An Oracle Approach for Interaction Neighborhood Estimation in Random Field

Matthieu Lerasle, D.Y. Takahashi

An Oracle Approach for Interaction Neighborhood Estimation in Random Fields

Matthieu Lerasle[*] **and Daniel, Y. Takahashi**[†]

Instituto de Matemática e Estatística
Universidade de São Paulo
Caixa Postal 66281
05315-970 São Paulo, Brasil
e-mail: lerasle@gmail.com

Instituto de Matemática e Estatística
Universidade de São Paulo
Caixa Postal 66281
05315-970 São Paulo, Brasil
currently at Princeton University
Department of Psychology and Neuroscience
Green Hall, Princeton, NJ 08648
e-mail: takahashiyd@gmail.com

Abstract: We consider the problem of interaction neighborhood estimation from the partial observation of a finite number of realizations of a random field. We introduce a model selection rule to choose estimators of conditional probabilities among natural candidates. Our main result is an oracle inequality satisfied by the resulting estimator. We use then this selection rule in a two-step procedure to evaluate the interacting neighborhoods. The selection rule selects a small prior set of possible interacting points and a cutting step remove from this prior set the irrelevant points.
We also prove that the Ising models satisfy the assumptions of the main theorems, without restrictions on the temperature, on the structure of the interacting graph or on the range of the interactions. It provides therefore a large class of applications for our results. We give a computationally efficient procedure in these models. We finally show the practical efficiency of our approach in a simulation study.

AMS 2000 subject classifications: Primary 62M40; secondary 62M45.
Keywords and phrases: Ising Model, Model Selection, Computationally efficient algorithm.

1. Introduction

Graphical models, also known as random fields, are used in a variety of domains, including computer vision [4, 21], image processing [9], neuroscience [19], and as a general model in spatial statistics [18]. The main motivation for our work comes from neuroscience where the advancement of multichannel and optical

[*]Supported by FAPESP grant 2009/09494-0.
[†]Supported by FAPESP grant 2008/08171-0.

technology enabled the scientists to study not only a unit of neurons per time, but tens to thousands of neurons simultaneously [20]. The very important question now in neuroscience is to understand how the neurons in this ensemble interact with each other and how this is related to the animal behavior [8, 19]. This question turns out to be hard for three reasons at least. First, the experimenter has always only access to a small part of the neural system. Moreover, there is no really good model for population of neurons in spite of the good models available for single neurons. Finally, strong long range interactions exist [15]. Our work tries to overcome some of these difficulties as will be shown.

A random field can be specified by a discrete set of sites G, possibly infinite, a finite alphabet of spins A, and a probability measure P on the set of configurations $\mathcal{X}(G) = A^G$. One of the objects of interest are the one-point specification probabilities, defined for all sites i in G and all configurations x in $\mathcal{X}(G)$ by a regular version of the conditional probability

$$P\left(x(i) \mid x(j),\ j \in G/\{i\} \right).$$

From a statistical point of view, two problems are of natural interest.

Interaction neighborhood identification problem (INI):

The INI problem is to identify, for all sites i in G, the minimal subset G_i of G necessary to describe the specification probabilities in site i (see Sections 2 and 3 for details). G_i is called the interaction neighborhood of i and the points in G_i are said to interact with i. G_i is not necessarily finite but only a finite subset $V_M \subset G$ of sites is observed. The observation set is a sample $X_{1:n}(V_M) = (X_1(j), ..., X_n(j))_{j \in V_M}$, where $(X_1, ..., X_n)$ are i.i.d with common law P. The question is then to recover from $X_{1:n}(V_M)$, for all i in V_M, the sets $G_i \cap V_M$.

Oracle neighborhood problem (ON):

The ON problem is to identify, for all i in G, a set $\hat{G}_i = \hat{G}_i(X_{1:n}(V_M))$, such that the estimation of the conditional probabilities $P\left(x(i)|x(j), j \in G/\{i\}\right)$ by the empirical conditional probabilities $\widehat{P}(x(i)|x(j), j \in \hat{G}_i)$ has a minimal risk (see Sections 2 and 3 for details). $\hat{G}_i$ is then said to satisfy an oracle inequality and it is also called oracle. We look for oracles among the subsets of V_M and we consider the L_∞-distance between conditional probabilities to measure the risk of the estimators. An oracle is in general smaller than G_i because it should balance approximation properties and parsimony.

The literature has mainly been focused in the INI problem, see [3, 7, 10, 11, 17] for examples. It requires in general strong assumptions on P to be solved. For example, the ℓ_1-penalization procedure proposed in [17] requires an incoherence assumption on the interaction neighborhoods that is very restrictive, as shown by [3]. Moreover, it is assumed in [3, 7, 17] that G is finite and that all the sites are observed, *i.e.* $V_M = G$. Csiszar and Talata [10] consider the case when $G = \mathbb{Z}^d$ but assume a uniform bound on the cardinality of G_i. The procedure proposed in [11] holds for infinite graph with each site having infinite neighborhoods, but requires that the main interactions belong to a known neighborhood

of i of order $O(\ln n)$. Moreover, the result is proved in the Ising model only when the interaction is sufficiently weak.

The first goal of this paper is to show that the ON problem can be solved without any of these hypotheses. We introduce in Section 3.2 a model selection criterion to choose a model $\hat{G}_i$ and prove that it is an oracle in Theorem 3.2. This result does not require any assumption on the structure of the interaction neighborhoods inside or outside V_M.

The second objective is to show that a selection rule provides also a useful tool to handle the INI problem. We introduce the following two steps procedure. First, we select, for all sites i in V_M, a small subset $\hat{V}_i$ of V_M with the model selection rule. We prove that this set contains the main interacting points inside V_M with large probability. Following the idea introduced in [11], we use then a test to remove from $\hat{V}_i$ the points of $(G/G_i) \cap \hat{V}_i$. The new test can be applied to all neighborhoods V_i that are smaller than $O(\ln n)$ and that contain the main interaction points in G_i. It requires less restrictive assumptions on the interactions outside V_i and on the measure P than the one of [11]. For example, it works in the Ising models without restrictions on the temperature parameter. Furthermore, the two-step method let us look for the interacting points inside all the observation set V_M (of order $O(e^{n^\beta})$ for some $0 \leq \beta < 1$), and not only inside a prior subset V_i (smaller than $O(\ln n)$) of V_M.

All the results hold under a key assumption **H1** that is not classical, but that is satisfied by Ising models, see Theorem 4.5. We obtain then a large class of models, widely used in practice, where our methods are efficient. We also provide for this model a computationally efficient version of our main algorithms.

The paper is organized as follows. In Section 2, we introduce notations and assumptions used all along the paper. Section 3 gives the main results, in a general framework. Section 4 shows the application to Ising models and Section 5 presents a large simulation study where the problem of the practical calibration of some parameters is addressed. Section 6 is a discussion of the results with an extensive comparison to existing papers. Section 7 gives the proofs of the main theorems and some technical results are recalled in an appendix in Section 9.

2. Notations and Main Assumptions

Let G be a discrete set of *sites*, possibly infinite, $A = \{-1, 1\}$ be the binary alphabet of *spins*, and P be a probability measure on the set of *configurations* $\mathcal{X}(G) = A^G$. More generally, for all subsets V of G, let $\mathcal{X}(V) = A^V$ be the set of configurations on V. In what follows, the triplet (G, A, P) will be called a *random field*. For all i in G, for all $V \subset G$, for all x in $\mathcal{X}(G)$, let $x(V) = (x(j))_{j \in V}$ and for all probability measures Q on $\mathcal{X}(V \cup \{i\})$, let

$$Q_{i|V}(x) = Q(x(i)|x(V/\{i\}))$$

be a regular version of the conditional probability. All along the paper, we will use the convention that, if V is a finite set, Q a probability measure on $\mathcal{X}(V)$ and x is a configuration such that $Q(x(V/\{i\})) = 0$, then $Q_{i|V}(x) = 1/2$.

For all x in $\mathcal{X}(G)$ and all j in G, let x_j be the configuration such that $x_j(k) = x(k)$ for all $k \neq j$ and $x_j(j) = -x(j)$. We say that there is a *pairwise interaction* from j to i if there exists x in $\mathcal{X}(G)$ such that $P_{i|G}(x_j) \neq P_{i|G}(x)$. For all subsets V of G, for all probability measures Q on $\mathcal{X}(V)$, let

$$\omega_{i,j}^V(Q) = \sup_{x \in \mathcal{X}(G)} \left\{ Q_{i|V}(x) - Q_{i|V}(x_j) \right\}.$$

With the above notations, there is a pairwise interaction from j to i if and only if $\omega_{i,j}^G(P) > 0$. Our second task in this paper is to recover the set G_i of sites having a pairwise interaction with i. This definition differs in general from the one suggested in introduction. However, it is easy to check that they coincide in the Ising models defined in Section 4.

Let $X_{1:n} = (X_1, ..., X_n)$ be i.i.d. with common law P. Let V_M be a finite subset of G of *observed sites*, with cardinality M. The *observation set* is then $X_{1:n}(V_M) = (X_1(V_M), ..., X_n(V_M))$. Let $\widehat{P}$ be the *empirical measure* on $\mathcal{X}(G)$ defined for all configurations x in $\mathcal{X}(G)$ by

$$\widehat{P}(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{X_i(G) = x(G)\}}.$$

For all real valued functions f defined on $\mathcal{X}(G)$, let $\|f\|_\infty = \sup_{x \in \mathcal{X}(G)} |f(x)|$. For all subsets V of V_M, the L_∞-*risk* of $\widehat{P}_{i|V}$ is defined by $\left\| \widehat{P}_{i|V} - P_{i|G} \right\|_\infty$. This risk is naturally decomposed into two terms. From the triangular inequality, we have

$$\left\| \widehat{P}_{i|V} - P_{i|G} \right\|_\infty \leq \left\| \widehat{P}_{i|V} - P_{i|V} \right\|_\infty + \left\| P_{i|V} - P_{i|G} \right\|_\infty.$$

We call variance term the random term $\left\| \widehat{P}_{i|V} - P_{i|V} \right\|_\infty$ and bias term the deterministic one $\left\| P_{i|V} - P_{i|G} \right\|_\infty$.

Let us finally present our general assumptions on the measure P. In the following ν and $\kappa_{\min}$ are positive constants. The two first assumptions are classical and will only be used to discuss the main results.

NN: *(Non-Nullness) For all x in $\mathcal{X}(G)$, $\nu^{-1} \leq P_{i|G}(x)$.*

CA:*(Continuity) For all growing sequences $(V_n)_{n \in \mathbb{N}^*}$ of subsets of G such that $\cup_{n \in \mathbb{N}^*} V_n = G$, for all i in G,*

$$\lim_{n \to \infty} \left\| P_{i|V_n} - P_{i|G} \right\|_\infty = 0.$$

The following last assumption is very important for the model selection criterion to work. It is satisfied for example by a generalized form of the Ising model as we will see in Section 4.

H1: *For all finite subsets V of G,*

$$\kappa_{\min} \left\| P_{i|G} - P_{i|V} \right\|_\infty \leq \left\| P_{i|G} \right\|_\infty - \left\| P_{i|V} \right\|_\infty.$$

3. General results

3.1. Control of the variance term of the L_∞-risk

Our first theorem provides a sharp control of the variance term of the risk of $\widehat{P}_{i|V}$. It holds without assumption on the measure P or the finite subset V.

Theorem 3.1. *Let P be a probability measure on $\mathcal{X}(G)$, let V be a finite subset of G. Let $p_-^V = \inf_{x\in\mathcal{X}(G),\ P(x(V))\neq 0} P(x(V))$. There exists an absolut constant c_1 such that, for all $\delta > 1$,*

$$P\left(\left\|\widehat{P}_{i|V} - P_{i|V}\right\|_\infty > c_1\sqrt{\frac{\ln(\delta/p_-^V)}{np_-^V}}\right) \leq \frac{1}{\delta}. \tag{1}$$

Moreover, let $\widehat{p}_-^V = n^{-1} \vee \inf_{x\in\mathcal{X}(V)} \widehat{P}(x(V/\{i\}))$. There exists an absolut constant $c_2 \leq 400$ such that, for all $\delta > 1$,

$$P\left(\left\|\widehat{P}_{i|V}(x) - P_{i|V}(x)\right\|_\infty > c_2\sqrt{\frac{\ln(\delta n)}{n\widehat{p}_-^V}}\right) \leq \frac{1}{\delta}. \tag{2}$$

Remark:

- Let $|V|$ denote the cardinality of V, if P satisfies **NN** we have $p_-^V \geq \nu^{-|V|}$. Hence, (1) implies that,

$$P\left(\left\|\widehat{P}_{i|V} - P_{i|V}\right\|_\infty \leq c_1\sqrt{\nu^{|V|}\frac{|V|\ln(\nu) + 2\ln(n)}{n}}\right) \geq 1 - n^{-2}.$$

  The variance term goes almost surely to 0 if $\nu^{|V|} << n(\ln n)^{-1}$. If in addition P satisfies **CA** and $(V_n)_{n\in\mathbb{N}^*}$ is a growing sequence of sets with limit G, the estimator $\widehat{P}_{i|V_n}$ is consistent.
- (1) is only interesting theoretically, because the parameter p_-^V is unknown in practice. We will use (2) for our model selection algorithm.

3.2. Model Selection

We deduce from Theorem 3.1 that the risk of the estimator $\widehat{P}_{i|V}$ is bounded in the following way. For all $\delta > 1$, for all subsets V,

$$P\left(\left\|P_{i|G} - \widehat{P}_{i|V}\right\|_\infty \leq \left\|P_{i|G} - P_{i|V}\right\|_\infty + c_2\sqrt{\frac{\ln(\delta n)}{n\widehat{p}_-^V}}\right) \geq 1 - \delta^{-1}. \tag{3}$$

The risk of $\widehat{P}_{i|V}$ depends on the approximation properties of V through the bias $\left\|P_{i|G} - P_{i|V}\right\|_\infty$ that is typically unknown in practice, and on the complexity of V, measured here by $\widehat{p}_-^V$. The aim of this section is to provide model selection

procedures in order to select a subset of V_M that optimizes the bound (3). In the following, we denote by $\mathcal{G}_n$ a finite collection of subsets of V_M, possibly random, and we call optimal or oracle in $\mathcal{G}_n$, any subset $\hat{G} = \hat{G}(X_{1:n}(\cup_{V\in\mathcal{G}_n} V))$ in $\mathcal{G}_n$, possibly random, such that,

$$P\left(\left\|P_{i|G} - \widehat{P}_{i|\hat{G}}\right\|_\infty \leq K \inf_{V\in\mathcal{G}_n}\left\{\left\|P_{i|G} - P_{i|V}\right\|_\infty + \sqrt{\frac{\ln(\delta n)}{n\widehat{p}_-^V}}\right\}\right) \geq 1-\delta^{-1}.$$

We introduce the following selection rule. Let N_n be an almost sure bound on the cardinality of $|\mathcal{G}_n|$. For all $\delta > 1$ and for all $C > c_2$, let

$$\hat{G}(C,\delta,\mathcal{G}_n) = \arg\min_{V\in\mathcal{G}_n}\left\{-\left\|\widehat{P}_{i|V}\right\|_\infty + C\mathrm{pen}(V)\right\}, \text{ where } \mathrm{pen}(V) \geq \sqrt{\frac{\ln(\delta n N_n)}{n\widehat{p}_-^V}}. \tag{4}$$

The following theorem states that $\hat{G}(C,\delta,\mathcal{G}_n)$ is almost an oracle.

