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

Matthieu Lerasle, D.Y. Takahashi

Submitted to the Bernoulli

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

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

[1] *Univ. Nice Sophia Antipolis, CNRS, LJAD, UMR 7351,*
06100 Nice, France E-mail: mlerasle@unice.fr

[2] *Princeton University; Department of Psychology; Neuroscience Institute,*
Princeton, NJ 08648, United States. 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.

AMS 2000 subject classifications: Primary 62M40; secondary 62M45.
Keywords: Model Selection, Penalization, Slope heuristic, Discrete random fields.

1. Introduction

The main motivation for our work comes from neuroscience where the advancement of multichannel and optical technology enables researchers to record signals from tens to thousands of neurons simultaneously [TSMI10]. The question is then to understand the interactions between neurons in the brain and their relationships with the animal behavior [SBSB06, BKM04].
Following [SBSB06], 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,

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

The specification probabilities are important in the applications as they encode the conditional independence between the sites, see for example [BM09, BMS08, CT06a, GOT10, RWL10, LT11]. 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 $X_{1:n} = X_1, ..., X_n$ with common distribution P, the data set is given by $(X_i(j))_{i=1,...,n;\ j\in V_M}$. Following [BM97, BBM99, BM01], we use a penalized criterion to select a subset $\widehat{V} \subset V_M$ with cardinality $O(\log n)$ and show that 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 [BM09, BMS08, CT06a, GOT10, RWL10] for example. This approach requires in general strong assumptions on the random field, *e.g.*, it is assumed that the data is generated by an Ising model with restrictive conditions on the temperature parameter [BM09, GOT10, RWL10]. In particular, [BM09, BMS08, RWL10] assumed that the set S is finite and that all the sites are observed, *i.e* that $V_M = S$. When V_M does not contain $V_\star$, the meaning of the estimators in these papers is not clear. [CT06a] considered $S = \mathbb{Z}^d$ but assumed that $V_\star$ is finite. Finally, [GOT10, LT11] 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, e.g., 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 [BM07] for Gaussian model selection and has been theoretically studied only for very few specific models [BM07, AM09, Ler12, Ler11, AB10, Sau13]. 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. The methods of this article can be adapted to the Küllback loss; the interested reader can find these developments in Section D of the appendix.

2. Setting

Let (S, A, P) be a discrete random field, *i.e.* 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 $X_1, ..., X_n$ 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 $P(x(V)) = \int_{y \in \mathcal{X}(S); y(V) = x(V)} dP(y(S))$, $\widehat{P}(x(V)) = \sum_{y \in \mathcal{X}(S); y(V) = x(V)} \widehat{P}(y)$. 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

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

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

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

The observation set is $X_{1:n}(V_M) = (X_1(j), ..., X_n(j))_{j \in V_M}$. 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 $\left\| \widehat{P}_{i|V} - P_{i|S} \right\|_P^2$. We can decompose the risk via Pythogoras relation (see Proposition B.11)

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

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

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

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

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

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

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

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

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

for some absolute constant C. This control depends on the approximation properties of V through the bias $\left\| P_{i|V} - P_{i|S} \right\|_P^2$ 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*

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

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

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

Comments:

- The bound can be integrated and yields

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

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

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

The parameter $1/\beta$ is also called the temperature parameter in the physic literature where the model was initially introduced, see [Geo88]. The Ising model is the triplet (S, A, P), where $A = \{-1, 1\}$ and P is given by its specifications by

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

It follows from Theorem 4.5 in [LT11] that

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

Rates of convergence can be obtained from this bound and our model selection theorem. For example, let $d_\infty(i,j) = \max\{|i_k - j_k| : k \in \{1, \dots, d\}\}$, assume that $s \log M = O((\log n)^2)$ 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 $\sum_{j>k} \left|J^*_{i,j}\right| \leq e^{-r'k}$, 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 $\{j \in \mathbb{Z} : d_\infty(i,j) \leq n^{\alpha_i}\} \subset V_M$, we have

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

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 $\left((J^{(k)}_{i,i_1,\dots,i_k})_{(i,i_1,\dots,i_k)\in S^{k+1}}\right)_{k\geq 0} \in \prod_{k\geq 0} \mathbb{R}^{k+1}$ 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:

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

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.

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

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

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

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,

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

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

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

5. Slope heuristic

The slope heuristic was introduced in [BM07]. Let

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

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 $\Delta_V = \left\|\widehat{P}_{i|V} - P_{i|V}\right\|_{\widehat{P}}^2$ 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 $\mathbb{E}\left[\Delta_V\right] \le C|A|^v/n$. 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*

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

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

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

Comments:

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

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

  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 $\delta > 5$, $r_2 \ge r_1 > 0$, $\epsilon > 0$ and assume that*

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

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 $\eta>0$,

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

Comments:

- In this theorem, following [AM09], the main task is to show that

$$\Delta_V \simeq \left\|\widehat{P}_{i|V}-P_{i|V}\right\|_P^2. \quad (5.5)$$

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

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

  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 $\mathrm{pen}_{\mathrm{min}}(V)=\Delta_V$. 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

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

  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, [AM09] proposed the following algorithm.

1. For all $K>0$, denote by $\widehat{V}(K)$ the model selected with $\mathrm{pen}(V)=K\Theta_V$.

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 $\mathrm{pen}_{\min}(V) = K_0\Theta_V$ 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 $2K_{\min}\Theta_V \simeq 2\mathrm{pen}_{\min}(V)$ 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 $S = \{1, \cdots, 9\}$ 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 1A. 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 [LT11]. The performances of the estimators are measured by the logarithm of the ratio between the risk of the estimated model and the oracle risk. Figure 1B 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

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

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 1C. We show the results for $i = 9$. We compute the risk ratio as in the first experiment (Figure 1D). The simulations are done for increasing sample sizes $n = 10$, 25, 50, 75, 100, 150, 200, 300, 400, 500 (Figure 1D). 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).

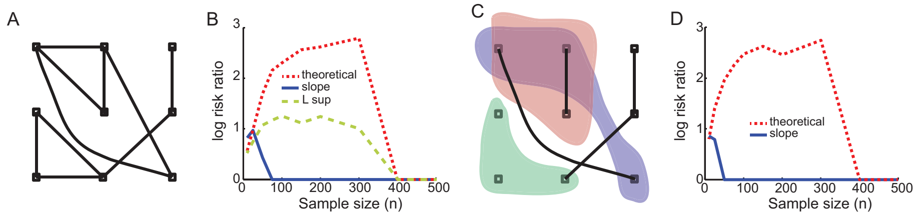

Figure 1. Simulation study. (A) Representation of the interacting pairs of the Ising model used in the first simulation experiment. The numbering of the sites increases from the top left to the bottom right.(B) Performance of the model selection for the first experiment. Plot of the log risk ratio for 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).

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 [SBSB06]. Modifications of the graph of interacting neurons for different tasks have been repeatedly shown [SBSB06]. 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 [KM].

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 con-

tains 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 20kHz. We binned the data with non-overlapping bins of size 10ms. 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 500ms, 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 ($n = 64099$, first 53.4min) and another sample for the second half of the experiment ($n = 64099$, second 53.4min).
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.

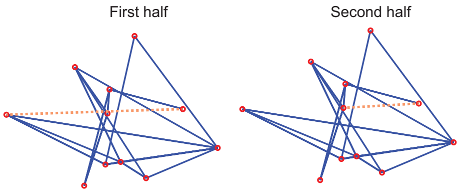

Figure 2. Representation of the interacting neuronal sites for the first half and second half of the experiment. The edges between sites indicate the interacting pairs. The dotted orange edges indicate the interactions that differed between both conditions. Observe that the interactions are represented 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 1D.

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 A: Supplementary Material**
(). 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

[AB10] S. Arlot and F. Bach. Data-driven calibration of linear estimators with minimal penalties. *Advances in Neural Information Processing Systems (NIPS)*, 22:46–54, 2010.

[AM09] S. Arlot and P. Massart. Data-driven calibration of penalties for least-squares regression. *Journal of Machine learning research*, 10:245–279, 2009.

[BBM99] A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413, 1999.

[BKM04] E.N. Brown, R.E. Kass, and P.P. Mitra. Multiple neural spike train data analysis: state-of-the-art and future challenges. *Nature Neuroscience*, 7:456–461, 2004.

[BM97] L. Birgé and P. Massart. From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam*, pages 55–87. Springer, New York, 1997.

[BM01] L. Birgé and P. Massart. Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3(3):203–268, 2001.

[BM07] L. Birgé and P. Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2):33–73, 2007.

[BM09] J. Bento and A. Montanari. Which graphical models are difficult to learn? *avalilable Markov random field texture modelsat http://arxiv.org/pdf/0910.5761*, 2009.

[BMS08] G. Bresler, E. Mossel, and A. Sly. *Approximation, Randomization and Combinatorial Optimization. Algorithms and Techniques*, chapter Reconstruction of Markov Random Fields from Samples: Some Easy Observations and Algorithms, pages 343–356. Springer, 2008.

[Bou02] O. Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Math. Acad. Sci. Paris*, 334(6):495–500, 2002.

[BS91] A. R. Barron and C.-H. Sheu. Approximation of density functions by sequences of exponential families. *Ann. Statist.*, 19(3):1347–1369, 1991.

[CT06a] I. Csiszar and Z. Talata. Consistent estimation of the basic neighborhood of markov random fields. *Ann. Statist.*, 34:123–145, 2006.

[CT06b] I. Csiszár and Z. Talata. Context tree estimation for not necessarily finite memory processes, via BIC and MDL. *IEEE Trans. Inform. Theory*, 52(3):1007–1016, 2006.

[Geo88] HO Georgii. *Gibbs measure and phase transitions*, volume 9 of *de Gruyter studies in mathematics*. de Gruyter, Berlin, 1988.

[GOT10] A. Galves, E. Orlandi, and D. Y. Takahashi. Identifying interacting pairs of sites in infinite range ising models. *Preprint, http://arxiv.org/abs/1006.0272*, 2010.

[KM] E Pastalkova G Buzsáki K Mizuseki, A Sirota. Theta oscillations provide temporal windows for local circuit computation in the entorhinal-hippocampal loop. *Neuron*, 64:267–280.

[Ler11] Matthieu Lerasle. Optimal model selection for stationary data under various mixing conditions. *Ann. Statist.*, 39(4):1852–1877, 2011.

[Ler12] Matthieu Lerasle. Optimal model selection in density estimation. *Ann. Inst. H. Poincaré Probab. Statist.*, 48(3):884–908, 2012.

[LT11] M. Lerasle and D.Yasumasa Takahashi. An oracle approach for interaction neighborhood estimation in random fields. *to appear in Electron. J. Statist. arXiv:1010.4783*, 2011.

[Mas07] P. Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

[RWL10] P. Ravikumar, M.J. Wainwright, and J. D. Lafferty. High-dimensional ising model selection using l_1-regularized logistic regression. *Ann. Statist.*, 38(3):1287–1319, 2010.

[Sau13] A. Saumard. The slope heuristics in heteroscedastic regression. *Electronic Journal of Statistics*, 2013.

[SBSB06] E. Schneidman, M.J. Berry, R. Segev, and W. Bialek. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, 440:1007–1012, 2006.

[TSMI10] N. Takahashi, T. Sasaki, W. Matsumoto, and Y. Ikegaya. Circuit topology for synchronizing neurons in spontaneously active networks. *Proceedings of National Academy of Science U.S.A.*, 107:10244–10249, 2010.

