Neighborhood radius estimation for Variable-neighborhood Random Fields

Eva Löcherbach, Enza Orlandi

October 4, 2010

Abstract

We consider random fields defined by finite-region conditional probabilities depending on a neighborhood of the region which changes with the boundary conditions. To predict the symbols within any finite region it is necessary to inspect a random number of neighborhood symbols which might change according to the value of them. In analogy to the one dimensional setting we call these neighborhood symbols the *context* of the region. This framework is a natural extension, to d-dimensional fields, of the notion of variable-length Markov chains introduced by Rissanen (1983) in his classical paper. We define an algorithm to estimate the radius of the smallest ball containing the context based on a realization of the field. We prove the consistency of this estimator. Our proofs are constructive and yield explicit upper bounds for the probability of wrong estimation of the radius of the context.

Key words: Gibbs measures, random lattice fields, Variable-neighborhood random fields, Context algorithm, consistent estimation.

AMS Classification: Primary: 60D05, 62F12 Secondary: 60G55, 60G60, 62M40

1 Introduction

We consider random fields on $\mathbb{Z}^d$ with finite state space defined by prescribing a family of conditional probabilities indexed by finite subsets Λ of $\mathbb{Z}^d$. We assume that these conditional probabilities depend on a finite neighborhood which changes according to the boundary conditions. Contrary to standard Markov random fields which are defined by a family of conditional probabilities depending on a *fixed* neighborhood and not sensitive to the boundary conditions (fixed order Markov dependence), the families of conditional probabilities considered here are not restricted to a predefined uniform depth. Rather, by examining the training data, a model is constructed that fits higher order Markov dependencies where needed, while using lower order Markov dependence elsewhere. We call these random fields *Variable-neighborhood random fields* or *Parsimonious Markov random fields*.

Adopting this parsimonious description means that we are aiming at reducing information by finding the minimal neighborhood of a given block of sites that is able to predict the sites within this block. The neighborhood changes when the outside configuration of the field changes, and the dependencies depend on the realization of the field.

Possible applications of Variable-neighborhood random fields are in image analysis, neuroscience's and in general in spatial statistics, whenever information reduction is needed.

The notion of Variable-neighborhood random fields has been inspired by Rissanen's Minimum Description Length principle for Markov chains, see Rissanen (1983), [21]. Rissanen calls the relevant neighborhood of a site, i.e. the sequence of symbols needed to predict the next symbol, given a finite sample, *context* of a site and proposes an estimator of the length of the context. Since this seminal paper, there have been several implementations and extensions of the method. We refer to the book of Grunwald (2007), [17], for a comprehensive introduction and to a review paper by Galves and Löcherbach (2008), [13]. All these generalizations are related to processes in dimension one. Our aim is to extend this method to more than one dimension and to define an estimator of the context in the framework of random fields.

This requires to define a random field which can predict the symbol at a given site by inspecting a "random" number of neighborhood symbols which might change according to the value of them. In analogy to the one dimensional setting we call this neighborhood, i.e. the subset of symbols needed to predict the symbol at the given site, the *context* of this site. For such random fields we estimate the radius of the context of a given site, i.e. the radius of the smallest ball containing the context of this site. It is enough to consider the contexts for one site, since in our setting the one point specification uniquely determines the specification for any other set. We apply a penalized pseudo-likelihood method, first introduced by Besag (1975), [2], and developed by Csiszàr and Talata (2006), [4], in order to construct our estimator. Our estimator is a function of the observed blocks or patterns appearing in the sample. It is based on a sequence of local decisions between two possible values of the radius of the context, lumping them together whenever their corresponding one point conditional probabilities are similar. We propose an estimator for any site within our observation window, depending on its local neighborhood. Hence we deal with a family of estimators indexed by the centers of observation patterns. For this family of estimators, we give in Theorem 3.5 and Theorem 3.9 explicit error-bounds for the probability of over- and underestimation. These bounds are non asymptotic with respect to the number of observed sites, i.e. the size of the observation window. As a consequence, we obtain the consistency of the neighborhood radius estimator.

Our results are based on several deviation inequalities which are interesting in its own right. They are collected in Sections 4 and 5. The first part of them (Section 4) is based on results obtained by Dedecker (2001), [5], on deviation inequalities for random fields, the second part (Section 5) is a rewriting of typicality results obtained by Csiszàr and Talata (2006), [4]. Csiszàr and Talata are only interested in consistency and they do not give explicit upper bounds for the error probabilities. We want to control the error bounds, for any fixed n, and so we carry on their ideas into non-asymptotic deviation inequalities.

We implement the estimates under the requirement that the one point conditional probabilities are strictly positive. This is enough for the overestimation. To implement the estimates for the underestimation, we need to assume that Dobrushin's contraction condition holds, see Dobrushin (1968), [6] and [7], and that there exists some finite order L, unknown to the statistician, such that the random field is Markov of order at most L. In the language of context-trees this means that we deal with finite trees only.

There is large number of papers devoted to parameter estimation for Markov random fields when the structure of the interaction is known, see for example Gidas (1993), [16], Comets (1992), [3], and many others. Typically, the parameter estimation addresses the problem of estimating parameters entering in determining the potential, but not directly the conditional probabilities. Quite recently, the non-parametric problem of model selection has been addressed, i.e. the statistical estimation of the interaction structure, see for example Ji and Seymour (1996), [18]. Csiszàr and Talata (2006), [4], propose to estimate the basic neighborhood of Markov random fields and estimate the support of the neighborhood (i.e. its geometrical structure) which is relevant to determine the conditional probabilities. In their framework this neighborhood does not depend on the configuration, hence they work in a strict Markovian frame. In Galves et al. (2010), [14], a related problem has been studied. The authors estimate for an Ising model having pairwise interactions of infinite range the pairs of interacting sites based on i.i.d. observations of the field. Our paper is not situated in the same framework. We do not address the problem of estimating the geometrical structure of the contexts, since this would require to introduce too many free parameters. We deal with a problem which is simpler and more difficult at the same time: we estimate only the radius of the basic neighborhood, but this neighborhood varies when the configuration changes. This last feature is the main difference from previous models which have appeared in the literature.

The paper is organized as follows. In Section 2 we define the Variable-neighborhood random fields, based on the prescription of a "variable-neighborhood"-specification and we provide two examples. In Section 3 we define the estimator of the radius of a single-site context and formulate the main results. In Theorem 3.5 we give the bound on the probability of overestimation and in Theorem 3.9 the bound on the probability of underestimation, under suitable assumptions on the decay of correlations in the field. In Section 4 we prove the deviation inequalities needed for the control of the underestimation. Section 5 is devoted to the proof of the deviation inequalities needed for the control of the overestimation. In Sections 6 and 7 we give the proof of the main results. We conclude with some final remarks in Section 8. In Section 9, the appendix, we collect some mathematical tools needed along the way. In particular we prove a relation between single site contexts and contexts of finite sets of sites.

Acknowledgments. The authors thank Antonio Galves for introducing them to the problem. Enza Orlandi thanks Antonio Galves and Roberto Fernández for helpful discussions at the starting of this project. She further acknowledges enlightening discussions with Anton Bovier, Stephan Luckhaus and Errico Presutti. The authors have been partially supported by Prin07: 20078XYHYS (E.O.), ANR-08-BLAN-0220-01 (E.L.) and Galileo 2008-09 (E.L. and E.O.), GREFI-MEFI.

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-08-BLAN-0220-01.

2 Variable-neighborhood random fields

We consider the d dimensional lattice $\mathbb{Z}^d$. The points $i \in \mathbb{Z}^d$ are called sites, $\|i\|$ denotes the maximum norm of i, i.e. for $i = (i_1, \dots, i_d)$, $\|i\| = \max(|i_1|, \dots, |i_d|)$ is the maximum of the absolute values of the coordinates of i. The cardinality of a finite set Δ is denoted by $|\Delta|$. The notations $\subset$ and $\Subset$ denote inclusion and strict inclusion. Subsets of $\mathbb{Z}^d$ will be denoted with uppercase Greek letters. If Λ is a finite set, we write $\Lambda \Subset \mathbb{Z}^d$.

A *random field* X is a family of random variables indexed by the sites i of the lattice, $\{X_i : i \in \mathbb{Z}^d\}$, where each X_i is a random variable taking values in a finite set $\mathcal{A}$.

We denote the set of all possible configurations of the random field by

$$\Omega = \mathcal{A}^{\mathbb{Z}^d},$$

where Ω is endowed with the product topology. We adopt the following notational conventions. We write $\omega_\Lambda \in \mathcal{A}^\Lambda$ for the restriction of the configuration ω to the subset Λ. If $\Lambda = \{i\}$ is a singleton, we shall write $\omega(i)$ for $\omega_{\{i\}}$. Configurations defined by regions are factorized with omitted subscripts indicating completion to the rest of the lattice: $\omega_\Lambda \eta_{\Lambda^c} = \omega_\Lambda \eta$. We call *local configurations* the elements of $\cup_{\Lambda \Subset \mathbb{Z}^d} \mathcal{A}^\Lambda$.

We identify the random field $\{X_i : i \in \mathbb{Z}^d\}$ with the coordinate maps X_i by $X_i(\omega) = \omega(i)$, for any $\omega \in \Omega$, and from now on we will use this *canonical version* of the random field. We define the following $\sigma-$algebras: For any $\Gamma \subset \mathbb{Z}^d$, let

$$\mathcal{F}_\Gamma = \sigma\{X_i : i \in \Gamma\} \text{ and } \mathcal{F} = \sigma\{X_i : i \in \mathbb{Z}^d\}.$$

In this set up a random field is just a probability measure on the product space $(\Omega, \mathcal{F})$. This measure is defined by local specifications. To define them, we recall the following well-known notions in statistical mechanics, see Georgii (1988), [15].

Definition 2.1 *A probability kernel on $(\Omega, \mathcal{F})$ is a function $\Gamma(\,\cdot \mid \cdot\,) : \mathcal{F} \times \Omega \longrightarrow [0,1]$ such that*

(a) $\Gamma(\,\cdot \mid \omega)$ is a probability measure on $(\Omega, \mathcal{F})$, for each $\omega \in \Omega$,

(b) $\Gamma(A \mid \cdot\,)$ is a $\mathcal{F}$-measurable function for each $A \in \mathcal{F}$.

Definition 2.2 *A specification on $(\Omega, \mathcal{F})$ is a family $\gamma = \{\gamma_\Lambda\}_{\Lambda \Subset \mathbb{Z}^d}$ of probability kernels on $(\Omega, \mathcal{F})$ such that*

(a) For each $\Lambda \Subset \mathbb{Z}^d$ and each $A \in \mathcal{F}$, the function $\gamma_\Lambda(A \mid \cdot\,)$ is $\mathcal{F}_{\Lambda^c}-$measurable,

(b) For each $\Lambda \Subset \mathbb{Z}^d$ and each $A \in \mathcal{F}_{\Lambda^c}$, $\gamma_\Lambda(A \mid \omega) = 1_A(\omega)$,

(c) For any pair of regions Λ and Δ, with $\Lambda \subset \Delta \Subset \mathbb{Z}^d$, and any measurable set A,

$$\int \gamma_\Delta(d\omega' \mid \omega)\, \gamma_\Lambda(A \mid \omega') \;=\; \gamma_\Delta(A \mid \omega) \tag{2.1}$$

for all $\omega \in \Omega$.

Definition 2.3 *A probability measure μ on $(\Omega, \mathcal{F})$ is consistent with a specification γ if for each $\Lambda \Subset \mathbb{Z}^d$ and for each $A \in \mathcal{F}$,*

$$\int \mu(d\omega)\, \gamma_\Lambda(A \mid \omega) \;=\; \mu(A). \tag{2.2}$$

We now define the Variable-neighborhood random fields.

Definition 2.4 Variable-neighborhood random field
Let μ be a probability measure on $(\Omega, \mathcal{F})$ consistent with the specification γ. Then μ is a variable-neighborhood random field if for any $\Lambda \Subset \mathbb{Z}^d$ and for $\mu-$almost all ω_{Λ^c} the following holds: there exists $\Gamma \Subset \mathbb{Z}^d$ such that

$$\gamma_\Lambda(\cdot \mid \omega_{\Lambda^c}) = \gamma_\Lambda(\cdot \mid \omega_\Gamma),$$

and for all $\tilde{\Gamma} \subset \mathbb{Z}^d$, if $\gamma_\Lambda(\cdot \mid \omega_{\Lambda^c}) = \gamma_\Lambda(\cdot \mid \omega_{\tilde{\Gamma}})$, then $\Gamma \subset \tilde{\Gamma}$. We define

$$\mathrm{sp}_\Lambda(\omega) = \Gamma.$$

Remark 2.5 *We call the Variable-neighborhood random fields also Parsimonious Markov random fields. Namely* $\gamma_\Lambda(\cdot|\omega_{\Lambda^c})$ *depends only on* $\omega_{\text{sp}_\Lambda(\omega)}$ *and we do not need to inspect the whole configuration* ω_{Λ^c} *in order to decide about the configuration of spins within* Λ. *Indeed it is sufficient to inspect* $\omega_{\text{sp}_\Lambda(\omega)}$.

According to Definition 2.4 there might be a set of realizations of $\mu-$measure zero so that $|\text{sp}_\Lambda(\omega)| = \infty$. From now on we assume that *for all* $\omega \in \Omega$, $\text{sp}_\Lambda(\omega)$ is a finite set. This means that for all $\omega \in \Omega$, $\gamma_\Lambda(\cdot \mid \omega_{\Lambda^c})$ does only depend on a finite, but random neighborhood of Λ. When for some $\Gamma_0 \Subset \mathbb{Z}^d$, $\text{sp}_\Lambda(\omega) = \Gamma_0$ for all ω, then μ (respectively, X) is a Markov field with basic neighborhood Γ_0. Define the $\sigma-$algebra

$$\mathcal{F}_{\text{sp}_\Lambda} = \Big\{A \in \mathcal{F} : \forall \Gamma \subset \mathbb{Z}^d : \{\text{sp}_\Lambda = \Gamma\} \cap \text{A} \in \mathcal{F}_\Gamma\Big\}. \tag{2.3}$$

Then for all $\omega_\Lambda \in \mathcal{A}^\Lambda$, $\gamma_\Lambda(\{\omega_\Lambda\}|\cdot)$ is a measurable function with respect to $\mathcal{F}_{\text{sp}_\Lambda}$.

In analogy to the terminology used for one dimensional variable length Markov chains we can rephrase the Definition 2.4 using the concept of family of contexts. This generalizes the notion of context trees to more than one dimension.

Definition 2.6 The family of contexts associated to the specification γ *For* $\Lambda \Subset \mathbb{Z}^d$ *and* $\omega \in \Omega$ *we denote by*

$$c_\Lambda(\omega) = \omega_{\text{sp}_\Lambda(\omega)}$$

the restriction of ω *on the set* $\text{sp}_\Lambda(\omega)$. *We call* $c_\Lambda(\omega)$ *the* $\Lambda-$*context of* ω *associated to the specification* γ. *We write* $\tau^{(\Lambda)} \equiv \tau_\gamma^{(\Lambda)} = \{\text{sp}_\Lambda(\omega), \omega \in \Omega\}$ *for the family of* Λ*-contexts. Under our assumptions,*

$$\tau^{(\Lambda)} \subset \bigcup_{\Gamma \Subset \mathbb{Z}^d \setminus \Lambda} \mathcal{A}^\Gamma. \tag{2.4}$$

We use the short-hand notation $c_i(\omega)$ for $c_{\{i\}}(\omega)$ and $\text{sp}_\text{i}(\omega)$ for $\text{sp}_{\{\text{i}\}}(\omega)$, for any $i \in \mathbb{Z}^d$. We shall also write $\gamma_i(a|\omega)$ instead of $\gamma_i(\{a\}|\omega)$.

Remark 2.7 *It is immediate to verify from Definition 2.4 that the family* $\tau^{(\Lambda)}$ *has the following properties:*

- *No element of* $\tau^{(\Lambda)}$ *is restriction of any other element of* $\tau^{(\Lambda)}$*: If* η_Γ *and* $\widetilde{\eta}_{\widetilde{\Gamma}}$ *both belong to* $\tau^{(\Lambda)}$, $\Gamma \subset \widetilde{\Gamma}$ *and* $\eta_\Gamma = \widetilde{\eta}_\Gamma$, *then* $\Gamma = \widetilde{\Gamma}$.
- $\tau^{(\Lambda)}$ *defines a partition of* $\mathcal{A}^{\mathbb{Z}^d \setminus \Lambda}$, *that is, for each* $\omega \in \mathcal{A}^{\mathbb{Z}^d \setminus \Lambda}$ *there exists a unique* $\Gamma \subset \mathbb{Z}^d \setminus \Lambda$ *such that* $\omega_\Gamma \in \tau^{(\Lambda)}$.

In this way the family of local specifications associated to μ is

$$\gamma = \{\gamma_\Lambda(\cdot|c_\Lambda(\omega)), \Lambda \Subset \mathbb{Z}^d; c_\Lambda(\omega)\} \tag{2.5}$$

which leads to a more parsimonious description than the original

$$\{\gamma_\Lambda(\cdot|\omega_{\Lambda^c})), \Lambda \Subset \mathbb{Z}^d; \omega_{\Lambda^c}\}. \tag{2.6}$$

We close the section with two examples. In the first one we embed a renewal process in a Variable-neighborhood random field. This example has been suggested by Ferrari and Wyner (2003), [10]. In the second example we construct a two dimensional Variable-neighborhood random field by specifying a variable-neighborhood interaction potential.

