}_{w'}(x_{1:n}) + c_1(\mathrm{df}(w) - \sum_{w'\in\tau^2_w}\mathrm{df}(w'))\log n \\
&\geqslant \sum_{w'\in\tau^2_w} \log \hat{P}_{w'}(x_{1:n}) + c_2(\mathrm{df}(w) - \sum_{w'\in\tau^2_w}\mathrm{df}(w'))\log n \\
&> \log \hat{P}_w(x_{1:n}),
\end{aligned}$$

which is a contradiction. The first inequality follows from the assumption that $\tau^1 = \hat{\tau}_{\mathrm{SMC}}(c_1)$ and the second equality in (2.22). To derive the second inequality we use the fact that $0 < c_1 < c_2$ and $\mathrm{df}(w) - \sum_{w'\in\tau^2_w}\mathrm{df}(w') < 0$. Finally, the last inequality leading to the contradiction follows from $\tau^2 = \hat{\tau}_{\mathrm{SMC}}(c_2)$ and again the second equality in (2.22). We conclude that $\tau^1 \succeq \tau^2$. $\square$

Proof of Theorem 2.23. The fact that C_n is totally ordered follows from Lemma 2.22. On the other hand, by Csiszár and Talata (2006b) we know that the BIC context tree estimator is strongly consistent for any constant $c > 0$. Therefore it follows that eventually almost surely $\tau^* \in C_n$ as $n \to \infty$. □

Proof of Theorem 2.24. To prove (2.28) let $\tau \in C_n$ be such that $\tau \prec \tau^*$. Then

$$\begin{aligned}
&\log \hat{P}_\tau(x_{1:n}) - \log \hat{P}_{\tau^*}(x_{1:n}) \\
&\quad = \sum_{w' \in \tau, a \in A} N_n(w', a) \log \hat{p}_n(a|w') - \sum_{w \in \tau^*, a \in A} N_n(w, a) \log \hat{p}_n(a|w) \\
&\quad = \sum_{w' \in \tau} \sum_{w \in \tau^*, w \succ w'} \sum_{a \in A} N_n(w, a) \log \frac{\hat{p}_n(a|w')}{\hat{p}_n(a|w)}.
\end{aligned}$$

Dividing by n and using Jensen's inequality we have that

$$\begin{aligned}
&\sum_{w' \in \tau} \sum_{w \in \tau^*, w \succ w'} \sum_{a \in A} \frac{N_n(w, a)}{n} \log \frac{\hat{p}_n(a|w')}{\hat{p}_n(a|w)} \\
&\qquad \longrightarrow \sum_{w' \in \tau'} \sum_{w \in \tau^*, w \succ w'} \sum_{a \in A} p(wa) \log \frac{p(a|w')}{p(a|w)} \; < \; 0,
\end{aligned}$$

as n goes to $+\infty$, by the minimality of τ^*. Then, for a sufficiently large n there exists a constant $c(\tau^*, \tau) > 0$ such that

$$\log \hat{P}_{\tau^*}(x_{1:n}) - \log \hat{P}_\tau(x_{1:n}) \geqslant c(\tau^*, \tau) n.$$

To prove (2.29) observe that

$$\begin{aligned}
&\log \hat{P}_{\tau'}(x_{1:n}) - \log \hat{P}_\tau(x_{1:n}) \\
&\quad = \sum_{w' \in \tau', a \in A} N_n(w', a) \log \hat{p}_n(a|w') - \sum_{w \in \tau, a \in A} N_n(w, a) \log \hat{p}_n(a|w) \\
&\quad \leqslant \sum_{w' \in \tau', a \in A} N_n(w', a) \log \hat{p}_n(a|w') - \sum_{w \in \tau, a \in A} N_n(w, a) \log p(a|w) \\
&\quad = \sum_{w \in \tau} \sum_{w' \in \tau', w' \succ w} \sum_{a \in A} N_n(w', a) \log \frac{\hat{p}_n(a|w')}{p(a|w)} \\
&\quad = \sum_{w \in \tau} \sum_{w' \in \tau', w' \succ w} N_n(w') D(\hat{p}_n(\cdot|w'); p(\cdot|w)).
\end{aligned}$$

By Csiszár and Talata (ibid., Lemmas 6.2 and 6.3) we have that, if n is sufficiently large, we can bound above the last term by

$$\sum_{w\in\tau}\sum_{w'\in\tau',w'\succ w} N_n(w')\sum_{a\in A}\frac{[\hat{p}_n(a|w')-p(a|w)]^2}{p(a|w)} \leqslant \sum_{w\in\tau}\sum_{w'\in\tau',w'\succ w} N_n(w')\frac{1}{p_{\min}}|A|\frac{\delta\log n}{N_n(w')},$$

where $p_{\min}=\min_{w\in\tau^*,a\in A}\{p(a|w)\colon p(a|w)>0\}$. This concludes the proof of Theorem 2.24. □

Proof of Theorem 2.25. Follows straightforward from Theorem 2.23 and Theorem 2.24. □

2.5 Exercises

Exercise 2.1. Let $\{X_n\colon n\in\mathbb{Z}\}$ be a Markov chain of order 2 assuming values in the alphabet $A=\{0,1\}$, with transition probabilities given by

$$P=\begin{array}{c} \\ 00\\01\\10\\11\end{array}\begin{array}{c}\begin{array}{cc}0 & 1\end{array}\\\begin{pmatrix}2/3 & 1/3\\1/2 & 1/2\\2/5 & 3/5\\2/5 & 3/5\end{pmatrix}\end{array}.$$

For example, for $w=01$ and $a=1$ we have $p(a|w)=p(1|01)=P_{2,2}=1/2$.

(a) Which is the context tree of this process?

(b) Compute the probability of the sequence $x_{3:12}=0001101010$ conditioned on $x_{1:2}=10$.

Exercise 2.2. Let $\{X_n\colon n\in\mathbb{Z}\}$ be a stationary Markov chain of order 1 assuming values in the alphabet $A=\{0,1\}$ with transition probabilities $0<p(1|0)<1$ and $0<p(1|1)<1$. Consider the sequence $\{\xi_n\colon n\in\mathbb{Z}\}$ of independent and identically distributed random variables, independent of $\{X_n\colon n\in\mathbb{Z}\}$, assuming values in $\{0,1\}$ and such that $\mathbb{P}(\xi_1=1)=q$, for some $q\in(0,1)$. Let $\{Y_n\colon n\in\mathbb{Z}\}$ defined by $Y_n=\xi_n X_n$.

(a) Find the context tree of the process $\{Y_n : n \in \mathbb{Z}\}$.

(b) Compute $\mathbb{P}(Y_n = 1 | Y_{n-1} = 1)$ and $\mathbb{P}(Y_n = 1 | Y_{n-1} = 0, Y_{n-2} = 1)$.

(c) Compute $\mathbb{P}(Y_n = 1 | Y_{n-1} = 0, Y_{n-1} = 0, Y_{n-3} = 1)$.

Exercise 2.3. Let $\{X_n : n \in \mathbb{Z}\}$ be a stochastic chain with memory of variable length with values in $A = \{0, 1\}$. Assume the context tree of the process is $\tau = \{10_{-k:-1} : k = 0, \dots\} \cup \{0_{-\infty:-1}\}$ and the transition probabilities are given by

$$p(1|10_{-k:-1}) = q_k \in (0, 1),\ k \geqslant 0, \quad \text{and} \quad p(1|0_{-\infty:-1}) = 0.$$

Compute $\mathbb{P}(X_1 = 1, X_2 = 0, X_3 = 0, X_4 = 1, X_5 = 0 \,|\, X_0 = 1)$.

Exercise 2.4. Prove that the set of contexts of a process with measure $\mathbb{P}$ is an irreducible tree, as specified in Definition 2.1.

Exercise 2.5. Let $x_{1:11} = 01101000100$ be a sample of a stochastic chain with memory of variable length assuming values in $A = \{0, 1\}$.

(a) Compute $N_{11}(0, 0)$, $N_{11}(0, 1)$, $N_{11}(1, 0)$ and $N_{11}(1, 1)$, taking $d = 1$ in (2.3).

(b) Compute the maximum likelihood estimators for the transition probabilities $p(0|0)$, $p(1|0)$, $p(0|1)$ and $p(1|1)$.

(c) Compute the maximum possible value of the likelihood function $L_\tau(q; x_{1:11})$ in (2.5) for $\tau = \{0, 1\}$.

Exercise 2.6. Prove that given a tree τ, the family of probability distributions defined in (2.6) corresponds to the maximum likelihood estimators of the transition probabilities associated to τ.

Exercise 2.7. Use the Ergodic Theorem to prove that for a given sequence $w \in \tau$ and symbol $a \in A$, the maximum likelihood estimator $\hat{p}_n(a|w)$ defined on (2.6) converges almost surely to $p(a|w)$ as $n \to \infty$.

Exercise 2.8. Prove that the set of maximal sequences in the set $\mathcal{V}_n$ defined by (2.11), i.e. the sequences that are not suffixes of other sequences in $\mathcal{V}_n$ is a tree. Give an example showing that this tree is not necessarily irreducible.

Exercise 2.9. Consider the sample $x_{1:11} = 00110100001$ over the alphabet $A = \{0, 1\}$. Compute the estimated trees $\hat{\tau}_{\text{C}}$ and $\hat{\tau}_{\text{PML}}$ given by (2.14) and (2.23), respectively, for $\delta_n = \text{pen}(n) = 0.15$.

Exercise 2.10. Prove that if τ^* is the context tree of the process generating the sample $x_{1:n}$, then for any $K \in \mathbb{N}$ and for a sufficiently large d, (2.15) holds for some $\epsilon > 0$.

3 Applications of stochastic chains to Biology and Linguistics

In this chapter we show how the stochastic chains with memory of variable length and some extended versions of this model can be applied in real data problems. We first present an application to the problem of classifying protein sequences into families, considered first in Bejerano and Yona (2001) using the *Probabilistic Suffix Tree* (PST) algorithm and generalized for *sparse stochastic chains* in Leonardi (2006) using an algorithm called *Sparse Probabilistic Suffix Trees* (SPST). We then present an application of the SMC algorithm described in Section 2.3.3 to detect differences in rhythmic patterns in European and Brazilian codified written texts, based on the work by Galves, Galves, et al. (2012).

3.1 Classification of protein sequences

Proteins are large biomolecules that are comprised of one or more long chains of amino acid residues[1]. So, the primarily structure of a protein or protein domain can be seen as a sequence of symbols in an alphabet of size 20, the different possible amino acids occurring in the genetic code.

A central problem in Bioinformatics is to determine the function of a protein using the information contained in its amino acid sequence. In general, this can be done by comparing the sequence of amino acids to that of proteins with a known function, or what is equivalent, to proteins in different protein families. In other words, given a protein family $\mathcal{F}$ and a new sequence of amino acids $x = x_{1:n}$, we want to know if x belongs to $\mathcal{F}$ or not. To answer this question, we can first construct a model for the family $\mathcal{F}$ using the sequences already classified on it, and then compute a *score* that is related to the probability that this model generates the sequence. Depending on this value, we classify the sequence as belonging to the family or not.

One possible model for the protein sequences on each family is the stochastic chains with memory of variable length introduced in Chapter 2. This was considered in the work Bejerano and Yona (2001), were the authors applied an algorithm called *Probabilistic Suffix Trees* (PST), originally introduced in Ron, Singer, and Tishby (1996), to the protein family classification task. Besides some minor differences, the PST algorithm is equivalent to the algorithm Context described in Section 2.3.1, as we show in the following section.

The PST algorithm was generalized to model *sparse chains* in Leonardi (2006), leading to the *Sparse Probabilistic Suffix Trees* (SPST). Sparse chains are those chains with memory of variable length were some contexts share the same associated probability distributions. In other words, at some positions of the past sequence, some symbols can be exchanged with others without modifying the transition probabilities. This property was particularly appealing in the protein family classification problem, were the alphabet size is considerably large, but some amino acids can be substituted by others with the same physicochemical properties. We describe this model and the SPST algorithm on Section 3.1.2. Finally in Section 3.1.3 we show the results of the PST and SPST algorithms on the protein classification task, for some families in the Pfam database (Bateman et al. 2004, release 1.0).

[1] `https://en.wikipedia.org/wiki/Protein`

3.1.1 The PST algorithm

Let A denotes the alphabet of twenty amino acids and let $x^1, x^2, \ldots, x^m$ be the sample of sequences belonging to a protein family $\mathcal{F}$. For each $i = 1, \ldots, m$ we denote by $x^i = x_{1:n_i}$ the i-th sequence in the sample, where n_i denotes the length of sequence x^i. As in Chapter 2, given a sequence w and a symbol a we define $N_{i,n_i}(w, a)$ as the number of occurrences of w followed by symbol a in the sample $x^i_{1:n_i}$, taking $d = \ell(w)$. That is

$$N_{i,n_i}(w,a) = \sum_{t=\ell(w)+1}^{n_i} \mathbf{1}\{x^i_{(t-\ell(w)):(t-1)} = w, x^i_t = a\}\,. \tag{3.1}$$

We define also $N_{i,n_i}(w)$ as the sum of $N_{i,n_i}(w,a)$ over all $a \in A$, that is

$$N_{i,n_i}(w) = \sum_{a\in A} N_{i,n_i}(w,a)\,. \tag{3.2}$$

As in this case we have possibly many sequences in the same family that are considered as samples of the model, we define the empirical transition probabilities as

$$\hat{p}(a|w) \;=\; \frac{\sum_{i=1}^m N_{i,n_i}(w,a)}{\sum_{i=1}^m N_{i,n_i}(w)} \qquad a \in A\,. \tag{3.3}$$

We also compute, for any sequence w, the relative frequency of the sequence in the sample, i.e

$$\hat{p}(w) \;=\; \frac{\sum_{i=1}^m N_{i,n_i}(w)}{\sum_{i=1}^m (n_i - \ell(w))} \qquad a \in A\,.$$

We now present the procedure for building a PST as presented in Bejerano and Yona (2001). The procedure uses five external parameters: L the maximal memory length (that is equivalent to the parameter d in the algorithm Context), $p_{\min}$ the minimal probability with which strings are required to occur, r which is a simple measure of the difference between the prediction of the candidate at hand and its direct suffix sequence, γ the smoothing factor, and α, a parameter that together with the smoothing probability defines the significance threshold for a conditional appearance of a symbol. We explain in more detail the roles played by these parameters below.

The PST algorithm, as described in Bejerano and Yona (ibid.), works as follows:

PST-Algorithm $(L,\ p_{\min}, r, \gamma, \alpha)$

1. *Initialization:* Let τ consist of a single root node (with label λ), and let $\bar{S} \leftarrow \{a : a \in A \text{ and } \hat{p}(a) \geqslant p_{\min}\}$.
2. *Building the PST skeleton:* while $\bar{S} \neq \emptyset$, pick any $s \in \bar{S}$ and do:
   (a) Remove s from $\bar{S}$.
   (b) If there exists a symbol $a \in A$ such that
   $$\hat{p}(a|s) \geqslant (1+\alpha)\gamma$$
   and
   $$\frac{\hat{p}(a|s)}{\hat{p}(a|suf(s))} \begin{cases} \geqslant r \\ \text{or} \\ \leqslant 1/r \end{cases}$$
   then add to τ the node corresponding to s and all the nodes on the path to s from the deepest node in τ that is a suffix of s.
   (c) If $\ell(s) < L$ then add the strings $\{bs : b \in A \text{ and } \hat{p}(bs) \geqslant p_{\min}\}$ (if any) to $\bar{S}$.
3. *Estimation of the transition probabilities*: assign to each node in τ, associated with a sequence s, the transition probability distribution over A given by (3.3).

One can show, by appropriately choosing the parameters L, $p_{\min}$, $\gamma_{\min}$, r and α, that the tree estimated by the PST algorithm is the same as the tree generated by the algorithm Context in (2.14), for appropriate values of d and δ_n, see Exercise 3.3. An example of a context tree generated by the PST algorithm is given in Figure 3.1.

In Bejerano and Yona (ibid.), the PST algorithm was used to classify proteins into families from the Pfam database. These results were compared with the state of art approaches at that time, that were Hidden Markov Models (HMM). We refer the reader to Bejerano and Yona (ibid.) for the results on the performance of these methods and further details on the use of PST for protein classification. In Section 3.1.3 we show some partial results of the PST algorithm in comparison to a generalization for sparse sequences, called SPST, that we describe in the following section.

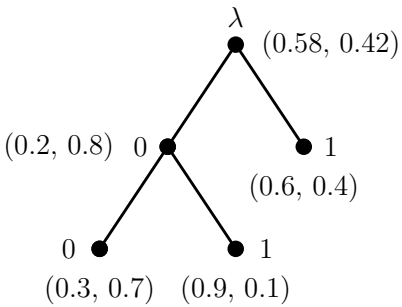


Figure 3.1: A probabilistic suffix tree over the alphabet $A = \{0, 1\}$ obtained by the PST algorithm. The set of contexts is $\{00, 01, 10, 11\}$.

3.1.2 SPST for sparse sequences

As mentioned before, the PST algorithm was successfully used for protein classification, but it was noticed that its performance decreases with less conserved families. For that reason, some attempts were made later to generalize this model for sparse sequences, as for example in Eskin, Noble, and Singer (2003) and García and González-López (2017). Although very attractive, these methods have the major disadvantage of having computationally expensive algorithms. Another generalization of PST to model sparse sequences was introduced in Leonardi (2006), where an algorithm called SPST was introduced. We describe this approach here.

An *Sparse Markov Chain* (SMC) is a stochastic chain with memory of variable length in which some contexts can be grouped together into an equivalence class. In an SMC the transition probabilities satisfies that

$$\begin{aligned}\mathbb{P}[X_n = x_n \mid X_0 = x_0, \ldots, X_{n-1} = x_{n-1}] \; = \\ P[X_n = x_n \mid X_{n-k} \in B_{n-k}, \ldots, X_{n-1} \in B_{n-1}]\,,\end{aligned}$$

where $B_i \subset A$ for all $i = n-k, \ldots, n-1$. This property induces a partition of the set of contexts of the process. Then, the contexts in an SMC are given by the equivalence classes denoted by the sequences of subsets $\bar{w} = (B_{-k}, \ldots, B_{-1})$. By an abuse of notation we write $w_{-k:-1} \in \bar{w}$ if the context $w_{-k:-1}$ belongs to the equivalence class $\bar{w}$, that is if $w_{-i} \in B_{-i}$ for all $i = 1, \ldots, k$. A tree

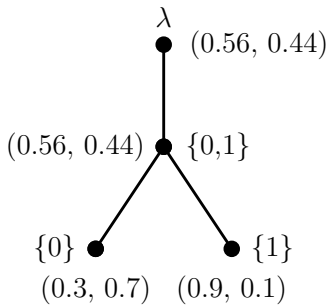


Figure 3.2: A sparse tree over the alphabet $A = \{0, 1\}$. The set of contexts is $\{00, 01, 10, 11\}$, and the equivalence classes are $\{00, 01\}$ and $\{10, 11\}$. This means that $p(a|00) = p(a|01)$ and $p(a|10) = p(a|11)$, for all $a \in A$.

representation for this type of contexts can be seen in Figure 3.2. To each node of the tree it is associated a conditional probability distribution over the next symbol given the sparse context represented by the node. For example, if we want to compute the probability of the sequence 01001 in the model given by Figure 3.2 we compute

$$\begin{aligned} p(01001) &= p(0) \times p(1|0) \times p(0|01) \times p(0|10) \times p(1|00) \\ &= 0.56 \times 0.44 \times 0.3 \times 0.9 \times 0.7 \,. \end{aligned} \tag{3.4}$$

Given a sample of protein sequences $x^1, \ldots, x^m$, each one of length n_i, $i = 1, \ldots, m$, a sparse context $\bar{w} = (B_{-k}, \ldots, B_{-1})$ and a symbol $a \in A$, we denote by $N(\bar{w}, a)$ the number of occurrences of the sparse context $\bar{w}$ followed by symbol a in the sample. That is,

$$N(\bar{w}, a) = \sum_{i=1}^{m} \sum_{j=k+1}^{n_i} \mathbf{1}\{x^i_{j-k} \in B_{-k}, \ldots, x^i_{j-1} \in B_{-1}\} \mathbf{1}\{x^i_j = a\} \,. \tag{3.5}$$

We also define

$$N(\bar{w}) = \sum_{a \in A} N(\bar{w}, a) \,. \tag{3.6}$$

Observe that $N(\bar{w}, a)$ and $N(\bar{w})$ in (3.5) and (3.6) correspond to the sum of the counters $N_{i,n_i}(w, a)$ and $N_{i,n_i}(w)$ in (3.1) and (3.2) for all $w \in \bar{w}$, respectively.

Given a sparse context $\bar{w} = (B_{-k}, \ldots, B_{-1})$ and symbols $a, b \in A$ we denote by $a\bar{w}$ and $b\bar{w}$ the sparse contexts $(\{a\}, B_{-k}, \ldots, B_{-1})$ and $(\{b\}, B_{-k}, \ldots, B_{-1})$, respectively. We also denote by $[a, b]\bar{w}$ the sparse context $(\{a, b\}, B_{-k}, \ldots, B_{-1})$. Using this notation we define the operator $\Delta_{\bar{w}}(a, b)$ as the logarithm of the ratio between the estimated probability of the sequences in the model that has the contexts $a\bar{w}$ and $b\bar{w}$ as equivalent and the model that distinguishes the two contexts as different. That is,

$$\begin{aligned}
\Delta(\bar{w}, a, b) &= \log\left[\prod_{c \in A} \frac{\hat{p}(c|[a,b]\bar{w})^{N([a,b]\bar{w},c)}}{\hat{p}(c|a\bar{w})^{N(a\bar{w},c)}\, \hat{p}(c|b\bar{w})^{N(b\bar{w},c)}} \right] \\
&= \sum_{c \in A} N([a,b]\bar{w}, c) \log\left(\frac{N([a,b]\bar{w}, c)}{N([a,b]\bar{w})} \right) \\
&\quad - \sum_{c \in A} N(a\bar{w}, c) \log\left(\frac{N(a\bar{w}, c)}{N(a\bar{w})} \right) \\
&\quad - \sum_{c \in A} N(b\bar{w}, c) \log\left(\frac{N(b\bar{w}, c)}{N(b\bar{w})} \right).
\end{aligned} \tag{3.7}$$

Note that $N([a, b]\bar{w}, c) = N(a\bar{w}, c) + N(b\bar{w}, c)$ and $N([a, b]\bar{w}) = N(a\bar{w}) + N(b\bar{w})$.

Using the preceding definitions we can specify how the SPST algorithm works. The free parameters that must be specified by the user are: the maximum depth of the tree, L, the minimum number of times that a sparse context has to be seen in the sample to be considered, $N_{\min}$, and a cutoff parameter that establishes the equivalence between two contexts, $r_{\max}$. The SPST algorithm works as follows. It starts with a tree τ consisting of a single root node. At each step, for every terminal node in τ labeled by a sparse sequence $\bar{w}$ with length less than L and for every symbol $a \in A$, the child a is added to node $\bar{w}$ if $N(a\bar{w}) \geqslant N_{\min}$. Then, for every pair of children of $\bar{w}$, a and b, we test the equivalence of the sparse contexts $a\bar{w}$ and $b\bar{w}$ using the Δ operator. That is, we compute $\Delta(\bar{w}, a, b)$ for every pair of symbols $(a, b) \in A^2$ added to node $\bar{w}$, and choose the minimum between all the pairs. If this minimum is smaller than $r_{\max}$, the corresponding nodes are merged together into a single node. This procedure is iterated with the new set of children of $\bar{w}$ until no more nodes can be merged. Taking the minimum of $\Delta(\bar{w}, a, b)$ between all the possible pairs (a, b) ensures the independence of the order in which the

tests are performed. To conclude the construction of the tree we assign to each node a transition probability distribution. This distribution gives the probability of a symbol in A given the sparse context between the node and the root of the tree. The transition probabilities are estimated, as usual, by the maximum likelihood estimators. That is, given a sparse context $\bar{w} = (B_{-k}, \ldots, B_{-1})$, the estimated probability of a symbol $a \in A$ given the sparse context $\bar{w}$ is given by

$$\hat{p}(a|\bar{w}) = \frac{N(\bar{w}, a)}{N(\bar{w})}. \tag{3.8}$$

We summarize the steps of the SPST algorithm below.

SPST-Algorithm $(N_{min}, r_{\max}, L)$

1. *Initialization*: let τ be a tree consisting of a single root node, and let

$$\bar{S} = \{a : a \in A \text{ and } N(a) \geqslant N_{\min}\}.$$

2. *Iteration*: while $\bar{S} \neq \emptyset$ do:

   (a) Remove $\bar{u}$ of $\bar{S}$ and add $\bar{u}$ to τ. Then remove all sparse contexts $\bar{w} \in \bar{S}$ such that $suf(\bar{w}) = suf(\bar{u})$ and add them to τ. Let C denote the set of contexts added to τ in this step.

   (b) Compute

$$r = \min\{\Delta(suf(\bar{u}), a, b) : a(suf(\bar{u})), b(suf(\bar{u})) \in C\},$$

   and

$$(a^*, b^*) = \underset{a,b \in A}{\operatorname{argmin}}\{\Delta(suf(\bar{u}), a, b) : a(suf(\bar{u})), b(suf(\bar{u})) \in C\}.$$

   (c) If $r < r_{max}$ merge a^* and b^* in a single node. Replace the contexts $a^*(suf(\bar{u}))$ and $b^*(suf(\bar{u}))$ in C by the context $[a^*, b^*]suf(\bar{u})$.

   (d) Repeat steps b. and c. until no more changes can be made in C.

   (e) For each sparse context $\bar{w} \in C$, if $\ell(w) < L$ then add the set $\{a\bar{w} : a \in A \text{ and } N(a\bar{w}) \geqslant N_{min}\}$ (if any) to $\bar{S}$.

3. *Estimation of the transition probabilities*: assign to each node in τ, associated with a sparse context $\bar{w}$, the transition probability distribution over A given by (3.8).

3.1.3 Prediction and results on the classification task

Once the context tree model (PST or SPST), that we will further denote by $\mathcal{M}$, has been constructed using the sequences already classified into a family $\mathcal{F}$, we need a way of predicting if a new observed sequence $x_{1:n}$ belongs to $\mathcal{F}$ or not. To do this we can compute a *score* given by

$$S(x_{1:n}) = \frac{1}{n} \log[\widetilde{p}_{\mathcal{M}}(x_{1:n})]\,, \tag{3.9}$$

where $\widetilde{p}_{\mathcal{M}}$ is the smoothed probability distribution derived from $\hat{p}$. That is

$$\widetilde{p}_{\mathcal{M}}(x_{1:n}) = \prod_{i=1}^{n} [(1 - |A|\gamma_{min})\hat{p}(x_i|w(x_1, \ldots, x_{i-1})) + \gamma_{min}]$$

where $w(x_1, \ldots, x_{i-1})$ is the context (respectively sparse context) corresponding to the sequence $x_1, \ldots, x_{i-1}$ in the estimated tree by the PST (respectively by the SPST) algorithm. The parameter $\gamma_{\min}$ is a smoothing parameter to avoid zero probabilities, and therefore, a $-\infty$ score.

Sometimes the region of high similarity between the sequences in a protein family is considerably smaller than the length of the sequences. This is because a protein sequence can be composed by several domains, performing different functions in the cell. Then, computing the score S over the entire sequence $x_{1:n}$ may not be appropriate. For this reason we propose a change in the computation of the score S, and called it S'. In this case we fix an integer M, and for sequences with length $n > M$ we compute the score $S'(x_{1:n})$ by

$$S'(x_{1:n}) = \max_{j=0,\ldots,n-M} \left\{ \frac{1}{M} \log[\widetilde{p}_{\mathcal{M}}(x_{j+1} \ldots x_{j+M})] \right\}.$$

In the case $n \leqslant M$, the score is computed using S as before. The algorithm that implements the score S' is called F-SPST.

The performance of the three approaches, PST, SPTS and F-SPST was compared in Leonardi (2006), on protein families from the Pfam database Bateman et al. (2004) release 1.0. The database contained at that time 175 families derived from the SWISSPROT 33 database Boeckmann et al. (2003). For each family in the selected set, a model where trained with PST or SPST. There were used $4/5$ of the sequences in each family for training, and then the resulting model where applied to classify all the sequences in the SWISSPROT 33 database. To establish the family membership threshold and following the approach in Bejerano and Yona

Family	Size	% true pos. detected by PST	No. PST false positives	% true pos. detected by SPST	No. SPST false positives	% true pos. detected by F-SPST	No. F-SPST false positives
7tm_1	515	93.0	36	96.3	19	97.7	12
7tm_2	36	94.4	2	97.2	1	100.0	0
7tm_3	12	83.3	2	100.0	0	100.0	0
AAA	66	87.9	8	90.9	6	93.9	4
ABC_tran	269	83.6	44	85.9	38	89.2	29
actin	142	97.2	4	97.2	4	99.3	1
adh_short	180	88.9	20	89.4	19	92.8	13
adh_zinc	129	95.3	6	91.5	11	95.4	6
aldedh	69	87.0	9	89.9	7	92.8	5
alpha-amylase	114	87.7	14	91.2	10	94.7	6
aminotran	63	88.9	7	88.9	7	90.5	6
ank	83	88.0	10	86.8	11	86.6	11
arf	43	90.7	4	93.0	3	93.0	3
asp	72	83.3	12	90.3	7	91.7	6
ATP-synt_A	79	92.4	6	94.9	4	97.5	2
ATP-synt_ab	180	96.7	6	96.7	6	98.3	3
ATP-synt_C	62	91.9	5	95.2	3	95.2	3
beta-lactamase	51	86.3	7	90.2	5	94.1	3
bZIP	95	89.5	10	90.5	9	93.7	6
C2	78	92.3	6	92.3	6	96.2	3

Table 3.1: Performance comparison between PST, SPST and F-SPST. Families are ordered alphabetically, and correspond to the first 20 families with more than 10 sequences in the Pfam database, version 1.0. The number of sequences in each family is given in the second column. The other six columns, two for each algorithm, indicate the percentage of true positives detected with respect to the size of each family and the number of false positives, when using the equivalence number criterion. This method sets the threshold at the point where the number of false positives equals the number of false negatives. PST results where taken from Bejerano and Yona (2001). The set of parameters to train the SPST and F-SPST algorithms where: $L = 20$, $N_{\min} = 3$, $\gamma_{\min} = 0.001$ and $r_{\max} = 3.8$. The value of M used in the F-SPST algorithm was $M = 80$ for all families.

(2001), it was used the *equivalence number criterion* (Pearson 1995). This method sets the threshold at the point where the number of false positives (the number of non member proteins with score above the threshold) equals the number of false negatives (the number of member proteins with score below the threshold), i.e. it is the point of balance between selectivity and sensitivity. A member protein that scores above the threshold (true positive) is considered successfully detected. The quality of the model is measured by the percentage of true positives detected with

respect to the total number of proteins in the family.

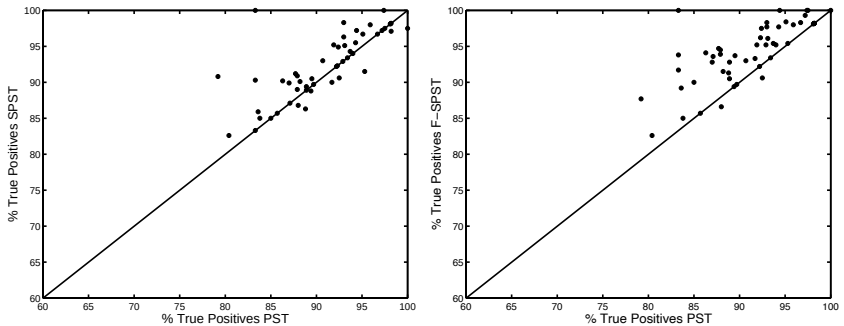


Figure 3.3: Scatter-plots of performances from PST, SPST and F-SPST protein classification methods. Above: SPST vs. PST. Below: F-SPST vs. PST

Table 3.1 shows the classification rates obtained with PST, SPST and F-SPST. The number of false positives of each algorithm is also shown. Because of the way of establishing the family membership threshold, the percentage of false positives is equal to 100% minus the percentage of true positives (with respect to the total number of sequences in the family). For example, in the case of family *7tm_1*, the percentage of true positives detected by the F-SPST algorithm is 97.7%, so the percentage of false positives is 2.3%. This gives 12 sequences erroneously classified as members of the *7tm_1* family. Figure 3.3 summarizes the classification rates of all the protein families considered in Leonardi (2006) in two scatter-plots, comparing the performance of PST with respect to SPTS and F-SPST. In general, both generalizations have better performance than the original PST algorithm. This is probably due to the fact that the sparse model mimics well the sparse nature of relevant domains in the amino acid chains. Another very interesting feature of SPST appears when comparing nodes in the estimated trees with the classes obtained by grouping the amino acids by their physical and chemical properties. For instance, the estimated tree for the *ATPase family associated with various cellular activities* (AAA) family has as a sparse node the set of amino acids $\{I, V, L\}$. This set corresponds exactly with the group of aliphatic amino acids, see Figure 3.4.