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

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

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

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

Hereafter in the proof of Theorem 3.1, we denote by

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

It comes from Lemma B.1 that

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

From Lemma B.1, we also have

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

Hence

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

Thus,

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

is smaller than

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

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 $1 - 2e^{-x}$,

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

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

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

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

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

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*

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

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

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

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

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

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

$$\frac{1}{|A|}\sum_{x\in\mathcal{X}(V\cup V')}\widehat{P}(x(V\cup V'))P_{i|V}(x) = \sum_{x\in\mathcal{X}(V)}\frac{\widehat{P}(x(V/\{i\}))}{|A|}\widehat{P}_{i|V}(x)P_{i|V}(x) = \left(\widehat{P}_{i|V}, P_{i|V}\right)_{\widehat{P}}.$$

$$\frac{1}{|A|}\sum_{x\in\mathcal{X}(V\cup V')}P(x(V\cup V'))P_{i|V}(x) = \sum_{x\in\mathcal{X}(V)}\frac{P(x(V/\{i\}))}{|A|}P_{i|V}^2(x) = \left\|P_{i|V}\right\|_P^2$$

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

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

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

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

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

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

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

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

is smaller than

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

We have also,

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

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

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

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

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

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

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

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

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

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

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

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

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 $\delta \ge 4$,

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

□

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 $\log_2(n) \geq |\mathcal{N}_i|$,
$$\mathbb{E}\left[\left\|P_{i|S} - \widehat{P}_{i|\hat{V}}\right\|_P^2\right] \leq C\frac{(\log(n)\log(M))^2}{n} + C_\beta\frac{2^{|\mathcal{N}_i|}}{n} \leq C_{\beta,|\mathcal{N}_i|}\frac{(\log(n)\log(M))^2}{n} \quad .$$
2. Assume that there exist constants r and r' such that $M = n^r$ and, for any $k \in \mathbb{N}$, $\sum_{j>k}\left|J^*_{i,j}\right| \leq e^{-r'k}$, then
$$\mathbb{E}\left[\left\|P_{i|S} - \widehat{P}_{i|\hat{V}}\right\|_P^2\right] \leq Cr^2\frac{\log(n)^4}{n} + C_\beta\left(\left(\sum_{j\notin V_M}|J_{i,j}|\right)^2 + n^{-\frac{2r'}{2r'+\log 2}}\right)$$
$$\leq C_{r,\beta}\left(\left(\sum_{j\notin V_M}|J_{i,j}|\right)^2 + n^{-\frac{2r'}{2r'+\log 2}}\right) \quad .$$

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

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

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

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

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

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

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

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

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

Therefore, the bias term is upper bounded by

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

Moreover, V has cardinality upper bounded by

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

A.4. Proof of Theorem 5.1:

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

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

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

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

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

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

Let $\Omega_{\mathrm{pen}} = \left\{ 0 \leq \mathrm{pen}(V) \leq (1-r) \left\| \widehat{P}_{i|V} - P_{i|V} \right\|_{\widehat{P}}^2 \right\}$ and let $\Omega^\delta_{\min \mathrm{pen}} = \Omega^\delta_3 \cap \Omega^\delta_4 \cap \Omega_{\mathrm{pen}}$, where Ω^δ_3 and Ω^δ_4 are respectively defined in (A.6) and (A.7). It comes from Lemma B.10 and our assumption on pen(V) that $P((\Omega^\delta_{\min \mathrm{pen}})^c) \leq \epsilon + 2\delta^{-1}$. Moreover, on $\Omega^\delta_{\min \mathrm{pen}}$, we have, for all $\eta > 0$,

$$|L(\widehat{V}) - L(V)| \leq \eta \left\| P_{i|S} - P_{i|\widehat{V}} \right\|_P^2 + \eta \left\| P_{i|S} - P_{i|V} \right\|_P^2 + \left(\frac{16}{\eta} + 1 \right) \frac{\log(N_s^2 \delta)}{3n}.$$

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

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

A.5. Proof of Theorem 5.2:

Let

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

let $\Omega_{\mathrm{comp}}^\delta=\Omega_3^\delta\cap\Omega_4^\delta\cap\Omega_{\mathrm{pen}}$, 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 $P((\Omega_{\min\mathrm{pen}}^\delta)^c)\leq\epsilon+2\delta^{-1}$. Moreover, on $\Omega_{\min\mathrm{pen}}^\delta$, we have, from (A.9), for all $\eta>0$,

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

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

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

It comes from Lemma B.5 that $P(\Omega_*)\geq1-\delta^{-1}$. Moreover, on $\Omega_{\mathrm{comp}}\cap\Omega_*$, we have, from (A.9), for all $0<\eta<1$,

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

Appendix B: Probabilistic Tools

Lemma B.1. *Let x in $\mathcal{X}(S)$, let V be a finite subset of S and let Q,R be two probability measures on $\mathcal{X}(V)$ such that $R(x(V/\{i\}))>0$. We have*

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

The lemma immediately follows from the fact that $Q_{i|V}(x)Q(x(V/\{i\})) = Q(x(V))$ and $R_{i|V}(x) = R(x(V))/R(x(V/\{i\}))$.
We recall the bound given by Bousquet [Bou02] for the deviation of the supremum of the empirical process.

Theorem B.2. *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 $x > 0$,*

$$P\left(Z > \mathbb{E}(Z) + \sqrt{\frac{2}{n}(v^2 + 2b\mathbb{E}(Z))x} + \frac{bx}{3n}\right) \leq e^{-x}. \tag{B.1}$$

Bousquet's result is a generalization of the elementary Benett's inequality.

Theorem B.3. *Let $X_1, ..., X_n$ be i.i.d. random variables, real valued and bounded by b. Let $v^2 = Var(X_1)$ and $\bar{X}_n = n^{-1}\sum_{i=1}^n (X_i - \mathbb{E}(X))$. Then, for all $x > 0$,*

$$P\left(\bar{X}_n > \sqrt{\frac{2v^2x}{n}} + \frac{bx}{3n}\right) \leq e^{-x}. \tag{B.2}$$

B.1. Concentration for Slope with quadratic risk

The aim of this section is to prove the following result.

Theorem B.4. *Let (S, A, P) be a random field and let V be a subspace in $\mathcal{V}_s$. Let $\mathcal{X}'(V) = \{x \in \mathcal{X}(V),\ P(x(V)) \neq 0\}$ and let $p_-^V = \inf_{x\in\mathcal{X}'(V)} P(x(V))$.*
Let $Z = \sup_{x\in\mathcal{X}'(V)} \frac{\left|\widehat{P}(x(V)) - P(x(V))\right|}{P(x(V))}$. For all $\delta > 1$, with probability larger than $1 - \delta^{-1}$,

$$Z \leq \frac{64\sqrt{2}}{\sqrt{np_-^V}}\sqrt{\log\left(\frac{16}{p_-^V}\right)} + \frac{2048}{np_-^V}\log\left(\frac{16}{p_-^V}\right) + \sqrt{\frac{2\log(\delta)}{np_-^V}} + 2\frac{\log(\delta)}{np_-^V}.$$

Let us state an important consequence of Theorem B.4.

Lemma B.5. *Assume that $\inf_{V\in\mathcal{V}_s} p_-^V \geq \varepsilon^{-2}n^{-1}\log(nN_s\delta)$. There exists an absolute constant C such that, with probability larger than $1 - \delta^{-1}$, for all V in $\mathcal{V}_s$,*

$$\left|\left\|\widehat{P}_{i|V} - P_{i|V}\right\|_P^2 - \left\|\widehat{P}_{i|V} - P_{i|V}\right\|_{\widehat{P}}^2\right| \leq C\varepsilon\left\|\widehat{P}_{i|V} - P_{i|V}\right\|_P^2.$$

Proof: Let V in $\mathcal{V}_s$ and let $\mathcal{X}'(V)=\{x\in\mathcal{X}(V),\ P(x(V/\{i\}))\neq 0\}$. We have

$$\left|\left\|\widehat{P}_{i|V}-P_{i|V}\right\|_{\widehat{P}}^2-\left\|\widehat{P}_{i|V}-P_{i|V}\right\|_P^2\right|$$
$$\leq \sum_{x\in\mathcal{X}'(V)}\frac{|\widehat{P}(x(V/\{i\}))-P(x(V/\{i\}))|}{|A|}\left(\widehat{P}_{i|V}(x)-P_{i|V}(x)\right)^2$$
$$\leq \sup_{x\in\mathcal{X}'(V)}\frac{|\widehat{P}(x(V/\{i\}))-P(x(V/\{i\}))|}{P(x(V/\{i\}))}\left\|\widehat{P}_{i|V}-P_{i|V}\right\|_P^2 .$$

We take a union bound in Theorem B.4 and we obtain, since $\inf_{V\in\mathcal{V}_s}p_-^V\geq\varepsilon^{-2}n^{-1}\log(nN_s\delta)$ that there exists an absolute constant C such that

$$\forall V\in\mathcal{V}_s,\ \sup_{x\in\mathcal{X}'(V)}\frac{|\widehat{P}(x(V/\{i\}))-P(x(V/\{i\}))|}{P(x(V/\{i\}))}\leq C\varepsilon .$$

In the remainder of this section, we state the results necessary to prove Theorem B.4.

Proposition B.6. *Let P be a probability measure on $\mathcal{X}(S)$ and let V be a finite subset of S. Let $\mathcal{X}'(V)=\{x\in\mathcal{X}(V),\ P(x(V))\neq 0\}$ and let $p_-^V=\inf_{x\in\mathcal{X}'(V)}P(x(V))$. Let $Z=\sup_{x\in\mathcal{X}'(V)}\frac{\left|\widehat{P}(x(V))-P(x(V))\right|}{P(x(V))}$. For all $\delta>0$, with probability larger than $1-\delta^{-1}$,*

$$Z\leq 2\mathbb{E}\left(Z\right)+\sqrt{\frac{2\log(\delta)}{np_-^V}}+2\frac{\log(\delta)}{np_-^V}.$$

Proposition B.6 is a straightforward consequence of Bousquet's version of Talagrand's inequality, that we apply to the class of functions $\mathcal{F}=\{(P(x(V)))^{-1}1_{x(V)}\}$.
The second proposition let us compute this expectation.

Proposition B.7. *Let P be a probability measure on $\mathcal{X}(S)$ and let V be a finite subset of S. Let $\mathcal{X}'(V)=\{x\in\mathcal{X}(V),\ P(x(V))\neq 0\}$ and let $p_-^V=\inf_{x\in\mathcal{X}'(V)}P(x(V))$.*

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

Proposition B.7 was proved in [LT11].

B.2. Concentration of the variance term in quadratic risk

The aim of this section is to prove the following concentration result, that is at the center of the main proofs.

Theorem B.8. *Let V be a finite subset of S. Let $b>0$ and let $\mathcal{X}^b(V)=\left\{x\in\mathcal{X}(V),\ P(x(V))\geq b^{-2}\right\}$. For all $x>0$, $\eta>0$, we have,*

$$P\left(\sum_{x\in\mathcal{X}^b(V)}\frac{\left(\widehat{P}(x(V))-P(x(V))\right)^2}{P(x(V))}\leq(1+\eta)^3\frac{|A|^v}{n}+\frac{4x}{\eta n}+\frac{32b^2x^2}{\eta^3n^2}\right)\geq 1-e^{-x}.$$

Proof: Let us first recall the following consequence of Cauchy-Schwarz inequality.

Lemma B.9. *Let I be a finite set and let $(b_i)_{i\in I}$ be a collection of real numbers. We have*

$$\sum_{i\in I}b_i^2=\left(\sup_{(a_i)_{i\in I},\ \sum_{i\in I}a_i^2\leq 1}\sum_{i\in I}a_ib_i\right)^2.$$

Proof: The lemma is obviously satisfied if all the $b_i=0$. Assume now that it is not the case. By Cauchy Schwarz inequality, we have, for all collection $(a_i)_{i\in I}$ such that $\sum_{i\in I}a_i^2\leq 1$,

$$\left(\sum_{i\in I}a_ib_i\right)^2\leq\sum_{i\in I}a_i^2\sum_{i\in I}b_i^2\leq\sum_{i\in I}b_i^2.$$

Moreover, consider for all i in I, $a_i=b_i/\sqrt{\sum_{i\in I}b_i^2}$, we have $\sum_{i\in I}a_i^2=1$ and $\sum_{i\in I}a_ib_i=\sqrt{\sum_{i\in I}b_i^2}$, which concludes the proof.
Let us now introduce the following set.

$$B_V^b=\left\{f:\mathcal{X}^b(V)\rightarrow\mathbb{R}\text{ such that }f=\sum_{x\in\mathcal{X}^b(V)}\frac{\alpha_x\mathbf{1}_{\{x\}}}{\sqrt{P(x(V))}},\text{ where }\sum_{x\in\mathcal{X}^b(V)}\alpha_x^2\leq 1.\right\}.$$

Let P and P_n be the following operators, defined for all functions f, by $P_nf=\frac{1}{n}\sum_{i=1}^nf(X_i)$ and, for all functions f in $L^1(P)$, by $Pf=\int f(x)dP(x)$. Using Lemma B.9 with $I=\mathcal{X}^b(V)$ and

$$b_x=\frac{\widehat{P}(x(V))-P(x(V))}{\sqrt{P(x(V))}}=(P_n-P)\left(\frac{1_{x(V)}}{\sqrt{P(x(V))}}\right),$$

we obtain,

$$\sum_{x\in\mathcal{X}^b(V)}\frac{\left(\widehat{P}(x(V))-P(x(V))\right)^2}{P(x(V))}=\left(\sup_{f\in B_V^b}(P_n-P)f\right)^2,$$

The functions f in B_V^p satisfy

$$\mathrm{Var}(f(X))\leq Pf^2=\sum_{x\in\mathcal{X}^b(V)}\frac{\alpha_x^2}{P(x(V))}P(x(V))\leq 1,\ \|f\|_\infty\leq\sup_{x\in\mathcal{X}^b(V)}\frac{1}{\sqrt{P(x(V))}}\leq b.$$