Theorem 3.2. *Let P be a probability measure on $\mathcal{X}(G)$ satisfying* **H1**. *Let $\mathcal{G}_n$ be a finite collection of finite subsets of G, possibly random, and let N_n be an almost sure bound on the cardinality of $\mathcal{G}_n$. For all $C > c_2$, $\delta > 1$, let $\hat{G}_\delta(C) = \hat{G}(C,\delta,\mathcal{G}_n)$ be the estimator given by (4). There exists a positive constant $K = K(c_2, C, \kappa_{\min})$ such that,*

$$P\left(\left\|\widehat{P}_{i|\hat{G}_\delta(C)} - P_{i|G}\right\|_\infty \leq K \inf_{V\in\mathcal{G}_n}\left\{\left\|P_{i|G} - P_{i|V}\right\|_\infty + \mathrm{pen}(V)\right\}\right) \geq 1 - \frac{1}{\delta}.$$

Remarks:

- Theorem 3.2 states that the risk of the estimator selected by the rule (4) is the best among the collection $\mathcal{G}_n$. It is the main result of the paper and we will discuss in what follows several applications.
- The key idea of the proof is that, by assumption **H1**, we have $\left\|P_{i|G}\right\|_\infty - \left\|P_{i|V}\right\|_\infty \simeq \left\|P_{i|G} - P_{i|V}\right\|_\infty$, hence, our decision rule consists essentially in minimizing the sum of the bias term and the variance term of the risk, and the selected estimator is then an oracle.
- The constant c_2 derived in Theorem 3.1 is very pessimistic. Hence, Theorem 3.2 is more interesting theoretically. In the simulations of Section 5, we will calibrate C with the slope algorithm introduced in [5] and illustrate the nice properties of the resulting $\hat{G}_\delta(C)$.

Let us go back to the ON problem. It is solved thanks to the following corollary.

Corollary 3.3. *Let (G, A, P) be a random field. Let V_M is a finite subset of G with cardinality M, let $\delta > 1$ and let $\Gamma_M(\delta) = \ln(n)(1 + \log_2(M)) + \ln(\delta)$. For all m, $e \leq m \leq M$, let $\mathcal{G}_{m,M} = \{V \subset V_M,\ |V| \leq m\}$. For all $V \subset V_M$, let $\mathrm{pen}(V) = (n\widehat{p}_-^V)^{-1/2}\sqrt{\Gamma_M(\delta)}$, let $\hat{G}_\delta(C) = \hat{G}(C,\delta,\mathcal{G}_{\log_2(n),M})$ be the estimator given by (4). With probability larger than $1-\delta^{-1}$, we have*

$$\left\|\widehat{P}_{i|\hat{G}_\delta(C))} - P_{i|G}\right\|_\infty \leq K \inf_{V\subset V_M}\left\{\left\|P_{i|G} - P_{i|V}\right\|_\infty + \sqrt{\frac{\Gamma_M(\delta)}{n\widehat{p}_-^V}}\right\}.$$

Remarks:

- The risk of the estimator $\hat{G}_\delta(C)$ optimizes the bound given in (2) among all the subsets of V_M, up to the $\log(M)$ term, it solves therefore the **ON** problem.
- $\hat{G}_\delta(C)$ is built using only the subsets of V_M with size smaller than $\log_2(n)$. This is due to the fact that, for a subset of size m of V_M, there is 2^m different configurations, hence, when $m > \log_2(n)$, there is at least one configuration that is not observed, hence $\widehat{p}_-^V = n^{-1}$ and the bound given in (2) or in Corollary 3.3 is useless.
- The complexity of the model selection algorithm for the collection $\mathcal{G}_{m,M}$ is $O(nM^m)$. This collection is used when a uniform bound m on the cardinalities of the $|G_i|$ is known. The complexity is the minimal necessary to recover the interaction graph in this problem [7].

3.3. Estimation of the interaction subgraph

Let M be an integer and let V_M be a finite subset of G, with cardinality M. For all subsets V_M of G, let us choose $v_n^V(\delta) \geq \sqrt{\ln(\delta n)}(n\widehat{p}_-^V)^{-1/2}$. Let V be a finite subset of G, we study in this section the estimators of G_i given by

$$\hat{G}_i^V(c) = \left\{ j \in V,\ \omega_{i,j}^V(\widehat{P}) > c v_n^V(\delta) \right\}. \tag{5}$$

We introduce the following function.

$$\Psi(v) = \inf_{V,\, \widehat{p}_-^V \geq v^{-2}} \left\| P_{i|V} - P_{i|G} \right\|_\infty .$$

Ψ represents the minimal value of the bias term at a given value of the variance term. Our assumption concerns the rate of convergence of Ψ to 0.

H2(ϵ_Ψ): *There exist $C_\Psi > 0$, $\alpha_\Psi > 0$ such that, for all $K > 1$, for all $v > 0$,*

$$P\left(\Psi(Kv) \leq C_\Psi K^{-\alpha_\Psi} \Psi(v) \right) \geq 1 - \epsilon_\Psi .$$

Theorem 3.4. *Let (G, A, P) be a random field satisfying* **H1, H2**. *Let $e \leq M$ be an integer, let V_M be a finite subset of G with cardinality M. Let $\delta > 1$ and let $\Gamma_M(\delta) = \ln(n)(1 + \log_2(M)) + \ln(\delta)$. Let $\mathcal{G}_n = \{ V \subset V_M,\ |V| \leq (\log_2 n) \}$. For all V in $\mathcal{G}_n$, let*

$$v_n^V(\delta) = \sqrt{\frac{\Gamma_M(\delta)}{n\widehat{p}_-^V}}.$$

Let $C \geq c_2$, $\mathrm{pen}(V) = v_n^V(\delta)$ and let $\hat{G}_\delta(C) = \hat{G}(C, \delta, \mathcal{G}_n)$ be the set selected by the selection rule (4). Let $c > 0$ and let $\hat{G}_i^{\hat{G}_\delta(C)}(c)$ be the associated set defined by (5). Let K be the constant defined in Theorem 3.2 and let $c_\infty =$

$2\left(c_2 + C_\Psi^{-1/\alpha_\Psi}(2K)^{1-1/\alpha_\Psi}\right)$. *We have*

$$P\Big(\Big\{j \in V_M,\ \omega_{i,j}^G(P) \geq (c+c_\infty)v_n^{\hat{G}}(\delta)\Big\} \subset \hat{G}_i^{\hat{G}_\delta(C)}(c) \subset \Big\{j \in V_M,\ \omega_{i,j}^G(P) \geq (c-c_\infty)v_n^{\hat{G}}(\delta)\Big\}\Big) \geq 1-\delta^{-1}-\epsilon_\Psi.$$

Remark:

- When $c > c_\infty$, $\hat{G}_i^{\hat{G}_\delta(C)}(c)$ contains exactly the sites that have a pairwise interaction with i of order the risk of an oracle. It provides a partial solution to the INI problem.
- Theorem 3.4 requires the extra assumption **H2** compared to Theorem 3.2. Moreover, the theoretical constant c_∞ depends on the constants $\kappa_{\min}$, C_Ψ, α_Ψ.

Let us conclude this section with the two steps algorithm suggested by Theorem 3.4 to estimate $G_i = \{j \in G,\ \omega_{i,j}^G(P) > 0\}$.

Estimation algorithm:

- Choose a large subgraph V_M of G, typically the M nearest neighbors of i in G.
- **Selection step.** Choose a model $\hat{G}$, applying the model selection algorithm of Theorem 3.2 to the collection of all subgraphs of V_M with cardinality smaller than $\log_2(n)$.
- **Cutting step.** Cut the edges of $\hat{G}$ such that $\omega_{i,j}^{\hat{G}}(\widehat{P}) > c_\infty v_n^{\hat{G}}$.

4. Ising Models

The remaining of the paper is devoted to Ising models. These models are very important in statistical mechanics [12] and neuroscience [19] where they represent the interactions respectively between particles and neurons. In this section, we prove that Ising models satisfy **H1**, so that all our general results apply in these models. We also define effective algorithms for the ON and INI problems, adapted to this special case.

4.1. Verification of H1.

Let us recall the definition of Ising models.

Definition 4.1. *Let $f : G^2 \times A^2 \to \mathbb{R}$, $(i,j,a,b) \mapsto f_{i,j}(a,b)$ be a real valued function. For all i,j in G and all a in A, let $\|f_{i,j}^a\| = \max_{b\in A}|f_{i,j}(a,b)|$. f is said to be a pairwise potential of interaction if, for all a,b in A, $f_{i,i}(a,b) = 0$ and if*

$$r := \sup_{i\in G}\sup_{a\in A}\sum_{j\in G}\|f_{i,j}^a\| < \infty.$$

In this case, $T = r^{-1}$ is called the temperature parameter of the pairwise potential f.

Definition 4.2. *A probability measure P on $\mathcal{X}(G)$ is called an Ising model with potential f if, for all $x \in \mathcal{X}(G)$,*

$$P_{i|G}(x) = \frac{e^{\sum_{j\in G} f_{i,j}(x(i),x(j))}}{\sum_{a\in A} e^{\sum_{j\in G} f_{i,j}(a,x(j))}} = \frac{1}{1+e^{\sum_{j\in G} f_{i,j}(x_i(i),x(j))-f_{i,j}(x(i),x(j))}}.$$

The existence of a such a measure is well known [12].

Remark:

- The classical Ising model has potential f defined by $f_{ij}(a,b) = J_{ij}ab + H_i a\mathbf{1}_{\{i=j\}}$, $J_{ij} \in \mathbb{R}$, $H_i \in \mathbb{R}$, for all $a, b \in A$ and $i, j \in G$.
- One of the fundamental questions studied for this class of models is the description of conditions on potential f that guarantees uniqueness and non-uniqueness of the Ising model. Usually, high temperature implies conditions for the uniqueness of the Ising model and low temperature implies non-uniqueness [12].

Let $g_{i,j}(a,b) = f_{i,j}(a,b) - f_{i,j}(-a,b)$, we have then

$$P_{i|G}(x) = \frac{1}{1+e^{-\sum_{j\in G} g_{i,j}(x(i),x(j))}}.$$

It is clear that Ising models satisfy **CA** and **NN** with $\nu = (1+e^{2r})^{-1}$.

Definition 4.3. *Let (G, A, P) be an Ising model, with potential f. For all i, j in G, for all a in A, let*

$$\omega_{i,j}(f) = \sup_{(a,b)\in A^2} \{g_{i,j}(a,b) - g_{i,j}(a,-b)\} = \sup_{b\in A} \{g_{i,j}(a,b) - g_{i,j}(a,-b)\}.$$

Let us first recall some elementary facts about Ising models.

Proposition 4.4. *Let (G, A, P) be an Ising model, with potential f. For all finite subsets V of G, for all i, j in G, we have*

1. $p_-^V \geq (1+e^{2r})^{-|V|}$.
2. $\frac{2e^{-2r}}{(1+e^{2r})^2}\omega_{i,j}(f) \leq \omega_{i,j}^G(P) \leq \frac{e^{2r}(e^{4r}-1)}{4r(1+e^{-2r})^2}\omega_{i,j}(f)$.

The following theorem states that all of our general results apply in Ising models. The key ingredient of the proof is the precise control of the bias term (6).

Theorem 4.5. *Let (G, A, P) be an Ising model, with potential f. There exist two positive constants $c_r^* \leq C_r^*$ such that, for all subsets V of G,*

$$c_r^* \sum_{j\notin V} \omega_{i,j}(f) \leq \left\|P_{i|G} - P_{i|V}\right\|_\infty \leq C_r^* \sum_{j\notin V} \omega_{i,j}(f). \tag{6}$$

P satisfies assumption **H1** *i.e. there exists a constant $\kappa_{\min} > 0$ such that, for all finite subsets V of G,*

$$\kappa_{\min} \left\|P_{i|G} - P_{i|V}\right\|_\infty \leq \left\|P_{i|G}\right\|_\infty - \left\|P_{i|V}\right\|_\infty.$$

4.2. A special strategy for Ising models

The model selection algorithm (4) might be computationally demanding in practice when the collection $\mathcal{G}_n$ is too large. This is the case of the collection $\mathcal{G}_{\log_2(n),M}$ used several times in Section 3, when the values of M and n are large. The purpose of this section is to show that a special strategy, computationally more attractive, can be adopted in Ising models. The idea comes from [7]. Let us describe the method.

Reduction of the number of sites. Let x_1 be the configuration in $\mathcal{X}(G)$ such that, for all j in G, $x_1(j) = 1$.

Step 1 Computation of the empirical probabilities. For all j in V_M, let

$$\widehat{p}(j) = \widehat{P}(x_1(j)),\ \widehat{p}(i,j) = \widehat{P}(x_1(i,j)).$$

Step 2 Reduction step. We keep the j in V_M such that

$$|\widehat{p}(i,j) - \widehat{p}(i)\widehat{p}(j)| > \eta.$$

Let also η_{ms} be the smallest $\eta > 3\sqrt{(2n)^{-1}\ln(6M\delta)}$ such that the number of j kept after Step 2 is smaller than $\kappa \log_2(n)$.

We denote by $\widehat{V}(\eta)$ the set of j kept after Step 2. It is clear that the reduction algorithm has a complexity $O(nM)$. Remark that the values $|\widehat{p}(i,j) - \widehat{p}(i)\widehat{p}(j)|$ do not depend on the configuration x_1 since the alphabet has only two letters.

Model selection algorithm. Let $\mathcal{G} = \left\{ V \subset \widehat{V}(\eta_{ms}) \right\}$.

Step 1 Computation of the conditional probabilities. For all V in $\widehat{V}(\eta_{ms})$, compute $\left\|P_{i|V}\right\|$, and $\mathrm{pen}(V)$.

