Learning latent structure of large random graphs

Roland Diel, Sylvain Le Corff, Matthieu Lerasle

Learning latent structure of large random graphs

Roland Diel* Sylvain Le Corff† Matthieu Lerasle†

Abstract

In this paper, we estimate the distribution of hidden nodes weights in large random graphs from the observation of very few edges weights. In this very sparse setting, the first non-asymptotic risk bounds for maximum likelihood estimators (MLE) are established. The proof relies on the construction of a graphical model encoding conditional dependencies that is extremely efficient to study n-regular graphs obtained using a round-robin scheduling. This graphical model allows to prove geometric loss of memory properties and deduce the asymptotic behavior of the likelihood function. Following a classical construction in learning theory, the asymptotic likelihood is used to define a measure of performance for the MLE. Risk bounds for the MLE are finally obtained by subgaussian deviation results derived from concentration inequalities for Markov chains applied to our graphical model.

1 Introduction

Inference in large random graphs is an important topic of interest due to its applications to many fields such as data science, sociology or neurobiology for instance. This paper focuses on large random graphs whose heterogeneity is described by latent data models. The nodes are associated with latent random weights, independent and with unknown distribution. The only available information is given by random weights associated with few edges in the graph which are independent conditionally on the nodes weights. The objective is to estimate the unknown distribution of the nodes weights from these observations. This latent data structure is appealing as it may be used to describe graphs in a wide range of applications. In sports tournaments, nodes represent contestants in a championship and each node weight is the "intrinsic value" of the corresponding player. An edge is drawn between players when they face each others, the result of a contest is the observed edge weight. The problem is to recover from a few games the distribution of the intrinsic values of the players to make early prediction on the issue of the championship for example. In social networks, nodes are members and their weights represent the "popularity" of each member. An edge is drawn between members if a "suggestion of friendship" has been made to one of them. The observed edge weight is 0 if these people are not connected and 1 otherwise. The problem here is to estimate the popularity density in a large population where only a few suggestions of friendship can be made compared to the global size of the network. In neurobiology, random graphs may be used to model neural functional connectivity inside the brain. In this case, nodes are neurons and their weights represent their efficiency to diffuse neural information. An edge between neurons is drawn if the activity of these neurons is observed simultaneously. The weight of this edge is a score

[1]Laboratoire J.A.Dieudonné UMR CNRS-UNS 6621 Université de Nice Sophia-Antipolis 06108 Nice Cedex 2.
[2]Laboratoire de mathématiques d'Orsay, Univ. Paris-Sud, CNRS, Université Paris-Saclay, 91405 Orsay.

representing the influence that these neurons exercise on each other. The problem is therefore to estimate the functional connectivity density inside the brain from these scores.

The problem studied in this paper has a long history, going back at least to [31] who considered the problem of *paired comparison* to evaluate performances of medicines. In [31] and later in [2], the problem was to recover the weights of a finite number of nodes when the number of measurements on every pair grows to infinity. Further extensions of the so-called Bradley-Terry model have then been studied, see for example [7] for a review. More recently, [22] considered the problem of estimating nodes weights in Bradley-Terry models based on one measurement per pair of nodes when the number N of nodes grows to infinity. This framework led to several developments in computational statistics for the Bradley-Terry model, see [14] and [4] for various extensions of this original model. A related problem was considered in [5] where an edge is inserted between each pair of nodes with a probability depending on the nodes weights. Each node has therefore a random degree and the observed degrees are used to infer nodes weights. When the graph is fully observed, [5] proved that with a probability of order $1 - 1/N^2$, there exists a unique maximum likelihood estimator of the nodes weights which is such that the supremum norm of the estimation error is upper bounded by $\sqrt{\log N/N}$.

This paper strongly departs from these settings where the all graph is observed, even from [30] where some edges are missing. We consider a very sparse alternative where only very few edges per nodes are observed. A reason why such a sparse setting has never been considered is probably due to [31] who proved that the estimation of the weights is actually impossible in the Bradley-Terry model in this situation. To overcome this issue, we consider the problem of estimating *the distribution of the weights* and not the weights themselves. There are several motivations to adopt this new approach. The Bradley-Terry model in "random environment" was applied with success to predict the issue of a championship by estimating the probability distribution of the teams weights (strengths) which were assumed to be uniformly distributed, see for example [23] and references therein. Moreover, [6] recently showed that the node with maximal weight can be recovered if the tail of the nodes weights distribution is sufficiently convex. More generally, the idea to use a Bayesian estimator when a frequentist approach is not available is rather standard. The performances of this estimator highly depend on the prior distribution of the parameters and providing a reasonable prior may have a great impact. The study of bayesian estimators with an estimated prior is known as empirical bayes theory [21] and is currently a subject of intense research, see for example [13] for a recent overview. The problem presented in this paper can be understood as finding a statistically efficient estimator of the prior to design an empirical bayes estimator for the nodes weights. The use of latent variables is also at the heart of mixed effect models widely spread in biostatistics, see [15].

This paper shows the first non-asymptotic risk bounds for non-parametric maximum likelihood estimators (MLE) of the distribution of nodes weights. Asymptotic properties of MLE rely heavily on a loss of memory property of the observed random graph. This can be analyzed using a graphical model describing the conditional dependencies between nodes and edges. This graphical model provides a natural parallel with hidden Markov models [3] which is used to study the asymptotic behavior of the likelihood, following [11] in particular. The limit likelihood defines a natural notion of risk to measure performances of MLE. These performances are obtained for finite values of the number N of nodes using concentration inequalities for Markov Chains [10]. The excess risk scales as the entropy of the underlying statistical model (in the sense of Dudley) normalized by a term of order $\sqrt{N}$ when n is fixed and $N \to \infty$. From a learning perspective, Dudley's entropy bound is known to be sub-optimal in general, it can be replaced by a majorizing measure bound [24] if needed

since the bound proposed in this paper is derived from a subgaussian concentration inequality for the underlying process, see Eq. (27).

More generally, we believe that the methodology introduced to prove our results leads the way to exciting research perspectives in various fields. For example, identifiability of non-parametric hidden Markov models with finite state spaces was established very recently along with the first convergence properties of estimators of the unknown distributions, see [8] for a penalized least-squares estimator of the emission densities, [9, 28, 29] for consistent estimation of the posterior distributions of the states and posterior concentration rates for the parameters or [16] for order estimation. However, very few theoretical results are available for the non-parametric estimation of general state spaces hidden Markov models. The arguments leading to our risk bound may probably be extended to this framework. In computational statistics, bayesian estimators of nodes weights have been studied in Bradley-Terry models and other extensions [4]. Designing new algorithms to compute MLE of the prior would therefore be of great interest to derive empirical bayes estimators of these weights.

The paper is organized as follows. Section 2 details the model and the maximum likelihood estimator of the unknown weights distribution. Section 3 presents preliminary results underlying our analysis. A graphical model encoding conditional dependencies in the original graph is built. The round-robin algorithm, a widely spread method in sports tournaments that builds sparse graphs for which our graphical model is stationary, is also presented. Our main results are finally given in Section 4. Convergence of the likelihood is established when the number N of nodes grows to $+\infty$ and risk bounds for the MLE are provided. Section 5 to 7 are devoted to the proofs of these results. Section 5 proves the fundamental properties of the graphical model associated with round-robin graphs. Section 6 proves the probabilistic tools required to establish the main results. These tools might be of independent interest, they are presented as independent results and hold for stationary processes with conditional dependencies encoded in the graphical model. Proofs of the main results are finally gathered in Section 7.

2 Setting

2.1 Random graphs with latent variables

Let n, N denote two positive integers and let $(\{1,\dots,N\}, E^{n,N})$ be a connected n-regular graph. Let $V_1,\dots,V_N$ denote independent and identically distributed (i.i.d.) random variables taking values in a measurable set $\mathcal{V}$ with common (unknown) distribution $\pi_\star$. For all $\{i,j\} \in E^{n,N}$, the observation $X_{i,j}$ takes values in a discrete set $\mathcal{X}$ and conditionally on $V = (V_1,\dots,V_N)$, the random variables $(X_{i,j})_{(i,j)\in E^{n,N}}$ are independent and such that the conditional distribution of $X_{i,j}$ is given by $k : \mathcal{X}\times\mathcal{V}\times\mathcal{V} \to [0,1]$:

$$\mathbb{P}(X_{i,j} = x|V) = k(x, V_i, V_j)\,.$$

This framework encompasses the following models.

Example 1 (Bradley-Terry model [2])**.** *In this example, $\mathcal{V} = \mathbb{R}_+^*$, $\mathcal{X} = \{0,1\}$ and for all $x \in \mathcal{X}$,*

$$k(x, V_i, V_j) = \left(\frac{V_i}{V_i + V_j}\right)^x \left(\frac{V_j}{V_i + V_j}\right)^{1-x}\,.$$

Example 2 (Extensions of Bradley-Terry model)**.** *In [4], the authors proposed several algorithms to perform Bayesian inference for generalized Bradley-Terry models which fit our framework.*

- *The Bradley-Terry model with home advantage introduces an additional parameter $\theta > 0$ to measure the home-field advantage. In this case, $\mathcal{V} = \mathbb{R}_+^*$, $\mathcal{X} = \{0, 1\}$ and, if the player i is home, for all $x \in \mathcal{X}$,*

$$k(x, V_i, V_j) = \left(\frac{\theta V_i}{\theta V_i + V_j}\right)^x \left(\frac{V_j}{\theta V_i + V_j}\right)^{1-x} .$$

- *The Bradley-Terry model with ties [20] introduces an additional parameter $\theta > 1$, $\mathcal{V} = \mathbb{R}_+^*$, $\mathcal{X} = \{-1, 0, 1\}$ and*

$$k(1, V_i, V_j) = \frac{V_i}{V_i + \theta V_j} \quad \text{and} \quad k(0, V_i, V_j) = \frac{(\theta^2 - 1) V_i V_j}{(\theta V_i + V_j)(V_i + \theta V_j)} .$$

Example 3 (Random graphs with a given degree sequence)**.** *[5] considers random graphs such that, for all $1 \le i < j \le N$, an edge is inserted between players i and j with probability $V_i V_j/(1 + V_i V_j)$ where $(V_1, \dots, V_N)$ are parameters to be estimated using the degrees of the vertices in the observed graph. Such random graphs fit our framework with $\mathcal{V} = \mathbb{R}_+^*$, $\mathcal{X} = \{0, 1\}$ ($X_{i,j} = 0$ in our framework representing $\{i, j\} \notin E$ in theirs) and for all $1 \le i < j \le N$, $x \in \mathcal{X}$,*

$$k(x, V_i, V_j) = \left(\frac{V_i V_j}{1 + V_i V_j}\right)^x \left(\frac{1}{1 + V_i V_j}\right)^{1-x} .$$

2.2 Maximum likelihood estimator

The weights $X^{n,N} = (X_{i,j})_{(i,j) \in E^{n,N}}$ are observed and the objective is to infer the distribution $\pi_\star$ of the hidden variables $V = (V_1, \dots, V_N)$ from these observations. Let Π be a set of probability measures on $\mathcal{V}$. For all $\pi \in \Pi \cup \{\pi_\star\}$, the joint distribution of $(X^{n,N}, V)$ is given by

$$\mathbb{P}_\pi^{n,N}(x^{n,N}, A) = \int \mathbb{1}_A(v) \prod_{(i,j) \in E^{n,N}} k(x_{i,j}^{n,N}, v_i, v_j) \pi^{\otimes N}(\mathrm{d}v) . \tag{1}$$

Using the convention $\log 0 = -\infty$, the log-likelihood is given, for all $\pi \in \Pi \cup \{\pi_\star\}$, by

$$\ell^{n,N}(\pi) = \log \mathbb{P}_\pi^{n,N}(X^{n,N}) , \qquad \text{where} \qquad \mathbb{P}_\pi^{n,N}(X^{n,N}) = \mathbb{P}_\pi^{n,N}(X^{n,N}, \mathcal{V}) .$$

In this paper, $\pi_\star$ is estimated by the standard maximum likelihood estimator $\widehat{\pi}^{n,N}$ defined as any maximizer of the log-likelihood:

$$\widehat{\pi}^{n,N} \in \operatorname*{argmax}_{\pi \in \Pi} \{\ell^{n,N}(\pi)\} .$$

3 Round-robin graphical model

Section 3.1 details a graphical model encoding the conditional dependences between the random variables $(X^{n,N}, V)$. This graphical model is studied in the particular case of round-robin graphs in Section 3.2.

3.1 Graphical model

Let $d_0^{n,N}$ denote the graph distance in $(\{1,\dots,N\}, E^{n,N})$, that is $d_0^{n,N}(i,j)$ is the minimal length of a path between nodes i and j. Write $\{V_1,\dots,V_N\} = \cup_{q=0}^{N} V_q^{n,N}$, where $V_0^{n,N} = \{V_1\}$ and, for any $q \geq 1$, $V_q^{n,N}$ is the set of V_i such that $d_0^{n,N}(1,i) = q$. Let $\mathfrak{q}_N^n + 1$ denote the maximal distance between 1 and $i \in \{1,\dots,N\}$:

$$\mathfrak{q}_N^n + 1 = \max_{1 \leq i \leq N} d_0^{n,N}(1,i)\,.$$

- For all $1 \leq q \leq \mathfrak{q}_N^n + 1$, let

$$X_{q \leftrightarrow q}^{n,N} = \{X_{i,j} : \{i,j\} \in E^{n,N},\ i \in V_q^{n,N},\ j \in V_q^{n,N}\}\,.$$

  The set $X_{q \leftrightarrow q}^{n,N}$ gathers all $X_{i,j}$ such that i and j satisfy $d_0^{n,N}(V_1,V_i) = d_0^{n,N}(V_1,V_j) = q$.