Example 2.8 *We consider $\mathcal{A} = \{0,1\}$. Let $\{X_n : n \in \mathbb{Z}\}$ be a stationary process taking values in $\mathcal{A}$ such that the times when the process switches between 1 and 0 or 0 and 1 are independent and identically distributed random variables. They have the same distribution as the random variable T defined through*

$$\mathbb{P}[T = j] = c_1\rho_1^j + c_2\rho_2^j, \quad 0 < \rho_2 < \rho_1 < 1, \quad j \in \mathbb{N}.$$

Put $\mu = \mathbb{E}[T]$. The process $\{X_n : n \in \mathbb{Z}\}$ is thus a stationary alternating renewal process. This process can be described as one dimensional Variable-neighborhood random field. To this aim define

$$R_b(\omega) = \inf\{n > b+1 : \omega(n) \neq \omega(b+1)\}, \quad L_a(\omega) = \sup\{n < a-1 : \omega(n) \neq \omega(a-1)\}, \tag{2.7}$$

where $\omega \in \mathcal{A}^{\mathbb{Z}}$. The family of local specifications $\gamma_{\{[a,b]\}}$ indexed by $[a,b] \subset \mathbb{Z}$, is given as follows:

$$\gamma_{\{[a,b]\}}(\cdot \mid c_{[a,b]}(\omega)) = \gamma_{\{[a,b]\}}(\cdot \mid L_a(\omega) = -k, R_b(\omega) = l).$$

The context $c_{[a,b]}(\omega)$ depends only on the neighbor sites of $[a,b]$ which are all of the same type 0 or 1. In the appendix we show that for the one point specification $\gamma_{\{0\}}(\cdot \mid c_0(\omega))$ we get the following formulas. Write $\mathrm{sp}_0(\omega) = [\mathrm{L}_0(\omega), \mathrm{R}_0(\omega)] \setminus \{0\}$. Then

$$\begin{aligned}&\gamma_{\{0\}}(1 \mid L_0(\omega) = -k, R_0(\omega) = l, \omega(1) = \omega(-1) = 1)\\ &= \frac{\left(c_1\varrho_1^{l+k-1} + c_2\varrho_2^{l+k-1}\right)}{\left(c_1\varrho_1^{l+k-1} + c_2\varrho_2^{l+k-1}\right) + \left(c_1\rho_1^{k-1} + c_2\rho_2^{k-1}\right)\left(c_1\rho_1^{l-1} + c_2\rho_2^{l-1}\right)\frac{c_1\varrho_1 + c_2\varrho_2}{\left(\frac{c_1}{1-\varrho_1}\varrho_1 + \frac{c_2}{1-\varrho_2}\varrho_2\right)^2}}\end{aligned} \tag{2.8}$$

and

$$\begin{aligned}&\gamma_{\{0\}}(1 \mid L_0(\omega) = -k, R_0(\omega) = l, \omega(1) = 0, \omega(-1) = 1)\\ &= \frac{\left(c_1\varrho_1^k + c_2\varrho_2^k\right)\left(c_1\varrho_1^{l-1} + c_2\varrho_2^{l-1}\right)}{\left(c_1\varrho_1^k + c_2\varrho_2^k\right)\left(c_1\varrho_1^{l-1} + c_2\varrho_2^{l-1}\right) + \left(c_1\varrho_1^{k-1} + c_2\varrho_2^{k-1}\right)\left(c_1\varrho_1^l + c_2\varrho_2^l\right)}.\end{aligned} \tag{2.9}$$

Due to the symmetry between 0 and 1, it is clear that with formulas (2.8) and (2.9), we have completely described the one-point specification.

In this example the context $c_0(\cdot)$ is $\mathbb{P}-$almost surely finite, i.e. there exists a subset of configurations of $\mathbb{P}-$measure zero for which $|c_0(\omega)| = \infty$.

We now give an example of a Variable-neighborhood random field in dimension $d = 2$. In analogy with the procedure used in statistical mechanics we define a variable-neighborhood specification by introducing a variable-neighborhood interaction potential.

Example 2.9 *We consider $\mathcal{A} = \{-1,1\}$ and $d = 2$. In order to define the support of the variable-neighborhood interaction potential it is convenient to embed $\mathbb{Z}^2$ into $\mathbb{R}^2$. We partition $\mathbb{R}^2$ into cubes of edge 1 centered at $\mathbb{Z}^2$. We say that two cubes are connected if they have one face in common. We denote by $\mathcal{R}$ the set of all connected unions of such cubes, by R an element of $\mathcal{R}$ and by $|\partial R|$ the topological surface of R.*
We say that $\Gamma \subset \mathbb{Z}^2$ is a polygon if there exists $R \in \mathcal{R}$ so that $\Gamma = R \cap \mathbb{Z}^2$. We denote by $\partial\Gamma = \{i \in \Gamma : d(i, \partial R) \leq \frac{1}{2}\}$, where $d(i, \partial R) = \inf\{\|i - j\| : j \in \partial R\}$ and $\|.\|$ is the

maximum norm introduced at the beginning of this section. Finally, let $\hat{\Gamma}$ be the interior of Γ, $\hat{\Gamma} = \Gamma \setminus \partial\Gamma$.

We say that Γ is a simple polygon if $\partial\Gamma$ is a path in $\mathbb{Z}^2$ which does not cross itself and $\hat{\Gamma} \neq \emptyset$. Note that $\partial\Gamma$ can be the union of disjoint connected paths.
Given $\omega \in \mathcal{A}^{\mathbb{Z}^2}$ we define for each $i \in \mathbb{Z}^2$

$$\Gamma_i^1(\omega) = \cap\{\Gamma \subset \mathbb{Z}^2, \Gamma \text{ simple polygon }, i \in \hat{\Gamma}, \omega_{\partial\Gamma} = 1\ \}.$$

Note that in the above definition we do not require Γ to be finite. Γ could be equal to $\mathbb{Z}^2$ in which case $\partial\Gamma = \emptyset$. In order to get a finite interaction range, we finally set

$$\Gamma_i(\omega) = V_i(L) \cap \Gamma_i^1(\omega) \text{ and } c_i^K(\omega) = \{\omega_j : j \in \Gamma_i(\omega)\},$$

where $V_i(L) = \{j \in \mathbb{Z}^2 : \|i - j\| \leq L\}$. Hence, $\Gamma_i(\omega)$ is a finite subset of $\mathbb{Z}^2$ of diameter at most $2L$. It will be the support of the variable neighborhood interaction which we define now. Let $\{J_n, n \in \mathbb{N}\}$ be a collection of real numbers and define for any $i \in \mathbb{Z}^2$,

$$K^i(\omega) = K^i(c_i^K(\omega)) = J_{|\Gamma_i(\omega)|} \prod_{j \in \Gamma_i(\omega)} \omega(j),$$

where $|\Gamma_i(\omega)|$ is the cardinal of $\Gamma_i(\omega)$. By construction, the interaction is summable:

$$\sup_{i \in \mathbb{Z}^2} \sum_{j \in \mathbb{Z}^2} \sup_{\omega : \Gamma_j(\omega) \ni i} |K^j(\omega)| \ < \ \infty\ . \tag{2.10}$$

Denote

$$H_\Lambda(\omega_\Lambda, \omega_{\Lambda^c}) = - \sum_{\{i \in \mathbb{Z}^2 : \Lambda \cap \Gamma_i(\omega) \neq \emptyset\}} K^i(\omega).$$

The Variable-neighborhood random field μ is determined by the following family of local specifications

$$\gamma_\Lambda(\{\omega_\Lambda\} \mid \omega_{\Lambda^c}) = \frac{1}{Z^{\omega_{\Lambda^c}}} \exp\{-\beta H_\Lambda(\omega_\Lambda, \omega_{\Lambda^c})\}, \qquad \omega_\Lambda \in \mathcal{A}^\Lambda, \omega_{\Lambda^c} \in \mathcal{A}^{\Lambda_c}, \tag{2.11}$$

where

$$Z^{\omega_{\Lambda^c}} = \sum_{\omega_\Lambda \in \{-1,1\}^\Lambda} \exp\{-\beta H_\Lambda(\omega_\Lambda, \omega_{\Lambda^c})\}.$$

The family of contexts $c_\Lambda(\omega) = \omega_{sp_\Lambda(\omega)}$ associated to $\{\gamma_\Lambda\}_\Lambda$, defined in (2.11), is determined by $c_i(\omega)$ for $i \in \Lambda$, therefore by the knowledge of ω only on $\mathrm{sp_i}(\omega)$. By Definition 2.4 and by (2.11) we have that

$$\mathrm{sp_i}(\omega) = \left[\bigcup_{\mathrm{j} \in \mathbb{Z}^2} \{\Gamma_\mathrm{j}(\omega) : \mathrm{i} \in \Gamma_\mathrm{j}(\omega)\} \cup \bigcup_{\mathrm{j} \in \mathbb{Z}^2} \{\Gamma_\mathrm{j}(\omega^\mathrm{i}) : \mathrm{i} \in \Gamma_\mathrm{j}(\omega^\mathrm{i})\} \right] \setminus \{\mathrm{i}\}, \tag{2.12}$$

where $\omega^i(j) = \omega(j)$ for all $j \neq i$, $\omega^i(i) = -\omega(i)$. This formula gives the relation between the support of the context of the specification and the support of the interaction. We show in the appendix that the following identity holds:

$$\mathrm{sp_i}(\omega) = \left(\Gamma_\mathrm{i}^1(\omega) \cap \mathrm{V_i(2L)}\right) \setminus \{\mathrm{i}\}. \tag{2.13}$$

In the following, we will be interested in estimating the support of the context $\mathrm{sp}_\Lambda(\omega)$ for a given set of sites $\Lambda \Subset \mathbb{Z}^d$ and a given observation ω.

Proposition 9.1 of the appendix shows that $\gamma_\Lambda(\,\cdot \mid c_\Lambda(\omega)\,)$ can be derived from the one point specification $\gamma_i(\,\cdot \mid c_i(\omega)\,)$ and that for $\Lambda \Subset \mathbb{Z}^d$, we have

$$\mathrm{sp}_\Lambda(\omega) = \cup_{\omega_\Lambda}\left(\cup_{\mathrm{i}\in\Lambda}\mathrm{sp}_{\{\mathrm{i}\}}(\omega)\right) \setminus \Lambda. \tag{2.14}$$

Hence, in order to estimate $\mathrm{sp}_\Lambda(\omega)$, it is sufficient to estimate the context for single sites, i.e. $\mathrm{sp}_\mathrm{i}(\omega)$. To implement the estimation procedure we need translation covariant models.

For any fixed $i \in \mathbb{Z}^d$, we denote by $\tau_i : \mathbb{Z}^d \to \mathbb{Z}^d$ the $i-$shift defined by $\tau_i(j) = i + j$. This naturally induces on Ω the $i-$shift $T_i : \Omega \to \Omega$ defined by

$$(T_i\omega)(j) = \omega(\tau_i(j)) = \omega(i+j) \quad \forall j \in \mathbb{Z}^d.$$

Definition 2.10 *A Variable-neighborhood random field μ on $(\Omega, \mathcal{F})$, determined by a family of local specifications $\{\gamma_\Lambda\}_\Lambda$, is translation covariant if for all $\Lambda \Subset \mathbb{Z}^d$ and for all $\omega \in \Omega$*

$$\gamma_\Lambda(\cdot|\omega) = \gamma_{\tau_i\Lambda}(\cdot|T_i\omega), \quad i \in \mathbb{Z}^d$$

where $\tau_i\Lambda = \Lambda + i$.

In the following we will consider only translation covariant Variable-neighborhood random fields. This implies that $\gamma_i(\cdot|c_i(T_i(\omega))) = \gamma_0(\cdot|c_0(\omega))$.

3 Main Results and Estimation procedure

In Section 2 we introduced the notion of Variable-neighborhood random fields. Such a random field is completely determined by the one point specification. It would therefore be interesting to estimate $\mathrm{sp}_\mathrm{i}(\omega)$, i.e. the set of points in $\mathbb{Z}^d$ which enables to determine the value of the spin at the site i. This requires, however, to estimate too many unknown parameters. Therefore we are less ambitious and estimate the radius of the smallest ball containing $\mathrm{sp}_\mathrm{i}(\omega)$. For $\ell \geq 1$ and $i \in \mathbb{Z}^d$, define

$$V_i(\ell) = \{j \in \mathbb{Z}^d : ||i-j|| \leq \ell\} \quad \text{and} \quad V_i^0(\ell) = V_i(\ell) \setminus \{i\}. \tag{3.1}$$

We also write

$$\partial V_i(\ell) = \{j \in \mathbb{Z}^d : \|i-j\| = \ell\}.$$

Then we define the length of the context of site i by

$$l_i(\omega) = \inf\{\ell > 0 : \mathrm{sp}_\mathrm{i}(\omega) \subset \mathrm{V}_\mathrm{i}(\ell)\}. \tag{3.2}$$

Note that $l_i(\omega)$ is a stopping time with respect to the filtration $(\mathcal{G}^i_n)_n = (\mathcal{F}_{V_i(n)})_n$.

Recall that $\omega \in \Omega = \mathcal{A}^{\mathbb{Z}^d}$ stands for a generic configuration of the field. In order to distinguish between generic configurations and observed data, we will denote the observed data by σ. Our statistical inference is based on observations of the Variable-neighborhood random field μ over an increasing and absorbing sequence of finite regions $\Lambda_n \subset \mathbb{Z}^d$, i.e. $\Lambda_n \subset \Lambda_{n+1} \subset \mathbb{Z}^d$ for all n, and for all $\Lambda' \subset \mathbb{Z}^d$, there exists n such that $\Lambda' \subset \Lambda_n$.

Hence, at step n, the sample is σ_{Λ_n}, where σ_{Λ_n} is the fixed realization of μ in restriction to Λ_n. We will construct our estimators based on sites within some security region $\bar{\Lambda}_n \subset \Lambda_n$, where

$$\bar{\Lambda}_n = \{i \in \Lambda_n : V_i(k(n)) \subset \Lambda_n\} \tag{3.3}$$

with

$$k(n) = (\log|\Lambda_n|)^{\frac{1}{2d}}\,. \tag{3.4}$$

In order to estimate $l_i(\omega)$, we have to compare the neighborhood configuration of site i with the neighborhood configurations of different sites j for all $j \in \bar{\Lambda}_n$. To do so we define for any fixed $i \in \bar{\Lambda}_n$ and any $1 \leq \ell \leq k(n)$,

$$X_i^\ell(\omega) = \{X_i^\ell(\omega)(j) = \omega(i+j), \quad j : 0 < ||j|| \leq \ell\}, \tag{3.5}$$

hence $X_i^\ell(\omega)$ is the configuration around i in a box of edge ℓ. In terms of the shift operator, $X_i^\ell(\omega) = (\omega \circ \tau_i))_{V_0^0(\ell)}$, i.e. this is the restriction of $T_i\omega$ to $V_0^0(\ell)$. We stress that X_i^ℓ does not depend on $\omega(i)$, the center of the observation window, and this is important to our purposes. We shall use the short-hand notation

$$X_i^\ell(\omega) = \omega_i^\ell.$$

For any $1 \leq \ell \leq k(n)$, for any fixed configuration $\eta \in \mathcal{A}^{V_0^0(\ell)}$, let

$$N_n(\eta) = \sum_{j \in \bar{\Lambda}_n} 1_{\{X_j^\ell = \eta\}} \tag{3.6}$$

be the total number of occurrences of η within $\bar{\Lambda}_n$. Moreover, for any fixed value $a \in \mathcal{A}$, we write

$$N_n(\eta, a) = \sum_{j \in \bar{\Lambda}_n} 1_{\{X_j^\ell = \eta, X_j = a\}}. \tag{3.7}$$

In particular for the observed data σ, σ_i^ℓ is the data observed around the site i in a ball of radius ℓ, $N_n(\sigma_i^\ell)$ is the total number of occurrences of the local pattern around i within $\bar{\Lambda}_n$. By construction $N_n(\sigma_i^\ell) \geq 1$. Note that $N_n(\sigma_i^\ell, a)$ could be zero.