From Theorem B.2, we have then, for all $\eta > 0$, for all $x > 0$,

$$P\left(\sup_{f\in B_V^b}(P_n-P)f > (1+\eta)\mathbb{E}\left(\sup_{f\in B_V^b}(P_n-P)f\right)+\sqrt{\frac{2x}{n}}+\left(\frac{1}{3}+\frac{1}{\eta}\right)\frac{bx}{n}\right)\leq e^{-x}.$$

From Cauchy-Schwarz inequality, we have then

$$\begin{aligned}\mathbb{E}\left(\sup_{f\in B_V^b}(P_n-P)f\right)&\leq\sqrt{\mathbb{E}\left(\left(\sup_{f\in B_V^b}(P_n-P)f\right)^2\right)}\\&=\sqrt{\sum_{x\in\mathcal{X}(V),\ P(x(V))\neq 0}\frac{\mathbb{E}\left(\left(\widehat{P}(x(V))-P(x(V))\right)^2\right)}{P(x(V/\{i\}))}}\\&=\sqrt{\sum_{x\in\mathcal{X}(V),\ P(x(V))\neq 0}\frac{\mathrm{Var}(1_{X(V)=x(V)})}{nP(x(V/\{i\}))}}\leq\sqrt{\frac{|A|^v}{n}}.\end{aligned}$$

We have obtain that

$$P\left(\sup_{f\in B_V^b}(P_n-P)f>(1+\eta)\sqrt{\frac{|A|^v}{n}}+\sqrt{\frac{2x}{n}}+\left(\frac{1}{3}+\frac{1}{\eta}\right)\frac{bx}{n}\right)\leq e^{-x}.$$

Since B_V^b is symmetric, $\sup_{f\in B_V^b}(P_n-P)f\geq 0$. We can therefore take the square in the previous inequality to conclude the proof of the Theorem.

B.3. Concentration of the remainder term in the quadratic case

Let us now give some important concentration inequalities.

Lemma B.10. *Let V, V' be two subsets in $\mathcal{V}_s$. For all $\delta > 0$, we have, with probability larger than $1-\delta$,*

$$\frac{1}{|A|}\sum_{x\in\mathcal{X}((V\cup V'))}(\widehat{P}(x((V\cup V')/\{i\}))-P(x((V\cup V')/\{i\})))\left(P^2_{i|V}(x)-P^2_{i|V'}(x)\right)$$
$$\leq 2\left\|P_{i|V}-P_{i|V'}\right\|_P\sqrt{2\frac{\log(\delta)}{n}}+\frac{\log(\delta)}{3n}. \quad (B.3)$$

$$\frac{1}{|A|}\sum_{x\in\mathcal{X}(V\cup V')}(\widehat{P}(x(V\cup V'))-P(x(V\cup V'))\left(P_{i|V'}(x)-P_{i|V}(x)\right)$$
$$\leq\left\|P_{i|V}-P_{i|V'}\right\|_P\sqrt{2\frac{\log(\delta)}{n}}+\frac{\log(\delta)}{3n}. \quad (B.4)$$

Proof of Lemma B.10: Let f_1 be the real valued function defined on $\mathcal{X}((V\cup V')/\{i\})$ by

$$f_1 = \sum_{x\in\mathcal{X}((V\cup V'))}\left(P^2_{i|V}(x)-P^2_{i|V'}(x)\right)1_{x((V\cup V')/\{i\}))}.$$

$$f_2 = \sum_{x\in\mathcal{X}((V\cup V'))}\left(P_{i|V}(x)-P_{i|V'}(x)\right)1_{x(V\cup V')}.$$

We have

$$f_1 = \sum_{x\in\mathcal{X}((V\cup V')/\{i\})}1_{x((V\cup V')/\{i\}))}\sum_{b\in A}\left(P^2_{i|V}(x_b)-P^2_{i|V'}(x_b)\right).$$

f_1 is upper bounded by $\max_{x\in\mathcal{X}((V\cup V')/\{i\})}\sum_{b\in A}\left(P^2_{i|V}(x_b)-P^2_{i|V'}(x_b)\right)\leq|A|$. f_2 is upper bounded by $\max_{x\in\mathcal{X}(V\cup V')}\left|P_{i|V}(x)-P_{i|V'}(x)\right|$. Since, for all $x\neq x'$ in $\mathcal{X}((V\cup V')/\{i\})$, $1_{x((V\cup V')/\{i\})}1_{x'((V\cup V')/\{i\})}=0$, we have

$$\begin{aligned}\mathrm{Var}(f_1(X)) &= \sum_{x\in\mathcal{X}((V\cup V')/\{i\})}P(x((V\cup V')/\{i\})))\left(\sum_{b\in A}P^2_{i|V}(x_b)-P^2_{i|V'}(x_b)\right)^2\\ &\leq |A|\sum_{x\in\mathcal{X}((V\cup V')/\{i\})}P(x((V\cup V')/\{i\})))\sum_{b\in A}\left(P^2_{i|V}(x_b)-P^2_{i|V'}(x_b)\right)^2\\ &\leq 4|A|\sum_{x\in\mathcal{X}(V\cup V')}\left(P_{i|V}(x)-P_{i|V'}(x)\right)^2P(x(V\cup V'/\{i\}))\\ &= 4|A|^2\left\|P_{i|V}-P_{i|V'}\right\|^2_P.\end{aligned}$$

Since, for all $x\neq x'$ in $\mathcal{X}(V\cup V')$, $1_{x(V\cup V')}1_{x'(V\cup V')}=0$, we have

$$\mathrm{Var}(f_2(X))\leq\sum_{x\in\mathcal{X}((V\cup V'))}\left(P_{i|V}(x)-P_{i|V'}(x)\right)^2P(V\cup V')\leq|A|\left\|P_{i|V}-P_{i|V'}\right\|^2_P.$$

Inequality B.3 is therefore a consequence of Benett's inequality, see Theorem B.3. We obtain Inequality B.4 exactly with the same arguments.□

Pythagoras relation

Let us give here Pythagoras relation that we used several times.

Proposition B.11. *Let (S,A,P) be a random field, let i in S and let V be a subset of S and let f be a function defined on $\mathcal{X}(V)$. Then, the following relations hold*

$$\begin{aligned}\int_{x\in\mathcal{X}(S)}f(x(V))P_{i|S}(x)dP(x(S/\{i\})) &= \int_{x\in\mathcal{X}(V)}f(x(V))P_{i|V}(x)dP(x(V/\{i\}))\\ &= \int_{x\in\mathcal{X}(S)}f(x(V))P_{i|V}(x)dP(x(S/\{i\})).\end{aligned}$$

In particular, we have

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

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

Proof: The first inequality comes from the following computations. For all x in $\mathcal{X}(V)$ and y in $\mathcal{X}(S/V)$, let $x(V)\oplus y(S/V)$ be the configuration on $\mathcal{X}(S)$ such that $(x(V)\oplus y(S/V))(j)=x(j)$ for all j in V and $(x(V)\oplus y(S/V))(j)=y(j)$ for all j in S/V. By definition of the conditional probabilities $P_{i|V}(x)$, we have

$$\begin{aligned}&\int_{x\in\mathcal{X}(S)}f(x(V))P_{i|S}(x)dP(x(S/\{i\}))\\&=\int_{x\in\mathcal{X}(V)}f(x(V))dP(x(V/\{i\}))\int_{y\in\mathcal{X}(S/V)}P_{i|S}(x(V)\oplus y(S/V))dP(y(S/V)|x(V))\\&=\int_{x\in\mathcal{X}(V)}f(x(V))P_{i|V}(x)dP(x(V/\{i\})).\end{aligned}$$

The second inequality is a straightforward consequence of the first one. For the third one, we apply the second inequality to $f(x(V))=\widehat{P}_{i|V}-P_{i|V}$, we have

$$\int_{x\in\mathcal{X}(S)}f(x(V))P_{i|S}(x)dP(x(S/\{i\}))=\int_{x\in\mathcal{X}(S)}f(x(V))P_{i|V}(x)dP(x(S/\{i\})).$$

Thus,

$$\begin{aligned}&\left\|\widehat{P}_{i|V}\ -P_{i|S}\right\|_P^2=\left\|f(x(V))+P_{i|V}-P_{i|S}\right\|_P^2=\left\|f(x(V))\right\|_P^2+\left\|P_{i|V}-P_{i|S}\right\|_P^2+\\&\quad\frac{2}{|A|}\left(\int_{x\in\mathcal{X}(S)}f(x(V))P_{i|V}(x)dP(x(S/\{i\}))-\int_{x\in\mathcal{X}(S)}f(x(V))P_{i|S}(x)dP(x(S/\{i\}))\right)\\&\quad=\left\|\widehat{P}_{i|V}-P_{i|V}\right\|_P^2+\left\|P_{i|V}-P_{i|S}\right\|_P^2 .\end{aligned}$$

For the last inequality, we use the second one with $f(x(V))=P_{i|V}(x)$, we have

$$\begin{aligned}\left\|P_{i|V}-P_{i|S}\right\|_P^2&=\left\|P_{i|V}\right\|_P^2+\left\|P_{i|S}\right\|_P^2-\frac{2}{|A|}\int_{x\in\mathcal{X}(S)}f(x(V))P_{i|S}(x)dP(x(S/\{i\}))\\&=\left\|P_{i|V}\right\|_P^2+\left\|P_{i|S}\right\|_P^2-\frac{2}{|A|}\int_{x\in\mathcal{X}(S)}f(x(V))P_{i|V}(x)dP(x(S/\{i\}))\\&=\left\|P_{i|V}\right\|_P^2+\left\|P_{i|S}\right\|_P^2-2\left\|P_{i|V}\right\|_P^2=\left\|P_{i|S}\right\|_P^2-\left\|P_{i|V}\right\|_P^2 .\end{aligned}$$

Appendix C: An example of oracle and slope heuristic

In this section we give an example of the interacting sites chosen by the oracle for the Ising model used in the first simulation study considered in Section 6. This allows the reader to compare the generating model (Figure 3A) and the models chosen by the oracle (Figure 3B) and by the slope heuristic (Figure 3C) when $n = 25$. Observe that even with this small n, the oracle and the slope heuristic seems to be able to identify some of the main interactions. We illustrate in Figure 3D how the complexity of $\widehat{V}_i, i = 9$, changes when increasing constant K in the slope heuristic described in Section 5. The black point indicates the constant $K_{\min}$ corresponding to the minimal penalty $\text{pen}_{\min}$ and the red cross indicates the constant $2K_{\min}$ which specifies $\widehat{V}(2K_{\min})$.

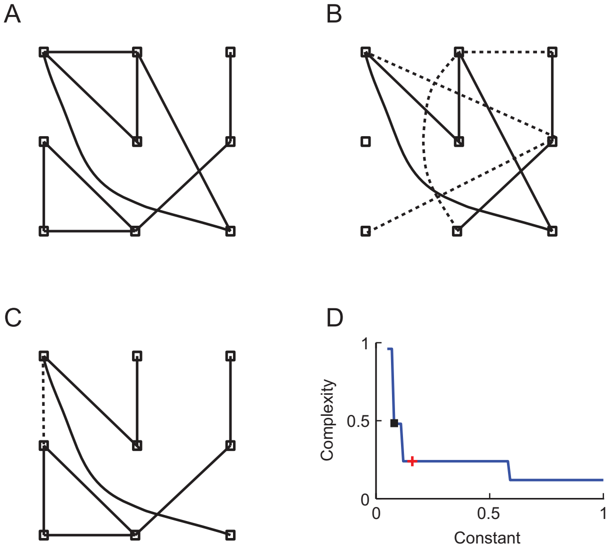

Figure 3. Comparison of the chosen models. (A) Representation of the interacting pairs of the Ising model used in the first simulation experiment. The numbering of the sites increase from the top left to the bottom right. This figure is the same of Figure fig:NewSimulationRiskRatioA. (B) Representation of the interacting pairs for the oracle. The solid and dashed lines indicate, respectively, the interactions that exist and doesn't exist in the generating model. (C) Representation of the interacting pairs for the model chosen by the slope heuristic. The legend is the same as in (B). (D) Graph showing the change in complexity when the constant K is increased. The black point indicates $K_{\min}$ and the red cross indicates $2K_{\min}$.

Appendix D: Results for Küllback loss

The purpose of this section is to show that the methods developed in the paper can be adapted to study the oracle approach with Küllback loss. This risk function is natural in information theory and therefore interesting in our applications in neuroscience. We postponed this section in appendix in order to avoid redundancy in the paper. We keep the notation of the paper except that we denote $a = |A|$.

