Electronic Journal of Statistics
Vol. 19 (2025) 3370–3448
ISSN: 1935-7524
https://doi.org/10.1214/25-EJS2418

Asymptotic properties of the maximum likelihood estimator for hidden Markov models indexed by binary trees

Julien Weibel

Institut Denis Poisson, Université d'Orléans, Université de Tours, CNRS, and CERMICS, École nationale des ponts et chaussées, e-mail: julien.weibel@normalesup.org

Abstract: We consider hidden Markov models indexed by a binary tree where the hidden state space is a general Polish space (i.e. a complete separable 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.

MSC2020 subject classifications: Primary 62M05, 60J80; secondary 62F12, 60J85.
Keywords and phrases: Hidden Markov tree (HMT), hidden Markov model (HMM), branching process, maximum likelihood estimator (MLE), asymptotic normality, consistency, geometric ergodicity.

Received December 2024.

Contents

arXiv: 2409.06295

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 Polish space (i.e. a complete separable 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 Polish 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 Baum and Petrie (1966) and were popularized by Rabiner's tutorial Rabiner (1989). Since then, HMMs have been used in a wide variety of applications such as speech recognition Yu and Deng (2015), bioinformatics Koski (2001), finance Mamon and Elliott (2014), and time-series analysis Zucchini and MacDonald (2009); see also Bouguila, Fan and Amayri (2022) for a more global reference on HMMs applications.

HMTs were first introduced in Crouse, Nowak and Baraniuk (1998) to account for the multi-scale dependency of wavelet coefficients in statistical signal processing with applications in wavelet-based image processing Romberg et al. (2000); Choi and Baraniuk (2001); Duarte,

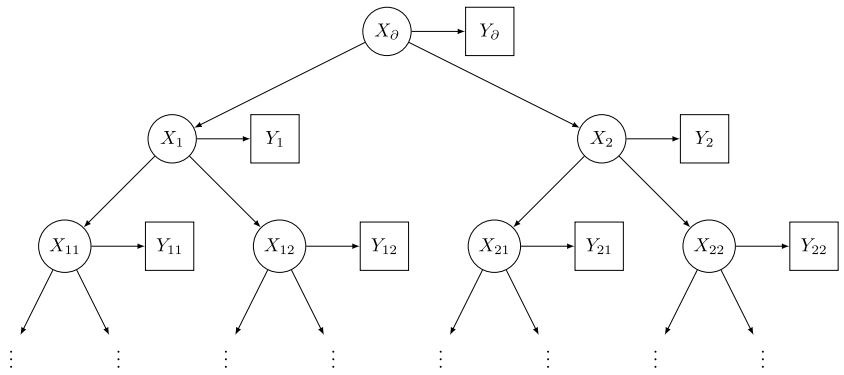

Fig 1. *Graph of dependance for variables of a HMT process indexed by the complete infinite rooted binary tree T. The observed variables are represented inside square, while the hidden variables are represented inside circles.*

Wakin and Baraniuk (2008); Shahdoosti and Hazavei (2017). After that, HMTs have been used in several application contexts such as natural language processing Grave, Obozinski and Bach (2013); Kondo, Duh and Matsumoto (2013), flood mapping Xie, Jiang and Sainju (2018), medical imaging Makhijani et al. (2012); Hu, Yang and Gao (2017); Hanzouli-Ben Salah et al. (2017), plant growth modeling Durand et al. (2005), and bioinformatics Olariu et al. (2009); Biesinger, Wang and Xie (2013); Nakashima, Sughiyama and Kobayashi (2020).

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. Dempster, Laird and Rubin (1977) 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. Baum et al. (1970), 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 (Cappé, Moulines and Rydén, 2005, Chapters 10 and 11). In the HMT case, the "forward-backward" algorithm must be replaced by the "upward-downward" algorithm developed in Crouse, Nowak and Baraniuk (1998). See also Durand, Gonçalves and Guédon (2004) 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 Baum and Petrie (1966) 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 Leroux (1992); Bickel, Ritov and Rydén (1998); Jensen and Petersen (1999); Le Gland and Mevel (2000); Douc and Matias (2001). 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 Douc, Moulines and Rydén (2004), 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 Douc, Moulines and Rydén (2004) rely 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 Douc, Moulines and Rydén (2004) was adapted in Kasahara and Shimotsu (2019) under similar assumptions to allow the transition densities of the hidden process to be zero valued. Since the article Douc, Moulines and Rydén (2004), the strong consistency of the MLE was proved under weaker assumptions in Genon-Catalot and Laredo (2006); Douc et al. (2011); Douc, Roueff and Sim (2016), but no generalization has been made for the asymptotic normality of the MLE.

