



**HAL**  
open science

# Asymptotic properties of the maximum likelihood estimator for Hidden Markov Models indexed by binary trees

Julien Weibel

► **To cite this version:**

Julien Weibel. Asymptotic properties of the maximum likelihood estimator for Hidden Markov Models indexed by binary trees. 2024. hal-04692082

**HAL Id: hal-04692082**

**<https://hal.science/hal-04692082v1>**

Preprint submitted on 9 Sep 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Asymptotic properties of the maximum likelihood estimator for Hidden Markov Models indexed by binary trees

Julien Weibel\*

September 8, 2024

## Abstract

We consider hidden Markov models indexed by a binary tree where the hidden state space is a general metric space. We study the maximum likelihood estimator (MLE) of the model parameters based only on the observed variables. In both stationary and non-stationary regimes, we prove strong consistency and asymptotic normality of the MLE under standard assumptions. Those standard assumptions imply uniform exponential memorylessness properties of the initial distribution conditional on the observations. The proofs rely on ergodic theorems for Markov chain indexed by trees with neighborhood-dependent functions.

**Key words and phrases**— Hidden Markov tree (HMT), hidden Markov model (HMM), branching process, maximum likelihood estimator (MLE), asymptotic normality, consistency, geometric ergodicity  
**2020 Mathematics Subject Classification**— 62M05, 62F12, 60J80, 60J85

## 1 Introduction

In this article, we consider a generalization of the hidden Markov chain/model (HMM) where the process is indexed by a binary tree, which we call hidden Markov tree (HMT). The HMT is composed of a hidden process and an observed process. The hidden process is a branching Markov process, that is, a random process  $X = (X_u, u \in T)$  with values in a metric space  $\mathcal{X}$  indexed by a rooted tree  $T$  with the Markov property: sibling nodes take independent and identically distributed values that depend only on the value of their parent node. Note that the hidden process is sometimes called latent process in the literature. Conditionally on the hidden process  $X$ , the observed process  $Y = (Y_u, u \in T)$ , with values in another metric space  $\mathcal{Y}$ , is composed of independent random variables  $Y_u$  which only depends on  $X_u$  for all  $u \in T$ . See Definitions 2.1 and 2.2 below for a complete formal definitions. In this article, we consider the case where the tree  $T$  is the (deterministic) complete infinite rooted binary tree, that is, each vertex has exactly two children. See Figure 1 for a graphical representation of the dependance between the variables composing the HMT process  $(X, Y)$  indexed by  $T$ .

### 1.1 Literature review

HMMs were first introduced by Baum and Petrie in [BP66] and were popularized by Rabiner's tutorial [Rab89]. Since then, HMMs have been used in a wide variety of applications such

---

\*Institut Denis Poisson, Université d'Orléans, Université de Tours, CNRS, France and CERMICS, École des Ponts, France. [julien.weibel@normalesup.org](mailto:julien.weibel@normalesup.org)

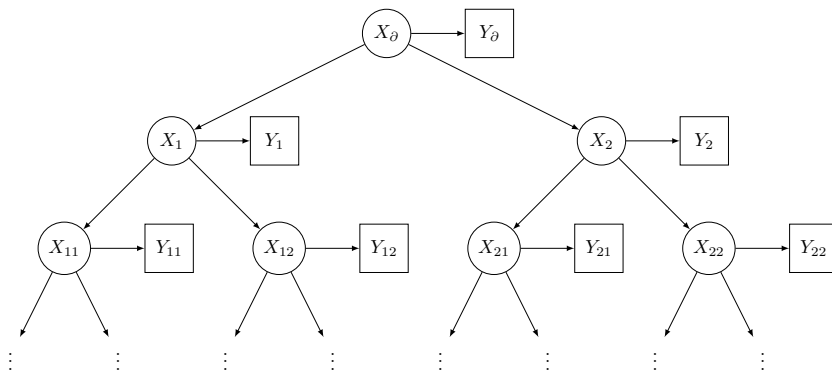


Figure 1: Graph of dependance for variables of a HMT process indexed by the complete infinite rooted binary tree  $T$ . The observed variables are represented inside square, while the hidden variables are represented inside circles.

as speech recognition [YD15], bioinformatics [Kos01], finance [ME14], and time-series analysis [ZM09]; see also [BFA22] for a more global reference on HMMs applications.

HMTs were first introduced in [CNB98] to account for the multi-scale dependency of wavelet coefficients in statistical signal processing with applications in wavelet-based image processing [RCBK00, CB01, DWB08, SH17]. After that, HMTs have been used in several application contexts such as natural language processing [GOB13, KDM13], flood mapping [XJS18], medical imaging [MBY<sup>+</sup>12, HYG17, HBSLLB<sup>+</sup>17], plant growth modeling [DGCC05], and bioinformatics [OCB<sup>+</sup>09, BWX13, NSK20].

In practice, maximum likelihood estimation for HMMs often relies on iterative numerical methods to approximate the maximum likelihood estimator (MLE). Those methods are often based on the expectation-maximization algorithm which is an algorithm for models with missing data and was popularized by Dempster et al. [DLR77] in a celebrated article. For HMMs with finite hidden state space, the first presentation of a complete expectation-maximization strategy is due to Baum et al. [BPSW70], and is the well-known “forward-backward” or Baum-Welch algorithm. For more details on the expectation-maximization and “forward-backward” algorithms and their stochastic approximations, see [CMR05, Chapters 10 and 11]. In the HMT case, the “forward-backward” algorithm must be replaced by the “upward-downward” algorithm developed in [CNB98]. See also [DGG04] for alternative “upward-downward” recursive formulae that can handle underflow issues implicitly.

The statistical properties of the MLE for the HMM were first studied in [BP66] which proved consistency and asymptotic normality in the case where both the hidden and the observed processes can only take finitely many values. Those results were then successively extended in a series of articles [Ler92, BRR98, JP99, LGM00, DM01]. An extension of all those results for HMMs with autoregression (that is, when conditionally on the hidden Markov chain, the observed process is an inhomogeneous  $s$ -order Markov chain for some  $s \in \mathbb{N}$ ) was later developed in [DMR04], which proved, using weaker assumptions, strong consistency and asymptotic normality of the MLE for auto-regressive HMMs with compact hidden state space and with possibly non-stationary regime. The methods used in [DMR04] relies on expressing the log-likelihood as an additive function of an extended Markov chain with infinite past thanks to stationarity and using geometric ergodicity of this extended chain (extension to non-stationary regime is then made separately). The method of [DMR04] was adapted in [KS19] under similar assumptions to allow the

transition densities of the hidden process to be zero valued. Since the article [DMR04], the strong consistency of the MLE was proved under weaker assumptions in [GL06, DMOVH11, DRS16], but no generalization has been made for the asymptotic normality of the MLE.

In this article, we will adapt the proof method of [DMR04] to the HMT case. We shall also refer to the monograph [CMR05] which exposes in details the theory of HMMs, and in particular to its Chapter 12 which covers the strong consistency and asymptotic normality of the MLE, under the same assumptions used in [DMR04], for HMMs where the hidden state space is a general metric space.

To adapt the proof method of [DMR04] to the HMT case, we will need almost sure (a.s.) and  $L^2$  ergodic convergence results for branching Markov chains under geometric ergodicity of the transition kernel as in [Guy07, Wei24]. Indeed, we will need variants of those results for neighborhood-dependent functions (that is, the function associated to each vertex  $u$  depends on variables  $X_v$  for vertices  $v$  in the neighborhood of  $u$ ) which we develop in Section 2.4 and Appendix A.

## 1.2 New contribution

In this article, we consider the case where the distribution of the HMT is parametrized by some vector  $\theta$ , that is, the transition kernel  $Q_\theta$  between the hidden variables and the transition kernel  $G_\theta$  from hidden variables to observed variables both depend on  $\theta$ . As an example, if the hidden state space  $\mathcal{X}$  is finite and  $Y_u$  conditioned on  $X_u$  is a Gaussian random variable for each  $u \in T$ , then  $\theta$  could parametrized the transition matrix of the hidden process and the mean and variances of the Gaussian distribution associated to each hidden state values. Our goal is to estimate the true parameter  $\theta^*$  of the HMT process among a compact set of possible parameters  $\Theta \subset \mathbb{R}^d$ , for some integer  $d$ , using only the knowledge of the observed process  $Y$  over  $n$  generations of the tree. Note that as our assumptions will imply uniform exponential memorylessness properties for the initial distribution, we cannot try estimate the initial distribution. Denote  $\partial$  the root of the tree  $T$ . Thus, we assume that the distribution of the hidden root variable  $X_\partial$  is some unknown measure  $\zeta$  which does not depend on  $\theta$ . Denote by  $\mathbb{P}_{\theta^*, \zeta}$  the probability distribution of the HMT under the true parameter  $\theta^*$  when the initial unknown distribution of  $X_\partial$  is  $\zeta$ . When  $\zeta$  is the unique invariant measure of  $Q_\theta$  (i.e. in the stationary case), we write  $\mathbb{P}_{\theta^*}$  instead of  $\mathbb{P}_{\theta^*, \zeta}$ .

To estimate the true parameter  $\theta^*$  of the HMT, we will use the maximum likelihood estimator (MLE). We will work with the likelihood conditioned on the hidden state of the root vertex  $X_\partial$ . The reason to do this is that the computation of the stationary distribution of the joint process  $(X, Y)$ , and thus also the true likelihood, is intractable in typical applications. Note that the idea of conditioning on the initial hidden state was already used in [DMR04] for HMMs with the same motivation, and conditioning on initial observations in time series goes back at least to [MW43]. Remind that  $T$  denote the (deterministic) complete infinite rooted binary tree. Denote  $T_n$  the tree  $T$  up to and including the  $n$ -th generation. Hence, for any value  $x \in \mathcal{X}$ , we denote by  $\ell_{n,x}(\theta)$  the log-likelihood under the parameter  $\theta$  of the observed process  $(Y_u, u \in T_n)$  until the  $n$ -th generation of the tree  $T$  conditionally on  $X_\partial = x$  (see (7) on page 11 for exact definition). Then, for any value  $x \in \mathcal{X}$ , we define the MLE  $\hat{\theta}_{n,x}$  as the maximizer of  $\ell_{n,x}$  over  $\Theta$  (see (33) on page 22 for exact definition).

Our goal is to study the asymptotic properties of the MLE. We prove the strong consistency and the asymptotic normality of the MLE in the stationary case in Sections 3 and 4, respectively. We then extend those results to the non-stationary case in Section 5. In our results, the hidden state space  $\mathcal{X}$  and the observed state space  $\mathcal{Y}$  are both general metric spaces. We prove our results under the same assumptions used in [DMR04] and in [CMR05, Chapter 12] for HMMs with  $L^1$

and  $L^2$  integrability assumptions replaced by  $L^2$  and  $L^4$  integrability assumptions, respectively, to accommodate the stronger assumptions needed in ergodic theorems for branching Markov chains. See Remark 1.6 below for a discussion on the main differences between the HMM case as in [DMR04, CMR05] and the HMT case we develop in this article.

We first state that strong consistency of the MLE holds under standard assumptions for HMMs. Following [DMR04], we assume a fully dominated model, that is, the transition kernels  $Q_\theta$  and  $G_\theta$  admits densities  $q_\theta$  and  $g_\theta$  w.r.t. to common measures  $\lambda$  and  $\mu$ , respectively (see Assumption 2). We also assume (see Assumption 3) :

$$0 < \sigma^- \leq \inf_{x, x' \in \mathcal{X}} q_\theta(x, x') \leq \sup_{x, x' \in \mathcal{X}} q_\theta(x, x') \leq \sigma^+ < \infty. \quad (1)$$

This assumption is rather strong as it imposes a full connection for the hidden space, see [KS19] for an extension of the method in [DMR04] for HMMs where  $q_\theta$  is allowed to be zero valued. Nevertheless, this assumption implies the uniform exponential memorylessness properties with mixing rate  $\rho := 1 - \sigma^- / \sigma^+$  of the initial distribution conditional on the observations  $(Y_u, u \in T_n)$ . The other assumptions are more standard regularity assumptions for the densities  $q_\theta$  and  $g_\theta$  (see Assumptions 2-6), and identifiability of the model. We can now state the strong consistency of the MLE under those assumptions, see Theorems 3.11 and 5.1 for the precise statements in the stationary and non-stationary case, respectively.

**Theorem 1.1** (Strong consistency of the MLE). *Under those assumptions of fully dominated model with density satisfying (1) and other more standard regularity assumptions, and under the assumption that the model is identifiable, for any  $x \in \mathcal{X}$ , the MLE  $\hat{\theta}_{n,x}$  is strongly consistent, that is, the sequence  $(\hat{\theta}_{n,x})_{n \in \mathbb{N}}$  converges  $\mathbb{P}_{\theta^*, \zeta}$ -almost surely to the true parameter  $\theta^* \in \Theta$ .*

To prove asymptotic normality of the MLE, in addition to the assumptions used in Theorem 1.1, we need existence and regularity assumptions for the gradient and the Hessian of the transition densities  $q_\theta$  and  $g_\theta$  (see Assumptions 7-9). Denote by  $\mathcal{I}(\theta^*)$  the limiting Fisher information matrix of the model (see (54) on page 32 for precise definition). The proof of asymptotic normality in the non-stationary case is an extension of the stationary case. The proof of asymptotic normality in the stationary case follows from a standard argument for asymptotic normality of the MLE that relies on Theorem 1.1 and Theorems 1.2 and 1.3 below.

The following theorem, which we only prove in the stationary case, states that the normalized score  $|T_n|^{-1/2} \nabla_\theta \ell_{n,x}(\theta^*)$  has asymptotic normal fluctuations with covariance matrix  $\mathcal{I}(\theta^*)$ , see Theorem 4.3 for the precise statement. Note that the extra assumption in Theorem 1.2 (not present in the case of HMMs) that  $\rho < 1/\sqrt{2}$  for the mixing rate  $\rho$  of the HMT process comes from the approximation bounds used in the proof of this theorem. See Remark 1.5 below for a discussion on this condition on  $\rho$ .

**Theorem 1.2** (Asymptotic normality of the normalized score). *Under the assumptions from Theorem 1.1 and existence and regularity assumptions for the gradient and the Hessian of the transition densities (see Assumptions 7-9), and under the assumption that  $\rho < 1/\sqrt{2}$  for the mixing rate  $\rho$  of the HMT process, in the stationary case we have:*

$$|T_n|^{-1/2} \nabla_\theta \ell_{n,x}(\theta^*) \xrightarrow[n \rightarrow \infty]{(d)} \mathcal{N}(0, \mathcal{I}(\theta^*)) \quad \text{under } \mathbb{P}_{\theta^*}.$$

The following theorem states the locally uniform convergence  $\mathbb{P}_{\theta^*, \zeta}$ -a.s. of the normalized observed information  $-|T_n|^{-1} \nabla_\theta^2 \ell_{n,x}(\theta)$  towards the Fisher information matrix  $\mathcal{I}(\theta^*)$ , see Theorems 4.6 and 5.2 for the precise statements in the stationary and non-stationary case, respectively. Note that in this theorem we need the stronger assumption  $\rho < 1/2$  for the mixing rate  $\rho$  of the

HMT process as we use more restrictive approximation bounds in the proof of this theorem than the ones used in the proof of Theorem 1.2.

**Theorem 1.3** (Convergence of the normalized observed information). *Under the assumptions from Theorem 1.2 on the HMT model, and under the assumption that  $\rho < 1/2$  for the mixing rate  $\rho$  of the HMT process, for all  $x \in \mathcal{X}$ , we have:*

$$\lim_{\delta \rightarrow 0} \lim_{n \rightarrow \infty} \sup_{\theta \in \Theta: \|\theta - \theta^*\| \leq \delta} \left\| -|T_n|^{-1} \nabla_{\theta}^2 \ell_{n,x}(\theta) - \mathcal{I}(\theta^*) \right\| = 0 \quad \mathbb{P}_{\theta^*, \zeta}\text{-a.s.}$$

In particular, combining Theorems 1.1 and 1.3, we get that the normalized observed information  $-|T_n|^{-1} \nabla_{\theta}^2 \ell_{n,x}(\hat{\theta}_{n,x})$  at the MLE  $\hat{\theta}_{n,x}$  is a strongly consistent estimator of the Fisher information matrix  $\mathcal{I}(\theta^*)$ .

As announced above, following a standard argument for asymptotic normality of the MLE, Theorems 1.1, 1.2 and 1.3 imply the following theorem which states that the MLE has asymptotic normal fluctuations with covariance matrix  $\mathcal{I}(\theta^*)^{-1}$ . See Theorems 4.7 and 5.5 for the precise statements in the stationary and non-stationary case, respectively.

**Theorem 1.4** (Asymptotic normality of the MLE). *Under the assumptions from Theorem 1.2 on the HMT model, that  $\theta^*$  is an interior point of  $\Theta$ , and the Fisher information matrix  $\mathcal{I}(\theta^*)$  is non-singular, and under the assumption that  $\rho < 1/2$  for the mixing rate  $\rho$  of the HMT process, we have the following convergence in distribution:*

$$|T_n|^{1/2} (\hat{\theta}_n - \theta^*) \xrightarrow[n \rightarrow \infty]{(d)} \mathcal{N}(0, \mathcal{I}(\theta^*)^{-1}) \quad \text{under } \mathbb{P}_{\theta^*, \zeta},$$

where  $\mathcal{N}(0, M)$  denotes the centered Gaussian distribution with covariance matrix  $M$ .

Note that the standard argument used in the proof of Theorem 1.4 implies that we have the following joint convergence in distribution:

$$\left( |T_n|^{1/2} (\hat{\theta}_n - \theta^*), |T_n|^{-1/2} \nabla_{\theta} \ell_{n,x}(\theta^*) \right) \xrightarrow[n \rightarrow \infty]{(d)} (\mathcal{I}(\theta^*)^{-1/2} G, \mathcal{I}(\theta^*)^{1/2} G) \quad \text{under } \mathbb{P}_{\theta^*},$$

where  $G$  is Gaussian random variable distributed as  $\mathcal{N}(0, I_d)$  with  $I_d$  the identity matrix of dimension  $d \times d$ , and  $\mathcal{I}(\theta^*)^{1/2}$  is a root matrix of  $\mathcal{I}(\theta^*)$ .

The following remark is a discussion on the condition on the mixing rate  $\rho$  of the HMT process  $(X, Y)$  that appear in Theorems 1.4, 1.2 and 1.3.

**Remark 1.5** (On the condition on the mixing rate  $\rho$ ). Note that in central limit theorems for branching Markov chains, three regimes with different asymptotic behaviors (and different normalization terms) for  $\rho < 1/\sqrt{2}$ ,  $\rho = 1/\sqrt{2}$  and  $\rho > 1/\sqrt{2}$  were observed in [BPD22a], corresponding to a competition between the ergodic mixing rate  $\rho$  and the branching rate 2 in  $T$ , see also [Ath69, BPDG14, BPD22b]. However, the condition on  $\rho$  disappears when we consider martingale increments in the central limit theorem for branching Markov chains, see [Guy07, BDSG09, DM10].

In our case, the condition  $\rho < 1/\sqrt{2}$  on the mixing rate  $\rho$  that appears in Theorem 1.2 is due to the coupling bounds and the grouping of terms used in the proof of Lemma 4.2 (the upper bounds at the end of the proof only add a constant multiplicative factor). It is an open question whether or not some convergence would still hold in Theorem 1.2 with  $\rho \geq 1/\sqrt{2}$  even with a possibly stronger normalization term and a possibly non Gaussian limit. Nevertheless, note that the proof of Theorem 1.2 relies on decomposing the score  $\nabla_{\theta} \ell_{n,x}(\theta)$  as a sum of martingale increments, which could indicate that convergence is possible for  $\rho \geq 1/\sqrt{2}$ .

Moreover, the stronger condition  $\rho < 1/2$  on the mixing rate  $\rho$  that appears in Theorem 1.3, and thus in Theorem 1.4, is due to the coupling bounds from Lemma 4.16 and the grouping of terms used in the proof of Lemma 4.17 (the upper bounds in the rest of the proof only add a constant multiplicative factor). It is an open question whether or not convergence would still hold in Theorem 1.3 and in Theorem 1.4 with  $\rho \geq 1/2$  even with a possibly stronger normalization term and a possibly non Gaussian limit in Theorem 1.4. Also note that the condition  $\rho < 1/2$  is used when proving that Theorem 1.4 extends to the non-stationary case to construct a coupling between a stationary HMT process and a non-stationary HMT process, see Lemma 5.3.

In the following remark, we discuss the main differences between the HMM case as in [DMR04, CMR05] and the HMT case we develop in this article.

**Remark 1.6** (On main differences with the HMM case). In both HMM and HMT cases, the study of the log-likelihood is based on decomposing it as a sum of increments, and then extending the “past” seen by each variable. However, while the extended “past” only spreads backwards in the HMM case, the extended “past” in the HMT case is a subtree that also spreads laterally due to the different topologies between the line  $\mathbb{Z}$  and the binary tree, see Figure 3 on page 12 for an illustration. See also Sections 2.4 and 3.1 for the definition of those “past” and extended “past”. Moreover, due to the enumeration of vertices in the tree in a breadth-first-search manner, those extended “past” do not have the same “shapes” for all vertices, see Section 2.4. Also note that the infinite expanded “past” of a vertex relies on a random infinite “backward spine” of left / right ancestors (see Figure 4 on page 15), which adds extra randomness to the “shape” of the “past”.

Furthermore, contrary to the HMM case, the lateral spreading of each vertex’s “past” in the HMT case implies that log-likelihood increments with infinite extended “pasts” do not form a branching Markov process. For this reason, we need to work with log-likelihood increments whose “past” is trimmed to a fixed common subtree height, and only expand to infinite “past” in the limit. To prove convergence for sums of log-likelihood increments with trimmed “pasts” which have different shapes, we need to develop new ergodic theorems for branching Markov chains and neighborhood-dependent functions, see Section 2.4 and Appendix A.

In the proof of asymptotic normality of the normalized score, the score is decomposed as a sum of martingale increments which is no longer stationary in the HMT case due to the “pasts” of vertices having different shapes. Thus, to apply the central limit theorem for martingales, we first need to verify convergence for the quadratic variations of the martingale increment sequences and Lindeberg’s condition. Moreover, the computation of the approximation bounds for the increments used to decompose the score and the observed information are more involved and impose conditions on the value of the mixing rate  $\rho$ , as already discussed in Remark 1.5. This also implies that the proof scheme for convergence of the observed information matrix needs to be modified as we cannot have almost sure convergence for all the increments simultaneously, and we must rely on  $L^2$  convergence instead.

Lastly, as discussed in Section 1.1, the results for HMMs in [DMR04] allowed for autoregression (remind, that is, when conditionally on the hidden Markov chain, the observed process is an inhomogeneous  $s$ -order Markov chain for some  $s \in \mathbb{N}$ ). Our results for HMTs are stated for processes without autoregression. However, as our approach adapts the proof scheme of [DMR04], note that with straightforward modifications of our proofs, we could allow for autoregression in HMT processes.

### 1.3 Organization of the paper

The rest of the paper is organized as follows. In Section 2, we define the notations used in this article, HMT processes and the log-likelihood for the HMT. For the stationary case, we prove the strong consistency of the MLE in Section 3, and its asymptotic normality in Section 4. In Section 5, we extend those results to the non-stationary case. In Appendix A, we develop the ergodic theorems for branching Markov chains with neighborhood-dependent functions needed in the proofs of the asymptotic properties of the MLE.

## 2 Definition of HMT and notations

In this section, we first define the notations we use for the infinite complete binary tree  $T$ . We then define branching Markov chains and hidden Markov models (HMMs) indexed by the binary tree  $T$ , which we will simply call Hidden Markov Tree (HMT) models. We continue with the basic assumptions we need to define the log-likelihood for the HMT. Lastly, we present the ergodic theorems for branching Markov chains and neighborhood-dependent functions needed in this article, whose proofs can be found in Appendix A.

### 2.1 Notations for trees

Let  $T = \cup_{n \in \mathbb{N}} \{0, 1\}^n$  denote the infinite complete plane rooted binary tree, that is the plane rooted tree where each vertex  $u$  has exactly two children  $u0$  and  $u1$ . Denote by  $\partial$  the root vertex of  $T$  (which is the unique point in  $\{0, 1\}^0$ ). If  $u$  is distinct from the root, we denote by  $p(u)$  its parent vertex. We denote by  $h(u)$  its height, i.e. the number of edges separating  $u$  from the root  $\partial$ . (The height of the root  $\partial$  is zero.) In particular, for  $k \leq h(u)$ , note that  $p^k(u)$  denotes the  $k$ -th ancestor of  $u$ . For two vertices  $u, v \in T$ , we denote by  $u \wedge v$  the most recent common ancestor of  $u$  and  $v$ , and by  $d(u, v)$  the graph-distance between  $u$  and  $v$  in  $T$ , that is  $d(u, v) = h(u) + h(v) - 2h(u \wedge v)$ . For all  $n \in \mathbb{N}$ , denote by  $G_n$  the  $n$ -th generation of the tree, that is vertices that are at distance  $n$  from the root, and denote by  $T_n = \cup_{0 \leq k \leq n} G_k$  the tree up to generation  $n$ . For a vertex  $u \in T$ , we denote by  $T(u)$  the subtree of  $T$  composed of descendants of  $u$ , and for all  $k \in \mathbb{N}$ , we denote by  $T(u, k) = T(u) \cap T_{h(u)+k}$  the subtree of  $T(u)$  composed of descendants of  $u$  at distance up to  $k$  from  $u$ . We will use the convention that for a subtree  $T_{\text{sub}}$  of  $T$ , we write  $T_{\text{sub}}^*$  for the subtree  $T_{\text{sub}}$  without its root vertex, for instance,  $T_n^* = T_n \setminus \{\partial\}$  and  $T(u)^* = T(u) \setminus \{u\}$ . For a finite subset  $A \subset T$ , we denote by  $|A|$  its cardinal.

We will sometimes use Neveu's notation, which we define recursively: a vertex  $u \in T$  with height  $h(u) = n$  can be represented as a sequence  $(u_{(j)})_{1 \leq j \leq n}$  where  $u$  is the  $u_{(n)}$ -th child of  $p(u)$  and  $p(u)$  can be represented by  $(u_{(j)})_{1 \leq j \leq n-1}$ ; and the representation of the root  $\partial$  is the empty sequence. Note that Neveu's notation can also be interpreted as encoding the path from the root  $\partial$  to the vertex  $u$ : starting from the root  $u_0 = \partial$ , at each generation  $j$  we go from  $u_j$  to its  $u_{(j+1)}$  child which we denote by  $u_{j+1}$ , and at generation  $n$  we get  $u_n = u$ .

For simplicity, we will write  $u_{(k:n)} = (u_{(j)})_{k \leq j \leq n}$  and  $u_{(k:n)} = (u_{(j)})_{k \leq j \leq n}$  for path sequences where  $k, n \in \mathbb{Z}$  with  $k < n$ .

As  $T$  is a plane rooted tree, we can order its vertices in a breadth-first-search manner, that is, the total order relation  $\leq$  on  $T$  is defined for all  $u, v \in T$  as  $u \leq v$  if  $h(u) < h(v)$  or  $h(u) = h(v)$  and  $u \leq_{\text{lex}} v$  (where  $\leq_{\text{lex}}$  is the lexicographical order on  $T$ ). Moreover, we denote by  $v < u$  if  $u \leq v$  and  $v \neq u$ .



## 2.2 Definition of HMT processes

For a sequence  $(x_u, u \in T)$ , for simplicity, we will write  $x_A = (x_u, u \in A)$  for all subsets  $A \subset T$ . For a metric space  $\mathcal{X}$ , we will always assume it is equipped with its Borel  $\sigma$ -field  $\mathcal{B}(\mathcal{X})$ .

For a measure  $\mu$  on a metric space  $\mathcal{X}$  and an integrable function  $f \in L^1(\mu)$ , we write  $\mu(f) = \int_{\mathcal{X}} f d\mu$ . For two probability measures  $\mu_1, \mu_2$  on a metric space  $\mathcal{X}$ , we denote the total variation norm between them by  $\|\mu_1 - \mu_2\|_{\text{TV}} = \sup_{A \subset \mathcal{X}} |\mu_1(A) - \mu_2(A)|$  (where  $A$  ranges over all measurable subsets of  $\mathcal{X}$ ). We also remind the identities  $\|\mu_1 - \mu_2\|_{\text{TV}} = \frac{1}{2} \sup_{f: |f| \leq 1} |\mu_1(f) - \mu_2(f)| = \sup_{f: 0 \leq f \leq 1} |\mu_1(f) - \mu_2(f)|$  (where  $f$  is a measurable function). Note that  $\|\mu_1 - \mu_2\|_{\text{TV}}$  takes value in  $[0, 1]$ .

Denote by  $X = (X_u, u \in T)$  the hidden (stochastic) process with values in a metric space  $\mathcal{X}$ , and by  $Y = (Y_u, u \in T)$  the observed (stochastic) process with values in a metric space  $\mathcal{Y}$ . We assume that the hidden process  $X$  is a branching Markov process.

**Definition 2.1** (Branching Markov process). *The stochastic process  $X$  is called a (branching) Markov process with transition (probability) kernel  $Q$  on  $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$  and initial (probability) distribution  $\nu$  on  $\mathcal{X}$  if for all  $n \in \mathbb{N}$ , we have:*

$$\mathbb{P}(X_{T_n} \in dx_{T_n}) = \nu(dx_{\partial}) \prod_{u \in T_n^*} Q(x_{p(u)}; dx_u).$$

We can now define the Hidden Markov Tree process.

**Definition 2.2** (Hidden Markov Tree process). *The stochastic process  $(X, Y) = ((X_u, Y_u), u \in T)$  is called a Hidden Markov Tree (HMT) with parameter  $(Q, G, \nu)$  if:*

- (i) *the hidden process  $X = (X_u, u \in T)$  is a branching Markov process with transition kernel  $Q$  and initial distribution  $\nu$ ,*
- (ii) *the observed process  $Y = (Y_u, u \in T)$  conditioned on the hidden process  $X$  is composed of independent variables whose laws factorize using the transition (probability) kernel  $G$  on  $(\mathcal{X}, \mathcal{B}(\mathcal{Y}))$ , that is for all  $n \in \mathbb{N}$ , we have:*

$$\mathbb{P}(Y_{T_n} = y_{T_n} \mid X_{T_n} = x_{T_n}) = \prod_{u \in T_n} G(x_u; dy_u).$$

Note that the definitions of branching Markov chains and HMT processes also work for non-plane rooted trees.

In particular, note that if  $(X, Y) = ((X_u, Y_u), u \in T)$  is a HMT process, then the joint process  $((X_u, Y_u), u \in T)$  is also a branching Markov chain (but the observed process  $Y$  is not necessarily Markov). The following fact, which we shall reuse later, illustrates the Markov property of the HMT process  $(X, Y)$ . For any  $k \in \mathbb{N}^*$ , any  $u \in T$  with height at least  $k$ , and any subset  $A \subset T$ , we have :

$$\mathcal{L}(X_u \mid Y_A, X_{p^k(u)}) = \mathcal{L}(X_u \mid Y_{A \cap T(p^{k-1}(u))}, X_{p^k(u)}) \quad (2)$$

where  $\mathcal{L}(R \mid S)$  denotes the distribution of a random variable  $R$  conditionally on another random variable  $S$ .

We say that a branching Markov process  $X$  is *stationary* if all its variables are identically distributed (that is, for all  $u \in T$ ,  $X_u$  has the same distribution as  $X_{\partial}$ ), or equivalently if its initial distribution  $\nu$  is invariant for its transition kernel  $Q$  (that is,  $\nu Q = \nu$ ). Moreover, if  $(X, Y)$  is a HMT process and the hidden process  $X$  is stationary, then the joint process  $(X, Y)$  is also stationary.

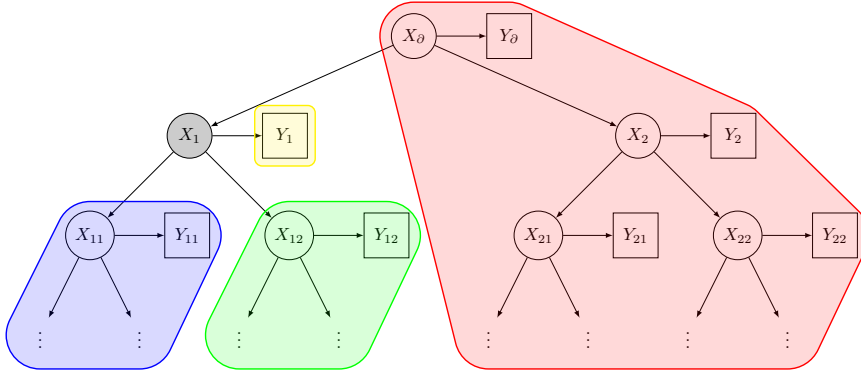


Figure 2: Illustration of the Markov property for the HMT. Conditioning on  $X_1$  (in grey) implies that the HMT process  $(X, Y)$  becomes independent between the four connected components of variable-dependence tree from Figure 1 where the vertex  $X_1$  is removed, that is,  $Y_1$ ,  $(X_{T(11)}, Y_{T(11)})$ ,  $(X_{T(12)}, Y_{T(12)})$  and  $(X_{T \setminus T(1)}, Y_{T \setminus T(1)})$  (respectively in yellow, blue, green and red) are independent conditionally on  $X_1$ .

### 2.3 Basic assumptions and definition of the log-likelihood

In this article, we consider the case where the kernels in the definition of the HMT  $(Q_\theta, G_\theta, \nu_\theta)$  are parametrized by some vector  $\theta$  that we want to estimate using only the knowledge of the observed process  $(Y_u, u \in T_n)$  up to generation  $n$ . We denote by  $\Theta$  the set of all possible vector parameters  $\theta$ , which we assume to be a subset of  $\mathbb{R}^d$  for some integer  $d$ . And we denote by  $\theta^*$  the true parameter of the HMT.

Through this article, with the exception of Section 5, we assume that the hidden process  $X$  is stationary.

**Assumption 1** (Stationarity). *The hidden process  $(X_u, u \in T)$  is stationary, (and thus the joint process  $((X_u, Y_u), u \in T)$  is also stationary).*

We denote by  $\mathbb{P}_\theta$  the probability distribution under the parameter  $\theta$  of the stationary joint process  $(X, Y)$ , and by  $\mathbb{E}_\theta$  the corresponding expectation.

To prove asymptotic properties of the MLE for the HMT, we will use assumptions similar to the HMM case in [CMR05, Chapter 12] and [DMR04]. We first assume that the HMT model is fully dominated. For two probability measures  $\lambda, \mu$  on the same space, we write  $\lambda \ll \mu$  to denote that  $\lambda$  is absolutely continuous w.r.t. to  $\mu$ .

**Assumption 2** (Fully dominated model, [CMR05, Assumption 12.0.1]).

- (i) *There exists a probability measure  $\lambda$  on  $\mathcal{X}$  such that for every  $x \in \mathcal{X}$  and every  $\theta \in \Theta$ ,  $Q_\theta(x, \cdot) \ll \lambda$ , with density  $q_\theta(x, \cdot)$ . That is,  $Q_\theta(x; A) = \int_A q_\theta(x, x') \lambda(dx')$  for all  $A \in \mathcal{B}(\mathcal{X})$ . Moreover, the density function  $q_\theta(\cdot, \cdot)$  is  $\mathcal{B}(\mathcal{X}) \otimes \mathcal{B}(\mathcal{X})$  measurable.*
- (ii) *There exists a  $\sigma$ -finite measure  $\mu$  on  $\mathcal{Y}$  such that for every  $x \in \mathcal{X}$  and every  $\theta \in \Theta$ ,  $G_\theta(x, \cdot) \ll \mu$ , with density  $g_\theta(x, \cdot)$ . That is,  $G_\theta(x; B) = \int_B g_\theta(x, y) \mu(dy)$  for all  $B \in \mathcal{B}(\mathcal{Y})$ . Moreover, the density function  $g_\theta(\cdot, \cdot)$  is  $\mathcal{B}(\mathcal{X}) \otimes \mathcal{B}(\mathcal{Y})$  measurable.*

In addition to Assumption 2, we use the following regularity assumptions on the density functions  $q_\theta$  and  $g_\theta$ .

**Assumption 3** (Regularity, [CMR05, Assumption 12.2.1]).

- (i) The transition density  $q_\theta$  is bounded: there exist  $\sigma^-, \sigma^+ \in (0, +\infty)$  such that  $\forall x, x' \in \mathcal{X}, \forall \theta \in \Theta, 0 < \sigma^- \leq q_\theta(x, x') \leq \sigma^+ < +\infty$ .
- (ii) For every  $y \in \mathcal{Y}$ , the function  $\theta \mapsto \int_{\mathcal{X}} g_\theta(x, y) \lambda(dx)$  is bounded away from 0 and  $\infty$  uniformly on  $\Theta$ , that is,  $\sup_{\theta \in \Theta} \int_{\mathcal{X}} g_\theta(x, y) \lambda(dx) < \infty$  and  $\inf_{\theta \in \Theta} \int_{\mathcal{X}} g_\theta(x, y) \lambda(dx) > 0$ .
- (iii) For every  $x \in \mathcal{X}$  and  $y \in \mathcal{Y}$ , we have  $g_\theta(x, y) > 0$ .

We will denote by  $\rho = 1 - \sigma^-/\sigma^+ \in (0, 1)$  the *mixing rate* of the hidden process  $X$ .

Note that Assumption 3-(iii) is due to our choice of conditioning on  $X_\partial = x$  for some  $x \in \mathcal{X}$  in the log-likelihood we analyse, we discuss how to get rid of this assumption after the definition of the log-likelihood at the end of this subsection.

As  $\lambda$  is a probability measure and  $q_\theta$  is the density of a probability kernel, Assumption 3-(i) implies that  $\sigma^- \leq 1 \leq \sigma^+$ . Moreover, Assumption 3-(i) also implies the following result.

**Lemma 2.3.** *Assume that Assumptions 2-(i) and 3-(i) hold. Then, for all  $\theta \in \Theta$ , the transition kernel  $Q_\theta$  has a unique invariant measure  $\pi_\theta$  and is uniformly geometrically ergodic, that is:*

$$\forall x \in \mathcal{X}, \forall \theta \in \Theta, \forall n \geq 0, \quad \|Q_\theta^n(x, \cdot) - \pi_\theta\|_{TV} \leq (1 - \sigma^-)^n \leq \rho^n.$$

Note that due to structure of the HMT, Lemma 2.3 extends naturally to the transition kernel of the joint process  $(X, Y)$  with the same mixing rate  $\rho$ . Moreover, note that Assumption 3-(i) also implies that  $\pi_\theta \ll \lambda$  with density  $\frac{d\pi_\theta}{d\lambda}$  taking value in  $[\sigma^-, \sigma^+]$ .

Lemma 2.3 is due to Assumption 3-(i) implying the Doeblin condition:

$$\forall \theta \in \Theta, \forall x \in \mathcal{X} \quad \sigma^- \lambda(\cdot) \leq Q_\theta(x; \cdot). \quad (3)$$

As we will reuse Doeblin conditions later, before proving Lemma 2.3, we give a quick summary on results for the Doeblin condition. For a transition kernel  $K$  on a metric space  $\mathcal{X}$  (to itself), we define its *Dobrushin coefficient*  $\delta(K)$  as:

$$\delta(K) = \sup_{x, x' \in \mathcal{X}} \|K(x; \cdot) - K(x'; \cdot)\|_{TV}. \quad (4)$$

The Dobrushin coefficient gives the following coupling bound in the total variation norm. (Note that the definition of the total variation norm  $\|\cdot\|_{TV}$  used in [CMR05, Chapter 4] differs by a factor 2 from ours, see [CMR05, Lemma 4.3.5].)

**Lemma 2.4** ([CMR05, Lemma 4.3.8]). *Let  $\mu_1, \mu_2$  be two probability measures on a metric space  $\mathcal{X}$ , and let  $K$  be a transition kernel on  $\mathcal{X}$ . Then, we have:*

$$\|(\mu_1 - \mu_2)K\|_{TV} \leq \delta(K) \|\mu_1 - \mu_2\|_{TV} \leq \delta(K).$$

Moreover, the Dobrushin coefficient is sub-multiplicative.

**Lemma 2.5** ([CMR05, Proposition 4.3.10]). *The Dobrushin coefficient is sub-multiplicative. That is, if  $K, R$  are two transition kernels on a metric space  $\mathcal{X}$ , then we have  $\delta(KR) \leq \delta(K)\delta(R)$ .*

We now define the Doeblin condition.

**Definition 2.6** (Doeblin condition, [CMR05, Definition 4.3.12]). *We say that a transition kernel  $K$  on a metric space  $\mathcal{X}$  satisfies a Doeblin condition if there exist  $\varepsilon > 0$  and a probability measure  $\nu$  on  $\mathcal{X}$  such that for all  $x \in \mathcal{X}$  and measurable subset  $A \subset \mathcal{X}$ , we have:*

$$K(x; A) \geq \varepsilon \nu(A).$$

The Doeblin condition gives an upper bound on the Dobrushin coefficient.

**Lemma 2.7** ([CMR05, Lemma 4.3.13]). *Let  $K$  be transition kernel (on a metric space  $\mathcal{X}$ ) that satisfies a Doeblin condition with  $(\varepsilon, \nu)$ . Then, we have  $\delta(K) \leq 1 - \varepsilon$ .*

Lastly, the Doeblin condition implies the existence of a unique invariant probability measure, as well as uniform geometric ergodicity.

**Lemma 2.8** ([CMR05, Theorem 4.3.16]). *Let  $K$  be a transition kernel on a metric space  $\mathcal{X}$  that satisfies a Doeblin condition with  $(\varepsilon, \nu)$ . Then,  $K$  admits a unique invariant probability measure  $\pi$ . Moreover, for any probability measure  $\zeta$  on  $\mathcal{X}$ , we have for all  $n \in \mathbb{N}$ :*

$$\|\zeta K^n - \pi\|_{TV} \leq (1 - \varepsilon)^n \|\zeta - \pi\|_{TV}.$$

Lemma 2.3 then follows immediately from Lemma 2.8 and the uniform Doeblin condition (3).

**Remark 2.9** (More properties of the transition kernel from the Doeblin condition). For a transition kernel  $K$  on a metric space  $\mathcal{X}$ , the Doeblin condition also implies that  $\mathcal{X}$  is an (accessible) 1-small set. In particular, we get that  $K$  satisfies some extra classical properties (that we will not use here):  $K$  is positive (i.e. irreducible and admits a unique invariant probability measure), strongly aperiodic and Harris recurrent (see [DMPS18, Chapter 9 and 10] for definitions and details).

We will use the letter  $p$  to denote (possibly conditional) probability density. For instance, for any  $\theta \in \Theta$ ,  $y_{T_n} \in \mathcal{Y}^{T_n}$  and  $x_\partial \in \mathcal{X}$ , we denote by:

$$p_\theta(y_{T_n} | X_\partial = x_\partial) = g_\theta(x_\partial, y_\partial) \int_{\mathcal{X}^{T_n^*}} \prod_{v \in T_n^*} q_\theta(x_{p(v)}, x_v) g_\theta(x_v, y_v) \lambda(dx_v), \quad (5)$$

the conditional density w.r.t.  $\mu^{\otimes T_n}$  under the parameter  $\theta$  of  $Y_{T_n} = y_{T_n}$  conditionally on  $X_\partial = x_\partial$ . Note that Assumption 3 guarantees that  $p_\theta(y_{T_n} | X_\partial = x_\partial)$  is positive for all  $y_{T_n} \in \mathcal{Y}^{T_n}$  and  $x_\partial \in \mathcal{X}$ .

We are now ready to define the log-likelihood. As discussed in Section 1.2, we will analyze the log-likelihood of the observed process  $(Y_u, u \in T_n)$  up to generation  $n$  conditioned on the hidden value of the root  $X_\partial = x$  for some  $x \in \mathcal{X}$ . Thus, for any  $x \in \mathcal{X}$ , we define the log-likelihood function as:

$$\ell_{n,x}(\theta; y_{T_n}) := \log(p_\theta(y_{T_n} | X_\partial = x)). \quad (6)$$

We then define the log-likelihood that we will analyze as the following random variable

$$\ell_{n,x}(\theta) := \ell_{n,x}(\theta; Y_{T_n}). \quad (7)$$

For simplicity, we will write  $\ell_{n,x}(\theta)$  instead of  $\ell_{n,x}(\theta; Y_{T_n})$  making the dependence on the observed variables  $(Y_u, u \in T_n)$  implicit. We will keep this convention for all quantities considered in this article, and only make the dependence explicit when necessary. The MLE is then the maximizer over  $\Theta$  of the log-likelihood  $\ell_{n,x}$ ; we post-pone the precise definition of the MLE to when we will first use it in Theorem 3.11.