The logarithmic loss of a non-negative function f is defined on $\mathcal{X}(S)$ by

$$L_Q(f) = \int \log\left(\frac{1}{f(x)}\right) dQ(x).$$

The Küllback loss of the estimator $\widehat{P}_{i|V}$ is then defined by

$$K(P_{i|S}, \widehat{P}_{i|V}) = L_P(\widehat{P}_{i|V}) - L_P(P_{i|S}).$$

The Küllback risk is decomposed in a variance term and a bias term thanks to the relation

$$\begin{aligned} K(P_{i|S}, \widehat{P}_{i|V}) &= \Big(L_P(\widehat{P}_{i|V}) - L_P(P_{i|V})\Big) + \big(L_P(P_{i|V}) - L_P(P_{i|S})\big) \\ &= \sum_{x\in\mathcal{X}(V)} P(x(V)) \log\left(\frac{P_{i|V}(x)}{\widehat{P}_{i|V}(x)}\right) + \int dP(x(S)) \log\left(\frac{P_{i|S}(x)}{P_{i|V}(x)}\right) \\ &= K(P_{i|V}, \widehat{P}_{i|V}) + K(P_{i|S}, P_{i|V}). \end{aligned}$$

Let $\Lambda \geq 100$, $\delta > 1$ and let

$$\mathcal{V}_{s,\Lambda} = \left\{V \in \mathcal{V}_s,\ \forall x \in \mathcal{X}(S), P(x(V)) = 0 \text{ or } \widehat{P}(x(V)) \geq \Lambda\frac{\log(2a^s N_s\delta)}{n}\right\}. \tag{D.1}$$

$$\mathcal{V}^{(2)}_{s,\Lambda} = \left\{V \in \mathcal{V}_s,\ \forall x \in \mathcal{X}(S), P(x(V)) = 0 \text{ or } P(x(V)) \geq \Lambda\frac{\log(2a^s N_s\delta)}{n}\right\}. \tag{D.2}$$

Let $p_* \geq 0$, and let

$$\mathcal{V}_{s,\Lambda,p_*} = \left\{V \in \mathcal{V}_{s,\Lambda},\ \forall x \in \mathcal{X}(S),\ P_{i|V}(x) = 0 \text{ or } \widehat{P}_{i|V}(x) \geq p_*\right\}. \tag{D.3}$$

$$\mathcal{V}^{(2)}_{s,\Lambda,p_*} = \left\{V \in \mathcal{V}^{(2)}_{s,\Lambda},\ \forall x \in \mathcal{X}(S),\ P_{i|V}(x) = 0 \text{ or } P_{i|V}(x) \geq p_*\right\}. \tag{D.4}$$

The idea of the sets $\mathcal{V}_{s,\Lambda,p_*}$ is that we restrict the collections of sets V to those where the possible configurations are sufficiently observed. The main advantage of the sets $\mathcal{V}_{s,\Lambda,p_*}$ is that the conditions can be verified in practice. In order to illustrate why we introduced $\mathcal{V}^{(2)}_{s,\Lambda,p_*}$, let us give the following weak non-nullness assumption.

NN *There exists $p_\star > 0$ such that, for all finite subsets V, for all sites i, and for all x in $\mathcal{X}(V)$,*

$$P(x(V)) = 0 \text{ or } P_{i|V}(x) \geq p_\star.$$

We have (see [Mas07] Proposition 2.5 p20)

$$N_s = \sum_{k=0}^{s} C_M^k \leq \left(\frac{eM}{s}\right)^s \leq M^s.$$

Hence, $\log(a^s N_s \delta) \leq s(\log(aM)) + \log\delta$. We have

$$\frac{nP(x(V))}{\Lambda\log(2a^s N_s\delta)} \geq \frac{e^{\log n - s\log(p_\star^{-1})}}{\Lambda(s\log(aM) + \log(\delta))} \to +\infty,$$

if $(\log n)^{-1}s < s_\star = (\log p_\star^{-1})^{-1}$ and $\Lambda\log(M\delta) = O(n^\alpha)$, where $\alpha \leq \alpha_\star = 1 - s_\star$. In that case, for all $n \geq n(p_\star)$, $\mathcal{V}_s = \mathcal{V}^{(2)}_{s,\Lambda} = \mathcal{V}^{(2)}_{s,\Lambda,p_\star}$.

D.1. Oracle properties in Küllback Loss

Our first result is a sharp control of the variance term of the Küllback risk.

Theorem D.1. *Let (S, A, P) be a random field, let $\Lambda \geq 100$, $\delta > 1$, $s > 0$. Let $\mathcal{V}_{s,\Lambda}$ be the collection defined in (D.1). Then, with probability larger than $1-\delta^{-1}$, for all V in $\mathcal{V}_{s,\Lambda}$, for all $\eta > 0$, we have*

$$\sum_{x\in\mathcal{X}(V)} P(x(V))\log\left(\frac{P_{i|V}(x)}{\widehat{P}_{i|V}(x)}\right) \leq 5\left((1+\eta)^3\frac{a^v}{n} + \left(1+\frac{64}{\eta^2\Lambda}\right)\frac{4\log(2N_s\delta)}{\eta n}\right).$$

$$\sum_{x\in\mathcal{X}(V)} \widehat{P}(x(V))\log\left(\frac{\widehat{P}_{i|V}(x)}{P_{i|V}(x)}\right) \leq 4\left((1+\eta)^3\frac{a^v}{n} + \left(1+\frac{64}{\eta^2\Lambda}\right)\frac{4\log(2N_s\delta)}{\eta n}\right).$$

Let $\mathcal{V}^{(2)}_{s,\Lambda}$ be the collection defined in (D.2). Then, with probability larger than $1-\delta^{-1}$, for all V in $\mathcal{V}^{(2)}_{s,\Lambda}$, for all $\eta > 0$, we have

$$\sum_{x\in\mathcal{X}(V)} P(x(V))\log\left(\frac{P_{i|V}(x)}{\widehat{P}_{i|V}(x)}\right) \leq 5\left((1+\eta)^3\frac{a^v}{n} + \left(1+\frac{64}{\eta^2\Lambda}\right)\frac{4\log(2N_s\delta)}{\eta n}\right).$$

$$\sum_{x\in\mathcal{X}(V)} \widehat{P}(x(V))\log\left(\frac{\widehat{P}_{i|V}(x)}{P_{i|V}(x)}\right) \leq 4\left((1+\eta)^3\frac{a^v}{n} + \left(1+\frac{64}{\eta^2\Lambda}\right)\frac{4\log(2N_s\delta)}{\eta n}\right).$$

Comments:

- The variance part of the Küllback risk is controlled as the variance part of the L_2 risk. We only have to restrict the study to the subset $\mathcal{V}_{s,\Lambda}$ of $\mathcal{V}_s$ where all the possible configurations are sufficiently observed. This restriction is not important when $s << n$, and our result holds also without restriction on the random field.

As in the previous section, we want to optimize the bound on the Küllback loss given by Theorem D.1 among $\mathcal{V}_{s,\Lambda}$. We introduce for this purpose the following penalized estimators.

$$\widehat{V} = \arg\min_{V\in\mathcal{V}_{s,\Lambda,p_*}} \left\{ -\sum_{x\in\mathcal{X}(V)} \widehat{P}(x(V))\log\left(\widehat{P}_{i|V}(x)\right) + \mathrm{pen}(V) \right\}. \tag{D.5}$$

$$\widehat{V}_{(2)} = \arg\min_{V\in\mathcal{V}^{(2)}_{s,\Lambda,p_*}} \left\{ -\sum_{x\in\mathcal{X}(V)} \widehat{P}(x(V))\log\left(\widehat{P}_{i|V}(x)\right) + \mathrm{pen}(V) \right\}. \tag{D.6}$$

The following theorem shows the oracle properties of the selected estimator when the penalty term is suitably chosen.

Theorem D.2. *Let $s > 0$, $\delta > 1$, $p_* > 0$, $\Lambda \geq 100$ and let $\mathcal{V}_{s,\Lambda,p_*}$ and $\mathcal{V}^{(2)}_{s,\Lambda,p_*}$ be the collections defined in (D.3) and (D.4). Let $K > 1$ and let $\widehat{V}$ and $\widehat{V}_{(2)}$ be the penalized estimators defined in (D.5) and (D.6) with*

$$\mathrm{pen}(V) \geq 9K\frac{a^v}{n}.$$

Then, we have, for all $\eta > 0$, with probability larger than $1-3\delta^{-1}$,

$$\frac{1-\eta}{1+\eta}K_P(P_{i|S},\widehat{P}_{i|\widehat{V}}) \leq \inf_{V\in\mathcal{V}_{s,\Lambda,p_*}}\left\{K_P(P_{i|S},P_{i|V}) + \mathrm{pen}(V)\right\} + \left(2\log n + \frac{C_{\Lambda,p_*,K}}{\eta}\right)\frac{\log(N_s^2\delta)}{n}.$$

Also, for all $\eta > 0$, with probability larger than $1-3\delta^{-1}$,

$$\frac{1-\eta}{1+\eta}K_P(P_{i|S},\widehat{P}_{i|\widehat{V}_{(2)}}) \leq \inf_{V\in\mathcal{V}^{(2)}_{s,\Lambda,p_*}}\left\{K_P(P_{i|S},P_{i|V}) + \mathrm{pen}(V)\right\} + \left(2\log n + \frac{C_{\Lambda,p_*,K}}{\eta}\right)\frac{\log(N_s^2\delta)}{n}.$$

Comments:

- We use the same kind of penalty as in the L_2 case. This is not surprising because the variance parts of the risks were controlled in the same way.
- We do not optimize the bound obtained in Theorem D.1 among all the sets in $\mathcal{V}_{s,\Lambda}$. We have to restrict ourselves to $\mathcal{V}_{s,\Lambda,p_*}$. However, the constant $C_{\Lambda,p_*,K}$ has the form $p_*^{-1}C_{\Lambda,K}$. Therefore, we can choose $p_* = (\log n)^{-1}$ and optimize the result asymptotically.
- We optimize the bound among all $\mathcal{V}_s$ under the weak Gibbs assumption **NN**.

D.2. Slope heuristic in the Küllback case

The purpose of this section is to give the equivalent of Theorems 5.1 and 5.2 in the case of Küllback loss.

Theorem D.3. *Let $s > 0$, $\delta > 1$, $\epsilon > 0$, $r > 0$, $p_* > 0$, $\Lambda \geq 100$ and let $\mathcal{V}_{s,\Lambda,p_*}$, $\mathcal{V}^{(2)}_{s,\Lambda,p_*}$ be the collections defined in (D.3) and (D.4). For all V in $\mathcal{V}_s$, let*

$$p_2(V) = \sum_{x \in \mathcal{X}(V)} \widehat{P}(x(V)) \log\left(\frac{\widehat{P}_{i|V}(x)}{P_{i|V}(x)}\right).$$

Let $\widehat{V}$ be the penalized estimator defined in (D.5) with a penalty term satisfying

$$P\left(\forall V \in \mathcal{V}_{s,\Lambda,p_*},\ 0 \leq \mathrm{pen}(V) \leq (1-r)p_2(V)\right) \geq 1 - \epsilon.$$

Then, we have, with probability larger than $1 - 2\delta^{-1} - \epsilon$,

$$p_2(\widehat{V}) \geq \max_{V \in \mathcal{V}_{s\Lambda,p_*}} \left\{ rp_2(V) - 2K(P_{i|S}, P_{i|V}) \right\} - \frac{\log(N_s^2\delta)}{n}\left(4\log n + \frac{3}{2p_*}\right)$$

Let $\widehat{V}_{(2)}$ be the penalized estimator defined in (D.6) with a penalty term satisfying

$$P\left(\forall V \in \mathcal{V}^{(2)}_{s,\Lambda,p_*},\ 0 \leq \mathrm{pen}(V) \leq (1-r)p_2(V)\right) \geq 1 - \epsilon.$$

Also, we have, with probability larger than $1 - 2\delta^{-1} - \epsilon$,

$$p_2(\widehat{V}_{(2)}) \geq \max_{V \in \mathcal{V}^{(2)}_{s\Lambda,p_*}} \left\{ rp_2(V) - 2K(P_{i|S}, P_{i|V}) \right\} - \frac{\log(N_s^2\delta)}{n}\left(4\log n + \frac{3}{2p_*}\right)$$

Comments:

- Theorem D.3 states that, when the penalty term is smaller than $p_2(V)$, the complexity $p_2(\widehat{V})$ is as large as possible. This is exactly **SH1**, with $\mathrm{pen}_{\min}(V) = \Delta_V = p_2(V)$.