In this article, we will adapt the proof method of Douc, Moulines and Rydén (2004) to the HMT case. We shall also refer to the monograph Cappé, Moulines and Rydén (2005) 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 Douc, Moulines and Rydén (2004), for HMMs where the hidden state space is a general metric space.

To adapt the proof method of Douc, Moulines and Rydén (2004) 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 Guyon (2007); Weibel (2025). 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 θ, that is, the transition kernel Q_θ between the hidden variables and the transition kernel G_θ from hidden variables to observed variables both depend on θ. 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 θ 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 $\theta^\star$ 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 to estimate the initial distribution. Denote by ∂ the root of the tree T. Thus, we assume that the distribution of the hidden root variable X_∂ is some unknown measure ζ which does not depend on θ. Denote by $\mathbb{P}_{\theta^\star,\zeta}$ the probability distribution of the HMT under the true parameter $\theta^\star$ when the initial unknown distribution of X_∂ is ζ. When ζ is the unique invariant measure of Q_θ (i.e. in the stationary case), we write $\mathbb{P}_{\theta^\star}$ instead of $\mathbb{P}_{\theta^\star,\zeta}$.

To estimate the true parameter $\theta^\star$ 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_∂. 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 Douc, Moulines and Rydén (2004) for HMMs with the same motivation, and conditioning on initial observations in time series goes back at least to Mann and Wald (1943). Remind that T denotes the (deterministic) complete infinite rooted binary tree. Denote by 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 θ 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 3382 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 Θ (see (33) on page 3394 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 Polish spaces. We prove our results under the same assumptions used in Douc, Moulines and Rydén (2004) and in (Cappé, Moulines and Rydén, 2005, 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 Douc, Moulines and Rydén (2004); Cappé, Moulines and Rydén (2005) and the HMT case we develop in this article.

Note that the assumption that $\mathcal{X}$ and $\mathcal{Y}$ are complete and separable will only be used to apply Kolmogorov's extension theorem, which holds for general Polish spaces and not just for the real line, see (Bogachev, 2007, Theorem 7.7.1 with Theorem 7.1.7). (Also note that those assumptions on $\mathcal{X}$ and $\mathcal{Y}$ could be replaced by asking for inner regularity w.r.t. compact sets of the HMT distribution on finite subtrees.)

We first state that strong consistency of the MLE holds under standard assumptions for HMMs. Following Douc, Moulines and Rydén (2004), we assume a fully dominated model, that is, the transition kernels Q_θ and G_θ admits densities q_θ and g_θ w.r.t. to common measures λ and μ, respectively (see Assumption 2). We also assume (see Assumption 3):

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

This assumption is rather strong as it imposes a full connection for the hidden space, see Kasahara and Shimotsu (2019) for an extension of the method in Douc, Moulines and Rydén (2004) for HMMs where q_θ 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_θ and g_θ (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 $\mathbb{P}_{\theta^\star,\zeta}$-almost surely to the true parameter $\theta^\star \in \Theta$.*

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_θ and g_θ (see Assumptions 7–9). Denote by $\mathcal{I}(\theta^\star)$ the limiting Fisher information matrix of the model (see (54) on page 3405 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 $|T_n|^{-1/2}\,\nabla_\theta \ell_{n,x}(\theta^\star)$ has asymptotic normal fluctuations with covariance matrix $\mathcal{I}(\theta^\star)$, 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 ρ 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 ρ.