- For all $0 \leq q \leq \mathfrak{q}_N^n$, let

$$X_{q \leftrightarrow q+1}^{n,N} = \{X_{i,j} : \{i,j\} \in E^{n,N},\ i \in V_q^{n,N},\ j \in V_{q+1}^{n,N}\}\,.$$

  Likewise, the set $X_{q \leftrightarrow q+1}^{n,N}$ gathers all $X_{i,j}$ such that $d_0^{n,N}(V_1,V_i) = q$ and $d_0^{n,N}(V_1,V_j) = q+1$.

Finally, for any $0 \leq q \leq \mathfrak{q}_N^n$, let

$$X_q^{n,N} = X_{q \leftrightarrow q+1}^{n,N} \cup X_{q+1 \leftrightarrow q+1}^{n,N}\,.$$

By (1) the joint distribution of $(V_q^{n,N})_{0 \leq q \leq \mathfrak{q}_N^n+1}$ and $(X_q^{n,N})_{0 \leq q \leq \mathfrak{q}_N^n}$ may be factorized using the conditional independence between some subsets of these variables. For all $0 \leq q \leq \mathfrak{q}_N^n$, and all $\pi \in \Pi$,

$$\mathbb{P}_\pi^{n,N}\left(X_q^{n,N} \middle| V, X_{0:q-1}^{n,N}\right) = \mathbb{P}_\pi^{n,N}\left(X_q^{n,N} \middle| V_q^{n,N}, V_{q+1}^{n,N}\right) = \prod_{\{i,j\}: X_{i,j} \in X_q^{n,N}} k\left(X_{i,j}, V_i, V_j\right)\,.$$

These conditional dependences are represented in the graphical model of Figure 1, where graph separations represent conditional independences. For all $0 \leq q \leq \mathfrak{q}_N^n$ any path between $X_q^{n,N}$ and other vertices except $V_q^{n,N}$ and $V_{q+1}^{n,N}$ goes through $V_q^{n,N}$ or $V_{q+1}^{n,N}$ which means that $X_q^{n,N}$ is independent of all other nodes given $V_q^{n,N}$ and $V_{q+1}^{n,N}$.

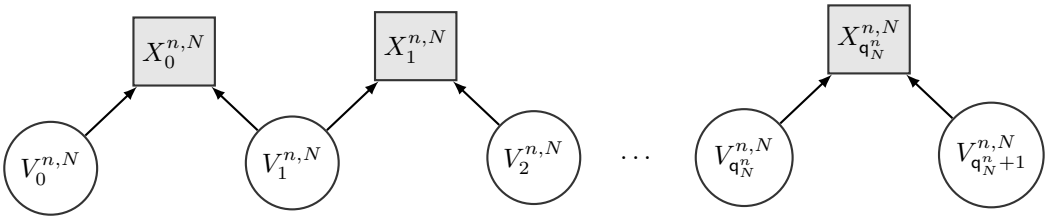
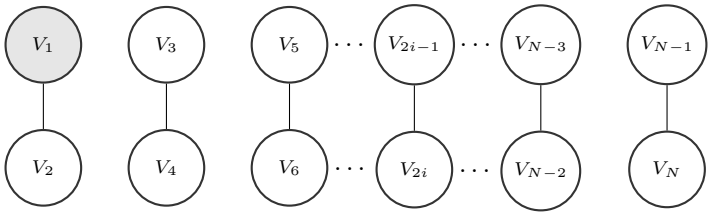


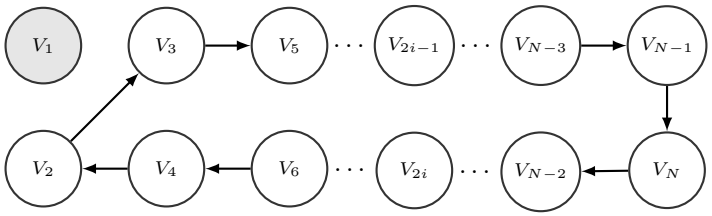
Figure 1: Graphical model of a paired comparisons based contest.

3.2 Round-Robin Scheduling

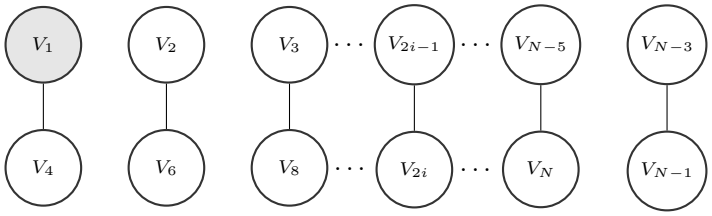
There is a large variety of n-regular graphs (even up to permutations of the indices), the results of this paper are obtained for the graph built using the round-robin scheduling. At time $t = 1$, this algorithm pairs nodes according to Figure 2a, that is $2i - 1$ is paired with $2i$, for all $i \in [N/2]$. At time $t = 2$, a node is fixed and all others are rotated clockwise as described in Figure 2b. Node 1 does not move, 2 takes the place of 3, each odd integer $2i - 1 < N - 1$ takes the place of $2i + 1$, $N - 1$ takes the place of N and each even integer $2i > 2$ takes the place of $2(i - 1)$. Then, each node is paired with the new node it faces as in Figure 2c. At each time $t > 2$, each node moves once according to the round-robin step detailed in Figure 2b and is paired with the new node it faces. The round-robin graph denoted by $E^{n,N}_{\mathrm{RR}}$ studied in detail in this paper contains all pairs collected in the first n rotations of the round-robin algorithm.



(a) Round-robin, $t = 1$.



(b) Round-robin moves.



(c) Round-robin, $t = 2$.

Figure 2: First two days using the round-robin algorithm.

Lemma 1 gathers results on the graphical model of Figure 1 when $E^{n,N} = E^{n,N}_{\mathrm{RR}}$ that are central in our analysis.

Lemma 1. *Let $n, N \geq 1$ and $(\{1, \ldots, N\}, E^{n,N}_{RR})$ be the round-robin graph. Assume that $2 \leq n <$*

$N/4$. Then, q_N^n is the quotient of the Euclidean division of $N/2-1$ by $n-1$, that is $N/2-1 = \mathsf{q}_N^n(n-1)+\mathsf{r}_N^n$ where $0 \le \mathsf{r}_N^n < n-1$. Moreover, $(V_q^{n,N}, X_q^{n,N})_{2\le q\le \mathsf{q}_N^n-1}$ is a stationary Markov chain such that for all $2 \le q \le \mathsf{q}_N^n - 1$,

$$\left|V_q^{n,N}\right| = 2(n-1)\,, \quad \left|X_q^{n,N}\right| = n(n-1)\,.$$

Lemma 1 is proved in Section 5.

4 Main results

Section 4.1 computes the limit likelihood function and define a natural risk function to evaluate the performances of the MLE. Risk bounds for the MLE are obtained in Section 4.2 using non-asymptotic concentration inequalities for Markov chains.

4.1 Convergence of the likelihood

The problem being reduced to the analysis of the graphical model of Figure 1, convergence results follow from the geometrically decaying mixing rates of the conditional law of $(V_k^{n,N})_{0\le k\le \mathsf{q}_N^n+1}$ given $X_{q:\mathsf{q}_N^n-1}^{n,N}$ for any $0 \le q < \mathsf{q}_N^n$. These rates derive from the following assumption.

H1 There exists $\varepsilon > 0$ such that for all $x \in \mathcal{X}, \pi \in \Pi \cup \{\pi_\star\}$ and $v_1, v_2 \in \mathrm{supp}(\pi)$, $k(x, v_1, v_2) \ge \varepsilon$.

When $E^{n,N} = E_{\mathrm{RR}}^{n,N}$, by Lemma 1, the joint sequence $(X_q^{n,N}, V_q^{n,N})_{2\le q\le \mathsf{q}_N^n-1}$ is a stationary Markov chain which may be extended to a stationary process indexed by $\mathbb{Z}$ with the same transition kernel. This extension is denoted by $(\mathbf{X^n}, \mathbf{V^n})$.

Define also the shift operator ϑ on $(\mathcal{X}^{n(n-1)})^{\mathbb{Z}}$ by $(\vartheta x)_k = x_{k+1}$ for all $k \in \mathbb{Z}$ and all $x \in (\mathcal{X}^{n(n-1)})^{\mathbb{Z}}$.

Theorem 2. *Assume H1 holds and $(\{1,\dots,N\}, E_{RR}^{n,N})$ is the round-robin graph. There exists a function ℓ_π^n such that for all $q \ge 2$,*

$$\sup_{\pi\in\Pi} \left|\log \mathbb{P}_\pi^{n,N}(X_q^{n,N} | X_{q+1:\mathsf{q}_N^n-1}^{n,N}) - \ell_\pi^n(\vartheta^q \mathbf{X^n})\right| \underset{N\to\infty}{\longrightarrow} 0, \qquad \mathbb{P}_{\pi_\star}\text{-a.s.}\,. \tag{2}$$

Moreover, for all $\pi \in \Pi$, $\mathbb{P}_{\pi_\star}$-a.s. and in $\mathrm{L}^1(\mathbb{P}_{\pi_\star})$,

$$\frac{1}{\mathsf{q}_N^n} \log \mathbb{P}_\pi^{n,N}\left(X^{n,N}\right) \underset{N\to\infty}{\longrightarrow} \mathsf{L}_{\pi_\star}^n(\pi) = \mathbb{E}_{\pi_\star}\left[\ell_\pi^n(\mathbf{X^n})\right]\,. \tag{3}$$

Theorem 2 establishes convergence to the limit likelihood $\mathsf{L}_{\pi_\star}^n(\pi)$ when the number of nodes N goes to ∞ while n remains fixed. The rate of almost sure convergence q_N^n is proportional to N by Lemma 1. Eq (3) is the key to understand the definition of the risk function used in the next section. We proceed as in Vapnick's learning theory [27, 26] described now to establish a parallel with our framework. Let $Y, Y_1, \dots, Y_N$ denote i.i.d. observations in $\mathcal{Y}$, let F denote a set of parameters, and let $\ell : F \times \mathcal{Y} \to \mathbb{R}$ denote a loss function. The empirical risk minimizer is defined in this context by

$$\hat{f}_N^{\mathrm{ERM}} = \operatorname*{argmin}_{f\in F} \sum_{i=1}^N \ell(f, Y_i)\,.$$

If $\mathbb{E}[\ell(f, Y_1)] < \infty$ for all $f \in F$, the risk of any $f \in F$ is measured by the *excess risk* [19]

$$R(f) = \mathbb{E}\left[\ell(f, Y)\right] - \mathbb{E}\left[\ell(f^*, Y)\right] ,$$

where Y is a copy of Y_1, independent of $Y_1, \dots, Y_N$ and f^* is the minimizer of $\mathbb{E}[\ell(f, Y)]$ over F. Note that when $\mathbb{E}[\ell(f, Y_1)] < \infty$, for all $f \in F$ the normalized empirical criterion satisfies almost surely,

$$\frac{1}{N}\sum_{i=1}^{N} \ell(f, Y_i) \to \mathbb{E}[\ell(f, Y_1)] .$$

Therefore the excess risk $R(f)$ is the difference between normalized asymptotic empirical criteria in f and its minimizer. In this paper, the MLE minimizes $-\log \mathbb{P}_{\pi}^{n,N}\left(X^{n,N}\right)$, which, properly normalized converges to $-\mathsf{L}_{\pi_\star}^{n}(\pi)$. This suggests to define the risk function

$$R_{\pi_\star}^{n}(\pi) = \mathsf{L}_{\pi_\star}^{n}(\pi_\star) - \mathsf{L}_{\pi_\star}^{n}(\pi), \qquad \forall \pi \in \Pi . \tag{4}$$

By Proposition 13, $\pi_\star$ is actually a minimizer of $-\mathsf{L}_{\pi_\star}^{n}(\pi)$ over $\Pi \cup \{\pi_\star\}$. Therefore, $R_{\pi_\star}^{n}$ is the excess risk associated with the likelihood function.

4.2 Risk bounds for the MLE

The following theorem provides non-asymptotic deviation bounds for the excess risk of the MLE. This is the main result of this paper.