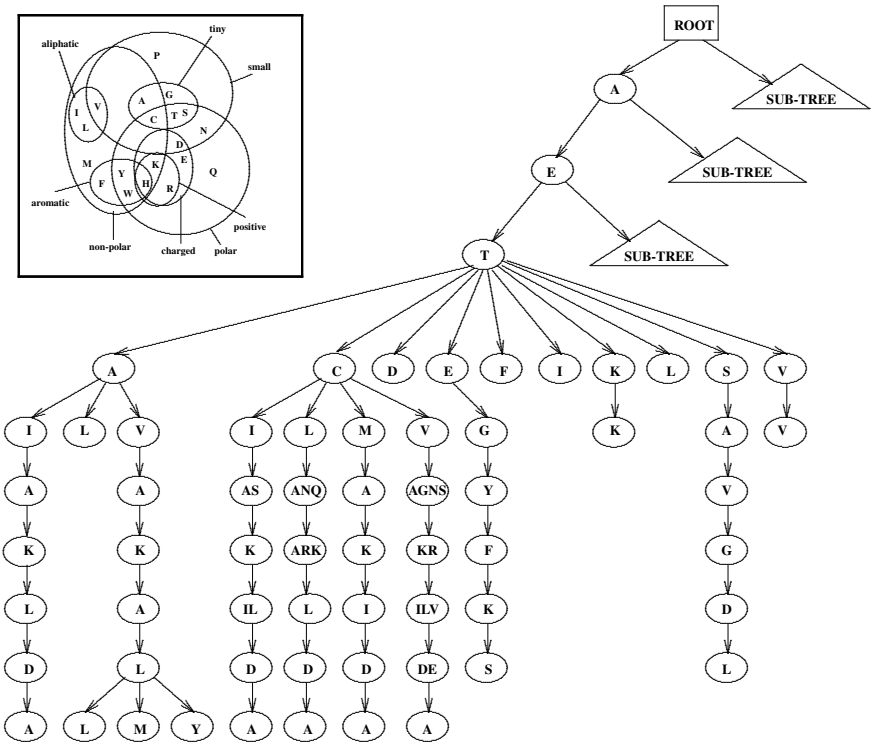


Figure 3.4: A SMC tree estimated with the SPST algorithm. This tree corresponds to sequences in the AAA family (*ATPase family associated with various cellular activities*). In each node we can see the subsets of amino acids corresponding to different positions of the sparse contexts (the curly brackets of the subsets were dropped). Some nodes of the tree are in correspondence with the physicochemical groups of the amino acids (shown in the square), as for example the set $\{I, L, V\}$ that corresponds to the set of aliphatic amino acids.

3.2 Rhythm in natural languages

In this section we describe the application of stochastic chains with memory of variable length to modeling real linguistic data, as presented in Galves, Galves, et al. (2012). The approach is based on the SMC algorithm introduced in Section 2.3.3. The linguistic problem refers to the determination of rhythmic patterns

in codified written texts of Brazilian and European Portuguese, as we describe in the following section.

3.2.1 The linguistic question

It has been conjectured in the linguistic literature that languages are divided into different rhythmic classes (Abercrombie 1967; James 1940; Pike 1945, among others). However, during half a century, neither a precise definition of each conjectured rhythmic class, nor any reliable phonetic evidence of the existence of these classes was presented in the linguistic literature (Dauer 1983).

The situation started changing at the end of the century. First of all, Mehler et al. (1996) gave empirical evidence that newborn babies are able to discriminate rhythmic classes. Then Ramus, Nespor, and Mehler (1999) gave, for the first time, evidence that simple statistics of the speech signal could discriminate between different rhythmic classes. A sound statistical basis to this descriptive analysis was given in Cuesta-Albertos et al. (2007) who used the projected Kolmogorov–Smirnov test to classify the sonority paths of the sentences in the sample analyzed in Ramus, Nespor, and Mehler (1999). We refer the reader to Ramus (2002) for an illuminating discussion of the rhythmic class conjecture.

The Brazilian and the European dialects of Contemporary Portuguese, from this point on referred as BP and EP, respectively, provide an interesting case to be analyzed from this point of view. In effect, BP and EP share the same lexicon. Moreover, from a descriptive point of view, most of the sentences they produce are superficially identical. However, it has been argued that they belong to different rhythmic classes (cf. for instance Brandão de Carvalho 1988; Frota and Vigário 2001; Sândalo et al. 2006).

All the analyses mentioned in the above paragraphs have been carried out on speech signal samples. The question addressed by Galves, Galves, et al. (2012) was whether it is possible to detect rhythmic differences in written texts. More specifically, the question raised was whether it is possible to identify in written texts rhythmic features characterizing and distinguishing BP and EP. In the absence of phonetic implementation, what kind of rhythmic evidence can be retrieved from the texts?

First of all, since the pioneer work by James (1940) and Abercrombie (1967), it has been conjectured that linguistic rhythm is characterized by the way stressed syllables interact in the sentence. Here by stressed syllables, we mean syllables carrying the main stress of the word. For instance, in the English word *linguistics*, which has three syllables *lin* - *guis* - *tics*, the main stress is on the second syllable

guis.

Second, it has also been conjectured that linguistic rhythm depends on the role played by the boundaries of phonological words (cf. Kleinhenz 1997). Here, by *phonological word* we mean a lexical word together with the functional non stressed words which precede it (cf. for instance Vigário 2003). For instance, in the sentence

The boy ate the candy

there are three phonological words: "*the boy*", "*ate*", and "*the candy*".

Finally, sentences themselves can be arguably considered as relevant units from the point of view of rhythm, since they correspond in written language to what has been called *Intonational Phrase* in the linguistic literature (cf. for instance Nespor and Vogel 2012).

This suggests to encode the texts by, first of all, assigning two symbols to each syllable of the text according to whether:

- the syllable is stressed or not;
- the syllable is the beginning of a phonological word or not,

This amounts to use $\{0,1\}^2$ as the set of symbols where the first symbol indicates if the syllable is the beginning or not of a prosodic word and the second symbol indicates if the syllable is stressed or not. To simplify the notation it is used the binary expansion to represent the pairs as integers as follows $(0,0) = 0$, $(0,1) = 1$, $(1,0) = 2$ and $(1,1) = 3$.

Additionally, it was added the extra symbol "4" to encode the periods marking the limits of each sentence. From now on, the alphabet $\{0,1,2,3,4\}$ obtained in this way is denoted by A.

Two examples will help understanding the codification. First of all, let us consider the encoding of the English sentence

The boy ate the candy.

This sentence is encoded as follows:

	The	boy	ate	the	can	dy	.
BPW	yes	no	yes	yes	no	no	
SS	no	yes	yes	no	yes	no	
ES	2	1	3	2	1	0	4

where BPW, SS and ES means *Beginning of a Phonological Word*, *Stressed Syllable* and *Encoded Sequence*, respectively. Let us now look at an example in Portuguese.

O menino já comeu o doce (The boy already ate the candy)

	O	me	ni	no	já	co	meu	o	do	ce	.
BPW	yes	no	no	no	yes	yes	no	yes	no	no	
SS	no	no	yes	no	yes	no	yes	no	yes	no	
ES	2	0	1	0	3	2	1	2	1	0	4

It is worth observing that BP and EP use the same spelling rules. These rules identify without ambiguity the syllables carrying the main stress in the words. Moreover, the set of non stressed functional words is well-defined. These two facts make it possible to encode both BP and EP texts in an automatic way. The Perl script "silaba2008.pl" was developed for this purpose. This script was included in the directory "SCRIPTS" which is part of the supplementary material of the paper Galves, Galves, et al. (2012).

With the encoded samples from BP and EP according to the mentioned rhythmic features, the class of stochastic chains with memory of variable length was used to detect rhythmic patterns in both languages. In effect, the linguistic conjectures reported above concerning the rhythmic role played by boundaries of sentences, boundaries of phonological words and stressed syllables can be translated using the notion of *context* which characterizes the model of stochastic chains with memory of variable length.

More precisely, the question at stake is whether the three rhythmic features considered in the coding of prosodic words play a role in the definition of the contexts identified through a statistical analysis of the BP and EP encoded data. If the linguistic conjecture concerning the rhythmic difference between BP and EP holds, then we expect to identify different context trees for the two languages. Moreover, this difference should reflect in some way the different role played in BP and EP by at least one of the three rhythmic features considered.

3.2.2 Results using the SMC

The data analyzed by Galves, Galves, et al. (ibid.) is a encoded corpus of newspaper articles extracted from Folha de São Paulo and Público, daily newspapers from Brazil and Portugal respectively. The sample consists of 80 articles randomly

selected from the 1994 and 1995 editions. Inside each edition the articles with less than 1000 words were discarded. There were also discarded interviews, synopsis, transcriptions of laws, whose peculiar characteristics made them unsuitable for the purposes of the analysis. The sample consists of 20 articles from each year for each newspaper randomly selected in the set of the remaining articles. The data is freely available as supplementary material to the article, see Galves, Galves, et al. (ibid.) for details. Before encoding each one of the selected texts, they were submitted to a linguistically oriented cleaning procedure. Hyphenated compound words were rewritten as two separate words, except when one of the components is unstressed. Suspension points, question marks and exclamation points were replaced by periods. Dates and special symbols like "%" were spelled out as words. All parentheses were removed.

To apply the SMC described in Section 2.3.3, the number of degrees of freedom of each candidate context tree has to be computed. This number takes into account the linguistic restrictions on the symbolic chain obtained after encoding. The restrictions are the following:

1. Due to Portuguese morphological constraints, a stressed syllable (encoded by 1 or 3) can be immediately followed by at most three unstressed syllables (encoded by 0).

2. Since by definition any phonological word must contain one and only one stressed syllable (encoded by 1 or 3), after a symbol 3 no symbol 1 is allowed, before a symbol 2 (non stressed syllable starting a phonological word) appears.

3. By the same reason, after a symbol 2 no symbols 2 or 3 are allowed before a symbol 1 appears.

4. As sentences are formed by the concatenation of phonological words, the only symbols allowed after 4 (end of sentence) are the symbols 2 or 3 (beginning of phonological word).

For each data set the first step in the SMC is to identify the set of champion trees for each penalizing constant given by (2.27). Then it is applied the Bootstrap Procedure explained in Section 2.3.3. For each data set there were used two different sample sizes: the first one, n_1, corresponding to 30% of the size of the sample and the second one, n_2, corresponding to 90% of the size of the sample. For each sample size, the number of resamples was $B = 200$.

n.l.	Champion trees
5	0 1 2 3 4
8	00 10 20 30 1 2 3 4
11	000 100 200 300 10 20 30 1 2 3 4
13	000 100 200 300 10 20 30 001 201 21 2 3 4
14	000 100 200 300 010 210 20 30 001 201 21 2 3 4
15	000 100 200 300 0010 2010 210 20 30 001 201 21 2 3 4
16	**0000 2000 100 200 300 0010 2010 210 20 30 001 201 21 2 3 4**
17	0000 2000 100 200 300 0010 2010 210 20 30 0001 2001 201 21 2 3 4

Table 3.2: Eight first BP champion trees, excluding the elementary root tree. The column n.l. indicates the number of leaves of each tree. The Smallest Maximizer Champion tree appears in bold face.

n.l.	Champion trees
5	0 1 2 3 4
8	00 10 20 30 1 2 3 4
11	000 100 200 300 10 20 30 1 2 3 4
13	000 100 200 300 10 20 30 001 201 21 2 3 4
14	000 100 200 300 010 210 20 30 001 201 21 2 3 4
17	**000 100 200 300 010 210 20 30 001 201 21 02 12 32 42 3 4**
20	000 100 200 300 010 0210 1210 3210 4210 20 30 001 201 21 02 12 32 42 3 4
21	000 100 200 300 0010 2010 0210 1210 3210 4210 20 30 001 201 21 02 12 32 42 3 4

Table 3.3: Eight first EP champion trees, excluding the elementary root tree. The column n.l. indicates the number of leaves of each tree. The Smallest Maximizer Champion tree appears in bold face.

In order to implement the bootstrap resampling, the authors of Galves, Galves, et al. (2012) took advantage of a striking feature which is present in all the champion trees, namely the fact that the symbol 4 is either a context by itself or appears as the final symbol of a context, as can be seen in Table 3.2 and Table 3.3. In other words, the successive occurrences of the symbol 4 are renewal points of the chain. Therefore, the blocks between consecutive occurrences of the symbol 4 are independent.

It follows that these independent blocks can be used to perform the usual Efron's bootstrap procedure (see Efron and Tibshirani 1993, for details). The final resamples of size n_j are obtained by the concatenation of the successively chosen independent blocks truncated at size n_j. The Perl script "G4L.pl" was developed to implement the SMC procedure, and is also available as supplementary

n.l.	c	New contexts
5	164.6	root → 0, 1, 2, 3, 4
8	30.1	0 → 00, 10, 20, 30
11	1.54	00 → 000, 100, 200, 300
13	1.037	1 → 001, 201, 21
14	0.75	10 → 010, 210
15	0.51	010 → 0010, 2010
16	0.357	000 → 0000, 2000
17	0.354	001 → 0001, 2001
19	0.30	210 → 0210, 3210, 4210

Table 3.4: Successive branchings producing the nine first BP champion trees. The first column n.l. indicates the total number of leaves of the new champion tree obtained by the new branching. The second column c indicates the largest value of the penalty constant making it worth selecting a tree with the new set of contexts.

material to the article Galves, Galves, et al. (2012) in the AOAS web site.

The results obtained with this approach are presented in the following figures and tables, taken from the original article Galves, Galves, et al. (ibid.). Table 3.2 and Table 3.3 show the eight first champion trees for Brazilian and European Portuguese, respectively. The smallest Maximizer Champion tree for each language appears in boldface. Successive branchings producing the successive champion trees in BP and EP are presented in Table 3.4 and Table 3.5, respectively. Figure 3.5 presents the log-likelihood corresponding to each champion tree for BP and EP according to the number of leaves. Finally, the selected trees for BP and EP are presented in Figure 3.6 and the corresponding families of transition probabilities are presented in Table 3.6.

Besides discriminating EP and BP, the selected trees have properties which are linguistically interpretable. First, 4 is a context or the ending symbol of a context, not only in the two selected trees, but actually in all the champion trees. This is a welcome result on linguistic grounds since it is reasonable to think that the successive sentences in a text are rhythmically, as well as syntactically, independent.

Second, in both trees, non stressed internal syllables provide poor information about the future. Three successive symbols zero are needed to constitute a context. This is consistent with linguistic common beliefs according to which non stressed non initial syllables do not play a salient role in rhythm by their own, but only as parts of bigger rhythmic units like phonological words.

Note that a stressed syllable alone is not enough to predict the next symbol ei-

n. l.	c	New contexts
5	177.1	root → 0, 1, 2, 3, 4
8	29.4	0 → 00, 10, 20, 30
11	1.70	00 → 000, 100, 200, 300
13	1.030	1 → 001, 201, 21
14	0.37	10 → 010, 210
17	0.34	2 → 02, 12, 32, 42
20	0.325	210 → 0210, 1210, 3210, 4210
21	0.321	010 → 0010, 2010
24	0.276	30 → 030, 130, 330, 430

Table 3.5: Successive branchings producing the nine first EP champion trees. The first column n.l. indicates the total number of leaves of the new champion tree obtained by the new branching. The second column c indicates the largest value of the penalty constant making it worth selecting a tree with the new set of contexts.

ther. Table 3.6 presenting the transition probabilities shows that in both languages the distribution of what follows a stressed syllable is dependent on the presence or absence of a preceding phonological word boundary in the two preceding steps. This fact, arguably derivable from the morphology of Portuguese, does not discriminate EP and BP. By morphology, we mean the way words of a particular language are constituted. This is not surprising since to a great extent EP and BP share the same lexicon.

Finally, according to the selected trees, the main difference between the two languages is that whereas in BP, both 2 (unstressed boundary of a phonological word) and 3 (stressed boundary of a phonological word) are contexts, in EP only 3 is a context. This means that in EP, as far as non-initial stress words are concerned, the choice of lexical items is dependent on the rhythmic properties of the preceding words. This is not true when the word begins with a stressed syllable. This does not occur in BP, where word boundaries are always contexts, and as such insensitive to what occurs before, independently of being stressed or not.

These statistical findings are compatible with the discussions in the linguistic literature concerning the different behavior of phonological words in the two languages (cf. Sândalo et al. 2006; Vigário 2003, among others).

It should be mentioned that Belloni and Oliveira (2017) presented a different point of view for model selection applied to a slightly generalized class of stochastic chain with memory of variable length, that they called *grouped context trees*. Their applied their method to the same linguistic data and their findings are coher-

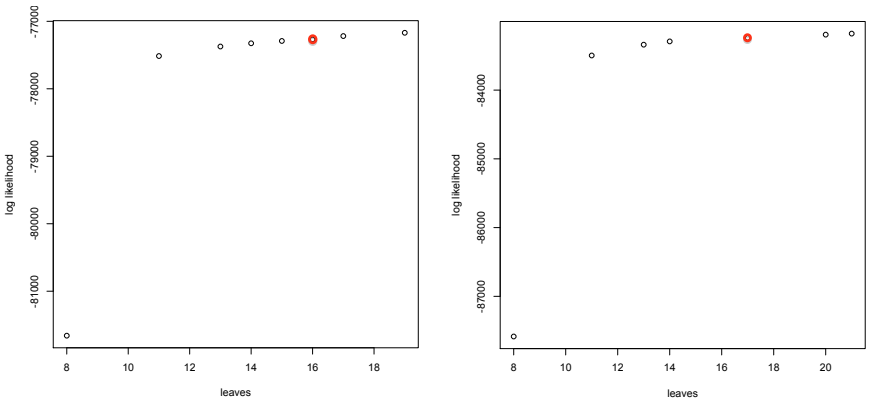


Figure 3.5: Log-likelihood corresponding to the champion trees for BP and EP according to the number of leaves

ent with ours.

3.3 Exercises

Exercise 3.1. Verify that $N(\bar{w}, a)$ and $N(\bar{w})$ in (3.5) and (3.6) correspond to the sum of the counters $N_{i,n_i}(w, a)$ and $N_{i,n_i}(w)$ in (3.1) and (3.2) for all $w \in \bar{w}$, respectively.

Exercise 3.2. Consider the alphabet $A = \{a, c, g, t\}$ and the DNA sequences given by

a a g t t a g c t a g a c g c g t a g c g a g t c c g c g
a a c t g a c c t a a a c g g g t g g c c a a t c t g g g
a c c g g a g c t a g a c a a g t a g c g a a g c t g a g

1. Compute the counters $N(w, y)$ and $N(w)$ for $w \in \{a, c, g, t\}^2$ and $y \in \{a, c, g, t\}$ given in (3.1)-(3.2).
2. Compute the PST context tree for $L = 2$ and $p_{\min} = r = \gamma = \alpha = 0.001$.
3. Compute the SPST context tree for $L = 2$, $N_{\min} = 1$ and $r_{\max} = 1$.

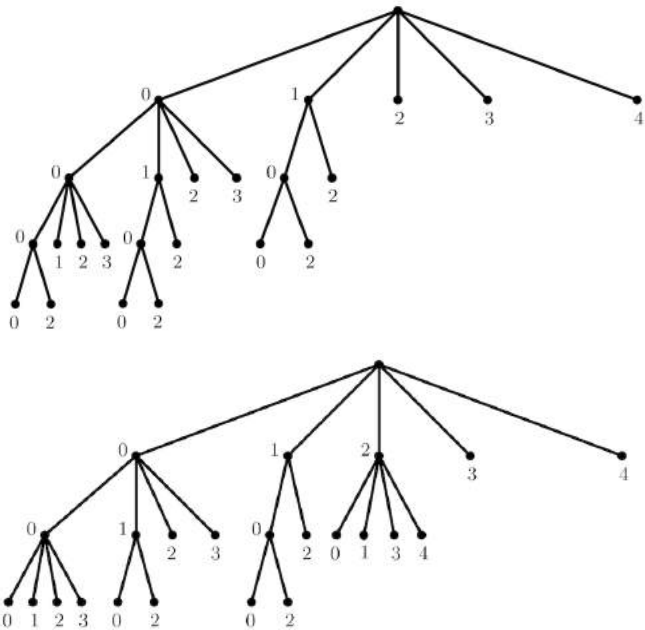


Figure 3.6: Smallest Maximizer trees for BP (top) and EP (bottom).

Exercise 3.3. Show that the tree generated by the algorithm Context in (2.14) is the same as the tree generated by the PST algorithm described in Section 3.1.1, for appropriate values of L, $p_{\min}$, $\gamma_{\min}$, r, α, d and δ_n. Discuss the similarities and the main differences between both algorithms.

Exercise 3.4. Given the sequence $x_{1:10} = 0001101010$, compute the scores of $x_{1:10}$ in (3.9) under the PST and SPST models given by Figure 3.1 and Figure 3.2, respectively, with $\gamma_{\min} = 0.001$.

BP

w	$p(0\|w)$	$p(1\|w)$	$p(2\|w)$	$p(3\|w)$
0000	0.28	0.72	0.00	0.00
2000	0.32	0.68	0.00	0.00
100	0.00	0.00	0.67	0.21
200	0.40	0.60	0.00	0.00
300	0.00	0.00	0.67	0.22
0010	0.03	0.00	0.67	0.20
2010	0.07	0.00	0.66	0.19 8
210	0.08	0.00	0.63	0.22
20	0.45	0.55	0.00	0.00
30	0.07	0.00	0.64	0.25
001	0.62	0.00	0.27	0.08
201	0.72	0.00	0.19	0.07
21	0.73	0.00	0.18	0.08
2	0.60	0.40	0.00	0.00
3	0.69	0.00	0.21	0.10
4	0.00	0.00	0.66	0.34

EP

w	$p(0\|w)$	$p(1\|w)$	$p(2\|w)$	$p(3\|w)$
000	0.27	0.73	0.00	0.00
100	0.00	0.00	0.67	0.25
200	0.36	0.64	0.00	0.00
300	0.00	0.00	0.70	0.20
010	0.05	0.00	0.67	0.19
210	0.08	0.00	0.63	0.22
20	0.45	0.55	0.00	0.00
30	0.05	0.00	0.64	0.27
001	0.61	0.00	0.28	0.07
201	0.72	0.00	0.19	0.07
21	0.72	0.00	0.19	0.07
02	0.59	0.41	0.00	0.00
12	0.55	0.45	0.00	0.00
32	0.50	0.50	0.00	0.00
42	0.52	0.48	0.00	0.00
3	0.69	0.00	0.19	0.12
4	0.00	0.00	0.65	0.35

Table 3.6: Transition probabilities associated to the contexts of BP and EP context trees given in Figure 3.6.

4 Stochastic systems of spiking neurons

In this chapter, we introduce an example of a space-time model, called *interacting chains with memory of variable length*. These chains describe the spiking activity of a neuronal network. In this network, the interactions between neurons are defined in terms of their interaction neighborhoods. The interaction neighborhood of a neuron is given by the set of all its presynaptic neurons. One important problem for such a network of neurons is to estimate these interaction neighborhoods. The main goal of this chapter is to present a simple method for interaction neighborhood estimation and prove its consistency. To illustrate the practical performance of this method, we present some empirical and simulation results. The empirical results are obtained by applying the neighborhood estimation method to a real data set from the first olfactory relay of the locust, *Schistocerca americana*. The material presented in this chapter are based on the article Duarte, Galves, Löcherbach, et al. (2019) and the manuscript Brochini, Hodara, et al. (2017).

4.1 Interacting chains with memory of variable length – a model for spiking neurons

Neurons communicate among themselves by firing sequences of short-lasting electrical pulses, called *spikes*. The sequence of spikes fired by a neuron is called *spike train* of the neuron. We adopt here a discrete time approach to model spike train data. In this approach, the time is discretized into bins of equal width ($10ms$ is a typical choice) and spikes are indicated by the symbol 1. The symbol 0 indicates the absence of a spike. In this way, the configuration of a network of neurons is described, for each time bin t, by a vector $X_t = (X_{1,t}, \dots, X_{d,t})$ with $\{0,1\}$-valued entries, where d denotes the size of the network. In the sequel, we assume that the bins are indexed by the set $\mathbb{Z}$ so that the network of neurons will be described by a collection of variables $\mathbf{X} = (X_{i,t})_{t\in\mathbb{Z}, i\in[d]}$ such that $X_t \in \{0,1\}^d$, where $[d] = \{1, \dots d\}$ denotes the set of neurons. Moreover, whenever we say time $t \in \mathbb{Z}$, it should be interpreted as time bin t.

In the network we consider, each neuron spikes with a probability which is an increasing function of its membrane potential. The membrane potential of a given neuron depends on the accumulated spikes coming from the presynaptic neurons since its last spike time. When a neuron spikes, its potential is reset to a resting level and at the same time postsynaptic current pulses are generated, modifying the membrane potential of all its postsynaptic neurons. To formalize this description, we need to introduce some notation.

Hereafter, for each $i \in [d]$, let $\varphi_i : \mathbb{R} \to [0,1]$ be a non-decreasing measurable function and $g_i = (g_i(t))_{t\in\mathbb{N}}$ be a sequence of strictly positive real numbers. The function φ_i and the sequence g_i are called *spike rate function* and *postsynaptic current pulse* of neuron i, respectively. Let also $(W_{j\to i})_{i,j\in[d]}$ be a collection of real numbers such that $W_{j\to j} = 0$ for all j. We call $W_{j\to i}$ the *synaptic weight of neuron j on neuron i*.

Recall that for each $i \in [d]$ and $t \in \mathbb{Z}$, $X_{i,t} = 1$ means that neuron i spiked at time t and $X_{i,t} = 0$, otherwise. For each $i \in [d]$ and $t \in \mathbb{Z}$, we write $L_{i,t}$ to denote the last spike time of neuron i before time t, defined as

$$L_{i,t} = \sup\{s \leqslant t : X_{i,s} = 1\}. \tag{4.1}$$

Here, we adopt the convention that $\sup\{\emptyset\} = -\infty$. Finally, for each $t \in \mathbb{Z}$, we write $X_{-\infty:t}$ to denote the past of the network up to time t, that is,

$$X_{-\infty:t} = (X_{j,s})_{j\in[d], s\leqslant t}.$$

In what follows, $\mathbb{P}$ denotes the law of neuronal network $\mathbf{X}$. The dynamics of the neuronal network $\mathbf{X}$ is defined as follows.

For each time $t \in \mathbb{Z}$ and any choice $a = (a_1, \ldots, a_d) \in \{0,1\}^d$,

$$\mathbb{P}(X_{t+1} = a | X_{-\infty:t}) = \prod_{i=1}^{d} \mathbb{P}(X_{i,t+1} = a_i | X_{-\infty:t}) \ \ \mathbb{P}\text{-a.s.}, \tag{4.2}$$

where for each $i \in [d]$,

$$\mathbb{P}(X_{i,t+1} = 1 | \mathcal{F}_t) = \varphi_i \Big(\sum_{j=1}^{d} W_{j\to i} \sum_{s=L_{i,t}+1}^{t} g_j(t-s) X_{j,s} \Big) \ \ \mathbb{P}\text{-a.s.}, \tag{4.3}$$

if $L_{i,t} < t$, and

$$\mathbb{P}(X_{i,t+1} = 1 | X_{-\infty:t}) = \varphi_i(0) \ \ \mathbb{P}\text{-a.s.},$$

otherwise.

Let us comment some features of the dynamics of this network. First, notice that by (4.2) the random variables $X_{1,t+1}, \ldots, X_{d,t+1}$ are conditionally independent given that past up to time t. Second, since the function φ_i is non-decreasing and g_i is a positive sequence, the spikes of j excite neuron i if $W_{j\to i} > 0$. In contrast, if $W_{j\to i} < 0$ the spikes of j inhibit neuron i. Moreover, if neuron i has spiked at time t ($L_{i,t} = t$), then it forgets its past in the sense that it spikes at time $t+1$ with a probability which does not depend on past up time t. On the other hand, if neuron i has spiked $k \geqslant 1$ units in the past with respect to time t ($L_{i,t} = t-k$), then its spiking probability at time $t+1$ depends only on past up to time t through the configuration $X_{[d],(t-k+1):t} := (X_{j,s})_{j\in[d],t-k-1\leqslant s\leqslant t}$. Hence, the random variables $L_{i,t}$'s introduce a structure of variable-length memory in the model. For this reason this stochastic model was introduced in Galves and Löcherbach (2013) under the name of *Systems of Interacting Chains with Memory of Variable Length*. In what follows, we call *GL neuron model* the stochastic chain $\mathbf{X}$ defined by (4.2) and (4.3).

The GL neuron model can be seen as a version of the Integrate and Fire (IF) model with random thresholds, but only in cases in which the postsynaptic current pulses are of the exponential type. Indeed, only in such cases, the time evolution of the family of membrane potentials is a Markov process, see Exercise 4.3. For general postsynaptic current pulses, this is not true, see Exercise 4.4. Therefore, the GL neuron model is a non-Markovian version of the IF model with random thresholds. This fact places the GL neuron model within a classical and widely

accepted framework of modern neuroscience. Indeed, IF models have a long and rich history, going back to the fundamental work Hodgkin and Huxley (1952). For more insights on IF-models we refer the interested reader to classical textbooks such as Dyan and Abbott (2001) and Gerstner and Kistler (2002).

The inherent randomness of the thresholds in the GL neuron model leads to random neuronal responses instead of deterministic ones. The idea that the spike activity is intrinsically random and not deterministic can be traced back to Adrian (1928), see also Adrian and Bronk (1929). Under the name of "escape noise"-models, this question has then been further emphasized by Gerstner and van Hemmen (1992) and Gerstner (1995).

To conclude this section, let us also mention here that continuous time versions of the GL neuron model have been studied in De Masi et al. (2015), Duarte and Ost (2016), Fournier and Löcherbach (2016), Robert and Touboul (2016), Hodara and Löcherbach (2017b), Duarte, Ost, and Rodríguez (2015) and Yaginuma (2016), Baccelli and Taillefumier (2019) and Baccelli, Davydov, and Taillefumier (2020). We also refer to Brochini, de Andrade Costa, et al. (2016) and the references therein for a simulation study and mean field analysis of the GL neuron model. All these papers deal with probabilistic aspects of the model, not with statistical model selection, which will be discussed in the next section.

4.2 Neighborhood estimation procedure

In the sequel, we write

$$x = (x_{i,t})_{-\infty<t\leqslant 0, i\in[d]},$$

for any configuration $x \in \{0,1\}^{[d]\times\{\ldots,-1,0\}}$, and for any $F \subseteq [d]$, we write

$$x_{F,t} = (x_{i,t}, i \in F).$$

Moreover, for any $x \in \{0,1\}^{[d]\times\{\ldots,-1,0\}}$, we write

$$X_{-\infty:0} = x, \text{ if } X_{i,t} = x_{i,t} \text{ for all } -\infty < t \leqslant 0 \text{ and } i \in [d].$$

Finally, for any $\ell \geqslant 1$, $t \in \mathbb{Z}$, $F \subseteq [d]$ and $w \in \{0,1\}^{\{-\ell,\ldots,-1\}\times F}$, we write

$$X_{F,t-\ell:t-1} = w, \text{ if } X_{j,t-s} = w_{j,-s}, \text{ for all } 1 \leqslant s \leqslant \ell \text{ and for all } j \in F$$

and

$$X_{i,t-\ell-1:t-1} = 10^{\ell}, \text{ if } X_{i,t-s} = 0, \text{ for all } 1 \leqslant s \leqslant \ell \text{ and } X_{i,t-\ell-1} = 1.$$

In what follows, $s, t \in \mathbb{Z}$ will be time indices, while $n \in \mathbb{N}$ will be saved for future use as the length of the time interval during which the neural network is observed.

Let

$$V_i = \{j \in [d] \setminus \{i\} : W_{j \to i} \neq 0\}, \tag{4.4}$$

be the set of presynaptic neurons of neuron i. The set V_i is called the *interaction neighborhood* of neuron i. The goal of our statistical selection procedure is to identify the set V_i from the data in a consistent way.

Let us briefly describe the statistical selection procedure we consider. We observe the process within a sampling region during a finite time interval. For each neuron i in the sampling region, we estimate its spiking probability given the spike trains of all other neurons since its last spike time. For each neuron $j \neq i$, we then introduce a measure of sensibility of this conditional spiking probability with respect to changes within the spike train of neuron j. If this measure of sensibility is *statistically small*, we conclude that neuron j does not belong to the interaction neighborhood of neuron i. In the sequel, we define rigorously this statistical procedure.

Let $x_{F,1}, \ldots, x_{F,n}$ be a sample where $F \subseteq [d]$ is a sampling region and $n \geqslant 3$ is the length of the time interval during which the network has been observed. For any fixed $i \in F$, we want to estimate its interaction neighborhood V_i.

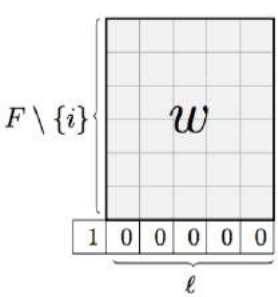


Figure 4.1: Local past $w \in \{0,1\}^{\{-\ell,\ldots,-1\} \times F \setminus \{i\}}$ outside of i with $\ell = 5$ and $|F| = 7$.

Our procedure is defined as follows. For each $1 \leqslant \ell \leqslant n-2$, local past $w \in \{0,1\}^{\{-\ell,\ldots,-1\} \times F \setminus \{i\}}$ outside of i (see Figure 4.1) and symbol $a \in \{0,1\}$, we

define

$$N_{(i,n)}(w,a) = \sum_{t=\ell+2}^{n} \mathbf{1}\{x_{i,t-\ell-1:t-1} = 10^{\ell}, x_{F\setminus\{i\},t-\ell:t-1} = w, x_{i,t} = a\}.$$

The random variable $N_{(i,n)}(w,a)$ counts the number of occurrences of w followed or not by a spike of neuron i ($a = 1$ or $a = 0$, respectively) in the sample $x_{F,1}, \ldots, x_{F,n}$, when the last spike of neuron i has occurred $\ell + 1$ time steps before in the past, see Figure 4.2.

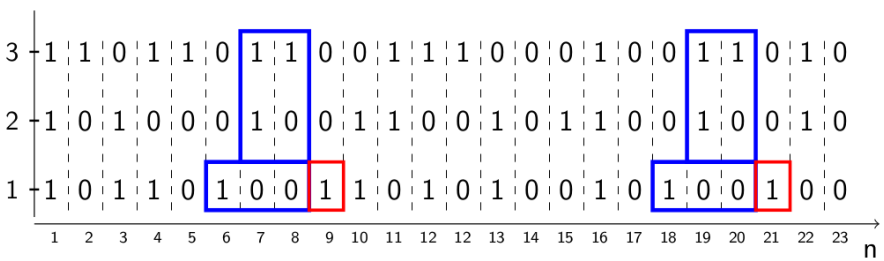


Figure 4.2: Example for $N_{(i,n)}(w,1) = 2$, where $i = 1$, for a given word w (in blue), $\ell = 2$, $|F| = 3$ and $n = 23$.

We define the empirical probability of neuron i having a spike at the next step given w by

$$\hat{p}_{(i,n)}(1|w) = \frac{N_{(i,n)}(w,1)}{N_{(i,n)}(w)}, \tag{4.5}$$

when $N_{(i,n)}(w) := N_{(i,n)}(w,0) + N_{(i,n)}(w,1) > 0$.