**Remark 2.10** (On Assumption 3-(iii)). Note that Assumption 3-(iii) is due to our choice of conditioning on  $X_\partial = x$  for some  $x \in \mathcal{X}$  in the log-likelihood  $\ell_{n,x}(\theta)$  we analyse. Indeed, without Assumption 3-(iii), there could be a non-zero probability under  $\mathbb{P}_{\theta^*}$  that  $g_{\theta^*}(x, Y_\partial) = 0$  for some  $x \in \mathcal{X}$ , implying  $\ell_{n,x}(\theta^*) = -\infty$ , and thus preventing the MLE to converge to  $\theta^*$ . Several modifications of the log-likelihood  $\ell_{n,x}(\theta)$  can be considered to get rid of Assumption 3-(iii). A first option would be to replace  $p_\theta(y_{T_n} | X_\partial = x)$  by  $p_\theta(y_{T_n^*} | X_\partial = x)$  in (6). A second option would be to extend the tree  $T$  and the HMT  $(X, Y)$  to add a parent vertex  $p(\partial)$  for the root vertex  $\partial$  (see Section 3.1.1), and then replace  $p_\theta(y_{T_n} | X_\partial = x)$  by  $p_\theta(y_{T_n} | X_{p(\partial)} = x)$  in (6).

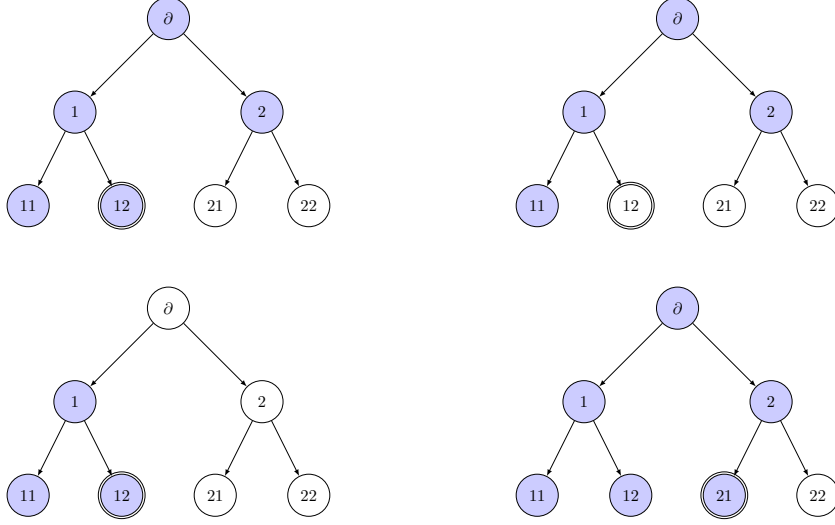


Figure 3: Illustration of the subtrees  $\Delta(u, k)$  and  $\Delta^*(u, k)$  where vertices in  $\Delta(u, k)$  and  $\Delta^*(u, k)$  are in blue and the vertex  $u$  is inside a double circle. From left to right and top to bottom, using Neveu's notation, we have  $\Delta(12) = \Delta(12, 2)$ ,  $\Delta^*(12) = \Delta^*(12, 2)$ ,  $\Delta(12, 1)$  and  $\Delta(21) = \Delta(21, 2)$ . Note that  $\Delta(12)$  and  $\Delta(21)$  have a different number of vertices, while vertices 12 and 21 are in the same generation.

## 2.4 Ergodic theorems with neighborhood-dependent functions

For all  $u \in T$  and  $0 \leq k \leq h(u)$ , define the subtrees of  $T$ :

$$\Delta^*(u, k) = \{v \in T(p^k(u)) : v < u\},$$

and  $\Delta(u, k) = \Delta^*(u, k) \cup \{u\}$ . In particular, note that when  $k = h(u)$ , we have that  $\Delta^*(u) := \Delta^*(u, h(u)) = \{v \in T : v < u\}$ , and we also write  $\Delta(u) := \Delta(u, h(u))$ . See Figure 3 for an illustration of those subtrees. The subtree  $\Delta^*(u)$  represents the past of the vertex  $u$ .

For the ergodic convergence results needed in this article, we will need to consider different functions for each vertex  $u \in T$  depending on the “shape” of the subtree  $\Delta(u, k)$  for some common  $k \in \mathbb{N}$ . For  $k \in \mathbb{N}$  and vertices  $u, v \in T$  both with height at least  $k$ , we say that  $\Delta(u, k)$  and  $\Delta(v, k)$  *have the same shape* if they are equal up to translation, that is, if they are isomorphic as (finite) rooted plane trees. For  $k \in \mathbb{N}$  and any vertex  $u \in T$  with  $h(u) \geq k$ , there exists a unique  $v_u \in G_k$  such that  $\Delta(u, k)$  and  $\Delta(v_u)$  have the same shape, and we thus define the *shape* of  $\Delta(u, k)$  as:

$$Sh(\Delta(u, k)) = \Delta(v_u). \quad (8)$$

Note that as  $|\Delta(v)|$  is different for each  $v \in G_k$ , thus the shape of  $\Delta(u, k)$  is characterized by its size. For any  $k \in \mathbb{N}$ , we define the (finite) set  $\mathcal{N}_k$  of possible shapes for  $\Delta(u, k)$  with  $u \in T$  as:

$$\mathcal{N}_k = \{\Delta(v) : v \in G_k\}. \quad (9)$$

For any  $k \in \mathbb{N}$ , we define a collection of *neighborhood-shape-dependent* functions as a collection of functions  $(f_{\mathcal{S}} : \mathcal{Z}^{\mathcal{S}} \rightarrow \mathbb{R})_{\mathcal{S} \in \mathcal{N}_k}$  where  $\mathcal{Z} \in \{\mathcal{X}, \mathcal{Y}, \mathcal{X} \times \mathcal{Y}\}$ . For such a collection of functions, we will simply write  $f_{\Delta(u, k)}$  instead of  $f_{Sh(\Delta(u, k))}$ . And we will also write  $f_{\Delta(u, k)}(Y_{\Delta(u, k)})$  for the evaluation of  $f_{\Delta(u, k)}$  on  $Y_{\Delta(u, k)}$ . Note that indexing such a collection of functions with  $G_k$  or with  $\mathcal{N}_k$  is equivalent in light of (9).

We prove the following ergodic convergence lemma for neighborhood-shape-dependent functions. The proof of this lemma relies on the theorems in Appendix A. Note that if  $U_n$  is uniformly distributed over  $G_n$  with  $n \geq k$ , then  $\mathcal{Sh}(\Delta(u, k))$  is uniformly distributed over  $\mathcal{N}_k$ .

**Lemma 2.11** (Ergodic theorem for neighborhood-dependent functions). *Assume that Assumptions 1–3 hold. Let  $k \geq 0$ . Let  $(f_S : \mathcal{Y}^S \rightarrow \mathbb{R})_{S \in \mathcal{N}_k}$  be a collection of neighborhood-shape-dependent Borel functions that are in  $L^2(\mathbb{P}_{\theta^*})$ . Then, we have:*

$$\lim_{n \rightarrow \infty} \frac{1}{|T_n|} \sum_{u \in T_n \setminus T_{k-1}} f_{\Delta(u, k)}(Y_{\Delta(u, k)}) = \mathbb{E}_{U_k} \otimes \mathbb{E}_{\theta^*} [f_{\Delta(U_k)}(Y_{\Delta(U_k)})] \quad \mathbb{P}_{\theta^*}\text{-a.s. and in } L^2(\mathbb{P}_{\theta^*}), \quad (10)$$

with the convention  $T_{-1} = \emptyset$ , and where  $U_k$  is uniformly distributed over  $G_k$  and independent of the process  $X$ , and  $\mathbb{E}_{U_k} \otimes \mathbb{E}_{\theta^*}$  denotes the joint expectation over  $U_k$  and  $X$  (under the true parameter  $\theta^*$ ).

Moreover, there exist finite constants  $C < \infty$  and  $\alpha \in (0, 1)$  such that:

$$\forall n \geq k, \quad \mathbb{E} \left[ \left( \frac{1}{|T_n|} \sum_{u \in T_n \setminus T_{k-1}} f_{\Delta(u, k)}(Y_{\Delta(u, k)}) - \mathbb{E}_{U_k} \otimes \mathbb{E}_{\theta^*} [f_{\Delta(U_k)}(Y_{\Delta(U_k)})] \right)^2 \right] \leq C\alpha^n. \quad (11)$$

Remark that in the left hand side of (10) the subtrees  $\Delta(u, k)$  are deterministic, while the subtree  $\Delta(U_k)$  is a random function of  $U_k$ .

*Proof.* Using Lemma 2.3, remind that under Assumptions 1–3, the branching Markov process  $(X, Y)$  is stationary and its transition kernel has a unique invariant probability and is uniformly geometrically ergodic. Hence, the lemma follows immediately from applying the ergodic Theorems A.2 and A.4 for neighborhood-shape-dependent functions from the appendix.  $\square$

As  $T$  is a plane rooted tree, we can enumerate its vertices as a sequence  $(v_j)_{j \in \mathbb{N}}$  in a breadth-first-search manner, that is, which is increasing for  $<$  (note that  $u_0 = \partial$ ). Note that if  $V_n$  is uniformly distributed over  $A_n := \{v_j : |T_{k-1}| < j \leq n\} = \Delta(v_n) \setminus T_{k-1}$ , then the distribution of  $\mathcal{Sh}(\Delta(V_n, k))$  converges to the uniform distribution over  $\mathcal{N}_k$  as  $n \rightarrow \infty$ . We will also need the following variant of Lemma 2.11 where  $T_n \setminus T_{k-1}$  is replaced by  $A_n$ .

**Lemma 2.12** (Another ergodic theorem for neighborhood-dependent functions). *Assume that Assumptions 1–3 hold. Let  $k \geq 0$ . Let  $(f_S : \mathcal{Y}^S \rightarrow \mathbb{R})_{S \in \mathcal{N}_k}$  be a collection of neighborhood-shape-dependent Borel functions that are in  $L^2(\mathbb{P}_{\theta^*})$ . Let  $(v_j)_{j \in \mathbb{N}}$  be the sequence enumerating the vertices in  $T$  in a breadth-first-search manner. For all  $n > |T_{k-1}|$ , define  $A_n = \Delta(v_n) \setminus T_{k-1}$ . Then, we have:*

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{u \in A_n} f_{\Delta(u, k)}(Y_{\Delta(u, k)}) = \mathbb{E}_{U_k} \otimes \mathbb{E}_{\theta^*} [f_{\Delta(U_k)}(Y_{\Delta(U_k)})] \quad \text{in } L^2(\mathbb{P}_{\theta^*}),$$

where  $U_k$  is uniformly distributed over  $G_k$  and independent of the process  $X$ , and  $\mathbb{E}_{U_k} \otimes \mathbb{E}_{\theta^*}$  denotes the joint expectation over  $U_k$  and  $X$  (under the true parameter  $\theta^*$ ).

*Proof.* Using Lemma 2.3, remind that under Assumptions 1–3, the branching Markov process  $(X, Y)$  is stationary and its transition kernel has a unique invariant probability and is uniformly geometrically ergodic. Hence, the lemma follows immediately from applying the ergodic Theorem A.2 for neighborhood-shape-dependent functions from the appendix.  $\square$

### 3 Strong consistency of the MLE

In this section, we first define the extended tree  $T^\infty$  to get an infinite past horizon and rewrite the log-likelihood as a sum of increments. Then, we construct the log-likelihood increments with infinite past, which allows to define the contrast function. We prove properties for this contrast function. Finally, we prove the strong consistency of the MLE.

#### 3.1 Decomposition of the log-likelihood into increments

##### 3.1.1 The extended tree $T^\infty$ to get an infinite past horizon

Remind that the subtree  $\Delta^*(u) = \Delta^*(u, h(u))$  represents the past of the vertex  $u$ .

To get an infinite past horizon, we will consider an extended version of the tree  $T$ . Thus, we are going to define a random (countable) plane rooted tree  $T^\infty$  that contains  $T$  as a subtree and is also rooted at  $\partial$  the root vertex of  $T$ , and where each vertex (including  $\partial$ ) has exactly one parent node and two children nodes. To construct  $T^\infty$ , we start from  $T$  and add a line  $L = \{u_{-j} : j \in \mathbb{N}^*\}$  of ancestors for  $\partial$  (that is,  $u_{-j} = p^j(\partial)$  for  $j \in \mathbb{N}$ , where  $u_0 = \partial$ ), and then for all  $j \in \mathbb{N}^*$ , we graft on  $u_{-j}$  a copy  $T^{(j)}$  of  $T$  (that is,  $u_{-j}$  is the parent of the root vertex  $\partial^{(j)}$  of  $T^{(j)}$ ). We extend the height function  $h$  from  $T$  to  $T^\infty$  as follows: for all  $j \in \mathbb{N}^*$ , we set  $h(u_{-j}) = -j$  and for all  $u \in T^{(j)}$ , we define  $h(u)$  as  $-j$  plus the number of edges separating  $u$  from  $u_{-j}$ . For  $u, v \in T^\infty$ , denote by  $u \wedge v$  their most recent common ancestor, and by  $d(u, v) = h(u) + h(v) - 2h(u \wedge v)$  the graph distance between  $u$  and  $v$ . The definition of the subtrees  $T^\infty(u)$  and  $T^\infty(u, k)$  then naturally extend to  $T^\infty$ .

Thus, we have constructed the deterministic non-plane version of the tree  $T^\infty$ , and we are left to define the random plane embedding of  $T^\infty$ . That is, for each vertex  $u \in T^\infty$ , we have to define a possibly random ordering of its children. As  $T$  is a plane rooted tree, note that if  $u \in T$  or  $u \in T^{(j)}$  for some  $j \in \mathbb{N}^*$ , then its children are already order deterministically. Let  $\mathcal{U} = (\mathcal{U}_{(j)})_{-\infty < j \leq 0}$  be a sequence of independent random variables with Bernoulli distribution of parameter  $1/2$ , and which is independent of the HMT process  $(X, Y)$ . For all  $j \in \mathbb{N}$ , we order the children of  $u_{-j-1}$ , that is  $u_{-j}$  and  $\delta^{(j+1)}$  (the root vertex of  $T^{(j)}$ ), as follows:  $u_{-j}$  is the left child of  $u_{-j-1}$  if  $\mathcal{U}_{(-j)} = 0$ , and is the right child otherwise. Hence, we have constructed the random plane rooted tree  $T^\infty$ . (Note that  $\mathcal{U}$  can be seen as the random shape of the backward spine of  $\partial$ .) See Figure 4 for an illustration of the extended random plane rooted tree  $T^\infty$ . We denote by  $\mathbb{P}_{\mathcal{U}}$  the distribution of the random sequence  $\mathcal{U}$ , and by  $\mathbb{E}_{\mathcal{U}}$  the corresponding expectation.

Note that the random plane embedding of  $T^\infty$  allows to use Neveu's notation to represent the random path between any vertex in the plane tree  $T^\infty$  and one of its descendants as a random sequence  $\mathcal{U}_{(k:n)}$  (which depends on  $\mathcal{U}$ ) for some  $k, n \in \mathbb{Z}$  with  $k < n$ . The random breadth-first-search order relation  $\leq := \leq_{\mathcal{U}}$  can then be naturally extend from  $T$  to  $T^\infty$  using the random plane embedding of  $T^\infty$  (which depends on  $\mathcal{U}$ ): we have  $u \leq v$  for  $u, v \in T^\infty$  if either  $h(u) < h(v)$ , or  $h(u) = h(v)$  and  $U_{(k:n)} \leq_{\text{lex}} V_{(k:n)}$  where  $U_{(k:n)}$  (resp.  $V_{(k:n)}$ ) is Neveu's notation for the random path (which depends on  $\mathcal{U}$ ) from  $u \wedge v$  to  $u$  (resp.  $v$ ) with  $k = h(u \wedge v) + 1$  and  $n = h(u)$ .

Thanks to the stationarity assumption, for all  $k \in \mathbb{N}$ , the HMT process  $(X, Y)$  can be defined on the (rooted) tree  $T^\infty(p^k(\partial))$ , and thus by Kolmogorov's extension theorem, the HMT process  $(X, Y)$  can be defined on the whole tree  $T^\infty$ . In particular, note that the stationarity assumption implies that the distribution of the HMT process  $(X, Y)$  is invariant by translation on  $T^\infty$ , that is, is the same (up to translation) on  $T$  and on  $T^\infty(u)$  for any  $u \in T^\infty$ . Note that the extended process does not depend on  $\mathcal{U}$ . Thus, we will now assume that the HMT process  $(X, Y)$  is defined on the whole tree  $T^\infty$ .

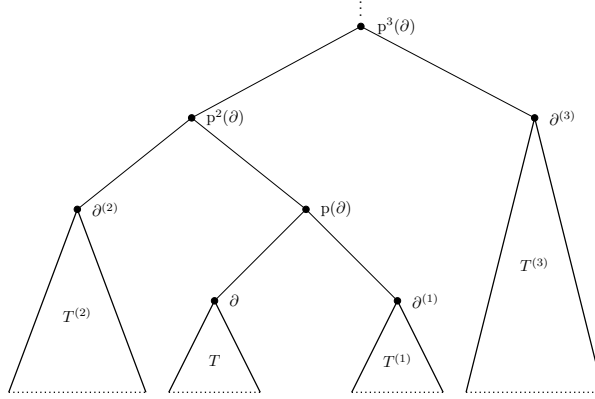


Figure 4: Illustration of the construction of the extended random plane rooted tree  $T^\infty$  (which is rooted at  $\partial$ ) for  $\mathcal{U}_{(-1)} = 0$  and  $\mathcal{U}_{(-2)} = 1$ .

For all  $u \in T^\infty$  and  $k \in \mathbb{N}$ , define the subtrees (which are measurable functions of  $\mathcal{U}$ ):

$$\Delta_{\mathcal{U}}^*(u, k) = \{v \in T^\infty(p^k(u)) : v <_{\mathcal{U}} u\},$$

and  $\Delta_{\mathcal{U}}(u, k) = \Delta_{\mathcal{U}}^*(u, k) \cup \{u\}$ . For simplicity, we will write instead  $\Delta^*(u, k)$  and  $\Delta(u, k)$ , making the dependence on the random variable  $\mathcal{U}$  implicit, and only make the dependence explicit when necessary. The following fact illustrates that the shape and size of the  $\Delta(u, k)$  do indeed depend on the value of  $\mathcal{U}$ : for  $u = \partial$  and  $k = 1$ , note that  $\Delta(\partial, 1)$  contains two vertices if  $\mathcal{U}_{(0)} = 0$ , and contains three vertices if  $\mathcal{U}_{(0)} = 1$ . Remark 3.1 below, which we shall reuse later, further illustrates the randomness of the set  $\Delta(u, k)$ . However, for  $u \in T$  and  $k \leq h(u)$ , we have that  $\Delta(u, k) = \Delta(u, k)$  and  $\Delta^*(u, k) = \Delta^*(u, k)$  are deterministic. Also note that we have the following inclusions:

$$T^\infty(u, k-1) \subset \Delta(u, k) \subset T^\infty(u, k), \quad (12)$$

where remind that the subtrees  $T^\infty(u, k-1)$  and  $T^\infty(u, k)$  are deterministic.

**Remark 3.1.** For a vertex  $u = u_{(1:n)}$  in  $T$  with  $h(u) = n \geq k$ , note that  $\Delta(u, k)$ , up to re-rooting (i.e. up to translation), can be identified with  $\Delta(\partial, k)$  conditioned on  $\mathcal{U}_{(-k+1:0)} = u_{(n-k+1:n)}$ . In particular, when  $U_n$  is a random vertex uniformly distributed over  $G_n$  for  $n \geq k$ , we get the following equality between the distribution of the shapes (that is, when the subtrees are seen up to translation / re-rooting) for the subtrees  $\Delta(\partial, k)$ ,  $\Delta(U_n, k)$  and  $\Delta(U_k)$ :

$$\mathcal{S}h(\Delta(\partial, k)) \stackrel{\mathcal{L}}{=} \mathcal{S}h(\Delta(U_n, k)) \stackrel{\mathcal{L}}{=} \Delta(U_k). \quad (13)$$

Moreover, if  $(f_S : \mathcal{Y}^S \rightarrow \mathbb{R})_{S \in \mathcal{N}_k}$  is a collection of neighborhood-shape-dependent Borel functions that are in  $L^2(\mathbb{P}_{\theta^*})$  (as in Lemmas 2.11 and 2.12), then we have:

$$\mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [f_{\Delta(\partial, k)}(Y_{\Delta(\partial, k)})] = \mathbb{E}_{U_k} \otimes \mathbb{E}_{\theta^*} [f_{\Delta(U_k)}(Y_{\Delta(U_k)})], \quad (14)$$

where  $\mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*}$  is the expectation corresponding to  $\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^*}$ .

### 3.1.2 The log-likelihood as a sum of increments

For any (possibly random) subtree  $\Delta$  of  $T^\infty$  with root vertex  $w$ , note that we have:

$$p_\theta(y_\Delta | X_w = x) = g_\theta(x, y_\Delta) \int_{\mathcal{X}^{|\Delta|-1}} \prod_{v \in \Delta \setminus \{w\}} q_\theta(x_{p(v)}, x_v) g_\theta(x_v, y_v) \lambda(dx_v). \quad (15)$$



We will use the convention  $p_\theta(Y_\Delta | X_w = x) = 1$  whenever  $\Delta = \emptyset$  and  $w$  is any vertex in  $T^\infty$ . For all  $u \in T$ ,  $k \in \mathbb{N}$ ,  $x \in \mathcal{X}$  and  $\theta \in \Theta$ , using the conditional probabilities formula, define:

$$\begin{aligned} \mathbb{H}_{u,k,x}(\theta) &= \int_{\mathcal{X}} g_\theta(x_u, Y_u) \mathbb{P}_\theta(X_u \in dx_u | Y_{\Delta^*(u,k)}, X_{\mathcal{P}^k(u)} = x) \\ &= \frac{p_\theta(Y_{\Delta(u,k)} | X_{\mathcal{P}^k(u)} = x)}{p_\theta(Y_{\Delta^*(u,k)} | X_{\mathcal{P}^k(u)} = x)}. \end{aligned} \quad (16)$$

We then define the log-likelihood contribution of node  $u \in T$  with past over  $k \in \mathbb{N}$  generation as:

$$h_{u,k,x}(\theta) = \log(\mathbb{H}_{u,k,x}(\theta)). \quad (17)$$

Note that  $h_{u,k,x}(\theta)$  (resp.  $\mathbb{H}_{u,k,x}(\theta)$ ) is a random variable as a function of  $Y_{\Delta(u,k)}$  with an implicit dependence on  $\mathcal{U}$  through  $\Delta(u,k)$ , and that  $h_{u,k,x}(\theta)$  (resp.  $\mathbb{H}_{u,k,x}(\theta)$ ) does not depend on  $\mathcal{U}$  is  $k \leq h(u)$ .

Hence, using (6), (7), (16) and (17) and a telescopic sum argument, the log-likelihood of the observed variables  $Y_{T_n}$  can be rewritten as the sum of the log-likelihood contributions defined in (17):

$$\ell_{n,x}(\theta) = \sum_{u \in T_n} h_{u,h(u),x}(\theta). \quad (18)$$

### 3.2 Construction of the log-likelihood increments with infinite past

In this subsection, we construct the log-likelihood increment functions with infinite past.

The following lemma states that, as the HMT is uniformly geometrically ergodic, the tree forgets exponentially fast its starting state. Recall the mixing ratio  $\rho = 1 - \sigma^-/\sigma^+ \in (0, 1)$  is defined just after Assumption 3.

**Lemma 3.2** (Exponential forgetting of the initial state). *Assume that Assumptions 2 and 3 hold. We have for all  $u \in T$ ,  $\theta \in \Theta$ ,  $n \in \mathbb{N}$  and  $y_{T_n} \in \mathcal{Y}^{T_n}$ , and all initials distributions  $\nu$  and  $\nu'$  on  $\mathcal{X}$ , that:*

$$\left\| \int_{\mathcal{X}} \mathbb{P}_\theta(X_u \in \cdot | Y_{T_n} = y_{T_n}, X_\partial = x) [\nu(dx) - \nu'(dx)] \right\|_{\text{TV}} \leq \rho^{h(u)}. \quad (19)$$

For simplicity, Lemma 3.2 is stated with  $\partial$  as the initial vertex, but note that the results still holds when replacing  $\partial$  and  $T_n$  by  $v$  and  $T(v, n)$  for any  $v \in T^\infty$ . We shall reuse this fact later.

*Proof.* Fix some  $u \in T$ ,  $\theta \in \Theta$ , an integer  $n$  and observables  $y_{T_n} \in \mathcal{Y}^{T_n}$ . Denote by  $u_0, \dots, u_k$  with  $k = h(u)$  the vertices on the path from  $\partial$  to  $u$ . The proof relies on the fact that conditionally on  $Y_{T_n} = y_{T_n}$ , the sequence  $(X_{u_j})_{0 \leq j \leq k}$  is an inhomogeneous Markov chain where for  $1 \leq j \leq k$ , the (forward smoothing) transition kernel  $F_j$  from  $X_{u_{j-1}}$  to  $X_{u_j}$  is defined if  $j \leq n$  as:

$$\begin{aligned} F_j[y_{T(u_j, n-j)}](x_{u_{j-1}}; f) &= \mathbb{E}_\theta[f(X_{u_j}) | Y_{T(u_j, n-j)} = y_{T(u_j, n-j)}, X_{u_{j-1}} = x_{u_{j-1}}] \\ &= \mathbb{E}_\theta[f(X_{u_j}) | Y_{T_n} = y_{T_n}, X_{u_{j-1}} = x_{u_{j-1}}] \\ &= \frac{\int_{\mathcal{X}} f(x_{u_j}) p_\theta(y_{T(u_j, n-j)} | X_{u_j} = x_{u_j}) q_\theta(x_{u_{j-1}}, x_{u_j}) \lambda(dx_{u_j})}{\int_{\mathcal{X}} p_\theta(y_{T(u_j, n-j)} | X_{u_j} = x_{u_j}) q_\theta(x_{u_{j-1}}, x_{u_j}) \lambda(dx_{u_j})}, \end{aligned}$$

for any  $x_{u_{j-1}} \in \mathcal{X}$  and any bounded Borel function  $f$  on  $\mathcal{X}$  (note that in the second equality, we used the Markov property of the HMT process, see (2)); and is defined as  $F_j = Q$  for  $j > n$ .

(Note that Assumption 3-(ii) is only used to insure that  $p_\theta(y_{T(u_j, n-j)} | X_{u_j} = x_{u_j})$  is positive, and thus the denominator in the last equality is also positive.)

Note that for all  $1 \leq j \leq k \wedge n$ , using Assumption 3-(i), the transition kernel  $F_j$  satisfies the following Doeblin condition:

$$\frac{\sigma^-}{\sigma^+} \nu_j[y_{T(u_j, n-j)}](f) \leq F_j[y_{T(u_j, n-j)}](x; f),$$

where for any bounded Borel function  $f$  on  $\mathcal{X}$ , we have:

$$\begin{aligned} \nu_j[y_{T(u_j, n-j)}](f) &= \mathbb{E}_\theta[f(X_{u_j}) | Y_{T(u_j, n-j)} = y_{T(u_j, n-j)}] \\ &= \frac{\int_{\mathcal{X}} f(x_{u_j}) p_\theta(y_{T(u_j, n-j)} | X_{u_j} = x_{u_j}) \lambda(dx_{u_j})}{\int_{\mathcal{X}} p_\theta(y_{T(u_j, n-j)} | X_{u_j} = x_{u_j}) \lambda(dx_{u_j})}. \end{aligned}$$

Note that the difference between the definitions of  $F_j$  and  $\nu_j$  is that the term  $q_\theta(x_{p(u_j)}, x_{u_j})$  has disappear from both the numerator and the denominator of  $\nu_j$ . Remark that (3) also implies the Doeblin condition  $\sigma^- \lambda(\cdot) \leq Q(x, \cdot)$  for the transition kernel  $Q$ . Thus, Lemma 2.7 shows that the Dobrushin coefficient of each transition kernel  $F_j$  for  $1 \leq j \leq k$  is upper bounded by  $\rho = 1 - \sigma^-/\sigma^+$ . Therefore, as the Dobrushin coefficient is sub-multiplicative (see Lemma 2.5), applying Lemma 2.4, we get that (19) holds. This concludes the proof.  $\square$

To construct the limit of the functions  $h_{u, k, x}(\theta)$  we first prove the following lemma which states some uniform bound about the asymptotic behavior of those functions when  $k \rightarrow \infty$ . For this lemma, we need the following assumption on the density function  $g_\theta$  that strengthens Assumption 3-(ii). Remind that  $\mathbb{P}_\theta$  denotes the stationary probability distribution under the parameter  $\theta \in \Theta$  of the HMT process  $(X, Y)$ , and by  $\mathbb{E}_\theta$  the corresponding expectation.

**Assumption 4** ( $L^1$  regularity, [CMR05, Assumption 12.3.1]). *Assume that we have:*

- (i)  $b^+ := 1 \wedge \sup_\theta \sup_{x, y} g_\theta(x, y) < \infty$ .
- (ii)  $\mathbb{E}_{\theta^*} |\log b^-(Y_\partial)| < \infty$ , where  $b^-(y) := \inf_\theta \int_{\mathcal{X}} g_\theta(x, y) \lambda(dx)$ .

Note that  $b^-(y) > 0$  for all  $y \in \mathcal{Y}$  by Assumption 2-(ii).

**Lemma 3.3** (Uniform bounds for  $h_{u, k, x}(\theta)$ ). *Assume that Assumptions 2-3 and 4-(ii) hold. For all vertices  $u \in T$  and all integers  $k, k' \in \mathbb{N}^*$ , the following assertions hold true:*

$$\sup_{\theta \in \Theta} \sup_{x, x' \in \mathcal{X}} |h_{u, k, x}(\theta) - h_{u, k', x'}(\theta)| \leq \frac{\rho^{(k \wedge k') - 1}}{1 - \rho}, \quad (20)$$

$$\sup_{\theta \in \Theta} \sup_{k \in \mathbb{N}^*} \sup_{x \in \mathcal{X}} |h_{u, k, x}(\theta)| \leq \log b^+ \vee |\log(\sigma^- b^-(Y_u))|. \quad (21)$$

*Proof.* [The proof is a straightforward adaptation of the proof of [CMR05, Lemma 12.3.2] using Lemma 3.2 for the coupling.] Remind the definition of  $H_{u, k, x}(\theta)$  in (16). Let  $k' \geq k \geq 1$ , and write  $w = p^k(u)$ ,  $w' = p^{k'}(u)$ . Then, write:

$$\begin{aligned} H_{u, k, x}(\theta) &= \int_{\mathcal{X} \times \mathcal{X}} \left[ \int_{\mathcal{X}} g_\theta(x_u, Y_u) q_\theta(x_{p(u)}, x_u) \lambda(dx_u) \right] \\ &\quad \times \mathbb{P}_\theta(X_{p(u)} \in dx_{p(u)} | Y_{\Delta^*(u, k)}, X_w = x_w) \times \delta_x(dx_w), \end{aligned} \quad (22)$$

and using the Markov property at  $X_w$ , write:

$$\begin{aligned} H_{u,k',x'}(\theta) &= \int_{\mathcal{X} \times \mathcal{X}} \left[ \int_{\mathcal{X}} g_\theta(x_u, Y_u) q_\theta(x_{p(u)}, x_u) \lambda(dx_u) \right] \\ &\quad \times \mathbb{P}_\theta(X_{p(u)} \in dx_{p(u)} | Y_{\Delta^*(u,k)}, X_w = x_w) \\ &\quad \times \mathbb{P}_\theta(X_w \in dx_w | Y_{\Delta^*(u,k')}, X_{w'} = x'). \end{aligned} \quad (23)$$

Applying Lemma 3.2, we get (note that the integrands in (22) and (23) are non-negative):

$$\begin{aligned} |H_{u,k,x}(\theta) - H_{u,k',x'}(\theta)| &\leq \rho^{k-1} \sup_{x_{p(u)} \in \mathcal{X}} \int_{\mathcal{X}} g_\theta(x_u, Y_u) q_\theta(x_{p(u)}, x_u) \lambda(dx_u) \\ &\leq \rho^{k-1} \sigma^+ \int_{\mathcal{X}} g_\theta(x_u, Y_u) \lambda(dx_u). \end{aligned} \quad (24)$$

The integral in (22) can be lower bounded giving us:

$$H_{u,k,x}(\theta) \geq \sigma^- \int_{\mathcal{X}} g_\theta(x_u, Y_u) \lambda(dx_u), \quad (25)$$

where the right hand side is positive by Assumption 3-(ii); and similarly for (23). Combining (24) with (25), and with the inequality  $|\log x - \log y| \leq |x - y|/(x \wedge y)$ , we get the first assertion of the lemma:

$$|h_{u,k,x}(\theta) - h_{u,k',x'}(\theta)| \leq \frac{\sigma^+}{\sigma^-} \rho^{k-1} = \frac{\rho^{k-1}}{1 - \rho}.$$

Combining (16) and (25), we get that  $\sigma^- b^-(Y_u) \leq H_{u,k,x}(\theta) \leq b^+$  (remind that  $b^-(Y_u) > 0$  by Assumption 3-(ii)), which yields the second assertion of the lemma.  $\square$

We are now ready to construct the limit of the functions  $h_{u,k,x}(\theta)$  and state some properties of this limit. Note that this result is stated for every  $u \in T$ , but we will only need it for  $u = \partial$ . Remind that we are in the stationary case, and that the HMT process  $(X, Y)$  is defined on  $T^\infty$ .

**Proposition 3.4** (Properties of the limit function  $h_{u,\infty}(\theta)$ ). *Assume that Assumptions 1–4 hold. For every  $u \in T$  and  $\theta \in \Theta$ , there exists  $h_{u,\infty}(\theta) \in L^1(\mathbb{P}_U \otimes \mathbb{P}_{\theta^*})$  such that for all  $x \in \mathcal{X}$ , the sequence  $(h_{u,k,x}(\theta))_{k \in \mathbb{N}}$  converges  $\mathbb{P}_U \otimes \mathbb{P}_{\theta^*}$ -a.s. and in  $L^1(\mathbb{P}_U \otimes \mathbb{P}_{\theta^*})$  to  $h_{u,\infty}(\theta)$ .*

*Furthermore, this convergence is uniform over  $\theta \in \Theta$  and  $x \in \mathcal{X}$ , that is, we have that  $\lim_{k \rightarrow \infty} \sup_{\theta \in \Theta} \sup_{x \in \mathcal{X}} |h_{u,k,x}(\theta) - h_{u,\infty}(\theta)| = 0$   $\mathbb{P}_U \otimes \mathbb{P}_{\theta^*}$ -a.s. and in  $L^1(\mathbb{P}_U \otimes \mathbb{P}_{\theta^*})$ .*

The limit function  $h_{u,\infty}(\theta)$  can be interpreted as  $\log p_\theta(Y_u | Y_{\Delta^*(u,\infty)})$ , where  $\Delta^*(u, \infty) = \{v \in T^\infty : v <_{\mathcal{U}} u\}$  is a random subset of vertices. Note that  $h_{u,\infty}(\theta)$  is a function of the random set of variables  $(Y_v, v \in \Delta(u, \infty))$ , where  $\Delta(u, \infty) = \Delta^*(u, \infty) \cup \{u\}$ , and thus implicitly depend on  $\mathcal{U}$  through  $\Delta(u, \infty)$ .

*Proof.* Fix some  $u \in T$ . Note that (20) shows that the sequence  $(h_{u,k,x}(\theta))_{k \in \mathbb{N}}$  is Cauchy uniformly in  $\theta$  and  $x$ , and thus has  $\mathbb{P}_U \otimes \mathbb{P}_{\theta^*}$ -almost surely a limit when  $k \rightarrow \infty$  which does not depend on  $x$ ; we denote this limit by  $h_{u,\infty}(\theta)$ . Furthermore, we get from (21) that  $(h_{u,k,x}(\theta))_{k \in \mathbb{N}}$  is uniformly bounded in  $L^1(\mathbb{P}_U \otimes \mathbb{P}_{\theta^*})$ , and thus  $h_{u,\infty}(\theta)$  is in  $L^1(\mathbb{P}_U \otimes \mathbb{P}_{\theta^*})$  and the convergence also holds in  $L^1(\mathbb{P}_U \otimes \mathbb{P}_{\theta^*})$ . Finally, as the bound in (20) is uniform in  $\theta$  and  $x$ , we get that the convergence holds uniform over  $\theta$  and  $x$  both  $\mathbb{P}_U \otimes \mathbb{P}_{\theta^*}$ -almost surely and in  $L^1(\mathbb{P}_U \otimes \mathbb{P}_{\theta^*})$ .  $\square$

### 3.3 Properties of the contrast function

As the functions  $h_{\partial,\infty}(\theta)$  are in  $L^1(\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^*})$  under the assumptions used in Proposition 3.4, we can now define the *contrast function*  $\ell$  (which is deterministic) as:

$$\ell(\theta) = \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [h_{\partial,\infty}(\theta)], \quad (26)$$

where remind  $\mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*}$  is the expectation corresponding to  $\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^*}$ .

We prove under the following  $L^2$  regularity assumption the convergence of the normalized log-likelihood to the contrast function. Remind that  $b^-(y) = \inf_{\theta} \int_{\mathcal{X}} g_{\theta}(x, y) \lambda(dx)$ . Also remind that  $\mathbb{P}_{\theta}$  denotes the stationary probability distribution under the parameter  $\theta \in \Theta$  of the HMT process  $(X, Y)$ , and by  $\mathbb{E}_{\theta}$  the corresponding expectation.

**Assumption 5** ( $L^2$  regularity). *Assume that  $\mathbb{E}_{\theta^*} [(\log b^-(Y_{\partial}))^2] < \infty$ .*

Remind that the log-likelihood  $\ell_{n,x}$  is defined in (7) on page 11.

**Proposition 3.5** (Ergodic convergence for the stationary log-likelihood). *Assume that Assumptions 1–5 hold. Then, for all  $x \in \mathcal{X}$ , the normalized log-likelihood  $|T_n|^{-1} \ell_{n,x}(\theta)$  converges  $\mathbb{P}_{\theta^*}$ -a.s. to the contrast function  $\ell(\theta)$  as  $n \rightarrow \infty$ .*

*Proof.* Let  $\theta \in \Theta$  be some parameter. Fix some  $k \in \mathbb{N}^*$  and  $x \in \mathcal{X}$ . Remind that  $\ell_{n,x}(\theta) = \sum_{u \in T_n} h_{u,h(u),x}(\theta)$ . Applying (20) for each vertex  $u \in T_n \setminus T_{k-1}$ , we get:

$$\frac{1}{|T_n|} \left| \ell_{n,x}(\theta) - \sum_{u \in T_n \setminus T_{k-1}} h_{u,k,x}(\theta) \right| \leq \frac{\rho^{k-1}}{1-\rho} + \frac{1}{|T_n|} \sum_{u \in T_{k-1}} |h_{u,h(u),x}(\theta)|. \quad (27)$$

Note that by (21), we have that  $|h_{u,h(u),x}(\theta)| < \infty$   $\mathbb{P}_{\theta^*}$ -a.s. for all  $u \in T \setminus \{\partial\}$ . For  $u = \partial$ , we have  $h_{\partial,0,x}(\theta) = \log g_{\theta}(x, Y_{\partial})$  which is finite  $\mathbb{P}_{\theta^*}$ -a.s. by Assumption 3-(iii).

For a vertex  $u$  in  $T \setminus T_{k-1}$ , let  $v_u \in G_k$  be the unique vertex that satisfies (8) (on page 12), then we have:

$$h_{u,k,x}(\theta; Y_{\Delta(u,k)} = y_{\Delta(u,k)}) = h_{v_u,k,x}(\theta; Y_{\Delta(v_u)} = y_{\Delta(u,k)}). \quad (28)$$

Moreover, using (21) together with Assumption 5, we get for every  $u \in T \setminus T_{k-1}$  that the random variable  $h_{u,k,x}(\theta; Y_{\Delta(u,k)})$  is in  $L^2(\mathbb{P}_{\theta^*})$ . Hence, applying Lemma 2.11 to the collection of neighborhood-shape-dependent functions  $(h_{v,k,x}(\theta; Y_{\Delta(v)} = \cdot))_{v \in G_k}$  (remind that indexing functions with  $G_k$  or with  $\mathcal{N}_k$  is equivalent by (9)), and using (28) and (14) (in Remark 3.1), we get:

$$|T_n|^{-1} \sum_{u \in T_n \setminus T_{k-1}} h_{u,k,x}(\theta) \xrightarrow{n \rightarrow \infty} \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [h_{\partial,k,x}(\theta)] \quad \mathbb{P}_{\theta^*}\text{-a.s. (and in } L^2(\mathbb{P}_{\theta^*})\text{)}. \quad (29)$$

Using (20) with Proposition 3.4, we get:

$$|\mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [h_{\partial,k,x}(\theta)] - \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [h_{\partial,\infty}(\theta)]| \leq \frac{\rho^{k-1}}{1-\rho}.$$

Thus, combining this bound with (27) and (29), we get  $\mathbb{P}_{\theta^*}$ -a.s. that:

$$\limsup_{n \rightarrow \infty} \left| |T_n|^{-1} \ell_{n,x}(\theta) - \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [h_{\partial,\infty}(\theta)] \right| \leq 2 \frac{\rho^{k-1}}{1-\rho}.$$

As the left hand side does not depend on  $k$ , letting  $k \rightarrow \infty$ , we get that  $|T_n|^{-1} \ell_{n,x}(\theta)$  converges  $\mathbb{P}_{\theta^*}$ -a.s. to  $\ell(\theta)$  as  $n \rightarrow \infty$ . This concludes the proof.  $\square$

We are going to prove that this convergence holds uniformly in  $\theta$ . First, we need to prove that the contrast function is continuous and has a unique global maximum at  $\theta^*$ . In order to get those results, we need a natural continuity assumption on the transition functions.

**Assumption 6** (Continuity, [CMR05, Assumption 12.3.5]). *For all  $(x, x') \in \mathcal{X} \times \mathcal{X}$  and  $y \in \mathcal{Y}$ , the functions  $\theta \mapsto q_\theta(x', x)$  and  $\theta \mapsto g_\theta(x, y)$  defined on  $\Theta \subset \mathbb{R}^d$  are continuous.*

We denote by  $\|\cdot\|$  the euclidean norm on  $\mathbb{R}^d$ .

**Proposition 3.6** ( $\ell$  is continuous). *Assume that Assumptions 1–4 and 6 hold. Then, for any  $n \in \mathbb{N}$  and  $x \in \mathcal{X}$ , the log-likelihood function  $\theta \mapsto \ell_{n,x}(\theta)$  is  $\mathbb{P}_{\theta^*}$ -a.s. continuous on  $\Theta$ .*

Moreover, for any  $\theta \in \Theta$ , we have:

$$\mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} \left[ \sup_{\theta' \in \Theta: \|\theta - \theta'\| \leq \delta} |h_{\partial, \infty}(\theta') - h_{\partial, \infty}(\theta)| \right] \rightarrow 0 \quad \text{as } \delta \rightarrow 0,$$

and the contrast function  $\theta \mapsto \ell(\theta)$  is continuous on  $\Theta$ .

*Proof.* This proof is a straightforward adaptation from the proof of [CMR05, Proposition 12.3.6].

Recall that  $h_{\partial, \infty}(\theta)$  is the limit of  $h_{\partial, k, x}(\theta)$  as  $k \rightarrow \infty$ . We first prove that, for every  $x \in \mathcal{X}$  and  $k \geq 0$ ,  $h_{\partial, k, x}(\theta)$  is a continuous function of  $\theta$ , and then use this to show continuity of the limit. Recall from (16) the second equality defining  $H_{u, k, x}(\theta)$ , which we remind for convenience for any  $u \in T$  and  $x \in \mathcal{X}$ :

$$H_{u, k, x}(\theta) = \frac{p_\theta(Y_{\Delta(u, k)} | X_{\mathcal{P}^k(u)} = x)}{p_\theta(Y_{\Delta^*(u, k)} | X_{\mathcal{P}^k(u)} = x)}.$$

Recall from (15) the definition of  $p_\theta(Y_{\Delta} | X_{\mathcal{P}^k(u)} = x)$  where here the possibly random subtree  $\Delta$  is either  $\Delta(u, k)$  or  $\Delta^*(u, k)$ . First note that the integrand in (15) is by assumption continuous w.r.t.  $\theta$  and upper bounded by  $(1 \vee \sigma^+ b^+)^{|\Delta|}$ . Thus, dominated convergence shows that  $p_\theta(Y_{\Delta} | X_{\mathcal{P}^k(u)} = x)$  is continuous w.r.t. to  $\theta$  (remind that  $\lambda$ , defined in Assumption 2, is finite). Moreover, note that  $p_\theta(Y_{\Delta^*(u, k)} | X_{\mathcal{P}^k(u)} = x)$  is lower bounded by  $\prod_{v \in \Delta^*(u, k) \setminus \{\mathcal{P}^k(u)\}} \sigma^- b^- (Y_v)$  which is positive  $\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^*}$ -a.s. (by Assumption 3). Thus,  $H_{u, k, x}(\theta)$  and  $h_{u, k, x}(\theta) = \log H_{u, k, x}(\theta)$  (remind (17)) are continuous w.r.t.  $\theta$   $\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^*}$ -a.s. as well. Hence, using (6), for all  $n \in \mathbb{N}$  and  $x \in \mathcal{X}$ , we get that  $\ell_{n,x}(\theta)$  is also continuous w.r.t.  $\theta$   $\mathbb{P}_{\theta^*}$ -a.s.