Theorem 3. *Assume H1 holds and* $(\{1, \dots, N\}, E_{RR}^{n,N})$ *is the round-robin graph. For any probability measures* π *and* π'*, let*

$$d(\pi, \pi') = \begin{cases} \|\pi - \pi'\|_{\mathsf{tv}} \log\left(\frac{1}{\|\pi - \pi'\|_{\mathsf{tv}}}\right) & \text{if } \|\pi - \pi'\|_{\mathsf{tv}} \le e^{-1} , \\ \|\pi - \pi'\|_{\mathsf{tv}} & \text{if } \|\pi - \pi'\|_{\mathsf{tv}} \ge e^{-1} . \end{cases} \tag{5}$$

Assume that Π *is a compact set for the topology induced by* d *and let* $\mathsf{N}(\Pi \cup \{\pi_\star\}, d, \epsilon)$ *be the minimal number of balls of* d*-radius* ϵ *necessary to cover* $\Pi \cup \{\pi_\star\}$*. Then, there exists* $c > 0$ *such that, for any* $t > 0$*,*

$$\mathbb{P}_{\pi_\star}^{n,N}\left(R_{\pi_\star}^{n}(\widehat{\pi}^{n,N}) > \frac{cn\varepsilon^{-6n^2}}{\sqrt{N}}\left[\int_0^{+\infty} \sqrt{\log \mathsf{N}(\Pi \cup \{\pi_\star\}, d, \epsilon)}\mathrm{d}\epsilon + t\right]\right) \le e^{-t^2} .$$

Theorem 3 is proved in Section 7.3. It provides the first non-asymptotic risk bounds for any estimator in a very sparse setting where the number of edges n observed for each node can be very small compared to the number of nodes N. It proves that the problem studied in this paper is fundamentally different from the problem of nodes weights estimation that is usually considered, at least in Bradley-Terry models. While estimating nodes weights is only possible when n is as large as N [31, 22, 30], some information on their distribution may be recovered when $n \ll N$. This difference is extremely relevant in sports tournaments for example, it means that one can start to make prediction on the final issue of a championship only after a few weeks, while predictions on the issue of each game can only be made when half the year has passed.

The distance d defined in (5) used to measure the entropy of Π is not intuitive. However, it is easy to check that $d(\pi, \pi') \le C_\alpha \|\pi - \pi'\|_{\mathrm{tv}}^{1-\alpha}$ for any $\alpha > 0$. It follows that, for any class Π with

polynomial entropy for the total variation distance, that is such that $\mathsf{N}(\Pi \cup \{\pi_\star\}, \|\cdot\|_{\mathrm{tv}}, \epsilon) \lesssim \epsilon^D$ for small ϵ, Dudley's entropy integral for distance d satisfies

$$\int_0^{+\infty} \sqrt{\log \mathsf{N}(\Pi \cup \{\pi_\star\}, d, \epsilon)} \mathrm{d}\epsilon \lesssim_\alpha \sqrt{D}\,.$$

Therefore, "slow rates" of convergence are obtained for the MLE. The polynomial growth $\mathsf{N}(\Pi \cup \{\pi_\star\}, \|\cdot\|_{\mathrm{tv}}, \epsilon) \lesssim \epsilon^D$ is extremely standard, see [25, p271–274] for various examples where this assumption is satisfied and our result applies. On the other hand, "fast" rates of convergence remain an open question. In particular, the margin condition [18] required to prove such rates would hold if the total variation distance between distributions of the nodes weights was bounded from above by the excess risk derived from the asymptotic of the likelihood.

The remaining of the paper is devoted to the proof of the main results. Section 5 proves Lemma 1, describing precisely the structure of the graphical model given in Figure 1 in the case of a round-robin scheduling. Then, Section 6 establishes central tools in the analysis of the likelihood of stationary processes whose conditional dependences are encoded in the graphical model of Figure 1. These results, that might be of independent interest, are therefore stated as independent lemmas. These tools are finally used in Section 7 to prove the main theorems.

5 Round-robin scheduling

This section details the sets $V_q^{n,N}$ and $X_q^{n,N}$ for $0 \le q \le \mathsf{q}_N^n + 1$ when $E^{n,N} = E_{\mathrm{RR}}^{n,N}$ (cf. Figures 2a-2c). In the following, notations for nodes and their weights are identified, i.e. i is identified with V_i for all $1 \le i \le N$. Lemma 1 follows directly from Lemma 4 and Lemma 5 below. To prove these lemmas, consider the following notations.

$$\mathcal{E} = \{4x - 1, 4x : x \in [\lfloor N/4 \rfloor]\} \quad \text{and} \quad \mathcal{O} = [N] \setminus \mathcal{E}\,.$$

The notation $\mathcal{E}$ (resp $\mathcal{O}$) comes from the fact that $\mathcal{E}$ (resp $\mathcal{O}$) contains all i paired with 1 after an *even* (resp *odd*) number $n \le N/4$ of rotations of the round-robin scheduling. For all $1 \le q \le \mathsf{q}_N^n$, let

$$V_{q,e}^{n,N} = V_q^{n,N} \cap \mathcal{E} \quad \text{and} \quad V_{q,o}^{n,N} = V_q^{n,N} \cap \mathcal{O}\,.$$

Lemma 4. *Let $n, N \ge 1$ and $(\{1, \dots, N\}, E_{RR}^{n,N})$ be the round-robin graph. Assume that $2 \le n < N/4$ and let $N/2 - 1 = \mathsf{q}_N^n(n-1) + \mathsf{r}_N^n$ where $0 \le \mathsf{r}_N^n < n - 1$. Then,*

$$V_1^{n,N} = \{V_{2x} : x = 1, \dots, n\}\,, \tag{6}$$

and, for any $2 \le q \le \mathsf{q}_N^n$,

$$V_q^{n,N} = \{V_{2x+1} : x \in [(q-2)(n-1) + 1, (q-1)(n-1)]\} \cup \{V_{2x} : x \in [2 + (q-1)(n-1), 1 + q(n-1)]\}\,. \tag{7}$$

Furthermore,

$$V_{\mathsf{q}_N^n+1}^{n,N} = \{V_{2x+1} : x \in [(\mathsf{q}_N^n - 1)(n-1) + 1, \mathsf{q}_N^n(n-1) + \mathsf{r}_N^n]\} \cup \{V_{2x} : x \in [2 + \mathsf{q}_N^n(n-1), 1 + \mathsf{r}_N^n + \mathsf{q}_N^n(n-1)]\}\,. \tag{8}$$

Therefore, $|V_0^{n,N}| = 1$, $|V_1^{n,N}| = n$ and for all $2 \le q \le \mathsf{q}_N^n$, $\left|V_q^{n,N}\right| = 2(n-1)$.

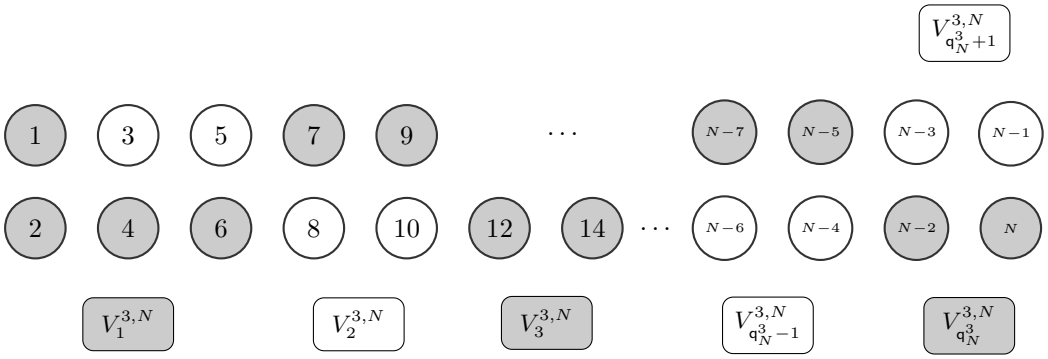


Figure 3: Elements of $\mathcal{V}^{3,N}$, case $\mathsf{r}^3_N = 0$.

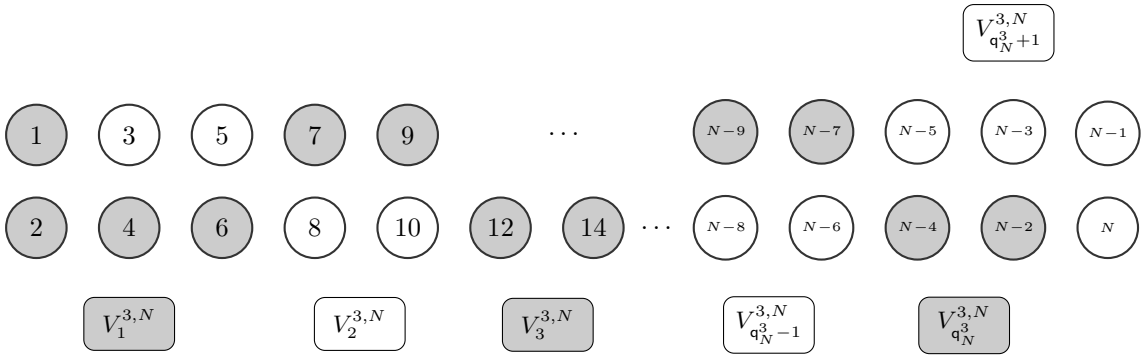


Figure 4: Elements of $\mathcal{V}^{3,N}$, case $\mathsf{r}^3_N = 1$.

Proof. To ease the reading of this proof, one can check its arguments on Figures 3 and 4.
We proceed by induction on q. The definition of $V_1^{n,N}$ given by (6) is straightforward. Then, $V_2^{n,N}$ contains:

- all V_i paired with some $V_j \in V_1^{n,N}$ on the first rotation of the algorithm besides V_1 that does not belong to $V_2^{n,N}$. These are all $\{V_{2x+1} : x = 1, \dots, n-1\}$;
- All V_i paired with V_2 and V_4 that are not in $V_0^{n,N} \cup V_1^{n,N}$. After n rotations of the round-robin algorithms, all V_i paired with V_2 are $\{V_1, V_{4x+2} : x = 1, \dots, n-1\}$ and those with V_4 are $\{V_1, V_3, V_{4x} : x = 2, \dots, n-2\}$.

Therefore,

$$V_2^{n,N} \supset \{V_{2x+1} : x = 1, \dots, n-1\} \cup \{V_{2x} : x = n+1, \dots, 2n-1\}\,.$$

On the other hand, by induction, for all $i \notin \{N-2x+1, x = 1, \dots, 2(n-1)\} \cup \{2x : x = 1, \dots, 2n-1\}$,

$$\begin{aligned} &\text{if } i \text{ is odd, it is paired with } \{V_{i+4x+1} : x = 0, \dots n-1\}\,, \\ &\text{if } i \text{ is even, it is paired with } \{V_{i-4x-1} : x = 0, \dots, n-1\}\,. \end{aligned} \tag{9}$$

This implies that there is no even number $i \geq 4n$ nor odd number $i > 2n-1$ such that $V_i \in V_2^{n,N}$, which yields:

$$V_2^{n,N} = \{V_{2x+1} : x = 1, \dots, n-1\} \cup \{V_{2x} : x = n+1, \dots, 2n-1\}\,.$$

(7) is obtained by induction using the same arguments and (8) is a direct consequence of the round-robin algorithm. The last claim follows by noting that for all $q \in [2, \mathsf{q}_N^n]$,

$$|V_{q,e}^{n,N}| = |V_{q,o}^{n,N}| = n-1\,.$$

Indeed, one of the following cases holds.

- $n-1 = 2p$ for some $p \in \mathbb{N}$. In this case,

$$|\{j : V_j \in V_{q,e}^{n,N}, j \in 2\mathbb{Z}\}| = |\{i : V_i \in V_{q,e}^{n,N}, i \in 2\mathbb{Z}+1\}| = p\,.$$

- $n-1 = 2p+1$ for some $p \in \mathbb{N}$. In this case, either

$$|\{j : V_j \in V_{q,e}^{n,N}, j \in 2\mathbb{Z}\}| = p, \qquad \text{and} \qquad |\{i : V_i \in V_{q,e}^{n,N}, i \in 2\mathbb{Z}+1\}| = p+1\,,$$

or

$$|\{j : V_j \in V_{q,e}^{n,N}, j \in 2\mathbb{Z}\}| = p+1, \qquad \text{and} \qquad |\{i : V_i \in V_{q,e}^{n,N}, i \in 2\mathbb{Z}+1\}| = p\,.$$

□

Lemma 5. *Let $n, N \geq 1$ and $(\{1, \dots, N\}, E_{RR}^{n,N})$ be the round-robin graph. Then, for all $2 \leq q \leq \mathsf{q}_N^n - 1$,*

$$|X_q^{n,N}| = n(n-1)\,.$$

Proof. The proof essentially consists in building the graphical model of Figure 5 from the one displayed in Figure 1.
Edges involving the first node are decomposed as:

$$X_{0\leftrightarrow 1,e}^{n,N} = \{X_{1,4x} : x = 1, \dots, \lfloor n/2 \rfloor\} = \{X_{1,i} : V_i \in V_{1,e}^{n,N}\} \quad \text{and} \quad X_{0\leftrightarrow 1,o}^{n,N} = \{X_{1,i} : V_i \in V_{1,o}^{n,N}\}\,.$$

Edges involving nodes in $V_1^{n,N}$ that are both different from 1 are described as follows.

- Edges between two nodes in $V_1^{n,N}$ denoted by:

$$X_{1\leftrightarrow 1,e}^{n,N} = \{X_{4x,4y} : (x,y) \in [\lfloor n/2 \rfloor], x < y\} = \{X_{i,j} : V_i, V_j \in V_{1,e}^{n,N}, i < j\}\,,$$
$$X_{1\leftrightarrow 1,o}^{n,N} = \{X_{i,j} : V_i, V_j \in V_{1,o}^{n,N}, i < j\}\,.$$

  Note that there is no edge between any $V_i \in V_{1,e}^{n,N}$ and a node $V_j \in V_{q,o}^{n,N}$ for any $q \geq 1$. In particular, there is no edge between any $V_i \in V_{1,e}^{n,N}$ and $V_j \in V_{1,o}^{n,N}$. Therefore, $X_{1\leftrightarrow 1,e}^{n,N} \cup X_{1\leftrightarrow 1,o}^{n,N}$ describes all edges between nodes in $V_1^{n,N}$.