Remark 4.1. Notice that the empirical probabilities defined in (4.5) are extensions to space-time configurations of the empirical transition probabilities defined in (2.6) in Chapter 2. As such, they enjoy similar properties as the empirical transition probabilities. For instance, one can check that the empirical probabilities defined in (4.5) are maximum likelihood estimators (see Exercise 4.2).

For any fixed parameter $\xi \in (0, 1/2)$, we consider the following set

$$\mathcal{T}_{(i,n)} = \left\{ w \in \bigcup_{\ell=1}^{n-2} \{0,1\}^{\{-\ell,\ldots,-1\}\times F\setminus\{i\}} : N_{(i,n)}(w) \geqslant n^{1/2+\xi} \right\}. \tag{4.6}$$

We use the notation $|w| = \ell$ whenever $w \in \{0,1\}^{\{-\ell,\dots,-1\}\times F\setminus\{i\}}$. If v, w both belong to $\{0,1\}^{\{-\ell,\dots,-1\}\times F\setminus\{i\}}$ we write

$$v_{\{j\}^c} = w_{\{j\}^c} \text{ if and only if } v_{F\setminus\{j\},-\ell:-1} = w_{F\setminus\{j\},-\ell:-1}.$$

In words, the equality $v_{\{j\}^c} = w_{\{j\}^c}$ means that v and w coincide on all but the j-th coordinate.

Finally, for each $w \in \mathcal{T}_{(i,n)}$ and for any $j \in F \setminus \{i\}$ we define the set

$$\mathcal{T}_{(i,n)}^{w,j} = \left\{v \in \mathcal{T}_{(i,n)} : |v| = |w|, v_{\{j\}^c} = w_{\{j\}^c}\right\}$$

and introduce the *measure of sensibility*

$$\Delta_{(i,n)}(j) = \max_{w\in\mathcal{T}_{(i,n)}} \max_{v\in\mathcal{T}_{(i,n)}^{w,j}} |\hat{p}_{(i,n)}(1|w) - \hat{p}_{(i,n)}(1|v)|.$$

Our interaction neighborhood estimator is defined as follows.

Definition 4.2. For any positive threshold parameter $\epsilon > 0$, the estimated interaction neighborhood of neuron $i \in F$, at accuracy ϵ, given the sample $x_{F,1}, \dots, x_{F,n}$, is defined as

$$\hat{V}_{(i,n)}^{(\epsilon)} = \{j \in F \setminus \{i\} : \Delta_{(i,n)}(j) > \epsilon\}. \tag{4.7}$$

The statistical selection criterion defined in (4.7) is, in a way, a spatial variant of the Algorithm Context discussed in Chapter 2. Indeed, we first compute, for each local past w outside of i, the empirical probability of neuron i to spike given w. We then compute, for each neuron $j \in F \setminus \{i\}$, the discrepancy measure $\Delta_{i,n}(j)$ between these empirical probabilities corresponding to local pasts coinciding on all but j-th coordinate. If the discrepancy measure $\Delta_{i,n}(j)$ is smaller than a given threshold, then the j-th coordinate of the local pasts w can be "pruned'. In this case, we conclude that neuron j is not in the interaction neighborhood of neuron i.

To conclude this section, let us stress that in the GL neuron model, the probability of a neuron to spike depends only on the history of the process since its last spike time. Therefore, temporal dependencies do not need to be estimated, making the estimation problem discussed here different from context tree estimation as considered in Chapter 2.

4.3 Results on simulations

Simulation and neighborhood estimation procedures used to produce the results presented here were implemented in Python 3.0. Information about the codes can be found in Brochini, Hodara, et al. (2017).

4.3.1 Searching for suitable parameter values

In this section, we use simulated data in order to fix the parameters ξ and ε involved in the estimation procedure. Recall that ξ is the parameter appearing in the definition of the set $\mathcal{T}_{(i,n)}$ in (4.6) and that ε appears in the definition of $\hat{V}^{(\varepsilon)}_{(i,n)}$ in (4.7). The role of ξ is to ensure that the observations contains enough repetitions of a given local past w in order to define the empirical probability $\hat{p}_{(i,n)}(1|w)$. The parameter ε can be seen as a significance threshold for the measure of sensitivity $\Delta_{(i,n)}(j)$.

The simulated samples have sample size $n = 10^6$ and correspond to a network with 5 neurons. The neuronal activity was simulated according to the dynamics described in (4.2) and (4.3). Synaptic weights $W_{j\to i}$ were arbitrarily distributed from 0 to 0.8 in this network for all possible pairs (i, j). We used an exponential postsynaptic current pulse $g_i(n) = \mu^n$ with parameter $\mu = 0.5$ for each neuron i. We used the spiking rate function $\varphi_i(u) = \min(u + q_i, 1)$, where $q_i = 0.02$ for each neuron i.

In Figure 4.3, we give the results of the estimation procedure for different values of the parameters ξ and ε. For each couple (ξ, ε) we present the result in a 5×5 matrix. For each line j and column i, the color of the square indicates the presence or absence of influence of neuron j on neuron i and the result of the estimation procedure. The color code is the following. Correct estimations are represented in black and white: black if $W_{j\to i} \neq 0$ and white if $W_{j\to i} = 0$. Incorrect estimations are represented in hatched cells. Hatched white cells correspond to false negatives, when $W_{j\to i} \neq 0$ but the estimator produced an absent connection. Hatched grey cells, on the other hand, indicate false positives, when $W_{j\to i} = 0$ but a connection was estimated to exist. Plain grey cells correspond to inconclusive results, a situation when the event $E^n_{i,j}$ is not realized, where $E^n_{i,j} := \left\{\exists w \in \mathcal{T}_{(i,n)} : \mathcal{T}^{w,j}_{(i,n)} \neq \emptyset\right\}$. This may happen due to the fact that the sample becomes relatively small as the cutoff parameter ξ increases, in which case the procedure will produce a smaller number of valid events to be considered by the estimator.

As expected, low values of the sensitivity threshold ε lead to more false pos-

itive whereas high values lead to more false negative. For this sample size, the estimation procedure correctly recovers the true connectivity graph for $\varepsilon = 0.05$ and $\xi = 0.001$ or 0.01.

4.3.2 Pruning

Since the estimator is well defined only on events of the type $\bigcap_{j\in F: j\neq i} E^n_{i,j}$, where $E^n_{i,j} := \left\{\exists w \in \mathcal{T}_{(i,n)} : \mathcal{T}^{w,j}_{(i,n)} \neq \emptyset\right\}$, we propose an iterative pruning procedure to deal with cases where this event is not realized. For neuron $j \in F \setminus \{i\}$ for which $E^n_{i,j}$ is not realized, the connection $j \to i$ will be called inconclusive. This may occur when the sample size is small. When $\mathcal{T}_{(i,n)} = \emptyset$, all connections leading to neuron i are considered inconclusive. In the case where $\mathcal{T}_{(i,n)} \neq \emptyset$, a connection $j \to i$ is considered inconclusive if $\mathcal{T}^{w,j}_{(i,n)} = \emptyset$ for all $w \in \mathcal{T}_{(i,n)}$.

The pruning procedure is described as follows. If there exist $j \in F \setminus \{i\}$ such that $E^n_{i,j}$ is not realized and $k \in F \setminus \{i, j\}$ such that $E^n_{i,k}$ is realized and $k \notin \hat{V}^{(\epsilon)}_{(i,n)}$, we say that the pruning condition is fulfilled. If so, we compute $\hat{V}^{(\epsilon)}_{i,n}$ considering the set $F \setminus \{i, k\}$ instead of $F \setminus \{i\}$. This step is repeated as long as the pruning condition is fulfilled. The consistency of this iterative pruning procedure is not discussed here and can be found in Brochini, Hodara, et al. (2017).

Inconclusive connections can be typically attributed to small sample sizes and/or data sparsity. Evidently, if we increase the number of neurons or decrease sample size while maintaining the same parameter values of ϵ and ξ, we expect a larger number of inconclusive connections. This is precisely what we did to illustrate the utility of the pruning procedure: we generated a sample of GL neuron model with 10 neurons and sample size of $n = 2 \times 10^5$, which is a larger number of neurons and smaller sample size as used in the previous section. All synapses have the same weight ($W_{i\to j} = W = 0.5$), the postsynaptic current pulses and the spiking rate functions are of the form $g_i(n) = \mu^n$ with parameter $\mu = 0.9$ and $\varphi_i(u) = \min(u + q_i, 1)$ with $q = 0.06$, respectively. In the analysis we used parameter values $\epsilon = 0.05$ and $\xi = 0.001$, determined in the previous section.

Notice that the first estimation obtained prior to any pruning (shown in Figure 4.4 A) produced a remarkably large number of inconclusive connections (grey cells). After the first estimation, the pruning procedure is used to help reduce the amount of inconclusive connections. For each postsynaptic neuron i, all neurons which are identified by the estimator as not preysynaptic to i are removed from the set of presynaptic candidates. Then the neighborhood estimating procedure is

repeated. The pruning and re-estimation is repeated while there are at least one inconclusive and one connection identified as null for the postsynaptic neuron i.

After the pruning procedure is performed for all postsynaptic neurons, we observe a dramatic improvement in the quality of the neighborhood estimations (Figure 4.4 B). The final estimation correctly identifies all existing connections for this network. The effectiveness of the pruning procedure is due to the reduction in the number of presynaptic candidate neurons while maintaining the same sample size, leading to the improvement of the estimation performance.

4.4 Results on a dataset recorded in vivo

Here we present results of the estimated interaction neighborhoods for a particular dataset that corresponds to a recording of about half an hour of spontaneous neural activity. Spike sorting procedure for this dataset can be found in Pouzat (2021). Through this procedure we obtain spike trains of 5 well isolated neurons, each neuron presenting the order of 10^4 total spikes in the sample.

In order to use the estimation procedure, we need to obtain a representation of the spike train in discrete time. We choose the largest binning window which produces less than 1% of overlaps. By overlap we mean when two or more spike events of the same neuron occur in the same time window. This leads to a binning window of about 10 milliseconds. The corresponding length of the time interval during which the network was observed is then $n = 18 \times 10^4$.

We fix for ξ and ε the values that fitted the simulations, i.e. $\xi = 0.001$ and $\varepsilon = 0.05$. Notice that the simulations presented in Section 4.3 were performed with the same number of neurons (5) and with the value of $n = 10^6$. We present in Figure 4.5 A the result of the estimation procedure. The color code is the following: black indicates we estimated that there is a connection $i \leftarrow j$, white indicates we estimated that there is no connection $i \leftarrow j$ and grey corresponds to an inconclusive connection. Notice that the results are mostly inconclusives for neurons 4 and 5, even with the pruning procedure described in Section 4.3.2.

In order to validate this estimation procedure, we split the dataset in two parts and proceed to the estimation for each part. The results are given in figures 4.5 B and 4.5 C. We can see that the estimation procedure gives us the same interaction neighborhoods for the two different parts of the dataset, except for pairs $4 \leftarrow 5$ and $5 \leftarrow 4$ where we have inconclusive results when data is split. As was already mentioned, the expected number of inconclusive connections should be very sensitive to sample size, so it is not surprising that two connections considered absent

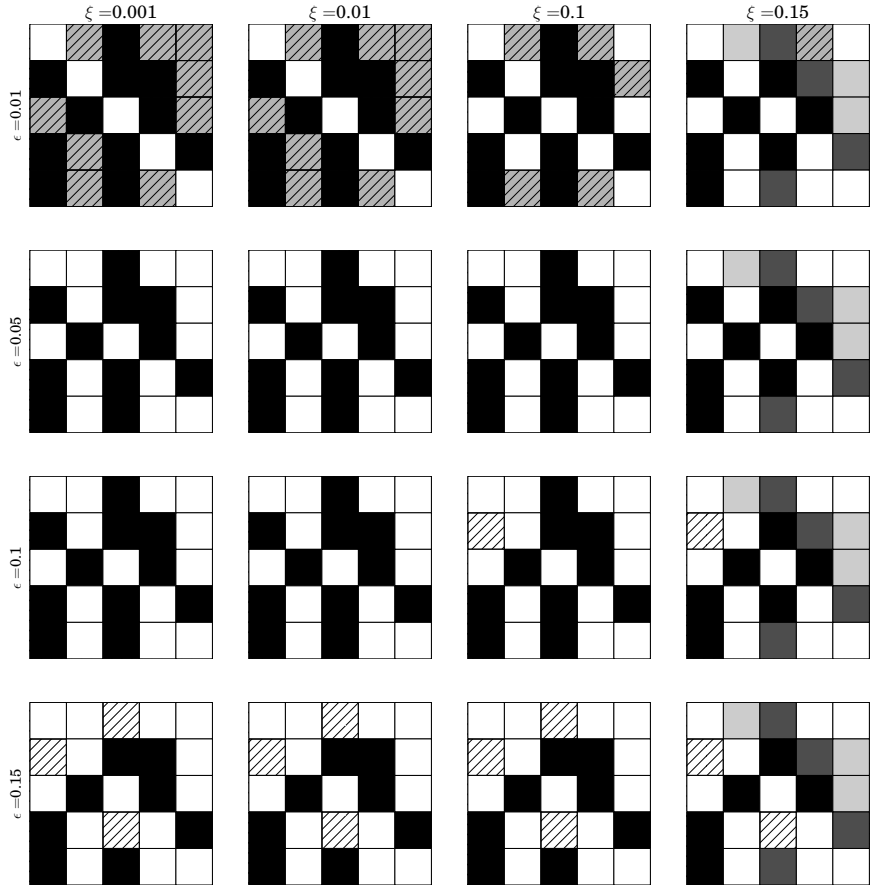


Figure 4.3: Interaction neighborhood estimations for a simulated dataset of a GL neuron model of 5 neurons for various values of parameters (ϵ, ξ), estimated for the same dataset with $n = 10^6$. Each element in the panel is a color coded representation of connections, where rows correspond to presynaptic neurons and columns to postsynaptic neurons. Colors indicate the comparison of the estimated interaction neighborhoods to the true ones used in the simulation to generate this dataset. Black cells correspond to a true connection correctly identified by the estimator. White ones indicate there is no connection, correctly identified as absent by the estimator. Hatched white and grey cells correspond to false negative and false positive, respectively. Grey cells correspond to inconclusive connections, where there is not enough repetitions of patterns to produce an estimation. Grey cells can be of two types: dark and light grey, in order to differentiate respectively the cases

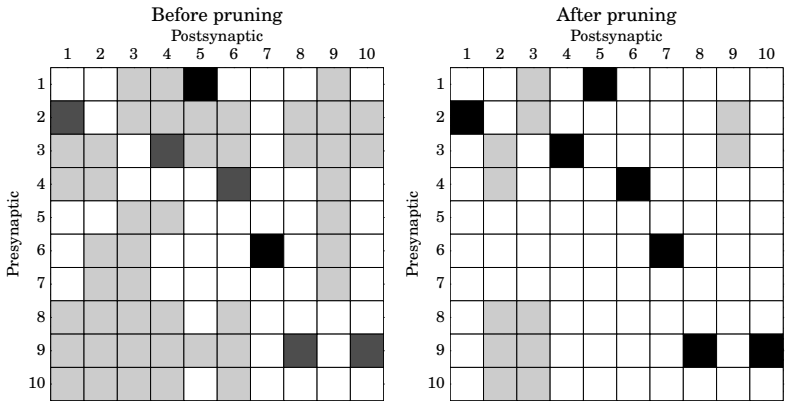


Figure 4.4: Color code: Black: existent connection correctly identified by estimator. White connection: non-existent connection correctly identified as such by estimator. Dark and light grey: inconclusive connections corresponding respectively to existent or non-existent connections. Original estimator produces too many inconclusives **(A)** for simulated dataset produced by a network of 10 neurons with $n = 2 \times 10^5$. After several pruning, we obtain a closer neighborhood estimation. The final estimation correctly identifies all existing connections for this network, but a few inconclusive connections remain where there are no connections.

when the whole data is analyzed appear as inconclusive ones when sample size is reduced by half. Having this considered, we can conclude that there is an overall agreement between the interaction neighborhoods obtained, and say that the estimation obtained is robust to data splitting for this dataset.

4.5 Consistency of the estimation procedure

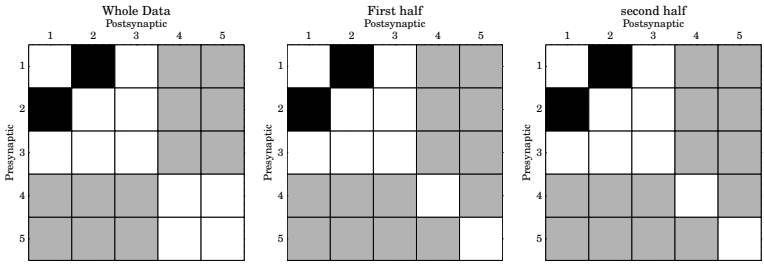


Figure 4.5: Neighborhood estimations for the whole dataset (left), for the first half (middle) and the second half of the dataset (right). Black: estimated connection. White: estimator produces no connection. Grey: inconclusive connections

4.5.1 Fully observed interaction neighborhoods

Let Ω^{adm} be the set of *admissible pasts*, defined as

$$\Omega^{adm} = \left\{ x \in \{0,1\}^{[d]\times\{\dots,-1,0\}} : \forall\, i \in [d],\ \exists\, \ell_i \leqslant 0 \text{ with } x_{i,\ell_i} = 1 \right\}. \quad (4.8)$$

Observe that if $X_{-\infty:0} = x \in \Omega^{adm}$, then $L_{i,0} > -\infty$ for all $i \in [d]$. In this case, we have that for each $i \in [d]$,

$$\sum_{j=1}^{d} W_{j\to i} \sum_{s=L_{i,0}+1}^{0} g_j(-s) X_{j,s} < \infty,$$

which implies that the transition probability $\mathbb{P}(X_{i,1} = 1 | X_{-\infty:0} = x)$ is well-defined. By induction, for each $t \geqslant 0$, the transition probabilities (4.3) are also well-defined. Thus, the existence of the stochastic chain $(X_t)_{t\in\mathbb{Z}}$, starting from $X_{-\infty:0} = x \in \Omega^{adm}$, follows immediately. Observe that we do not assume stationarity of the chain. To prove the consistency of our estimator we impose the following condition.

Assumption 4.3. For all $i \in [d]$, $\varphi_i \in C^1(\mathbb{R}, [0,1])$ is a strictly increasing function. Moreover, there exists a $p_* \in]0,1[$ such that for all $i \in [d]$ and $u \in \mathbb{R}$

$$p_* \leqslant \varphi_i(u) \leqslant 1 - p_*.$$

Define for $i \in [d]$,

$$K_i = \left[\sum_{j \in V_i^-} W_{j \to i} g_j(0), \sum_{j \in V_i^+} W_{j \to i} g_j(0) \right], \tag{4.9}$$

where $V_i^+ = \{j \in V_i : W_{j \to i} > 0\}$ and $V_i^- = \{j \in V_i : W_{j \to i} < 0\}$.

Notice that this interval is always bounded. Finally, we define

$$m_i = \inf_{u \in K_i} \left\{ \varphi_i'(u) \right\} \inf_{j \in V_i} \left\{ |W_{j \to i}| g_j(1) \right\}. \tag{4.10}$$

The following theorem is our first main result. It states the strong consistency of the interaction neighborhood estimator when the interaction neighborhood is fully observed, that is when $V_i \subset F$. By strong consistency we mean that the estimated interaction neighborhood of a fixed neuron i equals V_i eventually almost surely as $n \to \infty$.

Theorem 4.4. *Consider $F \subset [d]$ and let $x_{F,1}, \ldots, x_{F,n}$ be a sample produced by a the stochastic chain $(X_t)_{t \in \mathbb{Z}}$ compatible with (4.2) and (4.3), starting from $X_{-\infty:0} = x$ for some fixed $x \in \Omega^{adm}$. Under Assumption 4.3, for any $i \in F$ such that $V_i \subset F$, the following holds.*

1. **(Overestimation).** *For any $j \in F \setminus V_i$, we have that for any $\epsilon > 0$,*

$$\mathbb{P}\left(j \in \hat{V}_{(i,n)}^{(\epsilon)} \right) \leqslant 4n^{3/2-\xi} \exp\left\{ -\frac{\epsilon^2 n^{2\xi}}{2} \right\}.$$

2. **(Underestimation).** *The quantity m_i defined in (4.10) satisfies $m_i > 0$, and for any $j \in V_i$ and $0 < \epsilon < m_i$,*

$$\mathbb{P}\left(j \notin \hat{V}_{(i,n)}^{(\epsilon)} \right) \leqslant 4 \exp\left\{ -\frac{(m_i - \epsilon)^2 n^{2\xi}}{2} \right\} + \exp\left\{ -O\left(n^{1/2+\xi} \right) \right\}.$$

3. In particular, if we take $\epsilon_n \to 0$ as $n \to \infty$ such that $\epsilon_n \geqslant C n^{-\xi/2}$ for some constant $C > 0$, where ξ is the parameter appearing in (4.6), then

$$\hat{V}_{(i,n)}^{(\epsilon_n)} = V_i \text{ eventually almost surely.}$$

The proof of Theorem 4.4 is given in Section 4.7.

4.5.2 Extension to the case of partially observed interaction neighborhoods

We now discuss the case when V_i is not fully included in the sampling region F. In this case, we also impose the following assumptions.

Assumption 4.5. $\gamma = \sup_{j\in[d]} \|\varphi'_j\|_\infty < \infty$.

Assumption 4.6. There exists a positive constant C and $p \geqslant 1$, such that $g(t) = \sup_{j\in[d]} g_j(t) \leqslant C(1+t^p)$ for all $t \geqslant 1$.

Under Assumption 4.6, we may introduce, for each $t \geqslant 1$, the continuous operator $H(t) : \mathbb{R}^d \to \mathbb{R}^d$ defined by $(H(t)\xi)_j = \sum_{k=1}^d H_{j,k}(t)\xi_k$, for all $j \in [d]$, where

$$H_{j,k}(t) := \gamma |W_{k\to j}| g_k(t), \text{ for } j \neq i, \tag{4.11}$$

and, for p_* as in Assumption 4.3,

$$H_{i,i}(t) := (1-p_*)1_{\{t=1\}}.$$

By our assumptions, the norm of the operator $H(t)$ defined by

$$|||H(t)||| = \sup\{\|H(t)\xi\|_\infty : \xi \in \mathbb{R}^d, \|\xi\|_\infty = 1\},$$

satisfies

$$|||H(t)||| \leqslant C\gamma r(1+t^p) + (1-p_*)1_{\{t=1\}},$$

where $r = \sup_{i\in[d]} \sum_{j\in[d]} |W_{j\to i}| < \infty$. Then for any $\alpha > 0$, the linear operator

$$\Lambda(\alpha) = \sum_{t=1}^{\infty} e^{-\alpha t} H(t)$$

is well-defined and continuous as well. In particular, there exists $\alpha_0 \geqslant 0$ such that

$$|||\Lambda(\alpha_0)||| < 1. \tag{4.12}$$

We are now ready to state our second main result. It gives precise error bounds for the interaction neighborhood estimator when V_i is not fully observed. These error bounds depend on the tail of the series

$$\Sigma_i(F) := \sum_{j\notin V_i \cap F} |W_{j\to i}|. \tag{4.13}$$

To state the theorem we shall also need the definitions

$$K_i^{[F]} = \left[\sum_{j \in V_i^- \cap F} W_{j \to i} g_j(0), \sum_{j \in V_i^+ \cap F} W_{j \to i} g_j(0) \right]$$

and

$$m_i^{[F]} = \inf_{u \in K_i^{[F]}} \left\{ \varphi_i'(u) \right\} \inf_{j \in V_i \cap F} \left\{ |W_{j \to i}| g_j(1) \right\}.$$

Theorem 4.7. *Consider $F \subset I$ and let $x_1(F), \ldots, x_n(F)$ be a sample produced by a the stochastic chain $(X_t)_{t \in \mathbb{Z}}$ compatible with (4.2) and (4.3), starting from $X_{-\infty:0} = x$ for some fixed $x \in \Omega^{adm}$. Under Assumptions 4.3, 4.5 and 4.6, for any $i \in F$ such that $V_i \cap F \neq \emptyset$, the following assertions hold true.*
1. **(Overestimation).** *For any $j \in F \setminus V_i$, we have that for any $\epsilon > 0$,*

$$\mathbb{P}\left(j \in \hat{V}_{(i,n)}^{(\epsilon)} \right) \leqslant 4n^{3/2-\xi} \exp\left\{ -\frac{\epsilon^2 n^{2\xi}}{2} \right\} + C(e^{\alpha_0 n} \vee n) \Sigma_i(F).$$

2. **(Underestimation).** *We have that $m_i^{[F]} > 0$, and for any $j \in V_i \cap F$ and $0 < \epsilon < m_i^{[F]}$,*

$$\mathbb{P}\left(j \notin \hat{V}_{(i,n)}^{(\epsilon)} \right) \leqslant 4 \exp\left\{ -\frac{(m_i^{[F]} - \epsilon)^2 n^{2\xi}}{2} \right\} + \exp\left\{ -O\left(n^{1/2+\xi} \right) \right\} + C(e^{\alpha_0 n} \vee n) \Sigma_i(F).$$

The proof of Theorem 4.7 is given in Section 4.7.

4.6 Exponential inequalities

To prove Theorem 4.4 and Theorem 4.7 we need some exponential inequalities, including a new conditional Hoeffding-type inequality, stated in Proposition 4.8 below, which is interesting by itself.

For each $\ell \geqslant 1$, $F \subset [d]$ and $w \in \{0,1\}^{\{-\ell,\ldots,-1\} \times F \setminus \{i\}}$, we write

$$p_i(1|w) = \mathbb{P}(X_{i,\ell+2} = 1 | X_{i,1:\ell+1} = 10^\ell, X_{F \setminus \{i\},2:\ell+1} = w). \tag{4.14}$$

Notice that $p_i(1|w) = p_i(1|w(V_i))$ for any set $F \supset V_i$, where $w(V_i)$ is the configuration w restricted to the set V_i. Moreover, the time homogeneity of the transition probability (4.3) implies that, whenever $F \supset V_i$, for any $t \geqslant \ell + 2$,

$$p_i(1|w) = \mathbb{P}(X_{i,t} = 1 | X_{i,t-\ell-1:t-1} = 10^\ell, X_{F\setminus\{i\},t-\ell:t-1} = w).$$

Proposition 4.8. *Suppose $V_i \subset F$. Then, for any $w \in \{0,1\}^{\{-\ell,\dots,-1\}\times F\setminus\{i\}}$ with $\ell \geqslant 1$, $\lambda > 0$ and all $t > \ell + 1$,*

$$\mathbb{P}(|M_{(i,t)}(w)| > \lambda) \leqslant 2\exp\left\{-\frac{2\lambda^2}{t-\ell+1}\right\}\mathbb{P}(N_{(i,t)}(w) > 0), \tag{4.15}$$

where $M_{(i,t)}(w) := N_{(i,t)}(w,1) - p_i(1|w)N_{(i,t)}(w)$.

As a consequence of Proposition 4.8, we have the following result.

Proposition 4.9. *Suppose that $V_i \subset F$. Then for any $\ell \geqslant 1$, $t > \ell + 1$, $w \in \{0,1\}^{\{-\ell,\dots,-1\}\times F\setminus\{i\}}$, $\xi \in (0,1/2)$ and $\epsilon > 0$, we have*

$$\mathbb{P}\Big(|\hat{p}_{(i,t)}(1|w) - p_i(1|w)| > \epsilon, N_{(i,t)}(w) \geqslant t^{1/2+\xi}\Big) \leqslant 2\exp\left\{-2\epsilon^2 t^{2\xi}\right\}\mathbb{P}(N_{(i,t)}(w) > 0).$$

The next two results will be used to control the probability of underestimating V_i. We start with a simple lower bound which follows immediately from Assumption 4.3.

Lemma 4.10. *For any fixed $i \in F$, $t > \ell + 1$ and $w \in \{0,1\}^{\{-\ell,\dots,-1\}\times F\setminus\{i\}}$, we define for $1 \leqslant s \leqslant t - \ell$,*

$$Z_s = \mathbf{1}\{X_{i,s:s+\ell} = 10^\ell, X_{F\setminus\{i\},s+1:s+\ell} = w\}.$$

Under Assumption 4.3, it follows that

$$\mathbb{P}\left(Z_s = 1 | X_{-\infty:(s-1)}\right) \geqslant p_{min}^{|F|\ell+1},$$

where $p_{min} = \min\{p_*, (1-p_*)\} > 0$ with p_* as in Assumption 4.3.

Exercise 4.5 asks the reader to prove Lemma 4.10.

Lemma 4.11. *Suppose Assumption 4.3. For any $\xi \in (0,1/2)$, $i \in F$ and $w \in \{0,1\}^{\{-1\}\times F\setminus\{i\}}$, it holds that*

$$\mathbb{P}\left(N_{(i,t)}(w) < t^{1/2+\xi}\right) \leqslant \exp\left\{-O\left(t^{1/2+\xi}\right)\right\}.$$

4.7 Proofs of this chapter

Proof of Proposition 4.8. We denote $p = p_i(1|w)$ and for each $t \geqslant \ell + 1$, $N_{(i,t)}(w) = N_t$, $Y_t = \mathbf{1}\{X_{i,t} = 1\} - p$, $\chi_t = \mathbf{1}\{X_{F\setminus\{i\},t-\ell:t-1} = w, X_{i,t-\ell-1:t-1} = 10^\ell\}$ and also $M_{(i,t)}(w) = M_t$ with the convention that $M_{\ell+1} = 0$. Thus for $t \geqslant \ell + 2$,

$$M_t = M_{t-1} + \chi_t Y_t. \tag{4.16}$$

Since $\mathbb{P}(M_t > \lambda) = \mathbb{P}(M_t > \lambda, N_t > 0)$, the Markov inequality implies that

$$\mathbb{P}(M_t > \lambda) \leqslant e^{-\lambda\sigma} E\big[e^{\sigma M_t}\mathbf{1}\{N_t > 0\}\big],$$

for all $\sigma > 0$. Notice that $\{N_t > 0\} = \{N_{t-1} > 0\} \cup \{N_{t-1} = 0, \chi_t = 1\}$, so that by (4.16), it follows that $\mathbb{E}\big[e^{\sigma M_t}\mathbf{1}\{N_t > 0\}\big]$ can be rewritten as

$$\mathbb{E}\big[e^{\sigma M_{t-1}} e^{\sigma\chi_t Y_t}\mathbf{1}\{N_{t-1} > 0\}\big] + \mathbb{E}\big[e^{\sigma Y_t}\mathbf{1}\{N_{t-1} = 0, \chi_t = 1\}\big]. \tag{4.17}$$

From the assumption $V_i \subset F$ it follows that $p = p_i(1|w) = p_i(1|w(V_i))$ and $\mathbb{E}\big[\chi_t Y_t \big| \mathcal{F}_{t-1}\big] = 0$. Since $-p \leqslant \chi_t Y_t \leqslant 1 - p$, the classical Hoeffding bound implies that $\mathbb{E}\big[e^{\sigma\chi_t Y_t}\big|\mathcal{F}_{t-1}\big] \leqslant e^{\sigma^2/8}$ and therefore the expression (4.17) can be bounded above by

$$\mathbb{E}\big[e^{\sigma M_t}\mathbf{1}_{\{N_t>0\}}\big] \leqslant e^{\sigma^2/8}\mathbb{E}\big[e^{\sigma M_{t-1}}\mathbf{1}_{\{N_{t-1}>0\}}\big] + e^{\sigma^2/8}\mathbb{E}\big[\mathbf{1}_{\{N_{t-1}=0,\chi_t=1\}}\big].$$

By iterating the inequality above and using the identity

$$\mathbf{1}\{N_t > 0\} = \mathbf{1}\{N_{\ell+1} > 0\} + \sum_{s=\ell+2}^{t} \mathbf{1}\{N_{s-1} = 0, \chi_s = 1\},$$

we obtain that $\mathbb{E}\big[e^{\sigma M_t}\mathbf{1}\{N_t > 0\}\big] \leqslant e^{(t-\ell+1)\sigma^2/8}\mathbb{P}(N_t > 0)$. Thus, collecting all these estimates, we deduce, by taking $\sigma = 4\lambda(t - \ell + 1)^{-1}$, that

$$\mathbb{P}(M_t > \lambda) \leqslant \exp\left\{-\frac{2\lambda^2}{t-\ell+1}\right\}\mathbb{P}(N_t > 0).$$

The left-tail probability $\mathbb{P}(M_t < -\lambda)$ is treated likewise. □

Proof of Proposition 4.9. For any $w \in \{0,1\}^{\{-\ell,\dots,-1\}\times F\setminus\{i\}}$ and $t > \ell + 1$,

$$M_t(w) = \big(\hat{p}_{(i,t)}(1|w) - p_i(1|w)\big)N_{(i,t)}(w),$$

so that

$$\mathbb{P}\Big(|\hat{p}_{(i,t)}(1|w) - p_i(1|w)| > \epsilon, N_{(i,t)}(w) \geqslant t^{1/2+\xi}\Big) \leqslant \mathbb{P}(|M_t(w)| > \epsilon t^{1/2+\xi}).$$

Thus the result follows from Proposition 4.8 by taking $\lambda = \epsilon t^{1/2+\xi}$. □

Proof of Lemma 4.11. For each $1 \leqslant s \leqslant t-1$, let Z_s be the random variable defined as in Lemma 4.10 with $\ell = 1$. Now we define $Y_s = Z_{2(s-1)+1}$ for $1 \leqslant s \leqslant \lfloor t/2 \rfloor$ and observe that $\mathcal{G}_s := \sigma(Y_1, \ldots, Y_s) \subset \mathcal{F}_{2s}$. Thus, by Lemma 4.10,

$$\mathbb{P}(Y_s = 1|\mathcal{G}_{s-1}) = \mathbb{E}\Big[\mathbb{P}\Big(Z_{2(s-1)+1} = 1|\mathcal{F}_{2(s-1)}\Big)|\mathcal{G}_{s-1}\Big] \geqslant p_{min}^{|F|+1}.$$

Define $q_* = p_{min}^{|F|+1}$. Then Lemma A.3 of Csiszár and Talata (2006a) implies for every $\nu \in]0,1[$,

$$\mathbb{P}\left(\frac{1}{\lfloor t/2 \rfloor}\sum_{s=1}^{\lfloor t/2 \rfloor} Y_s < \nu q_*\right) \leqslant \exp\left\{-\lfloor t/2 \rfloor \frac{q_*}{4}\Big(1-\nu\Big)^2\right\}.$$

Clearly $N_{(i,t)}(w) = \sum_{s=1}^{t-1} Z_s \geqslant \sum_{s=1}^{\lfloor t/2 \rfloor} Y_s$, so that it follows from the inequality above that

$$\mathbb{P}\left(N_{(i,t)}(w) < \nu q_* \lfloor t/2 \rfloor\right) \leqslant \exp\left\{-\lfloor t/2 \rfloor \frac{q_*}{4}\Big(1-\nu\Big)^2\right\}.$$

Finally, for any fixed $\nu \in (0,1)$ and all t large enough, $\lfloor t/2 \rfloor \frac{q_*}{4}(1-\nu)^2 > t^{1/2+\xi}$ and $\nu q_* \lfloor t/2 \rfloor > t^{1/2+\xi}$, implying the assertion. □

Now, suppose that $V_i \subset F$. In this case, notice that for $\ell \geqslant 1$ and $w \in \{0,1\}^{\{-\ell,\ldots,-1\}\times V_i}$, it holds that

$$p_i(1|w) = \varphi_i\left(\sum_{j\in V_i} W_{j\to i} \sum_{t=-\ell+1}^{0} g_j(-t) w_{j,t-1}\right). \tag{4.18}$$

Proof of Item 1 of Theorem 4.4. Using the definition of $\hat{V}_{(i,n)}^{(\epsilon)}$ and applying the union bound, we deduce that

$$\begin{aligned}\mathbb{P}\big(j \in \hat{V}_{(i,n)}^{(\epsilon)}\big) &= \mathbb{P}(\Delta_{(i,n)}(j) > \epsilon) \\ &\leqslant \mathbb{E}\left[\sum_{(w,v)} \mathbf{1}\left\{A_{(i,n)}^{w,v,j}, |\hat{p}_{(i,n)}(1|w) - \hat{p}_{(i,n)}(1|v)| > \epsilon\right\}\right] \end{aligned} \tag{4.19}$$

where $A_{(i,n)}^{w,v,j} := \{(w,v) \in \mathcal{T}_{(i,n)} \times \mathcal{T}_{(i,n)}^{w,j}\}$. Since $j \notin V_i$ and $V_i \subset F$, the configurations of any pair $(w,v) \in \mathcal{T}_{(i,n)} \times \mathcal{T}_{(i,n)}^{w,j}$ coincide in restriction to the set V_i. In other words, $w(V_i) = v(V_i)$. In particular, it follows from (4.18) that $p_i(1|w) = p_i(1|w(V_i)) = p_i(1|v(V_i)) = p_i(1|v)$.

