Bernoulli **22**(1), 2016, 325–344
DOI: 10.3150/14-BEJ660

Sharp oracle inequalities and slope heuristic for specification probabilities estimation in discrete random fields

MATTHIEU LERASLE[1] and DANIEL Y. TAKAHASHI[2]

[1]*CNRS, LJAD, UMR 7351, Univ. Nice Sophia Antipolis, 06100 Nice, France. E-mail: mlerasle@unice.fr*
[2]*Department of Psychology, Neuroscience Institute, Princeton University, Princeton, NJ 08648, USA. E-mail: takahashiyd@gmail.com*

We study the problem of estimating the one-point specification probabilities in non-necessary finite discrete random fields from partially observed independent samples. Our procedures are based on model selection by minimization of a penalized empirical criterion. The selected estimators satisfy sharp oracle inequalities in L_2-risk.

We also obtain theoretical results on the slope heuristic for this problem, justifying the slope algorithm to calibrate the leading constant in the penalty. The practical performances of our methods are investigated in two simulation studies. We illustrate the usefulness of our approach by applying the methods to a multi-unit neuronal data from a rat hippocampus.

Keywords: model selection; penalization; slope heuristic; discrete random fields

1. Introduction

The main motivation for our work comes from neuroscience where the advancement of multi-channel and optical technology enables researchers to record signals from tens to thousands of neurons simultaneously [25]. The question is then to understand the interactions between neurons in the brain and their relationships with the animal behavior [11,24].

Following [24], we model interactions between neurons by discrete random fields. A discrete random field is a triplet (S, A, P) where S is a discrete set of *sites*, possibly infinite, A is a finite alphabet, and P is a probability measure on the set $\mathcal{X}(S) = A^S$ of *configurations* on S. Given a random field (S, A, P), we define the *one point specification probabilities* of P as regular versions of the following conditional probabilities,

$$\forall i \in S, \forall x \in \mathcal{X}(S), \qquad P_{i|S}(x) = P\big(x(i)|x(j), j \in S/\{i\}\big).$$

The specification probabilities are important in the applications as they encode the conditional independence between the sites, see, for example, [5,10,12,14,18,22]. The main goal of this paper is to provide good estimators of the specification probabilities, assuming that the configurations are only observed on a finite subset $V_M \subset S$. Consider i.i.d. random variables $X_{1:n} = X_1, \ldots, X_n$ with common distribution P, the data set is given by $(X_i(j))_{i=1,\ldots,n;\, j\in V_M}$. Following [3,6,7], we use a penalized criterion to select a subset $\widehat{V} \subset V_M$ with cardinality $\mathrm{O}(\log n)$ and show that

1350-7265 © 2016 ISI/BS

the empirical conditional probabilities $\widehat{P}_{i|\widehat{V}}$ satisfy a sharp oracle inequality (see Section 2 and Theorems 3.2 for details).

In most of the applications, the support $V_\star$ of $P_{i|S}$ (i.e., the minimal set $V_\star \subset S$ such that $P_{i|V_\star} = P_{i|S}$) is the object of interest and the literature focus on the estimation of $V_\star$, see [5,10,12,14,22] for example. This approach requires in general strong assumptions on the random field, for example, it is assumed that the data is generated by an Ising model with restrictive conditions on the temperature parameter [5,14,22]. In particular, [5,10,22] assumed that the set S is finite and that all the sites are observed, that is, that $V_M = S$. When V_M does not contain $V_\star$, the meaning of the estimators in these papers is not clear. [12] considered $S = \mathbb{Z}^d$ but assumed that $V_\star$ is finite. Finally, [14,18] worked with infinite sets of sites and without prior bounds on the number of interacting sites but required a two-letters alphabet A and some assumptions on P that the practitioner cannot easily verify. These restrictions are severe in practice, for example, in neuroscience, and cast doubt on the theoretical support for application of these methods. Our approach does not suffer from these drawbacks. In particular, the alphabet size $|A|$ can be larger than 2, P does not need to be an Ising or Potts model, and some configurations on V_M can be forbidden. Furthermore, $V_\star$ can be infinite and therefore not contained in V_M.

The second result of the paper is a proof of the slope heuristic for the estimation of one-point specification probabilities in discrete random fields. The slope heuristic was introduced in [8] for Gaussian model selection and has been theoretically studied only for very few specific models [1,2,8,16,17,23]. Our proof technique is novel and sheds new lights on this phenomenon.

The paper is organized as follows. Section 2 presents the framework and some notations used all along the paper. Section 3 introduces our estimators and the oracle inequalities that they satisfy. In Section 4, the bias for Gibbs models is computed and Section 5 is devoted to the slope heuristic. Section 6 illustrates the results of previous sections using two simulation experiments and in Section 7 our methods are applied on a neurophysiology data set. The proofs of the main theorems are postponed to the Appendix C. The methods of this article can be adapted to the Küllback loss; the interested reader can find these developments in Section C of the Appendix (Supplementary Materialy, [19]).

2. Setting

Let (S, A, P) be a discrete random field, that is, a triplet where S is a discrete set, A is a finite set, with cardinality $|A|$ and P is a probability measure on $\mathcal{X}(S) = A^S$. Let V_M be a finite subset of S with cardinality $M \geq 3$ and let $i \in S$ denote a fixed site so that we will often omit the dependence on i of some quantities when there is no confusion. For any $x \in \mathcal{X}(S)$ and any $V \subset V_M$, let $\mathcal{X}(V) = A^V$, $v = |V|$, $x(V) = (x(j))_{j\in V}$. Let $X_1, \ldots, X_n$ be i.i.d. random variables with distribution P. The empirical probability measure $\widehat{P}$ is defined for any $x \in \mathcal{X}(S)$ by $\widehat{P}(x) = \frac{1}{n}\sum_{k=1}^n \mathbf{1}_{\{X_k=x\}}$, where $\mathbf{1}_{\{X_k=x\}} = 1$ if $X_k = x$ and 0 otherwise. The measures P and $\widehat{P}$ define probability measures on $\mathcal{X}(V)$ by the formulas $P(x(V)) = \int_{y\in\mathcal{X}(S);y(V)=x(V)} \mathrm{d}P(y(S))$, $\widehat{P}(x(V)) = \sum_{y\in\mathcal{X}(S);y(V)=x(V)} \widehat{P}(y)$. Hereafter, Q always denotes either P or $\widehat{P}$. For any $V \subset V_M$, $x \in \mathcal{X}(S)$, let $Q_{i|V}(x) = \frac{Q(V\cup\{i\})}{Q(V\setminus\{i\})}$ if $Q(V \setminus \{i\}) \neq 0$, $|A|^{-1}$ otherwise. Let also

$$P_{i|S}(x) = P\big(x(i)|x\big(S \setminus \{i\}\big)\big)$$

be a regular version of the conditional distribution of P. For any function $f : \mathcal{X}(S) \to \mathbb{R}$, let

$$\|f\|_Q = \sqrt{\int f^2(x) \frac{\mathrm{d}Q(x(S/\{i\}))}{|A|}}.$$

The observation set is $X_{1:n}(V_M) = (X_1(j), \ldots, X_n(j))_{j \in V_M}$. Algebraic computations show

$$\forall y \in \mathcal{X}(V_M), \qquad \widehat{P}(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{X_i(V_M)=y\}},$$

and for any $V \subset V_M$, $\widehat{P}(x(V)) = \sum_{y \in \mathcal{X}(V_M); y(V)=x(V)} \widehat{P}(x(V))$ can be computed from the data set. Hence, for $V \subset V_M$ the empirical probability $\widehat{P}_{i|V}$ is an estimator of $P_{i|S}$. The $L_{2,P}$*-risk* of $\widehat{P}_{i|V}$ is defined by $\|\widehat{P}_{i|V} - P_{i|S}\|_P^2$. We can decompose the risk via Pythogoras relation (see Proposition B.11)

$$\|\widehat{P}_{i|V} - P_{i|S}\|_P^2 = \|\widehat{P}_{i|V} - P_{i|V}\|_P^2 + \|P_{i|V} - P_{i|S}\|_P^2.$$

The random term $\|\widehat{P}_{i|V} - P_{i|V}\|_P^2$ is called the *variance* and the deterministic term $\|P_{i|V} - P_{i|S}\|_P^2$ is called the *bias*. Let $s \geq 3$ be an integer and let

$$\mathcal{V}_s = \{V \subset V_M, v \leq s\}, \qquad N_s = \mathrm{Card}(\mathcal{V}_s).$$

An *oracle* is a set $V_o \in \mathcal{V}_s$ that minimizes the risk, *that is*,

$$\|\widehat{P}_{i|V_o} - P_{i|S}\|_P^2 = \min_{V \in \mathcal{V}_s} \|\widehat{P}_{i|V} - P_{i|S}\|_P^2$$

and the minimal risk is called *oracle risk*. We will show in the next section that we can obtain an estimator $\widehat{V}$ such that the risk of $\widehat{P}_{i|\widehat{V}}$ is close to the oracle risk.