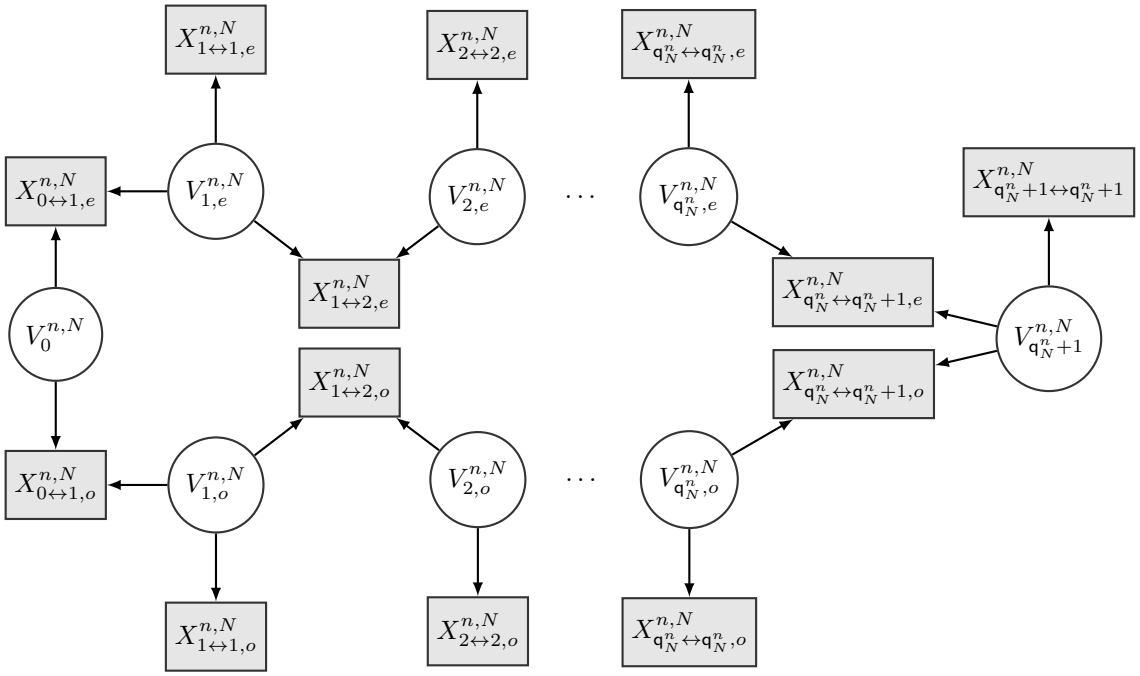


Figure 5: Graphical model of the round-robin algorithm.

- Edges between $V_i \in V_1^{n,N}$ and $V_j \in V_2^{n,N}$ are described as follows:

$$\begin{aligned} X^{n,N}_{1\leftrightarrow 2,e} &= \{X_{4y-1-4k,4y} : y \in [\lfloor n/2 \rfloor], k < y\} \cup \{X_{4x,4y} : x \in [\lfloor n/4 \rfloor], y \in [\lfloor n/2 \rfloor + 1, n - x]\} \\ &= \{X_{i,j} : V_i \in V^{n,N}_{1,e}, V_j \in V^{n,N}_{2,e}, j \in 2\mathbb{Z}+1, j > i\} \\ &\qquad \cup \{X_{i,j} : V_i \in V^{n,N}_{1,e}, V_j \in V^{n,N}_{2,e}, j \in 2\mathbb{Z} \cap [4n-i]\}, \\ X^{n,N}_{1\leftrightarrow 2,o} &= \{X_{i,j} : V_i \in V^{n,N}_{1,o}, V_j \in V^{n,N}_{2,o}, j \in 2\mathbb{Z}+1, j > i\} \\ &\qquad \cup \{X_{i,j} : V_i \in V^{n,N}_{1,o}, V_j \in V^{n,N}_{2,o}, j \in 2\mathbb{Z} \cap [4n-i]\}. \end{aligned}$$

By (9), for any $q \in [2, \mathsf{q}^n_N]$, edges between V_i and V_j both in $V^{n,N}_q$ are:

$$\begin{aligned} X^{n,N}_{q\leftrightarrow q,e} &= \{X_{i,j} : V_i \in V^{n,N}_{q,e}, i \in 2\mathbb{Z}+1, V_j \in V^{n,N}_{q,e}, j \in 2\mathbb{Z}\}, \\ X^{n,N}_{q\leftrightarrow q,o} &= \{X_{i,j} : V_i \in V^{n,N}_{q,o}, i \in 2\mathbb{Z}+1, V_j \in V^{n,N}_{q,o}, j \in 2\mathbb{Z}\}. \end{aligned}$$

Note that (9) shows also that there is no edge between $V_i \in V^{n,N}_{q,e}$ and $V_j \in V^{n,N}_{q,o}$. For all $2 \leq q \leq \mathsf{q}^n_N$

and all $V_i \in V_q^{n,N}$ and $V_j \in V_{q+1}^{n,N}$,

$$\begin{aligned}
X_{q\leftrightarrow q+1,e}^{n,N} =& \{X_{i,j} : V_i \in V_{q,e}^{n,N}, i \in (2\mathbb{Z}+1), V_j \in V_{q+1,e}^{n,N}, j \in 2\mathbb{Z} \cap [i+4n-3]\} \\
& \cup \{X_{i,j} : V_i \in V_{q,e}^{n,N}, i \in 2\mathbb{Z}, V_j \in V_{q+1,e}^{n,N}, j \in 2\mathbb{Z}+1 \cap [i]\}, \\
X_{q\leftrightarrow q+1,o}^{n,N} =& \{X_{i,j} : V_i \in V_{q,o}^{n,N}, i \in (2\mathbb{Z}+1), V_j \in V_{q+1,o}^{n,N}, j \in 2\mathbb{Z} \cap [i+4n-3]\} \\
& \cup \{X_{i,j} : V_i \in V_{q,o}^{n,N}, i \in 2\mathbb{Z}, V_j \in V_{q+1,o}^{n,N}, j \in (2\mathbb{Z}+1) \cap [i]\}.
\end{aligned}$$

Therefore, for all $2 \leq q \leq \mathsf{q}_N^n$,

$$\begin{aligned}
|X_{q\leftrightarrow q,e}^{n,N}| &= |\{i : V_i \in V_{q,e}^{n,N}, i \in 2\mathbb{Z}+1\}||\{j : V_j \in V_{q,e}^{n,N}, j \in 2\mathbb{Z}\}| \\
&= \begin{cases} p^2 & \text{if } n-1 = 2p, \\ p(p+1) & \text{if } n-1 = 2p+1. \end{cases}
\end{aligned}$$

The same holds for $|X_{q\leftrightarrow q,o}^{n,N}|$ so that $|X_{q\leftrightarrow q}^{n,N}| = 2p^2$ if $n-1 = 2p$ and $|X_{q\leftrightarrow q}^{n,N}| = 2p(p+1)$ if $n-1 = 2p+1$. On the other hand,

$$\begin{aligned}
|X_{q\leftrightarrow q+1,e}^{n,N}| =& \sum_{i:V_i \in V_{q,e}^{n,N}, i \in (2\mathbb{Z}+1)} |\{j : V_j \in V_{q+1,e}^{n,N} j \in 2\mathbb{Z} \cap [i+4n-3]\} \\
& + \sum_{i:V_i \in V_{q,e}^{n,N}, i \in 2\mathbb{Z}} |\{j : V_j \in V_{q+1,e}^{n,N}, j \in 2\mathbb{Z}+1 \cap [i]\}| \\
=& \begin{cases} 2\sum_{i=1}^{p} i = p(p+1) & \text{if } n-1 = 2p, \\ \sum_{i=1}^{p} i + \sum_{i=1}^{p+1} i = (p+1)^2 & \text{if } n-1 = 2p+1. \end{cases}
\end{aligned}$$

As the same holds for $|X_{q\leftrightarrow q+1,o}^{n,N}|$, $|X_{q\leftrightarrow q+1}^{n,N}| = 2p(p+1)$ if $n-1 = 2p$ and $|X_{q\leftrightarrow q+1}^{n,N}| = 2(p+1)^2$ if $n-1 = 2p+1$. The proof is completed by writing $|X_q^{n,N}| = |X_{q\leftrightarrow q+1}^{n,N}| + |X_{q+1\leftrightarrow q+1}^{n,N}|$. □

6 Probabilistic study of the graphical model

This section analyses stochastic processes whose conditional dependences are encoded in the graphical model of Figure 1. To ease applications of these general results to our problem, we focus on a restricted class of such stochastic processes.

Let $\mathsf{n} \in \mathbb{N} \setminus \{0\}$, π_V be a distribution on a measurable space $\mathbb{V}$ and $\mathbb{X}$ be a discrete space. Let K_i denote non-negative functions defined on $\mathbb{X} \times \mathbb{V}^2$ such that all $K_i(., v, w)$ are probability distributions on $\mathbb{X}$. Let $\mathbb{P}_{\pi_V}$ be the distribution on $\mathbb{V}^{\mathsf{n}+1} \times \mathbb{X}^{\mathsf{n}}$ defined by:

$$\mathbb{P}_{\pi_V}\left(V_{1:\mathsf{n}+1} \in A_{1:\mathsf{n}+1}, X_{1:\mathsf{n}} = x_{1:\mathsf{n}}\right) = \int \prod_{i=1}^{\mathsf{n}+1} \mathbb{1}_{A_i}(v_i) \prod_{i=1}^{\mathsf{n}+1} \pi_V(\mathrm{d}v_i) \prod_{i=1}^{\mathsf{n}} K_i(x_i, v_i, v_{i+1}). \tag{10}$$

The random variables $(V_i)_{i\in\{1,\ldots,\mathsf{n}+1\}}$ are i.i.d. taking values in $\mathbb{V}$ with common distribution π_V and $(X_i)_{i\in\{1,\ldots\mathsf{n}\}}$ is a stochastic process taking values in a discrete set $\mathbb{X}$ such that $(X_i)_{i\in\{1,\ldots,\mathsf{n}\}}$ are independent conditionally on V and

$$\mathbb{P}_{\pi_V}(X_i = x | V_{1:\mathsf{n}+1}) = \mathbb{P}_{\pi_V}(X_i = x | V_i, V_{i+1}) = K_i(x, V_i, V_{i+1}), \qquad \forall i \in \{1, \mathsf{n}\}, \forall x \in \mathbb{X}.$$

Therefore, $\mathbb{P}_{\pi_V}$ is a generic probability distribution with conditional dependences encoded by the graphical model of Figure 1. In this section, the following assumption is always granted.

H2 There exist $\nu_i > 0$ such that

$$\nu_i \leq K_i(x, v, w) \leq 1, \qquad \forall x \in \mathbb{X}, \forall i \in \mathbb{Z}, \forall v, w \in \mathbb{V}\,. \tag{11}$$

For some results, the following assumption is required.

$$\forall i \in \{1, \ldots, \mathsf{n}\}, \qquad K_i = K\,. \tag{12}$$

Whenever Assumption (12) holds, we shall denote by ν a real number such that

$$\nu \leq K(x, v, w) \leq 1, \qquad \forall x \in \mathbb{X}, \forall v, w \in \mathbb{V}\,.$$

Note that by (10), the sequence $(V_{k+1}, X_k)_{k\geq 0}$ is a Markov chain with transition kernel on $\mathbb{V} \times \mathbb{X}$ such that:

$$\mathbb{P}_{\pi_V}\left(V_{k+1} \in A, X_k | V_k, X_{k-1}\right) = \int \mathbb{1}_A(v_{k+1})\pi_V(\mathrm{d}v_{k+1})K_k(X_k, V_k, v_{k+1}) \geq \nu_k \pi_V(A)\,.$$

This uniform minorization condition ensures that the joint Markov chain $(V_{k+1}, X_k)_{k\geq 0}$ is geometrically ergodic and admits the whole space $\mathbb{V} \times \mathbb{X}$ as petite set. Note also that, as defined by (10), $\mathbb{P}_{\pi_V}$ is the law of this Markov chain started from stationarity, the stationary distribution on $\mathbb{V} \times \mathbb{X}$ being $(A, x_0) \mapsto \int \mathbb{1}_A(v_1)\pi_V(\mathrm{d}v_1)\pi_V(\mathrm{d}v_0)k(x_0, v_0, v_1)$.

Lemma 6 first shows that, conditionally on the observations, $V_1, \ldots, V_{\mathsf{n}}$ is a backward Markov chain admitting the all state space as petite set.