Remind from Proposition 3.4 that  $(h_{u, k, x}(\theta))_{k \in \mathbb{N}}$  converges to  $h_{u, \infty}(\theta)$  uniformly in  $\theta$   $\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^*}$ -a.s. Thus, the function  $\theta \mapsto h_{u, \infty}(\theta)$  is continuous  $\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^*}$ -a.s. Using the uniform bound (21), Assumption 4-(ii) and dominated convergence, we obtain the first part of the proposition.

We deduce the second part from the first one, as:

$$\begin{aligned} \sup_{\theta' \in \Theta: \|\theta' - \theta\| \leq \delta} |\ell(\theta') - \ell(\theta)| &= \sup_{\theta' \in \Theta: \|\theta' - \theta\| \leq \delta} |\mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [h_{\partial, \infty}(\theta') - h_{\partial, \infty}(\theta)]| \\ &\leq \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} \left[ \sup_{\theta' \in \Theta: \|\theta' - \theta\| \leq \delta} |h_{\partial, \infty}(\theta') - h_{\partial, \infty}(\theta)| \right]. \end{aligned}$$

This concludes the proof.  $\square$

We are now ready to state and prove that the convergence to the contrast function  $\ell$  holds uniformly in  $\theta$ .

**Proposition 3.7** (Uniform convergence to  $\ell$ ). *Assume that Assumptions 1–6 hold and  $\Theta$  is compact. Then, we have:*

$$\lim_{n \rightarrow \infty} \sup_{\theta \in \Theta} \left| |T_n|^{-1} \ell_{n,x}(\theta) - \ell(\theta) \right| = 0 \quad \mathbb{P}_{\theta^*}\text{-a.s.}$$

*Proof.* [We mimic the proof of [CMR05, Proposition 12.3.7].] As  $\Theta$  is compact, it is sufficient to prove that for every  $\theta \in \Theta$ :

$$\limsup_{\delta \rightarrow 0} \limsup_{n \rightarrow \infty} \sup_{\theta' \in \Theta: \|\theta' - \theta\| \leq \delta} \left| |T_n|^{-1} \ell_{n,x}(\theta') - \ell(\theta) \right| = 0 \quad \mathbb{P}_{\theta^*}\text{-a.s.} \quad (30)$$

As this claim is not proven in the proof of [CMR05, Proposition 12.3.7], we give a short proof. Indeed, assume that (30) holds for all  $\theta \in \Theta$ . Let  $\varepsilon > 0$ . By Proposition 3.6, the function  $\ell$  is continuous, and thus uniformly continuous as  $\Theta$  is compact. In particular, there exists  $\delta > 0$  such that for all  $\theta, \theta' \in \Theta$ , we have that  $\|\theta - \theta'\| \leq \delta$  implies  $|\ell(\theta) - \ell(\theta')| \leq \varepsilon$ . For every  $\theta \in \Theta$ , let  $\delta_\theta \in (0, \delta)$  be such that  $\limsup_{n \rightarrow \infty} \sup_{\theta' \in \Theta: \|\theta' - \theta\| \leq \delta_\theta} \left| |T_n|^{-1} \ell_{n,x}(\theta') - \ell(\theta) \right| < \varepsilon$ . As  $\cup_{\theta \in \Theta} \{\theta' : \|\theta' - \theta\| \leq \delta_\theta\}$  is an open cover of  $\Theta$  and as  $\Theta$  is compact, there exists a finite subset  $\{\theta_j : 1 \leq j \leq m\}$  of  $\Theta$  with  $m \geq 1$  such that  $\Theta = \cup_{j=1}^m \{\theta' : \|\theta' - \theta_j\| \leq \delta_{\theta_j}\}$ . Note that for  $n$  large enough, for all  $1 \leq j \leq m$ , we have that  $\sup_{\theta' \in \Theta: \|\theta' - \theta_j\| \leq \delta_{\theta_j}} \left| |T_n|^{-1} \ell_{n,x}(\theta') - \ell(\theta_j) \right| < \varepsilon$ . Thus, for  $n$  large enough, we have:

$$\sup_{\theta \in \Theta} \left| |T_n|^{-1} \ell_{n,x}(\theta) - \ell(\theta) \right| \leq \varepsilon + \max_{1 \leq j \leq m} \sup_{\theta' \in \Theta: \|\theta' - \theta_j\| \leq \delta_{\theta_j}} \left| |T_n|^{-1} \ell_{n,x}(\theta') - \ell(\theta_j) \right| \leq 2\varepsilon.$$

This being true for all  $\varepsilon > 0$ , we get that the statement in the proposition holds.

We now prove (30). Fix some  $\theta \in \Theta$ . Remind that by Proposition 3.5, we have that  $\lim_{n \rightarrow \infty} |T_n|^{-1} \ell_n(\theta) = \ell(\theta)$   $\mathbb{P}_{\theta^*}$ -a.s. Using this fact, we get:

$$\begin{aligned} & \limsup_{n \rightarrow \infty} \sup_{\theta' \in \Theta: \|\theta' - \theta\| \leq \delta} \left| |T_n|^{-1} \ell_{n,x}(\theta') - \ell(\theta) \right| \\ &= \limsup_{n \rightarrow \infty} \sup_{\theta' \in \Theta: \|\theta' - \theta\| \leq \delta} \left| |T_n|^{-1} \ell_{n,x}(\theta') - |T_n|^{-1} \ell_{n,x}(\theta) \right|. \end{aligned} \quad (31)$$

Using (27), for any  $k \geq 1$ , we get that (31) is  $\mathbb{P}_{\theta^*}$ -a.s. bounded by:

$$\begin{aligned} & 2 \limsup_{n \rightarrow \infty} \sup_{\theta' \in \Theta: \|\theta' - \theta\| \leq \delta} \left| |T_n|^{-1} \ell_{n,x}(\theta') - \sum_{u \in T_n \setminus T_{k-1}} h_{u,k,x}(\theta') \right| \\ &+ \limsup_{n \rightarrow \infty} |T_n|^{-1} \sum_{u \in T_n \setminus T_{k-1}} \sup_{\theta' \in \Theta: \|\theta' - \theta\| \leq \delta} \left| h_{u,k,x}(\theta') - h_{u,k,x}(\theta) \right| \\ &\leq 2 \frac{\rho^{k-1}}{1 - \rho} + \limsup_{n \rightarrow \infty} |T_n|^{-1} \sum_{u \in T_n \setminus T_{k-1}} \sup_{\theta' \in \Theta: \|\theta' - \theta\| \leq \delta} \left| h_{u,k,x}(\theta') - h_{u,k,x}(\theta) \right| \\ &= 2 \frac{\rho^{k-1}}{1 - \rho} + \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} \left[ \sup_{\theta' \in \Theta: \|\theta' - \theta\| \leq \delta} \left| h_{\partial,k,x}(\theta') - h_{\partial,k,x}(\theta) \right| \right] \\ &\leq 4 \frac{\rho^{k-1}}{1 - \rho} + \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} \left[ \sup_{\theta' \in \Theta: \|\theta' - \theta\| \leq \delta} \left| h_{\partial,\infty}(\theta') - h_{\partial,\infty}(\theta) \right| \right], \end{aligned} \quad (32)$$

where we used Lemma 2.11 for ergodic convergence (with  $L^2(\mathbb{P}_{\theta^*})$  boundedness given by (21)) in the equality, and we used (20) (with Proposition 3.4) in the second inequality. Then, letting  $k \rightarrow \infty$  in the upper bound of (32) (note that (31) does not depend on  $k$ ), and then letting  $\delta \rightarrow 0$  with Proposition 3.6, we get that (30) holds. This concludes the proof.  $\square$

**Remark 3.8** (Uniform convergence for the log-likelihood with general initial condition). Let  $\nu$  be a probability distribution on  $\mathcal{X}$  such that  $\sup_{\theta} |\int g_{\theta}(x, Y_{\partial}) \nu(dx)|$  is finite  $\mathbb{P}_{\theta^*}$ -a.s. The uniform convergence of  $|T_n|^{-1} \ell_{n,x}(\theta)$  to  $\ell(\theta)$  still holds when modifying the definition of the log-likelihood  $\ell_{n,x}(\theta)$  of the HMT to replace the Dirac mass  $\delta_x$  by  $\nu$  for the distribution of the root hidden variable  $X_{\partial}$ . When  $\nu$  is the stationary distribution  $\pi_{\theta}$  associated to  $q_{\theta}$ , uniform convergence holds without this extra regularity assumption by conditioning on the state of the root's parent  $X_{p(\partial)}$  instead (which allows to replace  $h_{\partial,0,x}(\theta) = g_{\theta}(x, Y_{\partial})$  in (27) by  $h_{\partial,1,\nu}(\theta) := \log \int_{\mathcal{X}} H_{\partial,1,x}(\theta) \nu(dx)$  for which  $\sup_{\theta} |h_{\partial,1,\nu}(\theta)|$  is finite by an immediate adaptation of (21)).

### 3.4 Identifiability and strong consistency

In this subsection, we prove the strong consistency of the MLE. We must first study the identifiability of the parameter of the HMT model. We start with a definition of equivalent parameters.

**Definition 3.9** (Equivalent parameters). *We say that two parameters  $\theta, \theta' \in \Theta$  are equivalent if they define the same distribution for the process  $(Y_u, u \in T)$ , i.e.  $\mathbb{P}_{\theta}(Y \in \cdot) = \mathbb{P}_{\theta'}(Y \in \cdot)$ .*

Note that by Kolmogorov's extension theorem,  $\theta$  and  $\theta'$  are equivalent if and only if they define the same law on every finite tree  $T_n$ , i.e. for  $(Y_u, u \in T_n)$ .

The following proposition characterizes global maxima of the contrast function  $\ell$ .

**Proposition 3.10** (Global maxima of the contrast function  $\ell$ ). *Assume that Assumptions 1–5 hold. Then a parameter  $\theta \in \Theta$  is a global maximum of  $\ell$  if and only if  $\theta$  is equivalent to  $\theta^*$ .*

We get as an immediate corollary that  $\theta^*$  is a global maximum of  $\ell$ .

The proof of Proposition 3.10, which is postponed to the end of this section, is an adaptation of the proof of [CMR05, Theorem 12.4.2]. This adaptation comes from the difference of topology between the tree and the line.

Remind that the log-likelihood function  $\theta \mapsto \ell_{n,x}(\theta)$  is continuous  $\mathbb{P}_{\theta^*}$ -a.s. under Assumptions 1–4 and 6. Thus, when we further assume that  $\Theta$  is compact, we get that the argmax set  $\operatorname{argmax}_{\theta \in \Theta} \ell_{n,x}(\theta)$  is non-empty. The maximum likelihood estimator (MLE) is then defined as the maximizer over  $\Theta$  of the log-likelihood  $\ell_{n,x}$ , that is as the following random variable (which depends on  $Y_{T_n}$ ):

$$\hat{\theta}_{n,x} = \hat{\theta}_{n,x}(Y_{T_n}) \in \operatorname{argmax}_{\theta \in \Theta} \ell_{n,x}(\theta). \quad (33)$$

Note that the argmax set in (33) is not necessarily unique, in which case we select one parameter  $\theta$  from the argmax set in a measurable manner (which is possible, see [BS96, Proposition 7.33]).

We are now ready to prove the following theorem that states the strong consistency of the MLE for the HMT model in the stationary case.

**Theorem 3.11** (Strong consistency of the MLE). *Assume that Assumptions 1–6 hold, the contrast function  $\ell$  has a unique maximum (which is then located at  $\theta^* \in \Theta$  by Proposition 3.10) and  $\Theta$  is compact. Then, for any  $x \in \mathcal{X}$ , the MLE  $\hat{\theta}_{n,x}$  (defined in (33)) converges  $\mathbb{P}_{\theta^*}$ -a.s. as  $n \rightarrow \infty$  to the true parameter  $\theta^* \in \Theta$ , i.e. the MLE is strongly consistent.*

*Proof.* [The proof is a straightforward adaptation of an argument for HMMs in [CMR05, Section 12.1], which itself adapts an argument that goes back to [Wal49].]

By definition of  $\hat{\theta}_n$ , we have that  $\ell_{n,x}(\hat{\theta}_n) \geq \ell_{n,x}(\theta)$  for every  $\theta \in \Theta$ . As the contrast function  $\ell$  has a unique maximum located at  $\theta^*$ , we have that  $\ell(\theta^*) \geq \ell(\theta)$  for every  $\theta \in \Theta$ , and in particular,  $\ell(\theta^*) \geq \ell(\hat{\theta}_n)$  for every  $n \in \mathbb{N}$ . Combining those two bounds, we get that:

$$\begin{aligned} 0 &\leq \ell(\theta^*) - \ell(\hat{\theta}_n) \\ &\leq \ell(\theta^*) - |T_n|^{-1} \ell_{n,x}(\theta^*) + |T_n|^{-1} \ell_{n,x}(\theta^*) - |T_n|^{-1} \ell_{n,x}(\hat{\theta}_n) + |T_n|^{-1} \ell_{n,x}(\hat{\theta}_n) - \ell(\hat{\theta}_n) \\ &\leq 2 \sup_{\theta \in \Theta} \left| \ell(\theta) - |T_n|^{-1} \ell_{n,x}(\theta) \right|, \end{aligned}$$

where the upper bound in the last line goes to zero  $\mathbb{P}_{\theta^*}$ -a.s. as  $n \rightarrow \infty$  by Proposition 3.7 as  $\Theta$  is compact. Hence, we get that  $\ell(\hat{\theta}_n) \rightarrow \ell(\theta^*)$   $\mathbb{P}_{\theta^*}$ -a.s. as  $n \rightarrow \infty$ . Consequently, as  $\ell$  is continuous (by Proposition 3.6) and has a unique global maximum located at  $\theta^*$ , and as  $\Theta$  is compact, we get that  $\hat{\theta}_n$  converges  $\mathbb{P}_{\theta^*}$ -a.s. to  $\theta^*$  as  $n \rightarrow \infty$ .  $\square$

We now prove Proposition 3.10.

*Proof of Proposition 3.10.* Remind that  $h_{u,k,x}(\theta)$  is defined in (17). By definition of  $\ell(\theta)$  (see (26)) and using the  $L^1(\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^*})$  convergence of  $(h_{\partial,k,x}(\theta))_{k \in \mathbb{N}}$  to  $h_{\partial,\infty}(\theta)$  (remind Proposition 3.4), we have:

$$\begin{aligned} \ell(\theta^*) - \ell(\theta) &= \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [h_{\partial,\infty}(\theta^*) - h_{\partial,\infty}(\theta)] \\ &= \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} \left[ \lim_{k \rightarrow \infty} (h_{\partial,k,x}(\theta^*) - h_{\partial,k,x}(\theta)) \right] \\ &= \lim_{k \rightarrow \infty} \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [h_{\partial,k,x}(\theta^*) - h_{\partial,k,x}(\theta)]. \end{aligned}$$

Remind that  $H_{u,k,x}(\theta)$  is defined in (16). Then, write:

$$\mathbb{E}_{\theta^*} [h_{\partial,k,x}(\theta^*) - h_{\partial,k,x}(\theta)] = \mathbb{E}_{\mathcal{U}} \left[ \mathbb{E}_{\theta^*} \left[ \mathbb{E}_{\theta^*} \left[ \log \frac{H_{\partial,k,x}(\theta^*)}{H_{\partial,k,x}(\theta)} \middle| Y_{\Delta^*(\partial,k)}, X_{\mathbb{P}^k(\partial)} = x \right] \right] \right], \quad (34)$$

where the inner expectation is on  $Y_{\partial}$  conditionally on  $X_{\mathbb{P}^k(\partial)} = x$  and  $Y_{\Delta^*(\partial,k)}$  (and thus also implicitly on  $\mathcal{U}$  as  $\Delta^*(\partial,k) = \Delta_{\mathcal{U}}^*(\partial,k)$ ). Recalling from (16) that  $H_{\partial,k,x}(\theta)$  is the conditional density of  $Y_{\partial}$  given  $Y_{\Delta^*(\partial,k)}$  and  $X_{\mathbb{P}^k(\partial)} = x$ , we see that the inner (conditional) expectation in the right hand side is a Kullback-Leibler divergence and thus is non-negative. Hence, the two outer expectations and the limit  $\ell(\theta^*) - \ell(\theta)$  as  $k \rightarrow \infty$  are non-negative as well, and thus  $\theta^*$  is a global maximum of  $\ell$ .

Remark that if  $\theta$  is equivalent to  $\theta^*$ , then as the process  $(Y_u, u \in T)$  is stationary and has same law under both parameters, the roles of  $\theta^*$  and  $\theta$  can be swapped in the argument above, and thus we get  $\ell(\theta) = \ell(\theta^*)$ . Hence, any  $\theta$  equivalent to  $\theta^*$  is a global maximum of  $\ell$ .

We now turn to prove that any global maximum  $\theta \in \Theta$  of  $\ell$  is equivalent to  $\theta^*$ .

Remind that we use the letter  $p$  to denote (possibly conditional) densities of random variables, e.g.  $p_{\theta}(Y_u | Y_{\Delta^*(u,k)}, X_{\mathbb{P}^k(u)} = x)$  denotes the *conditional density* (w.r.t. the measure  $\mu$  defined in Assumption 2-(i)) under the parameter  $\theta$  of  $Y_u$  conditionally on  $Y_{\Delta^*(u,k)}$  and  $X_{\mathbb{P}^k(u)} = x$ . Note that  $\mathbb{P}_{\theta}(Y_u \in \cdot | Y_{\Delta^*(u,k)}, X_{\mathbb{P}^k(u)} = x)$  denotes the *distribution* under the parameter  $\theta$  of  $Y_u$  conditionally on  $Y_{\Delta^*(u,k)}$  and  $X_{\mathbb{P}^k(u)} = x$ .

We first need a variant of the convergence in Proposition 3.4 where instead of considering one vertex  $u$  as in  $h_{u,k,x}(\theta)$  we consider a whole subtree  $T^{\infty}(u, m)$  for any  $m \geq 1$  (this can be seen as



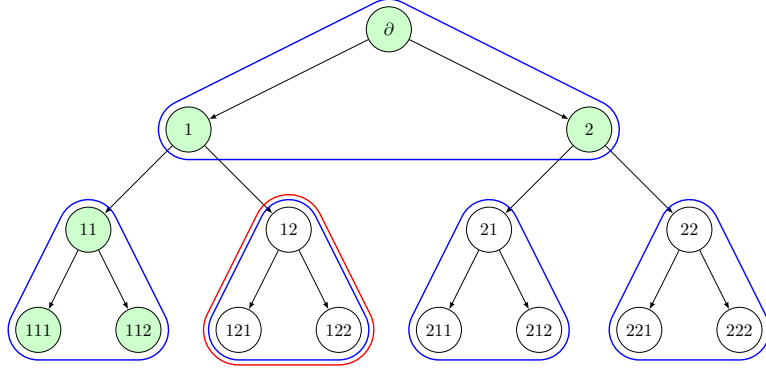


Figure 5: Illustration of the “past” subtree  $\Delta^*(T(u, m), k)$  of the block subtree  $T(u, m)$  for  $m = 1$ ,  $u = 12$  and  $k = 1$ . The block subtrees are circled with blue lines, and the block subtree  $T(12, 1)$  is circled a second time with a red line. The vertices in green are those in  $\Delta^*(T(12, 1), 1)$ . Note the difference with  $\Delta(u', k')$ , e.g. vertex 111 is in  $\Delta^*(T(12, 1), 1)$  but not in  $\Delta^*(12, 2)$ , and vertex 21 is in  $\Delta^*(121, 3)$  but not in  $\Delta^*(T(12, 1), 1)$ .

a convergence by block). To this end, we need to define an analogue of the breadth-first-search order relation  $<$  on  $T^\infty$  for subtree blocks of the form  $T^\infty(u, m)$ . Let  $m \geq 1$  be fixed. For  $u, v \in T^\infty$  with  $h(u) \equiv h(v) \pmod{m+1}$ , we write  $T^\infty(u, m) < T^\infty(v, m)$  if  $u < v$  (informally, “ $T^\infty(u, m)$  is above or on the left of  $T^\infty(v, m)$ ”). Note that the modulo congruence is there to insure the collection of block subtrees  $T^\infty(u, m)$  with  $h(u) \equiv h(\partial) \pmod{m+1}$  form a partition (i.e. a cover with non-overlapping subsets) of  $T^\infty$  (this still holds for any other class of congruence mod  $m+1$ ). Also note that in this congruence we have  $m+1$  and not  $m$ , because any subtree  $T^\infty(u, m)$  (e.g.  $T_m = T^\infty(\partial, m)$ ) spans over  $m+1$  different generations (remind that  $h(\partial) = 0$ ). We can then define the analogue of the subset  $\Delta^*(u, k)$  for subtree blocks, that is, for all  $u \in T^\infty$  and  $k \in \mathbb{N}$ , we define:

$$\Delta^*(T(u, m), k) = \bigcup \{T(v, m) : v \in \Delta^*(u, k(m+1)) \text{ such that } h(v) \equiv h(u) \pmod{m+1}\}.$$

See Figure 5 for an illustration of the “past” subtree  $\Delta^*(T(u, m), k)$  of the block subtree  $T(u, m)$ . (Informally, “the subset  $\Delta^*(T(u, m), k)$  is the union of the subtree blocks (with height  $m$ ) above and on the left of  $T(u, m)$  up to  $k$  block generations”. Note that we will not need to understand in details the geometry of the subset  $\Delta^*(T(u, m), k)$ , we only need to remember that all its vertices are upstream of the edge  $(p(u), u)$ , and we will then use the Markov property.) Remind that  $T^\infty(\partial, m) = T_m$ . Then, a straightforward adaptation of Lemma 3.3, and Propositions 3.4 and 3.5 to a decomposition of the log-likelihood into non-overlapping subtrees of height  $m$  instead of single vertices (see Appendix C for detailed proofs of those adaptations) give us for all  $\theta \in \Theta$ ,  $x \in \mathcal{X}$  and  $m \in \mathbb{N}^*$ :

$$\lim_{k \rightarrow \infty} \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [\log p_\theta(Y_{T_m} | Y_{\Delta^*(T_m, k)}, X_{p^{k(m+1)}(\partial)} = x)] = |T_m| \ell(\theta). \quad (35)$$

Let  $\mathcal{U}^+ = (\mathcal{U}_{(j)})_{1 \leq j < \infty}$  be a sequence of independent random variables with Bernoulli distribution of parameter  $1/2$  (note that  $\mathcal{U}^+$  can be seen as a random forward spine), which is independent of  $\mathcal{U}$  and of the HMT process  $(X, Y)$ . For all  $n \in \mathbb{N}^*$ , define the random vertex  $U_n$  as the unique vertex in  $G_n$  whose path from  $\partial$  is encoded by  $\mathcal{U}_{(1:n)}$  in Neveu’s notation. For all

$n \in \mathbb{N}$ , define the deterministic vertex  $U_{-n} = p^n(\partial)$ . Note that  $\partial = U_0$  and that  $U_{n-1}$  is the parent vertex of  $U_n$  for all  $n \in \mathbb{Z}$ . Moreover, using a similar argument as in Remark 3.1, note that for any  $m, k \in \mathbb{N}$ , the sequence of random shapes  $(\mathcal{S}h(\Delta(T^\infty(U_n, m), k)))_{n \in \mathbb{Z}}$  is stationary.

Now, pick  $\theta \in \Theta$  such that  $\ell(\theta) = \ell(\theta^*)$ . Thus for any positive integer  $n < m$ , we have:

$$\begin{aligned}
0 &= |T_m| (\ell(\theta^*) - \ell(\theta)) \\
&= \lim_{k \rightarrow \infty} \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} \left[ \log \frac{p_{\theta^*}(Y_{T_m} | Y_{\Delta^*}(T(\partial, m), k), X_{p^{k(m+1)}(\partial)} = x)}{p_{\theta}(Y_{T_m} | Y_{\Delta^*}(T(\partial, m), k), X_{p^{k(m+1)}(\partial)} = x)} \right] \\
&= \lim_{k \rightarrow \infty} \left\{ \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} \left[ \log \frac{p_{\theta^*}(Y_{T(U_{m-n}, n)} | Y_{\Delta^*}(T(\partial, m), k), X_{p^{k(m+1)}(\partial)} = x)}{p_{\theta}(Y_{T(U_{m-n}, n)} | Y_{\Delta^*}(T(\partial, m), k), X_{p^{k(m+1)}(\partial)} = x)} \right] \right. \\
&\quad \left. + \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} \left[ \log \frac{p_{\theta^*}(Y_{T_m \setminus T(U_{m-n}, n)} | Y_{\Delta^*}(T(\partial, m), k) \cup T(U_{m-n}, n), X_{p^{k(m+1)}(\partial)} = x)}{p_{\theta}(Y_{T_m \setminus T(U_{m-n}, n)} | Y_{\Delta^*}(T(\partial, m), k) \cup T(U_{m-n}, n), X_{p^{k(m+1)}(\partial)} = x)} \right] \right\} \\
&\geq \limsup_{k \rightarrow \infty} \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} \left[ \log \frac{p_{\theta^*}(Y_{T(U_{m-n}, n)} | Y_{\Delta^*}(T(\partial, m), k), X_{p^{k(m+1)}(\partial)} = x)}{p_{\theta}(Y_{T(U_{m-n}, n)} | Y_{\Delta^*}(T(\partial, m), k), X_{p^{k(m+1)}(\partial)} = x)} \right] \\
&= \limsup_{k \rightarrow \infty} \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} \left[ \log \frac{p_{\theta^*}(Y_{T_n} | Y_{\Delta^*}(T(U_{-m+n}, m), k), X_{p^{k(m+1)}(U_{-m+n})} = x)}{p_{\theta}(Y_{T_n} | Y_{\Delta^*}(T(U_{-m+n}, m), k), X_{p^{k(m+1)}(U_{-m+n})} = x)} \right], \tag{36}
\end{aligned}$$

where the inequality follows by noting that the second term is non-negative as an expectation of a (conditional) Kullback-Leibler divergence (using an argument similar as for (34) above), and the last equality follows by using stationarity of the HMT process  $(X, Y)$ , of the spinal process  $(U_n)_{n \in \mathbb{Z}}$ , and of the shape process  $(\mathcal{S}h(\Delta(T^\infty(U_n, m), k)))_{n \in \mathbb{Z}}$ . Note that the term in the lower bound is also non-negative as an expectation of a (conditional) Kullback-Leibler divergence.

Let  $n \in \mathbb{N}$  be fixed. Now, we define for all  $\theta \in \Theta$  and  $m, k \in \mathbb{N}^*$ :

$$\begin{aligned}
W_{m,k}(\theta) &= \log p_{\theta}(Y_{T_n} | Y_{\Delta^*}(T(U_{-m}, m+n), k), X_{p^{k(m+n+1)}(U_{-m})} = x), \\
\text{and} \quad W(\theta) &= \log p_{\theta}(Y_{T_n}). \tag{37}
\end{aligned}$$

Note that  $\log p_{\theta}(Y_{T_n})$  is well defined using an integral expression similar to (5) along Assumptions 2 and 3 and the comment on  $\pi_{\theta}$  after Lemma 2.3. From (36), we deduce that (where  $m$  in (37) and (38) below corresponds to  $m - n$  in (36)):

$$\forall m \in \mathbb{N}^*, \quad \lim_{k \rightarrow \infty} \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [W_{m,k}(\theta^*) - W_{m,k}(\theta)] = 0. \tag{38}$$

Hence, we have managed to insert a gap between the variables  $(Y_v, v \in T_n)$  whose density we examine and the variables  $(Y_v, v \in \Delta^*(T(U_{-m}, m+n), k))$  and  $X_{p^{k(m+n+1)}(U_{-m})}$  that appear in the conditioning. Remark that the following fact illustrates the gap between the variables: if  $u \in T_n$  and  $v \in \Delta^*(T(U_{-m}, m+n), k)$ , then the most recent common ancestor  $u \wedge v$  of  $u$  and  $v$  has height  $h(u \wedge v) < -m$ , that is  $u \wedge v$  is an ancestor of  $U_{-m}$ . See Figure 6 for a graphical illustration of this gap.

The idea is now to let this gap tend to infinity to show that in the limit the conditioning has no effect. Our next goal is thus to prove that:

$$\lim_{m \rightarrow \infty} \sup_{k \in \mathbb{N}} \left| \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [W_{m,k}(\theta^*) - W_{m,k}(\theta)] - \mathbb{E}_{\theta^*} [W(\theta^*) - W(\theta)] \right| = 0. \tag{39}$$

Combining (39) with (38), it is clear that if  $\theta \in \Theta$  is such that  $\ell(\theta) = \ell(\theta^*)$ , then we have that  $\mathbb{E}_{\theta^*} [\log [p_{\theta^*}(Y_{T_n}) / p_{\theta}(Y_{T_n})]] = 0$ , that is, the Kullback-Leibler divergence between the  $|T_n|$ -dimensional densities  $p_{\theta^*}(Y_{T_n})$  and  $p_{\theta}(Y_{T_n})$  is null. This implies, by the information inequality,

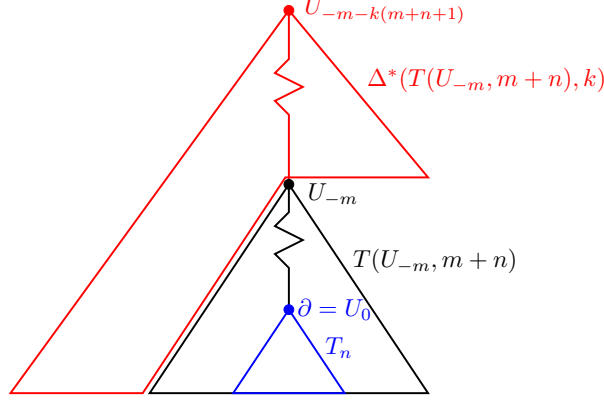


Figure 6: Illustration of the gap in (38) between the variables  $(Y_v, v \in T_n)$  (bottom triangle in blue) and the variables  $(Y_v, v \in \Delta^*(T(U_{-m}, m+n), k))$  and  $X_{\mathbf{p}^{k(m+n+1)}(U_{-m})}$  that appear in the conditioning (top partial triangle in red). Note that the two groups of variables are separated by the path from  $U_{-m-1} = \mathbf{p}(U_{-m})$  to  $\partial = U_0$ , which is of length  $m+1$ .

that these densities coincide except on a set with  $\mu^{\otimes |T_n|}$ -measure zero, so that the  $T_n$ -marginal laws of  $\mathbb{P}_{\theta^*}$  and  $\mathbb{P}_{\theta}$  agree. Because  $n$  was arbitrary, we find that  $\theta^*$  and  $\theta$  are equivalent.

What remains to do to complete the proof is thus to prove (39). Remind the definition of  $W_{m,k}(\theta)$  and  $W(\theta)$  in (37). Obviously, it is enough to prove that for all  $\theta \in \Theta$ , we have:

$$\lim_{m \rightarrow \infty} \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} \left[ \sup_{k \in \mathbb{N}} |W_{m,k}(\theta) - W(\theta)| \right] = 0. \quad (40)$$

Let  $\theta \in \Theta$  be fixed. To prove that (40) holds for  $\theta$ , we write (remind the discussion above on the gap between variables):

$$\begin{aligned} \exp(W_{m,k}(\theta)) &= p_{\theta}(Y_{T_n} | Y_{\Delta^*(T(U_{-m}, m+n), k)}, X_{\mathbf{p}^{k(m+n+1)}(U_{-m})} = x) \\ &= \int_{\mathcal{X} \times \mathcal{X}} p_{\theta}(Y_{T_n} | X_{\mathbf{p}(\partial)} = x_{\mathbf{p}(\partial)}) Q_{\theta}^{m-1}(x_{U_{-m}}; dx_{\mathbf{p}(\partial)}) \\ &\quad \times \mathbb{P}_{\theta}(X_{U_{-m}} \in dx_{U_{-m}} | Y_{\Delta^*(T(U_{-m}, m+n), k)}, X_{\mathbf{p}^{k(m+n+1)}(U_{-m})} = x), \end{aligned}$$

and

$$\exp(W(\theta)) = p_{\theta}(Y_{T_n}) = \int_{\mathcal{X} \times \mathcal{X}} p_{\theta}(Y_{T_n} | X_{\mathbf{p}(\partial)} = x_{\mathbf{p}(\partial)}) Q_{\theta}^{m-1}(x_{U_{-m}}; dx_{\mathbf{p}(\partial)}) \pi_{\theta}(dx_{U_{-m}}),$$

where remind from Lemma 2.3 that  $\pi_{\theta}$  is the stationary distribution of the process  $(X_u, u \in T^{\infty})$  with transition kernel  $Q_{\theta}$  (that is, under the distribution  $\mathbb{P}_{\theta}$ ). Note that we have the upper bound (remind that  $b^+$  is defined in Assumption 4):

$$p_{\theta}(Y_{T_n} | X_{\mathbf{p}(\partial)} = x_{\mathbf{p}(\partial)}) = \int_{\mathcal{X}^{T_n}} \prod_{u \in T_n} q_{\theta}(x_{\mathbf{p}(u)}, x_u) g_{\theta}(x_u, Y_u) \lambda(dx_u) \leq (b^+)^{|T_n|}. \quad (41)$$

Thus, as Assumptions 2 and 3 hold, applying the uniform geometric bound from Lemma 2.3 to the Markov chain  $(X_{U_j})_{j \in \mathbb{Z}}$  with transition kernel  $Q_{\theta}$ , we obtain  $\mathbb{P}_{\theta^*}$ -a.s. :

$$\sup_{k \in \mathbb{N}} \left| p_{\theta}(Y_{T_n} | Y_{\Delta^*(T(U_{-m}, m+n), k)}, X_{\mathbf{p}^{k(m+n+1)}(U_{-m})} = x) - p_{\theta}(Y_{T_n}) \right| \leq (b^+)^{|T_n|} (1 - \sigma^-)^{m-1}. \quad (42)$$

Moreover, as we have the lower bound:

$$\begin{aligned} p_\theta(Y_{T_n} | X_{\mathbb{P}(\partial)} = x_{\mathbb{P}(\partial)}) &= \int_{\mathcal{X}^{T_n}} \prod_{u \in T_n} q_\theta(x_{\mathbb{P}(u)}, x_u) g_\theta(x_u, Y_u) \lambda(dx_u) \\ &\geq \prod_{u \in T_n} \sigma^- b^-(Y_u), \end{aligned} \quad (43)$$

this implies that  $p_\theta(Y_{T_n} | Y_{\Delta^*(T(U_{-m}, m+n), k)}, X_{\mathbb{P}^{k(m+n+1)}(U_{-m})} = x)$  and  $p_\theta(Y_{T_n})$  both obey the same lower bound. This lower bound combined with the observation that  $b^-(Y_u) > 0$  for all  $u \in T_n$  (which follows from Assumption 3-(ii)), and the bound  $|\log(x) - \log(y)| \leq |x - y|/x \wedge y$ , (42) shows that:

$$\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^*}\text{-a.s.} \quad \lim_{m \rightarrow \infty} \sup_{k \in \mathbb{N}} |W_{m,k}(\theta) - W(\theta)| = 0.$$

Using the bounds (41) and (43) with Assumptions 3 and 4-(ii), we get:

$$\mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} \left[ \sup_{m \in \mathbb{N}^*} \sup_{k \in \mathbb{N}} |W_{m,k}(\theta)| \right] < \infty.$$

Hence, as this expectation is finite, (40) follows from dominated convergence. This concludes the proof.  $\square$

## 4 Asymptotic normality of the MLE

In this section, we prove that the MLE for the HMT has asymptotic normal fluctuations. We keep the assumptions used in Section 3. This section is divided in two parts: we first prove the asymptotic normality of the score, and then we prove a strong law of large numbers for the observed information. Together with the strong consistency, those two results imply the asymptotic normality of the MLE.

We will need the following assumption for existence and regularity of the gradient and Hessian of the transition kernels. Remind that  $\mathbb{P}_\theta$  denotes the stationary probability distribution under the parameter  $\theta \in \Theta$  of the HMT process  $(X, Y)$ , and by  $\mathbb{E}_\theta$  the corresponding expectation. Also remind that the measures  $\lambda$  and  $\mu$  are defined in Assumption 2. We denote by  $\nabla_\theta$  and  $\nabla_\theta^2$ , respectively, the gradient and Hessian operator w.r.t. the parameter  $\theta \in \Theta$ . With a slight abuse of notations, we denote by  $\|\cdot\|$  the euclidean norm on either  $\mathbb{R}^d$  or  $\mathbb{R}^{d \times d}$ .

**Assumption 7** (Regularity of the gradient, [CMR05, Assumption 12.5.1]). *There exists an open (for the trace topology on  $\Theta \subset \mathbb{R}^d$ ) neighborhood  $\mathcal{O} = \{\theta \in \Theta : \|\theta - \theta^*\| < \delta_0\}$  of  $\theta^*$  such that the following hold.*

(i) *For all  $(x, x') \in \mathcal{X} \times \mathcal{X}$  and all  $y \in \mathcal{Y}$ , the functions  $\theta \mapsto q_\theta(x, x')$  and  $\theta \mapsto g_\theta(x, y)$  are twice continuously differentiable on  $\mathcal{O}$ .*

(ii) *We have:*

$$\sup_{\theta \in \mathcal{O}} \sup_{x, x'} \|\nabla_\theta \log q_\theta(x, x')\| < \infty, \quad \text{and} \quad \sup_{\theta \in \mathcal{O}} \sup_{x, x'} \|\nabla_\theta^2 \log q_\theta(x, x')\| < \infty.$$

(iii) *We have:*

$$\mathbb{E}_{\theta^*} \left[ \sup_{\theta \in \mathcal{O}} \sup_x \|\nabla_\theta \log g_\theta(x, Y_\partial)\|^2 \right] < \infty, \quad \text{and} \quad \mathbb{E}_{\theta^*} \left[ \sup_{\theta \in \mathcal{O}} \sup_x \|\nabla_\theta^2 \log g_\theta(x, Y_\partial)\| \right] < \infty.$$

(iv) For  $\mu$ -almost all  $y \in \mathcal{Y}$ , there exists a function  $f_y : \mathcal{X} \rightarrow \mathbb{R}_+$  in  $L^1(\lambda)$  such that we have  $\sup_{\theta \in \mathcal{O}} g_\theta(x, y) \leq f_y(x)$ .

(v) For  $\lambda$ -almost all  $x \in \mathcal{X}$ , there exist functions  $f_x^1 : \mathcal{X} \rightarrow \mathbb{R}_+$  and  $f_x^2 : \mathcal{X} \rightarrow \mathbb{R}_+$  in  $L^1(\mu)$  such that  $\sup_{\theta \in \mathcal{O}} \|\nabla_\theta g_\theta(x, y)\| \leq f_x^1(y)$  and  $\sup_{\theta \in \mathcal{O}} \|\nabla_\theta^2 g_\theta(x, y)\| \leq f_x^2(y)$ .

These assumptions insures that the log-likelihood  $\ell_{n,x}$  is twice continuously differentiable, and that the score function  $\nabla_\theta \ell_{n,x}(\theta)$  and the observed information  $-\nabla_\theta^2 \ell_{n,x}(\theta)$  exist and are in  $L^2(\mathbb{P}_{\theta^*})$  and  $L^1(\mathbb{P}_{\theta^*})$ , respectively.

## 4.1 Asymptotic normality of the score

In this subsection, we prove the asymptotic normality of the score under the true parameter  $\theta^*$ . Note that the score function can be written for all  $n \in \mathbb{N}$  and  $x \in \mathcal{X}$  as:

$$\nabla_\theta \ell_{n,x}(\theta) = \sum_{u \in T_n} \nabla_\theta \log \left[ \int g_\theta(X_u, Y_u) \mathbb{P}_\theta(X_u \in dx_u \mid Y_{\Delta^*(u,h(u))}, X_\partial = x) \right],$$

and  $\nabla_\theta \ell_{n,x}(\theta)$  is implicitly a function of  $Y_{T_n}$ .

### 4.1.1 Decomposition of the score as a sum of increments

Remind that for  $u \in T$ , the subtrees  $\Delta^*(u, k)$  and  $\Delta(u, k)$  are defined in Section 2.4 for  $k \leq h(u)$  (with  $\Delta^*(u) = \Delta^*(u, h(u))$  and  $\Delta(u) = \Delta(u, h(u))$ ) and the random subtrees  $\Delta^*(u, k)$  and  $\Delta(u, k)$  are defined in Section 3.1 for  $k > h(u)$ . Also remind that we use the letter  $p$  to denote (possibly conditional) probability density, and in particular remind that  $p_\theta(Y_\Delta \mid X_\partial = x_\partial)$  for any subtree  $\Delta \subset T$  with root  $\partial$  is defined in (15) in Section 3.1.2 (with the convention  $p_\theta(Y_\emptyset \mid X_\partial = x_\partial) = 1$ ). Using (16) and (17) in Section 3.1.2, note that for any  $u \in T$  and  $x \in \mathcal{X}$ , we have:

$$h_{u,h(u),x}(\theta) = \log p_\theta(Y_{\Delta(u)} \mid X_\partial = x) - \log p_\theta(Y_{\Delta^*(u)} \mid X_\partial = x). \quad (44)$$

Using elementary computation along with permutations of the integral and the gradient operator which are valid under Assumption 7 (note that this result is also known as *Fisher identity*, see [CMR05, Proposition 10.1.6]), we get:

$$\begin{aligned} & \nabla_\theta \log p_\theta(Y_{\Delta(u)} \mid X_\partial = x) \\ &= \nabla_\theta \log g_\theta(x, Y_\partial) + \mathbb{E}_\theta \left[ \sum_{v \in \Delta(u) \setminus \{\partial\}} \phi_\theta(X_{p(v)}, X_v, Y_v) \middle| Y_{\Delta(u)}, X_\partial = x \right], \end{aligned} \quad (45)$$

where

$$\phi_\theta(x', x, y) = \nabla_\theta \log [q_\theta(x', x) g_\theta(x, y)]. \quad (46)$$

Note that under Assumption 7,  $\|\phi_\theta(X_{p(v)}, X_v, Y_v)\|$  is upper bounded by a square integrable function of  $Y_v$  (which does not depend on  $\theta$ ), and  $\phi_\theta(X_{p(v)}, X_v, Y_v)$  is thus integrable conditionally on  $Y_{\Delta(u)}$  and  $X_\partial = x$ . Also note that  $\nabla_\theta \log g_\theta(x, Y_\partial)$  is  $\mathbb{P}_{\theta^*}$ -a.s. finite by Assumption 7-(iii).

Combining those two equations with (18) in Section 3.1.2, we can express the score function as:

$$\begin{aligned} \nabla_\theta \ell_{n,x}(\theta) &= \nabla_\theta \log g_\theta(x, Y_\partial) + \sum_{u \in T_n^*} \mathbb{E}_\theta \left[ \sum_{\Delta(u) \setminus \{\partial\}} \phi_\theta(X_{p(v)}, X_v, Y_v) \middle| Y_{\Delta(u)}, X_\partial = x \right] \\ &\quad - \sum_{u \in T_n^*} \mathbb{E}_\theta \left[ \sum_{\Delta^*(u) \setminus \{\partial\}} \phi_\theta(X_{p(v)}, X_v, Y_v) \middle| Y_{\Delta^*(u)}, X_\partial = x \right]. \end{aligned}$$

We want to express the score function  $\nabla_{\theta} \ell_{n,x}(\theta)$  as a sum of increments (conditional scores) in order to apply a convergence result for the normalized score. To this end, define for every  $u \in T$ ,  $k \in \mathbb{N}$  and  $x \in \mathcal{X}$ , the function  $\dot{h}_{u,k,x}(\theta)$  by  $\dot{h}_{u,0,x}(\theta) = \nabla_{\theta} \log g_{\theta}(x, Y_u)$  if  $k = 0$ , and otherwise by:

$$\begin{aligned} \dot{h}_{u,k,x}(\theta) = & \mathbb{E}_{\theta} \left[ \sum_{v \in \Delta(u,k) \setminus \{p^k(u)\}} \phi_{\theta}(X_{p(v)}, X_v, Y_v) \middle| Y_{\Delta(u,k)}, X_{p^k(u)} = x \right] \\ & - \mathbb{E}_{\theta} \left[ \sum_{v \in \Delta^*(u,k) \setminus \{p^k(u)\}} \phi_{\theta}(X_{p(v)}, X_v, Y_v) \middle| Y_{\Delta^*(u,k)}, X_{p^k(u)} = x \right]. \end{aligned}$$

Note that  $\dot{h}_{u,k,x}(\theta)$  is well defined as  $\Delta(u,k)$  is finite and as  $\phi_{\theta}(X_{p(v)}, X_v, Y_v)$  is integrable conditionally on  $Y_{\Delta^*(u,k)}$  and  $X_{p^k(u)} = x$  under Assumption 7 (see the comment after (46)). Also note that  $\dot{h}_{u,k,x}(\theta)$  is the gradient w.r.t.  $\theta$  of  $h_{u,k,x}(\theta)$  defined in (17) (see (44) and (45) for the case  $k = h(u)$ ). Furthermore, note that  $\dot{h}_{u,k,x}(\theta)$  is a function of  $Y_{\Delta(u,k)}$  with an implicit dependence on  $\mathcal{U}$  through  $\Delta(u,k)$ , and that  $\dot{h}_{u,k,x}(\theta)$  does not depend on  $\mathcal{U}$  if  $k \leq h(u)$ .