Therefore, applying the triangle inequality, it follows that on $A_{(i,n)}^{w,v,j}$,

$$\mathbf{1}\big\{|\hat{p}_{(i,n)}(1|w) - \hat{p}_{(i,n)}(1|v)| > \epsilon\big\} \leqslant \sum_{u \in \{w,v\}} \mathbf{1}\big\{|\hat{p}_{(i,n)}(1|u) - p_i(1|u)| > \epsilon/2\big\},$$

so that the expectation in (4.19) can be bounded above by

$$2\mathbb{E}\left[\sum_w \mathbf{1}_{\{w \in \mathcal{T}_{(i,n)}\}} \sum_v \mathbf{1}_{\{v \in \mathcal{T}_{(i,n)}\}} \mathbf{1}\left\{|\hat{p}_{(i,n)}(1|v) - p_i(1|v)| > \epsilon/2\right\}\right]. \quad (4.20)$$

Now, since $\sum_w N_{(i,n)}(w) \leqslant n$, we have that

$$n \geqslant \sum_{w\colon N_{(i,n)}(w) \geqslant n^{1/2+\xi}} N_{(i,n)}(w) \geqslant n^{1/2+\xi} \, |\{w : N_{(i,n)}(w) \geqslant n^{1/2+\xi}\}|,$$

which implies that $|\mathcal{T}_{(i,n)}| \leqslant n^{1/2-\xi}$. From this last inequality and Proposition 4.9, which is stated in Section 4.6 below, we obtain the following upper bound for (4.20),

$$4n^{1/2-\xi} \exp\left\{-\frac{\epsilon^2 n^{2\xi}}{2}\right\} \mathbb{E}\left[\sum_w \mathbf{1}\big\{N_{(i,n)}(w) > 0\big\}\right]. \quad (4.21)$$

Since $\sum_w \mathbf{1}\big\{N_{(i,n)}(w) > 0\big\} \leqslant n$, the result follows from inequalities (4.19) and (4.21). □

Before proving Item 2 of Theorem 4.4, we will prove the following lemma.

Lemma 4.12. *Define for each $j \in V_i$,*

$$m_{i,j} := \max_{w,v \in \{0,1\}^{\{-1\} \times V_i} : w_{\{j\}^c} = v_{\{j\}^c}} |p_i(1|w) - p_i(1|v)|.$$

Then, under Assumption 4.3, we have that

$$\inf_{j \in V_i} m_{i,j} \geqslant \inf_{x \in K_i} \{\varphi_i'(x)\} \inf_{j \in V_i} \{|W_{j \to i}| g_j(0)\} = m_i > 0, \quad (4.22)$$

where K_i is defined in (4.9).

Proof. For each $j \in V_i$ take any pair $w, v \in \{0,1\}^{\{-1\}\times V_i}$ such that $w_{\{j\}^c} = v_{\{j\}^c}$ with $w_{j,-1} = 1$ and $v_{j,-1} = 0$. By Assumption 4.3, the function φ_i is differentiable such that, for $\ell = 1$,

$$|p_i(1|w) - p_i(1|v)| \geq \inf_{x\in K_i} \{\varphi_i'(x)\} \Big| \sum_{k\in V_i} W_{k\to i} \sum_{t=-\ell+1}^{0} g_k(-t)(w_{k,t} - v_{k,t})\Big|.$$

Since $\Big| \sum_{k\in V_i} W_{k\to i} \sum_{t=-\ell+1}^{0} g_k(-t)(w_{k,t} - v_{k,t})\Big| = |W_{j\to i}| g_j(0)$, the inequality above implies the first assertion of the lemma.

By Assumption 4.3, the function φ_i is strictly increasing ensuring that $\inf_{x\in K_i} \{\varphi_i'(x)\}$ 0. Thus, since for all $j \in [d]$ the sequence g_j is strictly positive and $V_i \neq \emptyset$ is finite, we clearly have that $m_i > 0$. □

We are now in position to conclude the proof of Theorem 4.4.

Proof of Item 2 of Theorem 4.4. Lemma 4.12 implies that m_i defined in (4.22) is positive. Let $0 < \epsilon < m_i$. If $j \in V_i$, Lemma 4.12 implies the existence of strings $w^*, v^* \in \{0,1\}^{\{-1\}\times F\setminus\{i\}}$ such that $w^*_{\{j\}^c} = v^*_{\{j\}^c}$ and

$$|p_i(1|w^*) - p_i(1|v^*)| = |p_i(1|w^*(V_i)) - p_i(1|v^*(V_i))| \geq m_i.$$

Denoting by $C_n = \{N_{(i,n)}(w^*) \geq n^{\xi+1/2}, N_{(i,n)}(v^*) \geq n^{\xi+1/2}\}$ it follows that

$$\mathbb{P}\left(j \notin \hat{V}^{(\epsilon)}_{(i,n)}\right) \leq \mathbb{P}(|\hat{p}_{(i,n)}(1|w^*) - \hat{p}_{(i,n)}(1|v^*)| < \epsilon, C_n) + \mathbb{P}(C_n^c). \quad (4.23)$$

Now notice that the first term on the right in (4.23) is upper bounded by

$$\sum_{u\in\{w^*,v^*\}} \mathbb{P}(|\hat{p}_{(i,n)}(1|u) - p_i(1|u)| > (m_i - \epsilon)/2, N_{(i,n)}(u) \geq n^{\xi+1/2}),$$

and since $m_i > \epsilon$, the result follows from Proposition 4.9 and Lemma 4.11, both stated in Section 4.6 above.

□

Proof of Item 3 of Theorem 4.4. Define for $n \in \mathbb{N}$ the sets

$$O_n = \left\{j \in F \setminus V_i : j \in V^{(\epsilon_n)}_{(i,n)}\right\} \text{ and } U_n = \left\{j \in V_i : j \notin V^{(\epsilon_n)}_{(i,n)}\right\}.$$

Applying the union bound and then Item 1, we infer that

$$\mathbb{P}(O_n) \leqslant 4\,(|F| - |V_i|)\, n^{3/2-\xi} \exp\left\{-\frac{\epsilon_n^2 n^{2\xi}}{2}\right\}.$$

Applying once more the union bound and then using Item 2, we also infer that

$$\mathbb{P}(U_n) \leqslant |V_i| \left(4 \exp\left\{-\frac{(m_i - \epsilon_n)^2 n^{2\xi}}{2}\right\} + \exp\left\{-O\left(n^{1/2+\xi}\right)\right\}\right).$$

Since $\{V_{(i,n)}^{(\epsilon)} \neq V_i\} = O_n \cup U_n$, we deduce that $\sum_{n=1}^{\infty} \mathbb{P}\left(V_{(i,n)}^{(\epsilon_n)} \neq V_i\right) < \infty$, so that the result follows from the Borel–Cantelli Lemma. □

Proof of Theorem 4.7. To deal with the case $V_i \not\subset F$, we couple the process $X = (X_t)_{t\in\mathbb{Z}}$ with the process $X^{[F]} = (X_t^{[F]})_{t\in\mathbb{Z}}$, where $X^{[F]}$ follows the same dynamics as X, defined in (4.2) and (4.3) for all $j \neq i$, except that (4.3) is replaced – for the fixed neuron i – by

$$\mathbb{P}(X_{i,t+1}^{[F]} = 1 | X_{-\infty:t}^{[F]}) = \varphi_i\Big(\sum_{j\in V_i\cap F} W_{j\to i} \sum_{s=L_{i,t}^{[F]}+1}^{t} g_j(t-s) X_{j,s}^{[F]}\Big). \quad (4.24)$$

Also, suppose that X and $X^{[F]}$ start from the same initial configuration, that is, $X_{-\infty:0} = X_{-\infty:0}^{[F]} = x$, where $x \in \Omega^{adm}$.

Proposition 4.13 below shows that Assumptions 4.5 and 4.6 imply the existence of a coupling between X and $X^{[F]}$ and of a constant $C > 0$ such that

$$\sup_{j\in[d]} \mathbb{P}\left(\exists t \in [1,n] : X_{j,t} \neq X_{j,t}^{[F]})\right) \leqslant C(e^{\alpha_0 n} \vee n) \sum_{j\notin V_i\cap F} |W_{j\to i}|. \quad (4.25)$$

Write

$$E_n = \bigcap_{1\leqslant s\leqslant n} \{X_{i,s} = X_{i,s}^{[F]}\}.$$

On E_n, instead of working with $X_{i,s}, 1 \leqslant s \leqslant n$, we can therefore work with its approximation $X_{i,s}^{[F]}, 1 \leqslant s \leqslant n$, having conditional transition probabilities (for neuron i) given by

$$p_i^{[F]}(1|w) = \varphi_i\Big(\sum_{j\in V_i\cap F} W_{j\to i} \sum_{s=-\ell+1}^{0} g_j(-s) w_{j,s-1}\Big)$$

which only depend on $w(V_i \cap F)$. As a consequence, on E_n the proof of Theorem 4.7 works as in the preceding section, except that we replace m_i by

$$m_i^{[F]} = \inf_{x \in K_i^{[F]}} \{\varphi_i'(x)\} \inf_{j \in V_i \cap F} \{|W_{j \to i}| g_j(0)\} > 0,$$

if $V_i \cap F \neq \emptyset$. Here $K_i^{[F]}$ is defined by

$$K_i^{[F]} = \Big[\sum_{j \in V_i^- \cap F} W_{j \to i} g_j(0), \sum_{j \in V_i^+ \cap F} W_{j \to i} g_j(0) \Big].$$

Finally, writing

$$O_n = \left\{ j \in F \setminus V_i : j \in V_{(i,n)}^{(\epsilon_n)} \right\} \text{ and } U_n = \left\{ j \in V_i \cap F : j \notin V_{(i,n)}^{(\epsilon_n)} \right\},$$

we obtain

$$\mathbb{P}(O_n) \leqslant \mathbb{P}(O_n \cap E_n) + \mathbb{P}(E_n^c), \ \mathbb{P}(U_n) \leqslant \mathbb{P}(U_n \cap E_n) + \mathbb{P}(E_n^c),$$

where as before in Theorem 4.4,

$$\mathbb{P}(O_n \cap E_n) \leqslant 4n^{3/2-\xi} \exp\left\{ -\frac{\epsilon_n^2 n^{2\xi}}{2} \right\}$$

and

$$\mathbb{P}(U_n \cap E_n)) \leqslant |F| \left(4 \exp\left\{ -\frac{(m_i^{[F]} - \epsilon_n)^2 n^{2\xi}}{2} \right\} + \exp\left\{ -O\left(n^{1/2+\xi} \right) \right\} \right).$$

Finally, by inequality (4.25),

$$\mathbb{P}(E_n^c) = \mathbb{P}(\exists t \in [1, n] : X_{i,t} \neq X_{i,t}^{[F]}) \leqslant C(e^{\alpha_0 n} \vee n) \sum_{j \notin V_i \cap F} |W_{j \to i}|,$$

for some constant C. This concludes the proof. □

In the remaining of this section, we prove the coupling result (4.25) used in the proof of Theorem 4.7. For that sake, let $F \subset [d]$, fix $i \in F$ and let $U_{j,t}$, $j \in [d], t \geqslant 1$, be an i.i.d. family of random variables uniformly distributed on $[0, 1]$.

The coupling is defined as follows. For any $x \in \Omega^{adm}$, we define $X_{j,t} = X^{[F]}_{j,t} = x_{j,t}$ for each $t \leqslant 0$ and $j \in [d]$. For each $t \geqslant 1$ and $j \in [d]$, we define

$$X_{j,t} = \begin{cases} 1, \text{ if } U_{j,t} > \varphi_j(\eta_{j,t-1}) \\ 0, \text{ if } U_{j,t} \leqslant \varphi_j(\eta_{j,t-1}) \end{cases}$$

and

$$X^{[F]}_{j,t} = \begin{cases} 1, \text{ if } U_{j,t} > \varphi_j(\eta^{[F]}_{j,t-1}) \\ 0, \text{ if } U_{j,t} \leqslant \varphi_j(\eta^{[F]}_{j,t-1}), \end{cases}$$

where for each $t \geqslant 0$ and $j \in [d]$,

$$\eta_{j,t} = \sum_{k \in V_j} W_{k \to j} \sum_{s=L_{j,t}+1}^{t} g_k(t-s) X_{k,s}$$

and, if $j \neq i$,

$$\eta^{[F]}_{j,t} = \sum_{k \in V_j} W_{k \to j} \sum_{s=L^{[F]}_{j,t}+1}^{t} g_k(t-s) X^{[F]}_{k,s}, \tag{4.26}$$

and finally

$$\eta^{[F]}_{i,t} = \sum_{k \in V_i \cap F} W_{k \to i} \sum_{s=L^{[F]}_{i,t}+1}^{t} g_k(t-s) X^{[F]}_{k,s}. \tag{4.27}$$

In other words, the process $X^{[F]}$ has exactly the same dynamics as the original process X, except that neuron i depends only on neurons belonging to $V_i \cap F$. Notice that we use the same uniform random variables $U_{j,t}$ to update the values of $X_{j,t}$ and of $X^{[F]}_{j,t}$. In this way we achieve a coupling between the two processes. We shall write $\mathbb{E}_x$ to denote the expectation with respect to this coupling. Then we have the following result.

Proposition 4.13. *Assume Assumptions 4.5 and 4.6, and let α_0 be defined as in* (4.12).
1. If $\alpha_0 > 0$, then

$$\sup_{j \in [d]} \mathbb{P}_x\Big(\bigcup_{s=1}^{t} \{X_{j,s} \neq X^{[F]}_{j,s}\}\Big) \leqslant C e^{\alpha_0 t} \sum_{k \notin V_i \cap F} |W_{k \to i}|. \tag{4.28}$$

2. Suppose now that $\alpha_0 = 0$ *and write for any* $j \in [d]$, $\varrho_j = \sum_{t=1}^{\infty} g_j(t)$, $\varrho = \sup_{j\in[d]} \varrho_i$. *Then*

$$\chi = (1 - p_*) + \gamma \sup_{j\in I} \sum_{k\in I} \varrho_k |W_{k\to I}| < 1, \tag{4.29}$$

and in this case

$$\sup_{j\in[d]} \mathbb{P}_x\Big(\bigcup_{s=1}^{t} \{X_{j,s} \neq X_{j,s}^{[F]}\}\Big) \leqslant \frac{\gamma\varrho t}{1-\chi} \sum_{k\notin V_i\cap F} |W_{k\to i}|. \tag{4.30}$$

Proof of Proposition 4.13. For notational convenience, let us assume that the starting configuration $x \in \Omega^{adm}$ satisfies $x_{i,0} = 1$ and extend the definition of g_j by defining $g_j(t) = 0$ for all $t \leqslant 0$ and $j \in [d]$.

We start proving Item 1. Recall the definition of the continuous operator $H(t)$ in (4.11). In the sequel, we set also $H(0) \equiv 0$.

Let for each $t \geqslant 0$,

$$D_{j,t} = 1\{L_{j,t} \neq L_{j,t}^{[F]}\},\ j \in [d],$$

and observe that

$$P_x(X_{j,t} \neq X_{j,t}^{[F]}) \leqslant E_x[D_{j,t}]. \tag{4.31}$$

Given $\mathcal{F}_t$, we update $D_{j,t}$ as follows. If neuron j spikes at time $t+1$ in both processes, then $D_{j,t+1} = 0$ regardless the value of $D_{j,t}$. By the definition of the coupling, this event occurs with probability $\varphi_j(\eta_{j,t} \wedge \eta_{j,t}^{[F]}) \geqslant p_*$. When $D_{j,t} = 1$, then $D_{j,t+1} = 1$ if and only if neuron j does not spike in both processes. Clearly, this event has probability $1-\varphi_j(\eta_{j,t} \wedge \eta_{j,t}^{[F]})$. Finally, if $D_{j,t} = 0$, then $D_{j,t+1} = 1$ if and only if neuron j spikes only in one of the two processes. This event occurs with probability $|\varphi_j(\eta_{j,t}) - \varphi_j(\eta_{j,t}^{[F]})|$. Thus for all $j \in [d]$, we have

$$\mathbb{E}_x(D_{j,t+1}|\mathcal{F}_t) = D_{j,t}(1 - \varphi_j(\eta_{j,t} \wedge \eta_{j,t}^{[F]}) + |\varphi_j(\eta_{j,t}) - \varphi_j(\eta_{j,t}^{[F]})|(1 - D_{j,t}). \tag{4.32}$$

Since φ_i is Lipschitz with Lipschitz constant γ and $L_{i,t} = L_{i,t}^{[F]}$ on $\{D_{i,t} = 0\}$,

we have on this event,

$$\begin{aligned}
&\frac{1}{\gamma}\,|\varphi_i(\eta_{i,t}) - \varphi_i(\eta_{i,t}^{[F]})| \\
&\quad \leqslant |\eta_{i,t} - \eta_{i,t}^{[F]}| \\
&\quad \leqslant \sum_{k\in V_i\cap F} |W_{k\to i}| \sum_{s=L_{i,t}+1}^{t} g_k(t-s)|X_{k,s} - X_{k,s}^{[F]}| \\
&\qquad\qquad + \sum_{k\notin V_i\cap F} |W_{k\to i}| \sum_{s=1}^{t} g_k(t-s) \\
&\quad \leqslant \sum_{k\in V_i\cap F} |W_{k\to i}| \sum_{s=1}^{t+1} g_k(t-s)|X_{k,s} - X_{k,s}^{[F]}| \\
&\qquad\qquad + \sum_{k\notin V_i\cap F} |W_{k\to i}| \sum_{s=1}^{t+1} g_k(t-s), \quad (4.33)
\end{aligned}$$

where we have used that $g_k(-1) = 0$ in order to replace the sum $\sum_{s=1}^{t}$ by $\sum_{s=1}^{t+1}$. Moreover, we have used that $L_{i,t} = L_{i,t}^{[F]} \geqslant 0$ for all $t \geqslant 0$, by our choice of x.

Similarly, for all $j \neq i$, we have on $\{D_{j,t} = 0\}$,

$$\frac{1}{\gamma}\,|\varphi_j(\eta_{j,t}) - \varphi_j(\eta_{j,t}^{[F]})| \leqslant \sum_{k\in V_j} |W_{k\to j}| \sum_{s=1}^{t+1} g_k(t-s)|X_{k,s} - X_{k,s}^{[F]}|. \quad (4.34)$$

For each $j \in [d]$, let $\delta_j(t) = \mathbb{E}_x(D_{j,t})$ and write $\delta(t) = (\delta_j(t))_{j\in[d]}$ for the associated column vector. Taking expectation in (4.32)–(4.34) and using that $\mathbb{E}_x|X_{k,s} - X_{k,s}^{[F]}| \leqslant \delta_k(s)$ (see (4.31)), we obtain

$$\delta(t+1) \leqslant H * \delta(t+1) + \gamma\Sigma_i(F)g * 1(t+1)e_i, \quad (4.35)$$

where e_i is the $i-$the unit vector. In the above formula,

$$(H * \delta(t))_j = \sum_{k\in[d]} \sum_{s=0}^{t} H_{j,k}(t-s)\delta_k(s)$$

is the operator convolution product, and the inequality in (4.35) has to be understood coordinate-wise.

Now let α_0 be as in (4.12) and introduce $\widetilde{H}(t) = e^{-\alpha_0 t} H(t)$, $\widetilde{\delta}(t) = e^{-\alpha_0 t}\delta(t)$ and $\widetilde{1}(t) = e^{-\alpha_0 t}$. Multiplying the above inequality with $e^{-\alpha_0(t+1)}$, we obtain

$$\widetilde{\delta}(t+1) \leqslant \widetilde{H} * \widetilde{\delta}(t+1) + \widetilde{g} * \widetilde{1}(t+1)\gamma\Sigma_i(F)e_i.$$

Let $\|\widetilde{\delta}\|_1 = (\|\widetilde{\delta}_i\|_1, i \in [d])$ be the column vector where each entry is given by $\|\widetilde{\delta}_i\|_1 = \sum_{t=0}^{\infty} \delta_i(t)$. Then we obtain, summing over $t \geqslant 0$,

$$\|\widetilde{\delta}\|_1 \leqslant \Lambda(\alpha_0)\|\widetilde{\delta}\|_1 + \frac{1}{1-e^{-\alpha_0}}\|\widetilde{g}\|_1 \Sigma_i(F)e_i,$$

implying that

$$(Id - \Lambda(\alpha_0))\|\widetilde{\delta}\|_1 \leqslant \frac{1}{1-e^{-\alpha_0}}\|\widetilde{g}\|_1 \Sigma_i(F)e_i. \tag{4.36}$$

By (4.12), $Id - \Lambda(\alpha_0)$ is invertible, and it is well-known that the operator norm of the inverse is bounded by

$$\|\!|(Id - \Lambda(\alpha_0))^{-1}|\!\| \leqslant (1 - \|\!|\Lambda(\alpha_0)|\!\|)^{-1} = C(\alpha_0).$$

Moreover, $(Id - \Lambda(\alpha_0))^{-1} : \mathbb{R}_+^d \to \mathbb{R}_+^d$, where $\mathbb{R}_+^d = \{(\xi_j)_{j\in[d]} : \xi_j \geqslant 0\}$. Therefore, (4.36) implies

$$\sup_{j\in[d]} \|\widetilde{\delta}_j\|_1 \leqslant \left[\frac{1}{1-e^{-\alpha_0}}\|\widetilde{g}\|_1 \Sigma_i(F)\right] C(\alpha_0). \tag{4.37}$$

By using the union bound and (4.31), it follows that

$$\sup_{j\in[d]} \mathbb{P}_x(\exists s \in [1,t] : X_{j,s} \neq X_{j,s}^{[F]}) \leqslant \sup_{j\in[d]} \sum_{s=1}^{t} \delta_j(s) \leqslant e^{\alpha_0 t} \sup_{j\in[d]} \|\widetilde{\delta}_j\|_1,$$

which implies the assertion of Item 1.

The proof of Item 2 is similar to the above argument, except that now it is possible to work directly with $g(t)$ instead of $\widetilde{g}(t)$. In this case, we write simply $\bar{\delta}(t) = \sup_{j\in[d]} \delta_j(t)$. (4.35) implies that

$$(\sup_{0\leqslant s\leqslant t} \bar{\delta}(s)) \leqslant \chi\,(\sup_{0\leqslant s\leqslant t} \bar{\delta}(s)) + \gamma\varrho \sum_{k\notin V_i\cap F} |W_{k\to i}|,$$

which implies the assertion.

□

Proof of Inequality (4.25). The coupling inequality (4.25) follows now directly from (4.28) ((4.30), respectively). □

4.8 Exercises

Exercise 4.1. Show that the GL neuron models is a stochastic chain with memory of variable length taking values in the alphabet $A = \{0,1\}^d$ and which associates to the past $x_{-\infty:-1} = (x_{i,t})_{i\in[d],t\in\mathbb{Z}_-}$ the context $x_{L(x):-1}$ where $L(x) = \min\{L_i(x) : i \in [d]\}$ and $L_i(x) = \sup\{t \leqslant -1 : x_{i,t} = 1\}$. Find the associated kernel of transition probabilities.

Exercise 4.2. Write the mathematical expression for the conditional likelihood of $X_{i,1:n} = x_{i,1:n}$ given $X_{V_i,1:(n-1)} = x_{V_i,1:(n-1)}$ and $X_{i,0} = 1$. Use this expression to show that $\hat{p}_{(i,n)}(1|w)$ defined in (4.5) are maximum likelihood estimators of the transition probabilities defined in (4.3).

Exercise 4.3. Let $U_{t,i} = \sum_{j=1}^{p} W_{j\to i} \sum_{s=L_{i,t}+1}^{t} g_j(t+1-s)X_{j,s}$ with $g_j(t) = e^{-\alpha_j t}$ where $\alpha_j \geqslant 0$ and denote $U_t = (U_{1,t},\dots,U_{p,t})$. Prove that the stochastic chain $(U_t)_{t\in\mathbb{Z}}$ is Markovian.

Exercise 4.4. Find a postsynaptic current pulses g_i for which the corresponding stochastic chain $(U_t)_{t\in\mathbb{Z}}$ as defined in Exercise 4.3 is not Markovian.

Exercise 4.5. Prove Lemma 4.10.

5 Sparse space-time stochastic systems

In this chapter, we will first introduce space-time stochastic systems in which the family of transition probabilities admits a space-time *Kalikow decomposition*. Next, we will show that when this space-time decomposition is sufficiently sparse, one can exhibit a perfect simulation algorithm to simulate samples of the space-time stochastic system from its (unique) invariant measure. As an application of this perfect simulation algorithm, we will then derive some Hoeffding-type concentration inequalities. These concentration inequalities are, in turn, crucial to deal with the statistical question of how to approximate transition probabilities of sparse space-time systems by linear combinations of a given dictionary, as presented in the last sections of this chapter. The materials presented in this section is based on the article Ost and Reynaud-Bouret (2020).

5.1 Stochastic framework and notation

We consider a stationary stochastic chain $\mathbf{X} = (X_{i,t})_{i\in I, t\in\mathbb{Z}}$ taking values in $\{0,1\}^{I\times\mathbb{Z}}$, defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where I is a countable (possi-

bly infinite) set. The configuration of $\mathbf{X}$ at time $t \in \mathbb{Z}$ is denoted by $X_t = (X_{i,t}, i \in I)$. For $s, t \in \mathbb{Z}$ with $s < t$, $X_{i,s:t}$ stands for the collection $(X_{i,s}, \ldots, X_{i,t})$ and $X_{s:t}$ for the collection $(X_{i,r})_{i\in I, s\leqslant r\leqslant t}$. We write $X_{-\infty:t}$ to denote the past history $(\ldots, X_{t-1}, X_t)$ of $\mathbf{X}$ at time $t+1$. Note that the past histories have space-time components. For $F \subset I$ and $t \in \mathbb{Z}$, $X_{F,t} = (X_{i,t}, i \in F)$ denotes the configuration of $\mathbf{X}$ at time t restricted to set F. More generally, X_S denotes the collection $(X_{i,t})_{(i,t)\in S}$ for any subset $S \subset I \times \mathbb{Z}$. We will use similar notation for deterministic space-time configurations belonging to either $\{0,1\}^{I\times\mathbb{Z}}$ or $\{0,1\}^{I\times\mathbb{Z}_-}$.

Remark 5.1. Many of the results presented in this chapter could also be formulated for space-time models taking values in $A^{I\times\mathbb{Z}}$, where the set A is finite. We consider only the case $A = \{0,1\}$ for two reasons. First, by doing so, we keep the exposition as simple as possible, making the argument more transparent. Moreover, our main motivation to consider space-time models come from networks of spiking neurons. In this context, the set I represents the neurons in the network. As already mentioned in Chapter 4, neurons talk to each other by firing sequences of spikes. By discretizing the time into bins of small width, one can model spike train data by assigning the symbol 1 to the bins in which a neuron has spiked and the symbol 0 to all the remaining ones. With the notation above, this means to set $X_{i,t} = 1$ if neuron $i \in I$ spike at the t-th bin and $X_{i,t} = 0$ otherwise.

Throughout this chapter, we will work under the following assumption.

Assumption 5.2. For each $t \in \mathbb{Z}$, the components of X_{t+1} are independent given the past history $X_{-\infty:t}$, that is,

$$\mathbb{P}(X_{J,t+1} = a_J | X_{-\infty:t} = x) = \prod_{i\in J} \mathbb{P}(X_{i,t+1} = a_i | X_{-\infty:t} = x),$$

for any finite $J \subset I$, $a_J := (a_i)_{i\in J} \in \{0,1\}^J$ and $\mathbb{P}$-a.e. $x \in \{0,1\}^{I\times\mathbb{Z}_-}$.

Since the stochastic chain $\mathbf{X}$ is stationary, Assumption 5.2 implies that the dynamics of $\mathbf{X}$ is fully characterized by the family of transition probabilities

$$p_i(x) = \mathbb{P}(X_{i,0} = 1 | X_{-\infty:-1} = x),\ x \in \{0,1\}^{I\times\mathbb{Z}_-},\ i \in I.$$

These transition probabilities are all assumed to be measurable functions of $x \in \{0,1\}^{I\times\mathbb{Z}_-}$.

Remark 5.3. In the context of stochastic modeling of spike train data, $p_i(x)$ models the probability of neuron i to spike at a given time given the spike history up

to that time of all the neurons in the network (including neuron i itself). The examples discussed in the next section provide different ways of describing the value of the spiking probability $p_i(x)$ as a function of spike history x.

Hereafter, we need the following notation. For any neighborhood $S \subset I \times \mathbb{Z}_-$ and $x, y \in \{0,1\}^{I\times\mathbb{Z}_-}$, we write $x \overset{S}{=} y$ to indicate $y_S = x_S$. For any real-valued function f on $\{0,1\}^{I\times\mathbb{Z}_-}$ and subset $S \subset I \times \mathbb{Z}_-$, we say f is *cylindrical* in S and write $f(x) = f(x_S)$, if $f(x) = f(y)$ for any $x, y \in \{0,1\}^{I\times\mathbb{Z}_-}$ such that $x \overset{S}{=} y$.

5.2 Space-time decomposition and perfect simulation

In this section, we first introduce the notion of space-time decomposition of a family of transition probabilities. We then given some examples of stochastic models for which the associated family of transition probabilities admit a space-time decomposition. Finally, we will show how to build perfect simulation algorithms for space-time models whose transition probabilities admit sparse space-time decompositions.

5.2.1 Definition

We denote by $\mathcal{V}$ the collection of finite neighborhoods, i.e. finite subsets of $I \times \mathbb{Z}_-$ and we consider processes for which the following decomposition holds.

Assumption 5.4 (Space-time decomposition). For all S in $\mathcal{V}$ and i in I, there exists a $[0,1]$-valued measurable function $p_i^S(.)$, cylindrical in S, and a non negative weight $\lambda_i(S)$, such that for all $x \in \{0,1\}^{I\times\mathbb{Z}_-}$ and $i \in I$,

$$\begin{cases} p_i(x) = \lambda_i(\emptyset)\, p_i^{\emptyset}(x) + \sum_{S\in\mathcal{V}, S\neq\emptyset} \lambda_i(S)\, p_i^S(x), \\ \\ \sum_{S\in\mathcal{V}} \lambda_i(S) = 1. \end{cases}$$

This decomposition can be interpreted as follows. At each time step, to decide which value to assign to site i, we first choose a random space-time neighborhood in $\mathcal{V}$ according to the distribution λ_i. Once this neighborhood is chosen, say S is the chosen neighborhood, we then assign the value 1 to the site i with probability $p_i^S(x_S)$. Note that $p_i^S(x_S)$ depends only on the past history restricted to S. Note also that $p_i^{\emptyset}(x)$ does not depend on x at all, and for this reason we denote this value $p_i^{\emptyset}$ in what follows.

Such a space-time decomposition of the transition probabilities $\{p_i(x), i \in I, x \in \{0,1\}^{I\times\mathbb{Z}-}\}$ generalizes the classical Kalikow decomposition introduced in Kalikow (1990) and further developed in Comets, Fernandéz, and Ferrari (2002), Galves, Garcia, et al. (2013) and Galves and Löcherbach (2013). The main difference consists in not forcing the neighborhoods S that lie in the support of λ_i to be nested. This helps us to exploit the fact that in many cases the distributions λ_i charge very few neighborhoods and that the cardinality of this neighborhood is usually very small, if the nesting is not forced. We speak in this case of *probabilistic sparsity*.