Let $\gamma : \mathcal{A} \times \mathcal{A}^{V_0^0(\ell)} \to [0,1]$. γ is interpreted as possible one-point specification of a hypothetical Markov random field for which the corresponding context is contained in $V_i(\ell)$. For any site i, under the hypothesis that its context is contained in $V_i(\ell)$, we define the *pseudo-likelihood of* γ as follows:

$$PL_n^{(i,\ell)}(\gamma) = \prod_{j \in \bar{\Lambda}_n,\, X_j^\ell = \sigma_i^\ell} \gamma(X_j | X_j^\ell) = \prod_{a \in \mathcal{A}} \gamma(a|\sigma_i^\ell)^{N_n(\sigma_i^\ell, a)}. \tag{3.8}$$

Maximizing (3.8) with respect to γ under the constraint

$$\sum_{a \in \mathcal{A}} \gamma(a|\sigma_i^\ell) = 1$$

gives the following estimator of the one-point specification

$$\hat{p}_n(a|\sigma_i^\ell) = \frac{N_n(\sigma_i^\ell, a)}{N_n(\sigma_i^\ell)}. \tag{3.9}$$

Analogously, we can define for any fixed configuration $\eta \in \mathcal{A}^{V_0^0(\ell)}$,

$$\hat{p}_n(a|\eta) = \begin{cases} \frac{N_n(\eta,a)}{N_n(\eta)} & \text{if } N_n(\eta) > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{3.10}$$

Remark 3.1 *Not all γ satisfying $\sum_{a\in\mathcal{A}}\gamma(a|\sigma_i^\ell)=1$ are possible one-point specifications; one point specifications have to satisfy additional conditions, which are collected in the appendix, see* (9.2)*, and which are not considered here. However, we define the pseudo-likelihood also for γ not satisfying these additional conditions.*

Thus, given the sample σ_{Λ_n}, the logarithm of the maximum pseudo-likelihood of γ is the following quantity:

$$\log MPL_n(i,\ell)=\sum_{a\in\mathcal{A}} N_n(\sigma_i^\ell,a)\log\hat{p}_n(a|\sigma_i^\ell). \tag{3.11}$$

The decision if for a given i the context has radius $\ell-1$ rather than ℓ is based on the Kullback-Leibler information. We introduce

$$\log L_n(i,\ell)=\sum_{v\in\mathcal{A}^{\partial V_0(\ell)}} N_n(\sigma_i^{\ell-1}v)D(\hat{p}_n(\cdot|\sigma_i^{\ell-1}v),\hat{p}_n(\cdot|\sigma_i^{\ell-1})), \tag{3.12}$$

where we sum over all possibilities of extending $\sigma_i^{\ell-1}$ to a configuration $\sigma_i^{\ell-1}v$ of radius ℓ and where

$$D(\hat{p}_n(\cdot|\sigma_i^{\ell-1}v),\hat{p}_n(\cdot|\sigma_i^{\ell-1}))=\sum_{a\in\mathcal{A}}\hat{p}_n(a|\sigma_i^{\ell-1}v)\log\left[\frac{\hat{p}_n(a|\sigma_i^{\ell-1}v)}{\hat{p}_n(a|\sigma_i^{\ell-1})}\right]$$

is the Kullback-Leibler information. Note that $\log L_n(i,\ell)$ is a function of $\sigma_i^{\ell-1}$, but not of σ_i^ℓ. We rewrite it as follows:

$$\log L_n(i,\ell)=\sum_{j:\,X_j^{\ell-1}=\sigma_i^{\ell-1}}\frac{1}{N_n(X_j^\ell)}\sum_{a\in\mathcal{A}}N_n(X_j^\ell,a)\log\left[\frac{\hat{p}_n(a|X_j^\ell)}{\hat{p}_n(a|X_j^{\ell-1})}\right]. \tag{3.13}$$

Finally note that

$$\log L_n(i,\ell)=\left[\sum_{j:\,X_j^{\ell-1}=\sigma_i^{\ell-1}}\frac{1}{N_n(X_j^\ell)}\log MPL_n(j,\ell)\right]-\log MPL_n(i,\ell-1). \tag{3.14}$$

For technical reasons we have to introduce the following security diameter

$$R_n=\left[\left(\log|\bar{\Lambda}_n|\right)^{\frac{1}{2d}}\right], \tag{3.15}$$

where $[\cdot]$ denotes the integer part of a number. Note however that

$$R_n/k(n)\to 1 \text{ as } n\to\infty,$$

where $k(n)$ was defined in (3.4). We are now able to define the estimator of the context length function.

Definition 3.2 The estimator *Given the observation σ_{Λ_n}, for any $i\in\bar{\Lambda}_n$, see* (3.3)*, the estimator of $l_i(\sigma)$, defined in* (3.2)*, is the following random variable*

$$\hat{l}_n(i)=\hat{l}_n(i,\sigma)=\max\{\ell=2,\ldots,R_n:\log L_n(i,\ell)>pen(\ell,n)\}, \tag{3.16}$$

whenever $\{\ell=2,\ldots,R_n:\log L_n(i,\ell)>pen(\ell,n)\}\neq\emptyset$. Otherwise we set $\hat{l}_n(i)=R_n$. In the above definition,

$$pen(\ell,n)=\kappa|\mathcal{A}||\mathcal{A}|^{|\partial V_0(\ell)|}\log|\Lambda_n|, \tag{3.17}$$

and κ is a positive constant that can be chosen freely, provided it is at least of the order given in (3.18)*.*

In other words, the above estimator chooses the minimal length ℓ such that all sites which are relevant to determine the value of the spin at site i belong to a ball of radius ℓ.

Remark 3.3 *In one dimension the above penalization term is independent of ℓ, since in this case $|\mathcal{A}|^{|\partial V_0(\ell)|} = |\mathcal{A}|^2$. This leads to a penalty term*

$$pen(n) = \kappa |\mathcal{A}|^3 \log |\Lambda_n|.$$

Once we have estimated the context length function, the underlying context $c_i(\sigma)$ is then estimated by

$$\hat{c}_{n,i}(\sigma) = \sigma_{V_i^0(\hat{l}_n(i))},$$

and the corresponding one point specification by $\hat{\gamma}_{n,i}(a|\sigma) = \hat{p}_n(a|\hat{c}_{n,i}(\sigma))$.

3.1 Main results

The estimator $\hat{l}_n(i)$ depends on the penalization term, (3.17), therefore on the choice of the constant κ. Choose $\delta > 2^d \log |\mathcal{A}| \frac{3e}{4q_{min}}$ and define

$$\kappa = \kappa(\delta) = 5^d \left(\frac{3}{2}\right)^{1/2} \delta. \tag{3.18}$$

For the estimator defined in this way the following theorems are our main results.

Assumption 3.4 *The local specification is positive. We define*

$$q_{min} = \inf_{a \in \mathcal{A}} \inf_{\omega \in \Omega} \gamma_0(a|\omega) > 0. \tag{3.19}$$

Theorem 3.5 (Overestimation) *Let μ be a translation covariant Variable-neighborhood random field for which Assumption 3.4 holds. For any $\epsilon > 0$ there exist $n_0 = n_0(\epsilon, \delta, q_{min})$ and $c(\delta) = c(\delta, q_{min})$, so that for any $n \geq n_0$ the probability of overestimation is bounded by*

$$\mu \left[\exists i \in \bar{\Lambda}_n : \hat{l}_n(i) > l_i(\sigma) \right] \leq$$
$$C(d) (\log |\bar{\Lambda}_n|)^{\frac{d+1}{2d}} \cdot \exp \left(-c(\delta) \sqrt{\log |\bar{\Lambda}_n|} \right) + C(d) \exp \left(-|\bar{\Lambda}_n|^{1-\epsilon} \right), \tag{3.20}$$

where $C(d)$ is a positive constant depending only on the dimension and where $\bar{\Lambda}_n$ is given in (3.3).

Remark 3.6 *To obtain an upper bound in* (3.20) *summable in n, we need a fast increase of the sampling regions of order for example*

$$\log |\bar{\Lambda}_n| \sim (\log n^{1+\varepsilon})^2,$$

which requires faster increase than choosing $\Lambda_n = [-\frac{n}{2}, \frac{n}{2}]^d$.

For bounding the probability of underestimation we need an additional assumption. To this aim define

$$r(i,j) = \sup_{\omega, \omega' : \omega_{\{j\}^c} = \omega'_{\{j\}^c}} \frac{1}{2} \| \gamma_i(\cdot|\omega) - \gamma_i(\cdot|\omega') \|_{TV}$$

where $\| \cdot \|_{TV}$ denotes the total variation norm. By translation covariance $r(i,j) = r(0, i-j)$. We denote

$$\beta(\ell) = \sum_{k \in \mathbb{Z}^d : \|i\| > \ell} r(0,k). \tag{3.21}$$

Assumption 3.7 *We assume that there exists $L > 0$ such that*

$$r(0, i) = 0 \quad \text{for all } \|i\| \geq L \tag{3.22}$$

and

$$r = \sum_{i \in \mathbb{Z}^d \setminus \{0\}} r(0, i) < 1. \tag{3.23}$$

Remark 3.8 *Condition (3.22) implies that the observed random field is actually a Markov random field of order L. The order L, however, is unknown. We do not propose to estimate this unknown order L. When passing to the parsimonious description (2.5), what we actually propose is to estimate, for every site i, given the observation σ, the minimal order $l_i(\sigma)$ that we need in order to determine the specification at that site, given σ. This is also called Minimum Description Length in the literature. However, if $l_i(\sigma)$ does not depend on the configuration, then our estimator naturally provides an estimator of L.*

Condition (3.23) *is the Dobrushin condition which implies uniqueness of the measure μ, see Dobrushin (1968), [6], [7].*

Theorem 3.9 (Underestimation) *Let μ be a translation covariant Variable-neighborhood random field for which Assumption 3.4 and Assumption 3.7 hold. Then for any $\epsilon > 0$ there exists $n_0 = n_0(\epsilon, \delta, q_{min}, L)$, so that for any $n \geq n_0$ the probability of underestimation is bounded by*

$$\mu\left[\exists i \in \bar{\Lambda}_n : \hat{l}_n(i) < l_i(\sigma)\right] \quad \leq \quad \exp\left(-|\bar{\Lambda}_n|^{1-\epsilon}\right). \tag{3.24}$$

Remark 3.10

1. *The above results are stated for all $n \geq n_0$ where n_0 depends on the (unknown) model parameter q_{min} and on the interaction through L. It is possible to write down upper bounds which hold for all n, but then the bounds become more complicated and depend on q_{min} and on L. We adopted the above way of writing in order to state the results in a more transparent way.*

2. *Note that the trade-off between the rates of the two kind of errors (exponential convergence for the probability of underestimation in* (3.24) *and (basically) polynomial convergence of the probability of overestimation in* (3.20)*) is a typical feature in problems of order estimation appearing already in the simpler problem of order estimation for Markov chains, see e.g. the papers by Finesso et al. (1996), [11], and Merhav et al. (1989), [19].*

   This represents the usual trade-off between type one and type two errors in statistical decision problems: Overestimation means that the estimate exceeds the true order and that we choose models that include the true data-generating mechanism. This choice is not optimal but does only lead to a higher cost. On the other hand underestimation leads to a restriction to lower order models that do not describe the observed data.

   So it is desirable to have an exponential control on the probability of underestimation while keeping some polynomial control on the probability of overestimation.

3. *The definition of our estimator depends on the parameter δ. This plays an important role only for the overestimation. Namely it appears in the exponent of the upper bound through the constant*

$$c(\delta) = \frac{2}{3}\frac{2q_{min}\delta}{e} - 2^d \log|\mathcal{A}|,$$

(see end of the proof of Lemma 5.2). To ensure the consistency of the estimator we need to choose δ sufficiently large, depending on the one-point specification and on q_{min} such that $c(\delta) > 0$. Therefore, our estimator is not universal, in the sense that for fixed δ it fails to be consistent for any random field such that $c(\delta) < 0$.

This problem appears even in the simpler case of order estimation for Markov chains, see for example Finesso et al. (1996), [11], and Merhav et al. (1989), [19]. As pointed out by Finesso et al. (1996), [11], it is not possible to have an exponential bound on the overestimation probability of an order estimator without rendering it inconsistent, for at least one model, for the underestimation.

4 Deviation inequalities for underestimation

The deviation inequalities needed for the underestimation are based on results obtained by Dedecker (2001), [5], on exponential inequalities for random fields. To adapt these results to our model we need Assumption 3.4 and Assumption 3.7. Under condition (3.22) of Assumption 3.7, the Variable-neighborhood random field is at most of order L, although the value of it is unknown to the statistician. Moreover, condition (3.23) implies that there exists a unique infinite volume Gibbs measure μ consistent with the specification γ.

In the next subsection we present a preliminary deviation inequality based on the results of Dedecker (2001), [5], and then give in the following subsection the deviation inequalities that shall be needed in order to control the probability of underestimation.

4.1 Preliminaries

Fix $\ell > 0$. For a given configuration $\eta \in \mathcal{A}^{V_0^0(\ell)}$, we define

$$p(\eta) = \mu(\{X_i = \eta_i,\ \forall\, i \in V_0^0(\ell)\}). \tag{4.1}$$

Recall that $N_n(\eta)$ is the total number of occurrences of η in the observation $\sigma_{\bar{\Lambda}_n}$. Then we get the following result which is an immediate consequence of Corollary 4 of Dedecker (2001), [5].

Proposition 4.1 *Under Assumption 3.4 and Assumption 3.7 there exists a constant $c(d, L)$ depending only on dimension and on the range L, such that for any configuration $\eta \in \mathcal{A}^{V_0^0(\ell)}$,*

$$\mu\left(\left|\frac{N_n(\eta)}{|\bar{\Lambda}_n|} - p(\eta)\right| \geq \delta\right) \leq e^{1/e} \exp\left(-\frac{c(d, L)|\bar{\Lambda}_n|\delta^2}{(2\ell)^{2d-1}e}\right). \tag{4.2}$$

The remainder of this section is devoted to show how this result can be obtained as a consequence of Corollary 4 of Dedecker (2001), [5]. We give this proof in detail since this shows at which extend Assumption 3.7 is needed.

Proof. For any i, let

$$Y_i = 1_{\{X_i^\ell = \eta\}}.$$

Then under μ, $\{Y_i : i \in \mathbb{Z}^d\}$ is a stationary random field. The associated filtration is defined as follows. For any $\Gamma \subset \mathbb{Z}^d$, let

$$\mathcal{G}^\ell{}_\Gamma = \sigma\{Y_i, i \in \Gamma\},$$

and define the $\Phi-$mixing coefficient

$$\Phi(\mathcal{G}^{\ell}{}_{\Gamma_1}, \mathcal{G}^{\ell}{}_{\Gamma_2}) = \sup\{||\mu(B|\mathcal{G}^{\ell}{}_{\Gamma_1}) - \mu(B)||_\infty, B \in \mathcal{G}^{\ell}{}_{\Gamma_2}\}.$$

Moreover, let

$$\Phi^{\ell}_{\infty,1}(n) = \sup\{\Phi(\mathcal{G}^{\ell}{}_{\Gamma_1}, \mathcal{G}^{\ell}{}_{\Gamma_2}) : |\Gamma_2| = 1, dist(\Gamma_1, \Gamma_2) \geq n\}, \tag{4.3}$$

where $dist(\Gamma_1, \Gamma_2) = \min\{\|j - i\|, i \in \Gamma_1, j \in \Gamma_2\}$. Let

$$\beta_\ell = 1 + \sum_{n \geq 1} \Phi^{\ell}_{\infty,1}(n)|\partial V_0(n)|. \tag{4.4}$$

Note that this quantity β_ℓ depends on ℓ through the filtration $\{\mathcal{G}^{\ell}_{\Gamma}, \Gamma \subset \mathbb{Z}^d\}$. To avoid confusion we warn the reader that β_ℓ defined in (4.4) is a different quantity from $\beta(\ell)$ defined in (3.21), although related.

Corollary 4 of Dedecker (2001), [5], implies the following exponential inequality

$$\mu\left(\left|\frac{N_n(\eta)}{|\bar{\Lambda}_n|} - p(\eta)\right| \geq \delta\right) \leq e^{1/e} \exp\left(-\frac{|\bar{\Lambda}_n|\delta^2}{4\beta_\ell e}\right). \tag{4.5}$$

To conclude the proof we have to estimate β_ℓ. This is done in Lemma 4.2, stated below. Then by defining $c(d, L) = \frac{1}{4C(d,L)}$, where $C(d, L)$ is the constant of Lemma 4.2, we obtain the desired result. •

Assumption 3.7 is essential for proving the following lemma.

Lemma 4.2 *Under Assumption 3.7, there exist constants $c^* = c^*(L)$ and $k = k(L)$ such that*

$$\Phi^{\ell}_{\infty,1}(n + 2\ell) \leq c^* \; |V_0^0(\ell)|e^{-kn},$$

and for β_ℓ defined in (4.4)*, we have*

$$\beta_\ell \leq C(d, L)(2\ell)^{2d-1}$$

where $C(d, L)$ is a positive constant depending on the dimension d and on L.