Theorem D.4. *Let $s > 0$, $\delta > 1$, $\epsilon > 0$, $r_1 > 0$, $r_2 > 0$, $p_* > 0$, $\Lambda \geq 100$ and let $\mathcal{V}_{s,\Lambda,p_*}$, $\mathcal{V}^{(2)}_{s,\Lambda,p_*}$ be the collections defined in (D.3) and (D.4). For all V in $\mathcal{V}_s$, let $p_2(V)$ be the quantity defined in Theorem D.3. Let $\widehat{V}$ be the penalized estimator defined in (D.5) with a penalty term satisfying*

$$P\left(\forall V \in \mathcal{V}_{s,\Lambda,p_*},\ (1+r_1)p_2(V) \leq \mathrm{pen}(V) \leq (1+r_2)p_2(V)\right) \geq 1 - \epsilon.$$

Then, there exists an absolute constant C such that, for all $\eta > 0$, with probability larger than $1 - 2\delta^{-1} - \epsilon$,

$$C_L K(P_{i|S}, \widehat{P}_{i|\widehat{V}}) \leq \inf_{V \in \mathcal{V}_{s,\Lambda,p_*}} \left\{ K(P_{i|S}, \widehat{P}_{i|V}) \right\} + \left(2\log(n) + \frac{C_{r_1,r_2,p_*}}{\eta}\right)\frac{\log(N_s^2\delta)}{n}, \tag{D.7}$$

where

$$C_L = \frac{(1-\eta) \wedge \big(r_1 - C(1+r_1)\Lambda^{-1/2}\big)}{(1+\eta) \vee \big(r_2 + C(1+r_2)\Lambda^{-1/2}\big)}.$$

Let $\widehat{V}_{(2)}$ *be the penalized estimator defined in (D.6) with a penalty term satisfying*

$$P\left(\forall V \in \mathcal{V}^{(2)}_{s,\Lambda,p_*},\ (1+r_1)p_2(V) \le \mathrm{pen}(V) \le (1+r_2)p_2(V)\right) \ge 1-\epsilon.$$

Also, there exists an absolute constant C *such that, for all* $\eta > 0$*, with probability larger than* $1 - 2\delta^{-1} - \epsilon$*,*

$$C_L K(P_{i|S}, \widehat{P}_{i|\widehat{V}_{(2)}}) \le \inf_{V \in \mathcal{V}^{(2)}_{s,\Lambda,p_*}} \left\{K(P_{i|S}, \widehat{P}_{i|V})\right\} + \left(2\log(n) + \frac{C_{r_1,r_2,p_*}}{\eta}\right)\frac{\log(N_s^2\delta)}{n}, \tag{D.8}$$

Comments:

- Let us take $\Lambda = 100 \vee \log(n)$. Take at first r_1 and r_2 slightly larger than 0 and therefore a penalty slightly larger than $\mathrm{pen}_{\min}$. Then (D.7) implies that, when n is sufficiently large $C_L > 0$, hence
$$p_2(\widehat{V}) \le K(P_{i|S}, \widehat{P}_{i|\widehat{V}}) \le C_L^{-1}\left(\inf_{V \in \mathcal{V}_{s,\Lambda,p_*}} K(P_{i|S}, \widehat{P}_{i|V}) + \left(2\log(n) + \frac{C_{r_1,r_2,p_*}}{\eta}\right)\frac{\log(N_s^2\delta)}{n}\right)$$
$$<< \sup_{V \in \mathcal{V}_{s,\Lambda,p_*}} K(P_{i|V}, \widehat{P}_{i|V}).$$
This justifies **SH2**.
- Take now r_1 and r_2 equal to 1, so that the penalty is equal to $2\mathrm{pen}_{\min}$. Then, we can take $C_L \to 1$ in (D.7). This justifies **SH3**.

D.3. Proof of Theorem D.1:

Let V in $\mathcal{V}_{s,\Lambda}$ or $\mathcal{V}^{(2)}_{s,\Lambda}$ and let us define

$$p_1(V) = \sum_{x \in \mathcal{X}(V)} P(x(V)) \log\left(\frac{P_{i|V}(x)}{\widehat{P}_{i|V}(x)}\right),\ p_2(V) = \sum_{x \in \mathcal{X}(V)} \widehat{P}(x(V)) \log\left(\frac{\widehat{P}_{i|V}(x)}{P_{i|V}(x)}\right).$$

From Lemma D.8 and we have

$$p_1(V) \le \frac{10}{3}\sum_{x \in \mathcal{X}(V)} \frac{\left(P(x(V)) - \widehat{P}(x(V))\right)^2}{P(x(V))} + \frac{14}{9}\sum_{x \in \mathcal{X}(V/\{i\})} \frac{\left(P(x(V/\{i\})) - \widehat{P}(x(V/\{i\}))\right)^2}{P(x(V/\{i\}))}.$$

$$p_2(V) \le \frac{3}{2}\sum_{x \in \mathcal{X}(V)} \frac{\left(P(x(V)) - \widehat{P}(x(V))\right)^2}{P(x(V))} + \frac{7}{3}\sum_{x \in \mathcal{X}(V/\{i\})} \frac{\left(P(x(V/\{i\})) - \widehat{P}(x(V/\{i\}))\right)^2}{P(x(V/\{i\}))}.$$

Let $V_* = V$ or $V/\{i\}$. On the event $\Omega_{prob}(\delta)$ defined in Lemma D.7, thanks to Lemma D.7, we have

$$\sup_{x\in\mathcal{X}(V_*)} \frac{1}{\sqrt{P(x(V_*))}} \le \sup_{x\in\mathcal{X}(V_*)} \frac{1}{\sqrt{P(x(V_*))}} \le \sup_{x\in\mathcal{X}(V_*)} \frac{1+2\Lambda^{-1/2}}{\sqrt{\widehat{P}(x(V_*))}} \le \frac{2\sqrt{n}}{\sqrt{\Lambda\log(2a^s N_s\delta)}}.$$

As this quantity is not random, the same bound holds on $\Omega_{prob}(\delta)^c$. We can apply Theorem B.8 to get that, for all $x>0$, for all $\eta>0$, with probability larger than $1-2e^{-x}$,

$$p_1(V) \le \frac{44}{9}\left((1+\eta)^3\frac{a^v}{n} + \frac{4x}{\eta n} + \frac{128x^2}{n\eta^3\Lambda\log(2a^sN_s\delta)}\right).$$

$$p_2(V) \le \frac{23}{6}\left((1+\eta)^3\frac{a^v}{n} + \frac{4x}{\eta n} + \frac{128x^2}{n\eta^3\Lambda\log(2a^sN_s\delta)}\right).$$

We use a union bound to obtain that, for all V in $\mathcal{V}_{s,\Lambda}$ or $\mathcal{V}^{(2)}_{s,\Lambda}$, with probability larger than $1-\delta$,

$$p_1(V) \le \frac{44}{9}\left((1+\eta)^3\frac{a^v}{n} + \left(\frac{4}{\eta}+\frac{128}{\eta^3\Lambda}\right)\frac{\log(2N_s\delta)}{n}\right).$$

$$p_2(V) \le \frac{23}{6}\left((1+\eta)^3\frac{a^v}{n} + \left(\frac{4}{\eta}+\frac{128}{\eta^3\Lambda}\right)\frac{\log(2N_s\delta)}{n}\right).$$

D.4. Proof of Theorem D.2:

Let us first decompose the selection criterion as follows.

$$-\sum_{x\in\mathcal{X}(V)} \widehat{P}(x(V))\log\left(\widehat{P}_{i|V}(x)\right) + \mathrm{pen}(V) = K(P_{i|S},\widehat{P}_{i|V}) + \mathrm{pen}(V) - p_1(V) - p_2(V) + L(V)$$
$$+\int dP(x(S)\log\left(\frac{1}{P_{i|S}(x)}\right). \qquad \text{(D.9)}$$

In the previous decomposition, we have

$$p_1(V) = \sum_{x\in\mathcal{X}(V)} P(x(V))\log\left(\frac{P_{i|V}(x)}{\widehat{P}_{i|V}(x)}\right).$$

$$p_2(V) = \sum_{x\in\mathcal{X}(V)} \widehat{P}(x(V))\log\left(\frac{\widehat{P}_{i|V}(x)}{P_{i|V}(x)}\right).$$

$$L(V) = \sum_{x\in\mathcal{X}(V)} (\widehat{P}(x(V)) - P(x(V)))\log\left(\frac{1}{P_{i|V}(x)}\right).$$

We deduce from (D.9) and the definition of $\widehat{V}$ that, for all V in $\mathcal{V}_{s,\Lambda,p_*}$,

$$K(P_{i|S},\widehat{P}_{i|\widehat{V}}) \leq K(P_{i|S},P_{i|V})+\mathrm{pen}(V)-p_2(V)-\Big(\mathrm{pen}(\widehat{V})-p_1(\widehat{V})-p_2(\widehat{V})\Big)+L(V)-L(\widehat{V}). \tag{D.10}$$

Let $\Omega_{prob}(\delta)$ defined in Lemma D.7. Let $\eta > 0$ and let $\Omega_{p1,p2}(\delta)$ be the event, for all V in $\mathcal{V}_{s,\Lambda,p_*}$

$$p_1(V) \leq 5\left((1+\eta)^3\frac{a^v}{n}+\left(1+\frac{64}{\eta^2\Lambda}\right)\frac{4\log(2N_s\delta)}{\eta n}\right).$$
$$p_2(V) \leq 4\left((1+\eta)^3\frac{a^v}{n}+\left(1+\frac{64}{\epsilon^2\Lambda}\right)\frac{4\log(2N_s\delta)}{\eta n}\right).$$

Let $\Omega_L(\delta)$ be the event, for all V, V' in $\mathcal{V}_{s,\Lambda,p_*}$,

$$(L(V)-L(V'))\mathbf{1}_{\{\Omega_{prob}(\delta)\}} \leq \eta(K(P_{i|S},P_{i|V})+K(P_{i|S},P_{i|V'}))+\frac{\log(N_s^2\delta)}{n}\left(4\log n+\frac{3}{2\eta p_*}\right).$$

Let $\Omega = \Omega_{prob}(\delta)\cap\Omega_{p1,p2}(\delta)\cap\Omega_L(\delta)$. It comes from Lemma D.7, Theorem D.1 and Lemma D.10 that $P(\Omega^c) \leq 3\delta$. Moreover, on Ω, we have, for all V in $\mathcal{V}_{s,\Lambda,p_*}$,

$$\begin{aligned} p_1(\widehat{V})+p_2(\widehat{V})+L(V)-L(\widehat{V}) &\leq \mathrm{pen}(\widehat{V})+\eta(K(P_{i|S},P_{i|V})+K(P_{i|S},P_{i|\widehat{V}})) \\ &\quad+\frac{\log(N_s^2\delta)}{n}\left(4\log n+\frac{3}{2\eta p_*}+1+\frac{64}{(K-1)^{2/3}\Lambda}\right). \end{aligned}$$

Hence, on Ω,

$$\frac{1-\eta}{1+\eta}K(P_{i|S},\widehat{P}_{i|\widehat{V}}) \leq K(P_{i|S},P_{i|V})+\mathrm{pen}(V)+\frac{\log(N_s^2\delta)}{n}\left(4\log n+\frac{3}{2\eta p_*}+1+\frac{64}{(K-1)^{2/3}\Lambda}\right).$$

We deduce from (D.9) and the definition of $\widehat{V}_{(2)}$ that, for all V in $\mathcal{V}^{(2)}_{s,\Lambda,p_*}$,

$$K(P_{i|S},\widehat{P}_{i|\widehat{V}_{(2)}}) \leq K(P_{i|S},P_{i|V})+\mathrm{pen}(V)-p_2(V)-\Big(\mathrm{pen}(\widehat{V}_{(2)})-p_1(\widehat{V}_{(2)})-p_2(\widehat{V}_{(2)})\Big)+L(V)-L(\widehat{V}_{(2)}). \tag{D.11}$$

Let $\Omega_{prob}(\delta)$ defined in Lemma D.7. Let $\eta > 0$ and let $\Omega^{(2)}_{p1,p2}(\delta)$ be the event, for all V in $\mathcal{V}^{(2)}_{s,\Lambda,p_*}$

$$p_1(V) \leq 5\left((1+\eta)^3\frac{a^v}{n}+\left(1+\frac{64}{\eta^2\Lambda}\right)\frac{4\log(2N_s\delta)}{\eta n}\right).$$
$$p_2(V) \leq 4\left((1+\eta)^3\frac{a^v}{n}+\left(1+\frac{64}{\epsilon^2\Lambda}\right)\frac{4\log(2N_s\delta)}{\eta n}\right).$$