Remark 5.5. If we denote $q_i(x) = \mathbb{P}(X_{i,0} = 0 | X_{-\infty:-1} = x) = 1 - p_i(x)$ and $q_i^S(x) = 1 - p_i^S(x)$ for all $i \in I$, $x \in \{0,1\}^{I\times\mathbb{Z}-}$, we can also write

$$q_i(x) = \lambda_i(\emptyset) q_i^{\emptyset}(x) + \sum_{S\in\mathcal{V}, v\neq\emptyset} \lambda_i(S) q_i^S(x),$$

where for each $S \in \mathcal{V}$, the function q_i^S is cylindrical in S.

Remark 5.6. For a given space-time decomposition, one can use Remark 5.5 to deduce that for all $i \in I$,

$$\inf_{x\in\{0,1\}^{I\times\mathbb{Z}-}} p_i(x) + \inf_{x\in\{0,1\}^{I\times\mathbb{Z}-}} q_i(x) \geqslant \lambda_i(\emptyset).$$

More generally, for any $S \in \mathcal{V}$, one can show that

$$\inf_{x\in\{0,1\}^{I\times\mathbb{Z}-}} \left\{ \inf_{y\in\{0,1\}^{I\times\mathbb{Z}-}:\, y \overset{S}{=} x} p_i(y) + \inf_{y\in\{0,1\}^{I\times\mathbb{Z}-}:\, y \overset{S}{=} x} q_i(y) \right\} \geqslant \lambda_i(\emptyset) + \sum_{V\subseteq S, V\neq\emptyset} \lambda_i(V).$$

One can also show that the space-time decomposition is not unique. This fact raises the question of whether there is an "optimal" decomposition of a given transition probability. Such a question, however, will not be discussed here.

5.2.2 Main examples

Example 5.7 (Chains of infinite order). Suppose I is a singleton, say $I = \{1\}$, and let us write X_t and $p(x)$ instead of $X_{1,t}$ and $p_1(x)$ for convenience. In this

case, the stochastic chain $\mathbf{X}$ is described by the transition probability $\{p(x), x \in \{0,1\}^{\mathbb{Z}-}\}$. Denote for $\ell \in \mathbb{Z}_+$, $\underline{\ell}$ the set $\{-\ell, ..., -1\}$ and

$$\beta_\ell = \sup_{x \in \{0,1\}^{\mathbb{Z}-}} \sup_{\substack{y,z \in \{0,1\}^{\mathbb{Z}-} \text{ s.t.} \\ y \overset{\underline{\ell}}{=} z \overset{\underline{\ell}}{=} x}} \{|p(y) - p(z)|\}.$$

If there exist $\ell_0 \geqslant 1$ such that $\beta_\ell = 0$ for all $\ell \geqslant \ell_0$, then the stochastic chain $\mathbf{X}$ is called Markov Chain of Order ℓ_0. Otherwise, $\mathbf{X}$ is called *chain of infinite Order*. We say that the transition probability $\{p(x), x \in \{0,1\}^{\mathbb{Z}-}\}$ of a chain of infinite order is *continuous* if $\beta_\ell \to 0$ as $\ell \to \infty$. In this case, the sequence $(\beta_\ell)_{\ell \in \mathbb{Z}_+}$ is called the *continuity rate* of the chain. We refer the reader to Ferrari, Fernandéz, and Galves (2001) for a comprehensive introduction to Chains of Infinite Order.

Denote $q(x) = 1 - p(x)$ for all $x \in \{0,1\}^{\mathbb{Z}-}$ and define for each $\ell \in \mathbb{Z}_+$,

$$\alpha(\ell) = \inf_{x \in \{0,1\}^{\mathbb{Z}-}} \left\{ \inf_{y \in \{0,1\}^{\mathbb{Z}-} \ s.t. \ y \overset{\underline{\ell}}{=} x} p(y) + \inf_{y \in \{0,1\}^{\mathbb{Z}-} \ s.t. \ y \overset{\underline{\ell}}{=} x} q(y) \right\}.$$

With this notation, let us denote the distribution λ which has support only on the sets $\underline{\ell}$'s, defined as

$$\lambda(\underline{\ell}) = \alpha(\ell) - \alpha(\ell - 1), \tag{5.1}$$

where $\alpha(0) = \lambda(\emptyset) = \inf_{x \in \{0,1\}^{\mathbb{Z}-}} p(x) + \inf_{x \in \{0,1\}^{\mathbb{Z}-}} q(x)$. One can show (see Exercise 5.2) that every continuous transition probability $\{p(x), x \in \{0,1\}^{\mathbb{Z}-}\}$ admits a decomposition of the form:

$$\begin{cases} p(x) = \lambda(\emptyset) p^{\emptyset} + \sum_{\ell \in \mathbb{Z}_+} \lambda(\underline{\ell}) p^{\underline{\ell}}(x) \\ \lambda(\emptyset) + \sum_{\ell \in \mathbb{Z}_+} \lambda(\underline{\ell}) = 1. \end{cases} \tag{5.2}$$

Moreover, (5.2) is a space-time decomposition since $p^{\emptyset} \in [0,1]$ and for each $\ell \in \mathbb{Z}_+$, $\{p^{\underline{\ell}}(x), x \in \{0,1\}^{\mathbb{Z}-}\}$ is a transition probability of a Markov chain of order ℓ.

Example 5.8 (Discrete-time linear Hawkes processes). For each $i \in I$ and $x \in \{0,1\}^{I \times \mathbb{Z}-}$, let

$$\psi_i(x) = \nu_i + \sum_{s \in \mathbb{Z}_-} \sum_{j \in I} h_{j \to i}(-s) x_{j,s}, \text{ and } \begin{cases} p_i(x) = \psi_i(x), & \text{if } \psi_i(x) \in [0,1], \\ p_i(x) = 1, & \text{if } \psi_i(x) > 1, \\ p_i(x) = 0, & \text{if } \psi_i(x) < 0. \end{cases} \tag{5.3}$$

Clearly, in this model, $p_i(x)$ is a function of the $\psi_i(x)$. Let us explain in the context of spike trains modeling (see Remark 5.1) each one of the terms defining the function $\psi_i(x)$. The parameter $\nu_i \geqslant 0$ represents the spontaneous activity of neuron i, that is its ability to produce spikes when there is no interaction. The interaction function $h_{j\to i}$ measures the amount of excitation (if positive) or inhibition (if negative) that a spike of neuron j has on neuron i after a delay $-s$ (a spike of neuron j with delay $-s$ corresponds to $x_{j,s} = 1$).

Remark 5.9. Hawkes processes are systems of interacting point processes on the real line. They have been introduced in Hawkes (1971) to model seismic shocks. More recently, the Hawkes processes have been used in areas such as finance Bacry, Mastromatteo, and Muzy (2015) and neuroscience Chen et al. (2019), Chevallier, Cáceres, et al. (2015), Chevallier, Duarte, et al. (2019), Chevallier and Ost (2020), Chornoboy, Schramm, and Karr (1988), Ditlevsen and Löherbach (2017), Hansen, Reynaud-Bouret, and Rivoirard (2015a), Hodara and Löcherbach (2017a), Johnson (1996), and Pernice et al. (2011).

The interaction between the different components of a Hawkes process is described by what is called the intensity function, which gives the probability of observing an arrival (spikes in context of spike train modeling) in a very short time interval, given the past history of the process. The function $\psi_i(x)$ corresponds to the discrete time description of this intensity function. For this reason, we call *discrete-time Hawkes processes* the space-time models $\mathbf{X}$ whose transition probabilities $p_i(x)$ are as above. The name linear comes from the fact that the function $\psi(x)$ is a linear function of x.

To check that $p_i(x)$ admits a space-time decomposition, we need to introduce some new notation. For each $i \in I$, we write

$$A_i^+ = \{(j,s) \in I \times \mathbb{Z}_- : h_{j\to i}(-s) > 0\},$$

and

$$A_i^- = \{(j,s) \in I \times \mathbb{Z}_- : h_{j\to i}(-s) < 0\},$$

and define the maximal excitatory (respectively inhibitory) strength by

$$\Sigma_i^+ = \sum_{(j,s)\in A_i^+} |h_{j\to i}(-s)| \text{ and } \Sigma_i^- = \sum_{(j,s)\in A_i^-} |h_{j\to i}(-s)|.$$

Let us assume that

$$0 \leqslant \nu_i - \Sigma_i^- \quad \text{and} \quad \nu_i + \Sigma_i^+ \leqslant 1, \tag{5.4}$$

which implies in particular that whatever the past configuration $x \in \{0,1\}^{I\times\mathbb{Z}_-}$, the transition probability $p_i(x) \in [0,1]$ is always equal to $\psi_i(x)$. It also implies that $\Sigma_i^+ + \Sigma_i^- \in [0,1]$.

With the notation and conditions above, one can easily check (see Exercise 5.3) that $p_i(x)$ admits a space-time decomposition where:

$$\begin{cases} \lambda_i(\emptyset) & = & 1-(\Sigma_i^+ + \Sigma_i^-) & \text{which is } \geqslant 0 \text{ since } 0 \leqslant \Sigma_i^+ + \Sigma_i^- \leqslant 1, \\ p_i^{\emptyset} & = & \frac{\nu_i - \Sigma_i^-}{\lambda_i(\emptyset)} & \text{which is } \leqslant 1 \text{ since } \nu_i + \Sigma_i^+ \leqslant 1, \\ \lambda_i(\{(j,s)\}) & = & |h_{j\to i}(-s)| & \text{for all } (j,s) \in A_i^+ \cup A_i^-, \\ p_i^{\{(j,s)\}}(x) & = & x_{j,s} & \text{for all } (j,s) \in A_i^+, \\ p_i^{\{(j,s)\}}(x) & = & (1-x_{j,s}) & \text{for all } (j,s) \in A_i^-. \end{cases} \tag{5.5}$$

It is moreover sufficient to assume that $\Sigma_i^+ + \Sigma_i^- < 1$ to have $\lambda_i(\emptyset) > 0$.

The discrete-time linear Hawkes model is an interesting example, because even if the true interaction graph, that is the set of edges $(j,i) \in I \times I$ for which $h_{j\to i}$ is non zero, is complete, the neighborhoods $S \in \mathcal{V}$ of the space-time decomposition have cardinality at most 1 almost surely. This *probabilistic sparsity* is exploited in the sequel to obtain concentration inequalities.

Example 5.10 (GL neuron model). Let $W_{j\to i} \in \mathbb{R}$ with $i, j \in I$, be a collection of real numbers such that $W_{j\to j} = 0$ for all j. For each $i \in I$, let $\varphi_i : \mathbb{R} \to [0,1]$ be a non-decreasing measurable function and $g_i = (g_i(\ell))_{\ell\in\mathbb{Z}_+}$ be a sequence of strictly positive real numbers.

For each $x \in \{0,1\}^{I\times\mathbb{Z}_-}$ and $i \in I$, we define $L_i(x) = \sup\{s \in \mathbb{Z}_- : x_{i,s} = 1\}$. The stochastic chain $\mathbf{X}$ satisfies a GL neuron model if the transition probabilities $\{p_i(x), i \in I, x \in \{0,1\}^{I\times\mathbb{Z}_-}\}$ are given by (cf. Equation (4.3))

$$p_i(x) = \begin{cases} \varphi_i(0), \text{if } L_i(x) = -1, \\ \varphi_i\left(\sum_{j\in I} W_{j\to i} \sum_{s=L_i(x)+1}^{-1} g_j(-s) x_{j,s}\right), \text{otherwise.} \end{cases} \tag{5.6}$$

Notice that here the set of neurons I is possibly infinite, whereas the GL neuron model as presented in Chapter 4 has a finite number d of neurons. This extension of the GL neuron models enables us to deal with arbitrarily high dimensional systems, taking into account the fact that the brain consists of a huge (about 10^{11}) number of interacting neurons. Notice that $L_i(x)$ appearing in Equation (5.6) corresponds to random variable $L_{i,-1}$ defined in Equation (4.1), given that $X_{-\infty:-1} = x$.

Linear spike rate functions. Let us consider the particular case where the parameters of the model are such that $\varphi_i(u) = \nu_i + u$ with $\nu_i \geqslant 0$ for each $i \in I$.

Similarly to Example 5.8, let us denote for each $i \in I$,

$$A_i^+ = \{(j,s) \in I \times \mathbb{Z}_- : W_{j\to i}\, g_j(-s) > 0\},$$

and

$$A_i^- = \{(j,s) \in I \times \mathbb{Z}_- : W_{j\to i}\, g_j(-s) < 0\},$$

and define the maximal excitatory (respectively inhibitory) strength by

$$\Sigma_i^+ = \sum_{(j,s)\in A_i^+} |W_{j\to i}| g_j(-s) \text{ and } \Sigma_i^- = \sum_{(j,s)\in A_i^-} |W_{j\to i}| g_j(-s).$$

We also assume that

$$0 \leqslant \nu_i - \Sigma_i^- \quad \text{and} \quad \nu_i + \Sigma_i^+ \leqslant 1. \tag{5.7}$$

Under these assumptions, one can check (see Exercise 5.4) that the transition probabilities (5.6) also satisfy Assumption 5.4. Specifically, we can use

$$\begin{cases} \lambda_i(\emptyset) &= 1 - (\Sigma_i^+ + \Sigma_i^-), \\ p_i^{\emptyset} &= \frac{\nu_i - \Sigma_i^-}{\lambda_i(\emptyset)}, \\ \lambda_i(\{(j,s)\}^{\downarrow i}) &= |W_{j\to i}| g_j(-s) & \text{for all } (j,s) \in A_i^+ \cup A_i^-, \\ p_i^{\{(j,s)\}^{\downarrow i}}(x) &= x_{j,s} 1_{x_{i,s:-1}=0} & \text{for all } (j,s) \in A_i^+, \\ p_i^{\{(j,s)\}^{\downarrow i}}(x) &= (1-x_{j,s}) 1_{x_{i,s:-1}=0} & \text{for all } (j,s) \in A_i^-, \end{cases} \tag{5.8}$$

where $\{(j,s)\}^{\downarrow i} = \{(j,s),(i,s),\ldots,(i,-1)\}$ is the augmentation of the set $\{(j,s)\}$ on the coordinate i for each $(j,s) \in A_i^+ \cup A_i^-$. Notice that $\lambda_i(\emptyset) \geqslant 0$ since $0 \leqslant \Sigma_i^+ + \Sigma_i^- \leqslant 1$ and $p_i^{\emptyset} \leqslant 1$ since $\nu_i + \Sigma_i^+ \leqslant 1$. In this case, the neighborhoods $S \in \mathcal{V}$ have cardinality either 0 (when $S = \emptyset$) or $s+1$ (when $S = \{(j,s)\}^{\downarrow i}$ with $j \neq i$) or s (when $S = \{(i,s)\}^{\downarrow i}$).

Non-linear spike rate functions. In the previous work of Galves and Löcherbach (2013), the space-time decomposition is restricted to growing sequences of neighborhoods S that are indexed by their range in time. For each $i \in I$, one assumes that there exists a growing sequence $J_i(1) = \{i\}$, $J_i(\ell) \subset J_i(\ell+1)$ of subsets of I that corresponds to the space positions that are needed when looking at a past of length ℓ, so that we can form $S_i(\ell) = J_i(\ell) \times \underline{\ell}$, defining a growing sequence of subsets of $I \times \mathbb{Z}_-$.

Next let us introduce the following quantities:

$$\alpha_i(\ell) = \inf_{x\in\{0,1\}^{I\times\mathbb{Z}-}} \left\{ \inf_{y\in\{0,1\}^{I\times\mathbb{Z}-}: y \overset{S_i(\ell)}{=} x} p_i(y) + \inf_{y\in\{0,1\}^{I\times\mathbb{Z}-}: y \overset{S_i(\ell)}{=} x} q_i(y) \right\}$$

and $\lambda_i(S_i(\ell)) = \alpha_i(\ell) - \alpha_i(\ell-1)$, where for each $i \in I$, $q_i(y) = 1 - p_i(y)$ and $\lambda_i(\emptyset) = \alpha_i(0) = \inf_{x\in\{0,1\}^{I\times\mathbb{Z}-}} p_i(x) + \inf_{x\in\{0,1\}^{I\times\mathbb{Z}-}} q_i(x)$.

Let us assume that

$$\sup_{i\in I}\sum_{j\in I}|W_{j\to i}| < \infty, \quad \sum_{\ell\in\mathbb{Z}_+}\sup_{i\in I} g_i(\ell) < \infty \text{ and } \sup_{i\in I}|\varphi_i(u) - \varphi_i(v)| \leqslant \gamma|u-v|, \tag{5.9}$$

where γ is a positive constant.

It has been proved in Galves and Löcherbach (2013) (see Proposition 2) that the transition probabilities $\{p_i(x), x \in \{0,1\}^{I\times\mathbb{Z}-}\}$ admit the following space-time decomposition:

$$\begin{cases} p_i(x) = \lambda_i(\emptyset)p_i^{\emptyset} + \sum_{\ell\in\mathbb{Z}_+}\lambda_i(S_i(\ell))p_i^{S_i(\ell)}(x), \\ \lambda_i(\emptyset) + \sum_{\ell=1}^{+\infty}\lambda_i(S_i(\ell)) = 1, \end{cases} \tag{5.10}$$

with, $p_i^{\emptyset} \in [0,1]$ and for $\ell \geqslant 1$, $p_i^{S_i(\ell)}(x)$ is a $[0,1]$-valued measurable function which is cylindrical in $S_i(\ell)$. Exercise 5.5 asks to prove this fact.

Hence, the transition probabilities p_i's also satisfy Assumption 5.4 in the non-linear case. The neighborhoods $S \in \mathcal{V}$ have cardinality either 0 (when $S = \emptyset$) or $\ell|J_i(\ell)|$ (when $S = S_i(\ell)$). Note that in the non-linear case the neighborhoods $S_i(\ell)$ are dense in time by construction, whereas in the linear case one can obtain a stronger *probabilistic sparsity.*

5.2.3 Main properties

In this section we present an algorithm to stimulate (construct) $X_{i,t}$ at any fixed site $(i,t) \in I \times \mathbb{Z}$. The first step of our algorithm uses the distribution λ_i to obtain a space-time neighborhood of (i,t). More precisely, because the distribution λ_i gives a neighborhood for site i at time 0, we need to shift it at time t to obtain a realization of the random neighborhood for site i at time t. Hence if for every $t \in \mathbb{Z}$ and subset A of $I \times \mathbb{Z}$,

$$A^{\to t} = \{(j, s+t) \text{ for } (j,s) \in A\},$$

with the convention that $\emptyset^{\to t} = \emptyset$, we can define the random neighborhood $K_{i,t}$ of site (i,t) as

$$K_{i,t} = V_{i,t}^{\to t},$$

where $V_{i,t}$ is drawn independently of anything else according to λ_i. We can proceed independently for all sites (j,s) and obtain $K_{j,s} = V_{j,s}^{\to s}$.

By looking recursively at the neighborhoods of the neighborhoods, we build a whole genealogy in space and time of the site (i,t), that is the list of sites that are really impacting on the variable $X_{i,t}$. This is the second step of our algorithm. Note that the genealogy is obtained by going backward in time and its construction depends only on the realizations of the neighborhoods, i.e. only on the distributions λ_i's.

The study of this space-time genealogy is of utmost importance. Indeed, if the genealogy is almost surely finite, then we can implement the two last steps of our algorithm. In these final steps, given the finite genealogy, we go forward in time and simulate first the variables $X_{j,s}$ in the genealogy by using the transition probabilities $p^{V_{j,s}}(X_{K_{j,s}})$. Once this is done, we can finally simulate the variable $X_{i,t}$ by using the transition probability $p^{V_{i,t}}(X_{K_{i,t}})$. This algorithm gives us a way of constructing a space-time stochastic chain by implementing the steps above to each pair of sites in $I \times \mathbb{Z}$. As we will see, the distribution of this process is stationary and compatible with the dynamics described in Section 5.1

Moreover, the study of the length of the genealogy enables us to cut time into almost independent blocks and therefore to have access to concentration inequalities, this second construction being inspired by Vienne t (1997), Reynaud-Bouret and Roy (2007) or Hansen, Reynaud-Bouret, and Rivoirard (2015b).

Sufficient condition for finite genealogies

For all sites $(i,t) \in I \times \mathbb{Z}$, let us define recursively $A_{i,t}^1 = K_{i,t}$ and for $n \geqslant 1$,

$$A_{i,t}^{n+1} = (\cup_{(j,s)\in A_{i,t}^n} K_{j,s}) \setminus \{A_{i,t}^1 \cup \ldots \cup A_{i,t}^n\},$$

the genealogy stopped after $n+1$ generations.

The complete genealogy is $G_{i,t} = \cup_{n=1}^{\infty} A_{i,t}^n$. It is finite if and only if

$$N_{i,t} = \inf\{n \geqslant 1 : A_{i,t}^n = \emptyset\},$$

is finite.

This is a consequence of the following property.

Assumption 5.11. For each $i \in I$, let

$$\bar{m}_i = \sum_{S \in \mathcal{V}} |S| \lambda_i(S). \tag{5.11}$$

We assume that

$$\bar{m} = \sup_{i \in I} \bar{m}_i < 1. \tag{5.12}$$

Probabilistic sparsity corresponds here to the fact that the mean size of the random neighborhoods $\bar{m}_i$ are strictly less than 1.

Thanks to this assumption, we can prove the following result.

Proposition 5.12. *For each $i \in I$, $t \in \mathbb{Z}$ and $\ell \in \mathbb{Z}_+$,*

$$\mathbb{P}\left(N_{i,t} > \ell\right) \leqslant (\bar{m})^{\ell}.$$

In particular, under Assumption 5.11, for all $i \in I$ and $t \in \mathbb{Z}$,

$$\mathbb{P}(N_{i,t} < \infty) = 1, \tag{5.13}$$

that is all genealogies are finite $\mathbb{P}$-almost surely.

Perfect Simulation Algorithm

Fix a site $(i,t) \in I \times \mathbb{Z}$. We want to simulate $X_{i,t}$.

Under Assumption 5.11, we know the genealogy is finite almost surely and it is possible to build this genealogy recursively without having to generate all the $V_{j,s}$. Once the genealogy is obtained by going backward in time, it is then sufficient to go forward and simulate the $X_{j,s}$'s in the genealogy according to the transitions $p^{V_{j,s}}(X_{\mathbf{K}_{j,s}})$.

More formally, we can use two independent fields of independent uniform random variables on $[0,1]$, $\mathbf{U^1} = (U^1_{i,t})_{i \in I, t \in \mathbb{Z}}$ and $\mathbf{U^2} = (U^2_{i,t})_{i \in I, t \in \mathbb{Z}}$, such that the whole randomness of the construction is encompassed in the field $\mathbf{U^1}$ for the genealogies and in the field $\mathbf{U^2}$ for the forward transitions and such that conditionally on these two fields, the whole simulation algorithm is deterministic. But in practice, we generate $U^1_{j,s}$ and $U^2_{j,s}$ only if we need it. This leads to the following algorithm

Step 1. Generate $U_{i,t}^1$ random uniform variable on $[0,1]$. Since $\mathcal{V}$ is countable, one can order its elements such that $\mathcal{V} = \{S_1, ..., S_n, ...\}$. Define the c.d.f. of λ_i by $F_i(0) = \lambda_i(\emptyset)$ and for $n \geqslant 1$,

$$F_i(n) = \lambda_i(\emptyset) + \sum_{k=1}^{n} \lambda_i(S_k)$$

and pick the random neighborhood of (i,t) as $K_{i,t} = V_{i,t}^{\rightarrow t}$ with

$$V_{i,t} = \begin{cases} \emptyset, \text{ if } U_{i,t}^1 \leqslant F_i(0), \\ S_n, \text{ if } F_i(n-1) < U_{i,t}^1 \leqslant F_i(n) \text{ for some } n \geqslant 1 \end{cases}.$$

Initialize $A_{i,t}^1 \leftarrow K_{i,t}$.

Step 2. Generate recursively $U_{j,s}^1$ for $j,s \in A_{i,t}^n$, compute the corresponding $V_{j,s}$ and $K_{j,s}$ as in Step 1 and actualize $A_{i,t}^{n+1} \leftarrow \left(\cup_{j,s \in A_{i,t}^n} K_{j,s}\right) \setminus \{A_{i,t}^1 \cup \ldots \cup A_{i,t}^n\}$. After a finite number of steps, $A_{i,t}^n$ is empty and [Step 2.] stops. Let $N_{i,t}$ be the final n of this recursive procedure and the genealogy of (i,t) is given by $G_{i,t} = \cup_{n=1}^{N_{i,t}} A_{i,t}^n$.

Step 3. Note that the (j,s)'s in $A_{i,t}^{N_{i,t}-1}$ have therefore an empty neighborhood. Generate i.i.d. uniform variables $U_{j,s}^2$ for (j,s) in $A_{i,t}^{N_{i,t}-1}$ and define

$$X_{j,s} = 1\{U_{j,s}^2 \leqslant p_j^{\emptyset}\}. \tag{5.14}$$

Step 4. Recursively generate $U_{j,s}^2$ for (j,s) in $A_{i,t}^{\ell}$ recursively from $\ell = N_{i,t} - 2$ to $\ell = 1$ and define

$$X_{j,s} = 1\{U_{j,s}^2 \leqslant p_j^{V_{j,s}}(X_{K_{j,s}})\}, \tag{5.15}$$

In particular arrived at $\ell = 1$, one generates

$$X_{i,t} = 1\{U_{i,t}^2 \leqslant p_j^{V_{i,t}}(X_{K_{i,t}})\}. \tag{5.16}$$

One can show that the algorithm described above not only shows the existence but also the uniqueness of a stationary stochastic chain $\mathbf{X}$ satisfying Assumption 5.2 which is compatible with a given family of transition probabilities

$\{p_i(x) : i \in I, x \in \{0,1\}^{I\times Z-}\}$ satisfying Assumptions 5.4 and 5.11. Such algorithms are called *Perfect simulation algorithms.* We refer the interested reader to Ferrari, Fernandéz, and Galves (2001) for an introduction to perfect simulation of stochastic chains and to Galves and Löcherbach (2013) for a rigorous proof of above result in the GL neuron model.

We conclude this section by noticing that the perfect simulation algorithm to simulate the linear Hawkes process is very simple. Indeed, since any non-empty neighborhood of the space-time decomposition has size 1, the algorithm reduces to a random walk in the past to find the genealogy, a random decision on the state $X_{j,s}$ at the end of the random walk and a forward decision of the other states $X_{j,s}$ which is then completely deterministic and just depends on the sign of $h_{j\to i}(-s)$.

Time length of a genealogy

We are now interested by the time length of a genealogy. Let, for each non-empty subset A of $I \times \mathbb{Z}$,

$$\mathbb{T}(A) = \min\{s \in \mathbb{Z} : (j,s) \in A\}.$$

We are interested by the variable $T_{i,t}$ which is equal to $t - \mathbb{T}(A_{i,t})$ if the genealogy $G_{i,t}$ is non empty and equal to 0 if $G_{i,t}$ is empty. By stationarity its distribution does not depend on t and the behavior of this variable is of course linked to the one of the variables $T(V_j) = -\mathbb{T}(V_j)$ for V_j obeying the distribution λ_j, with the convention that $T(\emptyset) = 0$. We are interested by conditions under which the variable $T_{i,t}$ has a Laplace transform, that is when

$$\theta \mapsto \Psi_i(\theta) = \mathbb{E}(e^{\theta T_{i,t}})$$

is finite for some positive θ. To do so, we are going to assume the following.

Assumption 5.13. There exists a strictly positive θ such that for all i,

$$\varphi_i(\theta) = \sum_{S\in\mathcal{V}} |S| e^{\theta T(S)} \lambda_i(S)$$

is finite and

$$\varphi(\theta) = \sup_{i\in I} \varphi_i(\theta) < 1. \tag{5.17}$$

Theorem 5.14. *Under Assumption 5.13, for all i in I, $\Psi_i(\theta)$ is finite and*

$$\Psi(\theta) = \sup_{i\in I} \Psi_i(\theta) \leqslant \frac{\sup_{i\in I} \lambda_i(\emptyset)}{1-\varphi(\theta)}.$$

Note that if $\varphi_i(\theta)$ is finite for some positive θ, $\lim_{\theta\to 0}\varphi_i(\theta) = \bar{m}_i$. Therefore if Assumption 5.11 is fulfilled, $\lim_{\theta\to 0}\varphi_i(\theta) < 1$ and it is possible to find $\theta > 0$ such that $\varphi_i(\theta) < 1$ as soon as λ_i has a Laplace transform. In this sense, and roughly speaking, Assumption 5.13 is a more stringent condition of *probabilistic sparsity* than Assumption 5.11.

Application on the main examples

Example 5.15 (Chains of infinite order). The space-time decomposition (5.2) implies that

$$\bar{m} = \sum_{\ell=1}^{\infty} \ell\lambda(\underline{\ell}).$$

Thus, the condition (5.12) is satisfied as soon as

$$\sum_{\ell=1}^{\infty} \ell\lambda(\underline{\ell}) < 1$$

and similarly the condition (5.17) is satisfied as soon as

$$\sum_{\ell=1}^{\infty} \ell e^{\theta\ell}\lambda(\underline{\ell}) < 1.$$

Hence both can be verified if λ is sufficiently exponentially decreasing. Typically one can have $\lambda(\underline{\ell}) = e^{-\lambda}\lambda^{\ell}/\ell!$ with $0 < \lambda < 1$ (Poisson distribution on the range) or $\lambda(\underline{\ell}) = (1-p)^{\ell}p$ with $1/2 < p \leqslant 1$ (Geometric distribution on the range).

Example 5.16 (Discrete-time linear Hawkes processes). According to the space-time decomposition (5.5), it follows that for each $i \in I$,

$$m_i = \Sigma_i^+ + \Sigma_i^-.$$

Therefore, the condition (5.12) reduces to

$$\sup_{i\in I}(\Sigma_i^+ + \Sigma_i^-) = \sup_{i\in I}\sum_{j,s}|h_{j\to i}(-s)| < 1.$$

Moreover the condition (5.17) becomes

$$\sup_{i\in I}\sum_{j,s} e^{\theta s}|h_{j\to i}(-s)| < 1.$$

So if for instance we can rewrite $h_{j\to i}(-s) = w_{j\to i}\, g(-s)$ for a fixed function g of mean 1, the condition (5.12) reduces to

$$\sup_{i\in I} \sum_{j\in I} |w_{j\to i}| < 1,$$

and the additional condition (5.17) is fulfilled for a small enough θ as soon as g has finite exponential moment.

Example 5.17 (GL neuron model). In the nonlinear case, it has been proved in Galves and Löcherbach (2013) (cf. inequalities (5.57) and (5.58)) that for each $i \in I$ the following estimates hold:

$$\lambda_i(\emptyset) \leqslant \gamma \sum_{j\in I} |W_{j\to i}| \sum_{s\geqslant 1} g_j(s), \tag{5.18}$$

and for $\ell \geqslant 1$,

$$\lambda_i(S_i(\ell)) \leqslant \gamma \left(\sum_{j\notin S_i(\ell)} |W_{j\to i}| \sum_{s\geqslant 1} g_j(s) + \sum_{j\in S_i(\ell)} |W_{j\to i}| \sum_{s\geqslant \ell} g_j(s) \right). \tag{5.19}$$

Therefore, a sufficient condition for Assumption 5.11 to hold (cf. inequality (2.9) of Galves and Löcherbach (ibid.)) is

$$\sup_{i\in I} \sum_{\ell\geqslant 1} \ell |S_i(\ell)| \left(\sum_{j\notin S_i(\ell)} |W_{j\to i}| \sum_{s\geqslant 1} g_j(s) + \sum_{j\in S_i(\ell)} |W_{j\to i}| \sum_{s\geqslant \ell} g_j(s) \right) < \frac{1}{\gamma}.$$

In the linear case (i.e. when $\varphi_i(u) = \nu_i + u$), the condition above reduces to

$$\sup_{i\in I} \sum_{\ell\geqslant 1} \ell |S_i(\ell)| \left(\sum_{j\notin S_i(\ell)} |W_{j\to i}| \sum_{s\geqslant 1} g_j(s) + \sum_{j\in S_i(\ell)} |W_{j\to i}| \sum_{s\geqslant \ell} g_j(s) \right) < 1. \tag{5.20}$$

Using the decomposition (5.8), one can verify that the condition (5.12) is, in the linear case, equivalent to

$$\sup_{i\in I} \sum_{\ell\geqslant 1} \left[\ell |W_{i\to i}| g_i(\ell) + \sum_{j\neq i, j\in I} (\ell+1) |W_{j\to i}| g_j(\ell) \right] < 1. \tag{5.21}$$

Note that condition (5.20) is usually much stronger than condition (5.21) and that a sparse space-time decomposition of the process allows us to derive existence of the linear process on a larger set of possible choices for $w_{j\to i}$ and g_j. Once again condition (5.17) is fulfilled under a very similar expression

$$\sup_{i\in I}\sum_{\ell\geqslant 1} e^{\theta\ell}\left[\ell|W_{i\to i}|g_i(\ell)+\sum_{j\neq i, j\in I}(\ell+1)|W_{j\to i}|g_j(\ell)\right]<1,$$

this can be easily fulfilled if $g_j(\ell) = g(\ell)$ is exponentially decreasing with $\sum_{\ell=1}^{\infty}(\ell+1)g(\ell)=1$. Indeed (5.21) is implied as in the Hawkes case by

$$\sup_{i\in I}\sum_{j\in I}|W_{j\to i}|<1$$

and it is easy to find by continuity a small $\theta>0$ such that (5.17) is fulfilled too.

5.3 Concentration inequalities

Block construction

Thanks to the control of the time length genealogy it is possible to cut the observations $X_{F,-(m-1):T}$ into (overlapping) blocks that form with high probability two families of independent variables. This is a key tool to derive concentration inequalities. This construction is inspired by Viennet (1997), who used as a central element, Berbee's lemma, which is replaced here by Theorem 5.14. Note that similar coupling arguments have been used in continuous and more restrictive settings (see Hansen, Reynaud-Bouret, and Rivoirard (2015b) and Reynaud-Bouret and Roy (2007) for linear Hawkes processes, Chen et al. (2019) for bounded Hawkes process and mixing arguments).