Proof. For any $\Gamma \Subset \mathbb{Z}^d$, let

$$\Gamma(\ell) = \{i \in \mathbb{Z}^d : d(i, \Gamma) \leq \ell\}.$$

We have that, whenever $|\Gamma| > 1$,

$$\mathcal{G}^{\ell}{}_{\Gamma} = \sigma\{Y_i, i \in \Gamma\} \subset \sigma\{X_i, i \in \Gamma(\ell)\} = \mathcal{F}_{\Gamma(\ell)}.$$

When $|\Gamma| = 1$, assuming $\Gamma = \{i\}$,

$$\mathcal{G}^{\ell}{}_{\Gamma} = \sigma\{Y_i\} \subset \sigma\{X_j, j \in \Gamma(\ell) \setminus \{i\}\} = \sigma\{X_j, j \in V_i^0(\ell)\}.$$

By translational covariance, if is sufficient to set $\Gamma_2 = \{0\}$, $|\Gamma_1| = \infty$ and $dist(\Gamma_1, \Gamma_2) \geq n+2\ell$, for $n \geq 1$. Now, take $B = \{X_0^\ell = \eta\}$ for some fixed $\eta \in \mathcal{A}^{V_0^0(\ell)}$. Since $\mathcal{G}^{\ell}_{\Gamma_1} \subset \mathcal{F}_{V_0(n+\ell)^c}$ and $\mu(B|\mathcal{G}^{\ell}_{\Gamma_1}) = \mu(\mu(B|\mathcal{F}_{V_0(n+\ell)^c})|\mathcal{G}^{\ell}_{\Gamma_1})$, in order to bound $\Phi^{\ell}_{\infty,1}(n + 2\ell)$, it is sufficient to bound

$$||\mu(B|\mathcal{G}^{\ell}_{\Gamma_1}) - \mu(B)||_\infty \leq ||\mu(B|\mathcal{F}_{V_0(n+\ell)^c}) - \mu(B)||_\infty.$$

But

$$\mu(B) = \mu(\mu(B|\mathcal{F}_{V_0(n+\ell)^c})).$$

Hence, using the specification γ defined in (2.1) and (2.2), by definition (4.3) we have

$$\begin{aligned}
&\Phi^{\ell}_{\infty,1}(n+2\ell)\\
&\le \sup_{\omega}\left\{\int d\mu(\omega')\left[|\gamma_{V_0(n+\ell)}(B|\omega)-\gamma_{V_0(n+\ell)}(B|\omega')|\right]\right\}\\
&\le \sup_{\omega,\omega'}\left[|\gamma_{V_0(n+\ell)}(B|\omega)-\gamma_{V_0(n+\ell)}(B|\omega')|\right]\\
&\le \sup_{\omega(V_0^0(\ell)),\omega(V_0(n+\ell)^c),\omega'(V_0(n+\ell)^c)} |\gamma_{V_0(n+\ell)}(\omega(V_0^0(\ell))|\omega)-\gamma_{V_0(n+\ell)}(\omega(V_0^0(\ell))|\omega')|.
\end{aligned} \tag{4.6}$$

To control this last term Assumption 3.7 is essential. Indeed, we need to show that, uniformly on boundary conditions outside $V_0(n+\ell)$, (4.6) is exponentially small in n. Applying Theorem 3.1.3.2 of Presutti (2009), [20], we obtain the following. There exists a function $u_{V_0(n+\ell)}:\mathbb{Z}^d\to\mathbb{R}_+$ such that

$$\sup_{\omega(V_0^0(\ell)),\omega(V_0(n+\ell)^c),\omega'(V_0(n+\ell)^c)} |\gamma_{V_0(n+\ell)}(\omega(V_0^0(\ell))|\omega)-\gamma_{V_0(n+\ell)}(\omega(V_0^0(\ell))|\omega')| \le \sum_{i\in V_0^0(\ell)} u_{V_0(n+\ell)}(i). \tag{4.7}$$

Moreover by Corollary 3.2.5.5. of Presutti (2009), [20], under (3.22), there exist $c^*=c^*(L)$ and $k=k(L)$ so that

$$u_{V_0(n+\ell)}(i)\le c^* e^{-kd(i,V_0(n+\ell)^c)},\, i\in V_0^0(\ell). \tag{4.8}$$

Therefore, we have

$$\Phi^{\ell}_{\infty,1}(n+2\ell)\le c^* e^{-kn}|V_0^0(\ell)|,$$

and thus

$$\begin{aligned}
\beta_\ell &\le |V_0^0(2\ell)|+\sum_{n\ge 2\ell+1}|\partial V_0(n)|\Phi^{\ell}_{\infty,1}(n)\\
&\le (4\ell)^d + c^*(4\ell)^d\sum_{n\ge 2\ell+1} n^{d-1}e^{-k(n-2\ell))}.
\end{aligned} \tag{4.9}$$

Hence we may conclude that

$$\beta_\ell\le C(d,L)(2\ell)^{2d-1},$$

where $C(d,L)$ is a constant depending on the dimension d and on the range of interaction L. •

Remark 4.3 *In Proposition 4.1 we obtain an exponential rate of convergence in the ergodic theorem. It is very likely that to our purposes polynomial or sub-exponential rates of convergence will be enough. This would allow to get the control for the probability of underestimation also in the regime of phase transition. This lies, however, outside the scope of the present paper.*

4.2 Deviation inequalities

We are now able to state the deviation inequalities needed to control the probability of underestimation. They are consequences of Proposition 4.1 and follow ideas of Galves and Leonardi (2008), [13]. Before doing so, we define for any $a\in\mathcal{A}$, $\eta\in\mathcal{A}^{V_0^0(\ell)}$,

$$p(a|\eta)=\frac{p((a,\eta))}{p(\eta)}=\frac{\mu(\{X_0=a, X_i=\eta_i,\ \forall\, i\in V_0^0(\ell)\})}{\mu(\{X_i=\eta_i,\ \forall\, i\in V_0^0(\ell)\})}. \tag{4.10}$$

By Assumption 3.4 we have that for any given configuration $\eta \in \mathcal{A}^{V_0^0(\ell)}$,

$$p(\eta) \geq q_{min}^{(2l)^d},$$

and

$$p(a|\eta) \geq q_{min}.$$

We are interested in configurations having support in a ball of radius at most L. Hence, writing

$$\alpha_0 = \inf_{\ell \leq L} \inf_{a \in \mathcal{A}, \eta \in \mathcal{A}^{V_0^0(\ell)}} \{p(a|\eta), p(\eta)\}, \tag{4.11}$$

we obtain that

$$\alpha_0 \geq q_{min}^{(2L)^d}. \tag{4.12}$$

We define the following quantity

$$\Delta_n(\eta) = \sum_{a \in \mathcal{A}} \left(\frac{N_n(\eta, a)}{|\bar{\Lambda}_n|} \log \hat{p}_n(a|\eta) - p((\eta, a)) \log p(a|\eta) \right), \tag{4.13}$$

where $\hat{p}_n(a|\eta)$ is the quantity defined in (3.10). We obtain the following deviation inequalities.

Corollary 4.4 *Let μ be a translation covariant Variable-neighborhood random field for which Assumption 3.4 and Assumption 3.7 hold. Let $t > 0$, $\ell \leq L$ where L is given in* (3.22)*, let $\eta \in \mathcal{A}^{V_0^0(\ell)}$, $\hat{p}_n(\cdot|\eta)$ defined in (3.10), $p(\cdot|\eta)$ in* (4.10) *and let $\Delta_n(\eta)$ as defined in* (4.13)*. Then there exists a constant $C(d, L)$ depending only on dimension and on L such that*

$$\mu\left(|\hat{p}_n(a|\eta) - p(a|\eta)| > t\right) \leq 2e^{1/e} exp\left(-C(d, L)\frac{|\bar{\Lambda}_n| t^2 \alpha_0}{4e}\right), \quad \forall a \in \mathcal{A}, \tag{4.14}$$

$$\mu\left(|\Delta_n(\eta)| > t\right) \leq 3|\mathcal{A}|e^{1/e} \exp\left(-C(d, L)\frac{|\bar{\Lambda}_n|(t \wedge t^2)\alpha_0^2}{8|\mathcal{A}|^2(\log^2 \alpha_0 \vee 1)e}\right), \tag{4.15}$$

where α_0 is given in (4.11) *and estimated in* (4.12).

Proof. Concerning (4.14) we obtain by inserting and subtracting the term $\frac{N_n(\eta,a)}{|\bar{\Lambda}_n| p(\eta)}$,

$$\begin{aligned} &|\hat{p}_n(a|\eta) - p(a|\eta)| \\ &\leq \left| \frac{N_n(\eta, a)}{N_n(\eta)} - \frac{N_n(\eta, a)}{|\bar{\Lambda}_n| p(\eta)} \right| + \left| \frac{1}{p(\eta)} \left(\frac{N_n(\eta, a)}{|\bar{\Lambda}_n|} - p((a, \eta)) \right) \right|. \end{aligned}$$

The first term in the last expression can be upper bounded by

$$\begin{aligned} \left| \frac{N_n(\eta, a)}{N_n(\eta)} - \frac{N_n(\eta, a)}{|\bar{\Lambda}_n| p(\eta)} \right| &= N_n(\eta, a) \left| \frac{|\bar{\Lambda}_n| p(\eta) - N_n(\eta)}{N_n(\eta) |\bar{\Lambda}_n| p(\eta)} \right| \\ = \frac{N_n(\eta, a)}{N_n(\eta)} \left| \frac{p(\eta) - \frac{N_n(\eta)}{|\bar{\Lambda}_n|}}{p(\eta)} \right| &\leq \left| \frac{p(\eta) - \frac{N_n(\eta)}{|\bar{\Lambda}_n|}}{p(\eta)} \right|. \end{aligned}$$

As a consequence we obtain that

$$\begin{aligned} &\mu\left(|\hat{p}_n(a|\eta) - p(a|\eta)| > t\right) \\ &\qquad \leq \mu\left(\left| p(\eta) - \frac{N_n(\eta)}{|\bar{\Lambda}_n|} \right| > \frac{t}{2} p(\eta) \right) \\ &\qquad\qquad + \mu\left(\left| \left(\frac{N_n(\eta, a)}{|\bar{\Lambda}_n|} - p((\eta, a)) \right) \right| > \frac{t}{2} p(\eta) \right). \end{aligned}$$

Then, applying (4.2), we get

$$\mu\left(|\hat{p}_n(a|\eta)-p(a|\eta)|>t\right)\leq 2e^{\frac{1}{e}}exp\left(-\frac{c(d,L)|\bar{\Lambda}_n|t^2p(\eta)^2}{4(2\ell)^{2d-1}e}\right)\leq 2e^{\frac{1}{e}}exp\left(-\frac{c(d,L)|\bar{\Lambda}_n|t^2\alpha_0^2}{4(2L)^{2d-1}e}\right).$$

Hence, writing $C(d,L)=c(d,L)/(2L)^{2d-1}$, where $c(d,L)$ is the constant of (4.2), assertion (4.14) follows.

To show (4.15) we subtract and add the term $\frac{N_n(\eta,a)}{|\bar{\Lambda}_n|}\log p(a|\eta)$ to $\Delta_n(\eta)$. We obtain

$$\begin{aligned}\Delta_n(\eta) &= \sum_{a\in\mathcal{A}}\frac{N_n(\eta,a)}{|\bar{\Lambda}_n|}\log\frac{\hat{p}_n(a|\eta)}{p(a|\eta)}\\ &\quad+\sum_{a\in\mathcal{A}}\left(\frac{N_n(\eta,a)}{|\bar{\Lambda}_n|}-p((\eta,a))\right)\log p(a|\eta)\\ &= \Delta_n^1(\eta)+\Delta_n^2(\eta).\end{aligned}$$

We rewrite $\Delta_n^1(\eta)$ in the following way and then apply the estimate (6.4):

$$\begin{aligned}\Delta_n^1(\eta) &= \frac{N_n(\eta)}{|\bar{\Lambda}_n|}\sum_{a\in\mathcal{A}}\hat{p}_n(a|\eta)\log\frac{\hat{p}_n(a|\eta)}{p(a|\eta)}\\ &\leq \frac{N_n(\eta)}{|\bar{\Lambda}_n|}\sum_{a\in\mathcal{A}}\frac{\left(\hat{p}_n(a|\eta)-p(a|\eta)\right)^2}{p(a|\eta)}\\ &\leq \sum_{a\in\mathcal{A}}\frac{\left(\hat{p}_n(a|\eta)-p(a|\eta)\right)^2}{p(a|\eta)}\\ &\leq \sum_{a\in\mathcal{A}}\frac{\left(\hat{p}_n(a|\eta)-p(a|\eta)\right)^2}{\alpha_0}.\end{aligned}$$

Therefore

$$\begin{aligned}&\mu\left(|\Delta_n^1(\eta)|>\frac{t}{2}\right)\\ &\leq \sum_{a\in\mathcal{A}}\mu\left(\left(\hat{p}_n(a|\eta)-p(a|\eta)\right)^2>\frac{1}{|\mathcal{A}|}\frac{t}{2}\alpha_0\right)\\ &\leq 2|\mathcal{A}|e^{1/e}\exp\left(-C(d,L)\frac{|\bar{\Lambda}_n|t\ \alpha_0^2}{8|\mathcal{A}|e}\right),\end{aligned}\tag{4.16}$$

by (4.14). We get for the second term

$$\begin{aligned}&\mu\left(|\Delta_n^2(\eta)|>\frac{t}{2}\right)\\ &\leq \sum_{a\in\mathcal{A}}\mu\left(\left|\frac{N_n(\eta,a)}{|\bar{\Lambda}_n|}-p((\eta,a))\right|>\frac{1}{|\mathcal{A}|}\frac{t}{2}\frac{1}{|\log\alpha_0|}\right)\\ &\leq |\mathcal{A}|e^{1/e}\exp\left(-C(d,L)\frac{|\bar{\Lambda}_n|t^2}{4|\mathcal{A}|^2\log^2\alpha_0 e}\right),\end{aligned}\tag{4.17}$$

by (4.2). This finishes the proof. •

5 Deviation inequalities for overestimation

In order to control the probability of overestimation we do not need as strong assumptions as for the control of the probability of underestimation. Indeed, we can avoid to impose Assumption 3.7. We mimic the method used by Csiszàr and Talata (2006), [4], see Proposition 3.1 and Lemma 3.3 of their paper. Their results are typicality results and they obtain the almost sure convergence of the empirical probabilities to the theoretical ones. We follow the way indicated by Csiszàr and Talata (2006), [4], but we quantify the errors and obtain in this way precise deviation inequalities. We will need only Assumption 3.4.

We partition the region $\bar{\Lambda}_n$ by intersecting it with a sub lattice of $\mathbb{Z}^d$ such that the distance between sites in the sub lattice is $4R_n + 1$. More precisely, let

$$\bar{\Lambda}_n^k = \{j \in \bar{\Lambda}_n,\ j = k + (4R_n + 1)l, l \in \mathbb{Z}^d\}, \quad \|k\| \leq 2R_n.$$

For any $\ell \leq R_n$ and any fixed configuration $\eta \in \mathcal{A}^{V_0^0(\ell)}$, let

$$N_n^k(\eta) = \sum_{j \in \bar{\Lambda}_n^k} 1_{\{X_j^\ell = \eta\}}$$

be the number of occurrences of η in the sample having center in $\bar{\Lambda}_n^k$. In the same way we denote

$$N_n^k(\eta, a) = \sum_{j \in \bar{\Lambda}_n^k} 1_{\{X_j^\ell = \eta, X_j = a\}}.$$

Note that we have

$$N_n(\eta) = \sum_{k : \|k\| \leq 2R_n} N_n^k(\eta),\ N_n(\eta, a) = \sum_{k : \|k\| \leq 2R_n} N_n^k(\eta, a).$$

Let

$$\mathcal{A}(n, \ell, k) = \{\frac{3}{2} \log N_n^k(\eta) \geq \log |\bar{\Lambda}_n|,\ \text{for all } \eta \in \mathcal{A}^{V_0^0(\ell)} \text{ s.t. } \ell \geq l_0(\eta)\} \tag{5.1}$$

and

$$\mathcal{B}(n, \ell) = \bigcap_{k : \|k\| \leq 2R_n} \mathcal{A}(n, \ell, k). \tag{5.2}$$

The probabilities $\mu(\mathcal{A}(n, \ell, k))$ and $\mu(\mathcal{B}(n, \ell))$ can be immediately obtained by Lemma 5.3 given at the end of this section. Recall the definition of $\hat{p}_n$ in (3.9).

Theorem 5.1 *For any*

$$\delta > 2^d \log |\mathcal{A}| \frac{3e}{4q_{min}}, \tag{5.3}$$

there exist a positive constant $c(\delta) = c(\delta, q_{\min})$ and n_0 (not depending on q_{min} nor on δ) such that for all $n \geq n_0$, for any $\ell \leq R_n$,

$$\mu\left[\ \exists\ \eta \in \mathcal{A}^{V_0^0(\ell)}, \ell \geq l_0(\eta) : \left|\hat{p}_n(a|\eta) - \gamma_{\{0\}}(a|\eta)\right| > \sqrt{\kappa(\delta)\gamma_{\{0\}}(a|\eta)\frac{\log N_n(\eta)}{N_n(\eta)}}\ ,\ \mathcal{B}(n, \ell)\right]$$
$$\leq 4\,(4R_n + 1)^d \exp\left(-c(\delta)\sqrt{\log |\bar{\Lambda}_n|}\right), \tag{5.4}$$

where $\kappa(\delta) > 0$ is as in (3.18).

The main ingredient to prove Theorem 5.1 is the following lemma.