Step 2 Selection Step. We choose $C > c_2$ and

$$\hat{G} = \arg\min_{V \in \mathcal{G}} \left\{ -\left\|P_{i|V}\right\| + C\sqrt{\frac{\ln(n^\kappa \delta)}{n\widehat{p}_-^V}} \right\}.$$

It is clear that, if $\widehat{m} = |\widehat{V}(\eta_{ms})| \leq \kappa(\log_2(n))$, hence

$$\widehat{N} = |\mathcal{G}| = \sum_{k=0}^{\widehat{m}} C_{\widehat{m}}^k \leq 2^{\widehat{m}} \leq n^\kappa.$$

Hence, the complexity of the model selection algorithm is $O(n^{\kappa+1})$. The global complexity of the algorithm is therefore $O(n^{\kappa+1} + nM)$. As a comparison, the model selection algorithm for $\mathcal{G}_n = \mathcal{G}_{\log_2(n),M}$ was $O(nM + n^{\log_2(M)})$.

4.2.1. Control of the risk of the resulting estimator

Theorem 4.6. *Let (G, A, P) be an Ising model, with potential f. Let*

$$C_1 = \frac{4r(1+e^{2r})^3}{e^{-6r}(e^{4r}-1)}, C_2 = \frac{4r(1+e^{2r})^2}{e^{6r}(e^{4r}-1)}.$$

With probability larger than $1-\delta$ *we have that*

$$\left\{ j \in V_M,\ |\omega_{i,j}(f)| \geq C_1 \left(\eta + 3\sqrt{\frac{\ln(6M\delta)}{2n}} \right) \right\}$$
$$\subset \widehat{V}(\eta) \subset \left\{ j \in V_M;\ |\omega_{i,j}(f)| \geq C_2 \left(\eta - 3\sqrt{\frac{\ln(6M\delta)}{2n}} \right) \right\}.$$

Furthermore, let us denote by

$$V(\delta, M) = \left\{ j \in V_M;\ |\omega_{i,j}(f)| \leq C_1(\eta_{ms} + 3\sqrt{(2n)^{-1}\ln(6M\delta)} \right\}.$$

With probability larger than $1-2\delta$*, we have,*

$$\frac{1}{K}\left\| \widehat{P}_{i|\hat{G}} - P_{i|G} \right\| \leq \sum_{j \in V(\delta,M)} |\omega_{i,j}(f)| + \inf_{V \in \mathcal{G}} \left\{ \sum_{j \in \widehat{V}(\eta)/V} |\omega_{i,j}(f)| + \sqrt{\frac{\ln(n^{\kappa}\delta)}{n\widehat{p}_{-}^{V}}} \right\}.$$

Remarks:

- The estimator of the interaction graph has better properties than the one obtained with selection and cutting procedure. The main difference is that there is no term $(\widehat{p}_{-}^{\hat{G}})^{-1/2}$ in the rate of convergence.
- The oracle inequality might be a little bit less sharp than the one obtained in (19). This is the price to pay to have a computationally efficient algorithm.
- Our result holds in the Ising model. However, [7] used a similar approach in more general random fields with some additional assumptions and obtained good properties for the INI problem.

5. Simulation studies

In this section we illustrate results obtained in Sections 3 and 4 using simulation experiments and introduce the slope heuristic. All these simulation experiments can be reproduced by a set of MATLAB® routines that can be downloaded from www.princeton.edu/∼ dtakahas/publications/LT10routines.zip.
Let $G = \{-1, 0, 1\} \times \{-1, 0, 1\}$. For the sections 5.1, 5.2, 5.3, 5.4, 5.5, 5.6, and 5.7, we consider an Ising model on A^G, with pairwise potential given by $f_{ij}(c, d) = J\mathbf{1}_{j \in V_i} cd$ for $i, j \in G$, $c, d \in A$, $J = 0.2$, and $V_i \subset G$. The pair of sites (i, j) where $j \in V_i$ is shown in Figure 1. For all these experiments, $i = (0, 0)$. We simulated independent samples of the Ising model with increasing sample sizes $n = 100k$, $k = 1, \ldots, 100$. For each sample size we have $N = 100$ independent replicas.

5.1. Variance term of the risk

In the following experiment we will verify Theorem 3.1 in a simulation. For each sample size we computed the normalized variance term $\sqrt{n}\left\| \hat{P}_{i|V_i} - P_{i|V_i} \right\|_{\infty}$ for

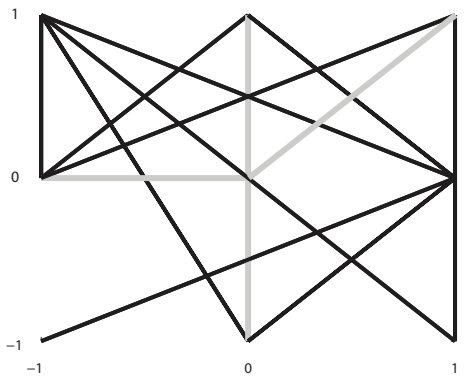

FIG 1. *Representation of the interacting pairs of the Ising model used in the simulation experiments. The edges between sites indicate the interacting pairs. The grey colored edges indicate the sites interacting with site* $(0, 0)$.

N different samples and obtained the average value. The result is summarized in Figure 2.

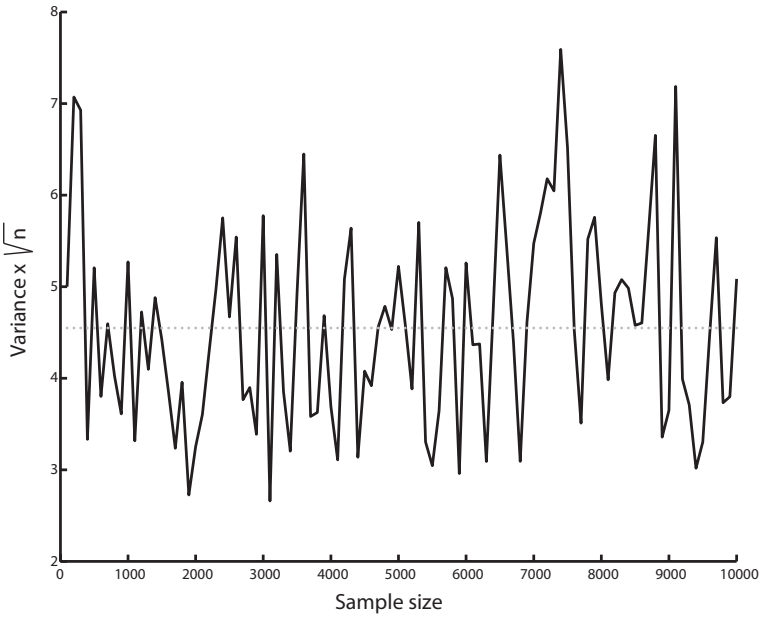

FIG 2. *Plot of the number of samples* n *against* $\sqrt{n}\left\|\hat{P}_{i|V_i} - P_{i|V_i}\right\|_\infty$. *The dotted line indicates the linear regression line. Observe that the regression line is essentially parallel to the abscissa.*

5.2. Slope heuristic

The constant c_2 derived from Theorem 3.1 is too pessimistic to be used in practice. The purpose of this section is to present a general method to design this constant. It is based on the slope heuristic, introduced in [5] and proved in several other frameworks in [1, 14]. We refer also to [2] for a large discussion on the practical use of this method. In order to describe it, let us introduce, for all V in $\mathcal{G}_{m,M}$, a quantity Δ_V, possibly random, measuring the complexity of the model V. The heuristic states the following facts.

1. There exists a positive constant $C_{\min}$ such that when $C < C_{\min}$, the complexity of the model selected by the rule (4) is as large as possible.
2. When C is slightly larger than $C_{\min}$ the complexity of the selected model is much smaller.
3. When $C = 2C_{\min}$ then the risk of the selected model is asymptotically the one of an oracle.

The heuristic yields the following algorithm, defined for all complexity measures Δ_V.

1. For all $C > 0$, compute $\Delta_{\hat{G}(C)}$, the complexity of the model selected by the rule (4).
2. Choose $\tilde{C}_{\min}$ such that $\Delta_{\hat{G}(C)}$ is very large for $C < \tilde{C}_{\min}$ and much smaller for $C > \tilde{C}_{\min}$.
3. Select the final $\hat{G} = \hat{G}(2\tilde{C}_{\min})$.

The algorithm is based on the idea that $\tilde{C}_{\min} \simeq C_{\min}$ and therefore that the final $\hat{G}$, selected by $2\tilde{C}_{\min}\Delta_V$ is an oracle by the third point of the slope heuristic. The actual efficiency of this approach depends highly on the choice of the complexity measure Δ_V and on the practical way to choose $\tilde{C}_{\min}$ in step 2 of the algorithm. We illustrate the dependence on Δ_V in the following experiences.

Δ_V is either the cardinality of V (the dimension) or the variance estimator $C(n\hat{p}_-^V)^{-1/2}$. $\tilde{C}_{\min}$ is selected with the maximum jump criteria [2]: fix an increasing sequence of positive numbers $C_0, \ldots, C_t$ and define

$$k = \arg\max_i \left\{ \Delta_{\hat{G}(C_i)} - \Delta_{\hat{G}(C_{i-1})} \right\}, \text{ and } \tilde{C}_{\min} = C_k.$$

If the maximum is achieved in more than one value, take the biggest of such k.

Remark: The calculation of $\tilde{C}_{\min}$ does not yield a significant increase of computational time compared to the evaluation of the model selection criteria for one fixed constant C. The only additional cost is due to the fact that one has to keep in the computer memory the conditional probabilities that must be computed only once.

5.3. Oracle risk compared to the risk of the estimated model

One way to verify the performance of the slope heuristic proposed in previous section is to compute the ratio

$$\frac{\left\|\hat{P}_{i|\hat{G}(2\tilde{C}_{\min})} - P_{i|G}\right\|_\infty}{\inf_{V\subset G}\left\|\hat{P}_{i|V} - P_{i|G}\right\|_\infty}. \tag{7}$$

With a reasonable procedure, we expect that the above quantity remains bounded. We applied the model selection procedure (4) with slope heuristic discussed above for the set $\{V \subset G \setminus \{i\} : |V| \leq 8\}$. For each sample size we computed the ratio (7) for 100 different samples and we obtained the average. The result is summarized in Figure 3.

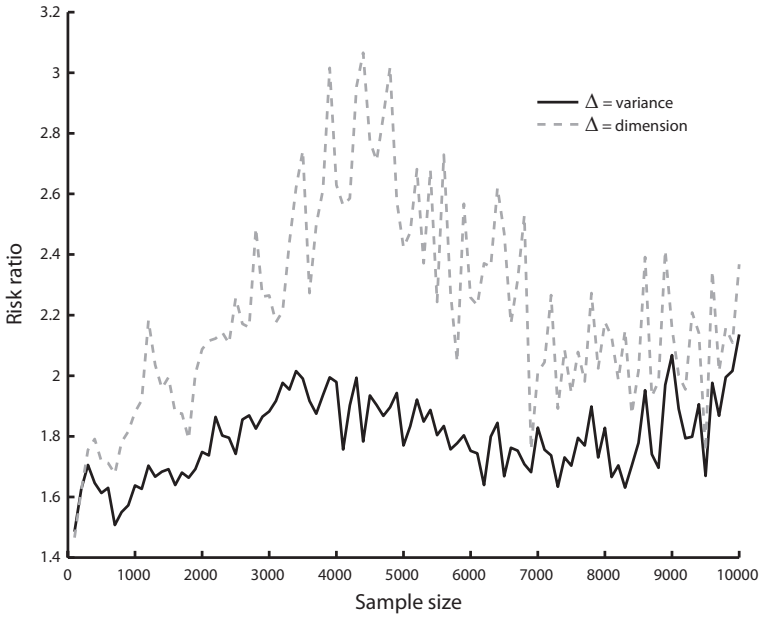

FIG 3. *Plot of the number of samples n against the average of ratio (7). Observe that the risk ratio remains bounded for both the variance (solid black) and the dimension (dashed grey) as the measure of complexity.*

5.4. Discovery rate of the model selection procedure for ON problem

Another way to measure the performance of our model selection procedure is to compute the positive discovery rate

$$\mathbb{E}\left[\frac{|\hat{G}(2\tilde{C}_{\min}) \cap \hat{G}_{oracle}|}{|\hat{G}_{oracle}|}\right] \tag{8}$$

and the negative discovery rate

$$\mathbb{E}\left[\frac{|G \setminus \left(\hat{G}(2\tilde{C}_{\min}) \cup \hat{G}_{oracle}\right)|}{|G \setminus \hat{G}_{oracle}|}\right]. \tag{9}$$

with respect to the oracle $\hat{G}_{oracle}$.
We estimated (8) and (9) and the result is summarized in Figure 4.

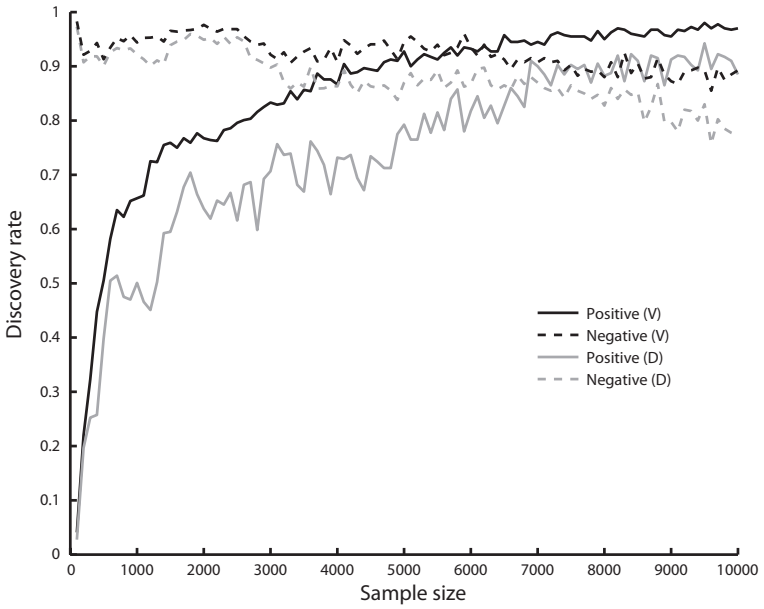

FIG 4. *Plot of positive and negative discovery rates with respect to the oracle against the sample size n. In solid/dashed black lines are represented the positive/negative discovery rates using the variance (V) as the complexity measure and in solid black/grey lines the positive/negative discovery rates using the dimension (D). Observe that the variance gives a better positive and negative discovery rates with respect to oracle when compared to the dimension.*

5.5. Performance of the model selection procedure for INI problem

A natural question is how well the proposed model selection procedure behaves for the INI problem. Observe that the model selection procedure was designed to solve the ON problem and in principle does not necessary work for the INI problem. To investigate this question for each sample size we estimated the positive discovery rate

$$\mathbb{E}\left[\frac{|\hat{G}(2\tilde{C}_{\min}) \cap V_i|}{|V_i|}\right]$$

and the negative discovery rate

$$\mathbb{E}\left[\frac{|G \setminus \left(\hat{G}(2\tilde{C}_{\min}) \cup V_i\right)|}{|G \setminus V_i|}\right],$$

with respect to the interaction neighborhood V_i. The result is summarized in Figure 5.