Using the increment functions  $\dot{h}_{u,k,x}(\theta)$ , we can rewrite the score function as:

$$\nabla_{\theta} \ell_{n,x}(\theta) = \sum_{u \in T_n} \dot{h}_{u,h(u),x}(\theta). \quad (47)$$

#### 4.1.2 Construction of score increments with infinite past

Our goal is to let  $k \rightarrow \infty$  as before to get a limit function  $\dot{h}_{u,\infty}$ . We now proceed to construct  $\dot{h}_{u,\infty}$ . First, we rewrite  $\dot{h}_{u,k,x}(\theta)$  (which is in  $L^2(\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^*})$  by Assumption 7), as:

$$\begin{aligned} \dot{h}_{u,k,x}(\theta) = & \mathbb{E}_{\theta}[\phi_{\theta}(X_{p(u)}, X_u, Y_u) | Y_{\Delta(u,k)}, X_{p^k(u)} = x] \\ & + \sum_{v \in \Delta^*(u,k) \setminus \{p^k(u)\}} \left( \mathbb{E}_{\theta}[\phi_{\theta}(X_{p(v)}, X_v, Y_v) | Y_{\Delta(u,k)}, X_{p^k(u)} = x] \right. \\ & \left. - \mathbb{E}_{\theta}[\phi_{\theta}(X_{p(v)}, X_v, Y_v) | Y_{\Delta^*(u,k)}, X_{p^k(u)} = x] \right). \quad (48) \end{aligned}$$

We will need the following lemma that states a coupling bound that works “backwards in time”, or rather along the path between a vertex  $v$  and the newly observed vertex  $u$ . Remind from (12) on page 15 that  $\Delta(u,k)$  is a random subtree of the deterministic subtree  $T^{\infty}(p^k(u), k)$ .

**Lemma 4.1** (Total variation bound “backwards in time”). *Assume that Assumptions 2–3 hold. Let  $k \in \mathbb{N}^*$ ,  $x \in \mathcal{X}$  and  $u \in T$ , and let  $v \in T^{\infty}(p^k(u), k) \setminus \{u\}$ . Then, we have:*

$$\left\| \mathbb{P}_{\theta}(X_v \in \cdot | Y_{\Delta(u,k)}, X_{p^k(u)} = x) - \mathbb{P}_{\theta}(X_v \in \cdot | Y_{\Delta^*(u,k)}, X_{p^k(u)} = x) \right\|_{TV} \leq \rho^{d(u,v)-1}.$$

The proof of Lemma 4.1, which is postponed to Appendix B, relies on a “backward in time” bound from  $p(u)$  to  $u \wedge v$  and then a “forward in time” bound from  $u \wedge v$  to  $v$ , and using the initial distributions  $\mathbb{P}_{\theta}(X_{p(u)} \in \cdot | Y_{\Delta}, X_{p^k(u)} = x)$  with  $\Delta$  equal to  $\Delta(u,k)$  and  $\Delta^*(u,k)$ , respectively. Note that this proof is similar to the proofs for Lemma 3.2 and [CMR05, Proposition 12.5.4].

The following lemma gives an  $L^2$ -bound on the difference between  $\dot{h}_{u,k,x}(\theta)$  and  $\dot{h}_{u,k',x'}(\theta)$  with a geometric decay. As we will reuse this result later with different functions, we state a more general version.

Note that the condition  $\rho < 1/\sqrt{2}$  on the mixing rate  $\rho$  of the HMT process  $(X, Y)$  is due to the coupling bounds and the grouping of terms used in the proof of Lemma 4.2 (the upper bounds at the end of the proof only add a constant multiplicative factor). See the discussion in Remark 1.5.

**Lemma 4.2.** *Assume that Assumptions 1–4. Further assume that  $\rho < 1/\sqrt{2}$ .*

*Let  $\Theta_0$  be a closed ball in  $\Theta$ , and let  $\psi$  be a Borel function from  $\Theta_0 \times \mathcal{X}^2 \times \mathcal{Y}$  to  $\mathbb{R}^d$  for some  $d \in \mathbb{N}$  such that for all  $x, x' \in \mathcal{X}$  and  $y \in \mathcal{Y}$ ,  $\theta \mapsto \psi(\theta, x, x', y) = \psi_\theta(x, x', y)$  is a continuous function on  $\Theta_0$ . Furthermore, assume that there exists  $b \in [1, +\infty)$  such that:*

$$\mathbb{E}_{\theta^*} \left[ \sup_{\theta \in \Theta_0} \sup_{x, x' \in \mathcal{X}} \|\psi_\theta(x, x', Y_\theta)\|^b \right] < \infty.$$

*Let  $\xi_{u,k,x}(\theta)$  be defined as in (48) (with  $\hat{h}_{u,k,x}(\theta)$  and  $\phi_\theta$  replaced by  $\xi_{u,k,x}(\theta)$  and  $\psi_\theta$ , respectively), and note that it is in  $L^b(\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^*})$ . Then, there exists a finite constant  $C < \infty$  such that for all  $u \in T$  and  $k' \geq k \geq 1$ , we have:*

$$\begin{aligned} & \left( \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} \left[ \sup_{\theta \in \Theta_0} \sup_{x, x' \in \mathcal{X}} \|\xi_{u,k,x}(\theta) - \xi_{u,k',x'}(\theta)\|^b \right] \right)^{1/b} \\ & \leq C \left( \mathbb{E}_{\theta^*} \left[ \sup_{\theta \in \Theta_0} \sup_{x, x' \in \mathcal{X}} \|\psi_\theta(x, x', Y_\theta)\|^b \right] \right)^{1/b} k (\max(\rho, 2\rho^2))^{k/2}. \end{aligned}$$

As a consequence of Lemma 4.2, for all  $u \in T$  and  $x \in \mathcal{X}$ , the sequence of function  $(\xi_{u,k,x}(\theta))_{n \in \mathbb{N}}$  converges in  $L^b(\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^*})$  to some limit function  $\xi_{u,\infty}(\theta)$  which does not depend on  $x$ . Moreover, the bound in Lemma 4.2 still holds when  $\xi_{u,k',x'}(\theta)$  is replaced by  $\xi_{u,\infty}(\theta)$ .

For the particular choice of  $\psi_\theta = \phi_\theta$ , under Assumptions-1-4 and 7, for all  $u \in T$ , we denote by  $\hat{h}_{u,\infty}(\theta)$  the limit function of the sequence  $(\hat{h}_{u,k,x}(\theta))_{n \in \mathbb{N}}$  (for all  $x \in \mathcal{X}$ ) which is in  $L^2(\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^*})$ .

As an immediate corollary of Lemma 4.2, there exists a finite constant  $C' < \infty$  such that for all  $x \in \mathcal{X}$ ,  $u \in T^*$  and  $k \geq 1$ , we have that  $(\mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [\sup_{x \in \mathcal{X}} \|\xi_{u,k,x}(\theta)\|^b])^{1/b} \leq \mathbb{E}_{\theta^*} [\sup_{x \in \mathcal{X}} \|\xi_{u,1,x}(\theta)\|^b]^{1/b} + C' < \infty$ , (note that by stationarity,  $\mathbb{E}_{\theta^*} [\sup_{x \in \mathcal{X}} \|\xi_{u,1,x}(\theta)\|^b]^{1/b} = \mathbb{E}_{\theta^*} [\sup_{x \in \mathcal{X}} \|\xi_{v,1,x}(\theta)\|^b]^{1/b}$  for any other  $v \in T^*$ ). Hence, we get:

$$\sup_{\theta \in \mathcal{O}} \sup_{u \in T} \sup_{k \geq 1} \left( \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} \left[ \sup_{x \in \mathcal{X}} \|\xi_{u,k,x}(\theta)\|^b \right] \right)^{1/b} < \infty. \quad (49)$$

*Proof.* We mimic the scheme of the proof of [CMR05, Lemma 12.5.3].

Let  $u \in T$  and  $k' \geq k \geq 1$  be fixed. The idea of the proof is to match, for each vertex index  $v$  of the sums expressing  $\xi_{u,k,x}(\theta)$  and  $\xi_{u,\infty}(\theta)$ , pairs of terms that are close. To be more precise, we match:

1. For  $v$  close to  $u$ ,

$$\mathbb{E}_\theta[\psi_\theta(X_{\mathbb{P}(v)}, X_v, Y_v) \mid Y_{\Delta(u,k)}, X_{\mathbb{P}^k(u)} = x]$$

and

$$\mathbb{E}_\theta[\psi_\theta(X_{\mathbb{P}(v)}, X_v, Y_v) \mid Y_{\Delta(u,k')}, X_{\mathbb{P}^{k'}(u)} = x'],$$

and similarly for the corresponding terms with  $\Delta(u, k)$  and  $\Delta(u, k')$  replaced by  $\Delta^*(u, k)$  and  $\Delta^*(u, k')$ , respectively;

2. For  $v$  far from  $u$ ,

$$\mathbb{E}_\theta[\psi_\theta(X_{p(v)}, X_v, Y_v) \mid Y_{\Delta(u,k)}, X_{p^k(u)} = x]$$

and

$$\mathbb{E}_\theta[\psi_\theta(X_{p(v)}, X_v, Y_v) \mid Y_{\Delta^*(u,k)}, X_{p^k(u)} = x],$$

and similarly for the corresponding terms with  $k$  and  $x$  replaced by  $k'$  and  $x'$ , respectively.

Remind from (12) on page 15 that  $\Delta(u, k) \subset T^\infty(p^k(u), k)$  and that the subtree  $\Delta(u, k)$  is random while the subtree  $T^\infty(p^k(u), k)$  is deterministic. Let  $(x, x') \in \mathcal{X} \times \mathcal{X}$  and let  $v \in T^\infty(p^k(u), k) \setminus \{p^k(u)\}$ , which implies that  $p(v) \in \Delta(u, k)$ .

We start with the first kind of matches. Using the Markov property (remind (2)), we have:

$$\begin{aligned} & \left\| \mathbb{E}_\theta[\psi_\theta(X_{p(v)}, X_v, Y_v) \mid Y_{\Delta(u,k)}, X_{p^k(u)} = x] - \mathbb{E}_\theta[\psi_\theta(X_{p(v)}, X_v, Y_v) \mid Y_{\Delta(u,k')}, X_{p^{k'}(u)} = x'] \right\| \\ &= \left\| \int_{\mathcal{X}^3} \psi_\theta(x_{p(v)}, x_v, Y_v) \mathbb{P}_\theta(X_v \in dx_v \mid Y_{\Delta(u,k) \cap T(v)}, X_{p(v)} = x_{p(v)}) \right. \\ & \quad \times \mathbb{P}_\theta(X_{p(v)} \in dx_{p(v)} \mid Y_{\Delta(u,k)}, X_{p^k(u)} = x_{p^k(u)}) \\ & \quad \left. \times [\delta_x(dx_{p^k(u)}) - \mathbb{P}_\theta(X_{p^k(u)} \in dx_{p^k(u)} \mid Y_{\Delta(u,k')}, X_{p^{k'}(u)} = x')] \right\| \\ & \leq 2 \sup_{x_1, x_2 \in \mathcal{X}} \|\psi_\theta(x_1, x_2, Y_v)\| \rho^{d(v, p^k(u))-1}, \end{aligned} \quad (50)$$

where the inequality is obtained using Lemma 3.2 (note that  $d(p(v), p^k(u)) = d(v, p^k(u)) - 1$ ). Note that this upper bound is a.s. finite as  $\sup_{x_1, x_2 \in \mathcal{X}} \|\phi_\theta(x_1, x_2, Y_v)\|$  is in  $L^b(\mathbb{P}_U \otimes \mathbb{P}_{\theta^*})$  by assumption (remind that the HMT process  $(X, Y)$  is stationary by Assumption 1). For  $v \neq u$ , note that this bound remains valid if  $\Delta(u, k)$  and  $\Delta(u, k')$  are replaced by  $\Delta^*(u, k)$  and  $\Delta^*(u, k')$ , respectively. Obviously, this bound is small if  $v$  is far away from  $p^k(u)$  (remind that  $k$  is fixed).

We now give a bound for the second kind of matches. Assume that  $v \neq u$ . If  $v$  is not an ancestor of  $u$  (then  $d(u, v) = d(u, p(v)) + 1$ ), using the Markov property (remind (2)) and Lemma 4.1, we get:

$$\begin{aligned} & \left\| \mathbb{E}_\theta[\psi_\theta(X_{p(v)}, X_v, Y_v) \mid Y_{\Delta(u,k)}, X_{p^k(u)} = x] - \mathbb{E}_\theta[\psi_\theta(X_{p(v)}, X_v, Y_v) \mid Y_{\Delta^*(u,k)}, X_{p^k(u)} = x] \right\| \\ &= \left\| \int_{\mathcal{X}^3} \psi_\theta(x_{p(v)}, x_v, Y_v) \mathbb{P}_\theta(X_v \in dx_v \mid Y_{\Delta(u,k) \cap T(v)}, X_{p(v)} = x_{p(v)}) \right. \\ & \quad \left. \times [\mathbb{P}_\theta(X_{p(v)} \in dx_{p(v)} \mid Y_{\Delta(u,k)}, X_{p^k(u)} = x) - \mathbb{P}_\theta(X_{p(v)} \in dx_{p(v)} \mid Y_{\Delta^*(u,k)}, X_{p^k(u)} = x)] \right\| \\ & \leq 2 \sup_{\theta \in \Theta_0} \sup_{x_1, x_2 \in \mathcal{X}} \|\psi_\theta(x_1, x_2, Y_v)\| \rho^{d(u,v)-2}. \end{aligned}$$

If  $v$  is an ancestor of  $u$  (then  $d(u, p(v)) = d(u, v) + 1$ ), using the Markov property (remind (2)) and Lemma 4.1, we get:

$$\begin{aligned} & \left\| \mathbb{E}_\theta[\psi_\theta(X_{p(v)}, X_v, Y_v) \mid Y_{\Delta(u,k)}, X_{p^k(u)} = x] - \mathbb{E}_\theta[\psi_\theta(X_{p(v)}, X_v, Y_v) \mid Y_{\Delta^*(u,k)}, X_{p^k(u)} = x] \right\| \\ &= \left\| \int_{\mathcal{X}^3} \psi_\theta(x_{p(v)}, x_v, Y_v) \mathbb{P}_\theta(X_{p(v)} \in dx_{p(v)} \mid Y_{\Delta(u,k) \setminus T(v)}, X_v = x_v) \right. \\ & \quad \left. \times [\mathbb{P}_\theta(X_v \in dx_v \mid Y_{\Delta(u,k)}, X_{p^k(u)} = x) - \mathbb{P}_\theta(X_v \in dx_v \mid Y_{\Delta^*(u,k)}, X_{p^k(u)} = x)] \right\| \\ & \leq 2 \sup_{\theta \in \Theta_0} \sup_{x_1, x_2 \in \mathcal{X}} \|\psi_\theta(x_1, x_2, Y_v)\| \rho^{d(u,v)-1}. \end{aligned}$$



In both cases, we get:

$$\begin{aligned} & \left\| \mathbb{E}_\theta[\psi_\theta(X_{\mathfrak{p}(v)}, X_v, Y_v) \mid Y_{\Delta(u,k)}, X_{\mathfrak{p}^k(u)} = x] - \mathbb{E}_\theta[\psi_\theta(X_{\mathfrak{p}(v)}, X_v, Y_v) \mid Y_{\Delta^*(u,k)}, X_{\mathfrak{p}^k(u)} = x] \right\| \\ & \leq 2 \sup_{\theta \in \Theta_0} \sup_{x_1, x_2 \in \mathcal{X}} \|\psi_\theta(x_1, x_2, Y_v)\| \rho^{d(u,v)-2}. \end{aligned} \quad (51)$$

Note that the same bound remain valid for the corresponding terms with  $k$  and  $x$  replaced by  $k'$  and  $x'$ , respectively, and with  $v \in \Delta(u, k') \setminus \{\mathfrak{p}^{k'}(u)\}$  instead of  $v \in \Delta(u, k) \setminus \{\mathfrak{p}^k(u)\}$ . This bound is small if  $v$  is far away from  $u$ .

Remind from (12) on page 15 that  $\Delta(u, k) \subset T^\infty(\mathfrak{p}^k(u), k)$  and as  $k' \geq k$  note that:

$$\Delta(u, k') \setminus \Delta(u, k) \subset T^\infty(\mathfrak{p}^{k'}(u), k') \setminus T^\infty(\mathfrak{p}^k(u), k).$$

For a vertex  $v \in T^\infty(\mathfrak{p}^k(u), k) \setminus \{\mathfrak{p}^k(u)\}$  (note that  $u \wedge v \in T^\infty(\mathfrak{p}^k(u))$ ), note that the term  $\rho^{d(v, \mathfrak{p}^k(u))-1}$  is smaller than  $\rho^{d(u,v)-2}$  whenever  $d(v, \mathfrak{p}^k(u)) > d(u, v) - 1$ , that is when  $d(u \wedge v, \mathfrak{p}^k(u)) \geq d(u \wedge v, u)$ , that is when  $d(u \wedge v, u) \leq k/2$ .

Combining those facts with the bounds (50) and (51), and using Minkowski's inequality for the  $L^b$ -norm, we find that  $(\mathbb{E}_U \otimes \mathbb{E}_{\theta^*} [\sup_{\theta \in \Theta_0} \sup_{x, x' \in \mathcal{X}} \|\xi_{u,k,x}(\theta) - \xi_{u,k',x'}(\theta)\|^b])^{1/b}$  is upper bounded by:

$$4 \sum_{v \in T^\infty(\mathfrak{p}^{\lfloor k/2 \rfloor}(u), \lfloor k/2 \rfloor)} \rho^{d(v, \mathfrak{p}^k(u))-1} + 4 \sum_{v \in T^\infty(\mathfrak{p}^{k'}(u), k') \setminus T^\infty(\mathfrak{p}^{\lfloor k/2 \rfloor}(u), \lfloor k/2 \rfloor)} \rho^{d(u,v)-2}, \quad (52)$$

up to the factor  $(\mathbb{E}_{\theta^*} [\sup_{\theta \in \Theta_0} \sup_{x_1, x_2 \in \mathcal{X}} \|\psi_\theta(x_1, x_2, Y_v)\|^b])^{1/b}$  (remind that the process  $(Y_u, u \in T^\infty)$  is stationary under Assumption 1). Denote by  $A_1$  and  $A_2$  respectively the first and second terms in (52). We are going to reindex those sums by  $j := d(u, u \wedge v)$  and  $q := d(u \wedge v, v)$  with  $q \leq j$ . Note that if  $q > 0$ , then the first vertex after  $u \wedge v = \mathfrak{p}^j(u)$  on the path from  $u$  to  $v$  cannot be  $\mathfrak{p}^{j-1}(u)$  and must be the other children of  $\mathfrak{p}^j(u)$ . Thus, there are  $2^{q-1}$  choices of  $v$  with the same coding  $(j, q)$  with  $0 < q \leq j$ . Hence, we get:

$$A_1 = 4 \sum_{j=0}^{\lfloor k/2 \rfloor} \rho^{k-j-1} \left( 1 + \sum_{q=1}^j 2^{q-1} \rho^q \right) \quad \text{and} \quad A_2 = 4 \sum_{j=\lfloor k/2 \rfloor+1}^{k'} \rho^{j-2} \left( 1 + \sum_{q=1}^j 2^{q-1} \rho^q \right).$$

Remark that there exists a finite constant  $C < \infty$  (which depends on the value of  $\rho$ ) such that for all  $j \in \mathbb{N}^*$  we have  $(1 + \sum_{q=1}^j 2^{q-1} \rho^q) \leq C \max(j, (2\rho)^j)$  and  $\sum_{q=j}^{\infty} q \rho^q \leq C \rho^j$ . Hence, there exists a finite constant  $C' < \infty$  (which depends only on the value of  $\rho$ ) such that (remind that  $\rho < 1/\sqrt{2}$ ):

$$A_1 \leq C' \left( k \rho^{k/2} + (2\rho^2)^{k/2} \right) \quad \text{and} \quad A_2 \leq C' \left( \rho^{k/2} + (2\rho^2)^{k/2} \right). \quad (53)$$

Combining (52) and (53), we get that the bound in the lemma holds. This concludes the proof of the lemma.  $\square$

### 4.1.3 Asymptotic normality of the score

Define the limiting Fisher information as:

$$\mathcal{I}(\theta^*) = \mathbb{E}_U \otimes \mathbb{E}_{\theta^*} \left[ \dot{\mathfrak{h}}_{\partial, \infty}(\theta^*) \dot{\mathfrak{h}}_{\partial, \infty}(\theta^*)^t \right], \quad (54)$$

where we see  $\dot{\mathfrak{h}}_{\partial, \infty}(\theta^*)$  as a column vector.

For the asymptotic normality of the score, we need the following extra regularity assumption of the gradient.

**Assumption 8** ( $L^4$  gradient regularity). *In addition to Assumption 7, we have:*

$$\mathbb{E}_{\theta^*} \left[ \sup_{\theta \in \mathcal{O}} \sup_{x \in \mathcal{X}} \|\nabla_{\theta} \log g_{\theta}(x, Y_{\partial})\|^4 \right] < \infty.$$

We are now ready to prove the following theorem stating the asymptotic normality of the normalized score towards a centered Gaussian random variable whose variance is the limiting Fisher information. Note that the condition  $\rho < 1/\sqrt{2}$  on the mixing rate  $\rho$  of the HMT process  $(X, Y)$  comes from the use of Lemma 4.2 in the proof of this theorem. See the discussion in Remark 1.5 for comments on this condition on  $\rho$ .

**Theorem 4.3** (Asymptotic normality of the normalized score). *Assume that Assumptions 1–4 and 7–8 hold with  $\theta^* \in \Theta$  given. Further assume that  $\rho < 1/\sqrt{2}$ . Then, for all  $x \in \mathcal{X}$ , we have:*

$$|T_n|^{-1/2} \nabla_{\theta} \ell_{n,x}(\theta^*) \xrightarrow[n \rightarrow \infty]{(d)} \mathcal{N}(0, \mathcal{I}(\theta^*)) \quad \text{under } \mathbb{P}_{\theta^*},$$

where  $\mathcal{N}(0, M)$  denotes the centered Gaussian distribution with covariance matrix  $M$ , and  $\mathcal{I}(\theta^*)$  is the limiting Fisher information defined in (54).

*Proof. Step 1: Approximation of the score by the stationary score.*

Remind from Lemma 2.3 that  $\pi_{\theta^*}$  denotes the invariant distribution for the hidden process  $X$  associated with  $Q_{\theta^*}$ . Define the stationary score  $\nabla_{\theta} \ell_{n,\pi_{\theta^*}}(\theta)$  as:

$$\nabla_{\theta} \ell_{n,\pi_{\theta^*}}(\theta) := \int_{\mathcal{X}} \nabla_{\theta} \ell_{n,x}(\theta) \pi_{\theta^*}(dx).$$

First, for all  $x, x' \in \mathcal{X}$  and  $\theta \in \mathcal{O}$ , write:

$$\nabla_{\theta} \ell_{n,x}(\theta) - \nabla_{\theta} \ell_{n,x'}(\theta) = \sum_{u \in T_n^*} \Phi(\theta; u, x, x'),$$

where:

$$\Phi(\theta; u, x, x') = \mathbb{E}_{\theta}[\phi_{\theta}(X_{p(u)}, X_u, Y_u) | Y_{T_n}, X_{\partial} = x] - \mathbb{E}_{\theta}[\phi_{\theta}(X_{p(u)}, X_u, Y_u) | Y_{T_n}, X_{\partial} = x'].$$

Using Minkowski's inequality and the upper bound (50) from the proof of Lemma 4.2, we get:

$$\begin{aligned} & \left( \mathbb{E}_{\theta^*} \left[ \sup_{x, x' \in \mathcal{X}} \frac{1}{|T_n|} \|\nabla_{\theta} \ell_{n,x}(\theta) - \nabla_{\theta} \ell_{n,x'}(\theta)\|^2 \right] \right)^{1/2} \\ & \leq \frac{1}{|T_n|^{1/2}} \sum_{u \in T_n^*} \left( \mathbb{E}_{\theta^*} \left[ \sup_{x, x' \in \mathcal{X}} \|\Phi(\theta; u, x, x')\|^2 \right] \right)^{1/2} \\ & \leq 2 \left( \mathbb{E}_{\theta^*} \left[ \sup_{\theta \in \Theta_0} \sup_{x, x' \in \mathcal{X}} \|\phi_{\theta}(x, x', Y_{\partial})\|^2 \right] \right)^{1/2} \frac{1}{|T_n|^{1/2}} \sum_{k=1}^n 2^k \rho^{k-1} \\ & \leq C \max(n2^{-n}, (2\rho^2)^{n/2}), \end{aligned}$$

where  $C < \infty$  is some finite constant. Thus (remind that  $\rho < 1/\sqrt{2}$ ), for any  $x \in \mathcal{X}$ , we have:

$$\lim_{n \rightarrow \infty} \frac{1}{|T_n|^{1/2}} \left( \nabla_{\theta} \ell_{n,x}(\theta^*) - \nabla_{\theta} \ell_{n,\pi_{\theta^*}}(\theta^*) \right) = 0 \quad \text{in } L^2(\mathbb{P}_{\theta^*}). \quad (55)$$

In particular, to prove asymptotic normality for the score  $\nabla_{\theta} \ell_{n,x}(\theta^*)$  for any  $x \in \mathcal{X}$ , it is enough to prove asymptotic normality for the stationary score  $\nabla_{\theta} \ell_{n,\pi_{\theta^*}}(\theta^*)$  (see for instance [Bil99, Theorem 3.1]).

For any  $u \in T$  and  $k \in \mathbb{N}$  and  $\theta \in \mathcal{O}$ , define:

$$\dot{h}_{u,k,\pi_{\theta^*}}(\theta) := \int_{\mathcal{X}} \dot{h}_{u,k,x}(\theta) \pi_{\theta^*}(dx). \quad (56)$$

In particular, note that, as the bound in Lemma 4.2 is uniform in  $x \in \mathcal{X}$ , this bound still holds with  $\dot{h}_{u,k,x}(\theta)$  replaced by  $\dot{h}_{u,k,\pi_{\theta^*}}(\theta)$ . Using (47), note that we have:

$$\nabla_{\theta} \ell_{n,\pi_{\theta^*}}(\theta) = \sum_{u \in T_n} \dot{h}_{u,h(u),\pi_{\theta^*}}(\theta).$$

Moreover, remark that for  $\theta = \theta^*$  and for any  $u \in T$  and  $k \in \mathbb{N}^*$ , we have:

$$\begin{aligned} \dot{h}_{u,k,\pi_{\theta^*}}(\theta^*) &= \mathbb{E}_{\theta^*}[\phi_{\theta^*}(X_{p(u)}, X_u, Y_u) | Y_{\Delta(u,k)}] \\ &+ \sum_{v \in \Delta^*(u,k) \setminus \{p^k(u)\}} \left( \mathbb{E}_{\theta^*}[\phi_{\theta^*}(X_{p(v)}, X_v, Y_v) | Y_{\Delta(u,k)}] \right. \\ &\quad \left. - \mathbb{E}_{\theta^*}[\phi_{\theta^*}(X_{p(v)}, X_v, Y_v) | Y_{\Delta^*(u,k)}] \right). \end{aligned} \quad (57)$$

**Step 2: The stationary score is a sum of martingale increments.**

As  $T$  is a plane rooted tree, we can enumerate its vertices in a breadth-first-search manner, that is, as a sequence  $(u_j)_{j \in \mathbb{N}}$  which is increasing for  $<$ . (Note that  $u_0 = \delta$ .) Remind that  $\Delta(u_{j-1}) = \Delta^*(u_j)$  for all  $j \geq 1$ . Define the filtration  $\mathcal{F}$  by  $\mathcal{F}_j = \sigma(Y_v : v \in T, v \leq u_j) = \sigma(Y_{\Delta(u_j)})$  for all  $j \in \mathbb{N}$ , and note that  $\mathcal{F}_j \subset \sigma(Y_T)$ . Let  $j \in \mathbb{N}^*$ ,  $1 \leq k \leq h(u_j)$ ,  $x \in \mathcal{X}$  and  $v \in Y_{\Delta^*(u_j,k)}$ . Note that we have:

$$\mathbb{E}_{\theta^*} \left[ \mathbb{E}_{\theta^*}[\phi_{\theta^*}(X_{p(v)}, X_v, Y_v) | Y_{\Delta(u_j,k)}] \mid \mathcal{F}_{j-1} \right] = \mathbb{E}_{\theta^*}[\phi_{\theta^*}(X_{p(v)}, X_v, Y_v) | Y_{\Delta^*(u_j,k)}].$$

Also note that Assumption 7 (on page 27) implies that:

$$\begin{aligned} &\mathbb{E}_{\theta^*}[\phi_{\theta^*}(X_{p(u_j)}, X_{u_j}, Y_{u_j}) | X_{p(u_j)}] \\ &= \int_{\mathcal{X} \times \mathcal{Y}} \nabla_{\theta} \log[q_{\theta}(X_{p(u_j)}, x)g_{\theta}(x, y)] q_{\theta}(X_{p(u_j)}, x)g_{\theta}(x, y) \lambda(dx)\mu(dy) \\ &= \int_{\mathcal{X} \times \mathcal{Y}} \nabla_{\theta}[q_{\theta}(X_{p(u_j)}, x)g_{\theta}(x, y)] \lambda(dx)\mu(dy) \\ &= \nabla_{\theta} \left[ \int_{\mathcal{X} \times \mathcal{Y}} q_{\theta}(X_{p(u_j)}, x)g_{\theta}(x, y) \lambda(dx)\mu(dy) \right] \\ &= 0. \end{aligned}$$

Thus, we have:

$$\begin{aligned} &\mathbb{E}_{\theta^*} \left[ \mathbb{E}_{\theta^*}[\phi_{\theta^*}(X_{p(u_j)}, X_{u_j}, Y_{u_j}) | Y_{\Delta(u_j,k)}] \mid \mathcal{F}_{j-1} \right] \\ &= \mathbb{E}_{\theta^*}[\phi_{\theta^*}(X_{p(u_j)}, X_{u_j}, Y_{u_j}) | Y_{\Delta^*(u_j,k)}] \\ &= \mathbb{E}_{\theta^*} \left[ \mathbb{E}_{\theta^*}[\phi_{\theta^*}(X_{p(u_j)}, X_{u_j}, Y_{u_j}) | Y_{\Delta^*(u_j,k)}, X_{p(u_j)}] \mid Y_{\Delta^*(u_j,k)} \right] \\ &= \mathbb{E}_{\theta^*} \left[ \mathbb{E}_{\theta^*}[\phi_{\theta^*}(X_{p(u_j)}, X_{u_j}, Y_{u_j}) | X_{p(u_j)}] \mid Y_{\Delta^*(u_j,k)} \right] \\ &= 0, \end{aligned}$$

where we used the Markov property for the inner expectation in the third equality. Moreover, it is immediate that  $\dot{h}_{u_j, k, \pi_{\theta^*}}(\theta^*)$  is  $\mathcal{F}_j$ -measurable for all  $j \in \mathbb{N}^*$  and  $1 \leq k \leq h(j)$ . Hence, we get that the sequence  $(\dot{h}_{u_j, h(u_j), \pi_{\theta^*}}(\theta^*))_{j \in \mathbb{N}^*}$  is a  $\mathbb{P}_{\theta^*}$ -martingale increment sequence adapted to the filtration  $\mathcal{F} = (\mathcal{F}_j)_{j \in \mathbb{N}}$  in  $L^2(\mathbb{P}_{\theta^*})$  (thanks to Assumption 7). We are going to apply a central limit theorem for martingales (see [Duf11, Corollary 2.1.10]). For all  $n \in \mathbb{N}$ , define  $M_n = \sum_{j=0}^n \dot{h}_{u_j, h(u_j), \pi_{\theta^*}}(\theta^*)$ . Note that  $M_0 = \dot{h}_{\partial, 0, \pi_{\theta^*}}(\theta^*) = \int_{\mathcal{X}} \nabla_{\theta} \log g_{\theta^*}(x, Y_{\partial}) \pi_{\theta^*}(dx)$  is in  $L^2(\mathbb{P}_{\theta^*})$  by Assumption 7. Hence, the sequence  $(M_n)_{n \in \mathbb{N}}$  is a  $\mathbb{P}_{\theta^*}$ -martingale sequence adapted to the filtration  $\mathcal{F} = (\mathcal{F}_j)_{j \in \mathbb{N}}$  in  $L^2(\mathbb{P}_{\theta^*})$ , and whose quadratic variation is:

$$\langle M \rangle_n = \sum_{j=1}^n \mathbb{E}_{\theta^*} \left[ \dot{h}_{u_j, h(u_j), \pi_{\theta^*}}(\theta^*) \dot{h}_{u_j, h(u_j), \pi_{\theta^*}}(\theta^*)^t \mid \mathcal{F}_{j-1} \right],$$

where, as in (54), we see  $\dot{h}_{u_j, h(u_j), \pi_{\theta^*}}(\theta^*)$  as a column vector. Note that for all  $n \in \mathbb{N}$ ,  $M_n$  and  $\langle M \rangle_n$  do not depend on  $\mathcal{U}$ .

**Step 3: Convergence of the quadratic variation.** Before applying the central limit theorem for martingales, we first need to prove that  $\lim_{n \rightarrow \infty} n^{-1} \langle M \rangle_n = \mathcal{I}(\theta^*)$  in  $\mathbb{P}_{\theta^*}$ -probability. Indeed, we will prove that this convergence holds in  $L^2(\mathbb{P}_{\theta^*})$ . Let  $k \in \mathbb{N}^*$  and  $x \in \mathcal{X}$ . Note that for  $u_j \in T \setminus T_{k-1}$  is equivalent to  $j \geq |T_{k-1}|$  (remind that  $u_0 = \partial$ ). Using (48) along with Assumption 8, we get that  $\sup_{x \in \mathcal{X}} \dot{h}_{u, k, x}(\theta^*, Y_{\Delta(u, k)})$  is in  $L^4(\mathbb{P}_{\theta^*})$ , and thus the random variable  $\sup_{x \in \mathcal{X}} \dot{h}_{u, k, x}(\theta^*, Y_{\Delta(u, k)}) \dot{h}_{u, k, x}(\theta^*, Y_{\Delta(u, k)})^t$  is in  $L^2(\mathbb{P}_{\theta^*})$  for every  $u \in T \setminus T_{k-1}$ . Thus, using (56) and Lemma 4.2 (remind that  $\rho < 1/\sqrt{2}$ ) for the first moment ( $b = 2$ ), there exists a finite constant  $C > 0$  and  $\alpha \in (0, 1)$  such that we have (remind (49)):

$$\mathbb{E}_{\theta^*} \left\| \left\| n^{-1} \langle M \rangle_n - \frac{1}{n} \sum_{j=|T_{k-1}|}^n \mathbb{E}_{\theta^*} \left[ \dot{h}_{u_j, k, x}(\theta^*) \dot{h}_{u_j, k, x}(\theta^*)^t \mid \mathcal{F}_{j-1} \right] \right\| \right\| \leq C \alpha^k + \frac{C |T_{k-1}|}{n}, \quad (58)$$

where remind that  $\|\cdot\|$  denotes the euclidean norm for  $d \times d$  matrices (or any other norm as they are all equivalent in finite dimension). To prove that the second term inside the expectation in the left hand side of (58) converges in  $L^2(\mathbb{P}_{\theta^*})$  as  $n \rightarrow \infty$ , we are going to apply the ergodic convergence Lemma 2.12 where the averages are done on the vertex subset  $\{u_j : |T_{k-1}| \leq j \leq n\}$ . Note that this lemma is stated for scalar-valued functions, but we can apply it individually for each of the matrix coefficients to get the equivalent for matrix-valued functions.

For all  $u \in T \setminus T_{k-1}$ , define the function:

$$\Psi_{u, k, x} : y_{\Delta^*(u, k)} \in \mathcal{Y}^{\Delta^*(u, k)} \mapsto \mathbb{E}_{\theta^*} \left[ \dot{h}_{u, k, x}(\theta^*; Y_{\Delta(u, k)}) \dot{h}_{u, k, x}(\theta^*; Y_{\Delta(u, k)})^t \mid Y_{\Delta^*(u, k)} = y_{\Delta^*(u, k)} \right].$$

For a vertex  $u$  in  $T \setminus T_{k-1}$ , let  $v_u \in G_k$  be the unique vertex that satisfies the shape equality constraint (8) (on page 12), then we have the equality between functions:

$$\Psi_{u, k, x} = \Psi_{v_u, k, x}. \quad (59)$$

Moreover, using (48) along with Assumption 8, we get that  $\dot{h}_{u, k, x}(\theta^*, Y_{\Delta(u, k)})$  is in  $L^4(\mathbb{P}_{\theta^*})$ , and thus the random variable  $\Psi_{u, k, x}(Y_{\Delta^*(u, k)})$  is in  $L^2(\mathbb{P}_{\theta^*})$  for every  $u \in T \setminus T_{k-1}$ . Hence, applying Lemma 2.12 to the collection of neighborhood-shape-dependent functions  $(\Psi_{v, k, x})_{v \in G_k}$  (remind that indexing functions with  $G_k$  or with  $\mathcal{N}_k$  is equivalent by (9)), and using (59) and (14) (in Remark 3.1), we get that the second term inside the expectation in the left hand side of (58) converges in  $L^2(\mathbb{P}_{\theta^*})$  to  $\mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [\dot{h}_{\partial, k, x}(\theta^*) \dot{h}_{\partial, k, x}(\theta^*)^t]$  as  $n \rightarrow \infty$ . Using Lemma 4.2, we have that  $\lim_{k \rightarrow \infty} \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [\dot{h}_{\partial, k, x}(\theta^*) \dot{h}_{\partial, k, x}(\theta^*)^t] = \mathcal{I}(\theta^*)$ . Combining those facts with (58), we get that  $\lim_{n \rightarrow \infty} n^{-1} \langle M \rangle_n = \mathcal{I}(\theta^*)$  in  $L^2(\mathbb{P}_{\theta^*})$ .

**Step 4: Lindeberg's condition holds.** We now need to verify that Lindeberg's condition holds (see [Duf11, Corollary 2.1.10]), that is, to prove for all  $\varepsilon > 0$  that  $\lim_{n \rightarrow \infty} F_n(\varepsilon\sqrt{n}) = 0$  in  $\mathbb{P}_{\theta^*}$ -probability where for all  $n \in \mathbb{N}^*$  and  $A \in \mathbb{R}_+$ :

$$F_n(A) = \frac{1}{n} \sum_{j=1}^n \mathbb{E}_{\theta^*} \left[ \|\dot{h}_{u_j, h(u_j), \pi_{\theta^*}}(\theta^*)\|^2 \mathbf{1}_{\{\|\dot{h}_{u_j, h(u_j), \pi_{\theta^*}}(\theta^*)\| \geq A\}} \mid \mathcal{F}_{j-1} \right]. \quad (60)$$

Remind that by Assumption 8 and Lemma 4.2 (remind that  $\rho < 1/\sqrt{2}$ ) for the fourth moment ( $b = 4$ ), we have:

$$C := \sup_{u \in T} \sup_{k \in \mathbb{N}^*} \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} \left[ \sup_{x \in \mathcal{X}} \|\dot{h}_{u, k, x}(\theta^*)\|^4 \right] < \infty.$$

Using Cauchy-Schwarz inequality and Markov inequality, we get:

$$\mathbb{E}_{\theta^*} [F_n(A)] \leq \frac{1}{n} \sum_{j=1}^n \frac{\mathbb{E}_{\theta^*} \left[ \sup_{x \in \mathcal{X}} \|\dot{h}_{u_j, h(u_j), x}(\theta^*)\|^4 \right]}{A^2} \leq \frac{C}{A^2}.$$

Let  $\varepsilon > 0$ . Then, setting  $A_n = \varepsilon\sqrt{n}$  for all  $n \in \mathbb{N}^*$ , we get that  $\lim_{n \rightarrow \infty} F_n(\varepsilon\sqrt{n}) = 0$  in  $L^1(\mathbb{P}_{\theta^*})$ , and thus in  $\mathbb{P}_{\theta^*}$ -probability. Hence, we get that Lindeberg's condition holds.

**Step 5: Applying the central limit theorem for martingales.** Hence, we can apply the central limit theorem for martingales (see [Duf11, Corollary 2.1.10]), which gives us that  $\mathbb{P}_{\theta^*}$ -a.s.  $\lim_{n \rightarrow \infty} n^{-1}M_n = 0$  and that the sequence  $(n^{-1/2}M_n)_{n \in \mathbb{N}^*}$  converges in  $\mathbb{P}_{\theta^*}$ -distribution towards a centered Gaussian distribution  $\mathcal{N}(0, \mathcal{I}(\theta^*))$  whose covariance matrix is  $\mathcal{I}(\theta^*)$ . In particular, using (55), we get that  $\mathbb{P}_{\theta^*}$ -a.s.  $\lim_{n \rightarrow \infty} |T_n|^{-1} \nabla_{\theta} \ell_{n, x}(\theta^*) = 0$  and that:

$$|T_n|^{-1/2} \nabla_{\theta} \ell_{n, x}(\theta^*) \xrightarrow[n \rightarrow \infty]{(d)} \mathcal{N}(0, \mathcal{I}(\theta^*)) \quad \text{under } \mathbb{P}_{\theta^*}.$$

This concludes the proof of the theorem.  $\square$

## 4.2 Law of large number for the normalized observed information

In this subsection, we prove that for all possibly random sequence  $(\theta_n)_{n \in \mathbb{N}}$  such that  $\lim_{n \rightarrow \infty} \theta_n = \theta^*$   $\mathbb{P}_{\theta^*}$ -a.s., then the normalized observed information  $-n^{-1} \nabla_{\theta}^2 \ell_{n, x}(\theta_n)$  converges  $\mathbb{P}_{\theta^*}$ -a.s. as  $n \rightarrow \infty$  to the limiting Fisher information matrix  $\mathcal{I}(\theta^*)$  which is defined in (54).

Remind the definition of the log-likelihood  $\ell_{n, x}(\theta)$  in (7) on page 11. We start by decomposing the Hessian of the log-likelihood  $\ell_{n, x}(\theta)$  as a sum of increment indexed by the tree  $T$ . Using elementary computation along with permutations of the integral and the gradient operator which are valid under Assumption 7 (note that this result is also known as *Louis missing information principle*, see [CMR05, Proposition 10.1.6]), we get for all  $\theta \in \mathcal{O}$  and  $x \in \mathcal{X}$ :

$$\begin{aligned} \nabla_{\theta}^2 \ell_{n, x}(\theta) &= \nabla_{\theta}^2 \log(g_{\theta}(X_{\partial}, Y_{\partial})) + \mathbb{E}_{\theta} \left[ \sum_{u \in T_n^*} \varphi_{\theta}(X_{\mathcal{P}(u)}, X_u, Y_u) \mid Y_{T_n}, X_{\partial} = x \right] \\ &\quad + \text{Var}_{\theta} \left[ \sum_{u \in T_n^*} \phi_{\theta}(X_{\mathcal{P}(u)}, X_u, Y_u) \mid Y_{T_n}, X_{\partial} = x \right], \end{aligned}$$

where remind that  $\phi_{\theta}$  is defined in (46) on page 28, and  $\varphi_{\theta}$  is defined as:

$$\varphi_{\theta}(x', x, y) = \nabla_{\theta}^2 \log(q_{\theta}(x', x)g_{\theta}(x, y)). \quad (61)$$

Note that similarly to the case of  $\phi_\theta$ , the random variable  $\varphi_\theta(X_{\mathbb{P}(u)}, X_u, Y_u)$  is integrable conditionally on  $Y_{\Delta(u)}$  and  $X_\partial = x$  (see the discussion after (46)). Also note that  $\nabla_\theta \log g_\theta(x, Y_\partial)$  is  $\mathbb{P}_{\theta^*}$ -a.s. finite by Assumption 7-(iii).