Lemma 5.18. *Let $m, T\in\mathbb{Z}_+$ such that $m\leqslant\lfloor T/2\rfloor$ and let $F\subset I$ be a finite subset. Let also B, the grid size, be an integer such that*

$$m\leqslant B\leqslant\lfloor T/2\rfloor,$$

and define $k=\lfloor\frac{T}{2B}\rfloor$. Let the $2k+1$ blocks be defined by, for $1\leqslant n\leqslant 2k$,

$$I_n=\{(n-1)B+1-m,\dots nB\} \text{ and } I_{2k+1}=\{2kB+1-m,\dots T\}.$$

There exist on a common probability space stochastic chains $\mathbf{X}$, $\mathbf{X^1}$,...,$\mathbf{X^{2k+1}}$ satisfying the following properties:

1. *All the chains* $\mathbf{X^n} = (X^n_{i,t})_{i\in I, t\in\mathbb{Z}}$ *have the same distribution as* $\mathbf{X}$ *which satisfies Assumption 5.2 and whose family* $\{p_i(x) : i \in I, x \in \{0,1\}^{I\times\mathbb{Z}-}\}$ *of transition probabilities satisfies Assumptions 5.4 and 5.13 for a given* $\theta > 0$.
2. *The odd chains* $\mathbf{X^1}, \mathbf{X^3}, ..., \mathbf{X^{2k+1}}$ *are independent.*
3. *The even chains* $\mathbf{X^2}, ..., \mathbf{X^{2k}}$ *are independent.*
4. *There exists an event,* Ω_{good}*, such that on* Ω_{good}*,* $X_{F,I_n} = X^n_{F,I_n}$ *for all* $n = 1, ..., 2k+1$ *and such that the probability of* Ω^c_{good}*, under the notation of Theorem 5.14, is at most*

$$|F|\,(2k+1)\,\frac{\Psi(\theta)}{(1-e^{-\theta})}e^{-\theta(B+1-m)}. \tag{5.22}$$

In particular, by choosing $B = m + \theta^{-1}(2\log(T) + \log(|F|))$*, we obtain that there exists a positive* $c'(\theta)$ *such that the probability of* Ω^c_{good} *is at most* $c'(\theta)T^{-1}$.

Applications

As an application of Lemma 5.18, we can derive the following Hoeffding type concentration inequality.

Theorem 5.19. *Let* $\mathbf{X} = (X_{i,t})_{i\in I, t\in\mathbb{Z}}$ *be a stationary sparse space-time process satisfying Assumptions 5.2, 5.4 and 5.13 for a given* $\theta > 0$*. For* $F \subset I$ *finite,* $m \in \mathbb{Z}_+$*, let* f *be a real-valued function of* $X_{F,t-m:t-1}$ *bounded by* M*. Let* $T \in \mathbb{Z}_+$ *such that*

$$m + \theta^{-1}(2\log(T) + \log(|F|)) \leqslant \lfloor T/2 \rfloor$$

and

$$Z(f) = \frac{1}{T}\sum_{t=1}^{T}\left(f(X_{F,t-m:t-1}) - \mathbb{E}\left[f(X_{F,t-m:t-1})\right]\right). \tag{5.23}$$

Then there exists nonnegative constant c', c", *which only depends on* θ *such that, for any* $x > 0$,

$$\mathbb{P}\left(Z(f) > \sqrt{c"(\theta)M^2\frac{m+\log T+\log|F|}{T}x}\right) \leqslant \frac{c'(\theta)}{T} + 2e^{-x}. \tag{5.24}$$

If there is a finite family $\mathcal{F}$ of such f, we also have that

$$\mathbb{P}\left(\exists f \in \mathcal{F}, Z(f) > \sqrt{c''(\theta) M^2 \frac{m + \log T + \log|F|}{T} x}\right) \leqslant \frac{c'(\theta)}{T} + 2|\mathcal{F}|e^{-x}.$$

There is a matrix counterpart to the previous inequality, which is an application of now classical results on random matrices (see Tropp (2012) and the references therein).

Theorem 5.20. *Let $\mathbf{X} = (X_{i,t})_{i\in I, t\in\mathbb{Z}}$ be a stationary sparse space-time process satisfying Assumptions 5.2, 5.4 and 5.13 for a given $\theta > 0$. For $F \subset I$ finite, $m \in \mathbb{Z}_+$, let $\mathcal{F}$ be a finite family of bounded real-valued functions of $X_{F,t-m:t-1}$ and denote $M = \max\{\|fg\|_\infty : f, g \in \mathcal{F}\}$. Let $T \in \mathbb{Z}_+$ such that*

$$m + \theta^{-1}(2\log(T) + \log(|F|)) \leqslant \lfloor T/2 \rfloor$$

and define the random matrix $Z = (Z(f,g))_{f,g\in\mathcal{F}}$ where for each $f, g \in \mathcal{F}$,

$$Z(f,g) = \frac{1}{T}\sum_{t=1}^{T}\big(f(X_{F,t-m:t-1})g(X_{F,t-m:t-1}) - \mathbb{E}\left[f(X_{F,t-m:t-1})g(X_{F,t-m:t-1})\right]\big). \quad (5.25)$$

Then there exists nonnegative constant c', c'', which only depends on θ such that, for any $x > 0$,

$$\mathbb{P}\left(\|Z\| > \sqrt{c''(\theta) M^4 |\mathcal{F}|^2 \frac{m + \log T + \log|F|}{T} x}\right) \leqslant \frac{c'(\theta)}{T} + 4|\mathcal{F}|e^{-x}, \quad (5.26)$$

where $\|Z\|$ corresponds to the spectral norm, that is the largest eigenvalue of the self-adjoint matrix Z.

5.4 On LASSO for sparse space-time systems

In this section we will address the statistical question of how to approximate transition probabilities of sparse space-time systems by linear combinations of a given dictionary. One possible approach to solve this question consists in choosing the

linear combination whose coefficients are selected via LASSO[1]. The main goal of this chapter is to discuss the theoretical guarantees of this approach. Under assumptions on the Gram matrix of the dictionary, we will show that the linear combination corresponding to the coefficients selected via LASSO performs almost as best as possible. Such results as called *Oracle*[2] inequalities. As an application of the concentration inequalities developed in Section 5.3, we show that the required assumptions on the Gram matrix hold with high probability for several examples of dictionaries. These results hold even if the system is only partially observed, making the methodology presented here suitable for applications in network of spiking neurons.

Although the results presented here could be applied more broadly to other settings, we will assume that the stationary space-time system $\mathbf{X} = (X_{i\in I, t\in\mathbb{Z}})$ models a network of spiking neurons. In this case, the countable (possibly infinite) set I represents the set of neurons in the network and

$$X_{i,t} = \begin{cases} 1, & \text{if neuron i spikes at time t,} \\ 0, & \text{otherwise} \end{cases} .$$

Recall that one of the basic assumptions (Assumption 5.2) on the space-time systems studied in this chapter is that the components of X_{t+1} (neurons in our current framework) are conditionally independent given the past $X_{-\infty:t}$, for each $t \in \mathbb{Z}$. In this case, as already mentioned in Section 5.1 , the dynamics of network $\mathbf{X}$ is completed characterized by the family of transition probabilities

$$p_i(x) = \mathbb{P}(X_{i,0} = 1 | X_{-\infty:-1} = x),\ x \in \{0,1\}^{I\times\mathbb{Z}_-},\ i \in I.$$

These transition probabilities are measurable functions of $x \in \{0,1\}^{I\times\mathbb{Z}_-}$.

For a finite $F \subset I$, subset of observed neurons, and integers $T > m \geqslant 1$ measuring the observation window, the aim is to estimate $x \mapsto p_i(x)$ for a fixed neuron $i \in F$, given the sample $x_{F,-(m-1):T}$. To that end, for each time $1 \leqslant t \leqslant T$, we compare the past $x_{F,(t-m):(t-1)}$ to the current observation $x_{i,t}$ to guess

[1]LASSO is an acronym for *Least Absolute Shrinkage and Selector Operator*. Proposed by Tibshirani in 1986, the LASSO is a very popular regularization method for high dimensional statistical settings. This popularity is due in part to its low computational cost. We refer the interested reader to Bühlmann and van de Geer (2011) for a comprehensive mathematical treatment of the LASSO and some its variants.

[2]In our setting, the Oracle corresponds to the best linear combination of elements of the fixed dictionary. The name oracle comes from the fact that it cannot be computed without knowledge of the unknown transition probabilities. This fact will be clarified in Section 5.4.2

what can be a good approximation of $p_i(x)$. The intuition behind this strategy is that for a well-chosen space-time neighborhood $S \subset I \times \mathbb{Z}_-$, it might be sufficient to know x_S and not the whole past configuration x to well approximate $p_i(x)$.

Given the sample $x_{F,-(m-1):T}$, one might consider several candidates to approximate $p_i(x)$. Here, we shall approximate $p_i(x)$ by linear combinations of a given dictionary Φ, i.e. a finite set of real-valued functions on $\{0,1\}^{I\times\mathbb{Z}_-}$ which are cylindrical in $F \times \underline{m}$ with $\underline{m} = \{-m, \dots, -1\}$. Let us assume that the size of the dictionary is $M \geqslant 1$ and denote $\Phi = \{\varphi_1, \dots, \varphi_M\}$. With this notation, for each vector $a = (a_1, \dots, a_M) \in \mathbb{R}^M$, we denote

$$x \mapsto f_a(x) = \sum_{j=1}^{M} a_j \varphi_j(x), \tag{5.27}$$

the candidate encoded by the vector a that should approximate $p_i(x)$. We assume that the functions in the dictionary are bounded in sup norm by $\|\Phi\|_\infty$, that is, $\max_{1\leqslant j\leqslant M} \|\varphi_j\|_\infty = \|\Phi\|_\infty < \infty$, where for each real-valued function φ on $\{0,1\}^{T\times\mathbb{Z}}$, we write $\|\varphi\|_\infty = \sup_{x\in\{0,1\}^{I\times\mathbb{Z}}} |\varphi(x)|$. to denote its sup norm.

For each candidate f_a with $a \in \mathbb{R}^M$, we compute its the least-squares contrast, given by

$$C(f_a) = -\frac{2}{T}\sum_{t=1}^{T} f_a(X_{F,(t-m):(t-1)})x_{i,t} + \frac{1}{T}\sum_{t=1}^{T} f_a^2(x_{F,(t-m):(t-1)}).$$

The least-square contrast can be interpreted as a data-fidelity term. Notice that, if for $j, k \in [M] = \{1, \dots, M\}$, we write

$$b_j = \frac{1}{T}\sum_{t=1}^{T} \varphi_j(x_{F,(t-m):(t-1)})x_{i,t}, \tag{5.28}$$

and

$$G_{jk} = \frac{1}{T}\sum_{t=1}^{T} \varphi_j(x_{F,(t-m):(t-1)})\varphi_k(x_{F,(t-m):(t-1)}), \tag{5.29}$$

then $C(f_a)$ can be rewritten as

$$-2a^\intercal b + a^\intercal G a,$$

where $b = (b_1, \dots, b_M)$ is a vector of $\mathbb{R}^{\Phi}$, $G = (G_{jk})_{j,k\in[M]}$ is the *Gram matrix* of the dictionary Φ and $a^\intercal$ is the transpose of vector $a \in \mathbb{R}^M$.

In the sequel, each vector $a \in \mathbb{R}^M$, let $|a| = (|a_1|, \dots, |a_M|)$, $|a|_\infty = \max_{j\in[M]} |a_j|$, $\|a\|_2 = \sqrt{a^\intercal a}$ and $\|a\|_1 = \mathbf{1}^\intercal |a|$ where $\mathbf{1}$ is the vector with all coordinates equal to 1.

To select the relevant coefficients, we minimize the least-square contrast $C(f_a)$ subject to a ℓ_1-penalization on the vector $a \in \mathbb{R}^M$ indexing the candidate functions f_a. Precisely, we choose the function $\hat{f} = f_{\hat{a}}$ where

$$\hat{a} \in \arg\min_{a\in\mathbb{R}^M} \{-2a^\intercal b + a^\intercal G a + \gamma d \|a\|_1\}, \tag{5.30}$$

[3]

for d a positive term controlling the random fluctuations and $\gamma > 0$, a tuning parameter. In (5.30) above, the vector $b \in \mathbb{R}^M$ and the matrix $G \in R^{M\times M}$ are defined in (5.28) and (5.29) respectively. The minimization problem (5.30) is called *LASSO program*.

The active set $S(a)$ of a vector $a \in \mathbb{R}^M$ is the set $S(a) = \{j \in [M] : a_j \neq 0\}$. We shall denote for any subset $J \subset [M]$ and any $a \in \mathbb{R}^M$, $a_J \in \mathbb{R}^M$ the vector whose coordinates in J are equal to the ones of a and 0 anywhere else. We also denote by $|J|$ the cardinality of J.

5.4.1 Examples of dictionaries

Let us present briefly some examples of dictionaries that might be useful.

Example 5.21 (Short memory effect). Let the dictionary Φ be defined by the set $\{\varphi_j : j \in F\}$, where

$$\varphi_j(x) = \begin{cases} 1, \text{ if } x_{j,s} = 1 \text{ for some } -m \leqslant s \leqslant -1 \\ 0, \text{ otherwise} \end{cases}, \ x \in \{0,1\}^{I\times\mathbb{Z}_-}.$$

With this dictionary, we are trying to explain the presence of a spike on neuron i at time t by a linear combination of the presence of a spike on neuron j in a small window just before time t.

[3] Notice that the multivariate function $\mathbb{R}^M \ni a \mapsto -2a^\intercal b + a^\intercal G a + \gamma d \|a\|_1$ is convex. By using techniques from convex analysis and optimization theory, one can propose efficiently algorithms (e.g. accelerated gradient descent algorithm) to solutions of the convex minimization problem (5.30). We refer the reader to Bubeck (2015) for more details on how to compute efficiently solutions of convex minimization problems.

Example 5.22 (Cumulative effect). We can also think that $m = \eta L$ is a much larger parameter and cut the past $\underline{m}$ into L small pieces of length η, where the effect of the spikes are different and cumulative. This leads to the dictionary Φ defined by the set $\{\varphi_{j,\ell} : j \in F \text{ and } 1 \leqslant \ell \leqslant L\}$ where

$$\varphi_{j,\ell}(x) = \sum_{s=-\eta\ell}^{-\eta(\ell-1)-1} x_{j,s}, \ x \in \{0,1\}^{I\times\mathbb{Z}_-}.$$

Example 5.23 (Cumulative effect with spontaneous apparition). It can be important to take into account a background activity, especially to explain the apparition of spikes due to the unobserved part of the network. To do so, we may add to the previous dictionary an extra function

$$\varphi_0 = 1,$$

whose corresponding coefficient corresponds to a spontaneous activity.

Example 5.24 (Hawkes dictionary). In both the cumulative effect and the cumulative effect with spontaneous part, one might be interested in a particular example where $\eta = 1$ and $L = m$. In particular, in the case with spontaneous part, we are therefore interested in approximating $p_i(x)$ by

$$f_a(x) = a_0 + \sum_{j\in F} \sum_{-m\leqslant s\leqslant -1} a_{j,-s} x_{j,s}, \ x \in \{0,1\}^{I\times\mathbb{Z}_-},$$

which is the exact form of a discrete Hawkes process restricted to $F \times \underline{m}$ (see Example 5.8) and this even if p_i is not of this shape.

Note that whatever the dictionary, m represents the maximal delay in the dictionary and $|F|$ the number of observed neurons. As we will see in Section 5.5, both these quantities have to be usually less than a certain increasing function of T, which depends on the dictionary (typically $\log(T)$), to derive an restricted eigenvalue property on the Gram matrix. In particular $|m|$ might grow slightly with T to ensure a good asymptotic approximation of the dependency in time. Mathematically speaking, the same holds for $|F|$, although the size of F is dictated by the neurophysiological experiment and, for practical purpose it is always thought to be a constant with respect to T.

5.4.2 Oracle inequality

It is classical, by now, to derive oracle inequalities for Lasso procedures if G satisfies some properties. We use two of them.

Definition 5.25. Let $\kappa > 0$. The matrix G satisfies Property **Inv**(κ) if

$$\forall a \in \mathbb{R}^M, \quad a^\intercal G a \geqslant \kappa \|a\|_2^2.$$

A weaker version is the restricted eigenvalue condition.

Definition 5.26. Let $c > 0$, $\kappa > 0$ and $s \in \mathbb{N}$. The matrix G satisfies Property **RE**(κ, c, s) if for all subset $J \subset [M]$ such that $|J| \leqslant s$ and for all $a \in \mathbb{R}^M$ such that

$$\|a_{J^c}\|_1 \leqslant c\|a_J\|_1,$$

the following holds

$$a^\intercal G a \geqslant \kappa \|a_J\|_2^2.$$

Notice that **Inv**(κ) corresponds to **RE**(κ, ∞, M) (see Exercise 5.6). Our first result establishes an oracle inequality for the estimator $\hat{f} = f_{\hat{a}}$ where $\hat{a}$ is defined by (5.30).

In the sequel, let us write for each $j \in [M]$,

$$\bar{b}_j = \frac{1}{T} \sum_{t=0}^{T-1} \varphi_j(x_{F,t-m:t}) p_i(x_{-\infty:t}). \tag{5.31}$$

Moreover, for real-valued functions f and g on $\{0,1\}^{I \times \mathbb{Z}_-}$, let us denote

$$\langle f, g \rangle_T = \frac{1}{T} \sum_{t=0}^{T-1} f(x_{-\infty:t}) g(x_{-\infty:t})^4, \tag{5.32}$$

and $\|f\|_T = \sqrt{\langle f, f \rangle_T}$ the corresponding norm.

Theorem 5.27. *Let $\gamma \geqslant 2$, $\kappa > 0$ and $s \in \mathbb{N}$. On the event on which*

[4] Note that $\langle f, g \rangle_T$ can be computed from the sample $x_{F,-(m-1):T}$ only for functions cylindrical in $F \times \underline{m}$, which is the case of all functions in the dictionary Φ. We measure the performance of $\hat{f}$ in terms of distance $\|\hat{f} - p_i(\cdot)\|_T$. Notice that if $\langle f, g \rangle_T$ were directly defined for functions cylindrical only, the distance $\|\hat{f} - p_i(\cdot)\|_T$ would not be well-defined since the function $p_i(\cdot)$ is not necessarily cylindrical.

(i) $|b_j - \bar{b}_j| \leqslant d$ *for all* $j \in [M]$, *and*

(ii) G *satisfies* $\mathbf{RE}(\kappa, c(\gamma), s)$ *with* $c(\gamma) = \frac{\gamma+2}{\gamma-2}$,

the following inequality is satisfied

$$\|\hat{f} - p_i(\cdot)\|_T^2 \leqslant \inf_{a \in \mathbb{R}^M : |S(a)| \leqslant s} \left\{ \|f_a - p_i(\cdot)\|_T^2 + \kappa^{-1}|S(a)|d^2 \frac{(\gamma+2)^2}{4} \right\}. \tag{5.33}$$

Moreover for any $0 < \delta < 1$, *if* $d = d_\delta$ *with*

$$d_\delta = \sqrt{\|\Phi\|_\infty^2 \frac{\log(M) + \log(2\delta^{-1})}{2T}},$$

then $\mathbb{P}(\exists \varphi \in [M] : |b_j - \bar{b}_j| > d) \leqslant \delta$.

Equation (5.33) is a classical oracle inequality (see for instance Hansen, Reynaud-Bouret, and Rivoirard (2015b) or Hunt et al. (2019) for close setups). This result means that the Lasso estimator gives the best s-sparse approximation of p_i based on the dictionary Φ and that the price to pay is of the order of $\kappa^{-1}sd^2$, if we assume that $\|\Phi\|_\infty \leqslant 1$. With the choice $d = d_\delta$, we have therefore a price of the order of $\kappa^{-1}s\frac{\log(M)+\log(2\delta^{-1})}{T}$. Note that if we knew that p_i can be indeed decomposed on Φ, meaning that the model is true and that in particular p_i only depends on s elements of the dictionary Φ, the price to pay to estimate p_i would be roughly of the order of s/T. The logarithmic factor is a classical loss for adaptation in (5.33). Therefore, it remains to see the order of κ, to see if (5.33) gives roughly the best possible rate.

Note that if G satisfies $\mathbf{Inv}(\kappa)$ then one can choose $\gamma = 2$ and $s = |\Phi|$ in Theorem 5.27 and (5.33) can be rewritten as

$$\|\hat{f} - p_i(\cdot)\|_T^2 \leqslant \inf_{a \in \mathbb{R}^\Phi} \left\{ \|f_a - p_i(\cdot)\|_T^2 + 4\kappa^{-1}|S(a)|d^2 \right\},$$

which is a sharper version of the result proved in Hansen, Reynaud-Bouret, and Rivoirard (2015b) in continuous time, up to the fact that they used more general weights which leads to a weighted ℓ_1 norm in the criterion. The same refinement would have been possible but since the focus is here on the Gram matrix, we have decided to use a classical ℓ_1 norm for sake of simplicity.

Note also that another (very easy) refinement consists in clipping $\hat{f}$ to ensure that it remains between 0 and 1. The same result holds for this clipped version.

Another way to find similar results for estimators that are forced to be in $[0, 1]$ is to use penalized maximum likelihood. Many works have used it (see for instance Mark, Raskutti, and Willett (2019) for Poisson counts or Basu and Michailidis (2015) and Gaïffas and Matulewicz (2019) in Gaussian Markovian setups). This comes with additional technicalities, in particular if the likelihood of the statistical model is not easy to compute, because the model is not Gaussian. In particular, Mark, Raskutti, and Willett (n.d.) use a setting very close to ours, but simpler and make use of Taylor expansion to approximate the criteria. Translated here, the approximation would depend on the dictionary we use and would be more complex for each dictionary. Once again, because the focus is here on the Gram matrix, we have decided to stick with the simplest Lasso result made for least-squares contrast.

Results for controlling Gram matrices are numerous (see for instance Basu and Michailidis (2015), Gaïffas and Matulewicz (2019), and Hunt et al. (2019) for simpler settings than the present one) but always assume that the whole network is observed and that *the target can be written on the dictionary*. In Hansen, Reynaud-Bouret, and Rivoirard (2015b), which is the closest framework to the present one, it has been proved for instance that, if one observes the whole finite network and if the spike trains are linear Hawkes processes, then G is invertible with large probability for well chosen dictionaries. In this case, the corresponding κ is roughly lower bounded by a quantity which is exponentially small in the total number of neurons in the network. Here we would like to go beyond these assumptions and prove that even if

- the model is wrong (i.e. p_i is not Hawkes for instance) and p_i cannot be written on the dictionary,
- the network is infinite,
- we only observe a very partial subnetwork,

it is still possible to find good κ with high probability and that the dependency in the number of neurons can be much better than these previous results.

The main idea consists in using very general Kalikow-type decomposition of the transition probability $p_i(x)$, that are available in discrete time (as discussed in Chapter 5) and that do not exist with such generality in continuous time (see however Hodara and Löcherbach (2017c) for promising results in this direction).

5.5 Back to the Gram matrices

To control the Gram matrix we need also that the following condition holds.

Assumption 5.28. There exists some positive μ such that for all $i \in I$, for all x,

$$\mu \leqslant p_i(x) \leqslant 1 - \mu,$$

Note that in each of the examples (strongly non-null chains of infinite order, discrete time Hawkes processes and GL neuron model), this assumption is easily fulfilled. For instance in the Hawkes case, this adds the condition $\mu \leqslant \nu_i - \Sigma_i^- \leqslant \nu_i + \Sigma_i^+ \leqslant (1-\mu)$ (see Example 5.8 to recall the notation.)

This assumption is useful to bound expectation by changing the underlying measure.

Lemma 5.29. *Under Assumptions 5.2 and 5.28, for all non negative function f cylindrical on a fixed finite space-time neighborhood $S \subset I \times \mathbb{Z}_-$,*

$$(2(1-\mu))^{|S|}\mathbb{E}^{\otimes\mathcal{V}}_{\mathcal{B}(1/2)}\left[f(X_S)\right] \geqslant \mathbb{E}\left[f(X_S)\right] \geqslant (2\mu)^{|S|}\mathbb{E}^{\otimes\mathcal{V}}_{\mathcal{B}(1/2)}\left[f(X_S)\right],$$

where $\mathbb{E}^{\otimes\mathcal{V}}_{\mathcal{B}(1/2)}$ means that the expectation is taken with respect to the measure where all $X_{i,t}$'s are i.i.d Bernoulli with parameter $1/2$.

5.5.1 Inv property for general dictionaries

In this section we prove that the $\mathrm{Inv}(\kappa)$ property holds on an event with high probability for the examples of dictionaries considered in Section 5.4.1. As a by product, we are able to derive oracle inequalities with high probability for these dictionaries. We start with the following result.

Theorem 5.30. *For a finite $F \subset I$ and integer $T > m \geqslant 1$, let $X_{F,-(m-1):T}$ be a sample produced by the stationary sparse space-time process $\mathbf{X} = (X_{i,t})_{i\in I, t\in\mathbb{Z}}$ satisfying Assumptions 5.2, 5.4 and 5.13. Let Φ denote a finite dictionary of bounded functions cylindrical in $F \times \underline{m}$ and G be the corresponding Gram matrix defined in (5.29). If the matrix $\mathbb{E}(G)$ satisfies property $Inv(\kappa')$ for some positive constant κ', then for any $\delta > 0$ and T sufficiently large, the Gram matrix G satisfies the property $Inv(\kappa)$ on an event of probability larger than $1 - \frac{c_1}{T} - \delta$ with*

$$\kappa = \kappa' - c_2 M \|\Phi\|_\infty^2 \sqrt{\frac{(m + \log(T) + \log|F|)(\log(M) + \log(\delta^{-1}))}{T}},$$

where c_1 and c_2 are positive constants which only depends on the underlying distribution of $\mathbf{X}$ and M is the size of the dictionary Φ.

To apply Theorem 5.30 to the dictionaries considered in Section 5.4.1 we must find the corresponding κ'. This is done below.

Example 5.31 (Short memory effect). To apply Theorem 5.30 we need first to find κ' for this class of models. This is done as follows. Let $Q = \mathcal{B}(1/2)^{\otimes \mathcal{V}}$ be the probability measure under which all $X_{i,t}$'s are i.i.d. Bernoulli with parameter $1/2$ and denote $p_j = Q(\varphi_j(X_{-\infty:-1}) = 1)$ for $j \in F$. Clearly, $p_j = 1 - (1/2)^m$ for all $j \in F$ and we write p to denote this common value. With this notation, one can check that (see Exercise 5.7),

$$\mathbb{E}^{\otimes \mathcal{V}}_{\mathcal{B}(1/2)}(G) = \begin{pmatrix} p & p^2 & p^2 & \dots & p^2 \\ p^2 & p & p^2 & \dots & p^2 \\ & & & \dots & \\ p^2 & p^2 & p^2 & \dots & p \end{pmatrix}. \tag{5.34}$$

Such a matrix has only two eigenvalues, namely, $p + (|F| - 1)p^2$ of multiplicity 1 and $p - p^2 = (1/2)^m(1 - (1/2)^m)$ with multiplicity $|F| - 1$. Indeed, ξ is an eigenvalue $\mathbb{E}^{\otimes \mathcal{V}}_{\mathcal{B}(1/2)}(G)$ if and only if there exists a non-null vector $u \in \mathbb{R}^F$ such that

$$(p - p^2)u + p^2 \sum_i u_i \mathbf{1} = \xi u.$$

On the one hand, by choosing the vector $u \neq 0$ such that $\sum_i u_i = 0$ gives that $\eta = p - p^2$ is an eigenvalue with multiplicity $|F| - 1$. On the other hand, the choice $\sum_i u_i = 1$ forces that $(p - p^2)u_i + p^2 = \xi u_i$ for all $i \in F$, ensuring that $\xi = p + p^2(|F| - 1)$ is the second eigenvalue. Its multiplicity is necessarily 1.

Note that if m is large, the smallest eigenvalue of $\mathbb{E}^{\otimes \mathcal{V}}_{\mathcal{B}(1/2)}(G)$ is really small. This can be interpreted in the following way : when m is large, one will find a "1" on every observed neuron in the past, therefore all the φ_j's will be equal with high probability and one cannot infer a dependence graph with this dictionary anymore.

Thus, Lemma 5.29 implies that eigenvalue of $\mathbb{E}(G)$ can be lower bounded by

$$\kappa' = (2\mu)^{m|F|}(1/2)^m(1 - (1/2)^m). \tag{5.35}$$

Choosing for a fixed integer η

$$m = \eta \text{ and } |F| \leqslant \log\log(T), \tag{5.36}$$

gives κ' of the order $(\log(T))^{-c_3}$ for some constant $c_3 > 0$ depending on μ and η.

Example 5.32 (Cumulative effect). In this case, let α denote the common value of $\mathbb{E}^{\otimes_{\mathcal{V}}}_{\mathcal{B}(1/2)}(\varphi^2_{j,\ell}(X_{-\infty:-1}))$ with $j \in F$ and $1 \leqslant \ell \leqslant L$, and β be the corresponding value of $\mathbb{E}^{\otimes_{\mathcal{V}}}_{\mathcal{B}(1/2)}(\varphi_{j,\ell}(X_{-\infty:-1})(\varphi_{k,n}(X_{-\infty:-1}))$ with $j,k \in F$ and $k \neq j$ and $1 \leqslant n, \ell \leqslant L$. With this notation, one can verify that (see Exercise 5.8)

$$\begin{aligned} \alpha &= \frac{\eta}{2} + \frac{\eta(\eta-1)}{4} = \frac{\eta}{4} + \frac{\eta^2}{4}, \\ \beta &= \frac{\eta^2}{4} \text{ and} \\ \mathbb{E}^{\otimes_{\mathcal{V}}}_{\mathcal{B}(1/2)}(G) &= \begin{pmatrix} \alpha & \beta & \beta & \dots & \beta \\ \beta & \alpha & \beta & \dots & \beta \\ & & & \dots & \\ \beta & \beta & \dots & \beta & \alpha \end{pmatrix}. \end{aligned} \tag{5.37}$$

Hence, the smallest eigenvalue of $\mathbb{E}^{\otimes_{\mathcal{V}}}_{\mathcal{B}(1/2)}(G)$ is $\alpha - \beta = \frac{\eta}{4}$ which grows with $\eta = \frac{m}{K}$. This seems also reasonable since once looking for cumulative effects, the larger the bin size η, the more points you see in it and the more diverse the situations are (hence the dictionary has many different functions) whereas if η is small there is a large probability to see all $\varphi_{j,\ell}$'s null.

Thus, Lemma 5.29 implies that eigenvalue of $\mathbb{E}(G)$ can be lower bounded by

$$\kappa' = \frac{\eta}{4}(2\mu)^{\eta K|F|}.$$

Choosing for some fixed integer η

$$m = \eta K \text{ with } K \leqslant \sqrt{\log\log T} \text{ and } |F| \leqslant \log\log T, \tag{5.38}$$

gives κ' of the order $(\log(T))^{-c_3}$ for some other constant $c_3 > 0$ depending on μ and η.

Example 5.33 (Cumulative effect with spontaneous apparition). With the same notation of the previous example, one can show that (see)

$$\mathbb{E}^{\otimes_{\mathcal{V}}}_{\mathcal{B}(1/2)}(G) = \begin{pmatrix} 1 & \eta/2 & \eta/2 & \dots & \eta/2 \\ \eta/2 & \alpha & \beta & \dots & \beta \\ & & \dots & & \\ \eta/2 & \beta & \dots & \beta & \alpha \end{pmatrix}. \tag{5.39}$$

Reasoning by block with the vector (μ, a) with $\mu \in \mathbb{R}$ and $a \in \mathbb{R}^{K|F|}$, we end up with

$$(\mu,a)^{\intercal}\mathbb{E}^{\otimes_{\mathcal{V}}}_{\mathcal{B}(1/2)}(G)(\mu,a) = \left(\mu + \frac{\eta}{2}\sum_{j\in F,k=1,\dots,K} a_{j,k}\right)^2 + \frac{\eta}{4}\|a\|_2^2.$$

But for all $0 < \theta < 1$,

$$\begin{aligned}\left(\mu + \frac{\eta}{2}\sum a_{j,k}\right)^2 &\geqslant (1-\theta)\mu^2 + \left(1-\frac{1}{\theta}\right)\frac{\eta^2}{4}\left(\sum_{j\in F,k=1,\dots,K} a_{j,k}\right)^2 \\ &\geqslant (1-\theta)\mu^2 - \frac{1-\theta}{\theta}\frac{K|F|\eta^2}{4}\|a\|_2^2,\end{aligned}$$

so that

$$\mathbb{E}^{\otimes_{\mathcal{V}}}_{\mathcal{B}(1/2)}(G) \geqslant (1-\theta)\mu^2 - \frac{1-\theta}{\theta}\frac{K|F|\eta^2}{4}\|a\|_2^2 + \frac{\eta}{4}\|a\|_2^2.$$

By choosing $\theta = \frac{2\eta K|F|}{1+2\eta K|F|}$ we conclude, thanks to Lemma 5.29, that the smallest eigenvalue of $\mathbb{E}(G)$ can be lower bounded by

$$\kappa' = (2\mu)^{\eta K|F|}\min\left(\frac{1}{1+2\eta K|F|}, \frac{\eta}{8}\right).$$

Once again choosing for some fixed integer η

$$m = \eta K \text{ with } K \leqslant \sqrt{\log\log T} \text{ and } |F| \leqslant \log\log T, \tag{5.40}$$

gives κ' roughly larger than $(\log(T))^{-c_3}$ for some other constant $c_3 > 0$ depending on μ and η.

Next, as a by product of Theorem 5.30 and Theorem 5.27, one can derive oracle inequalities for the dictionaries above.