3. Model selection results

Let start with a concentration inequality for the variance term of the risks.

Theorem 3.1. *Let $Q \in \{P, \widehat{P}\}$ and let $V \in \mathcal{V}_s$. Then, for all $\delta > 1$ and all $0 < \eta \leq 1$,*

$$\mathbb{P}\left(\|\widehat{P}_{i|V} - P_{i|V}\|_Q^2 > \frac{6}{|A|} \left((1 + 8\eta) \frac{|A|^v}{n} + \frac{4 \log(2\delta)}{\eta n} + \frac{9 \log(2\delta)^2}{\eta^4 n} \right) \right) \leq \frac{1}{\delta}. \tag{3.1}$$

Comment. The bound can be integrated to give the following control

$$\mathbb{E}\big[\|\widehat{P}_{i|V} - P_{i|S}\|_P^2\big] = \|P_{i|V} - P_{i|S}\|_P^2 + C \frac{|A|^{v-1}}{n},$$

for some absolute constant C. This control depends on the approximation properties of V through the bias $\|P_{i|V} - P_{i|S}\|_P^2$ and on the variance via the upper bound $|A|^{v-1}/n$. Our goal now is to find a subset V that balances these two terms. This is precisely the aim of the following result.

Theorem 3.2. *Let*

$$\widehat{V} = \arg\min_{V\in\mathcal{V}_s}\left\{-\|\widehat{P}_{i|V}\|_{\widehat{P}}^2 + \mathrm{pen}(V)\right\}, \qquad \textit{where } \mathrm{pen}(V) \geq 12\frac{|A|^{v-1}}{n}.$$

There exists a constant $\kappa = \kappa(|A|)$ such that, with probability larger than $1-\delta^{-1}$,

$$\|P_{i|S} - \widehat{P}_{i|\widehat{V}}\|_P^2 \leq \left(1+\frac{8}{\log(\delta)}\right)\inf_{V\in\mathcal{V}_s}\left\{\|P_{i|S} - P_{i|V}\|_P^2 + \mathrm{pen}(V)\right\} + \kappa\frac{(\log(N_s^2\delta))^2}{n}. \quad (3.2)$$

Comments.

- The bound can be integrated and yields

$$\mathbb{E}\left[\|P_{i|S} - \widehat{P}_{i|\widehat{V}}\|_P^2\right] \leq C_1 \inf_{V\in\mathcal{V}_s}\left\{\|P_{i|S} - P_{i|V}\|_P^2 + \frac{|A|^{v-1}}{n}\right\} + C_2\frac{(s\log M)^2}{n},$$

  for some absolute constant C_1 and a constant C_2 depending only on $|A|$. Therefore, $\widehat{V}$ optimizes the bound given by Theorem 3.1, up to the residual $(s\log(M))^2$ term, among all the subsets of $\mathcal{V}_s$.
- Enlarging the number of observed sites makes the control over all subsets in $\mathcal{V}_s$ harder, leading to a $(s\log M)^2$ loss in the rates. On the other hand, it is helpful to reduce the bias as will be shown in the next section.
- A very interesting feature of this result for the applications is that it holds without restrictions on P and the size of A or S in (S, A, P).

4. Computation of the bias

To complete the study of our estimator, it remains to understand the bias $\|P_{i|S} - P_{i|V}\|_P^2$. We present two important examples where explicit upper bounds can be obtained.

4.1. The Ising model

Let $S = \mathbb{Z}^d$ and let $(J_{i,j})_{(i,j)\in S^2}$ be an interaction potential, which is a collection of real numbers such that for any $i \neq j \in S$, $J_{i,i} = 0$, $J_{i,j} = J_{j,i}$ and

$$\beta := \sup_{i\in S}\sum_{j\in S}|J_{i,j}| < \infty.$$

The parameter $1/\beta$ is also called the temperature parameter in the physic literature where the model was initially introduced, see [15]. The Ising model is the triplet (S, A, P), where $A = \{-1, 1\}$ and P is given by its specifications by

$$P_{i|S}(x) = \frac{\mathrm{e}^{\sum_{j\in S}J_{i,j}x(i)x(j)}}{\mathrm{e}^{\sum_{j\in S}J_{i,j}x(i)x(j)} + \mathrm{e}^{-\sum_{j\in S}J_{i,j}x(i)x(j)}} = \frac{1}{1+\mathrm{e}^{-2\sum_{j\in S}J_{i,j}x(i)x(j)}}.$$

It follows from Theorem 4.5 in [18] that

$$\|P_{i|S} - P_{i|V}\|_P \leq \sup_{x \in \mathcal{X}(S)} \left|P_{i|S}(x) - P_{i|V}(x)\right| \leq C_\beta \sum_{j \notin V} |J_{i,j}|.$$

Rates of convergence can be obtained from this bound and our model selection theorem. For example, let $d_\infty(i, j) = \max\{|i_k - j_k|\colon k \in \{1, \dots, d\}\}$, assume that $s \log M = \mathrm{O}((\log n)^2)$ and that there exists constants r and r' such that $\sum_{j \in S : d_\infty(i,j) > k} |J_{i,j}| \leq k^{-r}$ and $\sum_{j > k} |J^*_{i,j}| \leq \mathrm{e}^{-r'k}$, where $J^*_{i,j}$ denote the rearrangement of the $J_{i,j}$ by decreasing absolute values. Then, for any $i \in V_M$, denoting by α_i the largest real number such that $\{j \in \mathbb{Z}\colon d_\infty(i, j) \leq n^{\alpha_i}\} \subset V_M$, we have

$$\begin{aligned}\mathbb{E}\big[\|P_{i|S} - P_{i|\hat{V}}\|_P^2\big] &\leq C\frac{(\log n)^4}{n} + C_\beta\big(n^{-\alpha_i r} + n^{-2r'/(2r'+\log 2)}\big)\\ &\leq C_\beta n^{-(\alpha_i r \wedge 2r'/(2r'+\log 2))}.\end{aligned}$$

Other consequences of this bound obtained under different assumptions on the $(J_{i,j})_{i,j \in S}$ are discussed in Section A.3.

4.2. The Gibbs model

Assume that A is a finite set of real numbers in $[-1, 1]$, $S = \mathbb{Z}^d$ for some $d \geq 1$. Let $((J^{(k)}_{i,i_1,\dots,i_k})_{(i,i_1,\dots,i_k) \in S^{k+1}})_{k \geq 0} \in \prod_{k \geq 0} \mathbb{R}^{k+1}$ be a collection of real numbers such that

$$\sum_{k \geq 0} \sum_{(i,i_1,\dots,i_k) \in S^{k+1}} \left|J^{(k)}_{i,i_1,\dots,i_k}\right| = \beta < \infty.$$

For any $x \in \mathcal{X}(S)$ and $i \in S$, denote by

$$J_i(x) = \sum_{k \geq 0} \sum_{(i_1,\dots,i_k) \in S^k} J^{(k)}_{i,i_1,\dots,i_k} \prod_{\ell=1}^{k} x(i_\ell).$$

Suppose that the conditional probabilities can be written in the following way:

$$P_{i|S}(x) = \frac{\mathrm{e}^{x(i)J_i(x)}}{\sum_{a \in A} \mathrm{e}^{aJ_i(x)}}.$$

The triplet (S, A, P) is called a *Gibbs* model, Ising models are special instances of Gibbs models where for all $k \geq 2$ and all $(j_1, \dots, j_k) \in S^k$, $J_{i,j_1,\dots,j_k} = 0$. For any $\ell \leq M$, denote by $(J^*_{i,\ell,n})_{n=1,\dots,M^\ell}$ the rearrangement of the $J^{(\ell)}_{i,i_1,\dots,i_\ell}$ by decreasing absolute values. We consider the following assumption.

$$\forall \ell, n \in \mathbb{N}^*, \qquad \sum_{r \geq n} \left|J^*_{i,\ell,r}\right| \leq \beta \mathrm{e}^{-\gamma \ell^{2+\alpha} n}, \tag{J}$$

for some constant γ and $\alpha > 0$. Under Assumption (J), we can build a set V with cardinality $v \leq \frac{1+2\alpha}{\gamma\alpha+\log|A|(1+2\alpha)} \log n$ such that the bias term is upper bounded by

$$\|P_{i|S} - P_{i|V}\|_P^2 \leq C_{\alpha,\beta,\gamma,|A|}\left(\frac{(\log n)^{1/(2+\alpha)}}{n^{\alpha\gamma/(\gamma\alpha+\log|A|(1+2\alpha))}} + \sum_{\ell\geq 1}\sum_{i_1,\dots i_\ell \in S:\exists j; i_j\notin V_M} \left|J^{(\ell)}_{i,i_1,\dots,i_\ell}\right|\right). \quad (4.1)$$