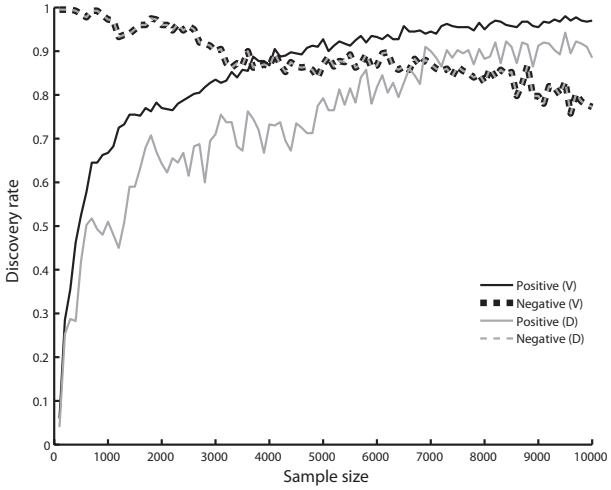

FIG 5. *Plot of positive and negative discovery rates with respect to V_i against the sample size n. In solid/dashed black lines are represented the positive/negative discovery rates using the variance (V) as the complexity measure and in solid black/grey lines the positive/negative discovery rates using the dimension (D). Observe that the variance gives higher positive discovery rates than the dimension as the measure of complexity although the negative discovery rates are the same.*

5.6. Relationship between the INI and ON problems

Another interesting question is to understand what is the relationship between the INI and ON problems. Useful quantities for this are the positive discovery rate

$$\mathbb{E}\left[\frac{|\hat{G}_{oracle} \cap V_i|}{|V_i|}\right] \tag{10}$$

and the negative discovery rate

$$\mathbb{E}\left[\frac{|G \setminus \left(\hat{G}_{oracle} \cup V_i\right)|}{|G \setminus V_i|}\right]. \tag{11}$$

We estimated these quantities and the results are summarized in Figure 6.

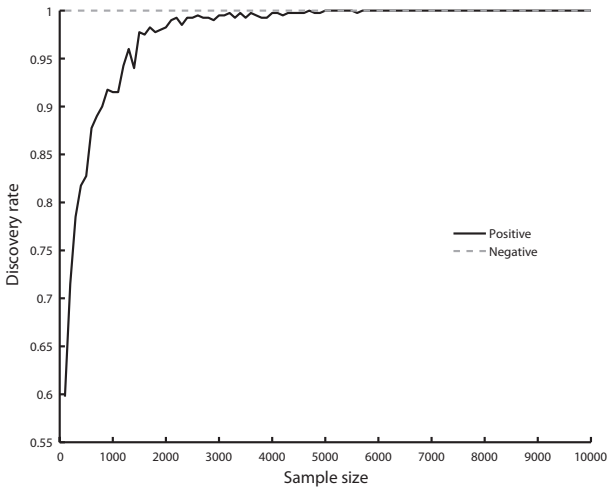

FIG 6. *Plot of positive and negative discovery rates of the oracle with respect to V_i against the sample size n. The solid black line represents the results for positive discovery rates and the dashed grey line represents the results for the negative discovery rates. Observe that in this example the oracle $\hat{G}_{oracle}$ matches the interaction neighborhood V_i quite fast. Also observe that in this example the oracle never included interactions not contained in V_i.*

5.7. *Select and cut procedure*

Here we will show the usefulness of the two-step procedure introduced in Theorem 3.4 by an example. We consider the same independent samples used in previous experiments. We also consider $i = (0, 0)$ and sample sizes $n = 100k$, $k = 1, \ldots, 100$ with 100 independent replicas for each sample size.

Let $\hat{G}(2\tilde{C}_{\min})$ be the subset of G chosen by first applying the model selection procedure for the set $\{V \subset G \setminus \{i\} : |V| \leq 8\}$. To choose the constant in the model selection procedure, we used the slope heuristic with variance as the complexity measure. Let $\hat{G}(SC)$ be the subset of G obtained by applying to the subset $\hat{G}(2\tilde{C}_{\min})$ the cutting procedure with $cv_n^V = 0.3(n\widehat{p}_-^V)^{-1}$. We first computed the average of the risk ratio

$$\frac{\left\|\hat{P}_{i|\hat{G}(SC)} - P_{i|G}\right\|_\infty}{\inf_{V \subset G}\left\|\hat{P}_{i|V} - P_{i|G}\right\|_\infty}. \tag{12}$$

for each sample size and compared them with the average of risk ratio (7). The results are summarized in Figure 7.

We also computed the positive and negative discovery rates of $\hat{G}(SC)$ and $\hat{G}(2\tilde{C}_{\min})$ with respect to V_i. The results are presented in Figure 8.

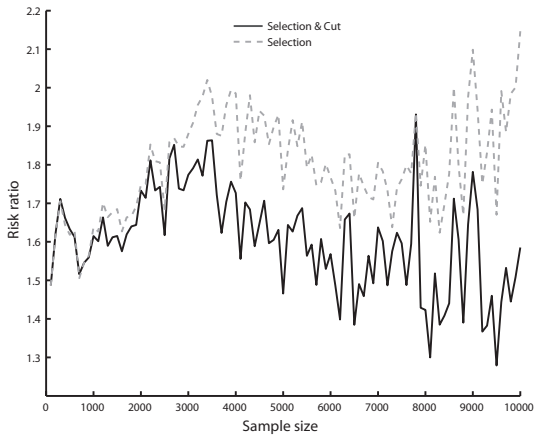

FIG 7. *Plot of the number of samples n against the average of risk ratio (12) and (7). In solid black is represented the risk ratio for the two-step procedure and in dashed grey the risk ratio for the model selection procedure alone. Observe that the risk ratio of the two-step procedure remains closer to one when compared to the model selection alone.*

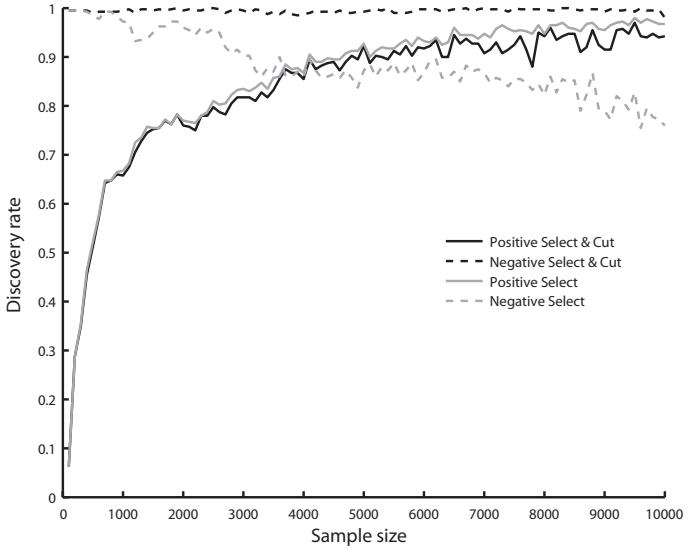

FIG 8. *Plot of positive and negative discovery rates of $\hat{G}(SC)$ and $\hat{G}(2\tilde{C}_{\min})$ with respect to V_i against the sample size n. The black solid/dashed lines represent the positive/negative discovery rates of the two-step procedure. The grey solid/dashed lines represent the positive/negative discovery rates of the model selection procedure alone. Observe that the two-step procedure has almost perfect negative discovery rates with increasing positive discovery rates.*

5.8. Computationally efficient algorithm

In this section we will illustrate the performance of the strategy introduced in Section 4.2 on the Ising model on A^G, where $G = \{1, \dots, 200\}$, with pairwise potential $f_{ij}(c,d) = |J_{ij}|\mathbf{1}_{j\in V_i}cd$ for $i, j \in G$, $c, d \in A$, $V_i \subset G$, and J_{ij} independently generated from a Gaussian distribution with $\mathbb{E}[J_{ij}] = 0$ and $\mathbb{E}[J_{ij}^2] = 4$. The pairs of sites (i, j) with $j \in V_i$ are represented in Figure 9.

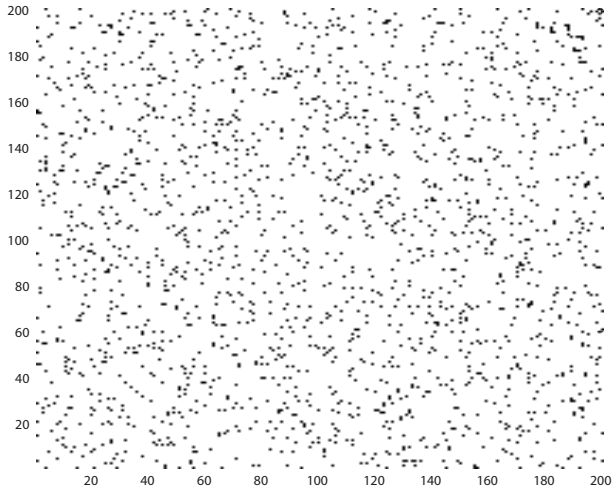

FIG 9. *Representation of the interacting sites in the Ising model described in 5.8. The positions (i, j) of the dots indicate the pair of sites (i, j) for which $j \in V_i$.*

For this experiment $i = 1$ and $|V_i| = 16$. We simulated independent samples of the Ising model with increasing sample sizes $n = 100k$, $k = 1, \dots, 100$. For each sample size we have $N = 50$ independent replicas. In this example, it is not practical to compute all candidates in collection $\mathcal{G}_{8,200}$ whereas the algorithm introduced in Section 4.2 is very efficient. We illustrate its performance in the case where the number of sites j kept after Step 2 of the reduction step in Section 4.2 is 10. We denote the model chosen by this algorithm by $\hat{G}_{\text{efficient}}$ We estimated the probability that the selected model $\hat{G}_{\text{efficient}}$ recover the largest, and second, third, fourth, fifth largest interaction potentials. Formally, let $\mathcal{J}_1 = \max\{|J_{ij}|\mathbf{1}_{j\in V_i} : i, j \in G\}$ and $\mathcal{J}_k = \max\{|J_{ij}|\mathbf{1}_{j\in V_i} : i, j \in G \setminus \mathcal{J}_{k-1}\}$, for $k = 1, \dots, 5$. We estimated

$$P(\hat{G}_{\text{efficient}} \ni \mathcal{J}_k), \tag{13}$$

for $k = 1, \dots, 5$. The result of the simulation is presented in Figure 10. By Monte Carlo simulation using a sample size of 100 000 we concluded that the considered Ising model at site $i = 1$ does not satisfy the incoherence condition in [17].

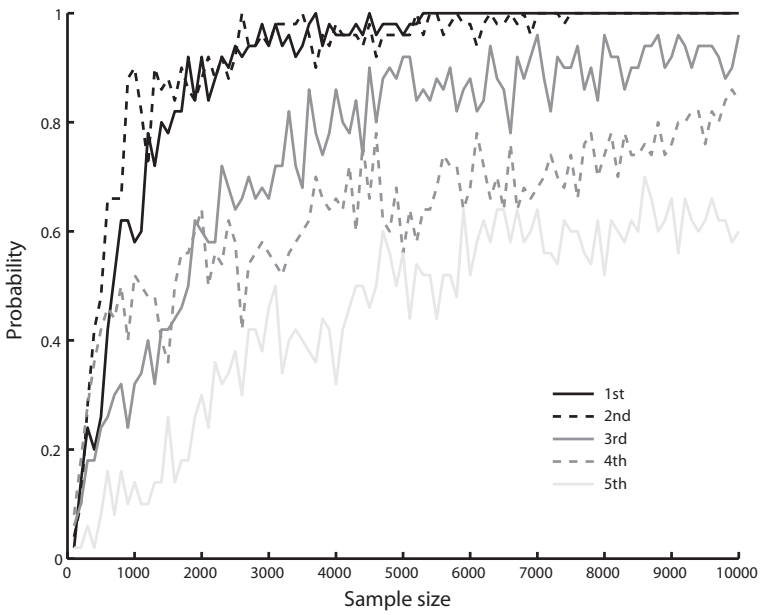

FIG 10. *Plot of the number of samples n against the probability that $\hat{G}_{efficient}$ includes the largest (solid black), and second (dashed black), third (solid gray), fourth (dashed gray), fifth (solid light gray) largest interaction potentials. Observe that the model selection procedure includes the sites with larger interaction potentials more often.*

6. Discussion and comparison with existing results

We introduced a model selection procedure for interaction neighborhood estimation in partially observed random fields. We proved that the proposed rule satisfies an oracle inequality. The results hold under general assumptions, which for instance, are satisfied by a generalized form of the Ising model.

Our model selection approach differs from other works [3, 7, 10, 11, 17] where only the INI problem is considered and more restrictive conditions are assumed. In particular, [3, 7, 17] consider the INI problem for finite random fields and assume that all the interacting sites are observed. This assumption is quite strong from practical point of view, *e.g.* in neuroscience, where the experimenter never has access to the whole set of neurons. Our result holds for partially observed random fields without any restriction on the range of the interactions.

Csiszar and Talata [10] considered a BIC like consistent model selection procedure for a homogeneous (translation invariant one point specification), finite range random field on $\mathbb{Z}^d$ based on one realization of the random field. Our main motivation is an application in neuroscience, where a priori we cannot assume homogeneity of interaction neither that every interacting sites are observed. Therefore, we consider a neighborhood estimation problem of an inhomogeneous, infinite range random field on arbitrary countable graph based on

n realizations of the random field. Because of these differences, the comparison between [10] and our work is not straightforward, but some differences are noteworthy.

1. Csiszar and Talata [10] consistency result is asymptotic whereas all our results are non-asymptotic and holds for all n.
2. Because they consider the finite range interaction random fields, the interaction neighborhood of each site will be included for large enough size of the the observed sites. As we consider the infinite range case, we have to have a control on the interactions between non-observed sites.
3. As stated in the discussion of [10], they don't exhibit a computationally efficient algorithm to compute the proposed statistics. We show a computationally efficient algorithm when the model is a generalized form of the Ising model.
4. The number of observed sites $|\Lambda|$ in [10] is the analogous quantity for the number of samples n in our article. From Theorem 2.1 in their article, this implies that the maximum size of the neighborhood is $o(\log^{1/2} n)$ which is selected from the $o(\log^{1/2} n)$ closest sites. Our model selection algorithm can be applied in high dimension situation and allows maximum neighborhood size of $O(\log n)$ selected from $O(e^{n^\beta})$, $0 \leq \beta < 1$, possible sites.