Let $\Omega^{(2)}_L(\delta)$ be the event, for all V, V' in $\mathcal{V}^{(2)}_{s,\Lambda,p_*}$,

$$(L(V)-L(V'))\mathbf{1}_{\{\Omega_{prob}(\delta)\}} \leq \eta(K(P_{i|S},P_{i|V})+K(P_{i|S},P_{i|V'}))+\frac{\log(N_s^2\delta)}{n}\left(4\log n+\frac{3}{2\eta p_*}\right).$$

Let $\Omega = \Omega_{prob}(\delta) \cap \Omega^{(2)}_{p1,p2}(\delta) \cap \Omega^{(2)}_{L}(\delta)$. It comes from Lemma D.7, Theorem D.1 and Lemma D.10 that $P(\Omega^{(2)}) \geq 1 - 3\delta$. Moreover, on $\Omega^{(2)}$, we have, for all V in $\mathcal{V}^{(2)}_{s,\Lambda,p_*}$,

$$p_1(\widehat{V}_{(2)}) + p_2(\widehat{V}_{(2)}) + L(V) - L(\widehat{V}_{(2)}) \leq \text{pen}(\widehat{V}_{(2)}) + \eta(K(P_{i|S}, P_{i|V}) + K(P_{i|S}, P_{i|\widehat{V}_{(2)}})) + \frac{\log(N_s^2\delta)}{n}\left(4\log n + \frac{3}{2\eta p_*} + 1 + \frac{64}{(K-1)^{2/3}\Lambda}\right).$$

Hence, on $\Omega^{(2)}$,

$$\frac{1-\eta}{1+\eta} K(P_{i|S}, \widehat{P}_{i|\widehat{V}_{(2)}}) \leq K(P_{i|S}, P_{i|V}) + \text{pen}(V) + \frac{\log(N_s^2\delta)}{n}\left(4\log n + \frac{3}{2\eta p_*} + 1 + \frac{64}{(K-1)^{2/3}\Lambda}\right).$$

D.5. Proof of Theorem D.3:

Let Ω_{pen} and $\Omega^{(2)}_{pen}$ be the events, for all V in $\mathcal{V}_{s,\Lambda,p_*}$, $0 \leq \text{pen}(V) \leq (1-r)p_2(V)$ and for all V in $\mathcal{V}^{(2)}_{s,\Lambda,p_*}$, $0 \leq \text{pen}(V) \leq (1-r)p_2(V)$. It comes from (D.10) that, on Ω_{pen}, for all V in $\mathcal{V}_{s,\Lambda,p_*}$,

$$K(P_{i|S}, P_{i|\widehat{V}}) - p_2(\widehat{V}) \leq K(P_{i|S}, P_{i|V}) - rp_2(V) + L(V) - L(\widehat{V}).$$

It comes from (D.10) that, on $\Omega^{(2)}_{\text{pen}}$, for all V in $\mathcal{V}^{(2)}_{s,\Lambda,p_*}$,

$$K(P_{i|S}, P_{i|\widehat{V}_{(2)}}) - p_2(\widehat{V}_{(2)}) \leq K(P_{i|S}, P_{i|V}) - rp_2(V) + L(V) - L(\widehat{V}_{(2)}).$$

Let $\Omega_{prob}(\delta)$ be the event defined on Lemma D.7 and $\Omega_L(\delta)$ be the event, for all V, V' in $\mathcal{V}_{s,\Lambda,p_*}$, for all $\eta > 0$

$$(L(V) - L(V')) \leq \eta(K(P_{i|S}, P_{i|V}) + K(P_{i|S}, P_{i|V'})) + \frac{\log(N_s^2\delta)}{n}\left(4\log n + \frac{3}{2\eta p_*}\right).$$

From Lemmas D.7 and D.10, we have $P(\Omega_{prob}(\delta) \cap \Omega_L(\delta)) \geq 1 - 2\delta$ and, on $\Omega_{prob}(\delta) \cap \Omega_L(\delta) \cap \Omega_{pen}$, we have, for $\eta = 1$,

$$-p_2(\widehat{V}) \leq 2K(P_{i|S}, P_{i|V}) - rp_2(V) + \frac{\log(N_s^2\delta)}{n}\left(4\log n + \frac{3}{2p_*}\right).$$

Let $\Omega^{(2)}_L(\delta)$ be the event, for all V, V' in $\mathcal{V}^{(2)}_{s,\Lambda,p_*}$, for all $\eta > 0$

$$(L(V) - L(V')) \leq \eta(K(P_{i|S}, P_{i|V}) + K(P_{i|S}, P_{i|V'})) + \frac{\log(N_s^2\delta)}{n}\left(4\log n + \frac{3}{2\eta p_*}\right).$$

From Lemmas D.7 and D.10, we have $P(\Omega_{prob}(\delta) \cap \Omega^{(2)}_L(\delta)) \geq 1 - 2\delta$ and, on $\Omega_{prob}(\delta) \cap \Omega^{(2)}_L(\delta) \cap \Omega^{(2)}_{pen}$, we have, for $\eta = 1$,

$$-p_2(\widehat{V}_{(2)}) \leq 2K(P_{i|S}, P_{i|V}) - rp_2(V) + \frac{\log(N_s^2\delta)}{n}\left(4\log n + \frac{3}{2p_*}\right).$$

D.6. Proof of Theorem D.4:

Let Ω_{pen} be the event, for all V in $\mathcal{V}_{s,\Lambda,p_*}$, $(1+r_1)p_2(V) \leq \mathrm{pen}(V) \leq (1+r_2)p_2(V)$. It comes from (D.10) that, on Ω_{pen}, for all V in $\mathcal{V}_{s,\Lambda,p_*}$,

$$\begin{aligned}&K(P_{i|S}, P_{i|\widehat{V}}) + r_1 p_1(\widehat{V}) + (1+r_1)(p_2(\widehat{V}) - p_1(\widehat{V}))\\ &\qquad\qquad \leq K(P_{i|S}, P_{i|V}) + r_2 p_1(V) + (1+r_2)(p_2(V) - p_1(V)) + L(V) - L(\widehat{V})\end{aligned}$$

Let $\Omega_{prob}(\delta)$ be the event defined on Lemma D.7 and $\Omega_L(\delta)$ be the event, for all V, V' in $\mathcal{V}_{s,\Lambda,p_*}$, for all $\eta > 0$,

$$(L(V) - L(V')) \leq \eta(K(P_{i|S}, P_{i|V}) + K(P_{i|S}, P_{i|V'})) + \frac{\log(N_s^2\delta)}{n}\left(4\log n + \frac{3}{2\eta p_*}\right).$$

From Lemmas D.7 and D.10, we have $P(\Omega_{prob}(\delta) \cap \Omega_L(\delta)) \geq 1 - 2\delta$ and, on $\Omega_{prob}(\delta) \cap \Omega_L(\delta) \cap \Omega_{\mathrm{pen}}$, we have, from Lemma D.9, for all V in $\mathcal{V}_{s,\Lambda,p_*}$,

$$|p_1(V) - p_2(V)| \leq \frac{C}{\sqrt{\Lambda}} p_1(V).$$

We obtain that, for all V in $\mathcal{V}_{s,\Lambda,p_*}$,

$$\begin{aligned}&(1-\eta)K(P_{i|S}, P_{i|\widehat{V}}) + \left(r_1 - \frac{C(1+r_1)}{\sqrt{\Lambda}}\right) p_1(\widehat{V})\\ &\qquad \leq (1+\eta)K(P_{i|S}, P_{i|V}) + \left(r_2 + \frac{C(1+r_2)}{\sqrt{\Lambda}}\right) p_1(V) + \frac{\log(N_s^2\delta)}{n}\left(4\log n + \frac{3}{2\eta p_*}\right).\end{aligned}$$

Let $\Omega_{\mathrm{pen}}^{(2)}$ be the event, for all V in $\mathcal{V}_{s,\Lambda,p_*}^{(2)}$, $(1+r_1)p_2(V) \leq \mathrm{pen}(V) \leq (1+r_2)p_2(V)$. It comes from (D.10) that, on $\Omega_{\mathrm{pen}}^{(2)}$, for all V in $\mathcal{V}_{s,\Lambda,p_*}^{(2)}$,

$$\begin{aligned}&K(P_{i|S}, P_{i|\widehat{V}_{(2)}}) + r_1 p_1(\widehat{V}_{(2)}) + (1+r_1)(p_2(\widehat{V}_{(2)}) - p_1(\widehat{V}_{(2)}))\\ &\qquad\qquad \leq K(P_{i|S}, P_{i|V}) + r_2 p_1(V) + (1+r_2)(p_2(V) - p_1(V)) + L(V) - L(\widehat{V}_{(2)})\end{aligned}$$

Let $\Omega_{prob}(\delta)$ be the event defined on Lemma D.7 and $\Omega_L^{(2)}(\delta)$ be the event, for all V, V' in $\mathcal{V}_{s,\Lambda,p_*}^{(2)}$, for all $\eta > 0$,

$$(L(V) - L(V')) \leq \eta(K(P_{i|S}, P_{i|V}) + K(P_{i|S}, P_{i|V'})) + \frac{\log(N_s^2\delta)}{n}\left(4\log n + \frac{3}{2\eta p_*}\right).$$

From Lemmas D.7 and D.10, we have $P(\Omega_{prob}(\delta) \cap \Omega_L^{(2)}(\delta)) \geq 1 - 2\delta$ and, on $\Omega_{prob}(\delta) \cap \Omega_L^{(2)}(\delta) \cap \Omega_{\mathrm{pen}}$, we have, from Lemma D.9, for all V in $\mathcal{V}_{s,\Lambda,p_*}^{(2)}$,

$$|p_1(V) - p_2(V)| \leq \frac{C}{\sqrt{\Lambda}} p_1(V).$$

We obtain that, for all V in $\mathcal{V}^{(2)}_{s,\Lambda,p_*}$,

$$(1-\eta)K(P_{i|S},P_{i|\widehat{V}_{(2)}})+\left(r_1-\frac{C(1+r_1)}{\sqrt{\Lambda}}\right)p_1(\widehat{V}_{(2)})$$
$$\leq(1+\eta)K(P_{i|S},P_{i|V})+\left(r_2+\frac{C(1+r_2)}{\sqrt{\Lambda}}\right)p_1(V)+\frac{\log(N_s^2\delta)}{n}\left(4\log n+\frac{3}{2\eta p_*}\right).$$

D.7. Basic tools for Küllback Loss

Let s be an integer larger than e. Let $\mathcal{V}_s$ be the collection of subsets of V with cardinality smaller than s. Let N_s be the cardinality of $\mathcal{V}_s$. Let i be a site in V. Let us first give an elementary lemma on Külback losses. It is a slightly sharper version of Lemma 6.3 in [CT06b].

Lemma D.5. *Let P, Q be two probability measures on a finite space A such that, for all a in A, $|P(a)-Q(a)|\leq\eta Q(a)$, with $\eta\leq 1/3$. Then*

$$\left(\frac{1}{2}-\frac{7\eta}{6}\right)\sum_{a\in A}\frac{(P(a)-Q(a))^2}{Q(a)}\leq\sum_{a\in A}P(a)\log\left(\frac{P(a)}{Q(a)}\right)\leq\left(\frac{1}{2}+\frac{5\eta}{6}\right)\sum_{a\in A}\frac{(P(a)-Q(a))^2}{Q(a)}.$$

Proof: Let us first prove the following inequality, that is valid for all $x\leq 1/3$.

$$x-x^2\left(\frac{1}{2}+\frac{\eta}{2}\right)\leq\log(1+x)\leq x-x^2\left(\frac{1}{2}-\frac{\eta}{2}\right).$$

It comes from the Taylor expansion.

$$\log(1+x)=x-\frac{x^2}{2}+\sum_{k\geq 3}\frac{(-1)^{k+1}x^k}{k}\leq x-\frac{x^2}{2}+\frac{x^2\eta}{3}\sum_{k\geq 0}\eta^k=x-x^2\left(\frac{1}{2}-\frac{\eta}{3(1-\eta)}\right).$$
$$\log(1+x)=x-\frac{x^2}{2}+\sum_{k\geq 3}\frac{(-1)^{k+1}x^k}{k}\geq x-\frac{x^2}{2}-\frac{x^2\eta}{3}\sum_{k\geq 0}\eta^k=x-x^2\left(\frac{1}{2}+\frac{\eta}{3(1-\eta)}\right).$$

We deduce from this inequality and the equality

$$\sum_{a\in A}P(a)\frac{P(a)-Q(a)}{Q(a)}=\sum_{a\in A}\frac{(P(a)-Q(a))^2}{Q(a)}$$

that

$$\sum_{a\in A} P(a)\log\left(\frac{P(a)}{Q(a)}\right) \le \sum_{a\in A} P(a)\frac{P(a)-Q(a)}{Q(a)} - \left(\frac{1}{2}-\frac{\eta}{2}\right)\sum_{a\in A}\frac{P(a)}{Q(a)}\frac{(P(a)-Q(a))^2}{Q(a)}$$

$$= \sum_{a\in A}\frac{(P(a)-Q(a))^2}{Q(a)}\left(\frac{1}{2}+\frac{\eta}{2}+\left|\frac{P(a)}{Q(a)}-1\right|\left(\frac{1}{2}-\frac{\eta}{2}\right)\right)$$

$$\sum_{a\in A} P(a)\log\left(\frac{P(a)}{Q(a)}\right) \ge \sum_{a\in A} P(a)\frac{P(a)-Q(a)}{Q(a)} - \left(\frac{1}{2}+\frac{\eta}{2}\right)\sum_{a\in A}\frac{P(a)}{Q(a)}\frac{(P(a)-Q(a))^2}{Q(a)}$$

$$= \sum_{a\in A}\frac{(P(a)-Q(a))^2}{Q(a)}\left(\frac{1}{2}-\frac{\eta}{2}-\left|\frac{P(a)}{Q(a)}-1\right|\left(\frac{1}{2}+\frac{\eta}{2}\right)\right)$$

D.8. Basic Concentration Inequality

Let us now give an elementary concentration results derived from Benett's inequality.