The bound (4.1) is proved in Section A.3. From Theorem 3.1 and $v \leq \frac{1+2\alpha}{\gamma\alpha+\log|A|(1+2\alpha)} \log n$, for some absolute constant C,

$$\mathbb{E}\left[\|\widehat{P}_{i|V} - P_{i|V}\|_P^2\right] \leq C\frac{|A|^{v-1}}{n} = \frac{C}{|A|n^{\alpha\gamma/(\gamma\alpha+\log|A|(1+2\alpha))}}.$$

Therefore, for some constant $C_{\alpha,\beta,\gamma,|A|}$ and rate $\theta = \frac{2\alpha\gamma}{2\alpha\gamma+(1+2\alpha)\log|A|}$,

$$\mathbb{E}\left[\|P_{i|S} - \widehat{P}_{i|\widehat{V}}\|_P^2\right] \leq C_{\alpha,\beta,\gamma,|A|}\left[\left(\frac{\log n}{n}\right)^{\theta} + \sum_{\ell\leq v}\sum_{i_1,\dots i_\ell \in S:\exists j; i_j\notin O} \left|J^{(\ell)}_{i,i_1,\dots,i_\ell}\right|\right].$$

5. Slope heuristic

The slope heuristic was introduced in [8]. Let

$$\widehat{V} = \arg\min_{V\in\mathcal{V}_s}\left\{-\|\widehat{P}_{i|V}\|_{\widehat{P}}^2 + \mathrm{pen}(V)\right\}. \quad (5.1)$$

The heuristic states that there exist a minimal penalty $\mathrm{pen}_{\min}$ and a complexity measure (to be defined) satisfying the following properties.

- SH1 When $\mathrm{pen}(V) < (1-\eta)\,\mathrm{pen}_{\min}(V)$, the complexity of $\widehat{V}$ is as large as possible.
- SH2 When $\mathrm{pen}(V) = (1+\eta)\,\mathrm{pen}_{\min}(V)$, the complexity of $\widehat{V}$ is much smaller.
- SH3 When $\mathrm{pen}(V) = 2\,\mathrm{pen}_{\min}(V)$, the risk of $\widehat{V}$ is equivalent to the oracle risk.

The purpose of this section is to justify this heuristic. We will show some theoretical evidence for the slope heuristic using $\Delta_V = \|\widehat{P}_{i|V} - P_{i|V}\|_{\widehat{P}}^2$ as a complexity measure for V and as a minimal penalty. It may be useful for the intuition to make the following approximation $n\Delta_V/|A|^v \approx C$ although it is only proved in Theorem 3.1 that $\mathbb{E}[\Delta_V] \leq C|A|^v/n$. For example, this explains why it's natural to consider Δ_V as a measure of complexity. The following theorem gives some theoretical grounds justifying SH1.

Theorem 5.1. *Let $r > 0$, $\epsilon > 0$. Let $\widehat{V}$ be defined by (5.1) and assume that*

$$\mathbb{P}\left(\forall V \in \mathcal{V}_s, 0 \leq \mathrm{pen}(V) \leq (1-r)\|\widehat{P}_{i|V} - P_{i|V}\|_{\widehat{P}}^2\right) \geq 1-\epsilon.$$

Then, for all $\delta > 2$, with probability larger than $1 - \epsilon - 2\delta^{-1}$,

$$\|P_{i|\widehat{V}} - \widehat{P}_{i|\widehat{V}}\|_{\widehat{P}}^2 \geq \sup_{V\in\mathcal{V}_s}\left\{r\|P_{i|V} - \widehat{P}_{i|V}\|_{\widehat{P}}^2 - 2\|P_{i|S} - P_{i|V}\|_P^2\right\} - \frac{17}{3}\frac{(\log(N_s^2\delta))^2}{n}.$$

Comments.

- Let us give some intuition on this result. Algebraic computations, see (A.9), show that $\widehat{V}$ minimizes, up to centered remainder terms, the quantity

$$\|P_{i|S} - P_{i|V}\|_P^2 + \mathrm{pen}(V) - \|\widehat{P}_{i|V} - P_{i|V}\|_{\widehat{P}}^2. \tag{5.2}$$

  We assume in Theorem 5.1 that $\mathrm{pen}(V) = (1-\eta)\Delta_V$, thus $\widehat{V}$ minimizes the bias minus $\eta\Delta_V$. When the bias term decreases with V, as in the models presented in Section 4 and when $n\Delta_V/|A|^v \approx C$, both terms decrease with V and the minimum is achieved for $\widehat{V} = V_M$. Thus, $\widehat{V}$ maximizes the complexity Δ_V.
- Theorem 5.1 makes this statement more precise, showing that this result actually holds when, for $V = V_M$, both the bias and the logarithmic remainder term are negligible compared to the variance part of the risk.

Let us now turn to the associated optimal penalty theorem which proves SH2 and SH3.

Theorem 5.2. *Let $\delta > 5$, $r_2 \geq r_1 > 0$, $\epsilon > 0$ and assume that*

$$\mathbb{P}\left(\forall V \in \mathcal{V}_s, (1+r_1) \leq \frac{\mathrm{pen}(V)}{\|\widehat{P}_{i|V} - P_{i|V}\|_{\widehat{P}}^2} \leq (1+r_2)\right) \geq 1-\epsilon. \tag{5.3}$$

Let $\widehat{V}$ be defined by (5.1). For all V in $\mathcal{V}_s$, let $p_-^V = \inf_{x\in\mathcal{X}(V), P(x(V))\neq 0} P(x(V))$ and assume that, for some $\varepsilon \leq 1$,

$$\inf_{V\in\mathcal{V}_s} p_-^V \geq \varepsilon^{-2}\frac{\log(nN_s\delta)}{n}.$$

Then, there exists an absolute constant C such that, with probability larger than $1 - 5\delta^{-1} - \epsilon$, for all V in $\mathcal{V}_s$, for all $\eta > 0$,

$$\frac{(1-\eta)\wedge(r_1 - C(1+r_1)\varepsilon)}{(1+\eta)\vee(r_2 + C(1+r_2)\varepsilon)}\|P_{i|S} - \widehat{P}_{i|\widehat{V}}\|_P^2 \leq \|P_{i|S} - \widehat{P}_{i|V}\|_P^2 + \frac{6}{\eta}\frac{(\log(N_s^2\delta))^2}{n}. \tag{5.4}$$

Comments.

- In this theorem, following [2], the main task is to show that

$$\Delta_V \simeq \|\widehat{P}_{i|V} - P_{i|V}\|_P^2. \tag{5.5}$$

  When (5.3) holds with $r_1 = r_2 = r$, then

$$\mathrm{pen}(V) = (1+r)\Delta_V \simeq \Delta_V + r\|\widehat{P}_{i|V} - P_{i|V}\|_P^2.$$

  From (5.2), $\widehat{V}$ minimizes the sum of the bias and r times the variance. The complexity should thus be much smaller, which proves SH2 for $\mathrm{pen}_{\min}(V) = \Delta_V$. Theorem 5.2 shows that the complexity of the selected model, that is bounded by the risk, is actually upper bounded by the supremum between the oracle risk and the remainder term, at least when ε is small enough.

- Take then $r_1 = r_2 = 1$, that is, a penalty equal to

$$\text{pen}(V) = 2\,\text{pen}_{\min}(V) \simeq \Delta_V + \|\widehat{P}_{i|V} - P_{i|V}\|_P^2.$$

  Then (5.2) shows that $\widehat{V}$ minimizes an approximately optimal criterion, and $\widehat{P}_{i|\widehat{V}}$ satisfies an oracle inequality that is asymptotically optimal, which proves SH3. Inequality (5.4) makes this result more precise, showing that the oracle inequality is indeed asymptotically optimal when the oracle rate of convergence is larger than the remainder term. Moreover, in this case, the rate of convergence of the leading quantity in the oracle is driven by the supremum of the rates η and ε.

Theorem 5.2 cannot be used directly to build an estimator since the complexity is unknown. Nevertheless, Theorem 3.1 shows that Δ_V is upper bounded by $K\Theta_V$, with $\Theta_V = |A|^{v-1}/n$ and some constant K that may not be optimal. This suggests to consider penalties of the form $K\Theta_V$, for some K that has to be optimized. To achieve this goal, [2] proposed the following algorithm.