Nevertheless, it is interesting to notice that their estimator is a penalized estimator and we can wonder if the BIC estimator might have oracle properties. The risk naturally associated to their procedure is not the L^∞-risk, as in this paper, but rather the Küllback loss, defined by

$$K(P_{i|G}, \widehat{P}_{i|V}) = \mathbb{E}\left(\ln\left(\frac{P(X(i)|X(j),\ j \neq i)}{\widehat{P}(X(i)|X(j),\ j \in V/\{i\})}\right)\right).$$

A complete study of the ON problem for Küllback loss, with a comparison with the approach in L^∞-risk presented here, is beyond the scope of this paper. However, we believe that this problem is of interest and we will address these questions in a forthcoming paper. Let us mention that Theorem 3.1 is related to the typicality result of [10] (Proposition 3.1 in this paper). Both results give the same asymptotic estimate for the L^∞-risk of the estimator $\widehat{P}_{i|V}$. Our Theorem 3.1 gives a non-asymptotic bound and holds even if V does not contain all the interacting points. An important difference between our work and [10] is that we work with independent realizations, whereas they have only one realization and they have to divide their observation set to obtain conditionally independent data. We also mention here that the typicality result is the starting point for the analysis in Küllback loss presented in [10].
In [11], it is considered the INI problem for infinite range Ising models in $\mathbb{Z}^d$. The main restriction in this last work is that it is assumed that the interactions between the sites are weak ("high temperature") and that a subset of the observed sites of size $O(\log(n))$, where n is the sample size, must be fixed to apply the proposed procedure. Our procedure has no restriction on the strength of

interaction and can be applied for example for low temperature Ising models, provided that the samples come from the same phase.

In [7] the analysis is restricted to finite random fields, where the maximum neighborhood size is known a priori. For infinite range random fields, their results are useless since the "constants" ϵ and δ, that should be positive, are both equal to 0 in general. More importantly, their procedure use the knowledge of the lower bound ϵ on the bias term. As this ϵ is unknown in practice, it is not clear how it should be evaluated. Our general result on neighborhood estimation Theorem 3.4 suffers the same kind of drawback since the constant c should be chosen larger than the constant c_∞, which depends on our assumptions, for our test to be efficient. However, in the case of Ising models, we have been able to remove this condition and propose in Theorem 4.6 a totally data-driven, efficient procedure. It is not straightforward to generalize their result to the case where the maximum neighborhood size is allowed to increase with n, this would probably require a careful analysis of the behavior of the quantities ϵ and δ. Nevertheless, in the specific case when the underling random field is the Ising model, from Theorem 3 in [7], it is not difficult to show that when the number of total sites is $O(e^{n^\beta})$, $0 \leq \beta < 1$, the maximum size of the allowed neighborhood is $O(\log n)$. We provide in Table 1 a comparative summary of the available results.

TABLE 1
Compartive table of related results

— Algorithm	This work	[7]	[10]	[11]	[17]
Infinite range interactions	yes	no	no	yes	no
Type of problem	ON/INI	INI	INI	INI	INI
Partially observed interacting sites	yes	yes, strong conditions	no	yes, high temperature	no
Ising model on $\mathbb{Z}^d$ below critical temperature	yes	no	yes	no	no
Restrictions on the interaction graph	none	none	none	none	Incoherence condition
Computationally efficient	yes, Ising at any temperature	yes, fast decay of correlation	no	yes	yes
Maximal size of the neighborhoods	$O(\ln n)$	$O(\ln n)$	$o(\ln n)$	$O(\ln n)$	$O(n^\alpha)$
Number of possible sites for the candidate neighborhoods	$O(e^{n^\beta})$	$O(e^{n^\beta})$	$o(\ln n)$	$O(\ln n)$	$O(e^{n^\beta})$

We also introduced a two-step procedure in which the model selection rule gives us a small set of candidate sites and a cutting procedure removes from this set the irrelevant interactions. This two step procedure can be understood as a combination of a model selection and a statistical test procedure in spirit

of [22].

Our first simulation experiment shows that the concentration bound for the variance term of the risk in Theorem 3.1 is sharp. We propose a slope heuristic with maximal jump criteria using the variance or the dimension as a measure of complexity to choose a good constant in the model selection procedure. In our simulation experiment, we measured the performance of the slope heuristic for ON and INI problems. We observed that the variance had a better behavior as a complexity measure than the dimension because 1) the risk ratio was always smaller for the variance compared to the dimension as the measure of complexity, although both risk ratios remained bounded, 2) the estimated positive and negative discovery rates with respect to the oracle were always higher for the variance compared to the dimension as the measure of complexity, 3) also the estimated positive and negative discovery rates with respect to the interaction neighborhood were always higher for the variance compared to the dimension as the measure of complexity.

Although at this point the variance seems to be a better choice for the complexity measure, a more comprehensive study must be carried to obtain a definitive conclusion and we recommend to consider both measures of complexity in practice. We addressed also in the simulation experiments the relationship between the ON and INI problems and observed that for sufficiently large sample size, both coincide. The two-step procedure introduced in this article was applied in an example where it clearly enhances the performance of the model selection procedure for both the INI and ON problems. Recently, multistep statistical procedures are gaining attention [22] although only few rigorous results exist. Our result for the two-step procedure is a contribution for this growing field. The main drawback of the proposed model selection procedure is its high computational cost which becomes prohibitive when a large number of sites are observed. We introduced a computationally efficient way to overcome this difficulty in the case of Ising model. The new procedure drastically reduces the set of models for which the model selection procedure must be applied, but still keeping the main interacting sites and a good oracle property. In the simulation experiment we show that the proposed algorithm has a good performance even when the number of the observed sites is as big as 200. It must be remarked that the Ising model considered for this experiment does not satisfy the incoherence condition [17] and therefore other computationally efficient algorithms as ℓ_1-penalizations are not guaranteed to be consistent. Finally, we provide a set of MATLAB® routines that can be used to reproduce our experimental results and to carry further simulation and applied studies.

Acknowledgements

We would like to thank Antonio Galves and Roberto Imbuzeiro for many discussions during the period in which this paper was written. We are also grateful to the referee and associate editor for their comments which led to an improved presentation of the paper.

This work is part of USP project "Mathematics, computation, language and the brain".

7. Proofs

7.1. Proof of Theorem 3.1:

For all x such that $P(x(V/\{i\})) = 0$, we have $\widehat{P}(x(V/\{i\})) = 0$, thus $P_{i|V}(x) = \widehat{P}_{i|V}(x)$. Hence, we can only consider the configurations x such that $P(x(V/\{i\})) > 0$. Let us first provide some inequalities about conditional probabilities.

Lemma 7.1. *Let $x \in \mathcal{X}(G)$, let V be a finite subset of G and let Q, R be two probability measures on $\mathcal{X}(V)$ such that $R(x(V/\{i\})) > 0$.*

$$Q_{i|V}(x) - R_{i|V}(x) = \frac{Q(x(V)) - R(x(V)) + Q_{i|V}(x)\left(R(x(V/\{i\})) - Q(x(V/\{i\}))\right)}{R(x(V/\{i\}))}.$$

$$|Q_{i|V}(x) - R_{i|V}(x)| \leq 3 \sup_{x\in\mathcal{X}(G),\ R(x(V/\{i\}))\neq 0} \frac{|Q(x(V)) - R(x(V))|}{R(x(V/\{i\})}.$$

Remark: In particular, we deduce from this lemma that

$$\left\|\widehat{P}_{i|V} - P_{i|V}\right\|_\infty \leq 3 \sup_{x\in\mathcal{X}(G),\ P(x(V/\{i\}))\neq 0} \frac{\left|\widehat{P}(x(V)) - P(x(V))\right|}{P(x(V/\{i\}))}.$$

The first inequality follows from the fact that $R_{i|V}(x) = R(x(V))/R(x(V/\{i\}))$ and $Q(x(V)) = Q_{i|V}(x)Q(x(V/\{i\}))$. The second one is consequence of the first one and the fact that

$$|R(x(V/\{i\})) - Q(x(V/\{i\}))| \leq |R(x(V)) - Q(x(V))| + |R(x_i(V)) - Q(x_i(V))|.$$

The proof of (1) is concluded thanks to the following Lemma.

Lemma 7.2. *Let P be a probability measure on $\mathcal{X}(G)$ and let V be a finite subset of G. Let $\mathcal{X}'(V) = \{x \in \mathcal{X}(G),\ P(x(V/\{i\})) \neq 0\}$, $p_-^V = \inf_{x\in\mathcal{X}'(G)} P(x(V/\{i\}))$. For all $\delta > 1$, with probability larger than $1 - \delta^{-1}$, we have*

$$\sup_{x\in\mathcal{X}'(G)} \frac{\left|\widehat{P}(x(V)) - P(x(V))\right|}{P(x(V/\{i\}))} \leq 64\sqrt{2}\sqrt{\frac{\ln(16\delta/p_-^V)}{np_-^V}} + 2048\frac{\ln(16\delta/p_-^V)}{np_-^V}.$$

Conclusion of the proof of (1). We deduce from Lemmas 7.1 and 7.2 that, with probability larger than $1 - \delta^{-1}$,

$$\left\|\widehat{P}_{i|V} - P_{i|V}\right\|_\infty \leq 192\sqrt{2}\sqrt{\frac{\ln(16\delta/p_-^V)}{np_-^V}} + 6144\frac{\ln(16\delta/p_-^V)}{np_-^V}.$$

As this result is trivial when $\frac{\ln(16\delta/p_-^V)}{np_-^V} > 1$, we can always assume that $\frac{\ln(16\delta/p_-^V)}{np_-^V} \leq 1$, hence that $\frac{\ln(16\delta/p_-^V)}{np_-^V} \leq \sqrt{\frac{\ln(16\delta/p_-^V)}{np_-^V}}$, thus, with probability larger than $1-\delta^{-1}$, for $c_1 = 6144 + 192\sqrt{2}$,

$$\left\| \widehat{P}_{i|V} - P_{i|V} \right\|_\infty \leq c_1 \sqrt{\frac{\ln(16\delta/p_-^V)}{np_-^V}}.$$

Proof of Lemma 7.2: We apply Bousquet's version of Talagrand's inequality to the class of functions $\mathcal{F} = \{(P(x(V/\{i\})))^{-1} 1_{x(V)}\}$. This inequality is recalled in Appendix. We have $v^2 \leq (p_-^V)^{-1}$, $b \leq (p_-^V)^{-1}$, hence, for all $\delta > 1$, with probability larger than $1 - \delta^{-1}$,

$$\sup_{x \in \mathcal{X}'(G)} \frac{\left| \widehat{P}(x(V)) - P(x(V)) \right|}{P(x(V/\{i\}))} \leq 2\mathbb{E}\left(\sup_{x \in \mathcal{X}'(G)} \frac{\left| \widehat{P}(x(V)) - P(x(V)) \right|}{P(x(V/\{i\}))} \right) + \sqrt{\frac{2\ln(\delta)}{np_-^V}} + 2\frac{\ln(\delta)}{np_-^V}. \quad (14)$$

We apply Lemma 9.6 with $A_x = x(V)$, $x \in \mathcal{X}(G)$, $\alpha_x = [P(x(V/\{i\}))]^{-1}$. We have

$$\alpha^* = \sup_{x \in \mathcal{X}'(G)} [P(x(V/\{i\}))]^{-1} = \frac{1}{p_-^V},\ p^* = \sup_{x \in \mathcal{X}'(G)} [P(x(V/\{i\}))]^{-2} P(x(V)) \leq \frac{1}{p_-^V}.$$

Hence,

$$\mathbb{E}\left(\sup_{x \in \mathcal{X}'(G)} \frac{\left| \widehat{P}(x(V)) - P(x(V)) \right|}{P(x(V/\{i\}))} \right) \leq \frac{32\sqrt{2}}{\sqrt{np_-^V}} \sqrt{\ln\left(\frac{16}{p_-^V}\right)} + \frac{1024}{np_-^V} \ln\left(\frac{16}{p_-^V}\right). \quad (15)$$

Lemma 7.2 is then obtained with (14) and (15).

Let us now turn to the proof of (2). Let V be a finite subspace of S. As (2) holds when $\widehat{p}_-^V = n^{-1}$, it remains to prove (2) when, for all x in $\mathcal{X}(V)$, $\widehat{P}(\mathcal{X}(V)) > 0$. This is done by the following Proposition.