For all  $u \in T$ ,  $k \in \mathbb{N}^*$  and  $x \in \mathcal{X}$ , we define:

$$\begin{aligned} \Lambda_{u,k,x}(\theta) = \mathbb{E}_\theta \left[ \sum_{v \in \Delta(u,k) \setminus \{\mathbb{P}^k(u)\}} \varphi_\theta(X_{\mathbb{P}(v)}, X_v, Y_v) \middle| Y_{\Delta(u,k)}, X_\partial = x_\partial \right] \\ - \mathbb{E}_\theta \left[ \sum_{v \in \Delta^*(u,k) \setminus \{\mathbb{P}^k(u)\}} \varphi_\theta(X_{\mathbb{P}(v)}, X_v, Y_v) \middle| Y_{\Delta^*(u,k)}, X_\partial = x_\partial \right], \end{aligned} \quad (62)$$

and:

$$\begin{aligned} \Gamma_{u,k,x}(\theta) = \text{Var}_\theta \left[ \sum_{v \in \Delta(u,k) \setminus \{\mathbb{P}^k(u)\}} \phi_\theta(X_{\mathbb{P}(v)}, X_v, Y_v) \middle| Y_{\Delta(u,k)}, X_\partial = x_\partial \right] \\ - \text{Var}_\theta \left[ \sum_{v \in \Delta^*(u,k) \setminus \{\mathbb{P}^k(u)\}} \phi_\theta(X_{\mathbb{P}(v)}, X_v, Y_v) \middle| Y_{\Delta^*(u,k)}, X_\partial = x_\partial \right], \end{aligned} \quad (63)$$

where  $\text{Var}_\theta$  (resp.  $\text{Cov}_\theta$ ) denotes the (possibly conditional) variance (resp. covariance) corresponding to  $\mathbb{P}_\mathcal{U} \otimes \mathbb{P}_\theta$ . Note that  $\Lambda_{u,k,x}(\theta)$  and  $\Gamma_{u,k,x}(\theta)$  are random variables which depend on  $Y_{\Delta(u,k)}$  with an implicit dependence on  $\mathcal{U}$ , and that they do not depend on  $\mathcal{U}$  if  $k \leq h(u)$ .

Then, using telescopic sums involving the quantities defined in (62) and (63), the Hessian of the log-likelihood  $\ell_{n,x}(\theta)$  can be rewritten for all  $\theta \in \mathcal{O}$  and  $x \in \mathcal{X}$  as:

$$\nabla_\theta^2 \ell_{n,x}(\theta) = \nabla_\theta^2 \log(g_\theta(X_\partial, Y_\partial)) + \sum_{u \in T_n^*} \Lambda_{u,h(u),x}(\theta) + \sum_{u \in T_n^*} \Gamma_{u,h(u),x}(\theta). \quad (64)$$

To prove the convergence of the two sums in the right hand side of (64), and thus get the convergence of the normalized observed information  $-n^{-1} \nabla_\theta^2 \ell_{n,x}(\theta)$ , we will need the following  $L^2$  regularity assumption on the Hessian of the transition kernel  $g_\theta$  of the HMT.

**Assumption 9** ( $L^2$  Hessian regularity). *In addition to Assumption 7, assume that we have:*

$$\mathbb{E}_{\theta^*} \left[ \sup_{\theta \in \mathcal{O}} \sup_x \|\nabla_\theta^2 \log g_\theta(x, Y_\partial)\|^2 \right] < \infty.$$

Propositions 4.4 and 4.5 below (whose proofs are given in Sections 4.2.1 and 4.2.2, respectively) state that  $\Lambda_{u,k,x}(\theta)$  and  $\Gamma_{u,k,x}(\theta)$  both have limits  $\mathbb{P}_{\theta^*}$ -a.s. and in  $L^2(\mathbb{P}_{\theta^*})$  when  $k \rightarrow \infty$ . Denote those limits by  $\Lambda_{u,\infty}(\theta)$  and  $\Gamma_{u,\infty}(\theta)$ , respectively. Furthermore, Propositions 4.4 and 4.5 also state that the two sums in the right hand side of (64) converge to  $\mathbb{E}_\mathcal{U} \otimes \mathbb{E}_{\theta^*}[\Lambda_{\partial,\infty}(\theta)]$  and  $\mathbb{E}_\mathcal{U} \otimes \mathbb{E}_{\theta^*}[\Gamma_{\partial,\infty}(\theta)]$ , respectively, with some uniformity in  $\theta$  near  $\theta^*$ .

We start with the proposition for the terms  $\Lambda_{u,k,x}(\theta)$ . Note that the condition  $\rho < 1/\sqrt{2}$  on the mixing rate  $\rho$  of the HMT process  $(X, Y)$  is due to the use of Lemma 4.2 in the proof of Proposition 4.4. See the discussion in Remark 1.5 for comments on this condition on  $\rho$ .

**Proposition 4.4** (Convergence for averages of  $\Lambda_{u,k,x}(\theta)$ ). *Assume that Assumptions 1-4, 6-7 and 9 hold. Assume that  $\rho < 1/\sqrt{2}$ . Then, for each  $\theta \in \mathcal{O}$ , we have that  $\Lambda_{u,k,x}(\theta)$  converges in*

$L^2(\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^*})$  to some limit  $\Lambda_{u,\infty}(\theta)$  (that does not depend on  $x$ ) as  $k \rightarrow \infty$ . Moreover, we have:

$$\lim_{n \rightarrow \infty} \mathbb{E}_{\theta^*} \left[ \sup_{x \in \mathcal{X}} \left| \frac{1}{|T_n|} \sum_{u \in T_n^*} \Lambda_{u,h(u),x}(\theta^*) - \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [\Lambda_{\partial,\infty}(\theta^*)] \right| \right] = 0. \quad (65)$$

Furthermore, the function  $\theta \mapsto \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [\Lambda_{\partial,\infty}(\theta)]$  is continuous on  $\mathcal{O}$ , and for all  $x \in \mathcal{X}$  and  $\theta \in \mathcal{O}$ , we have:

$$\lim_{\delta \rightarrow 0} \lim_{n \rightarrow \infty} \sup_{\theta' \in \mathcal{O}: \|\theta' - \theta\| \leq \delta} \left| |T_n|^{-1} \sum_{u \in T_n^*} \Lambda_{u,h(u),x}(\theta') - \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [\Lambda_{\partial,\infty}(\theta)] \right| = 0, \quad \mathbb{P}_{\theta^*}\text{-a.s.}$$

The following proposition is the equivalent of Proposition 4.4 for the terms  $\Gamma_{u,k,x}(\theta)$ . Note that the condition  $\rho < 1/2$  on the mixing rate  $\rho$  of the HMT process  $(X, Y)$  is due to the use of Lemma 4.17 in the proof of Proposition 4.5. See the discussion in Remark 1.5 for comments on this condition on  $\rho$ .

**Proposition 4.5** (Convergence for the averages of  $\Gamma_{u,k,x}(\theta)$ ). *Assume that Assumptions 1–4 and 6–8 hold. Assume that  $\rho < 1/2$ . Then, for each  $\theta \in \mathcal{O}$ , we have that  $\Gamma_{u,k,x}(\theta)$  converges in  $L^2(\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^*})$  to some limit  $\Gamma_{u,\infty}(\theta)$  (that does not depend on  $x$ ) as  $k \rightarrow \infty$ . Moreover, we have:*

$$\lim_{n \rightarrow \infty} \mathbb{E}_{\theta^*} \left[ \sup_{x \in \mathcal{X}} \left| \frac{1}{|T_n|} \sum_{u \in T_n^*} \Gamma_{u,h(u),x}(\theta^*) - \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [\Gamma_{\partial,\infty}(\theta^*)] \right| \right] = 0. \quad (66)$$

Furthermore, the function  $\theta \mapsto \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [\Gamma_{\partial,\infty}(\theta)]$  is continuous on  $\mathcal{O}$ , and for all  $x \in \mathcal{X}$  and  $\theta \in \mathcal{O}$ , we have:

$$\lim_{\delta \rightarrow 0} \lim_{n \rightarrow \infty} \sup_{\theta' \in \mathcal{O}: \|\theta' - \theta\| \leq \delta} \left| |T_n|^{-1} \sum_{u \in T_n^*} \Gamma_{u,h(u),x}(\theta') - \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [\Gamma_{\partial,\infty}(\theta)] \right| = 0, \quad \mathbb{P}_{\theta^*}\text{-a.s.}$$

With Propositions 4.4 and 4.5, we are now ready to prove the following theorem which states that the normalized observed information  $-|T_n|^{-1} \nabla_{\theta}^2 \ell_{n,x}(\theta_n)$  converges  $\mathbb{P}_{\theta^*}$ -a.s. locally uniformly to the limiting Fisher information  $\mathcal{I}(\theta^*)$  (which is defined in (54)). Note that the condition  $\rho < 1/2$  on the mixing rate  $\rho$  of the HMT process  $(X, Y)$  is inherited from Proposition 4.5. See the discussion in Remark 1.5 for comments on this condition on  $\rho$ .

**Theorem 4.6** (Convergence of the normalized observed information). *Assume that Assumptions 1–4 and 6–9 hold. Assume that  $\rho < 1/2$ . Assume that  $\Theta$  is compact. Then, for all  $x \in \mathcal{X}$ , we have:*

$$\lim_{\delta \rightarrow 0} \lim_{n \rightarrow \infty} \sup_{\theta \in \mathcal{O}: \|\theta - \theta^*\| \leq \delta} \left\| -|T_n|^{-1} \nabla_{\theta}^2 \ell_{n,x}(\theta) - \mathcal{I}(\theta^*) \right\| = 0 \quad \mathbb{P}_{\theta^*}\text{-a.s.} \quad (67)$$

As an immediate corollary, for any possibly random sequence  $(\theta_n)_{n \in \mathbb{N}}$  such that  $\lim_{n \rightarrow \infty} \theta_n = \theta^*$   $\mathbb{P}_{\theta^*}$ -a.s. and any  $x \in \mathcal{X}$ , we get that  $\mathbb{P}_{\theta^*}$ -a.s.  $\lim_{n \rightarrow \infty} -|T_n|^{-1} \nabla_{\theta}^2 \ell_{n,x}(\theta_n) = \mathcal{I}(\theta^*)$ . In particular, choosing  $\theta_n = \hat{\theta}_{n,x}$  for all  $n \in \mathbb{N}$  (remind that the MLE  $\hat{\theta}_{n,x}$  is defined in (33) on page 22), and combining Theorems 3.11 and 4.6, we get that the normalized observed information  $-|T_n|^{-1} \nabla_{\theta}^2 \ell_{n,x}(\hat{\theta}_{n,x})$  at the MLE  $\hat{\theta}_{n,x}$  is a strongly consistent estimator of the Fisher information matrix  $\mathcal{I}(\theta^*)$ .

*Proof.* Using (64) and Propositions 4.4 and 4.5, we get that (67) holds with  $\mathcal{I}(\theta^*)$  replaced by  $-\mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [\Lambda_{\partial, \infty}(\theta^*) + \Gamma_{\partial, \infty}(\theta^*)]$ . Thus, it remains to prove that this latter quantity is equal to  $\mathcal{I}(\theta^*)$ .

Using elementary computation along with permutations of the integral and the gradient operator which are valid under Assumption 7 (note that this result is also known as *Fisher information matrix identity*, see [RS18, p.21] or [CMR05, p.355]), we get for all  $\theta \in \mathcal{O}$  and  $x \in \mathcal{X}$ :

$$|T_n|^{-1} \mathbb{E}_{\theta} [\nabla_{\theta} \ell_{n,x}(\theta) \nabla_{\theta} \ell_{n,x}(\theta)^t \mid X_{\partial} = x] = -|T_n|^{-1} \mathbb{E}_{\theta} [\nabla_{\theta}^2 \ell_{n,x}(\theta) \mid X_{\partial} = x].$$

Setting  $\theta = \theta^*$  and taking the expectation over  $X_{\partial}$ , we get:

$$|T_n|^{-1} \mathbb{E}_{\theta^*} [\nabla_{\theta} \ell_{n,X_{\partial}}(\theta^*) \nabla_{\theta} \ell_{n,X_{\partial}}(\theta^*)^t] = -|T_n|^{-1} \mathbb{E}_{\theta^*} [\nabla_{\theta}^2 \ell_{n,X_{\partial}}(\theta^*)]. \quad (68)$$

Using (64) on page 37, Propositions 4.4 and 4.5 give us that the right hand side of (68) converges as  $n \rightarrow \infty$  to  $-\mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [\Lambda_{\partial, \infty}(\theta^*) + \Gamma_{\partial, \infty}(\theta^*)]$ .

Remind that using (48) along with Assumption 8, we get that  $\dot{h}_{u,k,x}(\theta^*, Y_{\Delta(u,k)})$  is in  $L^4(\mathbb{P}_{\theta^*})$ , and thus the random variable  $\dot{h}_{u,k,x}(\theta^*, Y_{\Delta(u,k)}) \dot{h}_{u,k,x}(\theta^*, Y_{\Delta(u,k)})^t$  is in  $L^2(\mathbb{P}_{\theta^*})$  for every  $u \in T \setminus T_{k-1}$ . Thus, using Lemma 4.2 for the first moment ( $b = 1$ ), there exists a finite constant  $C > 0$  and  $\alpha \in (0, 1)$  such that for any  $k \in \mathbb{N}^*$  and  $x \in \mathcal{X}$ , we have:

$$\mathbb{E}_{\theta^*} \left[ \frac{1}{|T_n|} \left\| \nabla_{\theta} \ell_{n,X_{\partial}}(\theta^*) \nabla_{\theta} \ell_{n,X_{\partial}}(\theta^*)^t - \sum_{u \in T_n \setminus T_{k-1}} \dot{h}_{u,k,x}(\theta^*) \dot{h}_{u,k,x}(\theta^*)^t \right\| \right] \leq C \alpha^k + \frac{C |T_{k-1}|}{|T_n|}, \quad (69)$$

where remind that we see  $\dot{h}_{u,k,x}(\theta^*)$  as a column vector. Then, using an ergodic convergence argument similar to the one used in Step 3 in the proof of Theorem 4.3, we get:

$$\lim_{n \rightarrow \infty} \frac{1}{|T_n|} \sum_{u \in T_n \setminus T_{k-1}} \dot{h}_{u,k,x}(\theta^*) \dot{h}_{u,k,x}(\theta^*)^t = \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [\dot{h}_{\partial,k,x}(\theta^*) \dot{h}_{\partial,k,x}(\theta^*)^t] \quad \text{in } L^2(\mathbb{P}_{\theta^*}).$$

Using Lemma 4.2, we have that  $\lim_{k \rightarrow \infty} \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [\dot{h}_{\partial,k,x}(\theta^*) \dot{h}_{\partial,k,x}(\theta^*)^t] = \mathcal{I}(\theta^*)$ . Combining those facts with (69), we get that the left hand side in (68) converges to  $\mathcal{I}(\theta^*)$  as  $n \rightarrow \infty$ .

Hence, we get  $\mathcal{I}(\theta^*) = -\mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [\Lambda_{\partial, \infty}(\theta^*) + \Gamma_{\partial, \infty}(\theta^*)]$ , which concludes the proof.  $\square$

Using Theorems 3.11, 4.3 and 4.6, we can prove the following theorem which states that the MLE has asymptotic normal fluctuations with covariance matrix  $\mathcal{I}(\theta^*)^{-1}$  where the Fisher information matrix  $\mathcal{I}(\theta^*)$  is defined in (54) on page 32. Recall that the contrast function  $\ell$  is defined in (26) on page 19, that the MLE  $\hat{\theta}_{n,x}$  is defined in (33) on page 22, and that the mixing rate  $\rho$  of the HMT process  $(X, Y)$  is defined after Assumption 3 on page 10.

Note that the condition  $\rho < 1/2$  on the mixing rate  $\rho$  of the HMT process  $(X, Y)$  is inherited from Theorem 4.6, and thus from Proposition 4.5. See the discussion in Remark 1.5 for comments on this condition on  $\rho$ .

**Theorem 4.7** (Asymptotic normality of the MLE). *Assume that Assumptions 1–9 hold. Assume that  $\rho < 1/2$ . Further assume that the contrast function  $\ell$  has a unique maximum (which is then located at  $\theta^* \in \Theta$  by Proposition 3.10) and that  $\Theta$  is compact,  $\theta^*$  is an interior point of  $\Theta$ , and the limiting Fisher information matrix  $\mathcal{I}(\theta^*)$  (which is defined in (54)) is non-singular. Then, for all  $x \in \mathcal{X}$ , we have:*

$$|T_n|^{1/2} (\hat{\theta}_{n,x} - \theta^*) \xrightarrow[n \rightarrow \infty]{(d)} \mathcal{N}(0, \mathcal{I}(\theta^*)^{-1}) \quad \text{under } \mathbb{P}_{\theta^*},$$

where  $\mathcal{N}(0, M)$  denotes the centered Gaussian distribution with covariance matrix  $M$ .



*Proof.* The proof is a standard argument and is similar to the proof of [BRR98, Theorem 1]. Remind that the gradient of  $\ell_{n,x}$  vanishes at the MLE  $\hat{\theta}_{n,x}$  by definition. Thus, using a Taylor expansion for  $\nabla_{\theta}\ell_{n,x}$  around  $\theta^*$ , we get:

$$0 = \nabla_{\theta}\ell_{n,x}(\hat{\theta}_{n,x}) = \nabla_{\theta}\ell_{n,x}(\theta^*) + \left( \int_0^1 \nabla_{\theta}^2\ell_{n,x}(\theta^* + t(\hat{\theta}_{n,x} - \theta^*)) dt \right) (\hat{\theta}_{n,x} - \theta^*),$$

where we see  $\hat{\theta}_{n,x}$  and  $\theta^*$  as column vectors. As  $\mathcal{I}(\theta^*)$  is non-singular (indeed definite positive), remark that Theorems 3.11 and 4.6 insure that  $\mathbb{P}_{\theta^*}$ -a.s. for  $n$  large enough the integrand in the integral of the above formula is non-singular (indeed definite positive) for all values of  $t$ , and thus the matrix-valued integral is non-singular. Thus, from the above equation, we obtain  $\mathbb{P}_{\theta^*}$ -a.s. for  $n$  large enough:

$$|T_n|^{1/2}(\hat{\theta}_{n,x} - \theta^*) = \left( -|T_n|^{-1} \int_0^1 \nabla_{\theta}^2\ell_{n,x}(\theta^* + t(\hat{\theta}_{n,x} - \theta^*)) dt \right)^{-1} |T_n|^{-1/2}\nabla_{\theta}\ell_{n,x}(\theta^*).$$

As by Theorem 3.11, we have that  $\mathbb{P}_{\theta^*}$ -a.s.  $\lim_{n \rightarrow \infty} \hat{\theta}_{n,x} = \theta^*$ , using Theorem 4.6, we get that the first factor in the right hand side  $\mathbb{P}_{\theta^*}$ -a.s. converges to  $\mathcal{I}(\theta^*)$  as  $n \rightarrow \infty$ . Using Theorem 4.3, we get that the second factor in the right hand side converges  $\mathbb{P}_{\theta^*}$ -weakly as  $n \rightarrow \infty$  to the Gaussian random distribution  $\mathcal{N}(0, \mathcal{I}(\theta^*))$ . Hence, using Cramér-Slutsky's theorem, we get that  $|T_n|^{1/2}(\hat{\theta}_{n,x} - \theta^*)$  converges  $\mathbb{P}_{\theta^*}$ -weakly as  $n \rightarrow \infty$  to the Gaussian random distribution  $\mathcal{N}(0, \mathcal{I}(\theta^*)^{-1})$ . This concludes the proof.  $\square$

#### 4.2.1 Proof of Proposition 4.4

We are going to prove a version of Proposition 4.4 where the functions  $\varphi_{\theta}$  used in (62) to define  $\Lambda_{u,k,x}(\theta)$  are replaced by scalar-valued functions, still denoted by  $\varphi_{\theta}$ , under more general assumptions. The extension to matrix-valued functions is then straightforward by applying the result coordinate-wise.

Let  $\Theta_0$  be a compact subset of  $\Theta$ , Let  $\Theta_0$  be a closed ball in  $\Theta$ , and let  $\varphi : \Theta_0 \times \mathcal{X}^2 \times \mathcal{Y} \rightarrow \mathbb{R}$  be a Borel function such that for all  $x', x \in \mathcal{X}$  and  $y \in \mathcal{Y}$ ,  $\theta \mapsto \varphi(\theta, x', x, y) = \varphi_{\theta}(x', x, y)$  is a continuous function on  $\Theta_0$ , and such that:

$$\mathbb{E}_{\theta^*} \left[ \sup_{\theta \in \Theta_0} \sup_{x, x' \in \mathcal{X}} |\varphi_{\theta}(x, x', Y_{\partial})|^2 \right] < \infty.$$

Let  $\Lambda_{u,k,x}(\theta)$  be defined as in (62) on page 37 and note that it is in  $L^2(\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^*})$ .

The proof of Proposition 4.4 is decomposed into several lemmas.

We start with the following lemma, stating a uniform  $L^2(\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^*})$  approximation bound on the quantities  $\Lambda_{u,k,x}(\theta)$ , and the existence of a limit function  $\Lambda_{u,\infty}(\theta)$  which does not depend on  $x$ . This lemma is an immediate consequence of Lemma 4.2 (remind that  $\rho < 1/\sqrt{2}$  under the assumptions of Proposition 4.4) for the second moment ( $b = 2$ ) with  $\psi_{\theta} = \varphi_{\theta}$ , see also the discussion after Lemma 4.2 for the existence of the limit function.

**Lemma 4.8.** *Under the assumptions of Proposition 4.4, there exist finite constants  $C < \infty$  and  $\alpha \in (0, 1)$  such that for all  $u \in T$  there exists some function  $\Lambda_{u,\infty}(\theta)$  in  $L^2(\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^*})$  such that for all  $k \in \mathbb{N}^*$ , we have:*

$$\left( \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} \left[ \sup_{\theta \in \Theta_0} \sup_{x, x' \in \mathcal{X}} |\Lambda_{u,k,x}(\theta) - \Lambda_{u,\infty}(\theta)|^2 \right] \right)^{1/2} \leq C\alpha^k.$$

In particular, for all  $u \in T$ ,  $\theta \in \Theta_0$  and  $x \in \mathcal{X}$ , the sequence  $(\Lambda_{u,k,x}(\theta))_{k \in \mathbb{N}^*}$  converges in  $L^2(\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^*})$  to  $\Lambda_{u,\infty}(\theta)$  which does not depend on  $x$ .

The following lemma gives an exponential bound on the  $L^2(\mathbb{P}_{\theta^*})$  norm uniformly in  $x \in \mathcal{X}$  for the the average of the quantities  $\Lambda_{u,h(u),x}(\theta^*)$  over  $u \in T_n^*$ .

**Lemma 4.9.** *Under the assumptions of Proposition 4.4, for all  $x \in \mathcal{X}$  and  $\theta \in \Theta_0$ , there exist finite constants  $C < \infty$  and  $\alpha \in (0, 1)$  such that for all  $n \in \mathbb{N}^*$  we have:*

$$\mathbb{E}_{\theta^*} \left[ \sup_{x \in \mathcal{X}} \left| \frac{1}{|T_n|} \sum_{u \in T_n^*} \Lambda_{u,h(u),x}(\theta) - \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [\Lambda_{\partial,\infty}(\theta)] \right|^2 \right]^{1/2} \leq C\alpha^n.$$

*Proof.* Let  $x' \in \mathcal{X}$  and  $\theta \in \Theta_0$ . Using Minkowski's inequality and Jensen's inequality, for all  $n, k \in \mathbb{N}^*$ , we get:

$$\begin{aligned} & \mathbb{E}_{\theta^*} \left[ \sup_{x \in \mathcal{X}} \left| \frac{1}{|T_n|} \sum_{u \in T_n^*} \Lambda_{u,h(u),x}(\theta) - \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [\Lambda_{\partial,\infty}(\theta)] \right|^2 \right]^{1/2} \\ & \leq \mathbb{E}_{\theta^*} \left[ \sup_{x,x' \in \mathcal{X}} \left| \frac{1}{|T_n|} \sum_{u \in T_{k-1}^*} \Lambda_{u,h(u),x}(\theta) \right|^2 \right]^{1/2} \\ & \quad + \mathbb{E}_{\theta^*} \left[ \sup_{x,x' \in \mathcal{X}} \left| \frac{1}{|T_n|} \sum_{u \in T_n \setminus T_{k-1}} \Lambda_{u,h(u),x}(\theta) - \Lambda_{u,k,x'}(\theta) \right|^2 \right]^{1/2} \\ & \quad + \mathbb{E}_{\theta^*} \left[ \left| \frac{1}{|T_n|} \sum_{u \in T_n \setminus T_{k-1}} \Lambda_{u,k,x'}(\theta) - \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [\Lambda_{\partial,k,x'}(\theta)] \right|^2 \right]^{1/2} \\ & \quad + \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [|\Lambda_{\partial,k,x'}(\theta) - \Lambda_{\partial,\infty}(\theta)|^2]^{1/2}. \end{aligned} \tag{70}$$

Using Lemma 4.8 together with (49) on page 30 (which, remind, are both immediate consequences of Lemma 4.2), there exists a finite constant  $C < \infty$  and  $\beta \in (0, 1)$  such that the first term in the right hand side of (70) is upper bounded by  $C2^{-(n-k)}$  (note that  $\frac{|T_{k-1}|}{|T_n|} \leq 2^{-(n-k)}$ ), and the second and fourth terms in the right hand side of (70) are both upper bounded by  $C\beta^{k/2}$ .

We now give an upper bound for the second term in the right hand side of (70). For a vertex  $u$  in  $T \setminus T_{k-1}$ , let  $v_u \in G_k$  be the unique vertex that satisfies the shape equality constraint (8) (on page 12), then we have:

$$\Lambda_{u,k,x'}(\theta; Y_{\Delta(u,k)} = y_{\Delta(u,k)}) = \Lambda_{v_u,k,x'}(\theta; Y_{\Delta(v_u)} = y_{\Delta(u,k)}). \tag{71}$$

Moreover, using the definition of  $\Lambda_{u,k,x}(\theta)$  in (62) together with the assumption on  $\varphi_\theta$  in Proposition 4.4, we get that the random variable  $\Lambda_{u,k,x'}(\theta; Y_{\Delta(u,k)} = y_{\Delta(u,k)})$  is in  $L^2(\mathbb{P}_{\theta^*})$  for every  $u \in T \setminus T_{k-1}$ . Thus, we can apply Lemma 2.11 (see in particular (11)) to the collection of neighborhood-shape-dependent functions  $(\Lambda_{v_u,k,x'}(\theta; Y_{\Delta(v)} = \cdot))_{v \in G_k}$  (remind that indexing functions with  $G_k$  or with  $\mathcal{N}_k$  is equivalent by (9)). Using (11) in Lemma 2.11 together with (28) and (14) in Remark 3.1, we get that there exist  $\gamma \in (0, 1)$  and a finite constant  $C' < \infty$  (note

that they both do not depend on  $k$  and  $n$ ) such that for all  $n, k \in \mathbb{N}^*$  with  $n \geq k$ , the second term in the right hand side of (70) is upper bounded by  $C'\gamma^{n-k}$ .

Hence, taking  $k = \lceil n/2 \rceil$ , we get that the left hand side of (70) is upper bounded by  $2C\beta^{n/4} + C'\alpha^{n/2} + C2^{-n/2+1}$ , and thus decays at exponential rate as desired. This concludes the proof.  $\square$

Lemma 4.9 implies as a corollary the convergence  $\mathbb{P}_{\theta^*}$ -a.s. and in  $L^2(\mathbb{P}_{\theta^*})$  uniformly in  $x \in \mathcal{X}$  for the the sum of the quantities  $\Lambda_{u, h(u), x}(\theta^*)$  over  $u \in T_n^*$ .

**Corollary 4.10.** *Under the assumptions of Proposition 4.4, for all  $x \in \mathcal{X}$  and  $\theta \in \Theta_0$ , we have:*

$$\lim_{n \rightarrow \infty} \sup_{x \in \mathcal{X}} \left| \frac{1}{|T_n|} \sum_{u \in T_n^*} \Lambda_{u, h(u), x}(\theta) - \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [\Lambda_{\partial, \infty}(\theta)] \right| = 0 \quad \mathbb{P}_{\theta^*}\text{-a.s. and in } L^2(\mathbb{P}_{\theta^*}).$$

*Proof.* The convergence in  $L^2(\mathbb{P}_{\theta^*})$  follows immediately from Lemma 4.9. Moreover, using again Lemma 4.9, we have:

$$\sum_{n \in \mathbb{N}^*} \mathbb{E}_{\theta^*} \left[ \sup_{x \in \mathcal{X}} \left| \frac{1}{|T_n|} \sum_{u \in T_n^*} \Lambda_{u, \infty}(\theta) - \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [\Lambda_{\partial, \infty}(\theta)] \right|^2 \right] < \infty.$$

Hence, Borel-Cantelli lemma and Markov's inequality imply that the convergence in the lemma also holds  $\mathbb{P}_{\theta^*}$ -a.s.  $\square$

The following lemma gives some continuity properties of the function  $\theta \mapsto \Lambda_{\partial, k, x}(\theta)$ .

**Lemma 4.11.** *Under the assumptions of Proposition 4.4, for all  $x \in \mathcal{X}$  and  $k \in \mathbb{N}$ , the random function  $\theta \mapsto \Lambda_{\partial, k, x}(\theta)$  is  $\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^*}$ -a.s. continuous on  $\Theta_0$ . Moreover, for all  $\theta \in \Theta_0$ , we have:*

$$\lim_{\delta \rightarrow 0} \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} \left[ \sup_{\theta' \in \Theta_0: \|\theta' - \theta\| \leq \delta} |\Lambda_{\partial, k, x}(\theta') - \Lambda_{\partial, k, x}(\theta)|^2 \right] = 0.$$

*Proof.* We mimic the proof of [DMR04, Lemma 14].

For all  $v \in T^\infty$ , define the random variable  $\|\varphi^v\|_\infty = \sup_{\theta' \in \Theta_0} \sup_{x, x' \in \mathcal{X}} |\varphi_{\theta'}(x', x, Y_v)|$ . Remind that under the assumptions of Proposition 4.4, the HMT process  $(X, Y)$  is stationary and the random variable  $\|\varphi^\partial\|_\infty$  is in  $L^2(\mathbb{P}_{\theta^*})$ . Thus, for all  $v \in T^\infty$ , the random variable  $\|\varphi^v\|_\infty$  is in  $L^2(\mathbb{P}_{\theta^*})$ . Remind from (12) on page 15 that  $\Delta(\partial, k)$  is a random subtree of the deterministic subtree  $T^\infty(\mathfrak{p}^k(u), k)$ . Then, note that we have:

$$\sup_{\theta \in \Theta_0} |\Lambda_{\partial, k, x}(\theta)| \leq 2 \sum_{v \in T^\infty(\mathfrak{p}^k(\partial), k)} \|\varphi^v\|_\infty,$$

where the upper bound is a random variable in  $L^2(\mathbb{P}_{\theta^*})$  (and thus in  $L^2(\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^*})$ ) which depends only on  $Y_{T^\infty(\mathfrak{p}^k(u), k)}$  but not on  $\mathcal{U}$ . Hence, to prove the lemma, it suffices to prove that for all  $v \in T^\infty(\mathfrak{p}^k(u), k) \setminus \{\mathfrak{p}^k(\partial)\}$  we have:

$$\lim_{\delta \rightarrow 0} \sup_{\theta' \in \Theta_0: \|\theta' - \theta\| \leq \delta} \left| \mathbb{E}_{\theta'} [\varphi_{\theta'}(X_{\mathfrak{p}(v)}, X_v, Y_v) \mid Y_{\Delta(\partial, k)}, X_{\mathfrak{p}^k(\partial)} = x] - \mathbb{E}_{\theta} [\varphi_{\theta}(X_{\mathfrak{p}(v)}, X_v, Y_v) \mid Y_{\Delta(\partial, k)}, X_{\mathfrak{p}^k(\partial)} = x] \right| = 0, \quad \mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^*}\text{-a.s.}$$

Denote  $x_{p^k(\partial)} = x$ , and write:

$$\begin{aligned} & \mathbb{E}_\theta[\varphi_\theta(X_{p(v)}, X_v, Y_v) \mid Y_{\Delta(\partial, k)}, X_{p^k(\partial)} = x] \\ &= \frac{\int_{\mathcal{X}^{\Delta(\partial, k) \setminus \{p^k(\partial)\}}} \varphi_\theta(x_{p(v)}, x_v, Y_v) \prod_{w \in \Delta(\partial, k) \setminus \{p^k(\partial)\}} q_\theta(x_{p(w)}, x_w) g_\theta(x_w, Y_w) \lambda(dx_w)}{\int_{\mathcal{X}^{\Delta(\partial, k) \setminus \{p^k(\partial)\}}} \prod_{w \in \Delta(\partial, k) \setminus \{p^k(\partial)\}} q_\theta(x_{p(w)}, x_w) g_\theta(x_w, Y_w) \lambda(dx_w)}. \end{aligned} \quad (72)$$

Using Assumptions 2-4 (which are part of the assumptions in Proposition 4.4), we know that the integrand in the numerator of the right hand side of (72) is continuous w.r.t.  $\theta$  and is upper bounded by the random variable  $\|\varphi^v\|_\infty (\sigma^+ b^+)^{|T^\infty(p^k(u), k)|-1}$  (remind that  $\sigma^+ \geq 1$  and  $b^+ \geq 1$ ). And similarly, the denominator is continuous w.r.t.  $\theta$ , and, using Assumption 3-(ii), is lower bounded by the random variable:

$$\prod_{w \in \Delta(\partial, k) \setminus \{p^k(\partial)\}} \sigma^- \inf_{\theta' \in \Theta} \int g_{\theta'}(x_w, Y_w) \lambda(dx_w) > 0.$$

Hence, using dominated convergence, we conclude that  $\mathbb{P}_\mathcal{U} \otimes \mathbb{P}_{\theta^*}$ -a.s. the left hand side of (72) is continuous w.r.t.  $\theta$ . This concludes the proof.  $\square$

As a corollary of Lemma 4.11, we get that the function  $\theta \mapsto \Lambda_{\partial, \infty}(\theta)$  is continuous in  $L^2(\mathbb{P}_{\theta^*})$ .

**Corollary 4.12.** *Under the assumptions of Proposition 4.4, for all  $\theta \in \Theta_0$ , we have:*

$$\lim_{\delta \rightarrow 0} \mathbb{E}_\mathcal{U} \otimes \mathbb{E}_{\theta^*} \left[ \sup_{\theta' \in \Theta_0: \|\theta' - \theta\| \leq \delta} |\Lambda_{\partial, \infty}(\theta') - \Lambda_{\partial, \infty}(\theta)|^2 \right] = 0.$$

In particular, the function  $\theta \mapsto \mathbb{E}_\mathcal{U} \otimes \mathbb{E}_{\theta^*}[\Lambda_{\partial, \infty}(\theta)]$  is continuous on  $\Theta_0$ .

*Proof.* Using Minkowski's inequality and Lemma 4.8, there exist a finite constant  $C < \infty$  and  $\beta \in (0, 1)$  such that for all  $x \in \mathcal{X}$  and  $k \in \mathbb{N}^*$ , we have:

$$\begin{aligned} & \mathbb{E}_\mathcal{U} \otimes \mathbb{E}_{\theta^*} \left[ \sup_{\theta' \in \Theta_0: \|\theta' - \theta\| \leq \delta} |\Lambda_{\partial, \infty}(\theta') - \Lambda_{\partial, \infty}(\theta)|^2 \right]^{1/2} \\ & \leq 2C\beta^{k/2} + \mathbb{E}_\mathcal{U} \otimes \mathbb{E}_{\theta^*} \left[ \sup_{\theta' \in \Theta_0: \|\theta' - \theta\| \leq \delta} |\Lambda_{\partial, k, x}(\theta') - \Lambda_{\partial, k, x}(\theta)|^2 \right]^{1/2}. \end{aligned} \quad (73)$$

Using Lemma 4.11, we get:

$$\limsup_{\delta \rightarrow 0} \mathbb{E}_\mathcal{U} \otimes \mathbb{E}_{\theta^*} \left[ \sup_{\theta' \in \Theta_0: \|\theta' - \theta\| \leq \delta} |\Lambda_{\partial, \infty}(\theta') - \Lambda_{\partial, \infty}(\theta)|^2 \right]^{1/2} \leq 2C\beta^{k/2},$$

and taking  $k \rightarrow \infty$ , the upper bound vanishes. This concludes the proof.  $\square$

We now prove a locally uniform law of large numbers for the quantities  $\Lambda_{u, k, x}(\theta)$ .

**Lemma 4.13.** *Under the assumptions of Proposition 4.4, for all  $x \in \mathcal{X}$ , we have:*

$$\lim_{\delta \rightarrow 0} \lim_{n \rightarrow \infty} \sup_{\theta' \in \Theta_0: \|\theta' - \theta\| \leq \delta} \left| \frac{1}{|T_n|} \sum_{u \in T_n^*} \Lambda_{u, h(u), x}(\theta') - \mathbb{E}_\mathcal{U} \otimes \mathbb{E}_{\theta^*}[\Lambda_{\partial, \infty}(\theta)] \right| = 0, \quad \mathbb{P}_{\theta^*}\text{-a.s.}$$

*Proof.* First, write:

$$\begin{aligned}
& \sup_{\theta' \in \Theta_0 : \|\theta' - \theta\| \leq \delta} \left| \frac{1}{|T_n|} \sum_{u \in T_n^*} \Lambda_{u, h(u), x}(\theta') - \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [\Lambda_{\partial, \infty}(\theta)] \right| \\
& \leq \frac{1}{|T_n|} \sum_{u \in T_n^*} \sup_{\theta' \in \Theta_0 : \|\theta' - \theta\| \leq \delta} |\Lambda_{u, h(u), x}(\theta') - \Lambda_{u, h(u), x}(\theta)| \\
& \quad + \left| \frac{1}{|T_n|} \sum_{u \in T_n^*} \Lambda_{u, h(u), x}(\theta) - \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [\Lambda_{\partial, \infty}(\theta)] \right|. \tag{74}
\end{aligned}$$

Then, we use the exact same argument as in the proofs of Lemma 4.9 and Corollary 4.10 where for all  $u \in T^*$ , the random variable  $\Lambda_{u, h(u), x}(\theta)$  is replaced by the random variable:

$$\sup_{\theta' \in \Theta_0 : \|\theta' - \theta\| \leq \delta} |\Lambda_{u, h(u), x}(\theta') - \Lambda_{u, h(u), x}(\theta)|,$$

which are in  $L^2(\mathbb{P}_{\theta^*})$  using the assumptions of Proposition 4.4. This gives us that the first term in the upper bound of (74) converges  $\mathbb{P}_{\theta^*}$ -a.s. as  $n \rightarrow \infty$  to:

$$\mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} \left[ \sup_{\theta' : \|\theta' - \theta\| \leq \delta} |\Lambda_{\partial, \infty}(\theta') - \Lambda_{\partial, \infty}(\theta)| \right],$$

which, by Corollary 4.12, vanishes when  $\delta \rightarrow 0$ . Corollary 4.10 implies that the second term in the upper bound of (74) vanishes  $\mathbb{P}_{\theta^*}$ -a.s. when  $n \rightarrow \infty$ . This concludes the proof.  $\square$

Combining the previous lemmas in this subsection, we are now ready to prove Proposition 4.4.

*Proof of Proposition 4.4.* By Lemma 4.8, for all  $u \in T$ , we have that  $(\Lambda_{u, k, x}(\theta))_{k \in \mathbb{N}^*}$  is a Cauchy sequence uniformly w.r.t.  $\theta \in \Theta_0$  in  $L^2(\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^*})$  that converges to some limit  $\Lambda_{u, \infty}(\theta)$  (that does not depend on  $x$ ). By Corollary 4.10, we have that  $\mathbb{P}_{\theta^*}$ -a.s. the convergence for the the average of the quantities  $\Lambda_{u, h(u), x}(\theta^*)$  over  $u \in T_n^*$  holds uniformly in  $x \in \mathcal{X}$ , that is, (65) in Proposition 4.4 holds. By Corollary 4.12, we have that the function  $\theta \mapsto \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [\Lambda_{\partial, \infty}(\theta)]$  is continuous on  $\Theta_0$ . Finally, the last part of the proposition is given by Lemma 4.13.  $\square$

#### 4.2.2 Proof of Proposition 4.5

Similarly to what we have done for Proposition 4.4, we are going to prove a version of Proposition 4.5 where the functions  $\phi_\theta$  used in (63) to define  $\Gamma_{u, k, x}(\theta)$  are replaced by scalar-valued functions, still denoted by  $\varphi_\theta$ , under more general assumptions. The extension to matrix-valued functions is then straightforward by applying the result coordinate-wise.

Let  $\Theta_0$  be a compact subset of  $\Theta$ , Let  $\Theta_0$  be a closed ball in  $\Theta$ , and let  $\phi : \Theta_0 \times \mathcal{X}^2 \times \mathcal{Y} \rightarrow \mathbb{R}$  be a Borel function such that for all  $x', x \in \mathcal{X}$  and  $y \in \mathcal{Y}$ ,  $\theta \mapsto \phi(\theta, x', x, y) = \phi_\theta(x', x, y)$  is a continuous function on  $\Theta_0$ , and such that:

$$\mathbb{E}_{\theta^*} \left[ \sup_{\theta \in \Theta_0} \sup_{x, x' \in \mathcal{X}} |\phi_\theta(x, x', Y_\partial)|^4 \right] < \infty.$$

Let  $\Gamma_{u, k, x}(\theta)$  be defined as in (63) on page 37 and note that it is in  $L^2(\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^*})$ .

The proof of Proposition 4.4 can be straightforwardly adapted to Proposition 4.5 except for Lemma 4.8. Thus, for brevity, we only present the adaptation of Lemma 4.8 to the terms

$\Gamma_{u,k,x}(\theta)$ . (The details of the adaptation for the rest of the proof of Proposition 4.4 to the terms  $\Gamma_{u,k,x}(\theta)$  can be found in Appendix D.)

We start with two lemmas giving coupling bounds that will be used to control the covariance terms that appear in the definition of the terms  $\Gamma_{u,k,x}(\theta)$ . The following lemma is a variant of Lemma 3.2 for two vertices of  $T$ .

**Lemma 4.14** (Forward coupling bound for two vertices). *Assume that Assumptions 2 and 3 hold. Then, for all  $u, v \in T$ , all  $y_{T_n} \in \mathcal{Y}^{|T_n|}$  (and  $n \in \mathbb{N}$ ) and all initials distributions  $\nu$  and  $\nu'$  on  $\mathcal{X}$ , we have:*

$$\left\| \int_{\mathcal{X}} \mathbb{P}_{\theta} \left( X_u \in \cdot, X_v \in \cdot \mid Y_{T_n} = y_{T_n}, X_{\partial} = x \right) [\nu(dx) - \nu'(dx)] \right\|_{\text{TV}} \leq 2\rho^{\min(h(u), h(v))}.$$

For simplicity, Lemma 4.14 is stated with  $\partial$  as the initial vertex, but note that the results still holds when replacing  $\partial$  and  $T_n$  by  $v$  and  $T(v, n)$  for any  $v \in T^{\infty}$ . We shall reuse this fact later.

*Proof.* We are going to construct a coupling that achieves this minimum. Denote by  $((X'_w, Y'_w), w \in T)$  the process started from the distribution  $\nu'$  (and similarly without the  $'$ ). Remark that we only need to define the (joint) coupling for the variables  $X_w$  and  $X'_w$  for  $w$  on the paths between the vertices  $\partial, u$  and  $v$ .

Lemma 3.2 applied to the vertex  $u \wedge v$  gives us a coupling for the variables  $X_w$  and  $X'_w$  for  $w$  on the path between  $\partial$  and  $u \wedge v$  with successful coupling probability upper bounded by  $1 - \rho^{h(u \wedge v)}$ . On this successful coupling event before or on  $u \wedge v$ , the two processes are still defined to be equal after the fork on both branches leading to  $u$  and  $v$ .

On the complementary event (no successful coupling before or on  $u \wedge v$ ), we get two new distributions  $\nu_{u \wedge v}$  and  $\nu'_{u \wedge v}$  for the variables  $X_{u \wedge v}$  and  $X'_{u \wedge v}$ , respectively. Note that conditioned on the value of  $X_{u \wedge v}$ , the two branches leading to  $u$  and  $v$  are independent. Thus, applying Lemma 3.2 to  $u$  (resp.  $v$ ) with the initial distributions  $\nu_{u \wedge v}$  and  $\nu'_{u \wedge v}$ , we construct a coupling of the processes  $X$  and  $X'$  on the branch from  $u \wedge v$  to  $u$  (resp.  $v$ ) with successful coupling probability upper bounded by  $1 - \rho^{h(u) - h(u \wedge v)}$  (resp.  $1 - \rho^{h(v) - h(u \wedge v)}$ ).

Hence, the probability that we do not get a successful coupling on at least one of the two variables  $X_u$  and  $X_v$  is upper bounded by  $\rho^{h(u \wedge v)}(\rho^{h(u) - h(u \wedge v)} + \rho^{h(v) - h(u \wedge v)}) \leq 2\rho^{\min(h(u), h(v))}$ . This concludes the proof.  $\square$

The following lemma is a variant of Lemma 4.1 giving a “backward in time” coupling bound for two vertices of  $T$ .