1. For all $K > 0$, denote by $\widehat{V}(K)$ the model selected with $\text{pen}(V) = K\Theta_V$.
2. Find $K_{\min}$ such that $\Theta_{\widehat{V}(K)}$ is very large for $K < K_{\min}$ and much smaller for $K > K_{\min}$.
3. Select $\widehat{V} = \widehat{V}(2K_{\min})$.

This algorithm is based on the slope heuristic. Indeed, assume that $\text{pen}_{\min}(V) = K_0\Theta_V$ for some unknown K_0. Then, $K_{\min}$ shall be close to K_0 because we observe a jump of the complexity $\Theta_{\widehat{V}}$ around $K_{\min}\Theta_V$ as expected by SH1, SH2. Therefore, $\widehat{V}$, chosen by $2K_{\min}\Theta_V \simeq 2\,\text{pen}_{\min}(V)$ shall be optimal from SH3. We did not prove that this algorithm improves the choice of K in theory but the simulation study of the next section presents examples where it does in practice.

6. Simulation studies

In this section, we illustrate the results obtained in previous ones using simulation experiments. All the simulations were implemented by a set of MATLAB® routines that can be downloaded from www.princeton.edu/~dtakahas/publications/LT11routines.zip.

Let $S = \{1, \ldots, 9\}$ and $A = \{-1, 1\}$. For the first simulation, we consider an Ising model (S, A, P), with one-point specification probabilities given by

$$\forall x \in \mathcal{X}(S), \qquad P_{i|S}(x) = \frac{1}{1 + \exp(-2\sum_{j\in S} J_{ij}x(i)x(j))},$$

where the J_{ij}'s are given by $J_{1,2} = J_{1,5} = -J_{2,5} = J_{1,9} = J_{2,9} = J_{3,6} = -J_{4,7} = -J_{4,8} = -J_{7,8} = J_{6,8} = 0.5$. The rest of J_{ij}'s are equal to zero. For each $i \in S$, the pair of sites (i, j) where $j \in V_i$ is shown in Figure 1(A). For the first experiment, we study the site $i = 9$ and its interaction sites. We simulate independent samples of the Ising model and compare the performances of the model selection procedures given by (1) the penalty given in Theorem 3.2 (theoretical), (2) the same penalty, but using the slope algorithm described in Section 5 to calibrate the constant in front of $|A|^{v-1}/n$, and (3) the L_∞-risk method with slope heuristic proposed in [18]. The performances of the estimators are measured by the logarithm of the ratio between the risk

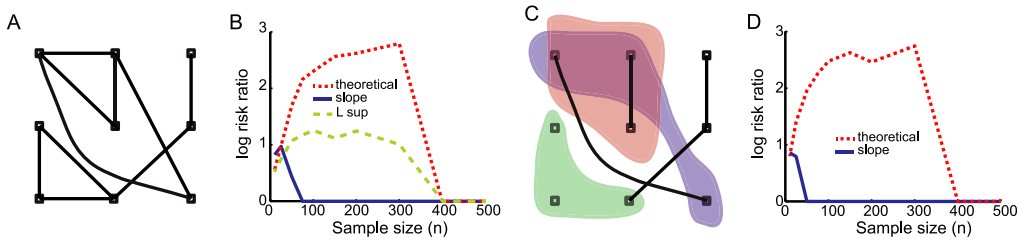

Figure 1. Simulation study. (A) Representation of the interacting pairs of the Ising model used in the first simulation experiment. The numbering of the sites increases from the top left to the bottom right. (B) Performance of the model selection for the first experiment. Plot of the log risk ratio for the model selection procedure using $K = 2$ (dotted red line), optimizing the constant using the slope heuristic (solid blue line), using the L_∞-risk method with slope heuristic (dashed yellow). (C) Representation of the interacting neurons of the Gibbs model used in the second simulation experiment. The colored regions represent the three-way interactions. (D) Performance of the model selection for the second experiment. The legend is the same as in (B).

of the estimated model and the oracle risk. Figure 1(B) shows the median value of the risk ratio calculated for 100 independent replicas. The maximum number of allowed interacting sites was set to $s = 5$. The simulations were done for increasing sample sizes $n = 10, 25, 50, 75, 100, 150, 200, 300, 400, 500$.

For the second simulation, we consider a Gibbs model (S, A, P), with one-point conditional probabilities given by

$$\forall x \in \mathcal{X}(S), \qquad P_{i|S}(x) = \frac{1}{1 + \exp(-2\sum_{j\in S} J_{ij}x(i)x(j) + \sum_{k\in S}\sum_{j\in S} J_{ijk}x(i)x(j)x(k))}.$$

The non-null pairwise interactions are given by $-J_{2,5} = J_{1,9} = J_{3,6} = J_{6,8} = 0.5$, and the three-way interactions are specified by $J_{1,2,5} = J_{1,2,9} = -J_{4,7,8} = 0.5$. The rest of J_{ij}'s and J_{ijk}'s are equal to zero. For each i, the interacting neighborhood V_i is shown in Figure 1(C). We show the results for $i = 9$. We compute the risk ratio as in the first experiment (Figure 1(D)). The simulations are done for increasing sample sizes $n = 10, 25, 50, 75, 100, 150, 200, 300, 400, 500$ (Figure 1(D)). Observe that in both experiments the slope heuristic improves the performance of the model selection, allowing to recover the oracle even for data set as small as 50 in our examples. For this example, any method that uses the Ising model to estimate the parameters has a non-null bias and therefore the risk will be strictly larger than the oracle risk. Further simulations are shown in the Appendix (Section C).

7. Application to multi-unit neuronal data

In this section, we illustrate the usefulness of the proposed methods on experimental data set. In neuroscience, it is conjectured that the set of interacting neurons represents different animal behaviors [24]. Modifications of the graph of interacting neurons for different tasks have been

repeatedly shown [24]. Nevertheless, if this hypothesis has any validity, we expect the set of interacting neurons to be the same when the same task is performed. We used our method here to test this hypothesis, which seems to be less verified in the literature.

The data set used contains multichannel simultaneous recordings made from layer CA1 of the right dorsal hippocampus of a Long-Evans rat during open field tasks in which the animal chased randomly placed drops of water while on a elevated square platform. It was downloaded from http://crcns.org/data-sets/hc/hc-2/about-hc-2. Details about the recording technique and experimental set up can be found at the website or in [21].

The spiking data set used is ec016.430.res.1, ec016.430.res.2, ec016.430.res.3, ec016.430.res.4, ec016.430.res.5, ec016.430.res.6, ec016.430.res.7, ec016.430.res.8. The full data set contains a total of 55 isolated neurons. For the analysis, we kept only the 11 neurons that showed more than 30 000 spikes during the experiment. The data set was sampled at 20 kHz. We binned the data with non-overlapping bins of size 10 ms. If there was at least one spike in the bin, we coded it as $+1$, otherwise we coded as -1. The spiking activity of the 11 neurons was recorded for 106.8 minutes. To ensure independence of the observations, we subsampled the data using one observation at each 500 ms, which is an order of magnitude larger than a typical decay of correlation (when the correlation becomes zero) between neurons in time. We then splitted the data into two parts, one sample for the first half of the experiment ($n = 64\,099$, first 53.4 min) and another sample for the second half of the experiment ($n = 64\,099$, second 53.4 min).

We computed our estimators of the interacting neurons and calibrate the constant in front of the penalty with the slope algorithm described in the end of Section 5. For each site, the maximum number of allowed interacting sites was $s = 3$. Figure 2 shows the results obtained for the first and second parts of the experiment. We clearly see that the interacting neuronal sites remained stable, with only one pair of interaction that changed between the two data sets. This result, together with those in the literature showing changes in interacting neighborhoods for different behaviors, corroborates the hypothesis that the set of interacting neurons can be related to specific animal behavior.

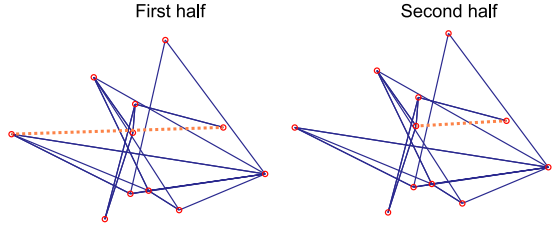

Figure 2. Representation of the interacting neuronal sites for the first half and second half of the experiment. The edges between sites indicate the interacting pairs. The dotted orange edges indicate the interactions that differed between both conditions. Observe that the interactions are represented by a graph for convenience of visualization, but for our method the interactions are not restricted to pairwise interaction as shown by our theoretical results and in Figure 1(D).