Proposition 7.3. *Let P be a probability measure on $\mathcal{X}(G)$, let V be a finite subset of G. Let $\mathcal{X}_n = \left\{ x \in \mathcal{X}(G),\ \widehat{P}(x(V/\{i\}) \neq 0 \right\}$, $\widehat{p}_-^V = \inf_{x \in \mathcal{X}_n} \widehat{P}(x(V))$. There exists an absolut constant $c_2 \leq 400$ such that, for all $\delta > 1$,*

$$P\left(\exists x \in \mathcal{X}_n,\ |\widehat{P}_{i|V}(x) - P_{i|V}(x)| > c_2 \sqrt{\frac{\ln(\delta n)}{n\widehat{P}(x(V))}} \right) \leq \frac{1}{\delta}. \quad (16)$$

In particular,

$$P\left(\sup_{x \in \mathcal{X}_n} |\widehat{P}_{i|V}(x) - P_{i|V}(x)| > c_2 \sqrt{\frac{\ln(\delta n)}{n\widehat{p}_-^V}} \right) \leq \frac{1}{\delta}.$$

Proof of Proposition 7.3. Let $n \geq 2$, $\delta > 1$, $c_2 = 400$ and let us first remark that we only have to prove (16) on the subset $\mathcal{X}_n' \subset \mathcal{X}_n$ of all the x in $\mathcal{X}_n$ such that $\widehat{P}(x(V)) \geq c_2^2 \ln(\delta n) n^{-1}$. Let x in $\mathcal{X}_n'$, then we also have $P(x(V/\{i\}) \neq 0$. From Lemma 7.1, we have

$$|\widehat{P}_{i|V}(x) - P_{i|V}(x)| \leq \frac{\left|\widehat{P}(x(V)) - P(x(V))\right| + \left|\widehat{P}(x(V/\{i\})) - P(x(V/\{i\}))\right|}{P(x(V/\{i\}))}.$$

From Lemma 7.1, we also have

$$|\widehat{P}_{i|V}(x) - P_{i|V}(x)| \leq \frac{\left|\widehat{P}(x(V)) - P(x(V))\right| + \left|\widehat{P}(x(V/\{i\})) - P(x(V/\{i\}))\right|}{\widehat{P}(x(V/\{i\})) \vee c_2^2 \ln(\delta n) n^{-1}}.$$

We deduce that

$$|\widehat{P}_{i|V}(x) - P_{i|V}(x)| \leq \frac{\left|\widehat{P}(x(V)) - P(x(V))\right| + \left|\widehat{P}(x(V/\{i\})) - P(x(V/\{i\}))\right|}{P(x(V/\{i\})) \vee \widehat{P}(x(V/\{i\})) \vee c_2^2 \ln(\delta n) n^{-1}}.$$

Hence, using the elementary inequality $a \vee b \geq \sqrt{ab}$ with $a = \widehat{P}(x(V/\{i\}))$, $b = P(x(V/\{i\})) \vee c_2^2 \ln(\delta n) n^{-1}$, we deduce that

$$|\widehat{P}_{i|V}(x) - P_{i|V}(x)| \leq \frac{\left|\widehat{P}(x(V)) - P(x(V))\right| + \left|\widehat{P}(x(V/\{i\})) - P(x(V/\{i\}))\right|}{\sqrt{\widehat{P}(x(V/\{i\}))\,(\,P(x(V/\{i\})) \vee c_2^2 \ln(\delta n) n^{-1}\,)}}.$$

We have obtain that, for all x in $\mathcal{X}_n'$,

$$\begin{aligned}
&\sqrt{\widehat{P}(x(V/\{i\}))}|\widehat{P}_{i|V}(x) - P_{i|V}(x)| \\
&\leq \frac{\left|\widehat{P}(x(V)) - P(x(V))\right| + \left|\widehat{P}(x(V/\{i\})) - P(x(V/\{i\}))\right|}{\sqrt{P(x(V/\{i\})) \vee c_2^2 \ln(\delta n) n^{-1}}} \\
&\leq 3 \sup_{x \in \mathcal{X}_n'} \frac{\left|\widehat{P}(x(V)) - P(x(V))\right|}{\sqrt{P(x(V/\{i\})) \vee \ln(\delta n) n^{-1}}} \leq 3 \sup_{x \in \mathcal{X}(G)} \frac{\left|\widehat{P}(x(V)) - P(x(V))\right|}{\sqrt{P(x(V/\{i\})) \vee c_2^2 \ln(\delta n) n^{-1}}}.
\end{aligned}$$

We apply Bousquet's version of Talagrand's inequality to the class of functions

$$\mathcal{F} = \left\{ f = \left(P(x(V/\{i\})) \vee c_2^2 \ln(\delta n) n^{-1} \right)^{-1/2} 1_{x(V)},\ x \in \mathcal{X}(G) \right\}$$

We have

$$v^2 = \sup_{f \in \mathcal{F}} \mathrm{Var}(f(X_1)) \leq 1,\ b = \sup_{f \in \mathcal{F}} \|f\|_\infty \leq \sqrt{c_2^{-2} (\ln(\delta n))^{-1} n}.$$

Hence, for all $\epsilon > 0$, with probability larger than $1 - \delta^{-1}$, we have

$$\begin{aligned}
&\sup_{x \in \mathcal{X}'_n} \sqrt{\widehat{P}(x(V/\{i\}))} |\widehat{P}_{i|V}(x) - P_{i|V}(x)| \\
&\quad \leq 3(1+\epsilon)\mathbb{E}\left(\sup_{f \in \mathcal{F}} |(P_n - P)f| \right) + 3\sqrt{\frac{2\ln(\delta)}{n}} + \left(1 + \frac{3}{\epsilon}\right) \frac{\ln(\delta)}{c_2 \sqrt{\ln(\delta n) n}} \\
&\quad \leq 3(1+\epsilon)\mathbb{E}\left(\sup_{f \in \mathcal{F}} |(P_n - P)f| \right) + \left(3\sqrt{2} + \frac{1}{c_2} + \frac{3}{c_2 \epsilon}\right) \sqrt{\frac{\ln(\delta)}{n}}.
\end{aligned}$$

We apply Lemma 9.6 to the sets $A_x = x(V)$ and the real numbers $\alpha_x = \left(P(x(V/\{i\})) \vee c_2^2 \ln(\delta n) n^{-1}\right)^{-1/2}$. We have

$$\alpha^* \leq \sqrt{n(c_2^2 \ln(\delta n))^{-1}} \; p^* = \sup_{x \in \mathcal{X}(G)} (P(x(V/\{i\})))^{-1} P(x(V)) \leq 1.$$

Hence,

$$\begin{aligned}
&\mathbb{E}\left(\sup_{f \in \mathcal{F}} |(P_n - P)f| \right) \leq \frac{64}{\sqrt{n}} \sqrt{\ln\left(4\sqrt{n(c_2^2 \ln(\delta n))^{-1}}\right)} \\
&\quad + 2048 \frac{\ln\left(4\sqrt{n(c_2^2 \ln(\delta n))^{-1}}\right)}{c_2 \sqrt{n(\ln(\delta n))^{-1}}} \leq \left(32\sqrt{2} + \frac{2048}{c_2}\right) \sqrt{\frac{\ln(n)}{n}}.
\end{aligned}$$

Thus, for all $\epsilon > 0$, with probability larger than $1 - \delta^{-1}$, we have

$$\begin{aligned}
&\sup_{x \in \mathcal{X}'_n} \sqrt{\widehat{P}(x(V/\{i\}))} |\widehat{P}_{i|V}(x) - P_{i|V}(x)| \\
&\quad \leq 2\left(\left(99\sqrt{2} + \frac{6144}{c_2}\right)(1+\epsilon) + \frac{1}{c_2}\left(1 + \frac{3}{\epsilon}\right) \right) \sqrt{\frac{\ln(\delta n)}{n}}.
\end{aligned}$$

We take $\epsilon = 0.001$ to conclude the proof.

7.2. Proof of Theorem 3.2:

It comes from Theorem 3.1 that, for all subsets V in $\mathcal{G}_n$, we have,

$$P\left(\left\| \widehat{P}_{i|V} - P_{i|V} \right\|_\infty \leq c_2 \sqrt{\frac{\ln(N_n \delta n)}{n \widehat{p}_-^V}} \right) \geq 1 - \frac{1}{N_n \delta}.$$

We use a union bound to get that,

$$P\left(\forall V \in \mathcal{G}_n, \; \left\| \widehat{P}_{i|V} - P_{i|V} \right\|_\infty \leq c_2 \sqrt{\frac{\ln(N_n \delta n)}{n \widehat{p}_-^V}} \right) \geq 1 - \delta^{-1}.$$

Hereafter in the proof of Theorem 3.2, we denote by $v_n^V = \sqrt{\ln(N_n \delta n)\left(n\widehat{p}_-^V\right)^{-1}}$ and by

$$\Omega = \left\{\forall V \in \mathcal{G}_n,\ \left\|\widehat{P}_{i|V} - P_{i|V}\right\|_\infty \leq c_2 v_n^V\right\}.$$

We have proved that $P(\Omega) \geq 1 - \delta^{-1}$. Let $C > c_2$ and denote, for short $\hat{G} = \hat{G}(C, \delta, \mathcal{G}_n)$. By definition of $\hat{G}$, for all $V \in \mathcal{G}_n$,

$$\left\|P_{i|G}\right\|_\infty - \left\|\widehat{P}_{i|\hat{G}}\right\|_\infty + C v_n^{\hat{G}} \leq \left\|P_{i|G}\right\|_\infty - \left\|\widehat{P}_{i|V}\right\|_\infty + C\mathrm{pen}(V).$$

Hence, on Ω, for all V in $\mathcal{G}_n$,

$$\left\|P_{i|G}\right\|_\infty - \left\|P_{i|\hat{G}}\right\|_\infty + (C - c_2) v_n^{\hat{G}} \leq \left\|P_{i|G}\right\|_\infty - \left\|P_{i|V}\right\|_\infty + (C + c_2)\mathrm{pen}(V). \tag{17}$$

From Assumption **H1**, $\left\|P_{i|G}\right\|_\infty - \left\|P_{i|\hat{G}}\right\|_\infty \geq \kappa_{\min}\left\|P_{i|G} - P_{i|\hat{G}}\right\|_\infty$ and from the triangular inequality, , $\left\|P_{i|G}\right\|_\infty - \left\|P_{i|V}\right\|_\infty \leq \left\|P_{i|G} - P_{i|V}\right\|_\infty$. Plugging these inequalities in (17), we obtain that, for all $V \in \mathcal{G}_n$,

$$\kappa_{\min}\left\|P_{i|G} - P_{i|\hat{G}}\right\|_\infty + (C - c_2) v_n^{\hat{G}} \leq \left\|P_{i|G} - P_{i|V}\right\|_\infty + (C + c_2)\mathrm{pen}(V). \tag{18}$$

On Ω, for all $V \in \mathcal{G}_n$, we have then

$$\begin{aligned}\left\|\widehat{P}_{i|\hat{G}} - P_{i|G}\right\|_\infty &\leq \left\|\widehat{P}_{i|\hat{G}} - P_{i|\hat{G}}\right\|_\infty + \left\|P_{i|\hat{G}} - P_{i|G}\right\|_\infty \leq c_2 v_n^{\hat{G}} + \left\|P_{i|G} - P_{i|\hat{G}}\right\|_\infty \\ &\leq \max\left(\frac{1}{\kappa_{\min}}, \frac{c_2}{C - c_2}\right)\left(\kappa_{\min}\left\|P_{i|G} - P_{i|\hat{G}}\right\|_\infty + (C - c_2) v_n^{\hat{G}}\right)\end{aligned}$$

$$\begin{aligned}\left\|\widehat{P}_{i|\hat{G}} - P_{i|G}\right\|_\infty &\leq \max\left(\frac{1}{\kappa_{\min}}, \frac{c_2}{C - c_2}\right)\left(\left\|P_{i|G} - P_{i|V}\right\|_\infty + (C + c_2)\mathrm{pen}(V)\right) \\ &\leq K(c_2, C, \kappa_{\min})\left(\left\|P_{i|G} - P_{i|V}\right\|_\infty + \mathrm{pen}(V)\right).\end{aligned}$$

7.3. Proof of Corollary 3.3:

It comes from [16] Proposition 2.5 p 20 that

$$N_{m,M} = |\mathcal{G}_{m,M}| = \sum_{k=0}^{m} C_M^k \leq \left(\frac{eM}{m}\right)^m \leq M^m \text{ hence } \ln(N_{m,M}) \leq m\ln(M).$$

Hence, from Theorem 3.2, with probability larger than $1 - \delta^{-1}$, we have

$$\left\|\widehat{P}_{i|\hat{G}_\delta(C))} - P_{i|G}\right\|_\infty \leq K \inf_{V \in \mathcal{G}_{m,M}}\left\{\left\|P_{i|G} - P_{i|V}\right\|_\infty + \sqrt{\frac{\ln(nM^m\delta)}{n\widehat{p}_-^V}}\right\}. \tag{19}$$

For all $|V| > \log_2(n)$, there is at least one configuration in $\mathcal{X}(V)$ that is not observed, hence $\widehat{p}_-^V = 1/n$. Therefore, for all $m \geq \log_2(n)$,

$$\inf_{V \in \mathcal{G}_{\log_2(n),M}} \left\{ \left\| P_{i|G} - P_{i|V} \right\|_\infty + \sqrt{\frac{\Gamma_M(\delta)}{n\widehat{p}_-^V}} \right\} = \inf_{V \in \mathcal{G}_{m,M}} \left\{ \left\| P_{i|G} - P_{i|V} \right\|_\infty + \sqrt{\frac{\Gamma_M(\delta)}{n\widehat{p}_-^V}} \right\}.$$

Taking $m = M$, (19) yields the corollary.

7.4. *Proof of Theorem 3.4:*

Let Ω be the event defined in the proof of Theorem 3.2 for the collection $\mathcal{G}_n$ and let $\hat{G} = \hat{G}_\delta(C)$. We have $P(\Omega^c) \leq \delta^{-1}$ and, on Ω, from Corollary 3.3,

$$\left\| P_{i|\hat{G}} - P_{i|G} \right\|_\infty \leq K \inf_{V \subset V_M} \left\{ \left\| P_{i|G} - P_{i|V} \right\|_\infty + v_n^V(\delta) \right\}.$$

By definition of Ψ, denoting by $l_n = \sqrt{n^{-1}\Gamma_M(\delta)}$, we have

$$\inf_{V \subset V_M} \left\{ \left\| P_{i|G} - P_{i|V} \right\|_\infty + v_n^V(\delta) \right\} = \inf_{v>0} \left\{ \Psi(v) + v l_n \right\}.$$

Let v^* be the smallest solution of the equation $v l_n = \Psi(v)$. As Ψ is non-increasing and $v \mapsto v l_n$ is non decreasing, we have

$$\Psi(v^*) \leq \inf_{v>0} \left\{ \Psi(v) + v l_n \right\} \leq 2\Psi(v^*).$$

Thus, on Ω, we have

$$\left\| P_{i|\hat{G}} - P_{i|G} \right\|_\infty \leq 2K\Psi(v^*).$$

Let Ω_2 be the event defined in **H2**. Let $r < 1$ and ω in $\Omega^* = \Omega \cap \Omega_2$ such that $\widehat{p}_-^{\hat{G}}(\omega) \geq (rv^*)^{-2}$. From assumption **H2** applied to $v = rv^*$, $K = r^{-1}$,

$$2K\Psi(v^*) \geq \left\| P_{i|\hat{G}}(\omega) - P_{i|G} \right\|_\infty \geq \Psi(rv^*) \geq C_\Psi^{-1} r^{-\alpha} \Psi(v^*).$$

Hence $r \geq (2C_\Psi K)^{-1/\alpha}$. Thus, on Ω^*, we have

$$\left\| P_{i|\hat{G}} - P_{i|G} \right\|_\infty \leq 2K l_n v^* \leq C_\Psi^{-1/\alpha} (2K)^{1-1/\alpha} v_n^{\hat{G}}.$$

By the triangular inequality, we have

$$\begin{aligned} \sup_{x \in \mathcal{X}(G)} & |(\widehat{P}_{i|V}(x) - \widehat{P}_{i|V}(x_j)) - (P_{i|G}(x) - P_{i|G}(x_j))| \\ & \leq 2 \left(\left\| \widehat{P}_{i|V} - P_{i|V} \right\|_\infty + \left\| P_{i|V} - P_{i|G} \right\|_\infty \right). \end{aligned}$$

Hence, on Ω^*,

$$\begin{aligned} \left| \omega_{i,j}^{\hat{G}}(\widehat{P}) - \omega_{i,j}^G(P) \right| & \leq 2 \left(c_2 v_n^{\hat{G}} + \left\| P_{i|\hat{G}} - P_{i|G} \right\|_\infty \right) \\ & \leq 2 \left(c_2 + C_\Psi^{-1/\alpha_\Psi} (2K)^{1-1/\alpha_\Psi} \right) v_n^{\hat{G}}. \end{aligned}$$