**Lemma 4.15** (Backward coupling bound for two vertices). *Assume that Assumptions 2–3 hold. Let  $k \in \mathbb{N}^*$ ,  $x \in \mathcal{X}$  and  $u \in T$ , and let  $v, w \in T^{\infty}(\mathfrak{p}^k(u), k) \setminus \{u\}$ . Then, we have:*

$$\left\| \mathbb{P}_{\theta}(X_v \in \cdot, X_w \in \cdot \mid Y_{\Delta(u,k)}, X_{\mathfrak{p}^k(u)} = x) - \mathbb{P}_{\theta}(X_v \in \cdot, X_w \in \cdot \mid Y_{\Delta^*(u,k)}, X_{\mathfrak{p}^k(u)} = x) \right\|_{\text{TV}} \leq 2\rho^{\min(d(u,v), d(u,w)) - 1}.$$

*Proof.* The idea of the proof is similar to that of Lemma 4.14. We explicitly construct a coupling with coupling failure probability upper bounded by  $2\rho^{\min(d(u,v), d(u,w)) - 1}$ . Denote by  $w_0$  the vertex on the path between  $v$  and  $w$  that is the closest to  $u$ , and note that we have  $w_0 \in T^{\infty}(\mathfrak{p}^k(u), k) \setminus \{u\}$ . Note that  $w_0$  is on the path from  $\mathfrak{p}(u)$  to  $v$ , and thus  $d(\mathfrak{p}(u), v) = d(\mathfrak{p}(u), w_0) + d(w_0, v)$ , and similarly when replacing  $v$  by  $w$ . On the path from  $\mathfrak{p}(u)$  to  $w_0$ , we use the coupling provided by the “backward in time” coupling bound from Lemma 4.1 with successful probability  $1 - \rho^{d(\mathfrak{p}(u), w_0)}$ . On the path from  $w_0$  to  $v$  and the other path from  $w_0$  to  $w$ , which are independent

using the Markov property, we use two independent couplings that are constructed using a similar coupling argument as in Lemma 4.1 with  $p(u)$  replaced by  $w_0$ . Those independent couplings have successful probabilities  $1 - \rho^{d(w_0, v)}$  and  $1 - \rho^{d(w_0, w)}$ , respectively. Note that the coupling we have constructed has a coupling failure probability upper bounded by  $2\rho^{\min(d(u, v), d(u, w)) - 1}$ .  $\square$

For brevity, for all  $u \in T$  we will denote  $\phi_{\theta, u} = \phi_{\theta}(X_{p(u)}, X_u, Y_u)$  and  $\|\phi_u\|_{\infty} = \sup_{\theta \in \Theta_0} \sup_{x, x' \in \mathcal{X}} |\phi_{\theta}(x', x, Y_u)|$ . The following lemma gives several upper bounds on the covariance terms that appear in the definition of the terms  $\Gamma_{u, k, x}(\theta)$ . Remind from (12) on page 15 that  $\Delta(\partial, k)$  is a random subtree of the deterministic subtree  $T^{\infty}(p^k(u), k)$ .

Note that this lemma is stated under the assumptions of Proposition 4.5, but here we do not need the assumption that  $\rho < 1/2$  for the mixing rate  $\rho$  of the HMT process  $(X, Y)$ .

**Lemma 4.16.** *Under the assumptions of Proposition 4.5 (without the need for the assumption that  $\rho < 1/2$ ), for all  $x, x' \in \mathcal{X}$ ,  $\theta \in \Theta_0$ ,  $k' \geq k > 0$  and  $u \in T$ , and for all  $v, w \in T^{\infty}(p^k(u), k) \setminus \{p^k(u)\}$ , we have:*

$$|\text{Cov}_{\theta}[\phi_{\theta, v}, \phi_{\theta, w} | Y_{\Delta(u, k)}, X_{p^k(u)} = x]| \leq 2 \|\phi_v\|_{\infty} \|\phi_w\|_{\infty} \rho^{d(v, w) - 2}, \quad (75)$$

and,

$$\begin{aligned} & |\text{Cov}_{\theta}[\phi_{\theta, v}, \phi_{\theta, w} | Y_{\Delta(u, k)}, X_{p^k(u)} = x] - \text{Cov}_{\theta}[\phi_{\theta, v}, \phi_{\theta, w} | Y_{\Delta(u, k')}, X_{p^{k'}(u)} = x']| \\ & \leq 8 \|\phi_v\|_{\infty} \|\phi_w\|_{\infty} \rho^{\min(d(p^k(u), v), d(p^k(u), w)) - 2}. \end{aligned} \quad (76)$$

Moreover, if  $v, w \in \Delta^*(u, k)$ , then we have:

$$\begin{aligned} & |\text{Cov}_{\theta}[\phi_{\theta, v}, \phi_{\theta, w} | Y_{\Delta(u, k)}, X_{p^k(u)} = x] - \text{Cov}_{\theta}[\phi_{\theta, v}, \phi_{\theta, w} | Y_{\Delta^*(u, k)}, X_{p^k(u)} = x]| \\ & \leq 8 \|\phi_v\|_{\infty} \|\phi_w\|_{\infty} \rho^{\min(d(u, v), d(u, w)) - 2}. \end{aligned} \quad (77)$$

*Proof.* We start by proving (75), that is, the first inequality in the lemma. Let  $A_1, A_2, B_1, B_2$  be measurable subsets of  $\mathcal{X}$ , and we write  $A = A_1 \times A_2$  and  $B = B_1 \times B_2$ . If one of the two vertices  $v$  and  $w$  is an ancestor of the other, say  $w$  is an ancestor of  $v$  (which implies that  $w$  is also an ancestor of  $p(v)$ ), then using the Markov property of the HMT process  $(X, Y)$ , we get:

$$\begin{aligned} & |\mathbb{P}_{\theta}(X_{\{p(v), v\}} \in A, X_{\{p(w), w\}} \in B | Y_{\Delta(u, k)}, X_{p^k(u)} = x) \\ & - \mathbb{P}_{\theta}(X_{\{p(v), v\}} \in A | Y_{\Delta(u, k)}, X_{p^k(u)} = x) \mathbb{P}_{\theta}(X_{\{p(w), w\}} \in B | Y_{\Delta(u, k)}, X_{p^k(u)} = x)| \\ & = \mathbb{P}_{\theta}(X_{\{p(w), w\}} \in B | Y_{\Delta(u, k)}, X_{p^k(u)} = x) \\ & \quad \times \left| \int_{\mathcal{X}^2} \mathbb{1}_{\{(x_{p(v)}, x_v) \in A\}} \mathbb{P}_{\theta}(X_v \in dx_v | Y_{\Delta(u, k)}, X_{p(v)} = x_{p(v)}) \right. \\ & \quad \times [\mathbb{P}_{\theta}(X_{p(v)} \in dx_{p(v)} | Y_{\Delta(u, k)}, X_w \in B_2, X_{p^k(u)} = x) \\ & \quad \left. - \mathbb{P}_{\theta}(X_{p(v)} \in dx_{p(v)} | Y_{\Delta(u, k)}, X_{p^k(u)} = x) \right| \\ & \leq \rho^{d(p(v), w)} \\ & = \rho^{d(v, w) - 1}, \end{aligned}$$

where the inequality follows using the same argument as in the proof of the ‘‘backward in time’’ coupling Lemma 4.1 (with the role of  $p(u)$  replaced by  $w$  and using the initial distributions  $\mathbb{P}((X_{p(w)}, X_w) \in \cdot | Y_{\Delta(u, k)}, X_w \in B_2, X_{p^k(u)} = x)$  with  $B' = B$  and  $\mathcal{X}$  respectively).

Otherwise, we have that both  $p(v)$  and  $p(w)$  are on the path between  $v$  and  $w$ , and similarly to the first case, we get:

$$\begin{aligned}
& |\mathbb{P}_\theta(X_{\{p(v),v\}} \in A, X_{\{p(w),w\}} \in B \mid Y_{\Delta(u,k)}, X_{p^k(u)} = x) \\
& \quad - \mathbb{P}_\theta(X_{\{p(v),v\}} \in A \mid Y_{\Delta(u,k)}, X_{p^k(u)} = x) \mathbb{P}_\theta(X_{\{p(w),w\}} \in B \mid Y_{\Delta(u,k)}, X_{p^k(u)} = x)| \\
& \quad = \mathbb{P}_\theta(X_{\{p(w),w\}} \in B \mid Y_{\Delta(u,k)}, X_{p^k(u)} = x) \\
& \quad \quad \times \left| \int_{\mathcal{X}^2} \mathbb{1}_{\{(x_{p(v)}, x_v) \in A\}} \mathbb{P}_\theta(X_v \in dx_v \mid Y_{\Delta(u,k)}, X_{p(v)} = x_{p(v)}) \right. \\
& \quad \quad \quad \times [\mathbb{P}_\theta(X_{p(v)} \in dx_{p(v)} \mid Y_{\Delta(u,k)}, X_{p(w)} \in B_1, X_{p^k(u)} = x) \\
& \quad \quad \quad \quad \left. - \mathbb{P}_\theta(X_{p(v)} \in dx_{p(v)} \mid Y_{\Delta(u,k)}, X_{p^k(u)} = x)] \right| \\
& \leq \rho^{d(p(v), p(w))} \\
& = \rho^{d(v,w)-2}.
\end{aligned}$$

Thus, in both case, we get:

$$\begin{aligned}
& |\mathbb{P}_\theta(X_{\{p(v),v\}} \in A, X_{\{p(w),w\}} \in B \mid Y_{\Delta(u,k)}, X_{p^k(u)} = x) \\
& \quad - \mathbb{P}_\theta(X_{\{p(v),v\}} \in A \mid Y_{\Delta(u,k)}, X_{p^k(u)} = x) \mathbb{P}_\theta(X_{\{p(w),w\}} \in B \mid Y_{\Delta(u,k)}, X_{p^k(u)} = x)| \\
& \leq \rho^{d(v,w)-2}.
\end{aligned}$$

This gives that (75) holds. (Note that the functions  $\phi_{\theta,v}$  and  $\phi_{\theta,w}$  can take positive, null or negative values.)

For (76), that is, the second inequality in the lemma, use the decomposition:

$$\begin{aligned}
& |\text{Cov}_\theta[\phi_{\theta,v}, \phi_{\theta,w} \mid Y_{\Delta(u,k)}, X_{p^k(u)} = x] - \text{Cov}_\theta[\phi_{\theta,v}, \phi_{\theta,w} \mid Y_{\Delta(u,k')}, X_{p^{k'}(u)} = x']| \\
& \leq |\mathbb{E}_\theta[\phi_{\theta,v} \phi_{\theta,w} \mid Y_{\Delta(u,k)}, X_{p^k(u)} = x] - \mathbb{E}_\theta[\phi_{\theta,v} \phi_{\theta,w} \mid Y_{\Delta(u,k')}, X_{p^{k'}(u)} = x']| \\
& \quad + |\mathbb{E}_\theta[\phi_{\theta,v} \mid Y_{\Delta(u,k)}, X_{p^k(u)} = x] - \mathbb{E}_\theta[\phi_{\theta,v} \mid Y_{\Delta(u,k')}, X_{p^{k'}(u)} = x']| \\
& \quad \quad \times |\mathbb{E}_\theta[\phi_{\theta,w} \mid Y_{\Delta(u,k)}, X_{p^k(u)} = x]| \\
& \quad + |\mathbb{E}_\theta[\phi_{\theta,w} \mid Y_{\Delta(u,k)}, X_{p^k(u)} = x] - \mathbb{E}_\theta[\phi_{\theta,w} \mid Y_{\Delta(u,k')}, X_{p^{k'}(u)} = x']| \\
& \quad \quad \times |\mathbb{E}_\theta[\phi_{\theta,v} \mid Y_{\Delta(u,k')}, X_{p^{k'}(u)} = x']|,
\end{aligned}$$

and then use the joint coupling Lemma 4.14 with  $v' = p(v)$  and  $w' = p(w)$  for the first term in the upper bound, and use the coupling Lemma 3.2 for the other two terms with  $p(v)$  and  $p(w)$ , respectively.

For (77), that is, the third inequality in the lemma, use a similar decomposition as for (76), and then use the “backward in time” coupling for two vertices from Lemma 4.15 for the first term in the upper bound, and use the “backward in time” coupling Lemma 4.1 for the other two terms. This gives an upper bound of  $8 \|\phi_v\|_\infty \|\phi_w\|_\infty \rho^m$  with  $m = \min\{d(p(u), w_0) : w_0 \in \{p(v), v, p(w), w\}\}$ . Noting  $m \leq \min(d(u, v), d(u, w)) - 2$ , we get that (77) holds. This concludes the proof of the lemma.  $\square$

We are now ready to prove the following lemma which is the adaptation of Lemma 4.8 to the terms  $\Gamma_{u,k,x}(\theta)$ , and which gives us a uniform  $L^2(\mathbb{P}_{\theta^*})$  approximation bound.

Note that the condition  $\rho < 1/2$  on the mixing rate  $\rho$  of the HMT process  $(X, Y)$  is due to the coupling bounds from Lemma 4.16 and the grouping of terms used in the proof of Lemma 4.17 (the upper bounds at the end of the proof only add a constant multiplicative factor). See the discussion in Remark 1.5.



**Lemma 4.17.** *Under the assumptions of Proposition 4.5, there exists a positive constant  $C < \infty$  such that for all  $u \in T$  and  $0 < k \leq k'$ , we have:*

$$\begin{aligned} \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} \left[ \sup_{\theta \in \Theta_0} \sup_{x, x' \in \mathcal{X}} |\Gamma_{u, k, x}(\theta) - \Gamma_{u, k', x'}(\theta)|^2 \right]^{1/2} \\ \leq C \mathbb{E}_{\theta^*} \left[ \sup_{\theta \in \Theta_0} \sup_{x, x' \in \mathcal{X}} |\varphi_{\theta}(x, x', Y_{\partial})|^4 \right]^{1/2} k^2 (2\rho)^{k/3}. \end{aligned}$$

*Proof.* Let  $u$ ,  $k$  and  $k'$  be as in the lemma. Similarly to the proof of Lemma 4.2, we use the bounds from Lemma 4.16 and Minkowski's inequality to bound the left hand side of the inequality in the lemma. For a finite subset  $I \subset T^\infty$ , we write  $S_I = \sum_{v \in I} \phi_{\theta, v}$  (the dependence on  $\theta$  is implicit). Similarly to the proof of [DMR04, Lemma 17], the difference  $\Gamma_{u, k, x}(\theta) - \Gamma_{u, k', x'}(\theta)$  may be rewritten as  $A + 2B + C + D + 2E + 2F$ , where all those terms are random variables which depend on  $Y_{\Delta(u, k')}$  and implicitly on  $\mathcal{U}$ , and are define as:

$$\begin{aligned} A &= \text{Var}_{\theta}[S_{\Delta^*(u, k)} | Y_{\Delta(u, k)}, X_{\mathfrak{p}^k(u)} = x] - \text{Var}_{\theta}[S_{\Delta^*(u, k)} | Y_{\Delta^*(u, k)}, X_{\mathfrak{p}^k(u)} = x] \\ &\quad - \text{Var}_{\theta}[S_{\Delta^*(u, k)} | Y_{\Delta(u, k')}, X_{\mathfrak{p}^{k'}(u)} = x'] + \text{Var}_{\theta}[S_{\Delta^*(u, k)} | Y_{\Delta^*(u, k')}, X_{\mathfrak{p}^{k'}(u)} = x'], \\ B &= \text{Cov}_{\theta}[S_{\Delta^*(u, k)}, \phi_{\theta, u} | Y_{\Delta(u, k)}, X_{\mathfrak{p}^k(u)} = x] \\ &\quad - \text{Cov}_{\theta}[S_{\Delta^*(u, k)}, \phi_{\theta, u} | Y_{\Delta(u, k')}, X_{\mathfrak{p}^{k'}(u)} = x'], \\ C &= \text{Var}_{\theta}[\phi_{\theta, u} | Y_{\Delta(u, k)}, X_{\mathfrak{p}^k(u)} = x] - \text{Var}_{\theta}[\phi_{\theta, u} | Y_{\Delta(u, k')}, X_{\mathfrak{p}^{k'}(u)} = x'], \\ D &= \text{Var}_{\theta}[S_{\Delta^*(u, k') \setminus \Delta^*(u, k)} | Y_{\Delta(u, k')}, X_{\mathfrak{p}^{k'}(u)} = x'] \\ &\quad - \text{Var}_{\theta}[S_{\Delta^*(u, k') \setminus \Delta^*(u, k)} | Y_{\Delta^*(u, k')}, X_{\mathfrak{p}^{k'}(u)} = x'], \\ E &= \text{Cov}_{\theta}[S_{\Delta^*(u, k') \setminus \Delta^*(u, k)}, S_{\Delta^*(u, k)} | Y_{\Delta(u, k')}, X_{\mathfrak{p}^{k'}(u)} = x'] \\ &\quad - \text{Cov}_{\theta}[S_{\Delta^*(u, k') \setminus \Delta^*(u, k)}, S_{\Delta^*(u, k)} | Y_{\Delta^*(u, k')}, X_{\mathfrak{p}^{k'}(u)} = x'], \\ F &= \text{Cov}_{\theta}[S_{\Delta^*(u, k') \setminus \Delta^*(u, k)}, \phi_{\theta, u} | Y_{\Delta(u, k')}, X_{\mathfrak{p}^{k'}(u)} = x'] \\ &\quad - \text{Cov}_{\theta}[S_{\Delta^*(u, k') \setminus \Delta^*(u, k)}, \phi_{\theta, u} | Y_{\Delta^*(u, k')}, X_{\mathfrak{p}^{k'}(u)} = x']. \end{aligned}$$

Using Minkowski's inequality, we will upper bound each of those six terms separately. First remark using Cauchy-Schwarz inequality, the stationarity of the process  $((X_u, Y_u), u \in T^\infty)$  and the assumptions in the proposition, that we have  $\mathbb{E}_{\theta^*}[\|\phi_v\|_\infty^2 \|\phi_w\|_\infty^2] \leq \mathbb{E}_{\theta^*}[\|\phi_\partial\|_\infty^4] < \infty$  for all  $v, w \in T^\infty$ .

Remind from (12) on page 15 that  $\Delta(u, k)$  is a random subtree of the deterministic subtree  $T^\infty(\mathfrak{p}^k(u), k)$ .

**Upper bound for A:** Applying the three inequalities in Lemma 4.16 and Minkowski's inequality, we get that  $\mathbb{E}_{\theta^*}[|A|^2]^{1/2}$  is upper bounded (up to the factor  $\mathbb{E}_{\theta^*}[\|\phi_\partial\|_\infty^4]$ ) by:

$$\begin{aligned} &2 \sum_{v, w \in T^\infty(\mathfrak{p}^k(u), k)} (2 \times 8\rho^{\min(d(v, u), d(w, u)) - 2} \wedge 2 \times 8\rho^{\min(d(v, \mathfrak{p}^k(u)), d(w, \mathfrak{p}^k(u))) - 2} \wedge 4 \times 2\rho^{d(v, w) - 2}) \\ &\leq \frac{32}{\rho^2} \sum_{v, w \in T^\infty(\mathfrak{p}^k(u), k)} (\rho^{\min(d(v, u), d(w, u))} \wedge \rho^{\min(d(v, \mathfrak{p}^k(u)), d(w, \mathfrak{p}^k(u)))} \wedge \rho^{d(v, w)}). \end{aligned} \quad (78)$$

Note that the value of this sum does not depend on the choice of  $u \in T$ .

For all  $j \in \mathbb{N}$ , denote  $u_j = \mathfrak{p}^j(u)$ . We will divide the sum in the upper bound of (78) according to four cases:  $v, w \in T^\infty(u_k, k) \setminus T^\infty(u_{\lfloor k/3 \rfloor}, \lfloor k/3 \rfloor)$ , or  $v, w \in T^\infty(u_{\lfloor 2k/3 \rfloor}, \lfloor 2k/3 \rfloor)$ ,

or  $v \in T^\infty(u_k, k) \setminus T^\infty(u_{\lfloor 2k/3 \rfloor}, \lfloor 2k/3 \rfloor)$  and  $w \in T^\infty(u_{\lfloor k/3 \rfloor}, \lfloor k/3 \rfloor)$  or similarly exchanging the roles of  $v$  and  $w$ . Note that those conditions are non-exclusive and we will count some vertices several times, but this is not a problem.

Let  $i, j \in \mathbb{N}$  be such that  $u \wedge v = u_i$  and  $u \wedge w = u_j$ , and let  $a, b \in \mathbb{N}$  be such that  $a = d(u_i, v)$  and  $b = d(u_j, w)$ . Note that for  $v, w$  in the first case, either  $\min(d(v, u), d(w, u))$  or  $d(v, w)$  is large, and thus using elementary computation we upper bound the sum for  $v, w$  in the first case by:

$$\begin{aligned} & \sum_{i=\lfloor k/3 \rfloor}^k \sum_{j=\lfloor k/3 \rfloor}^k \sum_{a=0}^i \sum_{b=0}^j 2^{a+b} (\rho^{\min(i+a, j+b)} \wedge \rho^{a+b+|j-i|}) \\ & \leq 2 \sum_{i=\lfloor k/3 \rfloor}^k \sum_{j=i}^k \sum_{a=0}^i \sum_{b=0}^j 2^{a+b} (\rho^{i+\min(a, b)} \wedge \rho^{a+b}) \\ & \leq \frac{8(1-\rho)}{(1-2\rho)^3} k (2\rho)^{\lfloor k/3 \rfloor + 1}. \end{aligned}$$

Note that for  $v, w$  in the second case, either  $\min(d(v, p^k(u)), d(w, p^k(u)))$  or  $d(v, w)$  is large, and thus using elementary computation we upper bound the sum for  $v, w$  in the second case by:

$$\sum_{i=0}^{\lfloor 2k/3 \rfloor} \sum_{j=0}^{\lfloor 2k/3 \rfloor} \sum_{a=0}^i \sum_{b=0}^j 2^{a+b} (\rho^{\min(k-i+a, k-j+b)} \wedge \rho^{a+b+|j-i|}) \leq \frac{8(1-\rho)}{(1-2\rho)^3} k (2\rho)^{\lfloor k/3 \rfloor + 1}.$$

Note that for  $v, w$  in the third and fourth case, either  $\min(d(v, p^k(u)), d(w, p^k(u)))$  or  $d(v, w)$  is large, and thus we upper bound the sum for  $v, w$  in the third and fourth case by:

$$2 \sum_{i=0}^{\lfloor k/3 \rfloor} \sum_{j=\lfloor 2k/3 \rfloor}^k \sum_{a=0}^i \sum_{b=0}^j 2^{a+b} \rho^{a+b+|j-i|} \leq \frac{2}{1-2\rho} k^2 \rho^{k/3-1}.$$

Putting those three upper bounds together, we get that  $\mathbb{E}_{\theta^*}[|A|^2]^{1/2}$  is upper bounded by an expression as in the lemma.

**Upper bound for B:** Using the first and second inequalities in Lemma 4.16 and Minkowski's inequality, we get that  $\mathbb{E}_{\theta^*}[|B|^2]^{1/2}$  is upper bounded (up to the factor  $\mathbb{E}_{\theta^*}[\|\phi_\partial\|_\infty^4]$ ) by:

$$8 \sum_{v \in \Delta^*(u, k)} (\rho^{d(v, u)-2} \wedge \rho^{d(v, p^k(u))-2}) \leq C k (\max(\rho, 2\rho^2))^{k/2},$$

where we used the same computation as in the proof of Lemma 4.2 and  $C < \infty$  is some finite constant (which depends only on  $\rho$ ).

**Upper bound for C:** Using the second equation in Lemma 4.16, we get that  $\mathbb{E}_{\theta^*}[|C|^2]^{1/2} \leq 8\rho^{k-2} \mathbb{E}_{\theta^*}[\|\phi_\partial\|_\infty^2]$ .

**Upper bound for D:** Using the first and third equation in Lemma 4.16, Minkowski's inequality and elementary computation, we get that  $\mathbb{E}_{\theta^*}[|D|^2]^{1/2}$  is upper bounded (up to the factor  $\mathbb{E}_{\theta^*}[\|\phi_\partial\|_\infty^4]$ ) by:

$$\sum_{v, w \in T^\infty(u_{k'}, k') \setminus T^\infty(u_k, k)} (2 \times 2\rho^{d(v, w)-2} \wedge 8\rho^{\min(d(v, u), d(w, u))-2}) \leq \frac{96}{\rho^2(1-2\rho)^4} k (2\rho)^k.$$

**Upper bound for E:** Using the first and third equation in Lemma 4.16, Minkowski's inequality and elementary computation, we get that  $\mathbb{E}_{\theta^*}[|E|^2]^{1/2}$  is upper bounded (up to the factor  $\mathbb{E}_{\theta^*}[\|\phi_{\partial}\|_{\infty}^4]$ ) by:

$$\sum_{v \in T^{\infty}(u_{k'}, k') \setminus T^{\infty}(u_k, k)} \sum_{w \in T^{\infty}(u_k, k)} (2 \times 2\rho^{d(v, w)-2} \wedge 8\rho^{\min(d(v, u), d(w, u))-2}) \leq \frac{64}{\rho^2(1-\rho)(1-2\rho)^2} k^2 (2\rho)^{\lfloor k/2 \rfloor}.$$

**Upper bound for F:** Using the first equation in Lemma 4.16 and Minkowski's inequality, we get that  $\mathbb{E}_{\theta^*}[|F|^2]^{1/2}$  is upper bounded (up to the factor  $\mathbb{E}_{\theta^*}[\|\phi_{\partial}\|_{\infty}^4]$ ) by:

$$2 \times 2 \sum_{v \in T^{\infty}(u_{k'}, k') \setminus T^{\infty}(u_k, k)} \rho^{d(v, u)-2} \leq \frac{4}{1-2\rho} \frac{\rho^{k-1}}{1-\rho}.$$

Hence, as the  $L^2(\mathbb{P}_{\theta^*})$  norm for the six terms  $A, B, C, D, E$  and  $F$  are all upper bounded by expressions as in the lemma, we get that the upper bound in the lemma holds. This concludes the proof.  $\square$

As announce at the beginning of this subsection, the rest of the proof of Proposition 4.5 closely follows the lines of the proof of Proposition 4.4.

## 5 Extension to the non-stationary case

In Sections 3 and 4, the stationarity assumption of the process  $(Y_u : u \in T)$  played a crucial role. In this section, we extend the strong consistency and the asymptotic normality of the MLE for the HMT to the case where this process is not stationary.

Hence, we assume that the HMT process  $(X', Y') = ((X'_u, Y'_u) : u \in T)$  has the same transition kernel  $Q_{\theta^*}$  and  $G_{\theta^*}$  that are parametrized by some  $\theta^* \in \Theta$  as before, and the hidden variable  $X'_{\partial}$  of the root vertex  $\partial$  has distribution  $\zeta$ . This initial distribution  $\zeta$  is unknown to us, may depend on  $\theta^*$ , and in general is different from the invariant distribution  $\pi_{\theta^*}$ . As before, we will denote by  $(X, Y) = ((X_u, Y_u) : u \in T)$  a stationary process distributed according to the same parameter  $\theta^*$ . Note that, in this section, we will use the convention that objects with an added  $'$  symbol are related to the non-stationary process  $(X', Y')$ , while those without the  $'$  symbol are their counterpart for the stationary process  $(X, Y)$ . Also note that due to the non-stationarity assumption, in this section, we will only consider the HMT process on the original tree  $T^{\infty} = T(\partial)$ .

For the non-stationary process  $(X', Y')$ , similarly to the stationary case in (7) on page 11, define its log-likelihood for all  $n \in \mathbb{N}$  and  $x \in \mathcal{X}$  as:

$$\ell'_{n,x}(\theta) := \ell_{n,x}(\theta; Y'_{T_n}), \quad (79)$$

where  $\ell_{n,x}(\theta; \cdot)$  is defined in (6) on page 11. Moreover, when Assumptions 1-4 and 6 hold and  $\Theta$  is compact, similarly to the stationary case in (33) on page 22, we define the MLE  $\hat{\theta}'_{n,x}$  for the non-stationary process as:

$$\hat{\theta}'_{n,x} = \hat{\theta}'_{n,x}(Y'_{T_n}) \in \operatorname{argmax}_{\theta \in \Theta} \ell'_{n,x}(\theta). \quad (80)$$

Denote by  $\mathbb{P}_{\theta^*, \zeta}$  the probability distribution of the non-stationary HMT process  $(X', Y')$ , and by  $\mathbb{E}_{\theta^*, \zeta}$  the corresponding expectation.

We can now prove the strong consistency of the MLE for a non-stationary HMT process.

**Theorem 5.1** (Strong consistency of the MLE, non-stationary case). *Assume that Assumptions 2–6 hold. the contrast function  $\ell$  has a unique maximum (which is then located at  $\theta^* \in \Theta$  by Proposition 3.10) and  $\Theta$  is compact. Then, the MLE is strongly consistent, that is, for all initial distributions  $\zeta$  and all  $x \in \mathcal{X}$ , the MLE  $\hat{\theta}'_{n,x}$  converges  $\mathbb{P}_{\theta^*, \zeta}$ -a.s. as  $n \rightarrow \infty$  to the true parameter  $\theta^* \in \Theta$ .*

*Proof.* We start by proving that for any  $n \in \mathbb{N}^*$ , the distribution of the non-stationary HMT process  $(X', Y')$  on  $T^*$  is absolutely continuous w.r.t. the distribution of the stationary HMT process  $(X, Y)$  on  $T^*$ , that is:

$$\mathbb{P}_{\theta^*, \zeta}(X'_{T^*} \in \cdot, Y'_{T^*} \in \cdot) \ll \mathbb{P}_{\theta^*, \pi_{\theta^*}}(X'_{T^*} \in \cdot, Y'_{T^*} \in \cdot) = \mathbb{P}_{\theta^*}(X_{T^*} \in \cdot, Y_{T^*} \in \cdot). \quad (81)$$

Remind that Assumption 3-(i) implies that  $\pi_{\theta^*} \ll \lambda$  with density  $\frac{d\pi_{\theta^*}}{d\lambda}$  taking value in  $[\sigma^-, \sigma^+]$ . Denote by  $u_1$  and  $u_2$  the two children vertices of  $\partial$ . Using Assumption 3, for any non-negative measurable function  $f$  from  $\mathcal{X}^2$  to  $\mathbb{R}_+$ , we get:

$$\begin{aligned} \int_{\mathcal{X}^2} f(x_{u_1}, x_{u_2}) \mathbb{P}_{\theta^*, \zeta}(X'_{u_1} \in dx_{u_1}, X'_{u_2} \in dx_{u_2}) \\ &= \int_{\mathcal{X}^3} f(x_{u_1}, x_{u_2}) q_{\theta^*}(x_{\partial}, x_{u_1}) q_{\theta^*}(x_{\partial}, x_{u_2}) \lambda(dx_{u_1}) \lambda(dx_{u_2}) \zeta(dx_{\partial}) \\ &\leq \left(\frac{\sigma^+}{\sigma^-}\right)^2 \int_{\mathcal{X}^2} f(x_{u_1}, x_{u_2}) \pi_{\theta^*}(dx_{u_1}) \pi_{\theta^*}(dx_{u_2}) \\ &= \left(\frac{\sigma^+}{\sigma^-}\right)^2 \int_{\mathcal{X}^2} f(x_{u_1}, x_{u_2}) \mathbb{P}_{\theta^*}(X_{u_1} \in dx_{u_1}, X_{u_2} \in dx_{u_2}). \end{aligned}$$

In particular, for any measurable subset  $A$  of  $\mathcal{X}^{T^*} \times \mathcal{Y}^{T^*}$ , we can choose  $f$  to be define as:

$$\begin{aligned} f(x_{u_1}, x_{u_2}) &= \mathbb{E}_{\theta^*, \zeta}[\mathbb{1}_A(X'_{T^*}, Y'_{T^*}) \mid X'_{u_1} = x_{u_1}, X'_{u_2} = x_{u_2}] \\ &= \mathbb{E}_{\theta^*}[\mathbb{1}_A(X_{T^*}, Y_{T^*}) \mid X_{u_1} = x_{u_1}, X_{u_2} = x_{u_2}] \end{aligned}$$

Hence, we get that (81) holds.

Using (81), we get that Proposition 3.7 also holds  $\mathbb{P}_{\theta^*, \zeta}$ -a.s. with  $\ell_{n,x}(\theta)$  replaced by  $\ell'_{n,x}(\theta)$ , that is, in the non-stationary case. Thus, the proof of Theorem 3.11 can be immediately adapted to the non-stationary case (note that Propositions 3.6 and 3.10 state properties of the contrast function  $\ell$ , which is the same in the stationary and non-stationary cases). This concludes the proof of the Theorem.  $\square$

Using a similar argument as for Theorem 5.1, we get that in the non-stationary case, the normalized observed information  $-|T_n|^{-1} \nabla_{\theta}^2 \ell'_{n,x}(\theta_n)$  converges  $\mathbb{P}_{\theta^*, \zeta}$ -a.s. locally uniformly to the limiting Fisher information  $\mathcal{I}(\theta^*)$  (which is defined in (54)). Note that the condition  $\rho < 1/2$  on the mixing rate  $\rho$  of the HMT process  $(X, Y)$  is inherited from Theorem 4.6 in the stationary case. See the discussion in Remark 1.5 for comments on this condition on  $\rho$ .

**Theorem 5.2** (Convergence of the normalized observed information, non-stationary case). *Assume that Assumptions 2–4 and 6–9 hold. Assume that  $\rho < 1/2$ . Assume that  $\Theta$  is compact. Then, for all initial distributions  $\zeta$  and all  $x \in \mathcal{X}$ , we have:*

$$\lim_{\delta \rightarrow 0} \lim_{n \rightarrow \infty} \sup_{\theta \in \mathcal{O} : \|\theta - \theta^*\| \leq \delta} \left\| -|T_n|^{-1} \nabla_{\theta}^2 \ell'_{n,x}(\theta) - \mathcal{I}(\theta^*) \right\| = 0 \quad \mathbb{P}_{\theta^*, \zeta}\text{-a.s.}$$

In particular, combining Theorems 5.1 and 5.2, we get that the normalized observed information  $-|T_n|^{-1}\nabla_{\theta}^2\ell_{n,x}(\hat{\theta}_{n,x})$  at the MLE  $\hat{\theta}_{n,x}$  is a strongly consistent estimator of the Fisher information matrix  $\mathcal{I}(\theta^*)$ .

Before proving the asymptotic normality of the MLE in the non-stationary case, we start with the following lemma which present a coupling construction for the two processes  $(X, Y)$  and  $(X', Y')$ .

**Lemma 5.3** (Coupling construction of two HMTs). *Assume that Assumptions 2 and 3 hold. Further assume that  $\sigma^- \geq 1/2$ . Then, it is possible to construct the two processes  $(X, Y)$  and  $(X', Y')$  on a common probability space such that there exists an a.s. finite random time  $N$ , which we call the coupling time, such that  $(X_u, Y_u) = (X'_u, Y'_u)$  for all  $u \in T_n$  with  $n \geq N$ .*

We will denote by  $\mathbb{P}_{\theta^* \bowtie \zeta}$  the probability distribution that realizes this coupling. Note that  $\rho \leq 1/2$  implies that  $\sigma^- \geq \sigma^+/2 \geq 1/2$  (see Assumption 3).

*Proof.* We first construct the coupling only for the process  $X$  and  $X'$ . We define the coupling construction inductively on the height of the tree. For the root vertex, we use an independent coupling construction for  $X_{\partial}$  and  $X'_{\partial}$ , which are distributed according to  $\pi_{\theta^*}$  and  $\zeta$  respectively (note that it is also possible to use a perfect coupling with probability error  $\|\pi_{\theta^*} - \zeta\|_{\text{TV}}$ ). Then, if the coupling has been constructed up to generation  $n \in \mathbb{N}$ , using the Markov property, we proceed to construct independently the coupling for each vertices in generation  $n+1$ . Let  $u \in G_{n+1}$ . If the variables were already coupled for the parent vertex  $p(u)$ , that is  $X_{p(u)} = X'_{p(u)}$ , then we choose the new value  $X_u = X'_u$  according to the transition kernel  $Q_{\theta^*}$ . Otherwise, if  $X_{p(u)} \neq X'_{p(u)}$ , then using the uniform geometric ergodicity (remind Assumption 3) of the transition kernel  $Q_{\theta^*}$ , we know that  $\sup_{x, x' \in \mathcal{X}} \|Q_{\theta^*}(x; \cdot) - Q_{\theta^*}(x'; \cdot)\|_{\text{TV}} \leq 1 - \sigma^-$ , and thus we can construct a coupling of  $X_u$  and  $X'_u$  conditionally on  $X_{p(u)} \neq X'_{p(u)}$  with exact matching probability at least  $1 - \sigma^-$ . We have constructing the matching for  $u$ , and thus for the whole generation  $n+1$ . Using Kolmogorov extension theorem, there exists a coupling measure for the whole tree  $T$  whose finite dimensional marginals are the ones given above.

Remark that the joint process  $(X, Y)$  satisfies a uniform geometric ergodicity bound with the same constant  $1 - \sigma^-$ . Thus, the construction above can be extended to the joint process  $(X, Y)$ . Denote by  $\mathbb{P}_{\theta^* \bowtie \zeta}$  the probability distribution of the coupling we have constructed for the joint process  $(X, Y)$ .

Define the random coupling time  $N = \inf\{n \in \mathbb{N} | \forall u \in G_n, X_u = X'_u\}$ , which is the first generation for which the exact coupling occurs for all vertices (and  $N = \infty$  is this never happens). We are left to prove that  $\mathbb{P}_{\theta^* \bowtie \zeta}$ -a.s.  $N < \infty$ . We say that a vertex  $u$  is a special vertex if  $X_u \neq X'_u$ . Note that if  $u$  is not special, then all its descendants are also not special. Also note that special vertices form a Bienaymé-Galton-Watson tree whose (homogeneous) offspring distribution takes the values: 0 with probability  $(\sigma^-)^2$ ; 1 with probability  $2\sigma^-(1 - \sigma^-)$ ; and 2 with probability  $(1 - \sigma^-)^2$ . The average of this offspring distribution is  $2(1 - \sigma^-)$ . Hence, the number of special vertices is finite  $\mathbb{P}_{\theta^* \bowtie \zeta}$ -a.s., that is,  $N$  is finite  $\mathbb{P}_{\theta^* \bowtie \zeta}$ -a.s., if and only if  $2(1 - \sigma^-) \leq 1$ , that is,  $\sigma^- \geq 1/2$ . This concludes the proof.  $\square$

Remind that the log-likelihood function  $\ell_{n,x}$  (resp.  $\ell'_{n,x}$ ), which is a random function depending on  $Y_{T_n}$  from the stationary HMT process (resp. on  $Y'_{T_n}$  from the non-stationary HMT process), is defined in (7) on page 11 (resp. just before (79) on page 50). For all  $\theta \in \Theta$ , define:

$$\begin{aligned} D_{n,x}(\theta) &= \ell'_{n,x}(\theta) - \ell_{n,x}(\theta) \\ &= \sum_{u \in T_n} \log p_{\theta}(Y'_u | Y'_{\Delta^*(u, h(u))}, X'_{\partial} = x) - \log p_{\theta}(Y_u | Y_{\Delta^*(u, h(u))}, X_{\partial} = x), \end{aligned}$$

where remind that  $p_\theta$  denotes possibly conditional density (see (5) on page 11).

Remind that when Assumptions 1-4 and 6 hold and  $\Theta$  is compact, for all  $x \in \mathcal{X}$ , the MLE  $\hat{\theta}_{n,x}$  (resp.  $\hat{\theta}'_{n,x}$ ) is a random variable which depends on  $Y_{T_n}$  from the stationary HMT process (resp. on  $Y'_{T_n}$  from the non-stationary HMT process) and is defined in (33) on page 22 (resp. in (80) on page 50).

To prove that  $\lim_{n \rightarrow \infty} |T_n|^{1/2}(\hat{\theta}'_{n,x} - \hat{\theta}_{n,x}) = 0$   $\mathbb{P}_{\theta^* \triangleright \zeta}$ -a.s., and thus that  $\hat{\theta}'_{n,x}$  and  $\hat{\theta}_{n,x}$  are asymptotically normal with the same covariance matrix (remind Theorem 4.7), we must first prove that the function  $\theta \mapsto D_{n,x}(\theta)$  satisfies some kind of continuity property. Note that we proved in the proof of Theorem 5.1 that Proposition 3.7 holds both in the stationary and the non-stationary cases, and thus we have:

$$\lim_{n \rightarrow \infty} \sup_{\theta \in \Theta} \left| |T_n|^{-1} D_{n,x}(\theta) \right| = 0 \quad \mathbb{P}_{\theta^* \triangleright \zeta}\text{-a.s.}$$

However, we need some kind of continuity property without the normalizing term  $|T_n|^{-1}$ , which is given by the following lemma.

**Lemma 5.4.** *Assume that Assumptions 2-6 hold. Further assume that  $\rho < 1/2$ . Then, for all initial distributions  $\zeta$  and all  $x \in \mathcal{X}$ , we have:*

$$\lim_{n \rightarrow \infty} |D_{n,x}(\hat{\theta}'_{n,x}) - D_{n,x}(\hat{\theta}_{n,x})| = 0, \quad \mathbb{P}_{\theta^* \triangleright \zeta}\text{-a.s.}$$

*Proof.* [The proof is a straightforward adaptation of the proof of [DMR04, Lemmas 11 and 12].]

Let  $N$  be the random time provided by Lemma 5.3. We first prove that  $\mathbb{P}_{\theta^* \triangleright \zeta}$ -a.s., we have:

$$\sum_{u \in T \setminus T_N} \sup_{\theta \in \Theta} |\log p_\theta(Y'_u | Y'_{\Delta^*(u,h(u))}, X'_\partial = x) - \log p_\theta(Y_u | Y_{\Delta^*(u,h(u))}, X_\partial = x)| < \infty. \quad (82)$$

Note that for all  $u \in T_n$  and  $v$  an ancestor of  $u$  (distinct of  $u$ ), we have:

$$\begin{aligned} & p_\theta(Y_u | Y_{\Delta^*(u,h(u))}, X_\partial = x) \\ &= \int_{\mathcal{X}^3} g_\theta(x_u, Y_u) q_\theta(x_{p(u)}, x_u) \lambda(dx_u) \mathbb{P}_\theta(X_{p(u)} \in dx_{p(u)} | X_v = x_v, Y_{\Delta^*(u,h(u)-h(v))}) \\ & \quad \times \mathbb{P}_\theta(X_v \in dx_v | Y_{\Delta^*(u,h(u))}, X_\partial = x), \end{aligned}$$

and similarly for  $p_\theta(Y'_u | Y'_{\Delta^*(u,h(u))}, X'_\partial = x)$ . Using the fact that for  $v \in T$  with height  $h(v) \geq N$ , we have  $Y_v = Y'_v$ , and using Lemma 3.2, we have for all  $u \in T$  with height  $h(u) > N$ :

$$\begin{aligned} & |p_\theta(Y'_u | Y'_{\Delta^*(u,h(u))}, X'_\partial = x) - p_\theta(Y_u | Y_{\Delta^*(u,h(u))}, X_\partial = x)| \\ & \leq 2\rho^{h(u)-N-1} \sigma^+ \int g_\theta(x, Y_u) \lambda(dx). \end{aligned}$$

Thus, using a similar argument as in the proof of Lemma 3.3, we get:

$$|\log p_\theta(Y'_u | Y'_{\Delta^*(u,h(u))}, X'_\partial = x) - \log p_\theta(Y_u | Y_{\Delta^*(u,h(u))}, X_\partial = x)| \leq \frac{\rho^{h(u)-N-1}}{1-\rho}.$$

Hence, the sum in (82) is  $\mathbb{P}_{\theta^* \triangleright \zeta}$ -a.s. upper bounded by a constant times  $\sum_{k=N+1}^{\infty} 2^k \rho^k \leq (2\rho)^{N+1}/(1-2\rho)$ , and is thus finite  $\mathbb{P}_{\theta^* \triangleright \zeta}$ -a.s.