Lemma 6. *For any $q \geq 1$, conditionally on $X_{q:\mathsf{n}}$, $(V_{\mathsf{n}}, \ldots, V_1)$ is a Markov chain. Its transition kernels $(K^{V|X}_{\pi_V,k,q})_{q\leq k<\mathsf{n}}$ are such that, for all $q \leq k < \mathsf{n}$, there exists a measure $\mu_{k,q}$ satisfying for all measurable set A:*

$$K^{V|X}_{\pi_V,k,q}(V_{k+1}, A) = \mathbb{P}_{\pi_V}\left(V_k \in A | V_{k+1:\mathsf{n}}, X_{q:\mathsf{n}}\right) = \mathbb{P}_{\pi_V}\left(V_k \in A | V_{k+1}, X_{q:\mathsf{n}}\right) \geq \nu_k \mu_{k,q}(A)\,.$$

On the other hand, for all $1 \leq k < q$,

$$K^{V|X}_{\pi_V,k,q}(V_{k+1}, A) = \mathbb{P}_{\pi_V}\left(V_k \in A | V_{k+1:\mathsf{n}}, X_{q:\mathsf{n}}\right) = \pi_V(A)\,.$$

Proof. The Markov property is immediate. The case $1 \leq k < q$ follows from the independence of V_k and $(V_{k+1:\mathsf{n}}, X_{q:\mathsf{n}})$. Then, for any $q \leq k < \mathsf{n}$ and all measurable set A,

$$\begin{aligned}\mathbb{P}_{\pi_V}\left(V_k \in A | V_{k+1:\mathsf{n}}, X_{q:\mathsf{n}}\right) &= \mathbb{P}_{\pi_V}\left(V_k \in A | V_{k+1}, X_{q:k}\right) \\ &= \frac{\int \mathbb{1}_A(v_k)\pi_V(\mathrm{d}v_k)K_k(X_k, v_k, v_{k+1})\mathbb{P}_{\pi_V}(X_{q:k-1}|v_k)}{\int \pi_V(\mathrm{d}v_k)K_k(X_k, v_k, v_{k+1})\mathbb{P}_{\pi_V}(X_{q:k-1}|v_k)}\,,\end{aligned}$$

with the conventions $\mathbb{P}_{\pi_V}(X_{q:q-1}|V_q) = 1$. By Assumption H-2,

$$\mathbb{P}_{\pi_V}\left(V_k \in A | V_{k+1}, X_{q:\mathsf{n}}\right) \geq \nu_k \frac{\int \mathbb{1}_A(v_k)\pi_V(\mathrm{d}v_k)\mathbb{P}_{\pi_V}(X_{q:k-1}|v_k)}{\int \pi_V(\mathrm{d}v_k)\mathbb{P}_{\pi_V}(X_{q:k-1}|v_k)}\,.$$

The proof is then completed by choosing:

$$\mu_{k,q}(A) = \frac{\int \mathbb{1}_A(v_k)\pi_V(\mathrm{d}v_k)\mathbb{P}_{\pi_V}(X_{q:k-1}|v_k)}{\int \pi_V(\mathrm{d}v_k)\mathbb{P}_{\pi_V}(X_{q:k-1}|v_k)}\,.$$

□

Lemma 7 shows the contraction properties of the Markov kernel of the chain V conditionally on the observations. It is a direct consequence of the minoration condition given in Lemma 6, see for instance [17, Sections III.9 to III.11] or [3, Corollary 4.3.9 and Lemma 4.3.13]. Let $\|\cdot\|_{\mathsf{tv}}$ be the total variation norm defined, for any measurable set $(\mathsf{Z}, \mathcal{Z})$ and any finite signed measure ξ on $(\mathsf{Z}, \mathcal{Z})$, by

$$\|\xi\|_{\mathsf{tv}} = \sup\left\{\int f(z)\xi(\mathrm{d}z)\ ;\ f \text{ measurable real function on } \mathsf{Z} \text{ such that } \|f\|_\infty = 1\right\}.$$

Lemma 7. *For all probability measures μ_1, μ_2 and all $1 \le q \le k < \mathsf{n}$,*

$$\left\|\int \mu_1(\mathrm{d}v) K^{V|X}_{\pi_V,k,q}(v,\cdot) - \int \mu_2(\mathrm{d}v) K^{V|X}_{\pi_V,k,q}(v,\cdot)\right\|_{\mathsf{tv}} \le (1-\nu_k)\,\|\mu_1-\mu_2\|_{\mathsf{tv}} \le (1-\nu_k)\ .$$

In particular, by induction,

$$\left\|\int \{\mu_1(\mathrm{d}v_{\mathsf{n}}) - \mu_2(\mathrm{d}v_{\mathsf{n}})\}\, K^{V|X}_{\pi_V,\mathsf{n}-1,q}(v_{\mathsf{n}},\mathrm{d}v_{\mathsf{n}-1})\ldots K^{V|X}_{\pi_V,k,q}(v_{k+1},\cdot)\right\|_{\mathsf{tv}} \le \prod_{i=k}^{\mathsf{n}-1}(1-\nu_i)\ . \tag{13}$$

Lemma 8 proves a key loss of memory property of the backward chain X_q, with geometric rate of convergence. Whenever it is necessary, we adopt the convention $\prod_{k=\ell}^{m} = 1$ for any $\ell > m$.

Lemma 8. *For any $1 \le q \le \mathsf{n}-1$,*

$$\left|\log \mathbb{P}_{\pi_V}\left(X_q | X_{q+1:\mathsf{n}}\right)\right| \le \log\left(\nu_q^{-1}\right)\ . \tag{14}$$

For all $\ell \ge 1$, $1 \le q \le \mathsf{n}-1$,

$$\left|\log \mathbb{P}_{\pi_V}\left(X_q | X_{q+1:\mathsf{n}}\right) - \log \mathbb{P}_{\pi_V}\left(X_q | X_{q+1:\mathsf{n}+\ell}\right)\right| \le \nu_q^{-1} \prod_{k=q+1}^{\mathsf{n}-1}(1-\nu_k)\,. \tag{15}$$

Proof. To prove (15), for $1 \le q < \mathsf{n}$, note that by Lemma 6,

$$\mathbb{P}_{\pi_V}\left(X_q | X_{q+1:\mathsf{n}}\right) = \int \mathbb{P}_{\pi_V}\left(\mathrm{d}v_{\mathsf{n}} | X_{q+1:\mathsf{n}}\right)\left(\prod_{k=q+1}^{\mathsf{n}-1} K^{V|X}_{\pi_V,k,q+1}(v_{k+1},\mathrm{d}v_k)\right)\pi_V(\mathrm{d}v_q)K_q(X_q,v_q,v_{q+1})\,. \tag{16}$$

Inequality (14) follows from (16). Likewise,

$$\mathbb{P}_{\pi_V}\left(X_q | X_{q+1:\mathsf{n}+\ell}\right) = \int \mathbb{P}_{\pi_V}\left(\mathrm{d}v_{\mathsf{n}} | X_{q+1:\mathsf{n}+\ell}\right)\left(\prod_{k=q+1}^{\mathsf{n}-1} K^{V|X}_{\pi_V,k,q+1}(v_{k+1},\mathrm{d}v_k)\right)\pi_V(\mathrm{d}v_q)K_q(X_q,v_q,v_{q+1})\,. \tag{17}$$

Then, by Lemma 7, combining (16) and (17) yields:

$$\left|\mathbb{P}_{\pi_V}\left(X_q | X_{q+1:\mathsf{n}+\ell}\right) - \mathbb{P}_{\pi_V}\left(X_q | X_{q+1:\mathsf{n}}\right)\right| \le \left(\prod_{k=q+1}^{\mathsf{n}-1}(1-\nu_k)\right)\sup_{v_{q+1}\in\mathbb{V}}\left|\int \pi_V(\mathrm{d}v_q)K_q(X_q,v_q,v_{q+1})\right| \le \prod_{k=q+1}^{\mathsf{n}-1}(1-\nu_k)\,.$$

(15) is then a direct consequence of (16), (17) and the fact that for all $x, y > 0$, $|\log x - \log y| \leq |x-y|/x \wedge y$. □

Lemma 9 is the crucial result to bound the increments of the log-likelihood.

Lemma 9. *For all distributions $\pi_V, \pi'_V \in \Pi \cup \{\pi^\star\}$ and any $1 \leq q \leq \mathsf{n}$,*

$$\left|\log \mathbb{P}_{\pi_V}(X_q|X_{q+1:\mathsf{n}}) - \log \mathbb{P}_{\pi'_V}(X_q|X_{q+1:\mathsf{n}})\right| \leq 2 \sum_{\ell=0}^{\mathsf{n}+1-q} (\nu_q \nu_{q+\ell-1} \nu_{q+\ell})^{-1} \left(\prod_{k=q+1}^{q+\ell-1} (1-\nu_k)\right) \|\pi_V - \pi'_V\|_{\mathsf{tv}} .$$

Proof. When $q = \mathsf{n}$,

$$\mathbb{P}_{\pi_V}(X_\mathsf{n}) - \mathbb{P}_{\pi'_V}(X_\mathsf{n}) = \int \left\{\pi_V'^{\otimes 2}(\mathrm{d}v_{\mathsf{n}:\mathsf{n}+1}) - \pi_V^{\otimes 2}(\mathrm{d}v_{\mathsf{n}:\mathsf{n}+1})\right\} K_\mathsf{n}(X_\mathsf{n}, v_\mathsf{n}, v_{\mathsf{n}+1}) .$$

Thus $|\mathbb{P}_{\pi_V}(X_\mathsf{n}) - \mathbb{P}_{\pi'_V}(X_\mathsf{n})| \leq 2\|\pi_V - \pi'_V\|_{\mathsf{tv}}$. When $1 \leq q \leq \mathsf{n}-1$,

$$\mathbb{P}_{\pi_V}(X_q|X_{q+1:\mathsf{n}}) - \mathbb{P}_{\pi'_V}(X_q|X_{q+1:\mathsf{n}}) = \sum_{\ell=0}^{\mathsf{n}+1-q} \left\{\mathbb{P}_\ell(X_q|X_{q+1:\mathsf{n}}) - \mathbb{P}_{\ell+1}(X_q|X_{q+1:\mathsf{n}})\right\} ,$$

where $\mathbb{P}_\ell$ is the joint distribution of $(X_{q:\mathsf{n}}, V_{q:\mathsf{n}+1})$ when $(V_q, \ldots, V_{q+\ell-1})$ are i.i.d. π'_V and $(V_{q+\ell}, \ldots, V_{\mathsf{n}+1})$ are i.i.d. π_V. The first term in the telescopic sum is given by:

$$\mathbb{P}_0(X_q|X_{q+1:\mathsf{n}}) - \mathbb{P}_1(X_q|X_{q+1:\mathsf{n}}) = \int \mathbb{P}_0\left(\mathrm{d}v_{q+1}|X_{q+1:\mathsf{n}}\right) \int \pi_V(\mathrm{d}v_q) K_q(X_q, v_q, v_{q+1}) - \int \mathbb{P}_0\left(\mathrm{d}v_{q+1}|X_{q+1:\mathsf{n}}\right) \int \pi'_V(\mathrm{d}v_q) K_q(X_q, v_q, v_{q+1}) ,$$

where $\mathbb{P}_0\left(V_{q+1}|X_{q+1:\mathsf{n}}\right)$ is the distribution of V_{q+1} conditionally on $X_{q+1:\mathsf{n}}$ when $(V_q, \ldots, V_{\mathsf{n}+1})$ are i.i.d. π_V. As V_q is independent of $(V_{q+1}, X_{q+1:\mathsf{n}})$, this distribution is the same as the distribution of V_{q+1} conditionally on $X_{q+1:\mathsf{n}}$ when $V_q \sim \pi'_V$ and $(V_{q+1}, \ldots, V_{\mathsf{n}+1})$ are i.i.d. π_V.

$$|\mathbb{P}_0(X_q|X_{q+1:\mathsf{n}}) - \mathbb{P}_1(X_q|X_{q+1:\mathsf{n}})| \leq \|\pi_V - \pi'_V\|_{\mathsf{tv}} .$$

Then, for all $1 \leq \ell \leq \mathsf{n}+2-q$,

$$\mathbb{P}_\ell\left(X_q|X_{q+1:\mathsf{n}}\right) = \int \mathbb{P}_\ell\left(\mathrm{d}v_{q+\ell}|X_{q+1:\mathsf{n}}\right) \left(\prod_{k=q+1}^{q+\ell-1} K^{V|X}_{\pi'_V,k,q+1}(v_{k+1}, \mathrm{d}v_k)\right) \int \pi'_V(\mathrm{d}v_q) K_q(X_q, v_q, v_{q+1}) .$$