Lemma D.6. *Let $\delta > 1$. With probability larger than $1-\delta^{-1}$, for all $(V \times x) \in (\mathcal{V}_s \times \mathcal{X}(S))$, we have*

$$\left|P(x(V)) - \widehat{P}(x(V))\right| \le \sqrt{\frac{2P(x(V))\log(2a^sN_s\delta)}{n}} + \frac{\log(2a^sN_s\delta)}{3n}.$$

Proof: Let V in $\mathcal{V}_s$ and x in $\mathcal{X}(V)$, we have from Benett's inequality, for all $t > 0$,

$$P\left(\left|P(x(V)) - \widehat{P}(x(V))\right| > \sqrt{\frac{2\mathrm{Var}(\mathbf{1}_{\{x(V)\}})t}{n}} + \frac{t}{3n}\right) \le 2e^{-t}.$$

We have $\mathrm{Var}(\mathbf{1}_{\{x(V)\}}) \le P(x(V))$. Hence, we conclude the proof with a union bound.
We deduce from Lemma D.6 the following typicality results.

Lemma D.7. *Let $\Lambda \ge 100$. Let $\Omega_{prob}(\delta)$ be the following event,*

$$\left\{\forall (V,x) \in \mathcal{V}_n \times \mathcal{X}(S),\ \left|P(x(V)) - \widehat{P}(x(V))\right| \le \sqrt{\frac{2P(x(V))\log(2a^sN_s\delta)}{n}} + \frac{\log(2a^sN_s\delta)}{3n}\right\}.$$

We have $P(\Omega_{prob}(\delta)) \ge 1-\delta^{-1}$ and, on $\Omega_{prob}(\delta)$, for all V in $\mathcal{V}_s$ and all x in $\mathcal{X}(S)$ such that

$$P(x(V)) \ge \Lambda\frac{\log(2a^sN_s\delta)}{n},$$

We have

$$\left|P(x(V)) - \widehat{P}(x(V))\right| \le 2\sqrt{\frac{P(x(V))\log(2a^sN_s\delta)}{n}} \le \frac{2P(x(V))}{\sqrt{\Lambda}}.$$

$$\left|P_{i|V}(x) - \widehat{P}_{i|V}(x)\right| \le \sqrt{\frac{11}{\Lambda}}P_{i|V}(x).$$

On $\Omega_{prob}(\delta)$, for all V in $\mathcal{V}_s$ and all x in $\mathcal{X}(S)$ such that

$$\widehat{P}(x(V)) \geq \Lambda \frac{\log(2a^s N_s \delta)}{n},$$

We have

$$\left|P(x(V)) - \widehat{P}(x(V))\right| \leq 2\sqrt{\frac{P(x(V))\log(2a^s N_s\delta)}{n}} \leq \frac{2P(x(V))}{\sqrt{\Lambda}}.$$

$$\left|P_{i|V}(x) - \widehat{P}_{i|V}(x)\right| \leq \sqrt{\frac{11}{\Lambda}} P_{i|V}(x).$$

Proof: When $P(x(V)) \geq \Lambda \frac{\log(2a^s N_s \delta)}{n}$, we have

$$\frac{\log(2a^s N_s\delta)}{3n} \leq \frac{1}{3\sqrt{2\Lambda}}\sqrt{\frac{2P(x(V))\log(2a^s N_s\delta)}{n}}, \text{ and } \sqrt{\frac{P(x(V))\log(2a^s N_s\delta)}{n}} \leq \frac{P(x(V))}{\sqrt{\Lambda}}.$$

This gives the first inequalities, as $\sqrt{2}+(3\sqrt{2\Lambda})^{-1} \leq 2$. We also have, since $P(x(V/\{i\})) \geq P(x(V))$,

$$\left|P(x(V/\{i\})) - \widehat{P}(x(V/\{i\}))\right| \leq 2\sqrt{\frac{P(x(V/\{i\}))\log(2a^s N_s\delta)}{n}} \leq \frac{2P(x(V/\{i\}))}{\sqrt{\Lambda}}.$$

From Lemma B.1, we have

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

Hence,

$$\left|P_{i|V}(x) - \widehat{P}_{i|V}(x)\right| \leq 2\sqrt{\frac{\log(2a^s N_s\delta)}{nP(x(V/\{i\}))}}\left(\sqrt{P_{i|V}(x)} + \widehat{P}_{i|V}(x)\right).$$

We just prove that

$$\frac{\widehat{P}_{i|V}(x)}{P_{i|V}(x)} \leq \left(\frac{1+2\Lambda^{-1/2}}{1-2\Lambda^{-1/2}}\right)^2, \text{ hence } \widehat{P}_{i|V}(x) \leq \sqrt{\widehat{P}_{i|V}(x)} \leq \frac{1+2\Lambda^{-1/2}}{1-2\Lambda^{-1/2}}\sqrt{P_{i|V}(x)}.$$

Therefore,

$$\begin{aligned}\left|P_{i|V}(x) - \widehat{P}_{i|V}(x)\right| &\leq \frac{4}{1-2\Lambda^{-1/2}}\sqrt{\frac{\log(2a^s N_s\delta)}{nP(x(V/\{i\}))}}\sqrt{P_{i|V}(x)} \\ &\leq \frac{4}{1-2\Lambda^{-1/2}}\sqrt{\frac{\log(2a^s N_s\delta)}{nP(x(V))}}P_{i|V}(x) \leq \frac{4}{\Lambda^{1/2}-2}P_{i|V}(x).\end{aligned}$$

Let $u^2 = n^{-1}\log(2a^s N_s \delta)$. On Ω_{prob}, we have

$$\widehat{P}(x(V)) \leq P(x(V)) + 2\frac{u}{\sqrt{2}}\sqrt{P(x(V))} + \frac{u^2}{3} = \left(\sqrt{P(x(V))} + \frac{u}{\sqrt{2}}\right)^2 + \frac{u^2}{12}.$$

Since $\widehat{P}(x(V)) \geq \Lambda u^2$, we deduce that

$$P(x(V)) \geq \left(\sqrt{\Lambda - \frac{1}{12}} - \frac{1}{\sqrt{2}}\right)^2 u^2 = \left(\Lambda + \frac{5}{12} - \sqrt{\frac{6\Lambda - 1}{3}}\right) u^2.$$

Since $\Lambda \geq 2$, we deduce that

$$\left|\widehat{P}(x(V)) - P(x(V))\right| \leq \left(\sqrt{2} + \frac{1}{3\sqrt{\Lambda + \frac{5}{12} - \sqrt{\frac{6\Lambda-1}{3}}}}\right) u\sqrt{P(x(V))} \leq 2u\sqrt{P(x(V))}.$$

From the same inequality, we also obtain

$$\left|\widehat{P}(x(V)) - P(x(V))\right| \leq \frac{2P(x(V))}{\sqrt{\Lambda - \frac{1}{12}} - \frac{1}{\sqrt{2}}} \leq \frac{2P(x(V))}{\sqrt{\Lambda}}.$$

Since $\widehat{P}(x(V/\{i\})) \geq \widehat{P}(x(V))$, we prove with the same arguments that

$$\left|P(x(V/\{i\})) - \widehat{P}(x(V/\{i\}))\right| \leq \sqrt{\frac{P(x(V/\{i\}))\log(2a^s N_s \delta)}{n}} \leq \frac{2P(x(V/\{i\}))}{\sqrt{\Lambda}}.$$

From Lemma B.1, we have

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

Hence,

$$\left|P_{i|V}(x) - \widehat{P}_{i|V}(x)\right| \leq 2\sqrt{\frac{\log(2a^s N_s \delta)}{nP(x(V/\{i\}))}}\left(\sqrt{P_{i|V}(x)} + \widehat{P}_{i|V}(x)\right).$$

If $\sqrt{P_{i|V}} \geq \widehat{P}_{i|V}$, we deduce that

$$\left|P_{i|V}(x) - \widehat{P}_{i|V}(x)\right| \leq 4\sqrt{\frac{\log(2a^s N_s \delta)P_{i|V}(x)}{nP(x(V/\{i\}))}}.$$

Otherwise, we have

$$\left(1 - \frac{4}{\sqrt{\Lambda}}\sqrt{1 + \frac{2}{\sqrt{\Lambda}}}\right)\widehat{P}_{i|V}(x) \leq \left(1 - 4\sqrt{\frac{\log(2a^s N_s \delta)}{nP(x(V/\{i\}))}}\right)\widehat{P}_{i|V}(x) \leq P_{i|V}(x).$$

Since $\Lambda \geq 100$, we obtain $\widehat{P}_{i|V}(x) \leq 2P_{i|V}(x)$. We deduce that

$$\left|P_{i|V}(x) - \widehat{P}_{i|V}(x)\right| \leq 2(1+\sqrt{2})\sqrt{\frac{\log(2a^s N_s \delta) P_{i|V}(x)}{nP(x(V/\{i\}))}} = 2(1+\sqrt{2})\sqrt{\frac{\log(2a^s N_s\delta)}{nP(x(V))}}P_{i|V}(x)$$

$$\leq \frac{2(1+\sqrt{2})}{\sqrt{\Lambda}}\sqrt{1+\frac{2}{\sqrt{\Lambda}}}P_{i|V}(x) \leq \sqrt{\frac{11}{\Lambda}}P_{i|V}(x).$$

D.9. Control of the variance terms in Küllback loss:

The following Lemma gives an important decomposition of the Küllback loss.

Lemma D.8. *Let $\Lambda \geq 100$ and let $\mathcal{V}_{s,\Lambda}$, $\mathcal{V}_{s,\Lambda}^{(2)}$ be respectively the collection of subsets V in $\mathcal{V}_s$ such that, for all x in $\mathcal{X}(V)$,*

$$P(x(V) = 0, \text{ or } \widehat{P}(x(V)) \geq \Lambda\frac{\log(2a^s N_s\delta)}{n}$$

and the collection of subsets V in $\mathcal{V}_s$ such that, for all x in $\mathcal{X}(V)$,

$$P(x(V) = 0, \text{ or } P(x(V)) \geq \Lambda\frac{\log(2a^s N_s\delta)}{n}.$$

Let $\Omega_{prob}(\delta)$ be the event defined on Lemma D.7. On $\Omega_{prob}(\delta)$, for all V in $\mathcal{V}_{s,\Lambda}$, we have

$$p_1(V) \leq \frac{20}{6}\sum_{x\in\mathcal{X}(V)}\frac{\left(P(x(V)) - \widehat{P}(x(V))\right)^2}{P(x(V))} + \frac{14}{9}\sum_{x\in\mathcal{X}(V/\{i\})}\frac{\left(P(x(V/\{i\})) - \widehat{P}(x(V/\{i\}))\right)^2}{P(x(V/\{i\}))}.$$

$$p_2(V) \leq \frac{3}{2}\sum_{x\in\mathcal{X}(V)}\frac{\left(P(x(V)) - \widehat{P}(x(V))\right)^2}{P(x(V))} + \frac{7}{3}\sum_{x\in\mathcal{X}(V/\{i\})}\frac{\left(P(x(V/\{i\})) - \widehat{P}(x(V/\{i\}))\right)^2}{P(x(V/\{i\}))}.$$

Proof: From Lemma D.7, for all V in $\mathcal{V}_{s,\Lambda}$ or $\mathcal{V}_{s,\Lambda}^{(2)}$, for all x in V, we have $|P_{i|V}(x), \widehat{P}_{i|V}(x)| \leq$