Let  $\varepsilon > 0$ . Using (82), there exists a random integer  $N_\varepsilon$  which  $\mathbb{P}_{\theta^* \triangleright \zeta}$ -a.s. is finite and satisfies:

$$\sum_{u \in T \setminus T_{N_\varepsilon}} \sup_{\theta \in \Theta} |\log p_\theta(Y'_u | Y'_{\Delta^*(u,h(u))}, X'_\partial = x) - \log p_\theta(Y_u | Y_{\Delta^*(u,h(u))}, X_\partial = x)| \leq \varepsilon.$$

Thus,  $\mathbb{P}_{\theta^* \times \zeta}$ -a.s., for all  $n \geq N_\varepsilon$ , we have:

$$|D_{n,x}(\hat{\theta}'_{n,x}) - D_{n,x}(\hat{\theta}_{n,x})| \leq 2\varepsilon + |\ell'_{N_\varepsilon,x}(\hat{\theta}'_{n,x}) - \ell'_{N_\varepsilon,x}(\hat{\theta}_{n,x})| + |\ell_{N_\varepsilon,x}(\hat{\theta}'_{n,x}) - \ell_{N_\varepsilon,x}(\hat{\theta}_{n,x})|.$$

Note that under the given assumptions, the functions  $\theta \mapsto \ell'_{N_\varepsilon,x}(\theta)$  and  $\theta \mapsto \ell_{N_\varepsilon,x}(\theta)$  are continuous  $\mathbb{P}_{\theta^* \times \zeta}$ -a.s. (see the proof of Proposition 3.6). Hence, the proof is complete upon observing that  $\hat{\theta}'_{n,x}$  and  $\hat{\theta}_{n,x}$  both converge  $\mathbb{P}_{\theta^* \times \zeta}$ -a.s. to  $\theta^*$  (see Theorem 5.1), and that  $\varepsilon$  was arbitrary.  $\square$

We can now prove the asymptotic normality of the MLE  $\hat{\theta}'_{n,x}$  in the non-stationary case. Remind that the contrast function  $\ell$  is defined in (26) on page 19.

**Theorem 5.5** (Asymptotic normality of the MLE, non-stationary case). *Assume that Assumptions 2–9 hold. Assume that  $\rho < 1/2$ . Further assume that the contrast function  $\ell$  has a unique maximum (which is then located at  $\theta^* \in \Theta$  by Proposition 3.10) and that  $\Theta$  is compact,  $\theta^*$  is an interior point of  $\Theta$ , and the limiting Fisher information matrix  $\mathcal{I}(\theta^*)$  (which is defined in (54)) is non-singular. Then, for all initial distributions  $\zeta$  and for all  $x \in \mathcal{X}$ , we have:*

$$|T_n|^{1/2}(\hat{\theta}'_{n,x} - \theta^*) \xrightarrow[n \rightarrow \infty]{(d)} \mathcal{N}(0, \mathcal{I}(\theta^*)^{-1}) \quad \text{under } \mathbb{P}_{\theta^*, \zeta},$$

where  $\mathcal{N}(0, M)$  denotes the centered Gaussian distribution with covariance matrix  $M$ .

*Proof.* [The proof is a straightforward adaptation of the proof of [DMR04, Theorem 6].]

Define  $\varepsilon_n = |T_n|^{1/2}(\hat{\theta}_{n,x} - \hat{\theta}'_{n,x})$  for all  $n \in \mathbb{N}$ , and remark that it is sufficient to prove that  $\lim_{n \rightarrow \infty} \varepsilon_n = 0$   $\mathbb{P}_{\theta^* \times \zeta}$ -a.s. Since  $\hat{\theta}'_{n,x}$  is the maximizer of the function  $\theta \mapsto \ell'_{n,x}(\theta)$ , we have that  $\ell'_{n,x}(\hat{\theta}'_{n,x}) \geq \ell'_{n,x}(\hat{\theta}_{n,x})$ . Thus, using a Taylor expansion of  $\ell_{n,x}$  around its maximizer  $\hat{\theta}_{n,x}$  (for which we have  $\nabla_{\theta} \ell_{n,x}(\hat{\theta}_{n,x}) = 0$ ), we get that there exists  $t_n \in [0, 1]$  such that:

$$\begin{aligned} D_{n,x}(\hat{\theta}'_{n,x}) - D_{n,x}(\hat{\theta}_{n,x}) &\geq \ell_{n,x}(\hat{\theta}_{n,x}) - \ell_{n,x}(\hat{\theta}'_{n,x}) \\ &= -\frac{1}{2}|T_n|^{-1} \varepsilon_n^t \nabla_{\theta}^2 \ell_{n,x}(t_n \hat{\theta}'_{n,x} + (1-t_n) \hat{\theta}_{n,x}) \varepsilon_n. \end{aligned}$$

Note that we have  $\lim_{n \rightarrow \infty} t_n \hat{\theta}'_{n,x} + (1-t_n) \hat{\theta}_{n,x} = \theta^*$   $\mathbb{P}_{\theta^* \times \zeta}$ -a.s. by Theorem 5.1. Thus, applying Theorem 4.6, we have:

$$\lim_{n \rightarrow \infty} -|T_n|^{-1} \nabla_{\theta}^2 \ell_{n,x}(t_n \hat{\theta}'_{n,x} + (1-t_n) \hat{\theta}_{n,x}) = \mathcal{I}(\theta^*), \quad \mathbb{P}_{\theta^* \times \zeta}\text{-a.s.}$$

As  $\mathcal{I}(\theta^*)$  is positive definite, there exists  $M > 0$  such that on a set with  $\mathbb{P}_{\theta^* \times \zeta}$ -probability one and for  $n$  sufficiently large, we have:

$$D_{n,x}(\hat{\theta}'_{n,x}) - D_{n,x}(\hat{\theta}_{n,x}) \geq M|\varepsilon_n|^2.$$

Then, the proof is complete by applying Lemma 5.4.  $\square$

## References

- [Ath69] Krishna Balasundaram Athreya. Limit theorems for multitype continuous time markov branching processes: I. The case of an eigenvector linear functional. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 12(4):320–332, 1969.

- [BDSG09] Bernard Bercu, Benoîte De Saporta, and Anne Gégout-Petit. Asymptotic Analysis for Bifurcating AutoRegressive Processes via a Martingale Approach. *Electronic Journal of Probability*, 14(none), January 2009.
- [BFA22] Nizar Bouguila, Wentao Fan, and Manar Amayri, editors. *Hidden Markov Models and Applications*. Unsupervised and Semi-Supervised Learning. Springer International Publishing, Cham, 2022.
- [Bil99] Patrick Billingsley. *Convergence of Probability Measures*. Wiley Series in Probability and Statistics. Probability and Statistics Section. Wiley, New York, 2nd ed edition, 1999.
- [BP66] Leonard E. Baum and Ted Petrie. Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *The Annals of Mathematical Statistics*, 37(6):1554–1563, December 1966.
- [BPD22a] S. Valère Bitseki Penda and Jean-François Delmas. Central limit theorem for bifurcating Markov chains under pointwise ergodic conditions. *The Annals of Applied Probability*, 32(5), October 2022.
- [BPD22b] S. Valère Bitseki Penda and Jean-François Delmas. Central Limit Theorem for Kernel Estimator of Invariant Density in Bifurcating Markov Chains Models. *Journal of Theoretical Probability*, October 2022.
- [BPDG14] S. Valère Bitseki Penda, Hacène Djellout, and Arnaud Guillin. Deviation inequalities, moderate deviations and some limit theorems for bifurcating Markov chains with application. *The Annals of Applied Probability*, 24(1), February 2014.
- [BPSW70] Leonard E. Baum, Ted Petrie, George Soules, and Norman Weiss. A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *The Annals of Mathematical Statistics*, 41(1):164–171, February 1970.
- [BRR98] Peter J. Bickel, Ya’acov Ritov, and Tobias Rydén. Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. *The Annals of Statistics*, 26(4), August 1998.
- [BS96] Dimitri P. Bertsekas and Steven E. Shreve. *Stochastic Optimal Control: The Discrete Time Case*. Optimization and Neural Computation Series. Athena Scientific, Belmont, Mass, 1996.
- [BWX13] Jacob Biesinger, Yuanfeng Wang, and Xiaohui Xie. Discovering and mapping chromatin states using a tree hidden Markov model. *BMC Bioinformatics*, 14(S5):S4, April 2013.
- [CB01] Hyeokho Choi and Richard G. Baraniuk. Multiscale image segmentation using wavelet-domain hidden Markov models. *IEEE Transactions on Image Processing*, 10(9):1309–1321, September 2001.
- [CMR05] Olivier Cappé, Éric Moulines, and Tobias Rydén. *Inference in Hidden Markov Models*. Springer Series in Statistics. Springer New York, New York, NY, 2005.
- [CNB98] Matt S. Crouse, Robert D. Nowak, and Richard G. Baraniuk. Wavelet-based statistical signal processing using hidden Markov models. *IEEE Transactions on Signal Processing*, 46(4):886–902, April 1998.



- [DGCC05] Jean-Baptiste Durand, Yann Guédon, Yves Caraglio, and Evelyne Costes. Analysis of the plant architecture via tree-structured statistical models: The hidden Markov tree models. *New Phytologist*, 166(3):813–825, June 2005.
- [DGG04] Jean-Baptiste Durand, Paulo Gonçalves, and Yann Guédon. Computational Methods for Hidden Markov Tree Models—An Application to Wavelet Trees. *IEEE Transactions on Signal Processing*, 52(9):2551–2560, September 2004.
- [DLR77] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum Likelihood from Incomplete Data Via the *EM* Algorithm. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 39(1):1–22, September 1977.
- [DM01] Randal Douc and Catherine Matias. Asymptotics of the Maximum Likelihood Estimator for General Hidden Markov Models. *Bernoulli*, 7(3):381, June 2001, 3318493.
- [DM10] Jean-François Delmas and Laurence Marsalle. Detection of cellular aging in a Galton–Watson process. *Stochastic Processes and their Applications*, 120(12):2495–2519, December 2010.
- [DMOVH11] Randal Douc, Eric Moulines, Jimmy Olsson, and Ramon Van Handel. Consistency of the maximum likelihood estimator for general hidden Markov models. *The Annals of Statistics*, 39(1), February 2011.
- [DMPS18] Randal Douc, Éric Moulines, Pierre Priouret, and Philippe Soulier. *Markov Chains*. Springer Series in Operations Research and Financial Engineering. Springer International Publishing, Cham, 2018.
- [DMR04] Randal Douc, Éric Moulines, and Tobias Rydén. Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. *The Annals of Statistics*, 32(5), October 2004.
- [DRS16] Randal Douc, François Roueff, and Tepmony Sim. The maximizing set of the asymptotic normalized log-likelihood for partially observed Markov chains. *The Annals of Applied Probability*, 26(4), August 2016.
- [Duf11] Marie Duflo. *Random Iterative Models*. Springer, Berlin, 2011.
- [DWB08] Marco F. Duarte, Michael B. Wakin, and Richard G. Baraniuk. Wavelet-domain compressive signal reconstruction using a Hidden Markov Tree model. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5137–5140, Las Vegas, NV, USA, March 2008. IEEE.
- [GL06] Valentine Genon-Catalot and Catherine Laredo. Leroux’s method for general hidden Markov models. *Stochastic Processes and their Applications*, 116(2):222–243, February 2006.
- [GOB13] Édouard Grave, Guillaume Obozinski, and Francis Bach. Hidden Markov tree models for semantic class induction. In *CoNLL*, volume Proceedings of the Seventeenth Conference on Computational Natural Language Learning, pages 94–103, Sofia, Bulgaria, 2013. Association for Computational Linguistics.
- [Guy07] Julien Guyon. Limit theorems for bifurcating Markov chains. Application to the detection of cellular aging. *The Annals of Applied Probability*, 17(5-6):1538–1569, October 2007, 0710.5434.

- [HBSLLB<sup>+</sup>17] Houda Hanzouli-Ben Salah, Jerome Lapuyade-Lahorgue, Julien Bert, Didier Benoit, Philippe Lambin, Angela Van Baardwijk, Emmanuel Monfrini, Wojciech Pieczynski, Dimitris Visvikis, and Mathieu Hatt. A framework based on hidden Markov trees for multimodal image co-segmentation. *Medical Physics*, 44(11):5835–5848, November 2017.
- [HYG17] Kai Hu, Wei Yang, and Xieping Gao. Microcalcification diagnosis in digital mammography using extreme learning machine based on hidden Markov tree model of dual-tree complex wavelet transform. *Expert Systems with Applications*, 86:135–144, November 2017.
- [JP99] Jens Ledet Jensen and Niels Væver Petersen. Asymptotic normality of the maximum likelihood estimator in state space models. *The Annals of Statistics*, 27(2), April 1999.
- [KDM13] Shuhei Kondo, Kevin Duh, and Yuji Matsumoto. Hidden markov tree model for word alignment. In *WMT*, volume Proceedings of the Eighth Workshop on Statistical Machine Translation, pages 503–511, Sofia, Bulgaria, 2013. Association for Computational Linguistics.
- [Kos01] Timo Koski. *Hidden Markov Models for Bioinformatics*. Number 2 in Computational Biology. Kluwer academic publ, Dordrecht [etc.], 2001.
- [KS19] Hiroyuki Kasahara and Katsumi Shimotsu. Asymptotic properties of the maximum likelihood estimator in regime switching econometric models. *Journal of Econometrics*, 208(2):442–467, February 2019.
- [Ler92] Brian G. Leroux. Maximum-likelihood estimation for hidden Markov models. *Stochastic Processes and their Applications*, 40(1):127–143, February 1992.
- [LGM00] François Le Gland and Laurent Mevel. Exponential Forgetting and Geometric Ergodicity in Hidden Markov Models. *Mathematics of Control, Signals, and Systems*, 13(1):63–93, February 2000.
- [MBY<sup>+</sup>12] Mahender K. Makhijani, Niranjan Balu, Kiyofumi Yamada, Chun Yuan, and Krishna S. Nayak. Accelerated 3D MERGE carotid imaging using compressed sensing with a hidden markov tree model. *Journal of Magnetic Resonance Imaging*, 36(5):1194–1202, November 2012.
- [ME14] Rogemar S. Mamon and Robert J. Elliott, editors. *Hidden Markov Models in Finance: Further Developments and Applications, Volume II*, volume 209 of *International Series in Operations Research & Management Science*. Springer US, Boston, MA, 2014.
- [MW43] Henry B. Mann and Abraham Wald. On the Statistical Treatment of Linear Stochastic Difference Equations. *Econometrica*, 11(3/4):173, July 1943, 1905674.
- [NSK20] So Nakashima, Yuki Sughiyama, and Tetsuya J Kobayashi. Lineage EM algorithm for inferring latent states from cellular lineage trees. *Bioinformatics*, 36(9):2829–2838, May 2020.
- [OCB<sup>+</sup>09] Victor Olariu, Daniel Coca, Stephen A. Billings, Peter Tonge, Paul Gokhale, Peter W. Andrews, and Visakan Kadirkamanathan. Modified variational Bayes EM estimation of hidden Markov tree model of cell lineages. *Bioinformatics*, 25(21):2824–2830, November 2009.

- [Rab89] Lawrence R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.
- [RCBK00] Justin Romberg, Hyeokho Choi, Richard G. Baraniuk, and Nicholas Kingsbury. Multiscale classification using complex wavelets and hidden Markov tree models. In *Proceedings 2000 International Conference on Image Processing (Cat. No.00CH37101)*, pages 371–374 vol.2, Vancouver, BC, Canada, 2000. IEEE.
- [RS18] Dieter Rasch and Dieter Schott. *Mathematical Statistics*. John Wiley & Sons, Hoboken, 2018.
- [SH17] Hamid Reza Shahdoosti and Seyede Mahya Hazavei. Image denoising in dual contourlet domain using hidden Markov tree models. *Digital Signal Processing*, 67:17–29, August 2017.
- [Wal49] Abraham Wald. Note on the Consistency of the Maximum Likelihood Estimate. *The Annals of Mathematical Statistics*, 20(4):595–601, December 1949.
- [Wei24] Julien Weibel. Ergodic theorem for branching Markov chains indexed by trees with arbitrary shape, March 2024, arXiv, 2403.16505, [math].
- [XJS18] Miao Xie, Zhe Jiang, and Arpan Man Sainju. Geographical Hidden Markov Tree for Flood Extent Mapping. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2545–2554, London United Kingdom, July 2018. ACM.
- [YD15] Dong Yu and Li Deng. *Automatic Speech Recognition: A Deep Learning Approach*. Signals and Communication Technology. Springer London, London, 2015.
- [ZM09] Walter Zucchini and Iain L. MacDonald. *Hidden Markov Models for Time Series: An Introduction Using R*. Chapman and Hall/CRC, 0 edition, April 2009.

## A Ergodic theorem for Markov processes indexed by trees with neighborhood-dependent functions

In this appendix, we prove generalization of the ergodic theorems in [Guy07] and in [Wei24], which give a.s. and  $L^2$  convergences for branching Markov chains, to allow for neighborhood-dependent functions. Those ergodic theorems are used to prove the ergodic convergence lemmas in Section 2.4 which are used in the main body of this article. Remind that we need those generalization as in the study of asymptotic property of the MLE for the HMT relies on the study of the likelihood contribution functions  $h_{u,k,x}(\theta; Y_{\Delta(u,k)})$  (defined in (17) on page 16) which are neighborhood-dependent.

Remind from Section 2 that if  $(X, Y)$  is a HMT process, then the joint process  $((X_u, Y_u), u \in T)$  is a branching Markov chain. Thus, it is enough to prove those ergodic theorems for branching Markov chains instead of HMT processes.

Let  $Q$  be a transition kernel on  $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$  where  $\mathcal{X}$  is a metric space. We assume that  $Q$  has a unique invariant probability distribution  $\pi$  and is uniformly geometrically ergodic, that is, there exists  $\rho \in (0, 1)$  and a finite positive constant  $C$  such that for all  $x \in \mathcal{X}$ , we have  $\|Q^n(x; \cdot) - \pi\|_{TV} \leq C\rho^n$ . Remind from Lemma 2.3 that this covers the case  $Q = Q_\theta$  for any  $\theta \in \Theta$ . Let  $X = (X_u, u \in T)$  be a branching Markov chain with transition kernel  $Q$  and initial

distribution  $\pi$ . Denote by  $\mathbb{P}_Q$  the probability distribution of the process  $X$ , and by  $\mathbb{E}_Q$  the corresponding expectation.

In this section, for a probability measure  $\nu$  on  $\mathcal{X}$ , a transition kernel  $Q$  on  $(\mathcal{X}, \mathcal{B}(\mathcal{Z}))$  and a Borel integrable function  $f$  on  $\mathcal{B}(\mathcal{Z})$  where  $\mathcal{Z} = \mathcal{X}^A$  for some finite subset  $A \subset T$ , we will write  $\nu Q$  for the image probability measure  $(\nu Q)(\cdot) = \int_{\mathcal{X}} Q(x; \cdot) \nu(dx)$ , and  $Qf$  for the Borel function  $(Qf)(x) = \int_{\mathcal{Z}} f(z) Q(x; dz)$ . For a probability measure  $\nu$  on  $\mathcal{X}$  and a Borel integrable function  $f$  on  $\mathcal{X}$ , we will write  $\langle \nu, f \rangle = \nu f = \int_{\mathcal{X}} f d\nu$ .

We will need the following lemma which states geometric convergence bounds for functions in  $L^2(\pi)$ .

**Lemma A.1** (Convergence bounds when  $Q$  is uniformly geometrically ergodic). *Assume that the transition kernel  $Q$  has a unique invariant measure  $\pi$ , and that  $Q$  is uniformly geometrically ergodic. Then, there exists finite positive constants  $\alpha \in (0, 1)$  and  $M < \infty$  such that for all functions  $f \in L^2(\pi)$ , we have:*

$$\forall n \in \mathbb{N}, \quad \sup_{k \in \mathbb{N}} \pi Q^k (Q^n f - \langle \pi, f \rangle)^2 = \pi (Q^n f - \langle \pi, f \rangle)^2 \leq M \alpha^{2n} \|f - \langle \pi, f \rangle\|_{L^2(\pi)}^2.$$

In particular, the function  $f$  satisfies  $\sup_{n \in \mathbb{N}} \pi Q^n f^2 < \infty$ .

Note, using Cauchy-Schwarz and Jensen's inequalities, that  $\sup_{n \in \mathbb{N}} \pi Q^n f^2 < \infty$  implies that  $Q^n f$ ,  $Q^n f^2$ , and  $Q^k(Q^n f \times Q^m f)$  (with  $n, m, k \in \mathbb{N}$ ) are well-defined and finite  $\pi$ -almost everywhere and are  $\pi$ -integrable.

*Proof.* Using [DMPS18, Proposition 22.3.5 and Definition 22.3.1], we get that there exists finite positive constants  $\alpha \in (0, 1)$  and  $M < \infty$  such that for all functions  $f \in L^2(\pi)$ , we have  $\pi (Q^n f - \langle \pi, f \rangle)^2 = \|Q^n(f - \langle \pi, f \rangle)\|_{L^2(\pi)}^2 \leq M \alpha^{2n} \|f - \langle \pi, f \rangle\|_{L^2(\pi)}^2$  for all  $n \in \mathbb{N}$ . In particular, we get that  $\sup_{n \in \mathbb{N}} \pi Q^n f^2 < \infty$ .  $\square$

Let  $k \in \mathbb{N}$  be fixed. Remind from Section 2.4 the definitions of the subtrees  $\Delta(u, k)$ , of their shapes  $\mathcal{S}h(\Delta(u, k))$  (defined in (8)), and of the finite set of possible shapes  $\mathcal{N}_k$  (defined in (9)). For simplicity, in this appendix we will write  $\mathcal{S}_u$  instead of  $\mathcal{S}h(\Delta(u, k))$ . Also remind from Section 2.4 the definition of a collection of neighborhood-shape-dependent functions  $(f_{\mathcal{S}} : \mathcal{X}^{\mathcal{S}} \rightarrow \mathbb{R})_{\mathcal{S} \in \mathcal{N}_k}$ . Remind that for such a collection of functions, we simply write  $f_{\Delta(u, k)}$  or  $f_{\mathcal{S}_u}$  instead of  $f_{\mathcal{S}h(\Delta(u, k))}$ . And also remind that we write  $f_{\mathcal{S}_u}(X_{\Delta(u, k)})$  for the evaluation of  $f_{\mathcal{S}_u} = f_{\Delta(u, k)}$  on  $X_{\Delta(u, k)}$ . Note that up to translation, we may identify  $\mathcal{X}^{\mathcal{S}}$  and  $\mathcal{X}^{\Delta(u, k)}$  for any  $u \in T \setminus T_{k-1}$  such that  $\mathcal{S}_u = \mathcal{S}$ .

Remind that any subset  $A \subset T$ , we denote by  $X_A$  the gathered variables  $(X_v : v \in A)$ . For a collection of neighborhood-shape-dependent functions  $f = (f_{\mathcal{S}} : \mathcal{X}^{\mathcal{S}} \rightarrow \mathbb{R})_{\mathcal{S} \in \mathcal{N}_k}$ , define the empirical average of  $f$  over a finite subset  $A \subset T \setminus T_{k-1}$  as:

$$\bar{M}_A(f) = |A|^{-1} \sum_{u \in A} f_{\mathcal{S}_u}(X_{\Delta(u, k)}). \quad (83)$$

For a neighborhood shape  $\mathcal{S} \in \mathcal{N}_k$ , let  $u \in G_k$  be the unique vertex such that  $\mathcal{S}_u = \mathcal{S}$ , and define the transition kernel  $Q^{\mathcal{S}}$  on  $(\mathcal{X}, \mathcal{B}(\mathcal{X}^{\mathcal{S}}))$  for any  $x \in \mathcal{X}$  and any Borel function  $f$  on  $\mathcal{X}^{\mathcal{S}} = \mathcal{X}^{\Delta(u)}$  which is in  $L^1(X_{\Delta(u)}) = L^1(X_{\mathcal{S}})$  by:

$$Q^{\mathcal{S}} f(x) = \mathbb{E}_Q \left[ f(X_{\Delta(u)}) \mid X_{\emptyset} = x \right]. \quad (84)$$

That is, from the value  $x \in \mathcal{X}$  of the root vertex  $v$  in  $\mathcal{S}$ , the transition kernel  $Q^{\mathcal{S}}$  returns the distribution of the Markov process  $X$  on  $\mathcal{S}$  with transition kernel  $Q$  conditioned on the value

$X_v$  of the vertex  $v$  being  $x$ . Note that (84) also extends to any vertex  $u \in T \setminus T_{k-1}$  such that  $\mathcal{S}_u = \mathcal{S}$ , which gives us:

$$Q^{\mathcal{S}_u} f(x) = \mathbb{E}_Q \left[ f(X_{\Delta(u,k)}) \mid X_{\mathcal{P}^k(u)} = x \right]. \quad (85)$$

Moreover, using Jensen's inequality, note that if  $f \in L^2(X_{\mathcal{S}})$ , then  $Q^{\mathcal{S}} f$  is in  $L^2(\pi) = L^2(X_{\partial})$ .

Remind that as  $T$  is a plane rooted tree, we can enumerate its vertices as a sequence  $(v_j)_{j \in \mathbb{N}}$  in a breadth-first-search manner, that is, which is increasing for  $<$  (note that  $u_0 = \partial$ ). Also remind that, for  $n \geq |T_{k-1}|$ , if  $V_n$  is uniformly distributed over  $A_n := \{v_j : |T_{k-1}| < j \leq n\} = \Delta(v_n) \setminus T_{k-1}$ , then the distribution of  $\mathcal{S}_{V_n}$  converges to the uniform distribution over  $\mathcal{N}_k$  as  $n \rightarrow \infty$ .

We are now ready to state the ergodic theorem with neighborhood-shape-dependent functions for branching Markov chains indexed by the infinite complete binary tree  $T$ . Remind that  $\bar{M}_{A_n}(f)$  is defined in (83).

**Theorem A.2** (Ergodic theorem with neighborhood-dependent functions). *Let  $k \in \mathbb{N}$  be fixed. Let  $(v_j)_{j \in \mathbb{N}}$  be the sequence enumerating the vertices in  $T$  in a breadth-first-search manner. For all  $n > |T_{k-1}|$ , define  $A_n = \Delta(v_n) \setminus T_{k-1}$ .*

*Let  $Q$  be a transition kernel on  $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$  which is uniformly geometrically ergodic and has a unique invariant probability measure  $\pi$ . Let  $X = (X_u, u \in T)$  be a branching Markov chain with transition kernel  $Q$  and initial distribution  $\pi$ .*

*Let  $f = (f_{\mathcal{S}} : \mathcal{X}^{\mathcal{S}} \rightarrow \mathbb{R})_{\mathcal{S} \in \mathcal{N}_k}$  be a collection of neighborhood-shape-dependent Borel functions that are in  $L^2(X)$ . Then, we have:*

$$\bar{M}_{A_n}(f) \xrightarrow[n \rightarrow \infty]{} \mathbb{E}_{U_k} \otimes \mathbb{E}_Q [f_{\mathcal{S}_{U_k}}(X_{\Delta(U_k)})] \quad \text{in } L^2(\pi) = L^2(X), \quad (86)$$

where  $U_k$  is uniformly distributed over  $G_k$  and independent of the process  $X$ , and  $\mathbb{E}_{U_k} \otimes \mathbb{E}_Q$  denotes the joint expectation over  $U_k$  and  $X$ .

As an immediate corollary, we get that the result still holds if  $A_n$  is replaced by  $T_n \setminus T_{k-1}$ .

**Remark A.3** (More general assumptions). Note that without changing the proof, we could replace the subtrees  $\Delta(u, k)$  by general subtrees (i.e. a connected subsets)  $\mathcal{O}_u$  of the the  $k$ -neighborhood  $B_T(u, k) := \{v \in T : d(u, v) \leq k\}$  of  $u$  such that  $\mathcal{O}_u$  contains  $u$ . In that case, we must assume that the distribution of the shape (i.e. when seen up to translation)  $\mathcal{S}h(\mathcal{O}_{V_n})$  converges to some limit distribution. Also note that we could allow more general choices as in [Wei24] for the averaging sets  $(A_n)_{n \in \mathbb{N}}$ , for the tree  $T$ , and for the transition kernel  $Q$  and initial distribution of the branching Markov chain  $X$ .

*Proof. First case: we only have constant functions  $f_{\mathcal{S}} \equiv c(\mathcal{S})$  for all  $\mathcal{S} \in \mathcal{N}_k$ .* Then, for every  $n \in \mathbb{N}$  we have:

$$\bar{M}_{A_n}(f) = \sum_{\mathcal{S} \in \mathcal{N}_k} c(\mathcal{S}) \frac{|u \in A_n : \mathcal{S}_u = \mathcal{S}|}{|A_n|}, \quad (87)$$

where the right hand side converges in distribution (and thus in  $L^2$ ) as  $n \rightarrow \infty$  to  $\mathbb{E}_{U_k} [c(\mathcal{S}_{U_k})]$  (remind that the distribution of  $\mathcal{S}_{V_n}$  converges to the uniform distribution on  $G_k$  when  $n \rightarrow \infty$ ). This concludes the proof in this first case.

**General case:** We adapt the proof of [Wei24, Theorem 2.2] to the case of neighborhood-shape-dependent functions.

Using the first case and the linearity in  $f$  of the empirical averages, and replacing  $f_{\mathcal{S}}$  by  $f_{\mathcal{S}} - \langle \pi, Q^{\mathcal{S}} f_{\mathcal{S}} \rangle$ , we may assume that  $\langle \pi, Q^{\mathcal{S}} f_{\mathcal{S}} \rangle = 0$  for all  $\mathcal{S} \in \mathcal{N}_k$ . For all  $n \in \mathbb{N}$ , we have:

$$\mathbb{E}_Q \left[ \bar{M}_{A_n}(f)^2 \right] = \frac{1}{|A_n|^2} \sum_{u,v \in A_n} \mathbb{E}_Q \left[ f_{\mathcal{S}_u}(X_{\Delta(u,k)}) f_{\mathcal{S}_v}(X_{\Delta(v,k)}) \right]. \quad (88)$$

Using Lemma A.1, as the transition kernel  $Q$  is uniformly geometrically ergodic and has unique invariant probability measure  $\pi$ , and as the function  $Q^{\mathcal{S}} f_{\mathcal{S}}$  for  $\mathcal{S} \in \mathcal{N}_k$  are all in  $L^2(\pi)$  (see the comment just after (85)), we have that  $C_{\mathcal{S}} := \sup_{n \in \mathbb{N}} \pi Q^n (Q^{\mathcal{S}} f_{\mathcal{S}})^2 < \infty$  for all  $\mathcal{S} \in \mathcal{N}_k$ . Define  $C := \max_{\mathcal{S} \in \mathcal{N}_k} C_{\mathcal{S}} < \infty$  (remind that  $\mathcal{N}_k$  is finite). Thus, for  $u \in T$ , we have:

$$\mathbb{E}_Q \left[ f_{\mathcal{S}_u}(X_{\Delta(u,k)})^2 \right] = \pi Q^{(h(u)-k)_+} (Q^{\mathcal{S}_u} f_{\mathcal{S}_u})^2 \leq C < \infty.$$

Hence, for all  $u, v \in T$ , using Cauchy-Schwarz inequality, we have:

$$\mathbb{E}_Q \left[ f_{\mathcal{S}_u}(X_{\Delta(u,k)}) f_{\mathcal{S}_v}(X_{\Delta(v,k)}) \right] \leq \left( \mathbb{E}_Q \left[ f_{\mathcal{S}_u}(X_{\Delta(u,k)})^2 \right] \mathbb{E}_Q \left[ f_{\mathcal{S}_v}(X_{\Delta(v,k)})^2 \right] \right)^{1/2} \leq C < \infty. \quad (89)$$

Let  $u, v \in T$  such that  $d(u, v) > 2k$ , which implies that  $\Delta(u, k) \cap \Delta(v, k) = \emptyset$ . Without loss of generality, assume that  $h(u) \geq h(v)$ . Then, we have that  $h(u \wedge v) < h(u) - k$ . Denote by  $v_0$  the last ancestor of  $u$  in  $\Delta(v, k) \cup \{u \wedge v\}$ . Remark that  $u \wedge v$  is an ancestor of  $v_0$ . Then, we have:

$$\begin{aligned} \mathbb{E}_Q \left[ f_{\mathcal{S}_u}(X_{\Delta(u,k)}) f_{\mathcal{S}_v}(X_{\Delta(v,k)}) \right] &= \mathbb{E}_Q \left[ \mathbb{E}_Q \left[ f_{\mathcal{S}_u}(X_{\Delta(u,k)}) \mid X_{\Delta(v,k)}, X_{u \wedge v} \right] f_{\mathcal{S}_v}(X_{\Delta(v,k)}) \right] \\ &\leq \left( \mathbb{E}_Q \left[ \mathbb{E}_Q \left[ f_{\mathcal{S}_u}(X_{\Delta(u,k)}) \mid X_{\Delta(v,k)}, X_{u \wedge v} \right]^2 \right] \mathbb{E}_Q \left[ f_{\mathcal{S}_v}(X_{\Delta(v,k)})^2 \right] \right)^{1/2} \\ &\leq C^{1/2} \left( \mathbb{E}_Q \left[ \mathbb{E}_Q \left[ f_{\mathcal{S}_u}(X_{\Delta(u,k)}) \mid X_{v_0} \right]^2 \right] \right)^{1/2} \\ &\leq C^{1/2} \left( \mathbb{E}_Q \left[ \mathbb{E}_Q \left[ f_{\mathcal{S}_u}(X_{\Delta(u,k)}) \mid X_{u \wedge v} \right]^2 \right] \right)^{1/2} \\ &= C^{1/2} \left( \pi Q^{h(u \wedge v)} \left( Q^{d(u \wedge v, u) - k} Q^{\mathcal{S}_u} f_{\mathcal{S}_u} \right)^2 \right)^{1/2} \\ &\leq C^{1/2} \left( \max_{\mathcal{S} \in \mathcal{N}_k} \pi Q^{h(u \wedge v)} \left( Q^{\tilde{d}(u, v) - k} Q^{\mathcal{S}} f_{\mathcal{S}} \right)^2 \right)^{1/2}, \end{aligned} \quad (90)$$

where we used Cauchy-Schwarz inequality in the first inequality, we used (89) and the Markov property of the process  $X$  in the second inequality, and we used Jensen's inequality in the third inequality. Remark that  $\tilde{d}(u, v) = \max(d(u \wedge v, u), d(u \wedge v, v))$  is a distance on  $T$  that satisfies  $d/2 \leq \tilde{d} \leq d$ .

Let  $U_n$  and  $V_n$  be uniformly distributed over  $A_n$ , and independent of each other and of the branching Markov chain  $X$ , and denote by  $\mathbb{P}_{U_n, V_n}$  their joint probability distribution. Using again Lemma A.1, there exist finite constants  $M < \infty$  and  $\alpha \in (0, 1)$  such that for all  $\mathcal{S} \in \mathcal{N}$  we have

$$\forall m \in \mathbb{N}, \quad \sup_{j \in \mathbb{N}} \pi Q^j (Q^m Q^{\mathcal{S}} f_{\mathcal{S}})^2 \leq M^2 \alpha^{2m}. \quad (91)$$

Hence, combining (88), (89), (90) and (91), we get for any  $K \geq k$ :

$$\begin{aligned}
\mathbb{E}_Q \left[ \bar{M}_{A_n}(f)^2 \right] &\leq C \mathbb{P}_{U_n, V_n}(\tilde{d}(U_n, V_n) \leq 2K) \\
&\quad + C^{1/2} |A_n|^{-2} \sum_{u, v \in A_n: \tilde{d}(u, v) > 2K} \left( \max_{\mathcal{S} \in \mathcal{N}_k} \pi Q^{h(u \wedge v)} \left( Q^{\tilde{d}(u, v) - K} Q^{\mathcal{S}} f_{\mathcal{S}} \right)^2 \right)^{1/2} \\
&\leq C \mathbb{P}_{U_n, V_n}(d(U_n, V_n) \leq 4K) \\
&\quad + C^{1/2} |A_n|^{-2} \sum_{u, v \in A_n: \tilde{d}(u, v) > 2K} M \alpha^{\tilde{d}(u, v) - K} \\
&\leq C \mathbb{P}_{U_n, V_n}(d(U_n, V_n) \leq 4K) + C^{1/2} M \alpha^K.
\end{aligned} \tag{92}$$

$$\leq C \mathbb{P}_{U_n, V_n}(d(U_n, V_n) \leq 4K) + C^{1/2} M \alpha^K. \tag{93}$$

Let  $\varepsilon > 0$ . Let  $K \geq k$  be such that  $C^{1/2} M \alpha^K < \varepsilon$ . Using [Wei24, Lemma 3.1], we get that the first term in the right hand side of (93) goes to zero as  $n \rightarrow \infty$ . Thus, for  $n$  large enough, the right hand side of (93) is upper bounded by  $2\varepsilon$ . This being true for all  $\varepsilon > 0$ , we get that  $\lim_{n \rightarrow \infty} \mathbb{E}_Q[\bar{M}_{A_n}(f)^2] = 0$ . This concludes the proof.  $\square$

We now state and prove a strong law of large numbers for branching Markov chains indexed by the infinite complete binary tree  $T$  and with neighborhood-shape-dependent functions. This result uses the same assumptions as in Theorem A.2.

**Theorem A.4** (Strong law of larger numbers with neighborhood-dependent function). *Let  $Q$  be a transition kernel on  $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$  which is uniformly geometrically ergodic and has a unique invariant probability measure  $\pi$ . Let  $X = (X_u, u \in T)$  be a branching Markov chain with transition kernel  $Q$  and initial distribution  $\pi$ .*

*Let  $k \in \mathbb{N}$  be fixed. Let  $U_k$  be uniformly distributed over  $G_k$  and independent of the process  $X$ , and let  $\mathbb{E}_{U_k} \otimes \mathbb{E}_Q$  denote the joint expectation over  $U_k$  and  $X$ . Let  $f = (f_{\mathcal{S}} : \mathcal{X}^{\mathcal{S}} \rightarrow \mathbb{R})_{\mathcal{S} \in \mathcal{N}_k}$  be a collection of neighborhood-shape-dependent Borel functions that are in  $L^2(X)$ .*

*Then, we have:*

$$\text{a.s.} \quad \lim_{n \rightarrow \infty} \bar{M}_{G_n}(f) = \lim_{n \rightarrow \infty} \bar{M}_{T_n \setminus T_{k-1}}(f) = \mathbb{E}_{U_k} \otimes \mathbb{E}_Q [f_{\mathcal{S}_{U_k}}(X_{\Delta(U_k)})].$$

Moreover, there exist finite constants  $C_0 < \infty$  and  $\beta \in (0, 1)$  such that:

$$\forall n \geq k, \quad \mathbb{E}_Q \left[ \left( \bar{M}_{G_n}(f) - \mathbb{E}_{U_k} \otimes \mathbb{E}_Q [f_{\mathcal{S}_{U_k}}(X_{\Delta(U_k)})] \right)^2 \right] \leq C_0 \beta^n. \tag{94}$$

*Proof.* After using (92), the proof is an easy adaptation of the proof of [Guy07, Theorem 14].

The case of  $\bar{M}_{T_n \setminus T_{k-1}}(f)$  follows directly from the case of  $\bar{M}_{G_n}(f)$  as:

$$|\bar{M}_{T_n}(f)| \leq \sum_{j=k}^n \frac{|G_j|}{|T_n \setminus T_{k-1}|} \bar{M}_{G_j}(f).$$

Thus, it is enough to treat the case of  $\bar{M}_{G_n}(f)$ . In the case where all functions  $f_{\mathcal{S}}$  for  $\mathcal{S} \in \mathcal{N}_k$  are constant, writing  $\bar{M}_{G_n}(f)$  as in (87), and using the convergence in distribution of  $(\mathcal{S}_{U_n})_{n \in \mathbb{N}}$  to the uniform distribution over  $\mathcal{N}_k$ , we get that the sought convergence holds a.s. for  $\bar{M}_{G_n}(f)$ . Thus, without loss of generality, we assume that  $\langle \pi, Q^{\mathcal{S}} f_{\mathcal{S}} \rangle = 0$  for all  $\mathcal{S} \in \mathcal{N}_k$ .

Remark that it is enough to prove that  $\sum_{n \geq k} \mathbb{E}[\bar{M}_{G_n}(f)^2] < \infty$ , as then we can immediately conclude using Borel-Cantelli lemma with Markov's inequality. Thus, for  $n \geq k$ , using (92) with

$n' = |T_n|$  (such that  $A_{n'} = G_n$ ) and  $K = k$ , we get:

$$\begin{aligned}
\mathbb{E}_Q[\bar{M}_{G_n}(f)^2] &\leq C 2^{-(n-2k)} + C^{1/2} M 2^{-2n} \sum_{u,v \in G_n: d(u,v) > 2k} \alpha^{\bar{d}(u,v)-k} \\
&= C 2^{-(n-2k)} + C^{1/2} M \sum_{j=k}^n 2^{-(n-j)-\mathbf{1}_{\{j>0\}}} \alpha^{j-k} \\
&\leq C 2^{-(n-2k)} + C^{1/2} M 2^{-(n-k)} \sum_{j=k}^n 2^{j-k} \alpha^{j-k} \\
&\leq C 2^{-(n-2k)} + C^{1/2} M 2^{-(n-k)} C' \max(n+1, (2\alpha)^{n-k}) \\
&\leq C 2^{-(n-2k)} + C^{1/2} M C' \max((n+1)2^{-(n-k)}, \alpha^{n-k}),
\end{aligned}$$

where  $C'$  is a constant whose value only depends on the value of  $2\alpha$ . Hence, there exist finite constants  $C_0 < \infty$  and  $\beta \in (0, 1)$  such that (94) holds. In particular, we get that  $\sum_{n \geq k} \mathbb{E}[\bar{M}_{G_n}(f)^2] < \infty$ . This concludes the proof of the theorem.  $\square$

## B Proof of the “backward” coupling Lemma 4.1

We now prove Lemma 4.1.

*Proof of Lemma 4.1.* The proof relies on a “backward in time” bound from  $u$  to  $u \wedge v$ , and then a “forward in time” bound from  $u \wedge v$  to  $v$ . We divide the proof in two cases: first when  $v$  is an ancestor of  $u$ , and then the general case.

For all  $j \leq k$ , define the vertex  $U_j = p^j(u)$  which is random for  $j > h(u)$  (in which case, it depends on  $\mathcal{U}$ ). Write  $x_{U_k} = x$ . For all  $j \in \{1, \dots, k\}$ , define the random set (which depends on  $\mathcal{U}$ ):

$$\Delta^-(u, k, j) = (\Delta^*(u, k) \setminus \{U_k\}) \cap (T^\infty(U_k) \setminus T^\infty(U_{j-1})),$$

and in particular remark that  $U_0 = u \notin \Delta^-(u, k, j)$  and  $U_k = p^k(u) \notin \Delta^-(u, k, j)$ .

**Case 1:  $v$  is an ancestor of  $u$ .** We mimic the proof of [CMR05, Proposition 12.5.4]. The proof of the first case relies on the observation that conditioned on  $X_{p^k(u)}$  and  $Y_{\Delta(u,k)}$ , the backward ancestral process  $X$  from  $U_0 = u$  to  $U_k = p^k(u)$  is a non-homogeneous Markov chain satisfying a uniform mixing condition. The fact that  $(X_{U_j})_{0 \leq j \leq k}$  is a Markov chain comes from the Markov property of the HMT  $(X, Y)$  (remind the discussion around (2) on page 8) which gives for all  $j \in \{1, \dots, k\}$ :

$$\begin{aligned}
\mathcal{L}(X_{U_j} | Y_{\Delta(u,k)}, X_{U_k}, X_{T^\infty(U_{j-1})}) &= \mathcal{L}(X_{U_j} | Y_{\Delta(u,k)}, X_{U_k}, X_{U_{j-1}}) \\
&= \mathcal{L}(X_{U_j} | Y_{\Delta^-(u,k,j)}, X_{U_k}, X_{U_{j-1}}).
\end{aligned} \tag{95}$$

For all integers  $j \in \{1, \dots, k\}$ , the backward transition kernel (which depends on  $\mathcal{U}$ ) from  $X_{U_{j-1}}$  to  $X_{U_j}$  is defined as:

$$B_{x_{U_k}, j}[y_{\Delta(u,k)}](x_{U_{j-1}}; f) = \mathbb{E}_\theta[f(X_{U_j}) \mid Y_{\Delta(u,k)} = y_{\Delta(u,k)}, X_{U_k} = x_{U_k}, X_{U_{j-1}} = x_{U_{j-1}}],$$

for any  $x_{U_{j-1}} \in \mathcal{X}$  and any bounded Borel function  $f$  on  $\mathcal{X}$ . By the Markov property (see (95)),



note that  $B_{x_{U_k}, j}[y_{\Delta(u, k)}](x_{U_{j-1}}, f)$  only depends on  $y_{\Delta^-(u, k, j)}$  instead of  $y_{\Delta(u, k)}$ , that is:

$$\begin{aligned} & B_{x_{U_k}, j}[y_{\Delta(u, k)}](x_{U_{j-1}}; f) \\ &= \mathbb{E}_\theta[f(X_{U_j}) \mid Y_{\Delta^-(u, k, j)} = y_{\Delta^-(u, k, j)}, X_{U_k} = x_{U_k}, X_{U_{j-1}} = x_{U_{j-1}}] \\ &= \frac{\int_{\mathcal{X}} f(x_{U_j}) p_\theta(y_{\Delta^-(u, k, j)}, x_{U_j} \mid X_{U_k} = x_{U_k}) q_\theta(x_{U_j}, x_{U_{j-1}}) \lambda(dx_{U_j})}{\int_{\mathcal{X}} p_\theta(y_{\Delta^-(u, k, j)}, x_{U_j} \mid X_{U_k} = x_{U_k}) q_\theta(x_{U_j}, x_{U_{j-1}}) \lambda(dx_{U_j})}, \end{aligned} \quad (96)$$

where:

$$\begin{aligned} & p_\theta(y_{\Delta^-(u, k, j)}, x_{U_j} \mid X_{U_k} = x_{U_k}) = \\ & \int_{\mathcal{X}^{\Delta^-(u, k, j)}} \prod_{w \in \Delta^-(u, k, j)} q_\theta(x_{p(w)}, x_w) g_\theta(x_w, y_w) \prod_{w \in \Delta^-(u, k, j) \setminus \{U_j\}} \lambda(dx_w). \end{aligned}$$

To simplify notations, we will keep the dependence on  $y_{\Delta(u, k)}$  for all indices  $j$ . Note that the integral in the denominator in the right hand side of (96) is lower bounded by:

$$\prod_{w \in \Delta^-(u, k, j)} \sigma^- \int_{\mathcal{X}} g_\theta(x_w, y_w) \lambda(dx_w),$$

and is thus positive  $\mathbb{P}_\theta$ -a.s. under Assumption 3.