Lemma 5.2 *For any δ as in (5.3) there exist n_0 (not depending on q_{min} nor on δ) and a positive constant $c(\delta) = c(\delta, q_{\min})$, such that for all $n \geq n_0$, for any $\ell \leq R_n$,*

$$\mu\Bigg[\exists\, \eta \in \mathcal{A}^{V_0^0(\ell)}, \ell \geq l_0(\eta) : \left|\frac{N_n^k(\eta,a)}{N_n^k(\eta)} - \gamma_{\{0\}}(a|\eta)\right| \geq \sqrt{\delta\gamma_{\{0\}}(a|\eta)\frac{(\log N_n^k(\eta))^{\frac{1}{2}}}{N_n^k(\eta)}}, \mathcal{A}(n,\ell,k)\Bigg]$$
$$\leq 4\exp\left(-c(\delta)\sqrt{\log|\bar{\Lambda}_n|}\right). \tag{5.5}$$

Proof. Fix $\eta \in \mathcal{A}^{V_0^0(\ell)}$ with $\ell \geq l_0(\eta)$ and set $\gamma(a) = \gamma_{\{0\}}(a|\eta)$. Recall that $\gamma(a) \geq q_{min}$. We first provide an upper bound for fixed η of

$$\mu\left[\left|N_n^k(\eta,a) - N_n^k(\eta)\gamma(a)\right| \geq \sqrt{\delta\gamma(a)N_n^k(\eta)(\log N_n^k(\eta))^{1/2}}\,, \mathcal{A}(n,\ell,k)\right].$$

By definition

$$N_n^k(\eta,a) - N_n^k(\eta)\gamma_{\{0\}}(a|\eta) = \sum_{j\in\bar{\Lambda}_n^k} 1_{\{X_j^\ell=\eta\}}\left[1_{\{X_j=a\}} - \gamma_{\{0\}}(a|\eta)\right]. \tag{5.6}$$

We order in some arbitrary way the points

$$\{j \in \bar{\Lambda}_n^k, X_j^\ell = \eta\} = \{j_l, 1 \leq l \leq N_n^k(\eta)\}.$$

Define

$$Z_l = \left[1_{\{X_{j_l}=a\}} - \gamma_{\{0\}}(a|\eta)\right], \quad l = 1,\ldots N_n^k(\eta).$$

The random variables $\{Z_l,\ l = 1,\ldots N_n^k(\eta)\}$ are identically distributed random variables, having mean zero; which are conditionally independent, in the sense that for $i \neq j$, $0 \leq |z_i| \leq 1$, $0 \leq |z_j| \leq 1$

$$\mu\left[Z_i = z_i, Z_j = z_j|\omega(\bar{\Lambda}_n \setminus \cup_{j\in\bar{\Lambda}_n^k} V_j(\ell))\right] = \mu\left[Z_i = z_i|\omega(\bar{\Lambda}_n \setminus \cup_{j\in\bar{\Lambda}_n^k} V_j(\ell))\right]$$
$$\cdot\mu\left[Z_j = z_j|\omega(\bar{\Lambda}_n \setminus \cup_{j\in\bar{\Lambda}_n^k} V_j(\ell))\right].$$

Take an independent copy $\{Z'_i, i \geq 1\}$ of i.i.d. random variables, having the same distribution as Z_1, independent of X. Then for $i > N_n^k(\eta)$ we let $Z_i = Z'_{i-N_n^k(\eta)}$. The important point of this definition is that in this way, the sequence of random variables $Z_1, Z_2, \ldots$ is independent of $N_n^k(\eta)$. Define partial sums

$$S_N = \sum_{j=1}^N Z_j, \quad S_N^* = \max\{S_j; j \leq N\}.$$

These are still independent of $N_n^k(\eta)$. We write the quantity in (5.6) as

$$N_n^k(\eta,a) - N_n^k(\eta)\gamma_{\{0\}}(a|\eta) = S_{N_n^k(\eta)} \leq S^*_{N_n^k(\eta)}. \tag{5.7}$$

We now use arguments similar to those in the proof of Lemma 3.3 of Csiszàr and Talata (2006), [4]. In the following,

$$\tilde{\mu} = \mu(\cdot|\omega(\bar{\Lambda}_n \setminus \cup_{j\in\bar{\Lambda}_n^k} V_j(\ell)))$$

denotes always conditional probability when conditioning with respect to $\omega(\bar{\Lambda}_n \setminus \cup_{j\in\bar{\Lambda}_n^k} V_j(\ell))$. Then,

$$\tilde{\mu}\left[S^*_{N_n^k(\eta)} \geq \sqrt{\delta\gamma(a)N_n^k(\eta)(\log N_n^k(\eta))^{1/2}} \quad, \mathcal{A}(n,\ell,k)\right] \leq \tag{5.8}$$
$$\sum_{j\in\mathbb{N}} \tilde{\mu}\left[S^*_{N_n^k(\eta)} \geq \sqrt{\delta\gamma(a)N_n^k(\eta)(\log N_n^k(\eta))^{1/2}}; e^j < N_n^k(\eta) \leq e^{j+1}, \mathcal{A}(n,\ell,k)\right].$$

Note that on $\mathcal{A}(n,\ell,k) \cap \{e^j < N_n^k(\eta) \leq e^{j+1}\}$, see (5.1), since $\log N_n^k(\eta) \leq \log|\bar{\Lambda}_n|$,

$$j < \log|\bar{\Lambda}_n| \leq \frac{3}{2}(j+1).$$

Hence by independence of $\{S^*_N, N \geq 1\}$ and $N_n^k(\eta)$, the last expression of (5.8) can be bounded from above as follows.

$$\sum_{j:j<\log|\bar{\Lambda}_n|\leq\frac{3}{2}(j+1)} \tilde{\mu}\left[S^*_{e^{j+1}} \geq \sqrt{\delta\gamma(a)e^j\sqrt{j}}\right] \leq \sum_{j:\log|\bar{\Lambda}_n|\leq\frac{3}{2}(j+1)} \tilde{\mu}\left[S^*_{e^{j+1}} \geq \sqrt{\delta\gamma(a)e^j\sqrt{j}}\right]. \tag{5.9}$$

Now, Bernstein's inequality, see Lemma 9.2, yields

$$\tilde{\mu}\left[S^*_N \geq c\right] \leq 2\exp\left(-\frac{2c^2}{N}\right).$$

This gives

$$\tilde{\mu}\left[S^*_{e^{j+1}} \geq \sqrt{\delta\gamma(a)e^j\sqrt{j}}\right] \leq 2\exp\left(-\frac{2q_{min}}{e}\delta\sqrt{j}\right).$$

Taking in account that

$$\int_{\sqrt{a}}^{\infty} e^{-by}y\,dy = \frac{1}{b}e^{-b\sqrt{a}}\sqrt{a} + \frac{1}{b^2}e^{-b\sqrt{a}},$$

setting $b = 2q_{min}\delta/e$ and $a = \frac{2}{3}\log|\bar{\Lambda}_n| - 1$, one can upper bound the sum over j in (5.9) obtaining

$$\tilde{\mu}\left[S^*_{N_n^k(\eta)} \geq \sqrt{\delta\gamma(a)N_n^k(\eta)(\log N_n^k(\eta))^{1/2}} \quad, \mathcal{A}(n,\ell,k)\right]$$
$$\leq \frac{2e}{\delta q_{min}}\left(\sqrt{\frac{2}{3}\log|\bar{\Lambda}_n| - 1} + \frac{e}{2\delta q_{min}}\right)\exp\left(-\frac{2q_{min}\delta}{e}\sqrt{\frac{2}{3}\log|\bar{\Lambda}_n| - 1}\right)$$
$$\leq 4\left(\sqrt{\frac{2}{3}\log|\bar{\Lambda}_n| - 1} + 1\right)\exp\left(-\frac{2q_{min}\delta}{e}\sqrt{\frac{2}{3}\log|\bar{\Lambda}_n| - 1}\right),$$

since by the choice of δ in (5.3), $\frac{e}{2\delta q_{min}} \leq 1$.

Now, there exists n_0 (not depending on q_{min} nor on δ) such that for all $n \geq n_0$, this last upper bound can be replaced by

$$\left(\sqrt{\frac{2}{3}\log|\bar{\Lambda}_n| - 1} + 1\right)\exp\left(-\frac{2q_{min}\delta}{e}\sqrt{\frac{2}{3}\log|\bar{\Lambda}_n| - 1}\right) \leq \exp\left(-\frac{2}{3}\frac{2q_{min}\delta}{e}\sqrt{\log|\bar{\Lambda}_n|}\right).$$

This upper bound holds also for the non-conditioned probability μ. Finally, in order to get the result uniformly over all possible configurations η having $l_0(\eta) \le \ell$, we need to sum over all possible choices of patterns η. This gives, by definition of R_n,

$$|\mathcal{A}|^{|V_0^0(\ell)|} = |\mathcal{A}|^{(2\ell)^d} \le |\mathcal{A}|^{(2R_n)^d} = e^{2^d \log|\mathcal{A}|\sqrt{\log|\bar{\Lambda}_n|}}$$

terms. Thus we can conclude that for all $n \ge n_0$, taking δ as in (5.3) we have

$$\begin{aligned}
&\mu\left[\ \exists\ \eta \in \mathcal{A}^{V_0^0(\ell)}, \ell \ge l_0(\eta) : \left|\frac{N_n^k(\eta,a)}{N_n^k(\eta)} - \gamma_{\{0\}}(a|\eta)\right| \ge \sqrt{\delta\gamma_{\{0\}}(a|\eta)\frac{(\log N_n^k(\eta))^{\frac{1}{2}}}{N_n^k(\eta)}}\right.\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\left., \mathcal{A}(n,\ell,k)\right]\\
&\le 4e^{2^d\log|\mathcal{A}|\sqrt{\log|\bar{\Lambda}_n|}}\exp\left(-\frac{2}{3}\frac{2q_{min}\delta}{e}\sqrt{\log|\bar{\Lambda}_n|}\right)\\
&= 4\exp\left(-c(\delta)\sqrt{\log|\bar{\Lambda}_n|}\right),
\end{aligned}$$

where $c(\delta) = \frac{2}{3}\frac{2q_{min}\delta}{e} - 2^d\log|\mathcal{A}| > 0$. This concludes the proof. •

We are now able to give the proof of Theorem 5.1.

Proof of Theorem 5.1 Fix $\eta \in \mathcal{A}^{V_0^0(\ell)}$ with $\ell \ge l_0(\eta)$, let $\gamma(a) = \gamma_{\{0\}}(a|\eta)$, δ as in (5.3) and denote by

$$E_n(\eta) = \bigcap_{k:\|k\|\le 2R_n}\left\{\left|\frac{N_n^k(\eta,a)}{N_n^k(\eta)} - \gamma(a)\right| \le \sqrt{\delta\gamma(a)[N_n^k(\eta)]^{-1}\sqrt{\log N_n^k(\eta)}}\right\}.$$

Then on $E_n(\eta)$, using Jensen's inequality, the definition of R_n and $N_n^k(\eta) \le N_n(\eta)$,

$$\begin{aligned}
&\left|\hat{p}_n(a|\eta) - \gamma_{\{0\}}(a|\eta)\right|\\
&\le \sum_{k:\|k\|\le 2R_n}\left|\frac{N_n^k(\eta,a)}{N_n^k(\eta)} - \gamma_{\{0\}}(a|\eta)\right|\cdot\frac{N_n^k(\eta)}{N_n(\eta)}\\
&\le \sum_{k:\|k\|\le 2R_n}\sqrt{\delta\gamma_{\{0\}}(a|\eta)\frac{\sqrt{\log N_n^k(\eta)}}{N_n^k(\eta)}}\cdot\frac{N_n^k(\eta)}{N_n(\eta)}\\
&\le \sqrt{\sum_{k:\|k\|\le 2R_n}\delta\gamma_{\{0\}}(a|\eta)\frac{\sqrt{\log N_n^k(\eta)}}{N_n(\eta)}}\\
&\le \frac{(4R_n+1)^{d/2}\delta^{1/2}\gamma_{\{0\}}(a|\eta)^{1/2}[\log N_n(\eta)]^{1/4}}{[N_n(\eta)]^{1/2}}\\
&\le 5^{d/2}\left(\log|\bar{\Lambda}_n|\right)^{\frac{1}{4}}\delta^{1/2}\gamma_{\{0\}}(a|\eta)^{1/2}\frac{[\log N_n(\eta)]^{1/4}}{[N_n(\eta)]^{1/2}}.
\end{aligned}\tag{5.10}$$

On $\{\log|\bar{\Lambda}_n| \le \frac{3}{2}\log N_n(\eta)\}$, this last expression can be bounded from above by

$$5^{d/2}(\frac{3}{2})^{1/4}\sqrt{\delta\gamma_{\{0\}}(a|\eta)\frac{\log N_n(\eta)}{N_n(\eta)}} = \sqrt{\kappa(\delta)\gamma_{\{0\}}(a|\eta)\frac{\log N_n(\eta)}{N_n(\eta)}},$$

where $\kappa(\delta)$ is chosen as in (3.18). Hence we get

$$\mu\left[\exists\, \eta \in \mathcal{A}^{V_0^0(\ell)}, \ell \geq l_0(\eta) : \left|\hat{p}_n(a|\eta) - \gamma_{\{0\}}(a|\eta)\right| > \sqrt{\kappa(\delta)\gamma_{\{0\}}(a|\eta)\frac{\log N_n(\eta)}{N_n(\eta)}}\,, \mathcal{B}(n,\ell)\right]$$

$$\leq \mu\left[\bigcup_{\eta\in\mathcal{A}^{V_0^0(\ell)}:\ \ell\geq l_0(\eta)} E_n(\eta)^c\,, \mathcal{B}(n,\ell)\right]. \tag{5.11}$$

But

$$\bigcup_{\eta\in\mathcal{A}^{V_0^0(\ell)}:\ \ell\geq l_0(\eta)} E_n(\eta)^c$$

$$= \bigcup_{k:\|k\|\leq 2R_n}\ \bigcup_{\eta\in\mathcal{A}^{V_0^0(\ell)}:\ \ell\geq l_0(\eta)} \left\{\left|\frac{N_n^k(\eta,a)}{N_n^k(\eta)} - \gamma(a)\right| > \sqrt{\delta\gamma(a)\frac{\sqrt{\log N_n^k(\eta)}}{N_n^k(\eta)}}\right\},$$

therefore applying Lemma 5.2 we can finally upper bound

$$\mu\left[\bigcup_{\eta:\ \ell\geq l_0(\eta)} E_n(\eta)^c\,, \mathcal{B}(n,\ell)\right] \leq 4(4R_n+1)^d \exp\left(-c(\delta)\sqrt{\log|\bar{\Lambda}_n|}\right),$$

for all $n \geq n_0$. This finishes the proof. •

The following lemma gives conditions ensuring that $\mu(\mathcal{B}(n,\ell)^c)$ converges to 0 by giving the precise rate of convergence.

Lemma 5.3 *For any $0 < \epsilon_1 < 1$, $0 < \epsilon_2 < 1$, and for any positive C_1 and C_2 there exists $n_0 = n_0(q_{min}, \min(\epsilon_1,\epsilon_2), \min(C_1,C_2))$ so that for $n \geq n_0$ and for any $\ell \leq R_n$, we have*

$$\mu\left(\exists\, \eta \in \mathcal{A}^{V_0^0(\ell)}, \ell \geq l_0(\eta) : N_n^k(\eta) < C_1|\bar{\Lambda}_n|^{1-\epsilon_1}\right) \leq \exp\left(-C_2|\bar{\Lambda}_n|^{1-\epsilon_2}\right). \tag{5.12}$$

Proof. Fix some η with $l_0(\eta) \leq \ell \leq R_n$. Then $\{1_\eta(X_j^\ell),\ j \in \bar{\Lambda}_n^k\}$ is a collection of conditional independent random variables, conditioned on fixing the configuration $\omega(\bar{\Lambda}_n \setminus \cup_{j\in\bar{\Lambda}_n^k} V_j(\ell))$. By Assumption 3.4, we have that

$$\mu(X_j^\ell = \eta) \geq q_{min}^{(2\ell)^d}.$$

Here we have used that $|V_0^0(\ell)| = (2\ell)^d$. Then a conditional version of the Hoeffding inequality, see for example Lemma A3 in Csiszàr and Talata (2006), [4], yields

$$\mu\left[\frac{N_n^k(\eta)}{|\bar{\Lambda}_n^k|} < \frac{1}{2}q_{min}^{(2\ell)^d}\Big|\omega(\bar{\Lambda}_n \setminus \cup_{j\in\bar{\Lambda}_n^k} V_j(\ell))\right] \leq e^{-|\bar{\Lambda}_n^k|\frac{q_{min}^{(2\ell)^d}}{16}}. \tag{5.13}$$

As a consequence, we obtain also for the unconditioned probability,

$$\mu\left[\frac{N_n^k(\eta)}{|\bar{\Lambda}_n^k|} < \frac{1}{2}q_{min}^{(2\ell)^d}\right] \leq e^{-|\bar{\Lambda}_n^k|\frac{q_{min}^{(2\ell)^d}}{16}}, \tag{5.14}$$

and thus

$$\mu\left[\exists\,\eta \in \mathcal{A}^{V_0^0(\ell)},\ \ell \geq l_0(\eta)\ :\ \frac{N_n^k(\eta)}{|\bar{\Lambda}_n^k|} < \frac{1}{2}q_{min}^{(2\ell)^d}\right] \leq |\mathcal{A}|^{(2R_n)^d} e^{-|\bar{\Lambda}_n^k|\frac{q_{min}^{(2\ell)^d}}{16}}. \tag{5.15}$$

To obtain (5.12) we need to compare $|\bar{\Lambda}_n|$ to $|\bar{\Lambda}_n^k|$. By construction we have for n sufficiently large,

$$|\bar{\Lambda}_n^k| \geq \frac{|\bar{\Lambda}_n|}{(4R_n+1)^d} \geq \frac{|\bar{\Lambda}_n|}{(5R_n)^d}. \tag{5.16}$$

This and (5.15) imply that

$$\mu\left[\exists\, \eta \in \mathcal{A}^{V_0^0(\ell)},\ \ell \geq l_0(\eta)\, :\, N_n^k(\eta) < \frac{1}{2} q_{min}^{(2\ell)^d} \frac{|\bar{\Lambda}_n|}{(5R_n)^d}\right] \leq |\mathcal{A}|^{(2R_n)^d} e^{-\frac{|\bar{\Lambda}_n|}{(5R_n)^d}\frac{q_{min}^{(2\ell)^d}}{16}}. \tag{5.17}$$