Corollary 5.34. *Let Φ be one of the dictionaries presented in Section 5.4.1, with the choices* (5.36)*,* (5.38) *or* (5.40)*. Assume one observes the sample $x_{F,-(m-1):T}$ generated by the underlying process $\mathbf{X}$ satisfying Assumptions 5.2, 5.4, 5.13 and 5.28. With the notation of Theorem 5.27, for T large enough, on an event with probability $1 - c_1/T$, the following oracle inequality holds*

$$\|\hat{f} - p_i(\cdot)\|_T^2 \leqslant \inf_{a \in \mathbb{R}^\Phi} \left\{ \|f_a - p_i(\cdot)\|_T^2 + c_2 |S(a)| \frac{(\log(T))^{c_3}}{T} \right\},$$

where the constant $c_1 > 0$ depends only on the underlying distribution of $\mathbf{X}$*, $c_2 > 0$ depends on η and γ and constant $c_3 > 0$ depends on both the underlying distribution of* $\mathbf{X}$ *and η.*

Note that the main improvement with respect to Hansen, Reynaud-Bouret, and Rivoirard (2015b), is that in all the examples , the constant κ is roughly of order $(\log(T))^{-c_3}$, that is asymptotically decreasing in roughly speaking the number of neurons used in the dictionary and not the total number of neurons in the network. The number of neurons that are used, which is bounded by the number of observed neurons, can very slowly grow with T.

5.5.2 Hawkes dictionary without spontaneous part

In this case the $\varphi(X_{F,-m:-1})$'s are just the $X_{j,s}$ for $j \in F$ and $s \in \underline{m}$ and one can prove the following result.

Theorem 5.35. *For a finite $F \subset I$ and integer $T > m \geqslant 1$, let $x_{F,-(m-1):T}$ be a sample produced by the stationary sparse space-time process $\mathbf{X} = (X_{i,t})_{i \in I, t \in \mathbb{Z}}$ satisfying Assumptions 5.2, 5.4, 5.13 and 5.28. For the Hawkes dictionary without spontaneous part, i.e. $\varphi = \varphi_{j,s}$ with $\varphi_{j,s}(X_{F,-m:-1}) = X_{j,s}$ for $j \in F$ and $s \in \underline{m}$, the corresponding Gram matrix G defined by (5.29) satisfies for all $c > 0$, $s \leqslant m|F|$ and T large enough, the property* $\mathbf{RE}(\kappa, c, s)$ *on an event of probability larger than $1 - \frac{c'}{T} - \delta$ with*

$$\kappa = \mu - \mu^2 - ((1 - 2\mu) + R_T)(1 + c)s,$$

where

$$R_T = \frac{c_1}{T^{1/2}} (m + \log T + \log |F|)^{1/2} (\log m + \log |F| + \log \delta^{-1})^{1/2},$$

for some positive constant c_1 and c_2 which only depends on the underlying distribution of $\mathbf{X}$*.*

The major point to note is that asymptotically, for slowly growing m and $|F|$ as functions of T, the constant κ does not depend at all on the number of observed

neurons and therefore the rate of convergence in Theorem 5.27 is not worsened by a huge number of observed neurons, $|F|$. This is a drastic improvement with respect to the previous result of Hansen, Reynaud-Bouret, and Rivoirard (2015b) which depends on the total number of neurons in the network. For each fixed c and s, we only need here μ to be close enough to $1/2$ to have $\kappa > 0$.

It also means that the size of the dictionary might be growing with T, much more rapidly than before: typically m the delay might grow like $\log(T)$ and the number of observed neurons might grow like T or even more rapidly as long as $\log|F| = o(T^{1/2})$. Therefore if one can reasonably well approximate p_i by a sparse combination in space and time for which the precise location is unknown, one might by a growing set of observations find the correct set in space and time.

5.6 Proofs of this chapter

Proof of Proposition 5.12. Since $\{N_{(i,t)} > \ell\} = \{|A^\ell_{i,t}| \geqslant 1\}$, the Markov inequality implies that

$$\mathbb{P}\left(N_{i,t} > \ell\right) \leqslant \mathbb{E}\left[|A^\ell_{i,t}|\right].$$

So let us prove by induction that $\mathbb{E}\left[|A^\ell_{i,t}|\right] \leqslant (\bar{m})^\ell$ for all $\ell \geqslant 1$. For $\ell = 1$, we have $\mathbb{E}\left[|A^1_{i,t}|\right] = \mathbb{E}\left[|V_{i,t}|\right] = \bar{m}_i \leqslant \bar{m}$. Next for $\ell > 1$,

$$\begin{aligned}\mathbb{E}\left[|A^\ell_{i,t}| \mid A^{\ell-1}_{i,t}\right] &\leqslant \sum_{(j,s)\in A^{\ell-1}_{i,t}} \mathbb{E}\left[|V^{\rightarrow s}_{j,s}|\right] \\ &\leqslant \sum_{(j,s)\in C_{i,t}(\ell-1)} \bar{m}_j \leqslant |A^{\ell-1}_{i,t}|\,\bar{m}.\end{aligned}$$

To conclude the proof take the overall expectation and use the induction assumption given by $\mathbb{E}\left[|A^{\ell-1}_{i,t}|\right] \leqslant (\bar{m})^{\ell-1}$. □

Proof of Theorem 5.14. For any fixed $n \geqslant 1$, for all site (i,t) let

$$G^n_{i,t} = \cup_{m=1}^n A^m_{i,t}$$

We adopt the convention that if $G^n_{i,t} = \emptyset$, $\mathbb{T}(G^n_{i,t}) = t$ and we consider the variable $T^n_{i,t} = t - \mathbb{T}(G^n_{i,t})$ as well as its Laplace transform $\Psi^n_i(\theta) = \mathbb{E}(e^{\theta T^n_{i,t}})$.

Let us prove by induction that $\Psi_i^n(\theta)$ is finite and that

$$\Psi^n(\theta) = \sup_i \Psi_i^n(\theta) \leqslant \bar{\lambda}(1+\varphi(\theta)+\ldots+\varphi(\theta)^{n-2})\mathbf{1}_{n>1}+\varphi(\theta)^{n-1}g(\theta), \quad (5.41)$$

where $\bar{\lambda} = \sup_{i\in I} \lambda_i(\emptyset)$ and

$$g(\theta) = \sup_{i\in I} \sum_{S\in\mathcal{V}} e^{\theta T(v)}\lambda_i(S).$$

Note that $g(\theta)$ is finite as soon as $\varphi(\theta)$ is and that $0 \leqslant \bar{\lambda} \leqslant 1$.

For $n = 1$, since for all i, $\mathbb{T}(G_{i,t}^1) = \mathbb{T}(A_{i,t}^1) = \mathbb{T}(K_{i,t}) = t - T(V_{i,t})$

$$\begin{aligned}\Psi_i^1(\theta) &= \mathbb{E}\left(\exp\left[\theta T(V_{i,t})\right]\right)\\ &= \sum_{S\in\mathcal{V}} e^{\theta T(S)}\lambda_i(S)\\ &\leqslant g(\theta).\end{aligned}$$

Next by induction, let us assume (5.41) at level n for all i and let us prove it at level $n+1$. Note that because the $G_{i,t}^n$ are computed recursively, we have that when $K_{i,t}$ is not empty,

$$\mathbb{T}(G_{i,t}^{n+1}) = \min_{(k,r)\in K_{i,t}} \mathbb{T}(G_{k,r}^n).$$

Therefore if $K_{i,t} = \emptyset$, $T_{i,t}^{n+1} = 0$ and

$$\mathbb{E}\left(\exp\left[\theta T_{i,t}^{n+1}\right] \mid K_{i,t}\right) = 1.$$

This happens with probability $\lambda_j(\emptyset)$. If $K_{i,t} \neq \emptyset$, then

$$\left(t - \mathbb{T}(G_{i,t}^{n+1}\right) = \max_{(k,r)\in K_{i,t}} \left(t - \mathbb{T}(G_{k,r}^n)\right),$$

and one can check that

$$\begin{aligned}&\mathbb{E}\left(\exp\left[\theta\left(t - \mathbb{T}(G_{i,t}^{n+1})\right)\right] \mid K_{j,t}\right)\\ &\qquad\leqslant \sum_{(k,r)\in K_{i,t}} e^{\theta(t-r)}\mathbb{E}\left(\exp\left[\theta\left(r - \mathbb{T}(G_{k,r}^n)\right)\right] \mid K_{i,t}\right).\end{aligned}$$

Since (see the algorithm) $K_{i,t}$ only depends on $U^1_{j,t}$ and $G^n_{k,r}$ only depends on the $U^1_{k',r'}$ for $k' \in I, r' \leqslant r$ and $r < t$, it follows that $\mathbb{T}(G^n_{k,r})$ is independent of $K_{i,t}$. Hence if $K_{i,t} \neq \emptyset$

$$\begin{aligned}
\mathbb{E}\left(\exp\left[\theta T^{n+1}_{i,t}\right] \mid K_{j,t}\right) &\leqslant \sum_{(k,r)\in K_{i,t}} e^{\theta(t-r)}\Psi^n_k(\theta) \\
&\leqslant \left[\sum_{(k,r)\in K_{i,t}} e^{\theta(t-r)}\right]\Psi^n(\theta) \\
&\leqslant \left[|K_{i,t}|e^{\theta(t-\mathbb{T}(K_{j,t}))}\right]\Psi^n(\theta) \\
&\leqslant |V_{i,t}|e^{\theta T(V_{i,t})}\Psi^n(\theta).
\end{aligned}$$

We obtain by taking the overall expectation that

$$\Psi^{n+1}_i(\theta) \leqslant \bar{\lambda} + \varphi(\theta)\Psi^n(\theta),$$

so that $\sup_{i\in I}\Psi^{n+1}_i(\theta)$ is finite and (5.41) holds at level $n+1$ by induction.

To conclude, it is sufficient to remark that by the monotone convergence theorem, $\Psi^n_i(\theta) \to_{n\to\infty} \Psi_i(\theta)$ which are therefore upper bounded by $\bar{\lambda}/(1-\varphi(\theta))$. This concludes the proof. □

Proof of Lemma 5.18. We use the perfect simulation algorithm to construct these chains. In what follows, let

$$\mathbf{U^0} = (U^{0,1}_{i,t}, U^{0,2}_{i,t})_{i\in I, t\in\mathbb{Z}}, \ldots, \mathbf{U^{2k+1}} = (U^{2k+1,1}_{i,t}, U^{2k+1,2}_{i,t})_{i\in I,t\in\mathbb{Z}}$$

be independent fields of independent random variables with uniform distribution on $[0,1]$. We assume that these sequences are defined in the same probability space and set $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{\mathbb{P}})$ to be this common probability space.

The perfect simulation algorithm performed with the same field $\mathbf{U^0}$ on each site (i,t) yields the construction of $\mathbf{X} = (X_{i,t})_{i\in I,t\in\mathbb{Z}}$.

For any n, the chain $\mathbf{X^n}$ is also built similarly via the perfect simulation algorithm but with the field $\mathbf{U^n}$ except on a small portion of time where we use $\mathbf{U^0}$. More precisely, we use the following variables

$$\left((U^{n,1}_{i,t}, U^{n,2}_{i,t})_{i\in I, t\leqslant(n-2)B}, (U^{0,1}_{i,t}, U^{0,2}_{i,t})_{i\in I,(n-2)B<t\leqslant nB}, (U^{n,1}_{i,t}, U^{n,2}_{i,t})_{i\in I,t>nB}\right),$$

for $1 \leqslant n \leqslant 2k$ and for $n = 2k+1$,

$$\left((U_{i,t}^{n,1}, U_{i,t}^{n,2})_{i\in I, t\leqslant(2k-1)B}, (U_{i,t}^{0,1}, U_{i,t}^{0,2})_{i\in I,(2k-1)B<t\leqslant T}, (U_{i,t}^{n,1}, U_{i,t}^{n,2})_{i\in I, t>T}\right).$$

Since all chains are simulated with the same set of weights $(\lambda_i)_{i\in I}$ and transitions $(p_i^S)_{i\in I, S\in\mathcal{V}}$, they have obviously the same distribution. Since the algorithms use disjoint sets of uniform variables for the odd (resp. even) chains, they are obviously independent and therefore Items 1-3 follows easily from the construction.

Let $G_{i,t}$ be the genealogy of site (i,t) in the chain $\mathbf{X}$ and $\mathbb{T}_{i,t} = \mathbb{T}(G_{i,t})$. For any n, any $i \in F$ and any $t \in I_n$, if $\mathbb{T}_{i,t} > (n-2)B$, then we use exactly the same set of uniform variables to produce the values of $X_{i,t}$ and $X_{i,t}^n$ and their values are equal.

Therefore on $\Omega_{good} = \cap_{i\in F}\cap_{n=1}^{2k+1}\cap_{t\in I_n}\{\mathbb{T}_{i,t} > (n-2)B\}$, $X_{F,I_n} = X_{F,I_n}^n$ for all $n = 1, ..., 2k+1$. Note that Ω_{good} only depends on $\mathbf{X}$.

It remains to control $\widetilde{\mathbb{P}}(\Omega_{good}^c)$. By a union bound, and the application of Theorem 5.14, we obtain

$$\begin{aligned}
\widetilde{\mathbb{P}}(\Omega_{good}^c) &\leqslant \sum_{i\in F}\sum_{n=1}^{2k+1}\sum_{t\in I_n}\mathbb{P}(\mathbb{T}_{i,t} \leqslant (n-2)B) \\
&\leqslant \sum_{i\in F}\sum_{n=1}^{2k+1}\sum_{t\in I_n}\mathbb{P}(t - \mathbb{T}_{i,t} \geqslant t - (n-2)B) \\
&\leqslant \sum_{i\in F}\sum_{n=1}^{2k+1}\sum_{t\in I_n} e^{-\theta(t-(n-2)B)}\Psi(\theta) \\
&\leqslant |F|(2k+1)\frac{e^{-\theta(B-m+1)}}{1-e^{-\theta}}\Psi(\theta).
\end{aligned}$$

In particular if we choose $B = m + \theta^{-1}(2\log(T) + \log(|F|)$,

$$\widetilde{\mathbb{P}}(\Omega_{good}^c) \leqslant \frac{2k+1}{T^2}\frac{\Psi(\theta)}{1-e^{-\theta}},$$

which concludes the proof.

□

Proof of Theorem 5.19. Take $B = m + \theta^{-1}(2\log(T) + \log(|F|))$, $k = \lfloor \frac{T}{2B} \rfloor$ and use the probability space $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{\mathbb{P}})$ and the stochastic chains $\mathbf{X}, \ldots, \mathbf{X^{2k+1}}$ given by Lemma 5.18. By Lemma 5.18-Item 1 we can assume that Z is also defined on $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{\mathbb{P}})$. Define also a partition $J_1, \ldots, J_{2k+1}$ of $1:T$ as follows:

$$J_n = \{1 + (n-1)B, \ldots, nB\} \text{ for } 1 \leqslant n \leqslant 2k, \text{ and } J_{2k+1} = \{1 + 2kB, \ldots, T\}.$$

For each $1 \leqslant n \leqslant 2k+1$, write $S_n = \frac{1}{T} \sum_{t \in J_n} f(X^n_{F,t-m:t-1})$ and note that S_n only depends on the t's in I_n as defined in Lemma 5.18. Since $|J_n| \leqslant B$ for all $1 \leqslant n \leqslant 2k+1$, it holds $|S_n| \leqslant MB/T$.

Observe that Lemma 5.18-Item 1 and 4 ensure that on Ω_{good},

$$Z = \sum_{n=1}^{2k+1} (S_n - \mathbb{E}(S_n)),$$

so that for any $w > 0$, we have

$$\begin{aligned}\widetilde{\mathbb{P}}(Z > w) &\leqslant \widetilde{\mathbb{P}}(\Omega^c_{good}) + \widetilde{\mathbb{P}}\left(\sum_{n=1}^{2k+1} (S_n - \mathbb{E}(S_n)) > w\right) \\ &\leqslant \frac{c'(\theta)}{T} + \widetilde{\mathbb{P}}\left(\sum_{n=1}^{2k+1} (S_n - \mathbb{E}(S_n)) > w\right).\end{aligned}$$

Now, if we denote $Z_1 = \sum_{n=1}^{k+1}(S_{2n-1} - \mathbb{E}(S_{2n-1}))$ and $Z_2 = \sum_{n=1}^{k}(S_{2n} - \mathbb{E}(S_{2n})$, then

$$\widetilde{\mathbb{P}}\left(\sum_{n=1}^{2k+1} (S_n - \mathbb{E}(S_n)) > u + v\right) \leqslant \widetilde{\mathbb{P}}\left(Z_1 > u\right) + \widetilde{\mathbb{P}}\left(Z_2 > v\right),$$

for all $u + v = w$.

Lemma 5.18-Item 3 implies that $S_2, \ldots, S_{2k}$ are independent, so that by the classical Hoeffding inequality, we have for any $x > 0$,

$$\widetilde{\mathbb{P}}\left(Z_1 > \sqrt{kB^2M^2T^{-2}x/2}\right) \leqslant e^{-x},$$

and similarly for $\widetilde{\mathbb{P}}\left(Z_1 > \sqrt{(k+1)B^2M^2T^{-2}x/2}\right) \leqslant e^{-x}$. Hence,

$$\widetilde{\mathbb{P}}\left(Z > \sqrt{kB^2M^2T^{-2}x/2} + \sqrt{(k+1)B^2M^2T^{-2}x/2}\right) \leqslant \frac{c'(\theta)}{T} + 2e^{-x}.$$

But $k \leqslant T(2B)^{-1}$ and $k+1 \leqslant (T+2B)(2B)^{-1} \leqslant T/B$. This leads directly to the first result.

For the second result, note that we can restrict ourselves to Ω_{good} once and for all at the beginning and use the union bound only on the auxiliary independent chains, which explains why we pay $|\mathcal{F}|$ only in front of the deviation e^{-x}. □

Proof of Theorem 5.20. In the sequel, let $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{\mathbb{P}})$ be the probability space and $\mathbf{X}, \ldots, \mathbf{X^{2k+1}}$ be the stochastic chains given by Lemma 5.18. By Lemma 5.18-Item 1 we can assume that Z is also defined on $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{\mathbb{P}})$. We write $\widetilde{\mathbb{E}}$ to denote the expectation taken with respect the probability measure $\widetilde{\mathbb{P}}$.

Now, let B, k, $J_1, \ldots, J_{2k+1}$ as in the proof of Theorem 5.19 and define for $1 \leqslant n \leqslant 2k+1$, the random matrix $\Sigma_n = ((\Sigma_n(f,g))_{f,g\in\mathcal{F}}$ as follows:

$$\Sigma_n(f,g) = \frac{1}{T}\sum_{t\in J_n}\Big(f(X^n_{F,t-m:t-1})g(X^n_{F,t-m:t-1}) - \mathbb{E}(f(X^n_{F,t-m:t-1})g(X^n_{F,t-m:t-1})\Big).$$

Clearly $\widetilde{\mathbb{E}}(\Sigma_n) = 0$. To apply Theorem 1.3 of Tropp (2012), we need to find a deterministic self-adjoint matrix A_n such that $A_n^2 - \Sigma_n^2$ is non negative. This means that for all vector $x \in \mathbb{R}^{\mathcal{F}}$,

$$x^\intercal[A_n^2 - \Sigma_n^2]x \geqslant 0.$$

By taking $A_n = \sigma I_n$, it is sufficient to prove that

$$x^\intercal \Sigma_n^2 x \leqslant \sigma^2\|x\|^2.$$

In the sequel, denote

$$A_{f,h}(X^n_{F,t-m:t-1}) = \Big(f(X^n_{F,t-m:t-1})h(X^n_{F,t-m:t-1}) - \mathbb{E}(f(X^n_{F,t-m:t-1})h(X^n_{F,t-m:t-1})\Big).$$

With this notation, we have

$$\begin{aligned}
x^\intercal \Sigma_n^2 x &= \sum_{f,g\in\mathcal{F}} x_f x_g \frac{1}{T^2} \sum_{t,t'\in J_n} \sum_{h\in\mathcal{F}} A_{f,h}(X^n_{F,t-m:t-1}) A_{g,h}(X^n_{F,t'-m:t'-1}) \\
&= \frac{1}{T^2} \sum_{t,t'\in J_n} \sum_{h\in\mathcal{F}} \left[\sum_f x_f A_{f,h}(X^n_{F,t-m:t-1}) \right] \times \\
&\qquad \left[\sum_g x_g A_{g,h}(X^n_{F,t'-m:t'-1}) \right] \\
&\leqslant \frac{1}{T^2} \sum_{t,t'\in J_n} \sum_{h\in\mathcal{F}} \|x\|^2 \sqrt{\sum_f A^2_{f,h}(X^n_{F,t-m:t-1})} \times \\
&\qquad \sqrt{\sum_g A^2_{g,h}(X^n_{F,t'-m:t'-1})} \\
&\leqslant \frac{4\|x\|^2|\mathcal{F}|}{T^2} \sum_{t,t'\in J_n} \sum_{h\in\mathcal{F}} M^4 \\
&\leqslant \frac{4|\mathcal{F}|^2 B^2 M^4}{T^2} \|x\|^2.
\end{aligned}$$

Hence $\sigma = \frac{2|\mathcal{F}|BM^2}{T}$ works. Denote $Z_1 = \sum_{n=1}^{k+1} \Sigma_{2n-1}$ and $Z_2 = \sum_{n=1}^{k} \Sigma_{2n}$. Lemma 5.18 implies that on Ω_{good},

$$Z = Z_1 + Z_2,$$

so that by the triangle inequality we have for any $u > 0$ and $v > 0$,

$$\widetilde{\mathbb{P}}(\|Z\| > u + v) \leqslant \widetilde{\mathbb{P}}(\Omega^c_{good}) + \widetilde{\mathbb{P}}(\|Z_1\| > u) + \widetilde{\mathbb{P}}(\|Z_2\| > v).$$

Since by Lemma 5.18-item 3, $\Sigma_2, \Sigma_4, \ldots, \Sigma_{2k}$ are i.i.d random matrices, we can apply Theorem 1.3 of Tropp (2012) to deduce that for any $v > 0$,

$$\widetilde{\mathbb{P}}\left(\|Z_2\| > \sqrt{8k\sigma^2 v}\right) \leqslant 2|\mathcal{F}|e^{-v},$$

Similarly, we have that for any $x > 0$,

$$\widetilde{\mathbb{P}}\left(\|Z_2\| > \sqrt{8(k+1)\sigma^2 u}\right) \leqslant 2|\mathcal{F}|e^{-u},$$

and as a consequence, it follows that for any $x > 0$,

$$\widetilde{\mathbb{P}}\left(\|Z\| > \sqrt{8k\sigma^2 x} + \sqrt{8(k+1)\sigma^2 x}\right) \leqslant \frac{c'(\theta)}{T} + 4|\mathcal{F}|e^{-x}.$$

Since $k^{1/2} + (k+1)^{1/2} \leqslant (4T/B)^{1/2}$, the result follows from the inequality above. □

To prove Theorem 5.27 we use arguments from Gaïffas and Guilloux (2012). During the proof of Theorem 5.27 we will need some technical lemmas, stated and proved below.

Lemma 5.36. *Let $\hat{f} = f_{\hat{a}}$ where $\hat{a}$ is defined by (5.30). For any vector $a \in \mathbb{R}^{\boldsymbol{M}}$, the following inequality holds*

$$2\langle \hat{f} - f_a, \hat{f} - p_i\rangle_T + \gamma d\,\|\hat{a}_{S^c(a)}\|_1 \leqslant \gamma d\,\|\hat{a}_{S(a)} - a_{S(a)}\|_1 + 2(b-\bar{b})^T(\hat{a}-a), \tag{5.42}$$

where $S(a) = \{j \in [M] : a_j \neq 0\}$ and the vectors $b, \bar{b} \in \mathbb{R}^{\Phi}$ are defined in (5.28) and (5.31) respectively.

Proof of Lemma 5.36. Throughout the proof we write $\partial g(p)$ to denote the subdifferential mapping of a convex function g at the point p. One can show that p is a global minimum of the convex function g if and only if $0 \in \partial g(p)$. Now since $\hat{a}$ is such that

$$\hat{a} \in \arg\min_{a \in \mathbb{R}^{\Phi}} \{a^T G a - 2a^T b + \gamma d\,\|a\|_1\},$$

it follows that

$$0 \in \partial(\hat{a}^T G\hat{a} - 2\hat{a}^T b + \gamma d\,\|\hat{a}\|_1) = 2G\hat{a} - 2b + \gamma d\,\partial\|\hat{a}\|_1.$$

Thus, it follows that for some $\hat{w} \in \partial\|\hat{a}\|_1$, the following equation holds

$$2G\hat{a} - 2b + \gamma d\,\hat{w} = 0,$$

which implies then

$$(2G\hat{a} - 2b + \gamma d\,\hat{w})^T(\hat{a}-a) = 0, \text{ for any } a \in \mathbb{R}^{\Phi}.$$

From the above equation we can deduce that for any vector $w \in \partial\|a\|_1$ and $a \in \mathbb{R}^{\boldsymbol{M}}$,

$$(2G\hat{a} - 2\bar{b})^T(\hat{a}-a) + \gamma d(\hat{w}-w)^T(\hat{a}-a) = -\gamma d w^T(\hat{a}-a) + 2(b-\bar{b})^T(\hat{a}-a). \tag{5.43}$$

One can easily show by the definition of subdifferentials that

$$(\hat{w} - w)^T(\hat{a} - a) \geqslant 0,$$

for all $\hat{w} \in \|\hat{a}\|_1$ and $w \in \|a\|_1$. Thus, using this fact in equation (5.43) together with the fact that $(2G\hat{a} - 2\bar{b})^T(\hat{a} - a) = 2\langle \hat{f} - f_a, \hat{f} - p_i\rangle_T$, we derive the following inequality

$$2\langle \hat{f} - f_a, \hat{f} - p_i\rangle_T \leqslant -\gamma d w^T(\hat{a} - a) + 2(b - \bar{b})^T(\hat{a} - a). \tag{5.44}$$

It is well know that

$$\partial\|a\|_1 = \{v : \|v\|_\infty \leqslant 1 \text{ and } v^T a = \|a\|_1\}.$$

In other words, $v \in \partial\|a\|_1$ if and only if $v_j = \text{sign}(a_j)$ for $j \in S(a)$ and $v_j \in [-1, 1]$ for all $j \in S^c(a)$. Now, take $w = (w_1, \ldots, w_M) \in \partial\|a\|_1$ of the following form

$$w_\varphi = \begin{cases} \text{sign}(a_j), & \text{if } j \in S(a) \\ \text{sign}(\hat{a}_j), & \text{if } j \in S^c(a) \end{cases},$$

and observe that $w^T(\hat{a} - a) = \sum_{j \in S(a)} \text{sign}(a_j)(\hat{a}_j - a_j) + |\hat{a}_{S^c(a)}|_1$. Thus, by plugging this identify into inequality (5.44), we obtain that

$$2\langle \hat{f} - f_a, \hat{f} - p_i\rangle_T + \gamma d|\hat{a}_{S^c(a)}|_1 \leqslant -\gamma d \sum_{j \in S(a)} \text{sign}(a_j)(\hat{a}_j - a_j) + 2(b - \bar{b})^T(\hat{a} - a),$$

and the result follows, because

$$\left| -\sum_{j \in S(a)} \text{sign}(a_j)(\hat{a}_j - a_j) \right| \leqslant \left|\hat{a}_{S(a)} - a_{S(a)}\right|_1.$$

□

Lemma 5.37. *Let $\hat{f} = f_{\hat{a}}$ where $\hat{a}$ defined by (5.30) with $\gamma \geqslant 2$ and $a \in \mathbb{R}^{\Phi}$. On an event on which*

(i) $\langle \hat{f} - f_a, \hat{f} - p_i\rangle_T \geqslant 0$,

(ii) $|b_j - \bar{b}_j| \leqslant d$ *for all* $j \in [M]$,

the following inequality is satisfied,

$$|\hat{a}_{S^c(a)}|_1 \leqslant \frac{\gamma+2}{\gamma-2}|\hat{a}_{S(a)} - a_{S(a)}|_1, \tag{5.45}$$

where $S(a) = \{j \in [M] : a_j \neq 0\}$.

Proof of Lemma 5.37. Suppose that $\langle \hat{f} - f_a, \hat{f} - p_i \rangle_T \geqslant 0$. In this case, one can use Lemma 5.36 to deduce that

$$\gamma d\,|\hat{a}_{S^c(a)}|_1 \leqslant \gamma d\,|\hat{a}_{S(a)} - a_{S(a)}|_1 + 2\sum_{j\in S(a)}(b_j - \bar{b}_j)(\hat{a}_j - a_j) + 2\sum_{j\in S^c(a)}(b_j - \bar{b}_j)\hat{a}_j.$$

On an event on which $|b_j - \bar{b}_j| \leqslant d$ for all $j \in [M]$, we then have that

$$\gamma d\,|\hat{a}_{S^c(a)}|_1 \leqslant (\gamma+2)d\,|\hat{a}_{S(a)} - a_{S(a)}|_1 + 2d\,|\hat{a}_{S^c(a)}|_1,$$

and the result follows. □

We now prove Theorem 5.27.

Proof of Theorem 5.27. To prove the first part of Theorem 5.27 we proceed as follows. First of all, on the event on which $\langle \hat{f} - f_a, \hat{f} - p_i \rangle_T < 0$, there is nothing to be proved, since in this case

$$\|\hat{f} - p_i\|_T^2 + \|\hat{f} - f_a\|_T^2 - \|f_a - p_i\|_T^2 = \langle \hat{f} - f_a, \hat{f} - p_i \rangle_T < 0.$$

Hence, in what follows, take $a \in \mathbb{R}^M$ such that $|S(a)| \leqslant s$ and $\langle \hat{f} - f_a, \hat{f} - p_i \rangle_T \geqslant 0$. In this case, thanks to Lemma 5.37, we can use Property $\mathbf{RE}(\kappa, c(\gamma), s)$ to the vector $\hat{a} - a$:

$$\|\hat{a}_{S(a)} - a_{S(a)}\|_2^2 \leqslant \kappa^{-1}(\hat{a} - a)^T G(\hat{a} - a).$$

Now, as in the proof of Lemma 5.37, we know that on an event on which $|b_j - \bar{b}_j| \leqslant d$ for all $j \in [M]$, the following bound holds:

$$2|(b - \bar{b})^T(\hat{a} - a)| \leqslant 2d\,\|(\hat{a}_{S(a)} - a_{S(a)})\|_1 + 2d\,\|\hat{a}_{S^c(a)}\|_1$$

By using this inequality together with Lemma 5.36, we conclude that

$$2\langle \hat{f} - f_a, \hat{f} - p_i\rangle_T + (\gamma - 2)d\,\|\hat{a}_{S^c(a)}\|_1 \leqslant (\gamma + 2)d\,\|\hat{a}_{S(a)} - a_{S(a)}\|_1. \quad (5.46)$$

Finally, by Cauchy–Schwartz inequality, we know that

$$\|\hat{a}_{S(a)} - a_{S(a)}\|_1 \leqslant \sqrt{S(a)}\|\hat{a}_{S(a)} - a_{S(a)}\|_2 \leqslant \sqrt{S(a)\kappa^{-1}(\hat{a} - a)^T G(\hat{a} - a)}.$$

Plugging this last inequality into (5.46), we deduce that

$$2\langle \hat{f} - f_a, \hat{f} - p_i\rangle_T + (\gamma - 2)d\,\|\hat{a}_{S^c(a)}\|_1 \leqslant (\gamma + 2)d\,\sqrt{S(a)\kappa^{-1}(\hat{a} - a)^T G(\hat{a} - a)}.$$

To conclude the proof of the first part, note that

$$\begin{cases} 2\langle \hat{f} - f_a, \hat{f} - p_i\rangle_T = \|\hat{f} - p_i\|_T^2 + \|\hat{f} - f_a\|_T^2 - \|f_a - p_i\|_T^2 \\ (\hat{a} - a)^T G(\hat{a} - a) = \|\hat{f} - f_a\|_T^2, \end{cases}$$

and use the inequality $qy - y^2 \leqslant q^2/4$, which is valid for any $q, y > 0$.

For the second part of the result, to control the fluctuations of $b_j - \bar{b}_j$, let us note that $b_j - \bar{b}_j = M_T$, where $(M_t)_{1 \leqslant t \leqslant T}$ is the martingale defined by

$$M_t = \sum_{i=1}^{t} \frac{\varphi(X_{-\infty:t-1})}{T} \left[X_{i,t} - p_i(X_{-\infty:t-1})\right].$$

We can apply the classical bound of Hoeffding's inequality on each increment of the martingale ΔM_t. Note that if $\varphi_j(X_{-\infty:t-1})$ is positive,

$$-\frac{\varphi_j(X_{-\infty:t-1})}{T} p_i(X_{-\infty:t-1}) \leqslant \Delta M_t \leqslant \frac{\varphi_j(X_{-\infty:t-1})}{T}[1 - p_i(X_{-\infty:t-1})],$$

and if $\varphi(X_{-\infty:t-1})$ is negative,

$$\frac{\varphi_j(X_{-\infty:t-1})}{T}[1 - p_i(X_{-\infty:t-1})] \leqslant \Delta M_t \leqslant -\frac{\varphi_j(X_{-\infty:t-1})}{T} p_i(X_{-\infty:t-1}).$$

This leads for every $\theta > 0$ to

$$\mathbb{E}(e^{\theta \Delta M_t} | X_{-\infty:t-1}) \leqslant \exp\left(\frac{\theta^2 \varphi_j(X_{-\infty:t-1})^2}{8T^2}\right) \leqslant \exp\left(\frac{\theta^2 \|\Phi\|_\infty^2}{8T^2}\right).$$

Therefore

$$\mathbb{E}(e^{\theta M_T}) \leqslant \exp\left(\frac{\theta^2 \|\Phi\|_\infty^2}{8T}\right).$$

Hence

$$\mathbb{P}(M_T \geqslant x) \leqslant \exp\left(\frac{\theta^2 \|\Phi\|_\infty^2}{8T} - \theta x\right).$$

By optimizing this in θ and applying the same inequality to $-\varphi$, we get for all positive u

$$\mathbb{P}\left(M_T \geqslant \sqrt{\frac{u\|\Phi\|_\infty^2}{2T}}\right) \leqslant e^{-u} \text{ and } \mathbb{P}\left(|b_j - \bar{b}_j| \geqslant \sqrt{\frac{u\|\Phi\|_\infty^2}{2T}}\right) \leqslant 2e^{-u}$$

Therefore taking $u = \log(M) + \log(2\delta^{-1})$ and then applying the union bound we obtain the result. □

Proof of Lemma 5.29. The proof is done for the lower bound. The argument is similar for the upper bound. We use induction on the time length of S. If $S = \emptyset$, f is constant and $\mathbb{E}(f(X_S)) = \mathbb{E}^{\otimes\mathcal{V}}_{\mathcal{B}(1/2)}(f(X_S))$. Let $Q = \mathcal{B}(1/2)^{\otimes S}$.