Using Assumption 3, we get that those backward transition kernels satisfy the following Doeblin condition (remind Definition 2.6):

$$\frac{\sigma^-}{\sigma^+} \nu_{x_{U_k}, j}[y_{\Delta(u, k)}](f) \leq B_{x_{U_k}, j}[y_{\Delta(u, k)}](x_{U_{j-1}}; f),$$

where for any bounded Borel function  $f$  on  $\mathcal{X}$ , we have:

$$\begin{aligned} \nu_{x_{U_k}, j}[y_{\Delta(u, k)}](f) &= \mathbb{E}_\theta[f(X_{U_j}) \mid Y_{\Delta^-(u, k, j)} = y_{\Delta^-(u, k, j)}, X_{U_k} = x_{U_k}] \\ &= \frac{\int_{\mathcal{X}} f(x_{U_j}) p_\theta(y_{\Delta^-(u, k, j)}, x_{U_j} \mid X_{U_k} = x_{U_k}) \lambda(dx_{U_j})}{\int_{\mathcal{X}} p_\theta(y_{\Delta^-(u, k, j)}, x_{U_j} \mid X_{U_k} = x_{U_k}) \lambda(dx_{U_j})}, \end{aligned}$$

where note that the only difference with the definition of  $B_{x_{U_k}, j}[y_{\Delta(u, k)}](x_{U_{j-1}}; f)$  is that the term  $q_\theta(x_{U_j}, x_{U_{j-1}})$  has disappeared in both the numerator and the denominator of  $\nu_{x_{U_k}, j}[y_{\Delta(u, k)}](f)$ . Thus, Lemma 2.7 shows that the Dobrushin coefficient  $\delta(B_{x_{U_k}, j})$  (defined in (4)) of the backward transition kernel  $B_{x_{U_k}, j}$  is upper bounded by  $\rho = 1 - \sigma^-/\sigma^+$ .

Note that the Markov property in (95) (with  $j = 1$ ) also gives us:

$$\begin{aligned} \mathcal{L}(X_{U_1} \mid Y_{\Delta(u, k)}, X_{U_k}, X_{U_0}) &= \mathcal{L}(X_{U_1} \mid Y_{\Delta^-(u, k, j)}, X_{U_k}, X_{U_0}) \\ &= \mathcal{L}(X_{U_1} \mid Y_{\Delta^*(u, k)}, X_{U_k}, X_{U_0}). \end{aligned} \quad (97)$$

Finally, if we write:

$$\begin{aligned} & \mathbb{P}_\theta(X_v \in \cdot \mid Y_{\Delta(u, k)} = y_{\Delta(u, k)}, X_{U_k} = x_{U_k}) \\ &= \int \mathbb{P}_\theta(X_v \in \cdot \mid Y_{\Delta^-(u, k, 1)} = y_{\Delta^-(u, k, 1)}, X_{U_k} = x_{U_k}, X_{U_1} = x_{U_1}) \\ & \quad \times \mathbb{P}_\theta(X_{U_1} \in dx_{U_1} \mid Y_{\Delta(u, k)} = y_{\Delta(u, k)}, X_{U_k} = x_{U_k}), \end{aligned} \quad (98)$$

and we also write (using (97)):

$$\begin{aligned} \mathbb{P}_\theta(X_v \in \cdot | Y_{\Delta^*(u,k)} = y_{\Delta^*(u,k)}, X_{U_k} = x_{U_k}) \\ = \int \mathbb{P}_\theta(X_v \in \cdot | Y_{\Delta^-(u,k,1)} = y_{\Delta^-(u,k,1)}, X_{U_k} = x_{U_k}, X_{U_1} = x_{U_1}) \\ \times \mathbb{P}_\theta(X_{U_1} \in dx_{U_1} | Y_{\Delta^*(u,k)} = y_{\Delta^*(u,k)}, X_{U_k} = x_{U_k}), \end{aligned} \quad (99)$$

then the two distributions (for  $X_v$ ) on the left hand sides of those displayed equations can be considered as obtained through running  $d(u, v) - 1$  iterations of the backward ancestral conditional Markov chain described above, using two different initial conditions. Therefore, as the Dobrushin coefficient is sub-multiplicative (remind Lemma 2.5), we get that those two probability distribution differ by at most  $2\rho^{d(u,v)-1}$  in total variation. This concludes the proof of the first case.

**Case 2: general case.** The proof of the second case relies on the observation that conditioned on  $X_{p^k(u)}$  and  $Y_{\Delta(u,k)}$ , if we consider the process  $X$  backward from  $u$  to  $u \wedge v$  (remind that  $v \in \Delta^*(u, k)$ ) and then forward from  $u \wedge v$  to  $v$ , we get a non-homogeneous Markov chain satisfying uniform mixing rate  $\rho$ . Note that as  $v \in T^\infty(p^k(u), k) \setminus \{u\}$ , we have that  $u \wedge v \in \{U_1, \dots, U_k\}$ . Using the first case, it only remains to check those observations for the forward segment, which were already proved in the proof of Lemma 3.2.

Hence, if we use the same decomposition as in (98) and (98), which corresponds to run  $d(u, v) - 1$  iterations of the backward-forward conditional chain described above ( $d(u, u \wedge v) - 1$  backward iterations and  $d(u \wedge v, v)$  forward iterations), we get as in the first case that those two probability distribution differ by at most  $2\rho^{d(u,v)-1}$  in total variation. This concludes the proof of the lemma.  $\square$

## C Proof of (35) (used in the proof of Proposition 3.10)

Let  $m \in \mathbb{N}^*$  be fixed through this section.

For ease of read, we restate some notation definitions used only in the proof of Proposition 3.10. For  $u, v \in T^\infty$  with  $h(u) \equiv h(v) \pmod{m+1}$ , we write  $T(u, m) < T(v, m)$  if  $u < v$ . Moreover for  $u, v \in T^\infty$ , we write  $u < T(v, m)$  if  $h(u) < h(v)$  or  $h(u) \leq h(v) + m$  and for all  $w \in T(v, m)$  with  $h(w) = h(u)$  (note that such  $w$  must exist), we have  $u < w$ . Informally  $u$  is “above or on the left of  $T(v, m)$ ”. For all  $u \in T$ ,  $k \in \mathbb{N}$ , define the random subtrees which depend on  $\mathcal{U}$ :

$$\Delta^*(T(u, m), k) = \bigcup \{T(v, m) : v \in \Delta^*(u, k(m+1)) \text{ such that } h(v) \equiv h(u) \pmod{m+1}\},$$

and  $\Delta(T(u, m), k) = \Delta^*(T(u, m), k) \cup T(u, m)$ . When  $h(u) \geq k(m+1)$ , then those subtrees do not depend on  $\mathcal{U}$ , and we write  $\Delta^*(T(u, m), k) = \Delta^*(T(u, m), k)$  and  $\Delta(T(u, m), k) = \Delta(T(u, m), k)$  to indicate it. See Figure 5 on page 24 for an illustration of the “past” subtree  $\Delta^*(T(u, m), k)$  of the block subtree  $T(u, m)$ .

The goal of this section is to prove (35) for all  $\theta \in \Theta$  and  $x \in \mathcal{X}$ , which we restate here for ease of read:

$$\lim_{k \rightarrow \infty} \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [\log p_\theta(Y_{T_m} | Y_{\Delta^*(T_m, k)}, X_{p^{k(m+1)}(\theta)} = x)] = |T_m| \ell(\theta).$$

## C.1 Decomposition of the log-likelihood into subtree increments

Following (16), for all  $u \in T$ ,  $k \in \mathbb{N}$ ,  $x \in \mathcal{X}$  and  $\theta \in \Theta$ , using the conditional probabilities formula, define:

$$\begin{aligned} \mathbb{H}_{T(u,m),k,x}(\theta) &= \frac{p_\theta(Y_{\Delta(T(u,m),k)} | X_{\mathbb{P}^k(u)} = x)}{p_\theta(Y_{\Delta^*(T(u,m),k)} | X_{\mathbb{P}^k(u)} = x)} \\ &= \int p_\theta(Y_{T(u,m)} | X_u = x_u) \mathbb{P}_\theta(X_u \in dx_u | Y_{\Delta^*(T(u,m),k)}, X_{\mathbb{P}^{(k-1)(m+1)}(u)} = x), \end{aligned} \quad (100)$$

where:

$$p_\theta(Y_{T(u,m)} | X_u = x_u) = \int_{\mathcal{X}^{|T(u,m)|}} g_\theta(x_u, Y_u) \prod_{w \in T(u,m) \setminus \{u\}} g_\theta(x_w, Y_w) q_\theta(x_{\mathbb{P}(w)}, x_w) \lambda(dx_w).$$

We then define the log-likelihood contribution of the subtree  $T(u, m)$  with past over  $k \in \mathbb{N}$  subtree generations (that is,  $k(m+1)$  (node) generations) as:

$$h_{T(u,m),k,x}(\theta) = \log \mathbb{H}_{T(u,m),k,x}(\theta)$$

For all  $n \in \mathbb{N}^*$ , we decompose the tree  $T_{n(m+1)-1}$  into subtrees of height  $m$  (such as  $T_m$ ), and we order those subtrees according to  $<$ . Hence, using (6), (7) and (100) and a telescopic sum argument, the log-likelihood of the observed variables  $Y_{T_{n(m+1)-1}}$  can be rewritten as:

$$\ell_{n(m+1)-1,x}(\theta) = \sum_{k=0}^{n-1} \sum_{u \in G_{k(m+1)}} h_{T(u,m),k,x}(\theta). \quad (101)$$

## C.2 Construction of the log-likelihood increments with infinite past for subtrees

In this subsection, we construct the log-likelihood increments with infinite past for subtrees.

To construct the limit of the functions  $h_{T(u,m),k,x}(\theta)$  we first prove the following lemma which states some uniform bound about the asymptotic behavior of those functions when  $k \rightarrow \infty$ .

**Lemma C.1** (Uniform bounds for  $h_{T(u,m),k,x}(\theta)$ ). *Assume that Assumptions 2–3 and 4-(ii) hold. For all vertices  $u \in T$  and all integers  $k, k' \in \mathbb{N}^*$ , the following assertions hold true:*

$$\sup_{\theta \in \Theta} \sup_{x, x' \in \mathcal{X}} |h_{T(u,m),k,x}(\theta) - h_{T(u,m),k',x'}(\theta)| \leq \frac{\rho^{(k \wedge k')(m+1)-1}}{(1-\rho)^{|T_m|}}, \quad (102)$$

$$\sup_{\theta \in \Theta} \sup_{k \in \mathbb{N}^*} \sup_{x \in \mathcal{X}} |h_{T(u,m),k,x}(\theta)| \leq (|T_m| \log b^+) \vee \left| \sum_{w \in T(u,m)} \log(\sigma^- b^- (Y_w)) \right|. \quad (103)$$

*Proof.* [The proof is a straightforward adaptation of the proof of [CMR05, Lemma 12.3.2] using Lemma 3.2 for the coupling.] Let  $k' \geq k \geq 1$ , and write  $v = \mathbb{P}^{k(m+1)}(u)$ ,  $v' = \mathbb{P}^{k'(m+1)}(u)$ . Then, write:

$$\begin{aligned} \mathbb{H}_{T(u,m),k,x}(\theta) &= \int_{\mathcal{X}^2} \left[ \int_{\mathcal{X}^{|T(u,m)|}} \prod_{w \in T(u,m)} g_\theta(x_w, Y_w) q_\theta(x_{\mathbb{P}(w)}, x_w) \lambda(dx_w) \right] \\ &\quad \times \mathbb{P}_\theta(X_{\mathbb{P}(u)} \in dx_{\mathbb{P}(u)} | Y_{\Delta^*(T(u,m),k)}, X_v = x_v) \times \delta_x(dx_v), \end{aligned} \quad (104)$$

and using the Markov property at  $X_v$ , write:

$$\begin{aligned} \mathbb{H}_{T(u,m),k',x'}(\theta) &= \int_{\mathcal{X}^2} \left[ \int_{\mathcal{X}^{T(u,m)}} \prod_{w \in T(u,m)} g_\theta(x_w, Y_w) q_\theta(x_{p(w)}, x_w) \lambda(dx_w) \right] \\ &\quad \times \mathbb{P}_\theta(X_{p(u)} \in dx_{p(u)} \mid Y_{\Delta^*(T(u,m),k)}, X_v = x_v) \\ &\quad \times \mathbb{P}_\theta(X_v \in dx_v \mid Y_{\Delta^*(T(u,m),k') \setminus \Delta(T(u,m),k)}, X_{v'} = x'). \end{aligned} \quad (105)$$

Applying Lemma 3.2, we get (note that the integrands in (104) and (105) are non-negative):

$$\begin{aligned} &|\mathbb{H}_{T(u,m),k,x}(\theta) - \mathbb{H}_{T(u,m),k',x'}(\theta)| \\ &\leq \rho^{k(m+1)-1} \sup_{x_{p(u)} \in \mathcal{X}} \int \prod_{w \in T(u,m)} g_\theta(x_w, Y_w) q_\theta(x_{p(w)}, x_w) \lambda(dx_w) \\ &\leq \rho^{k(m+1)-1} (\sigma^+)^{|T_m|} \prod_{w \in T(u,m)} \int g_\theta(x_w, Y_w) \lambda(dx_w). \end{aligned} \quad (106)$$

The integral in (104) can be lower bounded giving us:

$$\mathbb{H}_{u,k,x}(\theta) \geq (\sigma^-)^{|T_m|} \prod_{w \in T(u,m)} \int g_\theta(x_w, Y_w) \lambda(dx_w), \quad (107)$$

where the right hand side is positive by Assumption 3-(ii); and similarly for (105). Combining (106) with (107), and with the inequality  $|\log x - \log y| \leq |x - y|/(x \wedge y)$ , we get the first assertion of the lemma:

$$|\mathbb{h}_{T(u,m),k,x}(\theta) - \mathbb{h}_{T(u,m),k',x'}(\theta)| \leq \left( \frac{\sigma^+}{\sigma^-} \right)^{|T_m|} \rho^{k(m+1)-1} = \frac{\rho^{k(m+1)-1}}{(1-\rho)^{|T_m|}}.$$

Combining (100) and (107), we get:

$$\prod_{w \in T(u,m)} \sigma^- b^-(Y_w) \leq \mathbb{H}_{T(u,m),k,x}(\theta) \leq (b^+)^{|T_m|},$$

which yields the second assertion of the lemma (remind that  $b^-(Y_w) > 0$  for all  $w \in T^\infty$  by Assumption 3-(ii)).  $\square$

We are now ready to construct the limit of the functions  $\mathbb{h}_{T(u,m),k,x}(\theta)$  and state some properties of this limit. Note that this result is stated for every  $u \in T$ , but we will only need it for  $u = \partial$ . Remind that we are in the stationary case, and that the HMT process  $(X, Y)$  is defined on  $T^\infty$ .

**Proposition C.2** (Properties of the limit function  $\mathbb{h}_{T(u,m),\infty}(\theta)$ ). *Assume that Assumptions 1–4 hold. For every  $u \in T$  and  $\theta \in \Theta$ , there exists  $\mathbb{h}_{T(u,m),\infty}(\theta) \in L^1(\mathbb{P}_U \otimes \mathbb{P}_{\theta^*})$  such that for all  $x \in \mathcal{X}$ , the sequence  $(\mathbb{h}_{T(u,m),k,x}(\theta))_{k \in \mathbb{N}}$  converges  $\mathbb{P}_U \otimes \mathbb{P}_{\theta^*}$ -a.s. and in  $L^1(\mathbb{P}_U \otimes \mathbb{P}_{\theta^*})$  to  $\mathbb{h}_{T(u,m),\infty}(\theta)$ .*

*Furthermore, this convergence is uniform over  $\theta \in \Theta$  and  $x \in \mathcal{X}$ , that is, we have that  $\lim_{k \rightarrow \infty} \sup_{\theta \in \Theta} \sup_{x \in \mathcal{X}} |\mathbb{h}_{T(u,m),k,x}(\theta) - \mathbb{h}_{T(u,m),\infty}(\theta)| = 0$   $\mathbb{P}_U \otimes \mathbb{P}_{\theta^*}$ -a.s. and in  $L^1(\mathbb{P}_U \otimes \mathbb{P}_{\theta^*})$ .*

The limit function  $\mathbb{h}_{T(u,m),\infty}(\theta)$  can be interpreted as  $\log p_\theta(Y_{T(u,m)} \mid Y_{\Delta^*(T(u,m),\infty)})$ , where  $\Delta^*(T(u,m),\infty) = \{v \in T^\infty : v <_{\mathcal{U}} T(u,m)\}$  is a random subset of vertices. Note that  $\mathbb{h}_{T(u,m),\infty}(\theta)$  is a function of the random set of variables  $(Y_v, v \in \Delta(T(u,m),\infty))$ , where we define  $\Delta(T(u,m),\infty) = \Delta^*(T(u,m),\infty) \cup T(u,m)$ , and thus implicitly depend on  $\mathcal{U}$  through  $\Delta(T(u,m),\infty)$ .

*Proof.* Fix some  $u \in T$ . Note that (102) shows that the sequence  $(h_{T(u,m),k,x}(\theta))_{k \in \mathbb{N}}$  is Cauchy uniformly in  $\theta$  and  $x$ , and thus has  $\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^*}$ -almost surely a limit when  $k \rightarrow \infty$  which does not depend on  $x$ ; we denote this limit by  $h_{T(u,m),\infty}(\theta)$ . Furthermore, we get from (103) that  $(h_{T(u,m),k,x}(\theta))_{k \in \mathbb{N}}$  is uniformly bounded in  $L^1(\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^*})$ , and thus  $h_{T(u,m),\infty}(\theta)$  is in  $L^1(\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^*})$  and the convergence also holds in  $L^1(\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^*})$ . Finally, as the bound in (102) is uniform in  $\theta$  and  $x$ , we get that the convergence holds uniform over  $\theta$  and  $x$  both  $\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^*}$ -almost surely and in  $L^1(\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^*})$ .  $\square$

### C.3 Properties of the contrast function

As the functions  $h_{T(u,m),\infty}(\theta)$  are in  $L^1(\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^*})$  under the assumptions used in Proposition C.2, we can now define the *contrast function*  $\ell^{(m)}$  (which is deterministic) for block subtree of height  $m$  as:

$$\ell^{(m)}(\theta) = \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [h_{T_m,\infty}(\theta)],$$

where remind  $\mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*}$  is the expectation corresponding to  $\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^*}$ . We prove under the  $L^2$  regularity assumption Assumption 5 the convergence of the normalized log-likelihood to this contrast function.

**Proposition C.3** (Ergodic convergence for the log-likelihood). *Assume that Assumptions 1–5 hold. Then, for all  $x \in \mathcal{X}$ , the normalized log-likelihood  $|T_{n(m+1)-1}|^{-1} \ell_{n(m+1)-1,x}(\theta)$  converges  $\mathbb{P}_{\theta^*}$ -a.s. to the contrast function  $\ell^{(m)}(\theta)$  as  $n \rightarrow \infty$ .*

$$\lim_{n \rightarrow \infty} \frac{|T_m|}{|T_{n(m+1)-1}|} \ell_{n(m+1)-1,x}(\theta) = \ell^{(m)}(\theta) \quad \mathbb{P}_{\theta^*}\text{-a.s.} \quad (108)$$

In particular, we get that  $\ell^{(m)}(\theta) = |T_m| \ell(\theta)$ .

*Proof.* Let  $\theta \in \Theta$  be some parameter. Fix some  $k \in \mathbb{N}^*$  and  $x \in \mathcal{X}$ . Remind (101). Applying (102) for each vertex  $u \in G_{j(m+1)}$  with  $j \in \{k, \dots, n-1\}$ , we get:

$$\begin{aligned} \frac{|T_m|}{|T_{n(m+1)-1}|} \left| \ell_{n(m+1)-1,x}(\theta) - \sum_{j=k}^{n-1} \sum_{u \in G_{j(m+1)}} h_{T(u,m),k,x}(\theta) \right| \\ \leq \frac{\rho^{k(m+1)-1}}{(1-\rho)^{|T_m|}} + \frac{|T_m|}{|T_{n(m+1)-1}|} \sum_{j=0}^{k-1} \sum_{u \in G_{j(m+1)}} |h_{T(u,m),j,x}(\theta)|. \end{aligned} \quad (109)$$

Note that by (103), we have that  $|h_{T(u,m),j,x}(\theta)| < \infty$   $\mathbb{P}_{\theta^*}$ -a.s. for all  $j \in \mathbb{N}^*$  and  $u \in G_{j(m+1)}$ . For  $u = \partial$ , we have  $h_{T_m,0,x}(\theta) = \log p_{\theta}(Y_{T_m} | X_{\partial} = x)$  which is finite  $\mathbb{P}_{\theta^*}$ -a.s. by Assumption 3-(iii).

The definition of the *shape* from Section 2.4 can straightforwardly be adapted to the (deterministic) subtrees  $\Delta(T(u,m),k)$  for vertices  $u \in G_{j(m+1)}$  with  $j \geq k$ , where  $u$  is seen as a distinguished vertex of  $\Delta(T(u,m),k)$ . Following (8) (on page 12) in the initial vertex-by-vertex decomposition setting, for a vertex  $u \in G_{j(m+1)}$  with  $j \geq k$ , let  $v_u \in G_{k(m+1)}$  be the unique vertex  $G_{k(m+1)}$  such that  $\Delta(T(u,m),k)$  and  $\Delta(T(v_u,m),k)$  have the same shape. Then, we have:

$$\begin{aligned} h_{T(u,m),k,x}(\theta; Y_{\Delta(T(u,m),k)} = y_{\Delta(T(u,m),k)}) \\ = h_{T(v_u,m),k,x}(\theta; Y_{\Delta(T(v_u,m),k)} = y_{\Delta(T(u,m),k)}). \end{aligned} \quad (110)$$

Moreover, using (103) together with Assumption 5, we get for every  $u \in G_{j(m+1)}$  with  $j \geq k$  that the random variable  $h_{T(u,m),k,x}(\theta; Y_{\Delta(T(u,m),k)})$  is in  $L^2(\mathbb{P}_{\theta^*})$ . Hence, applying a straightforward modification of Lemma 2.11 for subtree blocks  $T(u, m)$  to the collection of neighborhood-shape-dependent functions  $(h_{T(v,m),k,x}(\theta; Y_{\Delta(T(v,m))} = \cdot))_{v \in G_{k(m+1)}}$  (remind that indexing functions with  $G_{k(m+1)}$  or with the set of possible shapes is equivalent by (9)), and using (110) and (14) (in Remark 3.1), we get:

$$\frac{|T_m|}{|T_{n(m+1)-1}|} \sum_{j=k}^{n-1} \sum_{u \in G_{j(m+1)}} h_{T(u,m),k,x}(\theta) \xrightarrow{n \rightarrow \infty} \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [h_{T_m,k,x}(\theta)] \quad \mathbb{P}_{\theta^*}\text{-a.s.} \quad (111)$$

Using (102) with Proposition C.2, we get:

$$|\mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [h_{T_m,k,x}(\theta)] - \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [h_{T_m,\infty}(\theta)]| \leq \frac{\rho^{k(m+1)-1}}{(1-\rho)^{|T_m|}}.$$

Thus, combining this bound with (109) and (111), we get  $\mathbb{P}_{\theta^*}$ -a.s. that:

$$\limsup_{n \rightarrow \infty} \left| \frac{|T_m|}{|T_{n(m+1)-1}|} \ell_{n(m+1)-1,x}(\theta) - \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [h_{T_m,\infty}(\theta)] \right| \leq 2 \frac{\rho^{k(m+1)-1}}{(1-\rho)^{|T_m|}}.$$

As the left hand side does not depend on  $k$ , letting  $k \rightarrow \infty$ , we get that (108) in the lemma holds. Lastly, as the limit must be the same as in Proposition 3.5, we get that  $\ell^{(m)}(\theta) = |T_m| \ell(\theta)$ . This concludes the proof.  $\square$

We are now ready to close this section by proving that (35) holds.

**Proposition C.4.** *Assume that Assumptions 1–5 hold. Then, (35) holds for all  $\theta \in \Theta$  and  $x \in \mathcal{X}$ .*

*Proof.* Applying Proposition C.2, we get that the left hand side of (35) is equal to  $\ell^{(m)}(\theta) = \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [h_{T(\partial,m),\infty}(\theta)]$ , which is equal to  $|T_m| \ell(\theta)$  by Proposition C.3.  $\square$

## D Details of the proof of Proposition 4.5

Remind that the proof of Proposition 4.4 can be straightforwardly adapted to Proposition 4.5 except for Lemma 4.8. In Section 4.2.2, for brevity, we have only presented the adaptation of Lemma 4.8 to the terms  $\Gamma_{u,k,x}(\theta)$ . In this appendix, we present all the details of the adaptation of the rest of the proof of Proposition 4.4 to the terms  $\Gamma_{u,k,x}(\theta)$ .

The following lemma gives an exponential bound on the  $L^2(\mathbb{P}_{\theta^*})$  norm uniformly in  $x \in \mathcal{X}$  for the the average of the quantities  $\Gamma_{u,h(u),x}(\theta^*)$  over  $u \in T_n^*$ .

**Lemma D.1.** *Under the assumptions of Proposition 4.5, for all  $x \in \mathcal{X}$  and  $\theta \in \Theta_0$ , there exist finite constants  $C < \infty$  and  $\alpha \in (0, 1)$  such that for all  $n \in \mathbb{N}^*$  we have:*

$$\mathbb{E}_{\theta^*} \left[ \sup_{x \in \mathcal{X}} \left| \frac{1}{|T_n|} \sum_{u \in T_n^*} \Gamma_{u,h(u),x}(\theta) - \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [\Gamma_{\partial,\infty}(\theta)] \right|^2 \right]^{1/2} \leq C \alpha^n. \quad (112)$$

*Proof.* Let  $x' \in \mathcal{X}$  and  $\theta \in \Theta_0$ . Using Minkowski's inequality and Jensen's inequality, for all  $n, k \in \mathbb{N}^*$ , we get:

$$\begin{aligned}
& \mathbb{E}_{\theta^*} \left[ \sup_{x \in \mathcal{X}} \left| \frac{1}{|T_n|} \sum_{u \in T_n^*} \Gamma_{u, h(u), x}(\theta) - \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [\Gamma_{\partial, \infty}(\theta)] \right|^2 \right]^{1/2} \\
& \leq \mathbb{E}_{\theta^*} \left[ \sup_{x, x' \in \mathcal{X}} \left| \frac{1}{|T_n|} \sum_{u \in T_{k-1}^*} \Gamma_{u, h(u), x}(\theta) \right|^2 \right]^{1/2} \\
& \quad + \mathbb{E}_{\theta^*} \left[ \sup_{x, x' \in \mathcal{X}} \left| \frac{1}{|T_n|} \sum_{u \in T_n \setminus T_{k-1}} \Gamma_{u, h(u), x}(\theta) - \Gamma_{u, k, x'}(\theta) \right|^2 \right]^{1/2} \\
& \quad + \mathbb{E}_{\theta^*} \left[ \left| \frac{1}{|T_n|} \sum_{u \in T_n \setminus T_{k-1}} \Gamma_{u, k, x'}(\theta) - \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [\Gamma_{\partial, k, x'}(\theta)] \right|^2 \right]^{1/2} \\
& \quad + \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [\Gamma_{\partial, k, x'}(\theta) - \Gamma_{\partial, \infty}(\theta)]^2]^{1/2}.
\end{aligned} \tag{113}$$

Using Lemma 4.17 together with (49) on page 30 (which, remind, are both immediate consequences of Lemma 4.2), there exists a finite constant  $C < \infty$  and  $\beta \in (0, 1)$  such that the first term in the right hand side of (113) is upper bounded by  $C2^{-(n-k)}$  (note that  $\frac{|T_{k-1}|}{|T_n|} \leq 2^{-(n-k)}$ ), and the second and fourth terms in the right hand side of (113) are both upper bounded by  $C\beta^{k/2}$ .

We now give an upper bound for the second term in the right hand side of (113). For a vertex  $u$  in  $T \setminus T_{k-1}$ , let  $v_u \in G_k$  be the unique vertex that satisfies the shape equality constraint (8) (on page 12), then we have:

$$\Gamma_{u, k, x'}(\theta; Y_{\Delta(u, k)}) = y_{\Delta(u, k)} = \Gamma_{v_u, k, x'}(\theta; Y_{\Delta(v_u)}) = y_{\Delta(v_u)}. \tag{114}$$

Moreover, using the definition of  $\Gamma_{u, k, x}(\theta)$  in (63) together with the assumption on  $\phi_\theta$  in Proposition 4.5, we get that the random variable  $\Gamma_{u, k, x'}(\theta; Y_{\Delta(u, k)})$  is in  $L^2(\mathbb{P}_{\theta^*})$  for every  $u \in T \setminus T_{k-1}$ . Thus, we can apply Lemma 2.11 (see in particular (11)) to the collection of neighborhood-shape-dependent functions  $(\Gamma_{v_u, k, x'}(\theta; Y_{\Delta(v_u)} = \cdot))_{v \in G_k}$  (remind that indexing functions with  $G_k$  or with  $\mathcal{N}_k$  is equivalent by (9)). Using (11) in Lemma 2.11 together with (28) and (14) in Remark 3.1, we get that there exist  $\gamma \in (0, 1)$  and a finite constant  $C' < \infty$  (note that they both do not depend on  $k$  and  $n$ ) such that for all  $n, k \in \mathbb{N}^*$  with  $n \geq k$ , the second term in the right hand side of (113) is upper bounded by  $C'\gamma^{n-k}$ .

Hence, taking  $k = \lceil n/2 \rceil$ , we get that the left hand side of (113) is upper bounded by  $2C\beta^{n/4} + C'\alpha^{n/2} + C2^{-n/2+1}$ , and thus decays at exponential rate as desired. This concludes the proof.  $\square$

Lemma D.1 implies as a corollary the convergence  $\mathbb{P}_{\theta^*}$ -a.s. and in  $L^2(\mathbb{P}_{\theta^*})$  uniformly in  $x \in \mathcal{X}$  for the the sum of the quantities  $\Gamma_{u, h(u), x}(\theta^*)$  over  $u \in T_n^*$ .

**Corollary D.2.** *Under the assumptions of Proposition 4.5, for all  $x \in \mathcal{X}$  and  $\theta \in \Theta_0$ , we have:*

$$\lim_{n \rightarrow \infty} \sup_{x \in \mathcal{X}} \left| \frac{1}{|T_n|} \sum_{u \in T_n^*} \Gamma_{u, h(u), x}(\theta) - \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [\Gamma_{\partial, \infty}(\theta)] \right| = 0 \quad \mathbb{P}_{\theta^*}\text{-a.s. and in } L^2(\mathbb{P}_{\theta^*}).$$

*Proof.* The convergence in  $L^2(\mathbb{P}_{\theta^*})$  follows immediately from Lemma D.1. Moreover, using again Lemma D.1, we have:

$$\sum_{n \in \mathbb{N}^*} \mathbb{E}_{\theta^*} \left[ \sup_{x \in \mathcal{X}} \left| \frac{1}{|T_n|} \sum_{u \in T_n^*} \Gamma_{u, \infty}(\theta) - \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [\Gamma_{\partial, \infty}(\theta)] \right|^2 \right] < \infty.$$

Hence, Borel-Cantelli lemma and Markov's inequality imply that the convergence in the lemma also holds  $\mathbb{P}_{\theta^*}$ -a.s.  $\square$

The following lemma gives some continuity properties of the function  $\theta \mapsto \Gamma_{\partial, k, x}(\theta)$ .

**Lemma D.3.** *Under the assumptions of Proposition 4.5, for all  $x \in \mathcal{X}$  and  $k \in \mathbb{N}$ , the random function  $\theta \mapsto \Gamma_{\partial, k, x}(\theta)$  is  $\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^*}$ -a.s. continuous on  $\Theta_0$ . Moreover, for all  $\theta \in \Theta_0$ , we have:*

$$\lim_{\delta \rightarrow 0} \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} \left[ \sup_{\theta' \in \Theta_0: \|\theta' - \theta\| \leq \delta} |\Gamma_{\partial, k, x}(\theta') - \Gamma_{\partial, k, x}(\theta)|^2 \right] = 0.$$

*Proof.* We mimic the proof of [DMR04, Lemma 14].

For all  $v \in T^\infty$ , define the random variable  $\|\phi^v\|_\infty = \sup_{\theta' \in \Theta_0} \sup_{x, x' \in \mathcal{X}} |\phi_{\theta'}(x', x, Y_v)|$ . Remind that under the assumptions of Proposition 4.5, the HMT process  $(X, Y)$  is stationary and the random variable  $\|\phi^\partial\|_\infty$  is in  $L^4(\mathbb{P}_{\theta^*})$ . Thus, for all  $v \in T^\infty$ , the random variable  $\|\phi^v\|_\infty$  is in  $L^4(\mathbb{P}_{\theta^*})$ . Remind from (12) on page 15 that  $\Delta(\partial, k)$  is a random subtree of the deterministic subtree  $T^\infty(\mathfrak{p}^k(u), k)$ . Then, note that we have:

$$\sup_{\theta \in \Theta_0} |\Gamma_{\partial, k, x}(\theta)| \leq 4 \left( \sum_{v \in T^\infty(\mathfrak{p}^k(\partial), k)} \|\phi^v\|_\infty \right)^2,$$

where the upper bound is a random variable in  $L^2(\mathbb{P}_{\theta^*})$  (and thus in  $L^2(\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^*})$ ) which depends on  $Y_{T^\infty(\mathfrak{p}^k(u), k)}$  but not on  $\mathcal{U}$ . Hence, to prove the lemma, it suffices to prove that for all  $v_1, v_2 \in T^\infty(\mathfrak{p}^k(u), k) \setminus \{\mathfrak{p}^k(\partial)\}$  and  $\epsilon \in \{0, 1\}$ , we have  $\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^*}$ -a.s. :

$$\begin{aligned} \lim_{\delta \rightarrow 0} \sup_{\theta' \in \Theta_0: \|\theta' - \theta\| \leq \delta} \left| \mathbb{E}_{\theta'}[\phi_{\theta'}^{(2, \epsilon)}(X_{\mathfrak{p}(v_1)}, X_{v_1}, Y_{v_1}, X_{\mathfrak{p}(v_2)}, X_{v_2}, Y_{v_2}) \mid Y_{\Delta(\partial, k)}, X_{\mathfrak{p}^k(\partial)} = x] \right. \\ \left. - \mathbb{E}_{\theta}[\phi_{\theta}^{(2, \epsilon)}(X_{\mathfrak{p}(v_1)}, X_{v_1}, Y_{v_1}, X_{\mathfrak{p}(v_2)}, X_{v_2}, Y_{v_2}) \mid Y_{\Delta(\partial, k)}, X_{\mathfrak{p}^k(\partial)} = x] \right| = 0, \end{aligned}$$

where:

$$\phi_{\theta'}^{(2, \epsilon)}(X_{\mathfrak{p}(v_1)}, X_{v_1}, Y_{v_1}, X_{\mathfrak{p}(v_2)}, X_{v_2}, Y_{v_2}) := \phi_{\theta'}(X_{\mathfrak{p}(v_1)}, X_{v_1}, Y_{v_1}) \phi_{\theta'}(X_{\mathfrak{p}(v_2)}, X_{v_2}, Y_{v_2})^\epsilon.$$

Denote  $x_{\mathfrak{p}^k(\partial)} = x$ , and write:

$$\begin{aligned} \mathbb{E}_{\theta}[\phi_{\theta}^{(2, \epsilon)}(X_{\mathfrak{p}(v_1)}, X_{v_1}, Y_{v_1}, X_{\mathfrak{p}(v_2)}, X_{v_2}, Y_{v_2}) \mid Y_{\Delta(\partial, k)}, X_{\mathfrak{p}^k(\partial)} = x] \\ = \frac{\int_{\mathcal{X}^{|\Delta(\partial, k)|-1}} \phi_{\theta}^{(2, \epsilon)}(x_{\mathfrak{p}(v_1)}, x_{v_1}, Y_{v_1}, x_{\mathfrak{p}(v_2)}, x_{v_2}, Y_{v_2}) \Psi(dx_{\Delta(\partial, k) \setminus \{\mathfrak{p}^k(\partial)\}})}{\int_{\mathcal{X}^{\Delta(\partial, k) \setminus \{\mathfrak{p}^k(\partial)\}}} \mathbf{1} \Psi(dx_{\Delta(\partial, k) \setminus \{\mathfrak{p}^k(\partial)\}})}. \end{aligned} \quad (115)$$

where:

$$\Psi(dx_{\Delta(\partial, k) \setminus \{\mathfrak{p}^k(\partial)\}}) := \prod_{w \in \Delta(\partial, k) \setminus \{\mathfrak{p}^k(\partial)\}} q_{\theta}(x_{\mathfrak{p}(w)}, x_w) g_{\theta}(x_w, Y_w) \lambda(dx_w).$$



Using Assumptions 2-4 (which are part of the assumptions in Proposition 4.5), we know that the integrand in the numerator of the right hand side of (115) is continuous w.r.t.  $\theta$  and is upper bounded by the random variable  $\|\phi^{v_1}\|_\infty(\|\phi^{v_2}\|_\infty)^\epsilon(\sigma^+b^+)^{|T^\infty(p^k(u),k)|-1}$  (remind that  $\sigma^+ \geq 1$  and  $b^+ \geq 1$ ). And similarly, the denominator is continuous w.r.t.  $\theta$ , and, using Assumption 3-(ii), is lower bounded by the random variable:

$$\prod_{w \in \Delta(\partial, k) \setminus \{p^k(\partial)\}} \sigma^- \inf_{\theta' \in \Theta} \int g_{\theta'}(x_w, Y_w) \lambda(dx_w) > 0.$$

Hence, using dominated convergence, we conclude that  $\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^*}$ -a.s. the left hand side of (115) is continuous w.r.t.  $\theta$ . This concludes the proof.  $\square$

As a corollary of Lemma D.3, we get that the function  $\theta \mapsto \Gamma_{\partial, \infty}(\theta)$  is continuous in  $L^2(\mathbb{P}_{\theta^*})$ .

**Corollary D.4.** *Under the assumptions of Proposition 4.5, for all  $\theta \in \Theta_0$ , we have:*

$$\lim_{\delta \rightarrow 0} \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} \left[ \sup_{\theta' \in \Theta_0: \|\theta' - \theta\| \leq \delta} |\Gamma_{\partial, \infty}(\theta') - \Gamma_{\partial, \infty}(\theta)|^2 \right] = 0.$$

*In particular, the function  $\theta \mapsto \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [\Gamma_{\partial, \infty}(\theta)]$  is continuous on  $\Theta_0$ .*

*Proof.* Using Minkowski's inequality and Lemma 4.17, there exist a finite constant  $C < \infty$  and  $\beta \in (0, 1)$  such that for all  $x \in \mathcal{X}$  and  $k \in \mathbb{N}^*$ , we have:

$$\begin{aligned} & \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} \left[ \sup_{\theta' \in \Theta_0: \|\theta' - \theta\| \leq \delta} |\Gamma_{\partial, \infty}(\theta') - \Gamma_{\partial, \infty}(\theta)|^2 \right]^{1/2} \\ & \leq 2C\beta^{k/2} + \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} \left[ \sup_{\theta' \in \Theta_0: \|\theta' - \theta\| \leq \delta} |\Gamma_{\partial, k, x}(\theta') - \Gamma_{\partial, k, x}(\theta)|^2 \right]^{1/2}. \end{aligned} \quad (116)$$

Using Lemma D.3, we get:

$$\limsup_{\delta \rightarrow 0} \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} \left[ \sup_{\theta' \in \Theta_0: \|\theta' - \theta\| \leq \delta} |\Gamma_{\partial, \infty}(\theta') - \Gamma_{\partial, \infty}(\theta)|^2 \right]^{1/2} \leq 2C\beta^{k/2},$$

and taking  $k \rightarrow \infty$ , the upper bound vanishes. This concludes the proof.  $\square$

We now prove a locally uniform law of large numbers for the quantities  $\Gamma_{u, k, x}(\theta)$ .

**Lemma D.5.** *Under the assumptions of Proposition 4.5, for all  $x \in \mathcal{X}$ , we have:*

$$\lim_{\delta \rightarrow 0} \lim_{n \rightarrow \infty} \sup_{\theta' \in \Theta_0: \|\theta' - \theta\| \leq \delta} \left| \frac{1}{|T_n^*|} \sum_{u \in T_n^*} \Gamma_{u, h(u), x}(\theta') - \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [\Gamma_{\partial, \infty}(\theta)] \right| = 0, \quad \mathbb{P}_{\theta^*}\text{-a.s.}$$

*Proof.* First, write:

$$\sup_{\theta' \in \Theta_0: \|\theta' - \theta\| \leq \delta} \left| \frac{1}{|T_n^*|} \sum_{u \in T_n^*} \Gamma_{u, h(u), x}(\theta') - \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [\Gamma_{\partial, \infty}(\theta)] \right|$$

$$\begin{aligned}
&\leq \frac{1}{|T_n|} \sum_{u \in T_n^*} \sup_{\theta' \in \Theta_0 : \|\theta' - \theta\| \leq \delta} |\Gamma_{u,h(u),x}(\theta') - \Gamma_{u,h(u),x}(\theta)| \\
&\quad + \left| \frac{1}{|T_n|} \sum_{u \in T_n^*} \Gamma_{u,h(u),x}(\theta) - \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [\Gamma_{\partial,\infty}(\theta)] \right|. \tag{117}
\end{aligned}$$

Then, we use the exact same argument as in the proofs of Lemma D.1 and Corollary D.2 where for all  $u \in T^*$ , the random variable  $\Gamma_{u,k,x}(\theta)$  is replaced by the random variable:

$$\sup_{\theta' \in \Theta_0 : \|\theta' - \theta\| \leq \delta} |\Gamma_{u,h(u),x}(\theta') - \Gamma_{u,h(u),x}(\theta)|,$$

which are in  $L^2(\mathbb{P}_{\theta^*})$  using the assumptions of Proposition 4.5. This gives us that the first term in the upper bound of (117) converges  $\mathbb{P}_{\theta^*}$ -a.s. as  $n \rightarrow \infty$  to:

$$\mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} \left[ \sup_{\theta' : \|\theta' - \theta\| \leq \delta} |\Gamma_{\partial,\infty}(\theta') - \Gamma_{\partial,\infty}(\theta)| \right],$$

which, by Corollary D.4, vanishes when  $\delta \rightarrow 0$ . Corollary D.2 implies that the second term in the upper bound of (117) vanishes  $\mathbb{P}_{\theta^*}$ -a.s. when  $n \rightarrow \infty$ . This concludes the proof.  $\square$

Combining the previous lemmas in this appendix and Lemma 4.17, we are now ready to prove Proposition 4.5.

*Proof of Proposition 4.5.* By Lemma 4.17, for all  $u \in T$ , we have that  $(\Gamma_{u,k,x}(\theta))_{k \in \mathbb{N}^*}$  is a Cauchy sequence uniformly w.r.t.  $\theta \in \Theta_0$  in  $L^2(\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^*})$  that converges to some limit  $\Gamma_{u,\infty}(\theta)$  (that does not depend on  $x$ ). By Corollary D.2, we have that  $\mathbb{P}_{\theta^*}$ -a.s. the convergence for the the average of the quantities  $\Gamma_{u,h(u),x}(\theta^*)$  over  $u \in T_n^*$  holds uniformly in  $x \in \mathcal{X}$ , that is, (66) in Proposition 4.5 holds. By Corollary D.4, we have that the function  $\theta \mapsto \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^*} [\Gamma_{\partial,\infty}(\theta)]$  is continuous on  $\Theta_0$ . Finally, the last part of the proposition is given by Lemma D.5.  $\square$