But

$$q_{min}^{(2\ell)^d} \frac{|\bar{\Lambda}_n|}{(5R_n)^d} \geq |\bar{\Lambda}_n| \frac{q_{min}^{(2R_n)^d}}{(5R_n)^d}. \tag{5.18}$$

By the definition of R_n in (3.15), $R_n^d = \sqrt{\log|\bar{\Lambda}_n|}$. Thus for any $C > 0$ and for any $\epsilon > 0$ there exists $n_0 = n_0(q_{min}, \epsilon, C)$ so that for $n \geq n_0$,

$$|\bar{\Lambda}_n|^\epsilon \frac{q_{min}^{(2R_n)^d}}{(5R_n)^d} = |\bar{\Lambda}_n|^\epsilon \frac{e^{2^d\sqrt{\log|\bar{\Lambda}_n|}\log q_{min}}}{5^d\sqrt{\log|\bar{\Lambda}_n|}} \geq C.$$

This and (5.17) imply that for any $\epsilon_1 > 0$ and $\epsilon_2 > 0$, positive C_1 and C_2, for $n \geq n_0$, we have

$$\mu\left[\exists\, \eta \in \mathcal{A}^{V_0^0(\ell)},\ \ell \geq l_0(\eta)\, :\, N_n^u(\eta) < C_1|\bar{\Lambda}_n|^{1-\epsilon_1}\right] \leq |\mathcal{A}|^{(2R_n)^d} e^{-2C_2|\bar{\Lambda}_n|^{1-\epsilon_2}}. \tag{5.19}$$

Finally, note that for $n \geq n_0$,

$$|\mathcal{A}|^{(2R_n)^d} = e^{2^d \log|\mathcal{A}|\sqrt{\log|\bar{\Lambda}_n|}} \leq e^{C_2|\bar{\Lambda}_n|^{1-\epsilon_2}}.$$

Thus we have proved the lemma. •

6 Proof of Theorem 3.5

We show the probability of overestimation (3.20). Recall the definition of the set $\mathcal{B}(n, R_n)$ given in (5.2). Clearly,

$$\begin{aligned}\mu(\exists\, i \in \bar{\Lambda}_n : \hat{l}_n(i) > l_i(\sigma)) \leq\ & \mu(\mathcal{B}(n, R_n)^c)\\ &+ \mu(\exists\, i \in \bar{\Lambda}_n : \hat{l}_n(i) > l_i(\sigma), \mathcal{B}(n, R_n)).\end{aligned} \tag{6.1}$$

The first term is estimated by Lemma 5.3, choosing $\epsilon_1 = \frac{1}{3}, \epsilon_2 = \epsilon$, $C_1 = 1, C_2 = 2$. This yields

$$\mu((\mathcal{B}(n, R_n))^c) \leq (4R_n+1)^d e^{-2|\bar{\Lambda}_n|^{1-\epsilon}},$$

for all $n \geq n_0$ where n_0 depends on the choices $\epsilon_1 = \frac{1}{3}, \epsilon_2 = \epsilon, C_1 = 1, C_2 = 2$ and q_{min}. Since

$$(4R_n+1)^d \leq C(d)\sqrt{\log|\bar{\Lambda}_n|} \leq C(d) e^{|\bar{\Lambda}_n|^{1-\epsilon}},$$

eventually, we have that for all $n \geq n_0$

$$\mu((\mathcal{B}(n, R_n))^c) \leq C(d) e^{-|\bar{\Lambda}_n|^{1-\epsilon}}. \tag{6.2}$$

We now study the last term of (6.1). We are interested in the event $\{\hat{l}_n(i) = \ell > l_i(\sigma)\}$. Note that $\ell > l_i(\sigma)$ implies that for any j such that $X_j^{\ell-1} = \sigma_i^{\ell-1}$, necessarily $l_j(\sigma) = l_i(\sigma) \leq \ell - 1$ and as a consequence $\gamma_j(\cdot|X_j^{\ell-1}) = \gamma_i(\cdot|\sigma_i^{\ell-1})$.

Hence, for any $\ell > l_i(\sigma)$, we have, by (3.14)

$$\begin{aligned}
\log L_n(i,\ell) &= \sum_{j:\, X_j^{\ell-1}=\sigma_i^{\ell-1}} \frac{1}{N_n(X_j^\ell)} \left[\log MPL_n(j,\ell) - \log MPL_n(i,\ell-1)\right] \\
&\leq \sum_{j:\, X_j^{\ell-1}=\sigma_i^{\ell-1}} \frac{1}{N_n(X_j^\ell)} \left[\log MPL_n(j,\ell) - \log PL_n^{(i,\ell-1)}(\gamma_i(\cdot|\sigma_i^{\ell-1}))\right] \\
&= \sum_{j:\, X_j^{\ell-1}=\sigma_i^{\ell-1}} \frac{1}{N_n(X_j^\ell)} \sum_{a\in A} \left(N_n(X_j^\ell, a) \log \left[\frac{\hat{p}_n(a|X_j^\ell)}{\gamma_i(a|\sigma_i^{\ell-1})}\right]\right) \\
&= \sum_{j:\, X_j^{\ell-1}=\sigma_i^{\ell-1}} \sum_{a\in A} \left(\hat{p}_n(a|X_j^\ell) \log \left[\frac{\hat{p}_n(a|X_j^\ell)}{\gamma_i(a|\sigma_i^{\ell-1})}\right]\right) \\
&\leq \sum_{j:\, X_j^{\ell-1}=\sigma_i^{\ell-1}} \sum_{a\in A} \frac{\left(\hat{p}_n(a|X_j^\ell) - \gamma_i(a|\sigma_i^{\ell-1})\right)^2}{\gamma_i(a|\sigma_i^{\ell-1})} \\
&= \sum_{j:\, X_j^{\ell-1}=\sigma_i^{\ell-1}} \sum_{a\in A} \frac{\left(\hat{p}_n(a|X_j^\ell) - \gamma_j(a|c_j(\sigma))\right)^2}{\gamma_j(a|c_j(\sigma))}. \qquad (6.3)
\end{aligned}$$

We used that for any two probability distributions P and Q on A,

$$\sum_a P(a) \log \frac{P(a)}{Q(a)} \leq \sum_a \frac{(P(a) - Q(a))^2}{Q(a)}, \qquad (6.4)$$

and in the last line the fact that $\{X_j^{\ell-1} = \sigma_i^{\ell-1}, \ell > l_i(\sigma)\}$ implies $\gamma_j(a|c_j(\sigma)) = \gamma_i(a|\sigma_i^{\ell-1})$. Hence, writing for short $\gamma_j(\cdot) = \gamma_j(\cdot|c_j(\sigma))$, define

$$E_\ell = \left\{\forall\, j \in \bar{\Lambda}_n, l_j(\sigma) < \ell, \forall a \in \mathcal{A} : \left|\hat{p}_n(a|\sigma_j^\ell) - \gamma_j(a)\right| \leq \sqrt{\kappa(\delta)\gamma_j(a)\frac{\log N_n(\sigma_j^\ell)}{N_n(\sigma_j^\ell)}}\right\},$$

where δ is as in (5.3) and $\kappa(\delta)$ is defined in (3.18). Then on E_ℓ, (6.3) can be bounded uniformly in $i \in \bar{\Lambda}_n$ from above by

$$\begin{aligned}
\sum_{j:\, X_j^{\ell-1}=\sigma_i^{\ell-1}} \sum_{a\in\mathcal{A}} \kappa(\delta) \frac{\log N_n(X_j^\ell)}{N_n(X_j^\ell)} &= \sum_{v\in\mathcal{A}^{\partial V_0(\ell)}} N_n(\sigma_i^{\ell-1}v) \sum_{a\in\mathcal{A}} \kappa(\delta) \frac{\log N_n(\sigma_i^{\ell-1}v)}{N_n(\sigma_i^{\ell-1}v)} \\
&\leq \kappa(\delta)|\mathcal{A}|\;|\mathcal{A}|^{|\partial V_0(\ell)|} \log|\bar{\Lambda}_n|.
\end{aligned}$$

Hence, on E_ℓ, for all $i \in \bar{\Lambda}_n$ having $l_i(\sigma) < \ell$,

$$\log L_n(i,\ell) \leq \kappa(\delta)|\mathcal{A}||\mathcal{A}|^{|\partial V_0(\ell)|} \log|\bar{\Lambda}_n| = pen(\ell, n).$$

This implies that $\hat{l}_n(i) < \ell$ for all $\ell > l_i(\sigma)$ and hence $\hat{l}_n(i) \leq l_i(\sigma)$. Thus

$$\mu(\exists\, i \in \bar{\Lambda}_n : \hat{l}_n(i) > l_i(\sigma), \mathcal{B}(n, R_n)) \leq \sum_{\ell=1}^{R_n} \mu(E_\ell^c, \mathcal{B}(n, R_n)).$$

But

$$E_\ell^c \subset \left\{ \exists a \in \mathcal{A}, \exists\, \eta \in \mathcal{A}^{V_0^0(\ell)} : l_0(\eta) \leq \ell, \left|\hat{p}_n(a|\eta) - \gamma_{\{0\}}(a|\eta)\right| > \sqrt{\kappa(\delta)\gamma_{\{0\}}(a|\eta)\frac{\log N_n(\eta)}{N_n(\eta)}} \right\}.$$

Hence by Theorem 5.1, for $n \geq n_0$, we have

$$\sum_{\ell=1}^{R_n} \mu(E_\ell^c, \mathcal{B}(n, R_n)) \leq |\mathcal{A}| C(d) R_n^{d+1} \exp\left(-c(\delta)\sqrt{\log|\bar{\Lambda}_n|}\right). \tag{6.5}$$

By definition of R_n,

$$R_n^{d+1} \leq (\log \bar{\Lambda}_n)^{\frac{d+1}{2d}}. \tag{6.6}$$

Taking into account (6.1), (6.2), (6.5) and (6.6) we get (3.20). This finishes the proof of Theorem 3.5.

7 Proof of Theorem 3.9

We now turn to the problem of underestimation. We suppose that n_0 is sufficiently large such that $R_n \geq L$. Fix $i \in \bar{\Lambda}_n$ and suppose that $\hat{l}_n(i) < l_i(\sigma)$. Since $l_i(\sigma) \leq L \leq R_n$, this implies by definition of the estimator that there exists $\ell \leq l_i(\sigma)$, such that

$$\log L_n(i, \ell) \leq pen(\ell, n). \tag{7.1}$$

Recall that by (3.14),

$$\log L_n(i,\ell) = \left[\sum_{j:\, X_j^{\ell-1} = \sigma_i^{\ell-1}} \frac{1}{N_n(X_j^\ell)} \log MPL_n(j,\ell)\right] - \log MPL_n(i, \ell-1).$$

By definition of $\Delta_n(\eta)$ in (4.13) we can write

$$\begin{aligned}
\frac{1}{|\bar{\Lambda}_n|}\log L_n(i,\ell) &= \left(\sum_{j:\, X_j^{\ell-1}=\sigma_i^{\ell-1}} \frac{1}{N_n(X_j^\ell)}\Delta_n(X_j^\ell)\right) - \Delta_n(\sigma_i^{\ell-1}) \\
&\quad + \left(\sum_{v\in\mathcal{A}^{\partial V_0(\ell)}} \sum_{a\in\mathcal{A}} p((\sigma_i^{\ell-1}, v, a)) \log p(a|\sigma_i^{\ell-1}, v)\right) \\
&\quad - \sum_{a\in\mathcal{A}} p((\sigma_i^{\ell-1}, a)) \log p(a|\sigma_i^{\ell-1}) \\
&= \left(\sum_{v\in\mathcal{A}^{\partial V_0(\ell)}} \Delta_n(\sigma_i^{\ell-1} v)\right) - \Delta_n(\sigma_i^{\ell-1}) \\
&\quad + \left(\sum_{v\in\mathcal{A}^{\partial V_0(\ell)}} \sum_{a\in\mathcal{A}} p((\sigma_i^{\ell-1}, v, a)) \log p(a|\sigma_i^{\ell-1}, v)\right) \\
&\quad - \sum_{a\in\mathcal{A}} p((\sigma_i^{\ell-1}, a)) \log p(a|\sigma_i^{\ell-1}).
\end{aligned}$$

Set

$$D(i,\ell,\sigma) = \sum_{v\in\mathcal{A}^{\partial V_0(\ell)}}\sum_{a\in\mathcal{A}} p((\sigma_i^{\ell-1},v,a))\log p(a|\sigma_i^{\ell-1},v) - \sum_{a\in\mathcal{A}} p((\sigma_i^{\ell-1},a))\log p(a|\sigma_i^{\ell-1}). \quad (7.2)$$

Moreover, for a constant $t>0$ that will be chosen later, define

$$E(t,\ell) = \{\forall\, \eta\in\mathcal{A}^{V_0^0(\ell-1)},\ \forall\, v\in\mathcal{A}^{\partial V_0(\ell)} : |\Delta_n(\eta v)| \le \frac{t}{2}\frac{1}{|\mathcal{A}|^{|\partial V_0(\ell)|}},\ \ |\Delta_n(\eta)|\le t/2\}.$$

Then on $E(t,\ell)$,

$$\frac{1}{|\bar\Lambda_n|}\log L_n(i,\ell) \ \ge\ D(i,\ell,\sigma)-t.$$

Next we show that $D(i,\ell,\sigma)$ can be bounded away from zero. Taking into account

$$\sum_{v\in\mathcal{A}^{\partial V_0(\ell)}} p((\sigma_i^{\ell-1},v,a)) = p((\sigma_i^{\ell-1},a)),$$

we can write

$$D(i,\ell,\sigma) = \sum_{v\in\mathcal{A}^{\partial V_0(\ell)}}\sum_{a\in\mathcal{A}} p((\sigma_i^{\ell-1},v,a))\log\frac{p(a|\sigma_i^{\ell-1},v)}{p(a|\sigma_i^{\ell-1})}.$$

By Pinsker's inequality for relative entropy (see for example Fedotov et al. (2003), [9]), we have that for P and Q probability distributions on $\mathcal{A}$,

$$\sum_a P(a)\log\frac{P(a)}{Q(a)} \ge \frac{1}{2}||P-Q||_{TV}^2.$$

Moreover,

$$||P-Q||_{TV}^2 \ge \sup_a (P(a)-Q(a))^2.$$

But, since $\ell\le l_i(\sigma)$, there exist v and a such that $p(a|\sigma_i^{\ell-1})\neq p(a|\sigma_i^{\ell-1},v)$. Hence we have that $D(i,\ell,\sigma)>0$. Since we are working under the assumption that $l_i(\omega)\le L$ for all ω (recall condition (3.22) of Assumption 3.7), we can thus conclude that

$$D(x,\ell,\sigma)\ge d_0>0, \quad (7.3)$$

where

$$d_0 = \inf_{i,\sigma}\inf_{\ell:\ell\le l_i(\sigma)} D(i,\ell,\sigma).$$

Choosing now $t=\frac{d_0}{2}$, we finally obtain that on $E(\frac{d_0}{2},\ell)$,

$$\log L_n(i,\ell) - pen(\ell,n) \ge |\bar\Lambda_n|\frac{d_0}{2} - pen(\ell,n) > 0$$

for $n\ge n_0(i)$, since $pen(\ell,n)=\kappa|\mathcal{A}||\mathcal{A}|^{|\partial V_0(\ell)|}\log|\Lambda_n| = O(\log|\Lambda_n|)$. This is in contradiction to (7.1) and implies that $\hat l_n(i)\ge\ell$. Hence we conclude that

$$\mu[\exists i\in\Lambda_n : \hat l_n(i) < l_i(\sigma)] \le \sum_{\ell\le L}\mu\left[E\left(\frac{d_0}{2},\ell\right)^c\right]. \quad (7.4)$$

We use (4.15) and sum over all possibilities of choosing $\eta \in \mathcal{A}^{V_0^0(\ell-1)}$ and of choosing patterns X_j^ℓ such that $X_j^{\ell-1} = \eta$, which gives $|\mathcal{A}|^{|V_0(\ell)|}$ terms. But since for $\ell \leq L$, $|\mathcal{A}|^{|V_0(\ell)|} \leq |\mathcal{A}|^{(2L+1)^d}$, we finally obtain

$$
\begin{aligned}
&\mu\left(\left(E(\frac{d_0}{2},\ell)\right)^c\right) \\
&\leq 3|\mathcal{A}|e^{1/e}\left(|\mathcal{A}|^{(2L+1)^d}\right)\exp\left(-C(d,L)\frac{|\bar{\Lambda}_n|d_0\alpha_0^2}{8|\mathcal{A}|^2[\log^2\alpha_0]|\mathcal{A}|^{|\partial V_0(\ell)|}e}\right) \\
&\leq 3|\mathcal{A}|e^{1/e}\left(|\mathcal{A}|^{(2L+1)^d}\right)\exp\left(-\frac{C(d,L)|\bar{\Lambda}_n|d_0\alpha_0^2}{8|\mathcal{A}|^2[\log^2\alpha_0]|\mathcal{A}|^{|\partial V_0(L)|}e}\right) \\
&= 3|\mathcal{A}|e^{1/e}\left(|\mathcal{A}|^{(2L+1)^d}\right)\exp\left(-C_2(d,L,q_{min})|\bar{\Lambda}_n|\right)
\end{aligned}
$$

(recall the control of α_0 given in (4.12)), where $C_2(d, L, q_{min})$ is another constant depending on the dimension d, on the interaction range L and on q_{min}.