Therefore, by (13),

$$\left|\mathbb{P}_\ell\left(X_q|X_{q+1:\mathsf{n}}\right) - \mathbb{P}_{\ell+1}\left(X_q|X_{q+1:\mathsf{n}}\right)\right| \leq \left(\prod_{k=q+1}^{q+\ell-1} (1-\nu_k)\right) \left\|\mathbb{P}_\ell\left(V_{q+\ell}|X_{q+1:\mathsf{n}}\right) - \mathbb{P}_{\ell+1}\left(V_{q+\ell}|X_{q+1:\mathsf{n}}\right)\right\|_{\mathsf{tv}} ,$$

where $\mathbb{P}_\ell\left(V_{q+\ell}|X_{q+1:\mathsf{n}}\right)$ is the distribution of $V_{q+\ell}$ conditionally on $X_{q+1:\mathsf{n}}$ when $(V_q,\ldots,V_{q+\ell-1})$ are i.i.d. π'_V and $(V_{q+\ell},\ldots,V_{\mathsf{n}+1})$ are i.i.d. π_V. It remains to show that

$$\left\|\mathbb{P}_\ell\left(V_{q+\ell}|X_{q+1:\mathsf{n}}\right)-\mathbb{P}_{\ell+1}\left(V_{q+\ell}|X_{q+1:\mathsf{n}}\right)\right\|_{\mathsf{tv}}\leq 2(\nu_q\nu_{q+\ell-1}\nu_{q+\ell})^{-1}\|\pi_V-\pi'_V\|_{\mathsf{tv}}$$

which amounts to showing that for all f such that $\|f\|_\infty\leq 1$,

$$\left|\int f(v_{q+\ell})\left\{\mathbb{P}_\ell\left(\mathrm{d}v_{q+\ell}|X_{q+1:\mathsf{n}}\right)-\mathbb{P}_{\ell+1}\left(\mathrm{d}v_{q+\ell}|X_{q+1:\mathsf{n}}\right)\right\}\right|\leq 2(\nu_q\nu_{q+\ell-1}\nu_{q+\ell})^{-1}\|\pi_V-\pi'_V\|_{\mathsf{tv}}\,.$$

Write, for all $1\leq\ell\leq\mathsf{n}+2-q$,

$$\mathrm{L}_\ell(\mathrm{d}v,X)=\prod_{m=q+1}^{q+\ell-1}\pi'_V(\mathrm{d}v_m)\prod_{m=q+\ell}^{\mathsf{n}+1}\pi_V(\mathrm{d}v_m)\prod_{m=q+1}^{\mathsf{n}}K_m(X_m,v_m,v_{m+1})\,. \tag{18}$$

We have

$$\int f(v_{q+\ell})\mathbb{P}_\ell\left(\mathrm{d}v_{q+\ell}|X_{q+1:\mathsf{n}}\right)=\frac{\int f(v_{q+\ell})L_\ell(\mathrm{d}v,X)}{\int L_\ell(\mathrm{d}v,X)}\,.$$

Therefore,

$$\begin{aligned}
&\int f(v_{q+\ell})\left\{\mathbb{P}_\ell\left(\mathrm{d}v_{q+\ell}|X_{q+1:\mathsf{n}}\right)-\mathbb{P}_{\ell+1}\left(\mathrm{d}v_{q+\ell}|X_{q+1:\mathsf{n}}\right)\right\}\\
&\qquad=\int f(v_{q+\ell})\left(\frac{L_\ell(\mathrm{d}v,X)}{\int L_\ell(\mathrm{d}v,X)}-\frac{L_{\ell+1}(\mathrm{d}v,X)}{\int L_{\ell+1}(\mathrm{d}v,X)}\right),\\
&\qquad=\int f(v_{q+\ell})\frac{L_\ell(\mathrm{d}v,X)-L_{\ell+1}(\mathrm{d}v,X)}{\int L_\ell(\mathrm{d}v,X)}\\
&\qquad\qquad+\int f(v_{q+\ell})\frac{L_{\ell+1}(\mathrm{d}v,X)}{\int L_{\ell+1}(\mathrm{d}v,X)}\frac{\int\left[L_{\ell+1}(\mathrm{d}v,X)-L_\ell(\mathrm{d}v,X)\right]}{\int L_\ell(\mathrm{d}v,X)}\,.
\end{aligned}$$

Thus,

$$\left|\int f(v_{q+\ell})\left\{\mathbb{P}_\ell\left(\mathrm{d}v_{q+\ell}|X_{q+1:\mathsf{n}}\right)-\mathbb{P}_{\ell+1}\left(\mathrm{d}v_{q+\ell}|X_{q+1:\mathsf{n}}\right)\right\}\right|\quad\leq\quad 2\frac{\|L_\ell(\cdot,X)-L_{\ell+1}(\cdot,X)\|_{\mathrm{tv}}}{\int L_\ell(\mathrm{d}v,X)}\,. \tag{19}$$

By (18), for any f such that $\|f\|_\infty\leq 1$, for any $1\leq\ell\leq\mathsf{n}+1-q$,

$$\begin{aligned}
&\left|\int f(v)(L_\ell(\mathrm{d}v,X)-L_{\ell+1}(\mathrm{d}v,X))\right|\\
&=\left|\int f(v)\prod_{m=q+1}^{q+\ell-1}\pi'_V(\mathrm{d}v_m)\left\{\pi_V(\mathrm{d}v_{q+\ell})-\pi'_V(\mathrm{d}v_{q+\ell})\right\}\prod_{m=q+\ell+1}^{\mathsf{n}+1}\pi_V(\mathrm{d}v_m)\prod_{m=q+1}^{\mathsf{n}}K_m(X_m,v_m,v_{m+1})\right|.
\end{aligned}$$

As $K_{q+\ell-1}$ and $K_{q+\ell}$ are upper bounded by 1,

$$\begin{aligned}
\left|\int f(v)L_\ell(\mathrm{d}v,X)-L_{\ell+1}(\mathrm{d}v,X)\right|&\leq\left(\int\prod_{m=q+1}^{q+\ell-1}\pi'_V(\mathrm{d}v_m)\prod_{m=q+1}^{q+\ell-2}K_m(X_m,v_m,v_{m+1})\right)\\
&\qquad\times\|\pi_V-\pi'_V\|_{\mathrm{tv}}\left(\int\prod_{m=q+\ell+1}^{n+1}\pi_V(\mathrm{d}v_m)\prod_{m=q+\ell+1}^{n}K_m(X_m,v_m,v_{m+1})\right).
\end{aligned}$$

Similarly, since $K_{q+\ell-1}$ and $K_{q+\ell}$ are respectively lower bounded by $\nu_{q+\ell-1}$ and $\nu_{q+\ell}$,

$$\int L_\ell(\mathrm{d}v, X) \geq \left(\int \prod_{m=q+1}^{q+\ell-1} \pi'_V(\mathrm{d}v_m) \prod_{m=q+1}^{q+\ell-2} K_m(X_m, v_m, v_{m+1}) \right) \times \nu_{q+\ell-1}\nu_{q+\ell} \left(\int \prod_{m=q+\ell+1}^{n+1} \pi_V(\mathrm{d}v_m) \prod_{m=q+\ell+1}^{n} K_m(X_m, v_m, v_{m+1}) \right) .$$

Plugging these bounds in (19) yields, for $1 \leq \ell \leq \mathsf{n}+1-q$,

$$\left| \int f(v_{q+\ell}) \left\{ \mathbb{P}_\ell \left(\mathrm{d}v_{q+\ell} | X_{q+1:\mathsf{n}} \right) - \mathbb{P}_{\ell+1} \left(\mathrm{d}v_{q+\ell} | X_{q+1:\mathsf{n}} \right) \right\} \right| \leq 2 \left(\nu_{q+\ell-1}\nu_{q+\ell} \right)^{-1} \left\| \pi_V - \pi'_V \right\|_{\mathrm{tv}} .$$

The proof is completed using the fact that for all $x, y > 0$, $|\log x - \log y| \leq |x-y|/x \wedge y$. □

Lemma 10 is a key ingredient to prove bounded difference properties for log-likelihood based processes.

Lemma 10. *For all $1 \leq q \leq \mathsf{n}$ and all $q \leq \tilde{q} \leq \mathsf{n}$, let $\widetilde{X}^{\tilde{q}}_{q:\mathsf{n}}$ be such that $\widetilde{X}^{\tilde{q}}_{\tilde{q}} \in \mathbb{X}$ and $\widetilde{X}^{\tilde{q}}_k = X_k$ for all $q \leq k \leq \mathsf{n}$ such that $k \neq \tilde{q}$. For any $1 \leq q \leq \tilde{q} \leq \mathsf{n}$,*

$$\left| \log \mathbb{P}_{\pi_V}(X_q | X_{q+1:\mathsf{n}}) - \log \mathbb{P}_{\pi_V}(\widetilde{X}^{\tilde{q}}_q | \widetilde{X}^{\tilde{q}}_{q+1:\mathsf{n}}) \right| \leq \nu_q^{-1} \prod_{k=q+1}^{\tilde{q}-1} (1-\nu_k) .$$

Proof. If $q = \tilde{q} = \mathsf{n}$, then

$$\begin{aligned} \left| \mathbb{P}_{\pi_V}(X_\mathsf{n}) - \mathbb{P}_{\pi_V}(\widetilde{X}^\mathsf{n}_\mathsf{n}) \right| &= \left| \int \pi_V(\mathrm{d}v_\mathsf{n}) \pi_V(\mathrm{d}v_{\mathsf{n}+1}) \left\{ K_\mathsf{n}(X_\mathsf{n}, v_\mathsf{n}, v_{\mathsf{n}+1}) - K_\mathsf{n}(\widetilde{X}^\mathsf{n}_\mathsf{n}, v_\mathsf{n}, v_{\mathsf{n}+1}) \right\} \right| , \\ &\leq 1 - \nu_\mathsf{n} \leq 1 . \end{aligned}$$

Assume now that $1 \leq q < \mathsf{n}$. When $\tilde{q} = q$,

$$\begin{aligned} &\mathbb{P}_{\pi_V}(X_q | X_{q+1:\mathsf{n}}) - \mathbb{P}_{\pi_V}(\widetilde{X}^q_q | \widetilde{X}^q_{q+1:\mathsf{n}}) \\ &\qquad = \int \mathbb{P}_{\pi_V} \left(\mathrm{d}v_{q+1} \middle| \widetilde{X}^q_{q+1:\mathsf{n}} \right) \pi_V(\mathrm{d}v_q) \left\{ K_q(X_q, v_q, v_{q+1}) - K_q(\widetilde{X}^q_q, v_q, v_{q+1}) \right\} , \end{aligned}$$

which ensures that $|\mathbb{P}_{\pi_V}(X_q | X_{q+1:\mathsf{n}}) - \mathbb{P}_{\pi_V}(\widetilde{X}^q_q | \widetilde{X}^q_{q+1:\mathsf{n}})| \leq 1 - \nu_q \leq 1$. When $\tilde{q} \geq q+1$, as for all $q+1 \leq k \leq \tilde{q}-1$ the Markov transition kernel $K^{V|X}_{\pi_V,k,q+1}$ depends only on π_V, K_k and $X_{q+1:k}$,

$$\mathbb{P}_{\pi_V} \left(\widetilde{X}^{\tilde{q}}_q \middle| \widetilde{X}^{\tilde{q}}_{q+1:\mathsf{n}} \right) = \int \mathbb{P}_{\pi_V} \left(\mathrm{d}v_{\tilde{q}} \middle| \widetilde{X}^{\tilde{q}}_{q+1:\mathsf{n}} \right) \left(\prod_{k=q+1}^{\tilde{q}-1} K^{V|X}_{\pi_V,k,q+1}(v_{k+1}, \mathrm{d}v_k) \right) \pi_V(\mathrm{d}v_q) K_q(X_q, v_q, v_{q+1}) .$$

By Lemma 7, it follows that

$$\begin{aligned} &\left| \mathbb{P}_{\pi_V}(X_q | X_{q+1:\mathsf{n}}) - \mathbb{P}_{\pi_V} \left(\widetilde{X}^{\tilde{q}}_q \middle| \widetilde{X}^{\tilde{q}}_{q+1:\mathsf{n}} \right) \right| \\ &\qquad \leq \left(\prod_{k=q+1}^{\tilde{q}-1} (1-\nu_k) \right) \sup_{v_{q+1} \in \mathbb{V}} \left| \int \pi_V(\mathrm{d}v_q) K_q(X_q, v_q, v_{q+1}) \right| . \end{aligned}$$

The proof is completed using the fact that for all $x, y > 0$, $|\log x - \log y| \leq |x-y|/x \wedge y$. □

Let π_V^* denote a probability distribution on $\mathbb{V}$ and let

$$Z_{\pi_V}(X_{1:\mathsf{n}}) = \frac{1}{\mathsf{n}} \sum_{q=1}^{\mathsf{n}} \left[\log \mathbb{P}_{\pi_V}(X_q | X_{q+1:\mathsf{n}}) - \mathbb{E}_{\pi_V^*} \left[\log \mathbb{P}_{\pi_V}(X_q | X_{q+1:\mathsf{n}}) \right] \right] .$$

Lemma 11 shows the concentration of $Z_{\pi_V}(X_{1:\mathsf{n}})$ around its expectation.