Let $c_\infty = 2\left(c_2 + C_\Psi^{-1/\alpha_\Psi}(2K)^{1-1/\alpha_\Psi}\right)$. It comes from this last inequality that, on Ω^*,

$$\left\{ j \in V_M,\ \omega_{i,j}^G(P) \geq (c + c_\infty) v_n^{\hat{G}} \right\} \subset \hat{G}_i^{\hat{G}}(c) \subset \left\{ j \in V_M,\ \omega_{i,j}^G(P) \geq (c - c_\infty) v_n^{\hat{G}} \right\}.$$

7.5. *Proof of Theorem 4.5:*

In all the proof, for all subsets V, V' of G such that $V \cap V' = \emptyset$, for all (x, y) in $\mathcal{X}(V) \times \mathcal{X}(V')$, let $x(V) \oplus y(V')$ be the configuration on $\mathcal{X}(V \cup V')$ such that, for all j in V $x(V) \oplus y(V')(j) = x(j)$ and for all j in V', $x(V) \oplus y(V')(j) = y(j)$. Let V be a finite subset of G and let x be a configuration on $\mathcal{X}(G)$.

$$P_{i|G}(x) - P_{i|V}(x) = \int (P_{i|G}(x) - P_{i|G}(x(V) \oplus y(G/V)))dP(y(G/V)|x(V/\{i\})) \tag{20}$$

From the definition of a Gibbs measure, we have

$$\begin{aligned} &P_{i|G}(x) - P_{i|G}(x(V) \oplus y(G/V)) \\ &= \frac{e^{-\sum_{j \in G} g_{i,j}(x(i),x(j))} \left(e^{\sum_{j \notin V} (g_{i,j}(x(i),x(j)) - g_{i,j}(x(i),y(j)))} - 1 \right)}{\left(1 + e^{-\sum_{j \in G} g_{i,j}(x(i),x(V) \oplus y(G/V)(j))}\right) \left(1 + e^{-\sum_{j \in G} g_{i,j}(x(i),x(j))}\right)} \end{aligned} \tag{21}$$

Hence,

$$\begin{aligned} &|P_{i|G}(x) - P_{i|G}(x(V) \oplus y(G/V))| \\ &\qquad \leq \frac{e^{2r}}{(1 + e^{-2r})^2} \left| e^{\sum_{j \notin V} (g_{i,j}(x(i),x(j)) - g_{i,j}(x(i),y(j)))} - 1 \right|. \end{aligned}$$

Let us now give the following lemma, whose proof is immediate from the convexity of $x \mapsto e^x$.

Lemma 7.4. *For all real numbers $r > 0$, for all x in $[-4r, 4r]$, we have*

$$\frac{1 - e^{-4r}}{4r}|x| \leq |e^x - 1| \leq \frac{e^{4r} - 1}{4r}|x|.$$

We deduce from Lemma 7.4 that

$$\begin{aligned} &|P_{i|G}(x) - P_{i|G}(x(V) \oplus y(G/V))| \\ &\qquad \leq \frac{(e^{4r} - 1)e^{2r}}{4r(1 + e^{-2r})^2} \left| \sum_{j \notin V} (g_{i,j}(x(i), x(j)) - g_{i,j}(x(i), y(j))) \right|. \end{aligned}$$

It is clear that, for all x in $\mathcal{X}(G)$,

$$\int \left| \sum_{j \notin V} (g_{i,j}(x(i), x(j)) - g_{i,j}(x(i), y(j))) \right| dP(y(G/V)|x(V/\{i\})) \leq \sum_{j \notin V} \omega_{i,j}(f) P(y(j) \neq x(j)|x(V/\{i\})).$$

The upper bound comes then from the inequality $P(y(j) \neq x(j)|x(V/\{i\})) \leq 1$.

For the lower bound, let, for all a in A, $x^a_{\max}$ be the configuration such that, for all j in G, $g_{i,j}(a, x^a_{\max}(j)) = \left\| g^a_{i,j} \right\|$ and let $x^a_{\min}$ be the configuration such that, for all j in G, $g_{i,j}(a, x^a_{\min}(j)) = \inf_{b \in A} g_{i,j}(a, b)$. From Lemma 7.4, we have

$$e^{\sum_{j \notin V} \left\| g^{x(i)}_{i,j} \right\| - g_{i,j}(x(i), y(j))} - 1 = \left| e^{\sum_{j \notin V} \left\| g^{x(i)}_{i,j} \right\| - g_{i,j}(x(i), y(j))} - 1 \right| \geq \frac{1 - e^{-4r}}{4r} \sum_{j \notin V} \left\| g^{x(i)}_{i,j} \right\| - g_{i,j}(x(i), y(j)). \quad (22)$$

Finally, we have

$$\int \left\| g^{x(i)}_{i,j} \right\| - g_{i,j}(x(i), y(j)) dP(y(G/V)|x(V/\{i\})) \geq P(y(j) = x^{x(i)}_{\min}(j)|x(V/\{i\})) \omega_{i,j}(f) \geq \frac{\omega_{i,j}(f)}{1 + e^{2r}}. \quad (23)$$

Using successively inequalities (20), (21), (22) and (23) with $x = x^{x(i)}_{\max}$, we obtain

$$\begin{aligned}
&\sup_{x \in \mathcal{X}(G)} P_{i|G}(x) - P_{i|V}(x) \\
&\geq \int (P_{i|G}(x^{x(i)}_{\max}) - P_{i|G}(x^{x(i)}_{\max}(V) \oplus y(G/V))) dP(y(G/V)|x(V/\{i\})) \\
&= \frac{e^{-2r}}{1 + e^{-2r}} \int \frac{e^{\sum_{j \notin V} \left\| g^{x(i)}_{i,j} \right\| - g_{i,j}(x(i), y(j))} - 1}{1 + e^{-\sum_{j \in V} \left\| g^{x(i)}_{i,j} \right\| - \sum_{j \notin V} g_{i,j}(x(i), y(j))}} dP(y(G/V)|x(V/\{i\})) \\
&\geq \frac{e^{-2r}}{(1 + e^{2r})^2} \int \left(e^{\sum_{j \notin V} \left\| g^{x(i)}_{i,j} \right\| - g_{i,j}(x(i), y(j))} - 1 \right) dP(y(G/V)|x(V/\{i\}))
\end{aligned}$$

Hence,

$$\begin{aligned}
&\sup_{x \in \mathcal{X}(G)} P_{i|G}(x) - P_{i|V}(x) \\
&\geq \frac{(1 - e^{-4r}) e^{-2r}}{4r(1 + e^{2r})^2} \int \left(\sum_{j \notin V} \left\| g^{x(i)}_{i,j} \right\| - g_{i,j}(x(i), y(j)) \right) dP(y(G/V)|x(V/\{i\})) \\
&\geq \frac{(1 - e^{-4r}) e^{-2r}}{4r(1 + e^{2r})^3} \sum_{j \notin V} \omega_{i,j}(f).
\end{aligned}$$

Let us now check that P satisfies assumption **H1**. Let x in $\mathcal{X}(V)$. Using successively inequalities (20), (21), (22) and (23) with $x = x(V) \oplus x^{x(i)}_{\max}(G/V)$, we obtain, as in the previous proof,

$$P_{i|G}(x(V) \oplus x^{x(i)}_{\max}(G/V)) - P_{i|V}(x) \geq \frac{(1-e^{-4r})e^{-2r}}{4r(1+e^{2r})^3} \sum_{j \notin V} \omega_{i,j}(f).$$

Taking x such that $P_{i|V}(x) = \left\|P_{i|V}\right\|_\infty$ and using that $P_{i|G}(x(V) \oplus x^{x(i)}_{\max}(G/V)) \leq \left\|P_{i|G}\right\|_\infty$, we obtain

$$\left\|P_{i|G}\right\|_\infty - \left\|P_{i|V}\right\|_\infty \geq \frac{(1-e^{-4r})e^{-2r}}{4r(1+e^{2r})^3} \sum_{j \notin V} \omega_{i,j}(f).$$

This yields the theorem thanks to inequality (6).

8. Proof of Theorem 4.6:

$$\begin{aligned}
&|\widehat{p}(i,j) - \widehat{p}(i)\widehat{p}(j) - (P(x_1(i,j)) - P(x_1(i))P(x_1(j)))| \\
&\leq \left|\widehat{P}(x_1(i,j)) - P(x_1(i,j))\right| + \left|\widehat{P}(x_1(j)) - P(x_1(j))\right| + \left|\widehat{P}(x_1(i)) - P(x_1(i))\right|
\end{aligned}$$

We use Hoeffding's inequality (see for example [16] Proposition 2.7) to the functions $t = 1_{x_1(i,j)},\ 1_{x_1(i)},\ 1_{x_1(j)}$, for all $x > 0$, we have

$$P\left(|(P_n - P)t| > \sqrt{\frac{x}{2n}}\right) \leq 2e^{-x}.$$

Hence, a union bound gives that, on a set $\Omega(\delta)$ satisfying $P(\Omega(\delta)^c) \leq \delta^{-1}$, for all j in V_M

$$\begin{aligned}
&\left|\widehat{P}(x_1(i,j)) - P(x_1(i,j))\right| + \left|\widehat{P}(x_1(j)) - P(x_1(j))\right| \\
&\qquad + \left|\widehat{P}(x_1(i)) - P(x_1(i))\right| \leq 3\sqrt{\frac{\ln(6M\delta)}{2n}}.
\end{aligned}$$

Moreover, we have

$$\begin{aligned}
&P_{i|j}(x_1) - P(x_1(i)) = \left(P_{i|j}(x_1) - P_{i|\emptyset}(x_1)\right) = \qquad (24)\\
&\int P_{i|G}(x_1(i,j) \oplus y(G/\{i,j\})) - P_{i|G}(x_1(i) \oplus y(G/\{i\}))dP(y(G/\{i\})|x_1(i)).
\end{aligned}$$

From the definition of a Gibbs measure, we have

$$\begin{aligned}
&P_{i|G}(x_1(i,j) \oplus y(G/\{i,j\})) - P_{i|G}(x_1(i) \oplus y(G/\{i\})) \qquad (25)\\
&= \frac{e^{-\sum_{j \in G} g_{i,j}(x_1(i),y(j))}\left(e^{(g_{i,j}(x_1(i),x_1(j)) - g_{i,j}(x_1(i),y(j)))} - 1\right)}{\left(1 + e^{-\sum_{j \in G} g_{i,j}(x_1(i),x_1(i,j) \oplus y(G/\{i,j\}))}\right)\left(1 + e^{-\sum_{j \in G} g_{i,j}(x_1(i),y(j))}\right)}.
\end{aligned}$$

We can assume that, without loss of generality that $g_{i,j}(x_1(i), x_1(j)) = \|g_{i,j}\|$ and therefore that

$$e^{g_{i,j}(x_1(i),x_1(j))-g_{i,j}(x_1(i),y(j))} - 1 = \left| e^{g_{i,j}(x_1(i),x_1(j))-g_{i,j}(x_1(i),y(j))} - 1 \right|.$$

It comes then from Lemma 7.4 that

$$\begin{aligned}
&\frac{1-e^{-4r}}{4r} \left| g_{i,j}(x_1(i), x_1(j)) - g_{i,j}(x_1(i), y(j)) \right| \\
&\qquad \leq e^{g_{i,j}(x_1(i),x_1(j))-g_{i,j}(x_1(i),y(j))} - 1 \\
&\qquad \leq \frac{e^{4r}-1}{4r} \left| g_{i,j}(x_1(i), x_1(j)) - g_{i,j}(x_1(i), y(j)) \right|
\end{aligned}$$

Hence

$$\begin{aligned}
&\frac{1-e^{-4r}}{4r} \left| \omega_{i,j}(f) \right| 1_{y(j)\neq x_1(j)} \\
&\qquad \leq e^{g_{i,j}(x_1(i),x_1(j))-g_{i,j}(x_1(i),y(j))} - 1 \leq \frac{e^{4r}-1}{4r} \left| \omega_{i,j}(f) \right| \qquad (26)
\end{aligned}$$

Using successively (24), (25), (26), we deduce that

$$\begin{aligned}
\frac{e^{-2r}(1-e^{-4r})}{4r(1+e^{2r})^3} \left| \omega_{i,j}(f) \right| &\leq \frac{e^{-2r}(1-e^{-4r})}{4r(1+e^{2r})^2} P(y(j) \neq x_1(j) | x_1(i)) \left| \omega_{i,j}(f) \right| \\
&\leq \left| P(x_1(i,j)) - P(x_1(i))P(x_1(j)) \right| \leq \frac{e^{2r}(e^{4r}-1)}{4r(1+e^{-2r})^2} \left| \omega_{i,j}(f) \right|.
\end{aligned}$$

We conclude that, on $\Omega(\delta)$,

$$\begin{aligned}
&\frac{e^{-2r}(1-e^{-4r})}{4r(1+e^{2r})^3} \left| \omega_{i,j}(f) \right| - 3\sqrt{\frac{\ln(6M\delta)}{2n}} \\
&\leq \left| \widehat{P}(x_1(i,j)) - \widehat{P}(x_1(i))\widehat{P}(x_1(j)) \right| \leq \frac{e^{2r}(e^{4r}-1)}{4r(1+e^{-2r})^2} \left| \omega_{i,j}(f) \right| + 3\sqrt{\frac{\ln(6M\delta)}{2n}}
\end{aligned}$$

All the sites j such that

$$\left| \omega_{i,j}(f) \right| \geq \frac{4r(1+e^{2r})^3}{e^{-2r}(1-e^{-4r})} \left(\eta + 3\sqrt{\frac{\ln(6M\delta)}{2n}} \right) \qquad (27)$$

belong to $\widehat{V}(\eta)$. All the sites such that

$$\left| \omega_{i,j}(f) \right| < \frac{4r(1+e^{-2r})^2}{e^{2r}(e^{4r}-1)} \left(\eta - 3\sqrt{\frac{\ln(6M\delta)}{2n}} \right) \qquad (28)$$

do not belong to $\widehat{V}(\eta)$.
We use then Theorem 3.2 with the collection $\mathcal{G} = \{ V \subset V(\eta_{ms}) \}$. Its cardinality

is bounded by n^κ. There exist a constant K and an event $\Omega_2(\delta)$, with probability $1-\delta$, such that, on $\Omega_2(\delta)$,

$$\left\|\widehat{P}_{i|\hat{G}} - P_{i|G}\right\| \leq K \inf_{V\in\mathcal{G}} \left\{ \left\|P_{i|V} - P_{i|G}\right\| + \sqrt{\frac{\ln(n^\kappa\delta)}{n\widehat{p}_-^V}} \right\}.$$

We use Theorem 4.5 to say that

$$\left\|P_{i|V} - P_{i|G}\right\| \leq C_r \sum_{j\notin V} |\omega_{i,j}(f)| = C_r \left(\sum_{j\notin \widehat{V}(\eta)} |\omega_{i,j}(f)| + \sum_{j\in \widehat{V}(\eta)/V} |\omega_{i,j}(f)| \right).$$

We deduce from (27) that, on $\Omega(\delta)$,

$$\sum_{j\notin \widehat{V}(\eta)} |\omega_{i,j}(f)| \leq \sum_{j\in V(\eta,\delta,M)} |\omega_{i,j}(f)|.$$

We choose $\Omega_*(\delta) = \Omega(\delta)\cap\Omega_2(\delta)$ to conclude the proof.