Appendix A: Proofs

A.1. Proof of Theorem 3.1

Let $\theta > 0$ to be chosen later and let Q denote either P or $\widehat{P}$. We decompose the risk as follows

$$\begin{aligned}\|\widehat{P}_{i|V} - P_{i|V}\|_Q^2 &= \sum_{x\in\mathcal{X}(V)} \frac{Q(x(V/\{i\}))}{|A|}\big(\widehat{P}_{i|V}(x) - P_{i|V}(x)\big)^2 \\ &= \sum_{x\in\mathcal{X}(V),\, Q(x(V/\{i\}))\le\theta(|A|^v n)^{-1}} \frac{Q(x(V/\{i\}))}{|A|}\big(\widehat{P}_{i|V}(x) - P_{i|V}(x)\big)^2 \\ &\quad + \sum_{x\in\mathcal{X}(V),\, Q(x(V/\{i\}))>\theta(|A|^v n)^{-1}} \frac{Q(x(V/\{i\}))}{|A|}\big(\widehat{P}_{i|V}(x) - P_{i|V}(x)\big)^2.\end{aligned}$$

As the cardinal of $\mathcal{X}(V)$ is $|A|^v$ and $(\widehat{P}_{i|V}(x) - P_{i|V}(x))^2 \le 1$, the first term in this decomposition is upper bounded by θn^{-1}. Hence

$$\|\widehat{P}_{i|V} - P_{i|V}\|_Q^2 = \frac{\theta}{n} + \sum_{x\in\mathcal{X}(V),\, Q(x(V/\{i\}))>\theta(|A|^v n)^{-1}} \frac{Q(x(V/\{i\}))}{|A|}(\widehat{P}_{i|V} - P_{i|V})^2. \tag{A.1}$$

Hereafter in the proof of Theorem 3.1, we denote by

$$\mathcal{X}^\theta(V) = \big\{x \in \mathcal{X}(V):\ Q\big(x\big(V/\{i\}\big)\big) > \theta\big(|A|^v n\big)^{-1}\big\}.$$

It comes from Lemma B.1 that

$$\begin{aligned}&\|\widehat{P}_{i|V} - P_{i|V}\|_P^2 - \frac{\theta}{n} \\ &= \sum_{x\in\mathcal{X}^\theta(V)} \frac{P(x(V/\{i\}))}{|A|}\big(\widehat{P}_{i|V}(x) - P_{i|V}(x)\big)^2 \\ &\le \sum_{x\in\mathcal{X}^\theta(V)} \frac{(|\widehat{P}(x(V)) - P(x(V))| + \widehat{P}_{i|V}(x)|(\widehat{P}(x(V/\{i\})) - P(x(V/\{i\})))|)^2}{|A|P(x(V/\{i\}))} \\ &\le \frac{2}{|A|}\Bigg(\sum_{x\in\mathcal{X}^\theta(V)} \frac{(\widehat{P}(x(V)) - P(x(V)))^2}{P(x(V/\{i\}))} + \sum_{x\in\mathcal{X}^\theta(V/\{i\})} \frac{(\widehat{P}(x(V/\{i\})) - P(x(V/\{i\})))^2}{P(x(V/\{i\}))}\Bigg).\end{aligned}$$

From Lemma B.1, we also have

$$\big|\widehat{P}_{i|V}(x) - P_{i|V}(x)\big| \le \frac{|\widehat{P}(x(V)) - P(x(V))| + P_{i|V}(x)|(\widehat{P}(x(V/\{i\})) - P(x(V/\{i\})))|}{|A|\widehat{P}(x(V/\{i\}))}.$$

Hence

$$\begin{aligned}&\big|\widehat{P}_{i|V}(x)-P_{i|V}(x)\big|\\&\quad\le\frac{|\widehat{P}(x(V))-P(x(V))|+(P_{i|V}(x)+\widehat{P}_{i|V}(x))|(\widehat{P}(x(V/\{i\}))-P(x(V/\{i\})))|}{|A|\sqrt{\widehat{P}(x(V/\{i\}))P(x(V/\{i\}))}}.\end{aligned}$$

Thus,

$$\|\widehat{P}_{i|V}-P_{i|V}\|_{\widehat{P}}^2-\frac{\theta}{n}=\sum_{x\in\mathcal{X}^\theta(V)}\frac{\widehat{P}(x(V/\{i\}))}{|A|}\big(\widehat{P}_{i|V}(x)-P_{i|V}(x)\big)^2$$

is smaller than

$$\begin{aligned}&\sum_{x\in\mathcal{X}^\theta(V)}\frac{(|\widehat{P}(x(V))-P(x(V))|+(\widehat{P}_{i|V}(x)+P_{i|V}(x))|(\widehat{P}(x(V/\{i\}))-P(x(V/\{i\})))|)^2}{|A|P(x(V/\{i\}))}\\&\quad\le\frac{2}{|A|}\bigg(\sum_{x\in\mathcal{X}^\theta(V)}\frac{(\widehat{P}(x(V))-P(x(V)))^2}{P(x(V/\{i\}))}+2\sum_{x\in\mathcal{X}^\theta(V/\{i\})}\frac{(\widehat{P}(x(V/\{i\}))-P(x(V/\{i\})))^2}{P(x(V/\{i\}))}\bigg).\end{aligned}$$

We use Theorem B.8 with $b=\sqrt{\theta^{-1}|A|^v n}$, for all $x>0$, for all $\eta>0$, we have, with probability larger than $1-2\mathrm{e}^{-x}$,

$$\|\widehat{P}_{i|V}-P_{i|V}\|_Q^2\le\frac{\theta}{n}+\frac{6}{|A|}\bigg((1+\eta)^3\frac{|A|^v}{n}+\frac{4x}{\eta n}+\frac{32|A|^v x^2}{\theta\eta^3 n}\bigg).$$

Take $\theta=8|A|^{v/2}x\eta^{-3/2}$, we obtain

$$\|\widehat{P}_{i|V}-P_{i|V}\|_Q^2\le\frac{6}{|A|}\bigg((1+\eta)^3\frac{|A|^v}{n}+\frac{4x}{\eta n}+\frac{6|A|^{v/2}x}{\eta^{3/2}n}\bigg).$$

Using $ab\le\eta a^2+(4\eta)^{-1}b^2$, we finally get

$$\|\widehat{P}_{i|V}-P_{i|V}\|_Q^2\le\frac{6}{|A|}\bigg((1+8\eta)\frac{|A|^v}{n}+\frac{4x}{\eta n}+\frac{9x^2}{\eta^4 n}\bigg).$$

A.2. Proof of Theorem 3.2

The theorem follows from the slightly more general following result.

Theorem A.1. *Let $K>1$ and let*

$$\widehat{V}=\arg\min_{V\in\mathcal{V}_s}\big\{-\|\widehat{P}_{i|V}\|_{\widehat{P}}^2+\mathrm{pen}(V)\big\},\qquad\textit{where }\mathrm{pen}(V)\ge 6K\frac{|A|^{v-1}}{n}.$$

Then, there exists a constant $\kappa = \kappa(|A|, K)$ *such that for all* $\delta \geq 1$, *with probability larger than* $1 - \delta^{-1}$,

$$\|P_{i|S} - \widehat{P}_{i|\widehat{V}}\|_P^2 \leq \kappa\left(\inf_{V \in \mathcal{V}_s}\left\{\|P_{i|S} - P_{i|V}\|_P^2 + \mathrm{pen}(V)\right\} + \frac{(\log(N_s^2\delta))^2}{n}\right). \tag{A.2}$$

Moreover, when $K \geq 2$, *there exists a constant* $\kappa = \kappa(|A|, K)$ *such that, with probability larger than* $1 - \delta^{-1}$,

$$\|P_{i|S} - \widehat{P}_{i|\widehat{V}}\|_P^2 \leq \left(1 + \frac{8}{\log(\delta)}\right)\inf_{V \in \mathcal{V}_s}\left\{\|P_{i|S} - P_{i|V}\|_P^2 + \mathrm{pen}(V)\right\} + \kappa\frac{(\log(N_s^2\delta))^2}{n}. \tag{A.3}$$

Proof. For $Q \in \{P, \widehat{P}\}$, let $(\cdot,\cdot)_Q$ be the scalar product associated to the $L_{2,Q}$-norm $\|\cdot\|_Q$. Let V and V' in the collection $\mathcal{V}_s$. We have

$$\begin{aligned}
&\frac{1}{|A|}\sum_{x \in \mathcal{X}(V \cup V')} \widehat{P}\big(x\big(V \cup V'\big)\big)P_{i|V}(x) \\
&\quad = \sum_{x \in \mathcal{X}(V)} \frac{\widehat{P}(x(V/\{i\}))}{|A|}\widehat{P}_{i|V}(x)P_{i|V}(x) = (\widehat{P}_{i|V}, P_{i|V})_{\widehat{P}}, \\
&\frac{1}{|A|}\sum_{x \in \mathcal{X}(V \cup V')} P\big(x\big(V \cup V'\big)\big)P_{i|V}(x) = \sum_{x \in \mathcal{X}(V)} \frac{P(x(V/\{i\}))}{|A|}P_{i|V}^2(x) = \|P_{i|V}\|_P^2.
\end{aligned}$$