If the time length of S is strictly positive, let t be the maximal time of S, let $w_t = \{(i,t) \text{ for } i \text{ such that } (i,t) \in S\}$ and denote $f_{x_{w_t}}(X_{S\setminus w_t}) = f((X_{S\setminus w_t}, x_{w_t}))$ for any $x_{w_t} \in \{0,1\}^{w_t}$. With this notation,

$$\begin{aligned}
\mathbb{E}(f(X_S)) &= \mathbb{E}[\mathbb{E}(f(X_S)|X_{-\infty:t-1})] \\
&= \mathbb{E}\left(\sum_{x_{w_t}\in\{0,1\}^{w_t}} f_{x_{w_t}}(X_{S\setminus w_t})\mathbb{P}(X_{w_t} = x_{w_t}|X_{-\infty:t-1})\right) \\
&= \mathbb{E}\left(\sum_{x_{w_t}\in\{0,1\}^{w_t}} f_{x_{w_t}}(X_{S\setminus w_t}) \prod_{i/(i,t)\in w_t} \mathbb{P}(X_{i,t} = x_{i,t}|X_{-\infty:t-1})\right) \\
&\geqslant (2\mu)^{|w_t|}\mathbb{E}\left(\sum_{x_{w_t}\in\{0,1\}^{w_t}} f_{x_{w_t}}(X_{S\setminus w_t})Q(X_{i,t} = x_{i,t})\right).
\end{aligned}$$

But $\sum_{x_{w_t}\in\{0,1\}^{w_t}} f_{x_{w_t}}(X_{S\setminus w_t})Q(X_{i,t} = x_{i,t})$ is a cylindrical function on $S \setminus w_t$ with time length strictly smaller than S, so by induction,

$$\mathbb{E}\left(\sum_{x_{w_t}\in\{0,1\}^{w_t}} f_{x_{w_t}}(X_{S\setminus w_t})Q(X_{i,t} = x_{i,t})\right)$$
$$\geqslant (2\mu)^{|S\setminus w_t|}\mathbb{E}^{\otimes\mathcal{V}}_{\mathcal{B}(1/2)}\left(\sum_{x_{w_t}\in\{0,1\}^{w_t}} f_{x_{w_t}}(X_{S\setminus w_t}))Q(X_{i,t} = x_{i,t})\right),$$

implying that

$$\mathbb{E}(f(X_S)) \geqslant (2\mu)^{|S|}\mathbb{E}^{\otimes\mathcal{V}}_{\mathcal{B}(1/2)}(f(X_S)),$$

and the result follows. □

Proof of Theorem 5.30. For any $a \in \mathbb{R}^M$ such that $\|a\|_2 = 1$, we have by Cauchy–Schwarz inequality

$$\kappa \leqslant a^\intercal \mathbb{E}(G)a \leqslant a^\intercal Ga + \|a\|_2\|(G - \mathbb{E}(G))a\| \leqslant a^\intercal Ga + \|G - \mathbb{E}(G)\|, \quad (5.47)$$

so that the result follows from Theorem 5.20 with $x = \log(4|\mathcal{F}|/\delta)$ and $\mathcal{F} = \Phi$. □

Proof of Theorem 5.35. First of all, remark that thanks to Lemma 5.29 and since φ_j in this case depends on a neighborhood of size 1, one has that

$$\mathbb{E}(G_{j,j}) = \mathbb{E}(\varphi_j(X)^2) \geqslant 2\mu 1/2 = \mu$$

and similarly for $j \neq k$, $\varphi_j\varphi_k$ is positive and depends on a neighborhood of size 2, hence

$$(1-\mu)^2 \geqslant \mathbb{E}(G_{jk}) \geqslant \mu^2.$$

Moreover let us apply our version of Hoeffding's inequality, i.e. the second result of Theorem 5.19 on all the $\varphi_j^2 = \varphi_j$, $\varphi_j\varphi_k$ and $-\varphi_j\varphi_k$ for $k \neq j$. Hence there exists and event of probability larger than $1 - \frac{c'(\theta)}{T} - \delta$ such that for all $j, k \in [M]$,

$$|G_{jk} - \mathbb{E}(G_{jk})| \leqslant R_T,$$

with

$$R_T = \sqrt{c''(\theta)\frac{(m + \log T + \log|F|)}{T}\log\left(\frac{4M^2}{\delta}\right)},$$

which means that there exists a constant c_1 depending only on the distribution such that for T large enough (depending on θ and $|F|$)

$$R_T = c_1 T^{-1/2}(m + \log T + \log|F|)^{1/2}(\log m + \log|F| + \log\delta^{-1})^{1/2}.$$

Therefore on this event, for all a and J such that $|J| \leqslant s$ and $\|a_{J^c}\|_1 \leqslant c\|a_J\|_1$, and if $\mu^2 \geqslant R_T$,

$$\begin{aligned}
a^\intercal G a &= \sum_{j\in[M]} a_j^2 G_{jj} + \sum_{j\neq k\in[M]} a_j a_k G_{jk} \\
&\geqslant (\mu - R_T)\sum_{j\in[M]} a_j^2 + (\mu^2 - R_T)\sum_{\substack{j\neq k\in[M]\\ a_j a_k \geqslant 0}} a_j a_k \\
&\qquad + ((1-\mu)^2 + R_T)\sum_{\substack{j\neq k\in[M]\\ a_j a_k < 0}} a_j a_k \\
&\geqslant (\mu - \mu^2)\|a\|_2^2 + \mu^2\sum_{j,k\in[M]} a_j a_k \\
&\qquad + (1-2\mu)\sum_{\substack{j\neq k\in[M]\\ a_j a_k < 0}} a_j a_k - R_T\|a\|_1^2 \\
&\geqslant (\mu - \mu^2)\|a\|_2^2 + \mu^2\left(\sum_{j\in[M]} a_j\right)^2 - ((1-2\mu) - R_T)\|a\|_1^2 \\
&\geqslant (\mu - \mu^2)\|a\|_2^2 - ((1-2\mu) + R_T)\left[\|a_J\|_1 + \|a_{J^c}\|_1\right]^2 \\
&\geqslant (\mu - \mu^2)\|a\|_2^2 - ((1-2\mu) + R_T))(1+c)^2\|a_J\|_1^2 \\
&\geqslant (\mu - \mu^2)\|a_J\|_2^2 - ((1-2\mu) + R_T)(1+c)s\|a_J\|_2^2,
\end{aligned}$$

which is the desired result. □

Proof of Corollary 5.34. We shall prove only for the short effect dictionary. The other cases are treated similarly. For this choice of dictionary $\|\Phi\|_\infty = 1$ and $M = |\Phi| = |F|$. Hence, by applying Theorem 5.30 and Theorem 5.27 both with $\delta = T^{-1}$ one deduces that, for T large enough, on an event of probability larger then 1-c_1/T, the following oracle inequality holds

$$\|\hat{f} - p_i(\cdot)\|_T^2 \leqslant \inf_{a\in\mathbb{R}^M} \left\{ \|f_a - p_i(\cdot)\|_T^2 + 4\kappa^{-1}|S(a)| \frac{(\log|F| + \log(2T))}{2T} \right\}, \tag{5.48}$$

where c_1 depends only on the distribution of $\mathbf{X}$ and

$$\kappa = \kappa' - c_1' T^{-1/2} |F|^{1/2} (m + \log(T) + \log|F|)^{1/2} (\log|F| + \log\delta^{-1})^{1/2},$$

with c_1' depending only on the distribution of $\mathbf{X}$ and κ' given by (5.35).

Now, for the choices given by (5.36), (5.38) and (5.40), then, as seen previously $\kappa' = c_2' \log(T))^{-c_3'}$, for positive constants c_2' and c_3' depending only on m and μ and

$$\kappa = \frac{c_2'}{(\log T)^{c_3'}}(1 - o(1)).$$

By plugging κ into (5.48), the result follows. □

5.7 Exercises

Exercise 5.1. Suppose I is a singleton, say $I = \{1\}$, and denote X_t instead of $X_{1,t}$ for convenience. Suppose also that for all $x \in \{0,1\}^{\mathbb{Z}-}$, $p(x) = \mathbb{P}(X_0 = 1|X_{-\infty:-1} = x) = \mathbb{P}(X_0 = 1|X_{-1} = x_{-1})$, that is, the stochastic chain $\mathbf{X} = (X_t)_{t\in\mathbb{Z}}$ is a Markov chain of order 1 taking values in $\{0,1\}$. Find a space-time decomposition for $p(x)$.

Exercise 5.2. Prove (5.2).

Exercise 5.3. Check the space-time decomposition (5.5).

Exercise 5.4. Verify the space-time decomposition (5.8).

Exercise 5.5. Show (5.10).

Exercise 5.6. Show that $\mathbf{Inv}(\kappa)$ corresponds to $\mathbf{RE}(\kappa, \infty, M)$.

Exercise 5.7. Verify that the matrix $\mathbb{E}^{\otimes\mathcal{V}}_{\mathcal{B}(1/2)}(G)$ associated to the short memory dictionary is given by (5.34).

Exercise 5.8. Check that the matrix $\mathbb{E}^{\otimes\mathcal{V}}_{\mathcal{B}(1/2)}(G)$ associated to the short cumulative effect dictionary is given by (5.37) and find its eigenvalues.

Exercise 5.9. Show that $\mathbb{E}^{\otimes\mathcal{V}}_{\mathcal{B}(1/2)}(G)$ associated to the short cumulative effect with spontaneous apparition dictionary is given by (5.39).

Bibliography

D. Abercrombie (1967). *Elements of General Phonetics*. Edinburgh University Press (cit. on p. 56).

E. D. Adrian (1928). *The basis of sensation : the action of the sense organs*. London: W. W. Norton, Incorporated (cit. on p. 69).

E. D. Adrian and D. Bronk (1929). "The discharge of impulses in motor nerve fibres. Part II. The frequency of discharge in reflex and voluntary contractions." *J. Physiol.* 67, pp. 119–151 (cit. on p. 69).

F. Baccelli, M. Davydov, and T. Taillefumier (2020). "Replica-Mean-Field Limits of Fragmentation-Interaction-Aggregation Processes." arXiv: 2005.07962 (cit. on p. 69).

F. Baccelli and T. Taillefumier (2019). "Replica-Mean-Field Limits for Intensity-Based Neural Networks." *SIAM Journal on Applied Dynamical Systems* 18.4, pp. 1756–1797. MR: 4016128. Zbl: 1435.92004 (cit. on p. 69).

E. Bacry, I. Mastromatteo, and J. F. Muzy (2015). "Hawkes Processes in Finance." *Market Microstructure and Liquidity* 01.01, p. 1550005 (cit. on p. 99).

S. Basu and G. Michailidis (2015). "Regularized Estimation in Sparse High-dimensional Time Series Models." *Ann. Stat.* 45.4, pp. 1535–1567. MR: 3357870. Zbl: 1317.62067 (cit. on p. 118).

A. Bateman et al. (Jan. 2004). "The Pfam protein families database." *Nucleic Acids Research* 32.suppl_1, pp. D138–D141 (cit. on pp. 45, 52).

G. Bejerano and G. Yona (Jan. 2001). "Variations on probabilistic suffix trees: statistical modeling and prediction of protein families." *Bioinformatics* 17.1, pp. 23–43 (cit. on pp. 44–47, 52, 53).

A. Belloni and R. I. Oliveira (2017). "Approximate group context tree." *The Annals of Statistics* 45.1, pp. 355–385. MR: 3611495. Zbl: 1426.62241 (cit. on p. 62).

B. Boeckmann et al. (Jan. 2003). "The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003." *Nucleic acids research* 31.1, pp. 365–370 (cit. on p. 52).

J. Brandão de Carvalho (1988). "Réduction vocalique, quantité et accentuation: pour une explication structurale de la divergence entre portugais lusitanien et portugais brésilien." *Boletim de filologia* 32, pp. 5–26 (cit. on p. 56).

L. Brochini, P. Hodara, C. Pouzat, and A. Galves (2017). "Estimation of neuronal interaction graph from spike train data." arXiv: 1612.05226 (cit. on pp. 66, 73, 74).

L. Brochini, A. de Andrade Costa, M. Abadi, A. C. Roque, J. Stolfi, and O. Kinouchi (2016). "Phase transitions and self-organized criticality in networks of stochastic spiking neurons." *Scientific Reports* 6.35831, pp. 1–15 (cit. on p. 69).

S. Bubeck (2015). *Convex Optimization: Algorithms and Complexity*. Vol. 8. Foundations and Trends in Machine Learning 3–4. Now publishers inc., pp. 231–357 (cit. on p. 114).

P. Bühlmann and S. van de Geer (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. 1st. Springer Publishing Company, Incorporated. MR: 2807761. Zbl: 1273.62015 (cit. on p. 112).

P. Bühlmann and A. J. Wyner (1999). "Variable length Markov chains." *Ann. Statist.* 27, pp. 480–513. MR: 1714720. Zbl: 0983.62048 (cit. on p. 15).

N. Cesa-Bianchi and G. Lugosi (2006). *Prediction, learning, and games*. Cambridge: Cambridge University Press, pp. xii+394. MR: 2409394(2009g : 91006). Zbl: 1114.91001 (cit. on pp. 29, 31).

S. Chen, A. Shojaie, E. Shea-Brown, and D. Witten (2019). "The Multivariate Hawkes Process in High Dimensions: Beyond Mutual Excitation." arXiv: 1707.04928 (cit. on pp. 99, 109).

J. Chevallier, M. J. Cáceres, M. Doumic, and P. Reynaud-Bouret (2015). "Microscopic approach of a time elapsed neural model." *Mathematical Models and Methods in Applied Sciences* 25.14, pp. 2669–2719. MR: 3411353. Zbl: 1325.35231 (cit. on p. 99).

J. Chevallier, A. Duarte, E. Löcherbach, and G. Ost (2019). "Mean field limits for nonlinear spatially extended Hawkes processes with exponential memory kernels." *Stochastic Processes and their Applications* 129.1, pp. 1–27. MR: 3906989. Zbl: 1404.60069 (cit. on p. 99).

J. Chevallier and G. Ost (2020). "Fluctuations for spatially extended Hawkes processes." *Stochastic Processes and their Applications* 130.9, pp. 5510–5542. MR: 4127337. Zbl: 1454.60062 (cit. on p. 99).

E. S. Chornoboy, L. P. Schramm, and A. F. Karr (1988). "Maximum likelihood identification of neural point process systems." English. *Biological Cybernetics* 59.4-5, pp. 265–275. MR: 0961117. Zbl: 0658.92007 (cit. on p. 99).

F. Comets, R. Fernandéz, and P. A. Ferrari (Aug. 2002). "Processes with long memory: Regenerative construction and perfect simulation." *Ann. Appl. Probab.* 12.3, pp. 921–943. Zbl: 1016.60061 (cit. on p. 97).

I. Csiszár and Z. Talata (Feb. 2006a). "Consistent estimation of the basic neighborhood of Markov random fields." *Ann. Statist.* 34.1, pp. 123–145. Zbl: 1102.62105 (cit. on p. 84).

— (2006b). "Context tree estimation for not necessarily finite memory processes, via BIC and MDL." *IEEE Trans. Inform. Theory* 52.3, pp. 1007–1016. MR: 2238067. Zbl: 1284.94027 (cit. on pp. 14, 17–19, 21, 23, 33, 35, 40, 41).

J. A. Cuesta-Albertos, R. Fraiman, A. Galves, J. E. García, and M. Svarc (2007). "Identifying rhythmic classes of languages using their sonority: a Kolmogorov-Smirnov approach." *Journal of Applied Statistics* 34.6, pp. 749–761. MR: 2410047 (cit. on p. 56).

R. Dauer (1983). "Stress-timing and syllable-timing reanalized." *Journal of Phonetics* 11, pp. 51–62 (cit. on p. 56).

A. De Masi, A. Galves, E. Löcherbach, and E. Presutti (2015). "Hydrodynamic Limit for Interacting Neurons." English. *Journal of Statistical Physics* 158.4, pp. 866–902. Zbl: 1315.35222 (cit. on p. 69).

J. Dedecker and P. Doukhan (2003). "A new covariance inequality and applications." *Stochastic Process. Appl.* 106.1, pp. 63–80. MR: 1983043. Zbl: 1075.60513 (cit. on p. 25).

J. Dedecker and C. Prieur (2005). "New dependence coefficients. Examples and applications to statistics." *Probab. Theory Relatated Fields* 132, pp. 203–236. MR: 2199291. Zbl: 1061.62058 (cit. on p. 26).

S. Ditlevsen and E. Löherbach (2017). "Multi-class oscillating systems of interacting neurons." *Stoch. Proc. Appl* 127.6, pp. 1840–1869. arXiv: 1512.00265. MR: 3646433. Zbl: 1367.92024 (cit. on p. 99).

A. Duarte, A. Galves, E. Löcherbach, and G. Ost (2019). "Estimating the interaction graph of stochastic neural dynamics." *Bernoulli* 25.1, pp. 771–792. MR: 3892336. Zbl: 1442.62214 (cit. on p. 66).

A. Duarte and G. Ost (2016). "A model for neural activity in the absence of external stimulus." *Markov Proc. Rel. Fields* 22.1, pp. 37–52. MR: 3523978. Zbl: 1342.60165 (cit. on p. 69).

A. Duarte, G. Ost, and A. A. Rodríguez (2015). "Hydrodynamic Limit for Spatially Structured Interacting Neurons." *Journal of Statistical Physics* 161.5, pp. 1163–1202. MR: 3422922. Zbl: 1333.82019 (cit. on p. 69).

D. Duarte, A. Galves, and N. L. Garcia (2006). "Markov approximation and consistent estimation of unbounded probabilistic suffix trees." *Bull. Braz. Math. Soc.* 37.4, pp. 581–592. MR: 2284889. Zbl: 1110.60092 (cit. on p. 17).

P. Dyan and L. F. Abbott (2001). *Theoretical neuroscience. Computational and mathematical modeling of neural systems*. MIT Press. MR: 1985615 (cit. on p. 69).

B. Efron and R. J. Tibshirani (1993). *An introduction to the bootstrap*. Vol. 57. Monographs on Statistics and Applied Probability. New York: Chapman and Hall, pp. xvi+436. MR: 1270903(95h:62077). Zbl: 0835.62038 (cit. on pp. 24, 60).

E. Eskin, W. S. Noble, and Y. Singer (2003). "Protein Family Classification Using Sparse Markov Transducers." *Journal of Computational Biology* 10.2. PMID: 12804091, pp. 187–213 (cit. on p. 48).

P. A. Ferrari, R. Fernandéz, and A. Galves (2001). "Coupling, renewal and perfect simulation of chains of infinite order." MR: 1978829 (cit. on pp. 98, 106).

N. Fournier and E. Löcherbach (2016). "On a toy model of interacting neurons." *Annales de l'IHP* 52, pp. 1844–1876. MR: 3573298. Zbl: 1355.92014 (cit. on p. 69).

S. Frota and M. Vigário (2001). "On the correlates of rhythm distinctions: the European/Brazilian Portuguese case." *Probus* 13, pp. 247–275 (cit. on p. 56).

S. Gaïffas and A. Guilloux (2012). "High-dimensional additive hazards models and the Lasso." *Electron. J. Statist.* 6, pp. 522–546. MR: 2988418. Zbl: 1274.62655 (cit. on p. 131).

S. Gaïffas and G. Matulewicz (2019). "Sparse inference of the drift of a high-dimensional Ornstein-Uhlenbeck process." *Journal of Multivariate Analysis* 169, pp. 1–20. MR: 3875583 (cit. on p. 118).

A. Galves, C. Galves, J. E. García, N. L. Garcia, and F. G. Leonardi (2012). "Context tree selection and linguistic rhythm retrieval from written texts." *Ann. Appl. Stat.* 6.1, pp. 186–209. MR: 2951534. Zbl: 1235.68305 (cit. on pp. 4, 6, 14, 24, 25, 44, 55, 56, 58–61).

A. Galves, N. L. Garcia, E. Löcherbach, and E. Orlandi (Aug. 2013). "Kalikow-type decomposition for multicolor infinite range particle systems." *Ann. Appl. Probab.* 23.4, pp. 1629–1659. MR: 3098444. Zbl: 1281.60079 (cit. on p. 97).

A. Galves and F. G. Leonardi (2008). "Exponential inequalities for empirical unbounded context trees." In: *In and out of equilibrium. 2*. Vol. 60. Progr. Probab. Basel: Birkhäuser, pp. 257–269. MR: 2477385(2010c:62269). Zbl: 1151.62343 (cit. on pp. 6, 12, 13, 26).

A. Galves and E. Löcherbach (2008). "Stochastic chains with memory of variable length." In: *Festschrift in honor of Jorma Rissanen on the occasion of his 75th birthday*. Ed. by P. Grunwald, P. Myllymäki, I. Tabus, M. Weinberger, and B. Yu. TICSP series 38. Finland: Tampere International Center for Signal Processing, pp. 117–133 (cit. on p. 17).

— (June 2013). "Infinite Systems of Interacting Chains with Memory of Variable Length—A Stochastic Model for Biological Neural Nets." *Journal of Statistical Physics* 151.5, pp. 896–921. Zbl: 1276.82046 (cit. on pp. 68, 97, 101, 102, 106, 108).

A. Galves, V. Maume-Deschamps, and B. Schmitt (2008). "Exponential inequalities for VLMC empirical trees." *ESAIM Probab. Stat* 12, pp. 43–45. MR: 2374639. Zbl: 1182.62165 (cit. on pp. 12, 15).

J. E. García and V. A. González-López (2017). "Consistent estimation of partition Markov models." *Entropy* 19.4, Paper No. 160, 15. MR: 3653184 (cit. on p. 48).

A. Garivier (2006). "Consistency of the unlimited BIC context tree estimator." *IEEE Trans. Inform. Theory* 52.10, pp. 4630–4635. MR: 2300844. Zbl: 1320.62014 (cit. on p. 21).

A. Garivier and F. G. Leonardi (2011). "Context tree selection: a unifying view." *Stochastic Process. Appl.* 121.11, pp. 2488–2506. MR: 2832411. Zbl: 1397.60130 (cit. on pp. 6, 13, 14, 17, 20).

W. Gerstner (Jan. 1995). "Time structure of the activity in neural network models." *Phys. Rev. E* 51 (1), pp. 738–758 (cit. on p. 69).

W. Gerstner and J. L. van Hemmen (1992). "Associative memory in a network of spiking neurons." *Network: Computation in Neural Systems* 3.2, pp. 139–164. MR: 1193432. Zbl: 0825.92051 (cit. on p. 69).

W. Gerstner and W. Kistler (2002). *Spiking Neuron Models: An Introduction*. New York, NY, USA: Cambridge University Press. MR: 1923120. Zbl: 1100.92501 (cit. on p. 69).

P. Guttorp (1995). *Stochastic modeling of scientific data*. Stochastic Modeling Series. Chapman & Hall, London, pp. xii+372. MR: 1358359. Zbl: 0862.60034 (cit. on p. 11).

N. R. Hansen, P. Reynaud-Bouret, and V. Rivoirard (2015a). "Lasso and probabilistic inequalities for multivariate point processes." *Bernoulli* 21, pp. 83–143. MR: 3322314. Zbl: 1375.60092 (cit. on p. 99).

— (Feb. 2015b). "Lasso and probabilistic inequalities for multivariate point processes." *Bernoulli* 21.1, pp. 83–143. MR: 3322314. Zbl: 1375.60092 (cit. on pp. 103, 109, 117, 118, 123, 124).

A. G. Hawkes (1971). "Spectra of Some Self-Exciting and Mutually Exciting Point Processes." *Biometrika* 58.1, pp. 83–90. MR: 0278410. Zbl: 0219.60029 (cit. on p. 99).

P. Hodara and E. Löcherbach (2017a). "Hawkes processes h variable length memory and an infinite number of components." *Advances in Applied Probability* 49.1, pp. 84–107. MR: 3631217 (cit. on p. 99).

— (2017b). "Hawkes Processes with variable length memory and an infinite number of components." *Adv. Appl. Probab.* 49.1, pp. 84–107. MR: 3631217 (cit. on p. 69).

P. Hodara and E. Löcherbach (2017c). "Hawkes processes with variable length memory and an infinite number of components." *Advances in Applied Probability* 49.1, pp. 84–107 (cit. on p. 118).

A. L. Hodgkin and A. F. Huxley (1952). "A quantitative description of membrane current and its application to conduction and excitation in nerve." *The Journal of Physiology* 117, pp. 500–544 (cit. on p. 69).

X. J. Hunt, P. Reynaud-Bouret, V. Rivoirard, L. Sansonnet, and R. Willett (2019). "A data-dependent weighted LASSO under Poisson noise." *IEEE Trans. Inform. Theory* 65.3, pp. 1589–1613. MR: 3923187 (cit. on pp. 117, 118).

A. James (1940). *Speech Signals in Telephony*. Sir I. Pitman & sons, Limited (cit. on p. 56).

D. H. Johnson (Dec. 1996). "Point process models of single-neuron discharges." *Journal of Computational Neuroscience* 3.4, pp. 275–299 (cit. on p. 99).

S. Kalikow (Dec. 1990). "Random markov processes and uniform martingales." *Israel Journal of Mathematics* 71.1, pp. 33–54. Zbl: 0711.60041 (cit. on p. 97).

U. Kleinhenz (1997). "Domain typology at the phonology-syntax interface." In: *Interfaces in Linguistic Theory*. Ed. by G. M. et al. Lisboa: APL/Colibri, pp. 201–220 (cit. on p. 57).

F. G. Leonardi (Mar. 2006). "A generalization of the PST algorithm: modeling the sparse nature of protein sequences." *Bioinformatics* 22.11, pp. 1302–1307 (cit. on pp. 44, 45, 48, 52, 54).

— (2010). "Some upper bounds for the rate of convergence of penalized likelihood context tree estimators." *Braz. J. Probab. Stat.* 24.2, pp. 321–336. MR: 2643569. Zbl: 1192.62193 (cit. on p. 6).

F. G. Leonardi, R. R. S. Carvalho, and I. Frondana (2021). "Strong structure recovery for partially observed discrete Markov random fields on graphs." arXiv: 1911.12198 (cit. on p. 14).

B. Mark, G. Raskutti, and R. Willett (2019). "Network Estimation From Point Process Data." *IEEE Transactions on Information Theory* 65.5, pp. 2953–2975. MR: 3951378 (cit. on p. 118).

— (n.d.). "Estimating Network Structure from Incomplete Event Data." In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics, 16–18 Apr 2019*. Ed. by K. Chaudhuri and M. Sugiyama. Vol. 89. Proceedings of Machine Learning Research. PMLR, pp. 2535–2544 (cit. on p. 118).

A. A. Markov (2006). "An Example of Statistical Investigation of the Text Eugene Onegin Concerning the Connection of Samples in Chains." *Science in Context* 19.4, pp. 591–600. MR: 2349178. Zbl: 1155.01005 (cit. on p. 2).

J. Mehler, E. Dupoux, T. Nazzi, and G. Dehaene-Lambertz (1996). "Coping with linguistic diversity: the infant's viewpoint." In: J. Morgan and K. Demuth (Eds.) Signal to syntax: bootstrapping from speech to grammar in early acquisition. Hillsdale, NJ: LEA, pp. 101–116 (cit. on p. 56).

M. Nespor and I. Vogel (2012). *Prosodic phonology*. Dordrecht: De Gruyter Mouton (cit. on p. 57).

G. Ost and P. Reynaud-Bouret (2020). "Sparse space–time models: Concentration inequalities and Lasso." *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* 56.4, pp. 2377–2405. MR: 4164841 (cit. on p. 94).

W. R. Pearson (1995). "Comparison of methods for searching protein sequence databases." *Protein Science* 4.6, pp. 1145–1160 (cit. on p. 53).

V. Pernice, B. Staude, S. Cardanobile, and S. Rotter (May 2011). "How Structure Determines Correlations in Neuronal Networks." *PLOS Computational Biology* 7.5, pp. 1–14. MR: 2821638 (cit. on p. 99).

K. L. Pike (1945). *The intonation of American English*. Ann Arbor: University of Michigan Press (cit. on p. 56).

C. Pouzat (2021). *GitHub of Christophe Pouzat* (cit. on p. 75).

F. Ramus (2002). "Acoustic correlates of linguistic rhythm: perspectives." In: *Proceedings of Speech Prosody 2002*, pp. 115–120 (cit. on p. 56).

F. Ramus, M. Nespor, and J. Mehler (1999). "Correlates of linguistic rhythm in the speech signal." *Cognition* 73, pp. 265–292 (cit. on p. 56).

P. Reynaud-Bouret and E. Roy (2007). "Some non asymptotic tail estimates for Hawkes processes." *Bulletin of the Belgian Mathematical Society-Simon Stevin* 13.5, pp. 883–896. MR: 2293215. Zbl: 1120.60052 (cit. on pp. 103, 109).

J. Rissanen (1983). "A universal data compression system." *IEEE Trans. Inform. Theory* 29.5, pp. 656–664. MR: 0730903. Zbl: 0521.94010 (cit. on pp. 14, 15).

P. Robert and J. Touboul (2016). "On the dynamics of random neuronal networks." *Journal of Statistical Physics* 165, pp. 545–584. MR: 3562424. Zbl: 1360.82071 (cit. on p. 69).

D. Ron, Y. Singer, and N. Tishby (1996). "The power of amnesia: Learning probabilistic automata with variable memory length." *Machine Learning* 25.2, pp. 117–149. MR: 4208209. Zbl: 0869.68066 (cit. on p. 45).

F. Sândalo, M. B. Abaurre, A. Mandel, and C. Galves (2006). "Secondary stress in two varieties of Portuguese and the Sotaq optimality based computer program." *Probus* 18, pp. 97–125 (cit. on pp. 56, 62).

G. Schwarz (1978). "Estimating the dimension of a model." *Ann. Statist.* 6, pp. 461–464. MR: 0468014. Zbl: 0379.62005 (cit. on p. 17).

P. C. Shields (1996). *The ergodic theory of discrete sample paths*. Vol. 13. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, pp. xii+249. MR: 1400225. Zbl: 0879.28031 (cit. on pp. 9, 12).

Z. Talata and T. Duncan (July 2009). "Unrestricted BIC context tree estimation for not necessarily finite memory processes." In: *Information Theory, 2009. ISIT 2009. IEEE International Symposium on*, pp. 724–728 (cit. on p. 16).

J. A. Tropp (Aug. 2012). "User-Friendly Tail Bounds for Sums of Random Matrices." *Foundations of Computational Mathematics* 12.4, pp. 389–434. Zbl: 1259.60008 (cit. on pp. 111, 129, 130).

G. Viennet (1997). "Inequalities for absolutely regular sequences: application to density estimation." *Probability Theory and Related Fields* 107, pp. 467–492. Zbl: 0933.62029 (cit. on pp. 103, 109).

M. Vigário (2003). *The prosodic word in European Portuguese*. Berlim/New York: Mouton Degruyter (cit. on pp. 57, 62).

F. M. J. Willems, Y. M. Shtarkov, and T. J. Tjalkens (1995). "The context-tree weighting method: Basic properties." *IEEE Trans. Inf. Theory* 41.3, pp. 653–664. Zbl: 0837.94011 (cit. on pp. 18, 19).

K. Yaginuma (2016). "A stochastic system with infinite interacting components to model the time evolution of the membrane potentials of a population of neurons." *Journal of Statistical Physics* 163.3, pp. 642–658. MR: 3483249. Zbl: 1346.82028 (cit. on p. 69).

Títulos Publicados — 33º Colóquio Brasileiro de Matemática

Geometria Lipschitz das singularidades – *Lev Birbrair e Edvalter Sena*

Combinatória – *Fábio Botler, Maurício Collares, Taísa Martins, Walner Mendonça, Rob Morris e Guilherme Mota*

Códigos geométricos, uma introdução via corpos de funções algébricas – *Gilberto Brito de Almeida Filho e Saeed Tafazolian*

Topologia e geometria de 3-variedades, uma agradável introdução – *André Salles de Carvalho e Rafał Marian Siejakowski*

Ciência de dados: algoritmos e aplicações – *Luerbio Faria, Fabiano de Souza Oliveira, Paulo Eustáquio Duarte Pinto e Jayme Luiz Szwarcfiter*

Discovering Poncelet invariants in the plane – *Ronaldo A. Garcia e Dan S. Reznik*

Introdução à geometria e topologia dos sistemas dinâmicos em superfícies e além – *Víctor León e Bruno Scárdua*

Equações diferenciais e modelos epidemiológicos – *Marlon M. López-Flores, Dan Marchesin, Vítor Matos e Stephen Schecter*

Differential Equation Models in Epidemiology – *Marlon M. López-Flores, Dan Marchesin, Vítor Matos e Stephen Schecter*

A friendly invitation to Fourier analysis on polytopes – *Sinai Robins*

PI-álgebras: uma introdução à PI-teoria – *Rafael Bezerra dos Santos e Ana Cristina Vieira*

First steps into Model Order Reduction – *Alessandro Alla*

The Einstein Constraint Equations – *Rodrigo Avalos e Jorge H. Lira*

Dynamics of Circle Mappings – *Edson de Faria e Pablo Guarino*

Statistical model selection for stochastic systems with applications to Bioinformatics, Linguistics and Neurobiology – *Antonio Galves, Florencia Leonardi e Guilherme Ost*

Transfer operators in Hyperbolic Dynamics - an introduction – *Mark F. Demers, Niloofar Kiamari e Carlangelo Liverani*

A course in Hodge Theory: Periods of Algebraic Cycles – *Hossein Movasati e Roberto Villaflor Loyola*

A dynamical system approach for Lane-Emden type problems – *Liliane Maia, Gabrielle Nornberg e Filomena Pacella*

Visualizing Thurston's geometries – *Tiago Novello, Vinícius da Silva e Luiz Velho*

Scaling problems, algorithms and applications to Computer Science and Statistics – *Rafael Oliveira e Akshay Ramachandran*

An introduction to Characteristic Classes – *Jean-Paul Brasselet*

Instituto de
Matemática
Pura e Aplicada

ISBN 978-65-89124-29-0