Lemma 11. *Assume that $K_i = K$ for all $i \in \mathbb{Z}$, let $\mathcal{P}$ denote a class of probability distributions on $\mathbb{V}$. There exists $c > 0$ such that for all $t > 0$,*

$$\mathbb{P}_{\pi_V^*} \left(\left| \sup_{\pi_V \in \mathcal{P}} \{ Z_{\pi_V}(X_{1:\mathsf{n}}) \} - \mathbb{E}_{\pi_V^*} \left[\sup_{\pi_V \in \mathcal{P}} \{ Z_{\pi_V}(X_{1:\mathsf{n}}) \} \right] \right| \geq c \nu^{-2} \frac{t}{\sqrt{\mathsf{n}}} \right) \leq 2 e^{-t^2} .$$

Proof. The proof relies on the bounded difference inequality for Markov chains [10, Theorem 0.2]. To apply this result, $\sup_{\pi_V \in \mathcal{P}} \{ Z_{\pi_V}(X_{1:\mathsf{n}}) \}$ has to be separately bounded. For all $1 \leq q \leq \mathsf{n}$ and all $q \leq \tilde{q} \leq \mathsf{n}$, let $\widetilde{X}_{1:\mathsf{n}}^{\tilde{q}}$ such that $\widetilde{X}_{\tilde{q}}^{\tilde{q}} \in \mathbb{X}$ and $\widetilde{X}_k^{\tilde{q}} = X_k$ for all $1 \leq k \leq \mathsf{n}$ such that $k \neq \tilde{q}$. Then,

$$\begin{aligned} &\left| \sup_{\pi_V \in \mathcal{P}} \{ Z_{\pi_V}(X_{1:\mathsf{n}}) \} - \sup_{\pi_V \in \mathcal{P}} \{ Z_{\pi_V}(\widetilde{X}_{1:\mathsf{n}}^{\tilde{q}}) \} \right| \\ &\qquad \leq \sup_{\pi_V \in \mathcal{P}} \left| \frac{1}{\mathsf{n}} \sum_{q=1}^{\mathsf{n}} \left[\log \mathbb{P}_{\pi_V}(X_q | X_{q+1:\mathsf{n}}) - \log \mathbb{P}_{\pi_V}(\widetilde{X}_q^{\tilde{q}} | \widetilde{X}_{q+1:\mathsf{n}}^{\tilde{q}}) \right] \right| \\ &\qquad \leq \sup_{\pi_V \in \mathcal{P}} \left| \frac{1}{\mathsf{n}} \sum_{q=1}^{\tilde{q}} \left[\log \mathbb{P}_{\pi_V}(X_q | X_{q+1:\mathsf{n}}) - \log \mathbb{P}_{\pi_V}(\widetilde{X}_q^{\tilde{q}} | \widetilde{X}_{q+1:\mathsf{n}}^{\tilde{q}}) \right] \right| . \end{aligned}$$

By Lemma 10, for any distribution $\pi_V \in \mathcal{P}$ and any $1 \leq q \leq \mathsf{n}$,

$$\left| \frac{1}{\mathsf{n}} \sum_{q=1}^{\mathsf{n}} \left[\log \mathbb{P}_{\pi_V}(X_q | X_{q+1:\mathsf{n}}) - \log \mathbb{P}_{\pi_V}(\widetilde{X}_q^{\tilde{q}} | \widetilde{X}_{q+1:\mathsf{n}}^{\tilde{q}}) \right] \right| \leq \frac{1}{\mathsf{n}} \sum_{q=1}^{\tilde{q}} \nu^{-1} (1-\nu)^{\tilde{q}-q-1} .$$

Hence, there exists $c > 0$ such that,

$$\left| \sup_{\pi_V \in \mathcal{P}} \{ Z_{\pi_V}(X_{1:\mathsf{n}}) \} - \sup_{\pi_V \in \mathcal{P}} \{ Z_{\pi_V}(\widetilde{X}_{1:\mathsf{n}}^{\tilde{q}}) \} \right| \leq \frac{c}{\nu^2 \mathsf{n}} .$$

The proof is concluded by [10, Theorem 0.2]. □

Lemma 12 shows the subgaussian concentration inequality of the increments of $Z_{\pi_V}(X_{1:\mathsf{n}})$.

Lemma 12. *Assume that $K_i = K$ for all $i \in \mathbb{Z}$, let π_V, π_V' denote two probability distributions on $\mathbb{V}$. Let d denote the distance defined in (5). Then, there exists $c > 0$ such that for all $n \geq 1$, $t > 0$,*

$$\mathbb{P}_{\pi_V^*} \left(\left| \sqrt{\mathsf{n}} \left\{ Z_{\pi_V}(X_{1:\mathsf{n}}) - Z_{\pi_V'}(X_{1:\mathsf{n}}) \right\} \right| > t \right) \leq \exp \left[- \frac{t^2}{\left(c \nu^{-5} d(\pi, \pi') \right)^2} \right] . \tag{20}$$

Proof. To prove that the increments $Z_{\pi_V} - Z_{\pi_V'}$ are separately bounded, consider, for all $1 \leq \tilde{q} \leq \mathsf{n}$, $\widetilde{X}_{1:\mathsf{n}}^{\tilde{q}}$ such that $\widetilde{X}_{\tilde{q}}^{\tilde{q}} \in \mathbb{X}$ and $\widetilde{X}_k^{\tilde{q}} = X_k$ for all $1 \leq k \leq \mathsf{n}$ such that $k \neq \tilde{q}$. On one hand, by Lemma 9,

$$\left| \log \mathbb{P}_{\pi_V}(X_q | X_{q+1:\mathsf{n}}) - \log \mathbb{P}_{\pi_V'}(X_q | X_{q+1:\mathsf{n}}) \right| \leq 2 \nu^{-4} \| \pi_V - \pi_V' \|_{\mathsf{tv}} .$$

On the other hand, by Lemma 10, for any $1 \le q \le \tilde{q} \le \mathsf{n}$,

$$\left|\log \mathbb{P}_{\pi_V}(X_q|X_{q+1:\mathsf{n}}) - \log \mathbb{P}_{\pi_V}(\widetilde{X}_q^{\tilde{q}}|\widetilde{X}_{q+1:\mathsf{n}}^{\tilde{q}})\right| \le \nu^{-1}(1-\nu)^{\tilde{q}-q-1}\,.$$

Since $\log \mathbb{P}_{\pi_V}(X_q|X_{q+1:\mathsf{n}}) = \log \mathbb{P}_{\pi_V}(\widetilde{X}_q^{\tilde{q}}|\widetilde{X}_{q+1:\mathsf{n}}^{\tilde{q}})$ for $q > \tilde{q}$,

$$\begin{aligned}\left|\left(Z_{\pi_V}(X_{1:\mathsf{n}}) - Z_{\pi'_V}(X_{1:\mathsf{n}})\right) - \left(Z_{\pi_V}(\widetilde{X}_{1:\mathsf{n}}^{\tilde{q}}) - Z_{\pi'_V}(\widetilde{X}_{1:\mathsf{n}}^{\tilde{q}})\right)\right| \\ \le \frac{2\nu^{-4}}{\mathsf{n}} \sum_{q=1}^{\tilde{q}} \left[\|\pi_V - \pi'_V\|_{\mathsf{tv}} \wedge (1-\nu)^{\tilde{q}-q}\right] \le \frac{2\nu^{-5}}{\mathsf{n}} d(\pi,\pi')\,.\end{aligned}$$

Eq (20) follows by plugging these bounded differences properties in [10, Theorem 0.2]. □

7 Proofs of the main results

When H1 holds and $E^{n,N} = E^{n,N}_{\mathrm{RR}}$, $(V^{n,N}_{2:\mathsf{q}^n_N}, X^{n,N}_{2:\mathsf{q}^n_N-1})$ satisfies the assumptions of Section 6 with

$$\pi_V = \pi^{\otimes n-1}, \qquad K_i(X_i^{n,N}, V_i^{n,N}, V_{i+1}^{n,N}) = \prod_{X_{i,j}\in X_i^{n,N}} k(X_{i,j}, V_i, V_j), \qquad \nu_i = \varepsilon^{|X_i^{n,N}|}\,.$$

Moreover, it is proved in Section 5 that $\left|X_q^{n,N}\right| = n(n-1)$ for $2 \le q \le \mathsf{q}^n_N - 1$, which implies that

$$\nu_i \ge \varepsilon^{n^2}\,.$$

Throughout the proofs, the following conventions are used. For all $0 \le k \le \mathsf{q}^n_N$,

$$v_k^{n,N} \in \mathcal{V}^{|V_k^{n,N}|}, \qquad \pi(\mathrm{d}v_k^{n,N}) = \prod_{i:V_i\in V_k^{n,N}} \pi(\mathrm{d}v_i)\,.$$

7.1 Proof of Theorem 2

Let ℓ denote an even number larger than 2 and let $Z^{n,N} = X_0^{n,N} \cup X_1^{n,N} \cup X^{n,N}_{\mathsf{q}^n_N}$. By Lemma 8,

$$\begin{aligned}\sup_{\pi\in\Pi} \left|\log \mathbb{P}^{n,N}_\pi\left(X_q^{n,N}\middle|X^{n,N}_{q+1:\mathsf{q}^n_N-1}\right) - \log \mathbb{P}^{n,N+\ell}_\pi\left(X_q^{(n,N+\ell)}\middle|X^{n,N+\ell}_{q+1:\mathsf{q}^n_{N+\ell}-1}\right)\right| \\ \le \varepsilon^{-n^2}\left(1-\varepsilon^{n^2}\right)^{\mathsf{q}^n_N-q-2}\,. \quad (21)\end{aligned}$$

This proves the first conclusion. The log-likelihood is decomposed as follows

$$\begin{aligned}\log \mathbb{P}^{n,N}_\pi\left(X^{n,N}\right) &= \log \mathbb{P}^{n,N}_\pi\left(X^{n,N}_{2:\mathsf{q}^n_N-1}\right) + \log \mathbb{P}^{n,N}_\pi\left(X_0^{n,N}, X_1^{n,N}, X^{n,N}_{\mathsf{q}^n_N}\middle|X^{n,N}_{2:\mathsf{q}^n_N-1}\right)\,, \\ &= \sum_{q=2}^{\mathsf{q}^n_N-1} \log \mathbb{P}^{n,N}_\pi\left(X_q^{n,N}\middle|X^{n,N}_{q+1:\mathsf{q}^n_N-1}\right) + \log \mathbb{P}^{n,N}_\pi\left(Z^{n,N}\middle|X^{n,N}_{2:\mathsf{q}^n_N-1}\right)\,. \quad (22)\end{aligned}$$

Let us first bound from above the last term in (22).

$$\mathbb{P}_\pi^{n,N}\left(Z^{n,N}\middle|X_{2:\mathsf{q}_N^n-1}^{n,N}\right) = \int \mathbb{P}_\pi^{n,N}\left(Z^{n,N}, \mathrm{d}v_{0:2}^{n,N}, \mathrm{d}v_{\mathsf{q}_N^n:\mathsf{q}_N^n+1}^{n,N}\middle|X_{2:\mathsf{q}_N^n-1}^{n,N}\right),$$
$$= \int \mathbb{P}_\pi^{n,N}\left(\mathrm{d}v_{0:2}^{n,N}, \mathrm{d}v_{\mathsf{q}_N^n:\mathsf{q}_N^n+1}^{n,N}\middle|X_{2:\mathsf{q}_N^n-1}^{n,N}\right)\left\{\prod_{X_{i,j}\in Z^{n,N}} k(X_{i,j}, v_i, v_j)\right\},$$

By Assumption 1,

$$\varepsilon^{3n^2} \le \mathbb{P}_\pi^{n,N}\left(Z^{n,N}\middle|X_{2:\mathsf{q}_N^n-1}^{n,N}\right) \le 1\,. \tag{23}$$

In particular, the last term in (22) is $o(\mathsf{q}_N^n)$ when N grows to infinity. Taking the limit as $\ell \to \infty$ in (21) yields, for any n and N,

$$\sup_{\pi\in\Pi} \frac{1}{\mathsf{q}_N^n}\left|\sum_{q=2}^{\mathsf{q}_N^n-1}\left\{\log \mathbb{P}_\pi^{n,N}\left(X_q^{n,N}\middle|X_{q+1:\mathsf{q}_N^n-1}^{n,N}\right) - \ell_\pi^n(\vartheta^q\mathbf{X}^\mathbf{n})\right\}\right| \le \frac{3\varepsilon^{-3n^2}}{\mathsf{q}_N^n}\,. \tag{24}$$

By (14), $|\ell_\pi^n(\mathbf{X}^\mathbf{n})| \le n^2\log(\varepsilon^{-1})$, thus ℓ_π^n is integrable. Therefore, the ergodic theorem [1, Theorem 24.1] can be applied to $\sum_{q=2}^{\mathsf{q}_N^n-1}\ell_\pi^n(\vartheta^q\mathbf{X}^\mathbf{n})/\mathsf{q}_N^n$ and (3) follows.

7.2 $R_{\pi_\star}$ is the excess risk function

The following result shows that $R_{\pi_\star}^n$ is a non-negative function.