Hence, for all V, V' in $\mathcal{V}_s$,

$$\begin{aligned}
\|\widehat{P}_{i|V}\|_{\widehat{P}}^2 &= \|P_{i|V}\|_{\widehat{P}}^2 + 2(\widehat{P}_{i|V} - P_{i|V}, P_{i|V})_{\widehat{P}} + \|\widehat{P}_{i|V} - P_{i|V}\|_{\widehat{P}}^2 \\
&= \|P_{i|V}\|_P^2 + \|\widehat{P}_{i|V} - P_{i|V}\|_{\widehat{P}}^2 - \big(\|P_{i|V}\|_{\widehat{P}}^2 - \|P_{i|V}\|_P^2\big) \\
&\quad + \frac{2}{|A|}\sum_{x \in \mathcal{X}(V \cup V')}\big(\widehat{P}\big(x\big(V \cup V'\big)\big) - P\big(x\big(V \cup V'\big)\big)\big)P_{i|V}(x).
\end{aligned} \tag{A.4}$$

Moreover, from Pythagoras relation see Proposition B.11, we have

$$\|P_{i|S} - P_{i|V}\|_P^2 = \|P_{i|S}\|_P^2 - \|P_{i|V}\|_P^2.$$

By definition of $\widehat{V}$, we have, for all V in $\mathcal{V}_s$,

$$\|P_{i|S}\|_P^2 - \|\widehat{P}_{i|\widehat{V}}\|_{\widehat{P}}^2 + \mathrm{pen}(\widehat{V}) \leq \|P_{i|S}\|_P^2 - \|\widehat{P}_{i|V}\|_{\widehat{P}}^2 + \mathrm{pen}(V).$$

Hence, for all $0 < \nu \leq 1$, from (A.4),

$$\nu\|P_{i|S} - \widehat{P}_{i|\widehat{V}}\|_P^2 \leq \|P_{i|S} - P_{i|\widehat{V}}\|_P^2 + \nu\|P_{i|\widehat{V}} - \widehat{P}_{i|\widehat{V}}\|_P^2$$

is smaller than

$$\begin{aligned}
&\|P_{i|S} - P_{i|V}\|_P^2 + \mathrm{pen}(V) - \|\widehat{P}_{i|V} - P_{i|V}\|_{\widehat{P}}^2 \\
&\quad - \big(\mathrm{pen}(\widehat{V}) - \|\widehat{P}_{i|\widehat{V}} - P_{i|\widehat{V}}\|_{\widehat{P}}^2 - \nu\|\widehat{P}_{i|\widehat{V}} - P_{i|\widehat{V}}\|_P^2\big) \\
&\quad + \big(\|P_{i|V}\|_{\widehat{P}}^2 - \|P_{i|V}\|_P^2 - \|P_{i|\widehat{V}}\|_{\widehat{P}}^2 + \|P_{i|\widehat{V}}\|_P^2\big) \\
&\quad + \frac{2}{|A|}\sum_{x\in\mathcal{X}(V\cup\widehat{V})}\big(\widehat{P}\big(x(V\cup\widehat{V})\big) - P\big(x(V\cup\widehat{V})\big)\big)\big(P_{i|\widehat{V}}(x) - P_{i|V}(x)\big).
\end{aligned} \tag{A.5}$$

We have also,

$$\begin{aligned}
&\|P_{i|V}\|_{\widehat{P}}^2 - \|P_{i|V}\|_P^2 - \|P_{i|\widehat{V}}\|_{\widehat{P}}^2 + \|P_{i|\widehat{V}}\|_P^2 \\
&\quad = \frac{1}{|A|}\sum_{x\in\mathcal{X}((V\cup\widehat{V}))}\big(\widehat{P}\big(x\big((V\cup\widehat{V})/\{i\}\big)\big) - P\big(x\big((V\cup\widehat{V})/\{i\}\big)\big)\big)\big(P_{i|V}^2(x) - P_{i|\widehat{V}}^2(x)\big).
\end{aligned}$$

Let $0 < \eta \le 1$, $\delta > 1$ and assume that, $N_s \ge 2$. Let Ω^δ be the intersection of the following events:

$$\Omega_1^\delta = \left\{\forall V \in \mathcal{V}_s, \|\widehat{P}_{i|V} - P_{i|V}\|_{\widehat{P}}^2 \le \frac{6}{|A|}\left((1+8\eta)\frac{|A|^v}{n} + \frac{13\log(2N_s\delta)^2}{\eta^4 n}\right)\right\},$$

$$\Omega_2^\delta = \left\{\forall V \in \mathcal{V}_s, \|\widehat{P}_{i|V} - P_{i|V}\|_P^2 \le \frac{6}{|A|}\left((1+8\eta)\frac{|A|^v}{n} + \frac{13\log(2N_s\delta)^2}{\eta^4 n}\right)\right\},$$

$$\begin{aligned}
\Omega_3^\delta = \Big\{&\forall V, V' \in \mathcal{V}_s^2, \|P_{i|V}\|_{\widehat{P}}^2 - \|P_{i|V}\|_P^2 - \|P_{i|V'}\|_{\widehat{P}}^2 + \|P_{i|V'}\|_P^2 \\
&\le 2\|P_{i|V} - P_{i|V'}\|_P\sqrt{2\frac{\log(N_s^2\delta)}{n}} + \frac{\log(N_s^2\delta)}{3n}\Big\},
\end{aligned} \tag{A.6}$$

$$\begin{aligned}
\Omega_4^\delta = \Big\{&\forall V, V' \in \mathcal{V}_s^2, \sum_{x\in\mathcal{X}(V\cup V')}\big(\widehat{P}\big(x\big(V\cup V'\big)\big) - P\big(x\big(V\cup V'\big)\big)\big)\frac{P_{i|V'}(x) - P_{i|V}(x)}{a} \\
&\le \|P_{i|V} - P_{i|V'}\|_P\sqrt{2\frac{\log(N_s^2\delta)}{n}} + \frac{\log(N_s^2\delta)}{3n}\Big\}.
\end{aligned} \tag{A.7}$$

Theorem 3.1, Lemma B.10 and union bounds give that

$$P\big((\Omega^\delta)^c\big) \le \frac{4}{\delta}.$$

For all V, V' in $\mathcal{V}_s$ and all $\xi > 0$, on Ω^δ, we have

$$2 \sum_{x \in \mathcal{X}(V \cup V')} \big(\widehat{P}\big(x\big(V \cup V'\big)\big) - P\big(x\big(V \cup V'\big)\big)\big) \frac{P_{i|V'}(x) - P_{i|V}(x)}{|A|} + \|P_{i|V}\|^2_{\widehat{P}}$$

$$- \|P_{i|V}\|^2_P - \|P_{i|V'}\|^2_{\widehat{P}} + \|P_{i|V'}\|^2_P \le \frac{\xi}{2}\|P_{i|V} - P_{i|V'}\|^2_P + \left(\frac{16}{\xi} + 1\right)\frac{\log(N_s^2\delta)}{3n}.$$

From (A.5), we deduce that, on Ω^δ, for all $0 < \xi < \eta$,

$$(\nu - \xi)\|P_{i|S} - \widehat{P}_{i|\widehat{V}}\|^2_P \le (1+\xi)\|P_{i|S} - P_{i|V}\|^2_P + \mathrm{pen}(V)$$
$$- \left(\mathrm{pen}(\widehat{V}) - (1+\nu)(1+\eta)^3 \frac{6}{|A|}\frac{|A|^{\widehat{v}}}{n}\right)$$
$$+ \frac{1}{n}\left(\frac{78(1+\nu)}{\eta^4|A|}\big(\log(2N_s\delta)\big)^2 + \left(\frac{16}{\xi} + 1\right)\log\big(N_s^2\delta\big)\right).$$

Take at first $0 < \xi < \nu$ and $0 < \eta$ sufficiently small to ensure that $(1+\nu)(1+\eta)^3 \le K$ to obtain (A.2). To obtain (A.3), choose $\nu = 1$ and $\eta > 0$ sufficiently small to ensure that $(1+\eta)^3 < K/2$ and $\xi = (\log(N_s^2\delta))^{-1}$. We conclude the proof, saying that the inequality is obvious when $\delta < 4$, and, when $\delta \ge 4$,

$$\frac{1 + (\log N_s^2\delta)^{-1}}{1 - (\log N_s^2\delta)^{-1}} = 1 + \frac{2(\log N_s^2\delta)^{-1}}{1 - (\log N_s^2\delta)^{-1}} \le 1 + \frac{2(\log\delta)^{-1}}{1 - (\log\delta)^{-1}} \le 1 + \frac{8}{\log\delta}. \qquad \square$$

A.3. Proof of the bias control

A.3.1. Discussion on the Ising model

In this section, we discuss some consequences of the bound given on the bias term in the Ising model, under additional assumptions on the $J'_{i,j}s$.