9. Appendix

In this Appendix, we recall the bound given by Bousquet [6] for the deviation of the supremum of the empirical process.

Theorem 9.1. *Let $X_1, ..., X_n$ be i.i.d. random variables valued in a measurable space $(A, \mathcal{X})$. Let $\mathcal{F}$ be a class of real valued functions, defined on A and bounded by b. Let $v^2 = \sup_{f\in\mathcal{F}} P[(f-Pf)^2]$ and $Z = \sup_{f\in\mathcal{F}}(P_n - P)f$. Then, for all $x > 0$,*

$$P\left(Z > \mathbb{E}(Z) + \sqrt{\frac{2}{n}(v^2 + 2b\mathbb{E}(Z))x} + \frac{bx}{3n} \right) \leq e^{-x}. \tag{29}$$

Let us recall some well known tools of empirical processes theory.

Definition 9.2. *The covering number $N(\epsilon, T, d)$ is the minimal number of balls of radius ϵ with centers in T needed to cover T. The entropy is the logarithm of the covering number $H(\epsilon, T, d) = \ln(N(\epsilon, T, d))$.*

Definition 9.3. *An ϵ-separated subset of T is a subset $\{t_k\}$ of elements of T whose pairwise distance is strictly larger than ϵ. The packing number $M(\epsilon, T, d)$ is the maximum size of an ϵ-separated subset of T.*

Those quantities are related by the famous following lemma.

Lemma 9.4. *(Kolmogorov and Tikhomirov [13]) Let (T, d) be a metric space and let $\epsilon > 0$,*

$$N(\epsilon, T, d) \leq M(\epsilon, T, d) \leq N(\epsilon/2, T, d).$$

The following result can be derived from classical chaining arguments (see for example [6]).

Lemma 9.5. *Let $\mathcal{F}$ be a class of functions, let $d_{2,P_n}(t,t') = \sqrt{P_n[(t-t')^2]}$ and $D_n = \sqrt{\sup_{t\in\mathcal{F}} P_n(t^2)}$ then*

$$\mathbb{E}\left(\sup_{t\in\mathcal{F}} |(P_n - P)t|\right) \leq \frac{16\sqrt{2}}{\sqrt{n}}\mathbb{E}\left(\int_0^{D_n/2} H^{1/2}(u,\mathcal{F},d_{2,P_n})du\right).$$

The next result was used to obtain our concentration inequalities.

Lemma 9.6. *Let $(A_i)_{i\in I}$ be a collection of sets such that, for all $i,j \in I$, $A_i \cap A_j = \emptyset$ and let $(\alpha_i)_{i\in I}$ be a collection of positive real numbers. Let $Z_I = \sup_{t\in\mathcal{F}_I} |(P_n - P)t|$, where $\mathcal{F}_I = \{t_i = \alpha_i 1_{A_i}\}$ and P_n is the empirical measure. Let $\alpha^* = \sup_{i\in I}\alpha_i$, $p_* = \sup_{i\in I}\alpha_i^2 P(A_i)$. We have*

$$\mathbb{E}\left(\sup_{t\in\mathcal{F}_I} |(P_n - P)t|\right) \leq \frac{64}{\sqrt{n}}\sqrt{p^* \ln\left(\frac{4\alpha^*}{\sqrt{p^*}}\right)} + \frac{2048}{n}\alpha^* \ln\left(\frac{4\alpha^*}{\sqrt{p^*}}\right). \tag{30}$$

In order to apply Lemma 9.5 to $\mathcal{F} = \mathcal{F}_I$, we compute the entropy of $\mathcal{F}_I$. For all $i \neq j$, since $A_i \cap A_j = \emptyset$,

$$(t_i - t_j)^2 = \left(\alpha_i 1_{A_i} - \alpha_j 1_{A_j}\right)^2 = \alpha_i^2 1_{A_i} + \alpha_j^2 1_{A_j}.$$

Hence $d_{2,P_n}(t_i,t_j) = \sqrt{\alpha_i^2 P_n(A_i) + \alpha_j^2 P_n(A_j)}$.
Consider an ϵ-separated set $T_\epsilon = \{t_{i_1},...,t_{i_N}\}$ in $(\mathcal{F}_I, d_{2,P_n})$ (see also the definition in the appendix), it comes from the previous computation that, for all $k \neq k'$,

$$\alpha_{i_k}^2 P_n(A_{i_k}) + \alpha_{i_{k'}}^2 P_n(A_{i_{k'}}) \geq \epsilon^2.$$

Hence, there is at least $N-1$ indexes $k \in \{1,...,N\}$ such that $\alpha_{i_k}^2 P_n(A_{i_k}) \geq \epsilon^2/2$. It follows that

$$1 = \sum_{i\in I} P_n(A_i) \geq \sum_{k=1}^{N} P_n(A_{i_k}) \geq \frac{\epsilon^2(N-1)}{2(\alpha^*)^2}.$$

Hence $N \leq 1 + 2(\alpha^*)^2\epsilon^{-2}$, thus $H(\epsilon,\mathcal{F}_I,d_{2,P_n}) \leq \ln\left(1 + 2(\alpha^*)^2\epsilon^{-2}\right)$. We deduce from this inequality and Lemma 9.5 that

$$\begin{aligned}\mathbb{E}\left(\sup_{t\in\mathcal{F}_I} |(P_n - P)t|\right) &\leq \frac{16\sqrt{2}}{\sqrt{n}}\mathbb{E}\left(\int_0^{\sqrt{\hat{p}_n^*}/2} \sqrt{\ln\left(1 + 2(\alpha^*)^2\epsilon^{-2}\right)}d\epsilon\right) \\ &\leq \frac{32}{\sqrt{n}}\mathbb{E}\left(\int_0^{\sqrt{\hat{p}_n^*}/2} \sqrt{\ln\left(2\alpha^*\epsilon^{-1}\right)}d\epsilon\right), \end{aligned} \tag{31}$$

where $\hat{p}_n^* = \sup_{i\in I}\alpha_i^2 P_n(A_i)$. Now, let us recall the following elementary lemma.

Lemma 9.7. *For all positive real numbers K, A such that $K/A > e$, we have*

$$\int_0^A \sqrt{\ln(Kx^{-1})}dx \leq 2A\sqrt{\ln\left(\frac{K}{A}\right)}$$

Actually,

$$\int_0^A \sqrt{\ln(Kx^{-1})}dx = K\int_{K/A}^{\infty} \frac{\sqrt{\ln(x)}}{x^2}dx = A\sqrt{\ln\left(\frac{K}{A}\right)} + \frac{K}{2}\int_{K/A}^{\infty} \frac{1}{u^2\sqrt{\ln u}}du.$$

Since $K/A > e$, $\frac{1}{u^2\sqrt{\ln u}} \leq \frac{\sqrt{\ln u}}{u^2}$ on $[K/A, \infty[$. The result follows.
By definition, $\hat{p}_n \leq (\alpha^*)^2$, hence $2\alpha^*/(\sqrt{\hat{p}_n^*}/2) \geq 4 > e$, we deduce from Lemma 9.7 that

$$\mathbb{E}\left(\sup_{t\in\mathcal{F}_I} |(P_n - P)t|\right) \leq \frac{32}{\sqrt{n}}\mathbb{E}\left[\sqrt{\hat{p}_n^*}\sqrt{\ln\left(\frac{4\alpha^*}{\sqrt{\hat{p}_n^*}}\right)}\right].$$

Let us now give another simple lemma.

Lemma 9.8. *The function $f : x \mapsto x\sqrt{\ln(K/x)}$, defined on $(0, K)$ is positive, non decreasing on $(0, K/e^{1/2})$ and strictly concave.*

The proof of the lemma is straightforward from the computations

$$f'(x) = \sqrt{\ln(K/x)} - \frac{1}{2\sqrt{\ln(K/x)}},\ f"(x) = -\frac{1}{2x\sqrt{\ln(K/x)}} - \frac{1}{4x(\sqrt{\ln(K/x)})^3}.$$

Applying Lemmas 9.8, 9.7, and Jensen's inequality to the right hand side of (31) we have that

$$\mathbb{E}\left(\sup_{t\in\mathcal{F}_I} |(P_n - P)t|\right) \leq \frac{32}{\sqrt{n}}\mathbb{E}\left(\sqrt{\hat{p}_n^*}\right)\sqrt{\ln\left(\frac{4\alpha^*}{\mathbb{E}\left(\sqrt{\hat{p}_n^*}\right)}\right)}.$$

Now it comes from Jensen inequality that

$$\mathbb{E}\left[\sqrt{\hat{p}_n^*}\right] \leq \sqrt{\mathbb{E}\left[\hat{p}_n^*\right]} \leq \sqrt{p^*} + \sqrt{\alpha^*\mathbb{E}\left(\sup_{t\in\mathcal{F}_I} |(P_n - P)t|\right)}.$$

It is clear from its definition that $p^* \leq (\alpha^*)^2$. Moreover, as P_n and P are probability measures, we have, for all t in $\mathcal{F}_I$, $|(P_n - P)t| \leq 2\alpha^*$. Hence, $\sqrt{\alpha^*\mathbb{E}\left(\sup_{t\in\mathcal{F}_I} |(P_n - P)t|\right)} \leq \sqrt{2}\alpha^*$. We deduce from these inequalities that

$$\sqrt{p^*} + \sqrt{\alpha^*\mathbb{E}\left(\sup_{t\in\mathcal{F}_I} |(P_n - P)t|\right)} \leq (1 + \sqrt{2})\alpha^* \leq (4\alpha^*)/e^{1/2}.$$

Hence, it comes from Lemma 9.8 that, if $E = \mathbb{E}\left(\sup_{t\in\mathcal{F}_I} |(P_n - P)t|\right)$

$$E \leq \frac{32}{\sqrt{n}}\left(\sqrt{p^*} + \sqrt{\alpha^* E}\right)\sqrt{\ln\left(\frac{4\alpha^*}{\sqrt{p^*} + \sqrt{\alpha^* E}}\right)}$$
$$\leq \frac{32}{\sqrt{n}}\left(\sqrt{p^*} + \sqrt{\alpha^* E}\right)\sqrt{\ln\left(\frac{4\alpha^*}{\sqrt{p^*}}\right)}.$$

It is then straightforward that (30) holds.

References

[1] Arlot, S. and Massart, P. (2009). Data-driven calibration of penalties for least-squares regression. *Journal of Machine learning research* **10** 245–279.

[2] Baudry, J.-P., Maugis, K. and Michel, B. (2010). Slope heuristics: overview and implementation. *INRIA report, available at http://hal.archives-ouvertes.fr/hal-00461639/fr/.*

[3] Bento, J. and Montanari, A. (2009). Which graphical models are difficult to learn? *available on Arxive http://arxiv.org/pdf/0910.5761.*

[4] Besag, L. (1993). Statistical analysis of dirty pictures. *Journal of applied statistics* **20** 63–87.

[5] Birgé, L. and Massart, P. (2007). Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields* **138** 33–73. MR2288064 (2008g:62070)

[6] Bousquet, O. (2002). A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Math. Acad. Sci. Paris* **334** 495–500. MR1890640 (2003f:60039)

[7] Bresler, G., Mossel, E. and Sly, A. (2008). *Approximation, Randomization and Combinatorial Optimization. Algorithms and Techniques* Reconstruction of Markov Random Fields from Samples: Some Easy Observations and Algorithms 343-356. Springer.

[8] Brown, E. N., Kass, R. E. and Mitra, P. P. (2004). Multiple neural spike train data analysis: state-of-the-art and future challenges. *Nature Neuroscience* **7** 456-461.

[9] Cross, G. and Jain, A. (1983). Markov Random field texture models. *IEEE Trans. PAMI* **5** 25–39.

[10] Csiszar, I. and Talata, Z. (2006). Consistent estimation of the basic neighborhood of Markov random fields. *Annals of Statistics* **34** 123-145.

[11] Galves, A., Orlandi, E. and Takahashi, D. Y. (2010). Identifying interacting pairs of sites in infinite range Ising models. *Preprint, http://arxiv.org/abs/1006.0272.*

[12] Georgii, H. (1988). *Gibbs measure and phase transitions. de Gruyter studies in mathematics* **9**. de Gruyter, Berlin.

[13] Kolmogorov, A. and Tikhomirov, V. (1963). -entropy and ϵ-capacity of sets in functional spacesϵ. *Amer.Math. Soc. Trans.* **1** 277–364.

[14] Lerasle, M. (2009). Optimal model selection in density estimation. *available on Arxive http://arxiv.org/abs/0910.1654*.

[15] Li, X., Ouyang, G., Usami, A., Ikegaya, Y. and Sik, A. (2010). Scale-free topology of the CA3 hippocampal network: a novel method to analyze functional neuronal assemblies. *Biophysics Journal* **98** 1733-1741.

[16] Massart, P. (2007). *Concentration inequalities and model selection. Lecture Notes in Mathematics* **1896**. Springer, Berlin. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard. MR2319879

[17] Ravikumar, P., Wainwright, M. J. and Lafferty, J. D. (2010). High-Dimensional Ising Model Selection Using l_1-regularized Logistic Regression. *Ann. Statist.* **38** 1287–1319.

[18] Ripley, B. D. (1981). *Spatial Statistics*. Wiley, New York.

[19] Schneidman, E., Berry, M. J., Segev, R. and Bialek, W. (2006). Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature* **440** 1007–1012.

[20] Takahashi, N., Sasaki, T., Matsumoto, W. and Ikegaya, Y. (2010). Circuit topology for synchronizing neurons in spontaneously active networks. *Proceedings of National Academy of Science U.S.A.* **107** 10244-10249.

[21] Woods, J. (1978). Markov Image Modeling. *IEEE Trans. Automat. Control* **23** 846–850.

[22] Zhou, S. (2010). Thresholded Lasso for high dimensional variable selection and statistical estimation. *arXiv:1002.158v2*.