Proposition 13. *For all $\pi \in \Pi$ and all $n \ge 1$, $R_{\pi_\star}^n(\pi) \ge 0$.*

Proof. Let $\pi \in \Pi$ and $n \ge 1$. By (2),

$$\mathsf{L}_{\pi_\star}^n(\pi) = \mathbb{E}_{\pi_\star}\left[\lim_{N\to\infty} \log \mathbb{P}_\pi^{n,N}(X_2^{n,N}|X_{3:\mathsf{q}_N^n-1}^{n,N})\right].$$

By Lebesgue's bounded convergence theorem

$$\mathsf{L}_{\pi_\star}^n(\pi) = \lim_{N\to\infty} \mathbb{E}_{\pi_\star}\left[\log \mathbb{P}_\pi^{n,N}(X_2^{n,N}|X_{3:\mathsf{q}_N^n-1}^{n,N})\right]$$
$$= \lim_{N\to\infty} \mathbb{E}_{\pi_\star}\left[\mathbb{E}_{\pi_\star}\left[\log \mathbb{P}_\pi^{n,N}(X_2^{n,N}|X_{3:\mathsf{q}_N^n-1}^{n,N})\middle|X_{3:\mathsf{q}_N^n-1}^{n,N}\right]\right].$$

Therefore,

$$R_{\pi_\star}^n(\pi) = \lim_{N\to\infty}\left\{\mathbb{E}_{\pi_\star}\left[\mathbb{E}_{\pi_\star}\left[\log \mathbb{P}_{\pi_\star}^{n,N}(X_2^{n,N}|X_{3:\mathsf{q}_N^n-1}^{n,N}) - \log \mathbb{P}_\pi^{n,N}(X_2^{n,N}|X_{3:\mathsf{q}_N^n-1}^{n,N})\middle|X_{3:\mathsf{q}_N^n-1}^{n,N}\right]\right]\right\},$$

and the latter is non negative since the term in the expectation is a Kullback-Leibler divergence. □

7.3 Proof of Theorem 3

As that for any $\pi \in \Pi \cup \{\pi_\star\}$, $\ell^{n,N}(\pi) = \log \mathbb{P}_\pi^{n,N}(X^{n,N})$, the excess loss satisfies:

$$\begin{aligned} R_{\pi_\star}^n(\widehat{\pi}^{n,N}) = \mathsf{L}_{\pi_\star}^n(\pi_\star) - \mathbb{E}_{\pi_\star}\left[\frac{1}{\mathsf{q}_N^n}\ell^{n,N}(\pi_\star)\right] + \mathbb{E}_{\pi_\star}\left[\frac{1}{\mathsf{q}_N^n}\ell^{n,N}(\pi_\star)\right] - \frac{1}{\mathsf{q}_N^n}\ell^{n,N}(\pi_\star) \\ + \frac{1}{\mathsf{q}_N^n}\ell^{n,N}(\pi_\star) - \frac{1}{\mathsf{q}_N^n}\ell^{n,N}\left(\widehat{\pi}^{n,N}\right) + \frac{1}{\mathsf{q}_N^n}\ell^{n,N}\left(\widehat{\pi}^{n,N}\right) - \mathbb{E}_{\pi_\star}\left[\frac{1}{\mathsf{q}_N^n}\ell^{n,N}\left(\widehat{\pi}^{n,N}\right)\right] \\ + \mathbb{E}_{\pi_\star}\left[\frac{1}{\mathsf{q}_N^n}\ell^{n,N}\left(\widehat{\pi}^{n,N}\right)\right] - \mathsf{L}_{\pi_\star}^n(\widehat{\pi}^{n,N}). \end{aligned}$$

By definition $\ell^{n,N}(\pi_\star) - \ell^{n,N}\left(\widehat{\pi}^{n,N}\right) \leq 0$. Thus,

$$R_{\pi_\star}^n(\widehat{\pi}^{n,N}) \leq 2\sup_{\pi\in\Pi\cup\{\pi^*\}}\left\{\left|\mathsf{L}^{\pi_\star}(\pi) - \frac{\mathbb{E}_{\pi_\star}\left[\ell^{n,N}(\pi)\right]}{\mathsf{q}_N^n}\right| + \left|\frac{1}{\mathsf{q}_N^n}\mathbb{E}_{\pi_\star}\left[\ell^{n,N}(\pi)\right] - \frac{\ell^{n,N}(\pi)}{\mathsf{q}_N^n}\right|\right\}.$$

Let $Z^{n,N} = X_0^{n,N} \cup X_1^{n,N} \cup X_{\mathsf{q}_N^n}^{n,N}$. For all $\pi \in \Pi$,

$$\begin{aligned} \left|\mathsf{L}^{\pi_\star}(\pi) - \frac{\mathbb{E}_{\pi_\star}\left[\ell^{n,N}(\pi)\right]}{\mathsf{q}_N^n}\right| \leq \frac{1}{\mathsf{q}_N^n}\mathbb{E}_{\pi_\star}\left[\sum_{q=2}^{\mathsf{q}_N^n-1}\left|\ell_\pi^n(\vartheta^q\mathbf{X}^{\mathbf{n}}) - \log\mathbb{P}_\pi^{n,N}\left(X_q^{n,N}\middle|X_{q+1:\mathsf{q}_N^n-1}^{n,N}\right)\right|\right] \\ + \frac{1}{\mathsf{q}_N^n}\mathbb{E}_{\pi_\star}\left[\left|2\ell_\pi^n(\mathbf{X}^{\mathbf{n}}) - \log\mathbb{P}_\pi^{n,N}\left(Z^{n,N}\middle|X_{2:\mathsf{q}_N^n-1}^{n,N}\right)\right|\right]. \end{aligned}$$

Then, by Lemma 8 and (24), for the round-robin scheduling, there exists c such that:

$$\sup_{\pi\in\Pi\cup\{\pi^*\}}\left|\mathsf{L}^{\pi_\star}(\pi) - \frac{\mathbb{E}_{\pi_\star}\left[\ell^{n,N}(\pi)\right]}{\mathsf{q}_N^n}\right| \leq \frac{c\varepsilon^{-3n^2}}{\mathsf{q}_N^n}.$$

This yields:

$$R_{\pi_\star}^{n,N}(\widehat{\pi}^{n,N}) \leq \frac{c\varepsilon^{-3n^2}}{\mathsf{q}_N^n} + \frac{2}{\mathsf{q}_N^n}\sup_{\pi\in\Pi\cup\{\pi^*\}}\left|\mathbb{E}_{\pi_\star}\left[\ell^{n,N}(\pi)\right] - \frac{1}{\mathsf{q}_N^n}\ell^{n,N}(\pi)\right|,$$

and therefore, by (23),

$$R_{\pi_\star}^{n,N}(\widehat{\pi}^{n,N}) \leq \frac{c\varepsilon^{-3n^2}}{\mathsf{q}_N^n} + 2\sup_{\pi\in\Pi\cup\{\pi^*\}}|Z_{\pi_V}|, \tag{25}$$

where

$$Z_\pi = \frac{1}{\mathsf{q}_N^n}\sum_{q=2}^{\mathsf{q}_N^n-1}\left[\log\mathbb{P}_\pi^{n,N}(X_q^{n,N}|X_{q+1:\mathsf{n}}^{n,N}) - \mathbb{E}_{\pi_\star}\left[\log\mathbb{P}_\pi^{n,N}(X_q^{n,N}|X_{q+1:\mathsf{n}}^{n,N})\right]\right].$$

Lemma 11 applies by assumption H1 since $E^{n,N} = E_{\mathrm{RR}}^{n,N}$, therefore, there exists $c > 0$ such that, for all $t > 0$,

$$\mathbb{P}_{\pi_\star}\left(\left|\sup_{\pi\in\Pi\cup\{\pi^*\}}Z_\pi - \mathbb{E}_{\pi_\star}\left[\sup_{\pi\in\Pi\cup\{\pi^*\}}Z_\pi\right]\right| > c\varepsilon^{-2n^2}\frac{t}{\sqrt{\mathsf{q}_N^n}}\right) \leq e^{-t^2},. \tag{26}$$

Furthermore, by Lemma 12, the increments of Z_π have subgaussian tails. Since by Lemma 1, $|V_2^{n,N}| = 2(n-1)$, for all $t > 0$,

$$\mathbb{P}_{\pi_\star}\left(\sqrt{\mathsf{q}_N^n}\,|Z_\pi - Z_{\pi'}| > t\right) \leq \exp\left(-\frac{t^2}{\left(c\varepsilon^{-5n^2} d(\pi^{\otimes 2(n-1)}, (\pi')^{\otimes 2(n-1)})\right)^2}\right).$$

Now it is easy to check that

$$\left\|\pi^{\otimes 2(n-1)} - (\pi')^{\otimes 2(n-1)}\right\|_{\mathsf{tv}} \leq 2(n-1)\left\|\pi - \pi'\right\|_{\mathsf{tv}}.$$

Therefore, $d(\pi^{\otimes 2(n-1)}, (\pi')^{\otimes |V_2^{n,N}|}) \leq cn^2 d(\pi, \pi') \leq c\varepsilon^{-n^2} d(\pi, \pi')$, thus for all $t > 0$,

$$\mathbb{P}_{\pi_\star}\left(\sqrt{\mathsf{q}_N^n}\,|Z_\pi - Z_{\pi'}| > t\right) \leq \exp\left(-\frac{t^2}{\left(c\varepsilon^{-6n^2} d(\pi, \pi')\right)^2}\right). \tag{27}$$

Then, by Dudley's entropy bound, see [12] or [24, Proposition 2.1],

$$\mathbb{E}_{\pi_\star}\left[\sup_{\pi\in\Pi\cup\{\pi_\star\}} Z_\pi(X^{n,N})\right] \leq \frac{ce^{-6n^2}}{\sqrt{\mathsf{q}_N^n}} \int_0^{+\infty} \sqrt{\log \mathsf{N}(\Pi \cup \{\pi_\star\}, d, \epsilon)}\mathrm{d}\epsilon\,. \tag{28}$$

Plugging (26) and (28) into (25) concludes the proof.

References

[1] P. Billingsley. *Probability and Measure*. Wiley, 1995.

[2] R. Bradley and M. Terry. Rank analysis of incomplete block designs: I. the method of pair comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[3] O. Cappe, E. Moulines, and T. Ryden. *Inference in hidden Markov models*. Springer, 2005.

[4] D. Caron and A. Doucet. Efficient Bayesian inference for generalized Bradley-Terry models. *Journal of Computational and Graphical Statistics*, 21(1):174–196, 2012.

[5] S. Chatterjee, P. Diaconis, and A. Sly. Random graphs with a given degree sequence. *The Annals of Applied Probability*, 21(4):1400–1435, 2011.

[6] R. Chetrite, R. Diel, and M. Lerasle. The number of potential winners in bradley-terry models in random environment. *To appear in The Annals of Applied Probability*, 2017.

[7] H. A. David. *The method of paired comparisons*, volume 41 of *Griffin's Statistical Monographs & Courses*. Charles Griffin & Co., Ltd., London; The Clarendon Press, Oxford University Press, New York, second edition, 1988.

[8] Y. De Castro, E. Gassiat, and C. Lacour. Minimax adaptative estimation of nonparametric hidden Markov models. *Journal of Machine Learning Research*, 17:1–43, 2016.

[9] Y. De Castro, E. Gassiat, and S. Le Corff. Consistent estimation of the filtering and marginal smoothing distributions in nonparametric hidden Markov models. *To appear in IEEE Transactions on Information Theory*, 2017.

[10] J. Dedecker and S. Gouezel. Subgaussian concentration inequalities for geometrically ergodic Markov chains. *Electronic Communications in Probability*, 20:1–12, 2015.

[11] R. Douc and É. Moulines. Asymptotic properties of the maximum likelihood estimation in misspecified hidden Markov models. *The Annals of Statistics*, 40(5):2697–2732, 2012.

[12] R. M. Dudley. The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *Journal of Functional Analysis*, 1:290–330, 1967.

[13] B. Efron. *Large-scale inference*, volume 1 of *Institute of Mathematical Statistics (IMS) Monographs*. Cambridge University Press, Cambridge, 2010. Empirical Bayes methods for estimation, testing, and prediction.

[14] D.R. Hunter. MM algorithms for generalized Bradley-Terry models. *The Annals of Statistics*, 32(1):384–406, 2004.

[15] M. Lavielle. *Mixed effects models for the population approach*. Chapman & Hall/CRC Biostatistics Series. CRC Press, Boca Raton, FL, 2015. Models, tasks, methods and tools, With contributions by Kevin Bleakley.

[16] L. Lehéricy. Consistent order estimation for nonparametric hidden Markov models. *ArXiv:1606.00622*, 2017.

[17] T. Lindvall. *Lectures on the coupling method*. Wiley, 1992.

[18] E. Mammen and A.B. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829, 1999.

[19] P. Massart and É. Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 34(5):2326–2366, 2006.

[20] P. Rap and L. Kupper. Ties in paired-comparison experiments: a generalization of the Bradley-Terry model. *Journal of the American Statistical Association*, 62:194–204, 1967.

[21] H. Robbins. An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, vol. I*, pages 157–163. University of California Press, Berkeley and Los Angeles, 1956.

[22] G. Simons and Y.-C. Yao. Asymptotics when the number of parameters tends to infinity in the Bradley-Terry model for paired comparisons. *The Annals of Statistics*, 27(3):1041–1060, 1999.

[23] C. Sire and S. Redner. Understanding baseball team standings and streaks. *Eur. Phys. J. B*, 67:473–481, 2009.

[24] M. Talagrand. *Upper and lower bounds for stochastic processes*, volume 60. Springer, Heidelberg, 2014. Modern methods and classical problems.

[25] A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.

[26] V. N. Vapnik and A. Ya. Chervonenkis. *Teoriya raspoznavaniya obrazov. Statisticheskie problemy obucheniya.*

[27] V.N. Vapnik. *Statistical learning theory.* Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons, Inc., New York, 1998. A Wiley-Interscience Publication.

[28] É. Vernet. Posterior consistency for nonparametric hidden Markov models with finite state space. *Electronic Journal of Statistics*, 9, 2015.

[29] É. Vernet. Nonparametric hidden Markov models with finite state space: Posterior concentration rates. *ArXiv:1511.08624*, 2017.

[30] T. Yan, Y. Yang, and J. Xu. Sparse paired comparisons in the Bradley-Terry model. *Statistica Sinica*, 22(3):1305–1318, 2012.

[31] E. Zemerlo. Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 29(1):436–460, 1929.