1. Assume that the set of $j \in S$ such that $J_{i,j} \neq 0$, $\mathcal{N}_i$ is finite and that $\mathcal{N}_i \subset V_M$. The bound (A.3) implies that, when $\log_2(n) \ge |\mathcal{N}_i|$,

$$\mathbb{E}\big[\|P_{i|S} - \widehat{P}_{i|\hat{V}}\|^2_P\big] \le C\frac{(\log(n)\log(M))^2}{n} + C_\beta\frac{2^{|\mathcal{N}_i|}}{n} \le C_{\beta,|\mathcal{N}_i|}\frac{(\log(n)\log(M))^2}{n}.$$

2. Assume that there exist constants r and r' such that $M = n^r$ and, for any $k \in \mathbb{N}$, $\sum_{j>k}|J^*_{i,j}| \le \mathrm{e}^{-r'k}$, then

$$\mathbb{E}\big[\|P_{i|S} - \widehat{P}_{i|\hat{V}}\|^2_P\big] \le Cr^2\frac{\log(n)^4}{n} + C_\beta\left(\left(\sum_{j\notin V_M}|J_{i,j}|\right)^2 + n^{-2r'/(2r'+\log 2)}\right)$$
$$\le C_{r,\beta}\left(\left(\sum_{j\notin V_M}|J_{i,j}|\right)^2 + n^{-2r'/(2r'+\log 2)}\right).$$

A.3.2. Proof of the bound on the bias in the Gibbs case

In order to bound the bias term $\|P_{i|S} - P_{i|V}\|_P^2$, we still use the inequalities

$$\|P_{i|S} - P_{i|V}\|_P \leq \|P_{i|S} - P_{i|V}\|_\infty \leq \sup_{x,y\in\mathcal{X}(S):x(V\cup\{i\})=y(V\cup\{i\})} \left|P_{i|S}(x) - P_{i|S}(y)\right|.$$

Now, we will build an approximation set $V = \bigcup_{\ell=0}^{\log_{|A|} n} \mathcal{N}_\ell$ and bound the bias of $P_{i|V}$, using the inequality for any $v \leq |V|$,

$$\begin{aligned}
&\frac{|J_i(x) - J_i(y)|}{2} \\
&\leq \sum_{\ell\leq v} \sum_{i_1,\dots i_\ell \in S:\exists j; i_j\notin\mathcal{N}_\ell} \left|J^{(\ell)}_{i,i_1,\dots,i_\ell}\right| + \sup_{z\in\mathcal{X}(S)} \sum_{\ell>v} \left|J^{(\ell)}_i(z)\right| \\
&\leq \sum_{\ell\leq v} \sum_{i_1,\dots i_\ell \in S:\exists j; i_j\notin V_M} \left|J^{(\ell)}_{i,i_1,\dots,i_\ell}\right| + \sum_{i_1,\dots i_\ell \in V_M:\exists j; i_j\notin\mathcal{N}_\ell} \left|J^{(\ell)}_{i,i_1,\dots,i_\ell}\right| + \frac{\beta}{1-\mathrm{e}^{-\gamma}} \mathrm{e}^{-rv^{2+\alpha}}.
\end{aligned}$$

Let $\mathcal{N}_\ell$ denote the union of the K_ℓ ℓ-tuples $i_1,\dots,i_\ell$ such that $(J^*_{i,\ell,r})_{r=1,\dots,K_\ell}$ are indexed by the $\{(i,i_1,\dots,i_\ell), \text{s.t.}(i_1,\dots,i_\ell)\in\mathcal{N}_\ell\}$. $\mathcal{N}_\ell$ has a cardinality smaller than $K_\ell\ell$ and by assumption (J), we have

$$\sum_{i_1,\dots i_\ell \in O:\exists j; i_j\notin V_\ell} \left|J^{(\ell)}_{i,i_1,\dots,i_\ell}\right| \leq \beta \mathrm{e}^{-\gamma\ell^{2+\alpha}K_\ell}. \tag{A.8}$$

Now, let us fix some $\nu > 0$ and let $K_\ell = 1 + \lfloor \nu\ell^{-2-\alpha}\log n\rfloor$ for any $\ell \leq (\nu\log n)^{1/(2+\alpha)}$ and $K_\ell = 0$ when $\ell > (\nu\log n)^{1/(2+\alpha)}$. In particular, $K_\ell \geq \nu\ell^{-2-\alpha}\log n$ when $\ell \leq (\nu\log n)^{1/(2+\alpha)}$, hence, from (A.8), for any $1 \leq \ell \leq (\nu\log n)^{1/(2+\alpha)}$, we have

$$\sum_{i_1,\dots i_\ell \in V_M:\exists j; i_j\notin\mathcal{N}_\ell} \left|J^{(\ell)}_{i,i_1,\dots,i_\ell}\right| \leq \frac{\beta}{(1-\mathrm{e}^{-\gamma})n^{\nu\gamma}}.$$

Therefore, the bias term is upper bounded by

$$\|P_{i|S} - P_{i|V}\|_P^2 \leq C_{\alpha,\beta,\gamma,|A|}\left(\frac{\log n}{n^{\nu\gamma}} + \sum_{\ell\geq 1} \sum_{i_1,\dots i_\ell \in S:\exists j; i_j\notin O} \left|J^{(\ell)}_{i,i_1,\dots,i_\ell}\right|\right).$$

Moreover, V has cardinality upper bounded by

$$\sum_{\ell=1}^{(\nu\log n)^{1/(2+\alpha)}} \ell K_\ell \leq \sum_{\ell=1}^{(\nu\log n)^{1/(2+\alpha)}} \left(\ell + \frac{\nu\log n}{\ell^{1+\alpha}}\right) \leq \frac{1+2\alpha}{\alpha}\nu\log n.$$

A.4. Proof of Theorem 5.1

Let us introduce, for all V in $\mathcal{V}_s$,

$$L(V) = \|P_{i|V}\|_P^2 - \|P_{i|V}\|_{\widehat{P}}^2 + \frac{2}{|A|}\sum_{x\in\mathcal{X}(V)}\big(\widehat{P}\big(x(V)\big) - P\big(x(V)\big)\big)P_{i|V}(x).$$

By definition of $\widehat{V}$, we have, for all V in $\mathcal{V}_s$,

$$\|P_{i|S}\|_P^2 - \|\widehat{P}_{i|\widehat{V}}\|_{\widehat{P}}^2 + \mathrm{pen}(\widehat{V}) \le \|P_{i|S}\|_P^2 - \|\widehat{P}_{i|V}\|_{\widehat{P}}^2 + \mathrm{pen}(V).$$

Hence from inequality (A.4) in the proof of Theorem 3.2, we have, for all V in $\mathcal{V}_s$,

$$\begin{aligned}&\|P_{i|S} - P_{i|\widehat{V}}\|_P^2 + \big(\mathrm{pen}(\widehat{V}) - \|\widehat{P}_{i|\widehat{V}} - P_{i|\widehat{V}}\|_{\widehat{P}}^2\big) - L(\widehat{V})\\ &\quad\le \|P_{i|S} - P_{i|V}\|_P^2 + \big(\mathrm{pen}(V) - \|\widehat{P}_{i|V} - P_{i|V}\|_{\widehat{P}}^2\big) - L(V).\end{aligned} \tag{A.9}$$

Let $\Omega_{\mathrm{pen}} = \{0 \le \mathrm{pen}(V) \le (1-r)\|\widehat{P}_{i|V} - P_{i|V}\|_{\widehat{P}}^2\}$ and let $\Omega^\delta_{\text{min pen}} = \Omega_3^\delta \cap \Omega_4^\delta \cap \Omega_{\mathrm{pen}}$, where Ω_3^δ and Ω_4^δ are respectively defined in (A.6) and (A.7). It comes from Lemma B.10 and our assumption on $\mathrm{pen}(V)$ that $P((\Omega^\delta_{\text{min pen}})^c) \le \epsilon + 2\delta^{-1}$. Moreover, on $\Omega^\delta_{\text{min pen}}$, we have, for all $\eta > 0$,

$$\begin{aligned}&\big|L(\widehat{V}) - L(V)\big|\\ &\quad\le \eta\|P_{i|S} - P_{i|\widehat{V}}\|_P^2 + \eta\|P_{i|S} - P_{i|V}\|_P^2 + \left(\frac{16}{\eta} + 1\right)\frac{\log(N_s^2\delta)}{3n},\\ &(1-\eta)\|P_{i|S} - P_{i|\widehat{V}}\|_P^2 - \|\widehat{P}_{i|\widehat{V}} - P_{i|\widehat{V}}\|_{\widehat{P}}^2\\ &\quad\le (1+\eta)\|P_{i|S} - P_{i|V}\|_P^2 - r\|\widehat{P}_{i|V} - P_{i|V}\|_{\widehat{P}}^2 + \left(\frac{16}{\eta} + 1\right)\frac{\log(N_s^2\delta)}{3n}.\end{aligned}$$

We conclude the proof choosing $\eta = 1$.