$\sqrt{11\Lambda^{-1}}P_{i|V}(x)$. Hence, from Lemma D.5,

$$\begin{aligned}
p_1(V) &= \sum_{x\in\mathcal{X}(V)} P(x(V))\log\left(\frac{P_{i|V}(x)}{\widehat{P}_{i|V}(x)}\right)\\
&\le \frac{1}{2}\left(1+\frac{5\sqrt{11}}{3\sqrt{\Lambda}}\right)\sum_{x\in\mathcal{X}(V)} P(x(V/\{i\}))\frac{(P_{i|V}(x)-\widehat{P}_{i|V}(x))^2}{\widehat{P}_{i|V}(x)}.\\
p_2(V) &= \sum_{x\in\mathcal{X}(V)} \widehat{P}(x(V))\log\left(\frac{\widehat{P}_{i|V}(x)}{P_{i|V}(x)}\right)\\
&\le \frac{1}{2}\left(1+\frac{5\sqrt{11}}{3\sqrt{\Lambda}}\right)\sum_{x\in\mathcal{X}(V)} \widehat{P}(x(V/\{i\}))\frac{(P_{i|V}(x)-\widehat{P}_{i|V}(x))^2}{P_{i|V}(x)}.
\end{aligned}$$

From Lemma B.1, we have

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

The second inequality gives that $p_1(V)$ is smaller than

$$\begin{aligned}
&\le \left(1+\frac{5\sqrt{11}}{3\sqrt{\Lambda}}\right)\sum_{x\in\mathcal{X}(V)}\frac{\left(P(x(V))-\widehat{P}(x(V))\right)^2}{P(x(V/\{i\}))\widehat{P}_{i|V}(x)}+\widehat{P}_{i|V}(x)\frac{\left(P(x(V/\{i\}))-\widehat{P}(x(V/\{i\}))\right)^2}{P(x(V/\{i\}))}\\
&\le \frac{1+\frac{5\sqrt{11}}{3\sqrt{\Lambda}}}{1-\sqrt{\frac{11}{\Lambda}}}\sum_{x\in\mathcal{X}(V)}\frac{\left(P(x(V))-\widehat{P}(x(V))\right)^2}{P(x(V))}\\
&+\left(1+\frac{5\sqrt{11}}{3\sqrt{\Lambda}}\right)\sum_{x\in\mathcal{X}(V/\{i\})}\frac{\left(P(x(V/\{i\}))-\widehat{P}(x(V/\{i\}))\right)^2}{P(x(V/\{i\}))}.
\end{aligned}$$

We also have

$$\left(P_{i|V}(x)-\widehat{P}_{i|V}(x)\right)^2 \le \frac{\left(P(x(V))-\widehat{P}(x(V))\right)^2+\widehat{P}_{i|V}(x)P_{i|V}(x))P(x(V/\{i\}))-\widehat{P}(x(V/\{i\}))^2}{P(x(V/\{i\}))\widehat{P}(x(V/\{i\}))}$$

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

$$\le \frac{3\left(P(x(V))-\widehat{P}(x(V))\right)^2+2(\widehat{P}_{i|V}(x)+P_{i|V}(x))^2\left(P(x(V/\{i\}))-\widehat{P}(x(V/\{i\}))\right)^2}{2P(x(V/\{i\}))\widehat{P}(x(V/\{i\}))}.$$

Hence,

$$p_2(V) \le \frac{3}{2}\sum_{x\in\mathcal{X}(V)}\frac{\left(P(x(V))-\widehat{P}(x(V))\right)^2}{P(x(V))}$$

$$+\sum_{x\in\mathcal{X}(V)}\frac{\left(P(x(V/\{i\}))-\widehat{P}(x(V/\{i\}))\right)^2}{P(x(V/\{i\}))}\frac{(\widehat{P}_{i|V}(x)+P_{i|V}(x))^2}{P_{i|V}(x)}$$

is smaller than

$$\frac{3}{2}\sum_{x\in\mathcal{X}(V)}\frac{\left(P(x(V))-\widehat{P}(x(V))\right)^2}{P(x(V))}+\left(2+\sqrt{\frac{11}{\Lambda}}\right)\sum_{x\in\mathcal{X}(V/\{i\})}\frac{\left(P(x(V/\{i\}))-\widehat{P}(x(V/\{i\}))\right)^2}{P(x(V/\{i\}))}.$$

D.10. Concentration for the slope heuristic in the Küllback case

Lemma D.9. *Let $\Lambda \ge 100$ and let $\mathcal{V}_{s,\Lambda}$ and $\mathcal{V}_{s,\Lambda}^{(2)}$ be respectively the collection of subsets V in $\mathcal{V}_s$ such that, for all x in $\mathcal{X}(V)$,*

$$P(x(V)=0,\ \text{or}\ \widehat{P}(x(V))\ge\Lambda\frac{\log(2a^sN_s\delta)}{n}$$

and the collection of subsets V in $\mathcal{V}_s$ such that, for all x in $\mathcal{X}(V)$,

$$P(x(V)=0,\ \text{or}\ P(x(V))\ge\Lambda\frac{\log(2a^sN_s\delta)}{n}.$$

Let $\Omega_{prob}(\delta)$ be the event defined on Lemma D.7. On $\Omega_{prob}(\delta)$, there exists an absolute constant $C>0$ such that

$$|p_1(V)-p_2(V)|\le\frac{C}{\sqrt{\Lambda}}p_1(V).$$

Proof: We use Lemmas D.5 and D.7. On $\Omega_{prob}(\delta)$, we have

$$\frac{1}{2}\left(1-\frac{7\sqrt{11}}{3\sqrt{\Lambda}}\right)\sum_{x\in\mathcal{X}(V)} P(x(V/\{i\}))\frac{(P_{i|V}(x)-\widehat{P}_{i|V}(x))^2}{\widehat{P}_{i|V}(x)} \leq p_1(V).$$

$$p_1(V) \leq \frac{1}{2}\left(1+\frac{5\sqrt{11}}{3\sqrt{\Lambda}}\right)\sum_{x\in\mathcal{X}(V)} P(x(V/\{i\}))\frac{(P_{i|V}(x)-\widehat{P}_{i|V}(x))^2}{\widehat{P}_{i|V}(x)}.$$

$$\begin{aligned}
&\frac{1}{2}\left(1-\frac{7\sqrt{11}}{3\sqrt{\Lambda}}\right)\left(1-\sqrt{\frac{11}{\Lambda}}\right)\left(1-\sqrt{\frac{4}{\Lambda}}\right)\sum_{x\in\mathcal{X}(V)} P(x(V/\{i\}))\frac{(P_{i|V}(x)-\widehat{P}_{i|V}(x))^2}{\widehat{P}_{i|V}(x)}\\
&\leq p_2(V) = \sum_{x\in\mathcal{X}(V)} \widehat{P}(x(V))\log\left(\frac{\widehat{P}_{i|V}(x)}{P_{i|V}(x)}\right)\\
&\leq \frac{1}{2}\left(1+\frac{5\sqrt{11}}{3\sqrt{\Lambda}}\right)\left(1+\sqrt{\frac{11}{\Lambda}}\right)\left(1+\sqrt{\frac{4}{\Lambda}}\right)\sum_{x\in\mathcal{X}(V)} P(x(V/\{i\}))\frac{(P_{i|V}(x)-\widehat{P}_{i|V}(x))^2}{\widehat{P}_{i|V}(x)}.
\end{aligned}$$

D.11. Concentration of L(V)-L(V')

The following Lemma let us control the remainder term in the oracle inequality.

Lemma D.10. *Let $\delta > 1$ and let $\mathcal{V}_{s,\Lambda,p_*}$ be the subset of $\mathcal{V}_{s,\Lambda}$ of the sets V such that, for all x in $\mathcal{X}(V$, $P_{i|V}(x) = 0$ or $\widehat{P}_{i|V}(x) \geq p_*$. With probability at least $1-\delta$, for all V, V' in $\mathcal{V}_{s,\Lambda,p_*}$, for all $\eta > 0$, we have,*

$$\begin{aligned}
&\sum_{x\in\mathcal{X}(V\cup V')} (\widehat{P}(x(V\cup V')) - P(x(V\cup V'))\log\left(\frac{P_{i|V}(x)}{P_{i|V'}(x)}\right)\\
&\qquad \leq \eta(K(P_{i|S}, P_{i|V}) + K(P_{i|S}, P_{i|V'})) + \frac{\log(N_s^2\delta)}{n}\left(4\log n + \frac{3}{2\eta p_*}\right).
\end{aligned}$$

Let $\mathcal{V}^{(2)}_{s,\Lambda,p_*}$ be the subset of $\mathcal{V}^{(2)}_{s,\Lambda}$ of the sets V such that, for all x in $\mathcal{X}(V$, $P_{i|V}(x) = 0$ or $P_{i|V}(x) \geq p_*$. With probability at least $1-\delta$, for all V, V' in $\mathcal{V}^{(2)}_{s,\Lambda,p_*}$, for all $\eta > 0$, we have,

$$\begin{aligned}
&\sum_{x\in\mathcal{X}(V\cup V')} (\widehat{P}(x(V\cup V')) - P(x(V\cup V'))\log\left(\frac{P_{i|V}(x)}{P_{i|V'}(x)}\right)\\
&\qquad \leq \eta(K(P_{i|S}, P_{i|V}) + K(P_{i|S}, P_{i|V'})) + \frac{\log(N_s^2\delta)}{n}\left(4\log n + \frac{3}{2\eta p_*}\right).
\end{aligned}$$

Proof: Let us first write

$$\sum_{x\in\mathcal{X}(V\cup V')}(\widehat{P}(x(V\cup V'))-P(x(V\cup V'))\log\left(\frac{P_{i|V}(x)}{P_{i|V'}(x)}\right)$$
$$\leq\sum_{x\in\mathcal{X}(V\cup V')}(\widehat{P}(x(V\cup V'))-P(x(V\cup V'))\left(\log\left(\frac{P_{i|V\cup V'}(x)}{P_{i|V'}(x)}\right)-\log\left(\frac{P_{i|V\cup V'}(x)}{P_{i|V}(x)}\right)\right).$$

Let us now write V_* for V or V'. We have

$$\sum_{x\in\mathcal{X}(V\cup V')}(\widehat{P}(x(V\cup V'))-P(x(V\cup V'))\log\left(\frac{P_{i|V\cup V'}(x)}{P_{i|V_*}(x)}\right)$$
$$=(P_n-P)\left(\sum_{x\in\mathcal{X}(V\cup V')}\log\left(\frac{P_{i|V\cup V'}(x)}{P_{i|V_*}(x)}\right)\mathbf{1}_{\{x(V\cup V')\}}\right).$$

The function $f:\mathcal{X}(V\cup V')\to\mathbb{R}$, $x\mapsto\log\left(\frac{P_{i|V\cup V'}(x)}{P_{i|V_*}(x)}\right)$ is upper bounded on $\Omega_{prob}(\delta)$ by $2\log n$. Since it is not random, the bound also holds on $\Omega_{prob}(\delta)^c$. Let us evaluate its variance

$$\mathrm{Var}(f(X))\leq Pf^2=\sum_{x\in\mathcal{X}(V\cup V')}P(x(V\cup V'))\left(\log\left(\frac{P_{i|V\cup V'}(x)}{P_{i|V_*}(x)}\right)\right)^2$$

Let us recall also here the following Lemma see [Mas07], Lemma 7.24 p 275 or [BS91])

Lemma D.11. *For all probability measures P and Q, with $P << Q$,*

$$\frac{1}{2}\int(dP\wedge dQ)\left(\log\left(\frac{dP}{dQ}\right)\right)^2\leq K(P,Q)\leq\frac{1}{2}\int(dP\vee dQ)\left(\log\left(\frac{dP}{dQ}\right)\right)^2.$$

Since $P_{i|V_*}(x)\geq 2\widehat{P}_{i|V_*}(x)/3\geq 2p_*/3$, we deduce that

$$\mathrm{Var}(f(X))\leq\frac{3}{p_*}K(P_{i|V\cup V'},P_{i|V_*}).$$

Applying Benett's inequality to f, we obtain that, with probability $1-2e^{-t}$,

$$(P_n-P)\left(\sum_{x\in\mathcal{X}(V\cup V')}\log\left(\frac{P_{i|V\cup V'}(x)}{P_{i|V_*}(x)}\right)\mathbf{1}_{\{x(V\cup V')\}}\right)\leq\sqrt{\frac{6}{p_*}K(P_{i|V\cup V'},P_{i|V_*})\frac{t}{n}}+\frac{2t\log n}{n}.$$

We conclude the proof with a union bound and the classical inequality $2ab\leq\eta a^2+\eta^{-1}b^2$.