Thus, we can conclude that for any $0 < \epsilon < 1$ there exists $n_0 = n_0(\epsilon, q_{min}, L, d)$ such that for all $n \geq n_0$,

$$
\sum_{\ell \leq L}\mu\left(E\left(\frac{d_0}{2},\ell\right)^c\right) \leq \exp\left(-|\bar{\Lambda}_n|^{1-\epsilon}\right). \tag{7.5}
$$

(7.4) and (7.5) together conclude the proof of Theorem 3.9. •

8 Final comments

In the present paper we generalize the concept of chains with memory of variable length to the multidimensional case of random fields. The main aim of this concept is to adopt a parsimonious way of describing data: the spin at site i is influenced by a finite, but random number of spins; it is important to distinguish between relevant and irrelevant states. As in the case of one dimensional models, the set of relevant neighbor states of site i is called the *context*. The radius of the smallest ball containing the support of the context is the *length of the context* of site i.

We presented in Section 3 an estimator of the context length function based on a sequence of local decisions between two possible context lengths. These decisions are performed using the log likelihood ratio function. In the case of dimension one, our estimator is simply the context length estimator of variable length chains which has been classically considered in the literature. We refer the interested reader to Galves and Löcherbach (2008), [13], for a survey and bibliographic comments.

Let us finally compare our results in the case of dimension one to the results of Ferrari and Wyner (2003), [10]. They consider stationary chains taking values in a finite alphabet without imposing any a priori bound on the memory. Hence, they are dealing with infinite trees. They overcome this difficulty by approximating the possibly infinite memory chain by a sequence of finite range Markov chains of growing order. The price they have to pay in order to deal with these general processes is to impose geometrically $\alpha-$mixing conditions both for the control of the over- and the underestimation.

In comparison to their results, to control the underestimation, we need a slightly stronger assumption. We require geometrically $\Phi-$mixing which implies geometrically $\alpha-$mixing. This is crucial to obtain Theorem 3.9. We use Condition (3.22) as sufficient condition to obtain the geometrically $\Phi-$mixing.

Condition (3.22), which implies that the random field is of finite range, could probably be relaxed. It should be possible to deal with infinite range models, provided one finds other sufficient conditions implying mixing.

Note however, that mixing implies automatically the uniqueness of the underlying measure μ. Hence, using this kind of technique always implies that the Dobrushin uniqueness condition (3.23) must be satisfied. There is some hope to deal with the underestimation even in the case of phase transition, see Remark 4.3.

Concerning the control of the overestimation, we are able to deal with the general long range case without requiring mixing. Hence we can do better than Ferrari and Wyner (2003), [10], in this aspect. We need to impose the positivity condition on the specification, see Assumption 3.4. Ferrari and Wyner do only impose positivity within each step of the canonical Markov approximation, allowing these lower bounds to tend to zero at a certain rate.

9 Appendix

At the end of Section 2 we argued that in order to estimate the context of a finite set of sites $\Lambda \Subset \mathbb{Z}^d$ it is sufficient to estimate the contexts of the one-point specification. In particular, we stated formula (2.14), which relates $c_\Lambda(\omega)$ to $c_i(\omega)$. In the first subsection of this appendix we show this. In the second subsection, we complete the computations for the examples 2.8 and 2.9. Finally we state a deviation inequality needed in Section 5.

9.1 From one point specifications to several points

It is well known in Statistical Mechanics that the positive one point specification uniquely determines the family of specifications, see Theorem 1.33 of Georgii (1988), [15]. This result still holds for Variable-neighborhood random fields, since they can be embedded into classical random fields. But we would like to determine if and how the context of one single site determines the $\Lambda-$contexts of the specification, for any $\Lambda \Subset \mathbb{Z}^d$. Proposition 9.1 gives an answer.

We consider local specifications γ which are positive, i.e.

$$\gamma_\Lambda(\omega_\Lambda|\cdot) > 0, \quad \text{for all } \omega_\Lambda \in \mathcal{A}^\Lambda \quad \text{and} \quad \Lambda \Subset \mathbb{Z}^d.$$

In the following it will be convenient to write

$$\gamma_\Lambda(\{\omega_\Lambda\}|\sigma) = \varrho_\Lambda(\omega_\Lambda\sigma).$$

This family $\{\varrho_\Lambda, \Lambda \Subset \mathbb{Z}^d\}$ is a family of functions $\varrho_\Lambda : \Omega \to [0,1]$ satisfying the following two conditions:

$$\sum_{\omega_\Lambda \in \mathcal{A}^\Lambda} \rho_\Lambda(\omega_\Lambda\sigma) \ = \ 1, \quad \forall \sigma \in \Omega, \tag{9.1}$$

and for every $\Lambda \subset \Delta \Subset \mathbb{Z}^d$, all ω, η, σ in Ω we have

$$\frac{\rho_\Delta(\omega_\Lambda\sigma_{\Delta\setminus\Lambda}\sigma_{\Delta^c})}{\rho_\Delta(\eta_\Lambda\sigma_{\Delta\setminus\Lambda}\sigma_{\Delta^c})} = \frac{\rho_\Lambda(\omega_\Lambda\sigma_{\Lambda^c})}{\rho_\Lambda(\eta_\Lambda\sigma_{\Lambda^c})}. \tag{9.2}$$

Proposition 9.1 *Assume that the family of local specifications γ defined in* (2.5) *is positive* [1]. *We have the following:*

[1] The positivity requirement can be relaxed, under some minor modifications of the proof, see Georgii (1988), [15], Theorem 1.33.

- *γ is uniquely determined by $\{\gamma_{\{i\}}(\cdot|c_i(\omega)),\ i \in \mathbb{Z}^d, \omega \in \Omega\}$.*
- *For $\Lambda \Subset \mathbb{Z}^d$,*

$$\mathrm{sp}_\Lambda(\omega) = \cup_{\omega_\Lambda}\left(\cup_{\mathrm{i}\in\Lambda}\mathrm{sp}_{\{\mathrm{i}\}}(\omega)\right) \setminus \Lambda. \tag{9.3}$$

Proof. Recall that we set $\gamma_{\{i\}}(\{\omega(i)\}|c_i(\omega)) = \rho_{\{i\}}(\omega)$. Further $\rho_{\{i\}}(\omega) > 0$ for $\omega \in \Omega$ since we assumed that γ is positive. For each fixed $\omega(i)$, $\omega_{\{i\}^c} \mapsto \rho_{\{i\}}(\omega)$ is a measurable function with respect to $\mathcal{F}_{\mathrm{sp}_{\{\mathrm{i}\}}}$, see (2.3). For each Λ, Georgii (1988), [15], shows in the proof of Theorem 1.33 how to determine ρ_Λ in terms of $\{\rho_{\{i\}}, i \in \mathbb{Z}^d\}$ such that for any measurable function f we have that

$$\int f(\omega)d\mu(\omega) = \int d\mu(\omega_{\Lambda^c}) \sum_{\omega_\Lambda \in \mathcal{A}^\Lambda} f(\omega)\rho_\Lambda(\omega), \quad \forall \Lambda \Subset \mathbb{Z}^d, \tag{9.4}$$

where μ is any measure on Ω so that

$$\int f(\omega)d\mu(\omega) = \int d\mu(\omega_{\{i\}^c}) \sum_{\omega_{\{i\}} \in \mathcal{A}} f(\omega)\rho_i(\omega).$$

This immediately shows that γ is uniquely determined by $\rho_{\{i\}}$. To construct ρ_Λ and to prove (2.14), one proceeds by induction on $|\Lambda|$. The case $|\Lambda| = 1$ is trivial. Suppose then that ρ_{Λ_1} and ρ_{Λ_2} have been constructed. Let Λ be the union of two disjoint sets, $\Lambda = \Lambda_1 \cup \Lambda_2$, $\Lambda_1 \cap \Lambda_2 = \emptyset$. Define

$$\rho_\Lambda(\omega) = \frac{\rho_{\Lambda_1}(\omega)}{\sum_{\bar\omega_{\Lambda_1}} \frac{\rho_{\Lambda_1}(\bar\omega_{\Lambda_1}\omega_{\Lambda_1^c})}{\rho_{\Lambda_2}(\bar\omega_{\Lambda_1}\omega_{\Lambda_1^c})}}. \tag{9.5}$$

By induction, for any given ω_{Λ_1}, $\omega_{\Lambda_1^c} \mapsto \rho_{\Lambda_1}(\omega) = \rho_{\Lambda_1}(\omega_{\Lambda_1}, \omega_{\Lambda_1^c})$ depends only on $\mathrm{sp}_{\Lambda_1}(\omega)$, and for any given ω_{Λ_2}, $\omega_{\Lambda_2^c} \mapsto \rho_{\Lambda_2}(\omega_{\Lambda_2}, \omega_{\Lambda_2^c})$ only on $\mathrm{sp}_{\Lambda_2}(\omega)$. Hence (9.5) implies that for any given ω_Λ, the function $\omega_\Lambda \mapsto \rho_\Lambda(\omega)$ depends, by construction, on the $\sigma-$algebra generated by $\mathrm{sp}_{\Lambda_1}(\omega) \cup \bigcup_{\bar\omega_{\Lambda_1}} \mathrm{sp}_{\Lambda_2}(\bar\omega_{\Lambda_1}\omega_{\Lambda_1^c})$. Note that in general $\mathrm{sp}_{\Lambda_1}(\omega) \cap \mathrm{sp}_{\Lambda_2}(\omega) \neq \emptyset$. Therefore the value of ω_{Λ_1} might be relevant for determining $\mathrm{sp}_{\Lambda_2}(\omega)$ and the value of ω_{Λ_2} might be relevant for determining $\mathrm{sp}_{\Lambda_1}(\omega)$. To have a function $\rho_\Lambda(\omega)$ measurable for any choice of ω_Λ we set

$$\mathrm{sp}_\Lambda(\omega) = \left[\cup_{\omega_{\Lambda_1}} \cup_{\omega_{\Lambda_2}} (\mathrm{sp}_{\Lambda_1}(\omega) \cup \mathrm{sp}_{\Lambda_2}(\omega))\right] \setminus (\Lambda_1 \cup \Lambda_2).$$

In this way, for any choice of ω_Λ, $\rho_\Lambda(\omega)$ is $\mathcal{F}_{\mathrm{sp}_\Lambda}-$measurable. It is immediate to verify by induction that one has

$$\mathrm{sp}_\Lambda(\omega) = \left[\cup_{\omega_\Lambda} \cup_{\mathrm{i}\in\Lambda} \mathrm{sp}_{\{\mathrm{i}\}}(\omega)\right] \setminus \Lambda.$$

We need to show that (9.4) holds. By induction, taking in account that $\omega = (\omega_{\Lambda_k}, \omega_{\Lambda_k^c})$,

$$\int f(\omega)d\mu(\omega) = \int d\mu(\omega_{\Lambda_k^c}) \sum_{\omega_{\Lambda_k}} f(\omega)\rho_{\Lambda_k}(\omega), \quad k = 1, 2 \tag{9.6}$$

holds. To show that this holds for ρ_Λ take a positive measurable function f defined on Ω. We have

$$\begin{aligned}
&\int d\mu(\bar\omega) \sum_{\omega_\Lambda} f(\omega_\Lambda \bar\omega_{\Lambda^c}) \\
&= \int d\mu(\bar\omega) \sum_{\omega_{\Lambda_2}} \sum_{\omega_{\Lambda_1}} f(\omega_{\Lambda_1}\omega_{\Lambda_2}\bar\omega_{\Lambda^c}) \\
&= \int d\mu(\bar\omega) \sum_{\omega_{\Lambda_2}} \rho_{\Lambda_2}(\omega_{\Lambda_2}\bar\omega_{\Lambda_2^c})\rho_{\Lambda_2}^{-1}(\omega_{\Lambda_2}\bar\omega_{\Lambda_2^c}) \sum_{\omega_{\Lambda_1}} f(\omega_{\Lambda_1}\omega_{\Lambda_2}\bar\omega_{\Lambda^c}).
\end{aligned}$$

But applying (9.6) first to Λ_2, then to Λ_1, this last line can be written as

$$\int d\mu(\bar\omega) \sum_{\omega_{\Lambda_2}} \rho_{\Lambda_2}(\omega_{\Lambda_2}\bar\omega_{\Lambda_2^c})\rho_{\Lambda_2}^{-1}(\omega_{\Lambda_2}\bar\omega_{\Lambda_2^c}) \sum_{\omega_{\Lambda_1}} f(\omega_{\Lambda_1}\omega_{\Lambda_2}\bar\omega_{\Lambda^c})$$

$$= \int d\mu(\bar\omega)\, \rho_{\Lambda_2}^{-1}(\bar\omega) \left[\sum_{\omega_{\Lambda_1}} f(\omega_{\Lambda_1}\bar\omega_{\Lambda_1^c})\right]$$

$$= \int d\mu(\bar\omega) \sum_{\tilde\omega_{\Lambda_1}} \rho_{\Lambda_1}(\tilde\omega_{\Lambda_1}\bar\omega_{\Lambda_1^c})\rho_{\Lambda_2}^{-1}(\tilde\omega_{\Lambda_1}\bar\omega_{\Lambda_1^c}) \left[\sum_{\omega_{\Lambda_1}} f(\omega_{\Lambda_1}\bar\omega_{\Lambda_1^c})\right]$$

$$= \int d\mu(\bar\omega) \sum_{\omega_{\Lambda_1}} f(\omega_{\Lambda_1}\bar\omega_{\Lambda_1^c})\rho_{\Lambda_1}(\omega_{\Lambda_1}\bar\omega_{\Lambda_1^c})\rho_{\Lambda_1}^{-1}(\omega_{\Lambda_1}\bar\omega_{\Lambda_1^c}) \left[\sum_{\tilde\omega_{\Lambda_1}} \rho_{\Lambda_1}(\tilde\omega_{\Lambda_1}\bar\omega_{\Lambda_1^c})\rho_{\Lambda_2}^{-1}(\tilde\omega_{\Lambda_1}\bar\omega_{\Lambda_1^c})\right]$$

$$= \int d\mu(\bar\omega) \sum_{\omega_{\Lambda_1}} f(\omega_{\Lambda_1}\bar\omega_{\Lambda_1^c})\rho_{\Lambda_1}(\tilde\omega_{\Lambda_1}\bar\omega_{\Lambda_1^c})\rho_{\Lambda}^{-1}(\tilde\omega_{\Lambda_1}\bar\omega_{\Lambda_1^c})$$

$$= \int d\mu(\omega_{\Lambda_1^c}) \sum_{\omega_{\Lambda_1}} f(\omega)\rho_{\Lambda_1}(\omega)\rho_{\Lambda}^{-1}(\omega).$$

Applying once more (9.6), we obtain

$$\int d\mu(\omega_{\Lambda_1^c}) \sum_{\omega_{\Lambda_1}} f(\omega)\rho_{\Lambda_1}(\omega)\rho_{\Lambda}^{-1}(\omega) = \int d\mu(\omega) f(\omega)\rho_{\Lambda}^{-1}(\omega).$$

By applying the above equality to $f(\omega)\rho_\Lambda(\omega)$ instead of $f(\omega)$ we get the result. The above definition of ρ_Λ depends on the choice of Λ_1 and Λ_2; one needs to obtain an unambiguous definition of ρ_Λ to choose a definite strategy to exhausting Λ site by site. •

9.2 Continuation of the examples 2.8 and 2.9

Continuation of example 2.8

We explicitly compute the one-point specification $\gamma_{\{0\}}(\cdot \mid c_0(\omega))$ of example 2.8 given in Section 2. According to (2.7) let

$$R_0(\omega) = \inf\{n > 1 : \omega(n) \neq \omega(1)\}, \quad L_0(\omega) = \sup\{n < -1 : \omega(n) \neq \omega(-1)\}. \tag{9.7}$$

Write $\mathrm{sp}_0(\omega) = [\mathrm{L}_0(\omega), \mathrm{R}_0(\omega)] \setminus \{0\}$.