A.5. Proof of Theorem 5.2

Let

$$\Omega_{\mathrm{pen}} = \big\{\forall V \in \mathcal{V}_s, (1+r_1)\|\widehat{P}_{i|V} - P_{i|V}\|_{\widehat{P}}^2 \le \mathrm{pen}(V) \le (1+r_2)\|\widehat{P}_{i|V} - P_{i|V}\|_{\widehat{P}}^2\big\},$$

let $\Omega^\delta_{\mathrm{comp}} = \Omega_3^\delta \cap \Omega_4^\delta \cap \Omega_{\mathrm{pen}}$, where Ω_3^δ and Ω_4^δ are respectively defined in (A.6) and (A.7). It comes from Lemma B.10 and our assumption on $\mathrm{pen}(V)$ that $P((\Omega^\delta_{\text{min pen}})^c) \le \epsilon + 2\delta^{-1}$.

Moreover, on $\Omega^{\delta}_{\text{min pen}}$, we have, from (A.9), for all $\eta > 0$,

$$(1-\eta)\|P_{i|S} - P_{i|\widehat{V}}\|_P^2 + r_1\|\widehat{P}_{i|\widehat{V}} - P_{i|\widehat{V}}\|_P^2 + (1+r_1)\big(\|\widehat{P}_{i|\widehat{V}} - P_{i|\widehat{V}}\|_{\widehat{P}}^2 - \|\widehat{P}_{i|\widehat{V}} - P_{i|\widehat{V}}\|_P^2\big)$$
$$\leq (1+\eta)\|P_{i|S} - P_{i|V}\|_P^2 + r_2\|\widehat{P}_{i|V} - P_{i|V}\|_P^2$$
$$+ (1+r_2)\big(\|\widehat{P}_{i|V} - P_{i|V}\|_{\widehat{P}}^2 - \|\widehat{P}_{i|V} - P_{i|V}\|_P^2\big) + \left(\frac{17}{\eta} + 1\right)\frac{\log(N_s^2\delta)}{3n}.$$

Let C be the constant given by Lemma B.5 and let

$$\Omega_* = \big\{\forall V \in \mathcal{V}_s, \big|\|\widehat{P}_{i|V} - P_{i|V}\|_{\widehat{P}}^2 - \|\widehat{P}_{i|V} - P_{i|V}\|_P^2\big| \leq C\varepsilon\|\widehat{P}_{i|V} - P_{i|V}\|_P^2\big\}.$$

It comes from Lemma B.5 that $P(\Omega_*) \geq 1 - \delta^{-1}$. Moreover, on $\Omega_{\text{comp}} \cap \Omega_*$, we have, from (A.9), for all $0 < \eta < 1$,

$$(1-\eta)\|P_{i|S} - P_{i|\widehat{V}}\|_P^2 + \big(r_1 - C(1+r_1)\varepsilon\big)\|\widehat{P}_{i|\widehat{V}} - P_{i|\widehat{V}}\|_P^2$$
$$\leq (1+\eta)\|P_{i|S} - P_{i|V}\|_P^2 + \big(r_2 + C(1+r_2)\varepsilon\big)\|\widehat{P}_{i|V} - P_{i|V}\|_P^2 + \frac{6}{\eta}\frac{\log(N_s^2\delta)}{n}.$$

Acknowledgements

We are grateful to Antonio Galves for many discussions and fruitful advices during the redaction of the paper. We also would like to thank the referees for the comments that considerably improved the manuscript.

ML was supported by FAPESP Grant 2009/09494-0. DYT was partially supported by FAPESP Grant 2008/08171-0, Pew Latin American Fellowship, and Ciência sem Fronteiras Fellowship (CNPq Grant 246778/2012-1). This work is part of USP project "Mathematics, computation, language and the brain".

Supplementary Material

Supplement to "Sharp oracle inequalities and slope heuristic for specification probabilities estimation in discrete random fields" (DOI: 10.3150/14-BEJ660SUPP; .pdf). On this supplementary material available on-line, we prove the probabilistic tools needed in the proofs of the main results. The second part provides additional simulation results. The last one is devoted to the extension of all our results to the Küllback loss.

References

[1] Arlot, S. and Bach, F. (2010). Data-driven calibration of linear estimators with minimal penalties. In *Advances in Neural Information Processing Systems* (*NIPS*) (Y. Bengio, D. Schuurmans, J.D. Lafferty, C.K.I. Williams and A. Culotta, eds.) **22** 46–54. Available at http://papers.nips.cc/book/advances-in-neural-information-processing-systems-22-2009.

[2] Arlot, S. and Massart, P. (2009). Data-driven calibration of penalties for least-squares regression. *J. Mach. Learn. Res.* **10** 245–279.
[3] Barron, A., Birgé, L. and Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields* **113** 301–413. MR1679028
[4] Barron, A.R. and Sheu, C.-H. (1991). Approximation of density functions by sequences of exponential families. *Ann. Statist.* **19** 1347–1369. MR1126328
[5] Bento, J. and Montanari, A. (2009). Which graphical models are difficult to learn? Available at http://arxiv.org/pdf/0910.5761.
[6] Birgé, L. and Massart, P. (1997). From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam* 55–87. New York: Springer. MR1462939
[7] Birgé, L. and Massart, P. (2001). Gaussian model selection. *J. Eur. Math. Soc.* (*JEMS*) **3** 203–268. MR1848946
[8] Birgé, L. and Massart, P. (2007). Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields* **138** 33–73. MR2288064
[9] 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
[10] Bresler, G., Mossel, E. and Sly, A. (2008). Reconstruction of Markov random fields from samples: Some observations and algorithms. In *Approximation, Randomization and Combinatorial Optimization. Lecture Notes in Computer Science* **5171** 343–356. Berlin: Springer. MR2538799
[11] 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.
[12] Csiszár, I. and Talata, Z. (2006). Consistent estimation of the basic neighborhood of Markov random fields. *Ann. Statist.* **34** 123–145. MR2275237
[13] Csiszár, I. and Talata, Z. (2006). Context tree estimation for not necessarily finite memory processes, via BIC and MDL. *IEEE Trans. Inform. Theory* **52** 1007–1016. MR2238067
[14] Galves, A., Orlandi, E. and Takahashi, D.Y. (2010). Identifying interacting pairs of sites in infinite range ising models. Preprint. Available at http://arxiv.org/abs/1006.0272.
[15] Georgii, H.-O. (1988). *Gibbs Measures and Phase Transitions. de Gruyter Studies in Mathematics* **9**. Berlin: de Gruyter. MR0956646
[16] Lerasle, M. (2011). Optimal model selection for density estimation of stationary data under various mixing conditions. *Ann. Statist.* **39** 1852–1877. MR2893855
[17] Lerasle, M. (2012). Optimal model selection in density estimation. *Ann. Inst. Henri Poincaré Probab. Stat.* **48** 884–908. MR2976568
[18] Lerasle, M. and Takahashi, D.Y. (2011). An oracle approach for interaction neighborhood estimation in random fields. *Electron. J. Stat.* **5** 534–571. MR2813554
[19] Lerasle M. and Takahashi D. Y. (2014). Supplement to "Sharp oracle inequalities and slope heuristic for specification probabilities estimation in discrete random fields." DOI:10.3150/14-BEJ660SUPP.
[20] Massart, P. (2007). *Concentration Inequalities and Model Selection. Lecture Notes in Math.* **1896**. Berlin: Springer. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003. With a foreword by Jean Picard. MR2319879
[21] Pastalkova, E., Buzsáki, G., Mizuseki, K. and Sirota, A. Theta oscillations provide temporal windows for local circuit computation in the entorhinal-hippocampal loop. *Neuron* **64** 267–280.
[22] 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. MR2662343
[23] Saumard, A. (2013). The slope heuristics in heteroscedastic regression. *Electron. J. Stat.* **7** 1184–1223. MR3056072
[24] 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.

[25] Takahashi, N., Sasaki, T., Matsumoto, W. and Ikegaya, Y. (2010). Circuit topology for synchronizing neurons in spontaneously active networks. *Proc. Natl. Acad. Sci. USA* **107** 10244–10249.

Received December 2011 and revised June 2014