Proof of (2.8) **and** (2.9)**:** In a first step we calculate for $k, l > 1$,

$$\mathbb{P}[X(0) = 1 |\, X(1) = \dots X(l-1) = 1, X(l) = 0, X(-1) = \dots X(-k+1) = 1, X(-k) = 0]. \tag{9.8}$$

Put

$$R = R(\omega) = \min\{l \geq 1 : X(l) = 0\}, L = L(\omega) = \min\{|k| \geq 1 : X(-k) = 0\}.$$

Remark that $R(\omega)$ and $L(\omega)$ just defined are different from the one defined in (9.7). They coincide when $\omega(1) = \omega(-1) = 1$. Hence we have to compute

$$\frac{\mathbb{P}[X(0) = 1, X(1) = \dots X(l-1) = 1, X(l) = 0, X(-1) = \dots X(-k+1) = 1, X(-k) = 0]}{\mathbb{P}[X(1) = \dots X(l-1) = 1, X(l) = 0, X(-1) = \dots X(-k+1) = 1, X(-k) = 0]}$$
$$= \frac{\mathbb{P}[X(0) = 1, R = l, L = -k]}{\mathbb{P}[R = l, L = -k]}. \tag{9.9}$$

In the denominator there is no restriction on the value $X(0)$ at 0. To compute the numerator, we use

$$
\begin{aligned}
&\mathbb{P}[X(0)=1, R=l, L=-k] \\
&= \mathbb{P}[X(l-1)=\ldots=X(0)=X(-1)=X(-2)=\ldots=X(-k+1)=1, X(-k)=0] \\
&- \mathbb{P}[X(l)=1=\ldots=X(0)=1, X(-1)=X(-2=\ldots=X(-k+1)=1, X(-k)=0].
\end{aligned} \tag{9.10}
$$

We have from formula (22) of Ferrari and Wyner (2003), [10], that

$$
\begin{aligned}
&\mathbb{P}[X(0)=X(-1)=X(-2)=\ldots=X(-n+1)=1, X(-n)=0] \\
&\qquad = \frac{1}{2\mu}\left(\frac{c_1}{1-\rho_1}\rho_1^n + \frac{c_2}{1-\rho_2}\rho_2^n\right)
\end{aligned} \tag{9.11}
$$

for any $n \geq 1$. Here, $\mu = \mathbb{E}[T]$. Therefore (9.10) is equal to

$$
\begin{aligned}
&\mathbb{P}[X(l)=0, X(l-1)=\ldots=X(0)=X(-1)=\ldots=X(-k+1)=1, X(-k)=0] \\
&= \frac{1}{2\mu}\left(\frac{c_1}{1-\rho_1}\rho_1^{l+k-1} + \frac{c_2}{1-\rho_2}\rho_2^{l+k-1}\right) - \frac{1}{2\mu}\left(\frac{c_1}{1-\rho_1}\rho_1^{l+k} + \frac{c_2}{1-\rho_2}\rho_2^{l+k}\right) \\
&= \frac{1}{2\mu}\left(c_1\rho_1^{l+k-1} + c_2\rho_2^{l+k-1}\right).
\end{aligned} \tag{9.12}
$$

We need to compute the denominator of (9.8) We have

$$
\begin{aligned}
\mathbb{P}[R=l, L=-k] &= \mathbb{P}[R=l, X(0)=1, L=-k] + \mathbb{P}[R=l, X(0)=0, L=-k] \\
&= \frac{1}{2\mu}\left(c_1\rho_1^{l+k-1}k + c_2\rho_2^{l+k-1}\right) + \mathbb{P}[R=l, X(0)=0, L=-k].
\end{aligned} \tag{9.13}
$$

We still have to calculate

$$
\begin{aligned}
&\mathbb{P}[X(0)=0; R=l, L=-k] \\
&\qquad = \mathbb{P}[X(l)=0 \mid X(1)=\ldots=X(l-1)=1, X(0)=0, L=-k] \\
&\qquad\qquad \cdot \mathbb{P}[X(1)=\ldots=X(l-1)=1, X(0)=0, L=-k].
\end{aligned} \tag{9.14}
$$

Concerning the first term, due to the renewal structure and to (9.12),

$$
\begin{aligned}
&\mathbb{P}[X(l)=0 \mid X(1)=\ldots=X(l-1)=1, X(0)=0, L=-k] \\
&\qquad = \mathbb{P}[X(l)=0 \mid X(1)=\ldots=X(l-1)=1, X(0)=0] \\
&\qquad = \frac{\mathbb{P}[X(l)=0, X(1)=\ldots=X(l-1)=1, X(0)=0]}{\mathbb{P}[X(1)=\ldots=X(l-1)=1, X(0)=0]} \\
&\qquad = \frac{c_1\varrho_1^{l-1} + c_2\varrho_2^{l-1}}{\frac{c_1}{1-\varrho_1}\varrho_1^{l-1} + \frac{c_2}{1-\varrho_2}\varrho_2^{l-1}}.
\end{aligned} \tag{9.15}
$$

Moreover, for the second term of (9.14) we can write similarly,

$$
\begin{aligned}
&\mathbb{P}[X(1)=\ldots=X(l-1)=1, X(0)=0, L=-k] \\
&\qquad = \mathbb{P}[X(2)=\ldots=X(l-1)=1 \mid X(1)=1, X(0)=0] \\
&\qquad\qquad \cdot \mathbb{P}[X(1)=1, X(0)=0, L=-k].
\end{aligned} \tag{9.16}
$$

Here,

$$\mathbb{P}[X(2) = \ldots = X(l-1) = 1 \mid X(1) = 1, X(0) = 0] = \frac{\frac{c_1}{1-\varrho_1}\varrho_1^{l-1} + \frac{c_2}{1-\varrho_2}\varrho_2^{l-1}}{\frac{c_1}{1-\varrho_1}\varrho_1 + \frac{c_2}{1-\varrho_2}\varrho_2}. \quad (9.17)$$

Finally,

$$\mathbb{P}[X(1) = 1, X(0) = 0, L = -k] = \mathbb{P}[X(1) = 1 \mid X(0) = 0, X(-1) = 1] \cdot \mathbb{P}[X(0) = 0, L = -k], \quad (9.18)$$

where

$$\mathbb{P}[X(0) = 0, L = -k] = \frac{1}{2\mu}\left(c_1\rho_1^{k-1} + c_2\rho_2^{k-1}\right).$$

Moreover,

$$\mathbb{P}[X(1) = 1 \mid X(0) = 0, X(-1) = 1] = \frac{\mathbb{P}[X(1) = 1, X(0) = 0, X(-1) = 1]}{\mathbb{P}[X(0) = 0, X(-1) = 1]} = \frac{c_1\rho_1 + c_2\rho_2}{\frac{c_1}{1-\varrho_1}\varrho_1 + \frac{c_2}{1-\varrho_2}\varrho_2}. \quad (9.19)$$

Putting all things together, we thus obtain

$$\mathbb{P}[X(0) = 0, R = l, L = -k] = \frac{1}{2\mu}\left(c_1\rho_1^{k-1} + c_2\rho_2^{k-1}\right) \left(c_1\rho_1^{l-1} + c_2\rho_2^{l-1}\right)\frac{c_1\varrho_1 + c_2\varrho_2}{\left(\frac{c_1}{1-\varrho_1}\varrho_1 + \frac{c_2}{1-\varrho_2}\varrho_2\right)^2}. \quad (9.20)$$

Thus, as a consequence,

$$\mathbb{P}[X(0) = 1 \mid R = l, L = -k] = \frac{\left(c_1\varrho_1^{l+k-1} + c_2\varrho_2^{l+k-1}\right)}{\left(c_1\varrho_1^{l+k-1} + c_2\varrho_2^{l+k-1}\right) + \left(c_1\rho_1^{k-1} + c_2\rho_2^{k-1}\right)\left(c_1\rho_1^{l-1} + c_2\rho_2^{l-1}\right)\frac{c_1\varrho_1 + c_2\varrho_2}{\left(\frac{c_1}{1-\varrho_1}\varrho_1 + \frac{c_2}{1-\varrho_2}\varrho_2\right)^2}}. \quad (9.21)$$

2) Finally we propose to compute

$$\mathbb{P}[X(0) = 1 | \ X(1) = \ldots X(l-1) = 0, X(l) = 1, X(-1) = \ldots X(-k+1) = 1, X(-k) = 0]. \quad (9.22)$$

Here, the numerator is given by

$$\begin{aligned}&\mathbb{P}[X(0) = 1, X(1) = \ldots X(l-1) = 0, X(l) = 1, X(-1) = \ldots X(-k+1) = 1, X(-k) = 0]\\ &= \mathbb{P}[X(0) = 1, X(1) = \ldots X(l-1) = 0, X(-1) = \ldots X(-k+1) = 1, X(-k) = 0]\\ &- \mathbb{P}[X(0) = 1, X(1) = \ldots X(l) = 0, X(-1) = \ldots X(-k+1) = 1, X(-k) = 0].\end{aligned} \quad (9.23)$$

Let us calculate

$$\begin{aligned}&\mathbb{P}[X(0) = 1, X(1) = \ldots = X(l) = 0, X(-1) = \ldots X(-k+1) = 1, X(-k) = 0]\\ &= \mathbb{P}[X(2) = \ldots = X(l) = 0 \mid X(0) = 1, X(1) = 0]\\ &\quad \mathbb{P}[X(0) = 1, X(1) = 0, X(-1) = \ldots X(-k+1) = 1, X(-k) = 0].\end{aligned}$$

But by (9.12),

$$\mathbb{P}[X(0)=1, X(1)=0, X(-1)=\dots X(-k+1)=1, X(-k)=0] = \frac{1}{2\mu}\left(c_1\varrho_1^k + c_2\varrho_2^k\right).$$

Moreover, it is easy to calculate that

$$\mathbb{P}[X(2)=\dots=X(l)=0 \mid X(0)=1, X(1)=0] = \frac{\frac{c_1}{1-\varrho_1}\varrho_1^l + \frac{c_2}{1-\varrho_2}\varrho_2^l}{\frac{c_1}{1-\varrho_1}\varrho_1 + \frac{c_2}{1-\varrho_2}\varrho_2}.$$

Coming back to (9.23), we obtain finally

$$\begin{aligned}\mathbb{P}[X(0)=1, X(1)=\dots X(l-1)=0, X(l)=1, X(-1)=\dots X(-k+1)=1, X(-k)=0]\\ = \frac{1}{2\mu}\frac{\left(c_1\varrho_1^k + c_2\varrho_2^k\right)\left(c_1\varrho_1^{l-1} + c_2\varrho_2^{l-1}\right)}{\frac{c_1}{1-\varrho_1}\varrho_1 + \frac{c_2}{1-\varrho_2}\varrho_2}. \quad (9.24)\end{aligned}$$

Applying the same argument as above, we have

$$\begin{aligned}\mathbb{P}[X(0)=0, X(1)=\dots X(l-1)=0, X(l)=1, X(-1)=\dots X(-k+1)=1, X(-k)=0]\\ = \frac{1}{2\mu}\frac{\left(c_1\varrho_1^{k-1} + c_2\varrho_2^{k-1}\right)\left(c_1\varrho_1^{l} + c_2\varrho_2^{l}\right)}{\frac{c_1}{1-\varrho_1}\varrho_1 + \frac{c_2}{1-\varrho_2}\varrho_2}. \quad (9.25)\end{aligned}$$

Thus we can conclude that

$$\begin{aligned}\mathbb{P}[X(0)=1 \mid X(1)=\dots=X(l-1)=0, X(l)=1, X(-1)=\dots X(-k+1)=1, X(-k)=0]\\ = \frac{\left(c_1\varrho_1^k + c_2\varrho_2^k\right)\left(c_1\varrho_1^{l-1} + c_2\varrho_2^{l-1}\right)}{\left(c_1\varrho_1^k + c_2\varrho_2^k\right)\left(c_1\varrho_1^{l-1} + c_2\varrho_2^{l-1}\right) + \left(c_1\varrho_1^{k-1} + c_2\varrho_2^{k-1}\right)\left(c_1\varrho_1^{l} + c_2\varrho_2^{l}\right)}. \quad (9.26)\end{aligned}$$

•

Continuation of example 2.9
We prove formula (2.13), using (2.12). First note that $i \in \Gamma_j(\omega)$ implies that $\|i-j\| \le L$. Now if $i \in \Gamma_j(\omega)$, we have two cases. Either $i \in \hat{\Gamma}_i^1(\omega)$, in which case $\Gamma_i^1(\omega) = \Gamma_j^1(\omega)$. Or $i \in \partial\Gamma_j^1(\omega)$. Then $\omega(i)i = 1$, and in this case, $i \in \hat{\Gamma}_j(\omega^i)$. Then the same arguments as above show that

$$\Gamma_j^1(\omega^i) = \Gamma_i^1(\omega).$$

Hence

$$\left[\bigcup_{j\in\mathbb{Z}^2}\{\Gamma_j^1(\omega) : i \in \Gamma_j(\omega)\} \cup \bigcup_{j\in\mathbb{Z}^2}\{\Gamma_j^1(\omega^i) : i \in \Gamma_j(\omega^i)\}\right] = \Gamma_i^1(\omega).$$

Finally, by definition of $\Gamma_j(\omega)$,

$$\begin{aligned}&\left[\bigcup_{j\in\mathbb{Z}^2}\{\Gamma_j(\omega) : i \in \Gamma_j(\omega)\} \cup \bigcup_{j\in\mathbb{Z}^2}\{\Gamma_j(\omega^i) : i \in \Gamma_j(\omega^i)\}\right]\\ &= \left[\bigcup_{j\in\mathbb{Z}^2}\{\Gamma_j^1(\omega) : i \in \Gamma_j(\omega)\} \cup \bigcup_{j\in\mathbb{Z}^2}\{\Gamma_j^1(\omega^i) : i \in \Gamma_j(\omega^i)\}\right] \cap \bigcup_{j\in V_i(L)} V_j(L)\\ &= \Gamma_i^1(\omega) \cap V_i(2L).\end{aligned}$$

This concludes the proof. •

We close with the following version of Bernstein's inequality obtained by Friedman (1975), for discrete-time martingales having bounded jumps, see for instance Dzhaparidze and van Zanten[8].

Lemma 9.2 *Let $M_n = \xi_1 + \ldots + \xi_n$ be a discrete martingale with respect to some filtration $(\mathcal{F}_n)_{n \geq 0}$ having bounded jumps $|\xi_n| \leq a$. Let*

$$< M >_n = \sum_{i=1}^{n} E(\xi_i^2 | \mathcal{F}_{i-1}).$$

Then

$$P(\max_{k \leq n} |M_k| > z; < M >_n \leq L) \leq 2 \exp\left(-\frac{1}{2}\left(\frac{z^2}{L} + \frac{az}{3}\right)\right).$$

References

[1] Azencott, R. *Image analysis and Markov fields.* ICIAM 87: Proceedings of the first international conference on industrial and applied mathematics (Paris 1987), SIAM Philadelphia, PA. 53-61 (1987).

[2] Besag, J. *Spatial interaction and the statistical analysis of lattice systems.* J. Ro. Stat. Soc. Ser. B 36, 192-236 (1974).

[3] Comets, F. *On Consistency of a Class of Estimators for Exponential Families of Markov Random Fields on the Lattice.* Ann. Statist. 20, No. 1, 455-468 (1992).

[4] Csiszàr, I.,Talata, Z. *Consistent estimation of the basic neighborhood of Markov random fields.* Ann. Statist. 34, No. 1, 123-145 (2006).

[5] Dedecker, J. *Exponential inequalities and functional central limit theorems for a random field.* ESAIM, Probab. Stat. 5, 77-104 (2001).

[6] Dobrushin, R. L., *Gibbsian random fields for lattice systems with pairwise interactions.* Funct. Anal. Appl. 2, 292-301, (1968).

[7] Dobrushin, R. L., *The problem of uniqueness of Gibbs random field and the problem of phase transition.* Funct. Anal. Appl. 2, 302-312, (1968).

[8] Dzhaparidze, K., van Zanten, J.H. *On Bernstein-type inequalities for martingales.* Stochastic Processes Appl. 93, 109-117 (2001).

[9] Fedotov, A., Harremoes, P. and Topsøe, F. *Refinements of Pinsker's Inequality.* IEEE Trans. Inf. Theory, vol. 49, pp. 1491–1498 (2003).

[10] Ferrari, F.F., Wyner, A. *Estimation of general stationary processes by variable length Markov Chains.* Scand. J. Stat. 30, No. 3, 459–480 (2003).

[11] Finesso L., Liu C.C., Narayan P. *The optimal error exponent for Markov order estimation.* IEEE Trans. Inf. Theory, vol. 42, pp. 1488–1497 (1996).

[12] Galves, A., Leonardi, F. *Exponential inequalities for empirical unbounded context trees.* Sidoravicius, Vladas (ed.) et al., In and out of equilibrium 2. Papers celebrating the 10th edition of the Brazilian school of probability (EBP), Rio de Janiero, Brazil, July 30 to August 4, 2006. Basel: Birkhäuser. Progress in Probability 60, 257-269 (2008).

[13] Galves, A., Löcherbach, E. *Stochastic chains with memory of variable length.* "Festschrift in honour of the 75th birthday of Jorma Rissanen", Tampere University Press, 2008, arxiv.org/abs/0804.2050.

[14] Galves, A., Orlandi, E., Takahashi, D.Y *Identifying interacting pairs of sites in infinite range Ising models.* arxiv.org/abs/1006.0272

[15] Georgii, Hans-Otto *Gibbs measures and phase transitions.* De Gruyter Studies in Mathematics, 9. Berlin etc.: Walter de Gruyter. (1988).

[16] Gidas, B. *Parameter estimation for Gibbs distributions from fully observed data* In *Markov Random Fields: Theory and Application* 471-498. Academic Press, Boston

[17] Grunwald P. D. *The minimum description length principle.* MIT Press (2007).

[18] Ji, C., Seymour, L. *A consistent model selection procedure for Markov random fields based on penalized pseudolikelihood.* Ann. App. Probab. 6, 423-443 (1996).

[19] Merhav, N., Gutman, M., Ziv, J. *On the estimation of the order of a Markov chain and universal data compression.* IEEE Trans. Inf. Theory, vol. 35, pp. 1014–1019 (1989).

[20] Presutti, E. *Scaling Limits in Statistical Mechanics and Microstructures in Continuum Mechanics.* Springer Berlin Heidelberg, Series: Theoretical and Mathematical Physics, (2009).

[21] Rissanen, J. *A universal data compression system.* IEEE, Trans. Inform. Theory, 29, 656-664 (1983).

Eva Löcherbach
CNRS UMR 8088
Département de Mathématiques
Université de Cergy-Pontoise
95 000 CERGY-PONTOISE, France
email: `eva.loecherbach@u-cergy.fr`

Enza Orlandi
Dipartimento di Matematica
Università di Roma Tre
L.go S.Murialdo 1, 00146 Roma, Italy.
email: `orlandi@mat.uniroma3.it`