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 ρ of the HMT process, in the stationary case we have:*

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

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

The following theorem states the locally uniform convergence $\mathbb{P}_{\theta^\star,\zeta}$-a.s. of the normalized observed information $-|T_n|^{-1}\nabla^2_\theta \ell_{n,x}(\theta)$ towards the Fisher information matrix $\mathcal{I}(\theta^\star)$, 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 ρ 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 ρ of the HMT process, for all $x \in \mathcal{X}$, we have:*

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

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

As announced above, following a standard argument for asymptotic normality of the MLE, Theorems 1.1 (see, e.g., (Bickel, Ritov and Rydén, 1998, Theorem 1) or (van der Vaart, 2014, Theorem 7.12)). 1.2 and 1.3 imply the following theorem which states that the MLE has asymptotic normal fluctuations with covariance matrix $\mathcal{I}(\theta^\star)^{-1}$. 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 $\theta^\star$ is an interior point of Θ, and the Fisher information matrix $\mathcal{I}(\theta^\star)$ is non-singular, and under the assumption that $\rho < 1/2$ for the mixing rate ρ of the HMT process, we have the following convergence in distribution:*

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

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

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

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 $\mathcal{I}(\theta^\star)^{1/2}$ is a root matrix of $\mathcal{I}(\theta^\star)$.

The following remark is a discussion on the condition on the mixing rate ρ 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 ρ)**.** 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 Bitseki Penda and Delmas (2022a), corresponding to a competition between the ergodic mixing rate ρ and the branching rate 2 in T, see also Athreya (1969); Bitseki Penda, Djellout and Guillin (2014); Bitseki Penda and Delmas (2022b). However, the condition on ρ disappears when we consider martingale increments in the central limit theorem for branching Markov chains, see Guyon (2007); Bercu, De Saporta and Gégout-Petit (2009); Delmas and Marsalle (2010).

In our case, the condition $\rho < 1/\sqrt{2}$ on the mixing rate ρ 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 ρ 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 Douc, Moulines and Rydén (2004); Cappé, Moulines and Rydén (2005) 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 3383 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 3386), 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 ρ, 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 Douc, Moulines and Rydén (2004) 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 autogression. However, as our approach adapts the proof scheme of Douc, Moulines and Rydén (2004), 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. Appendices B, C and D collect proofs that are redundant from the main body of the article.

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 ∂ the root vertex of T (which is the unique point in $\{0,1\}^0$). If u is distinct from the root, we denote by $\mathrm{p}(u)$ its parent vertex. We denote by $h(u)$ its height, i.e. the number of edges separating u from the root ∂. (The height of the root ∂ is zero.) In particular, for $k \le h(u)$, note that $\mathrm{p}^k(u)$ 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 $T_n = \cup_{0\le k\le n} G_k$ 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_{sub} of T, we write T^*_{sub} for the subtree T_{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 $(u_{(j)})_{1\le j\le n}$ where u is the $u_{(n)}$-th child of $\mathrm{p}(u)$ and $\mathrm{p}(u)$ can be represented by $(u_{(j)})_{1\le j\le n-1}$; and the representation of the root ∂ is the empty sequence. Note that Neveu's notation can also be interpreted as encoding the path from the root ∂ 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 $u_{(k:n)} = (u_{(j)})_{k\le j\le n}$ and $u_{(k:n)} = (u_{(j)})_{k\le j\le n}$ 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 $\le$ on T is defined for all $u, v \in T$ as $u \le v$ if $h(u) < h(v)$ or $h(u) = h(v)$ and $u \le_{\mathrm{lex}} v$ (where $\le_{\mathrm{lex}}$ is the lexicographical order on T). Moreover, we denote by $u < v$ if $u \le v$ and $v \ne u$.

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 σ-field $\mathcal{B}(\mathcal{X})$.

For a measure μ on a metric space $\mathcal{X}$ and an integrable function $f \in L^1(\mu)$, we write $\mu(f) = \int_{\mathcal{X}} f \,\mathrm{d}\mu$. For two probability measures μ_1, μ_2 on a metric space $\mathcal{X}$, we denote the total variation norm between them by $\|\mu_1 - \mu_2\|_{\mathrm{TV}} = \sup_{A\subset\mathcal{X}} |\mu_1(A) - \mu_2(A)|$ (where A ranges over all measurable subsets of $\mathcal{X}$). We also remind the identities $\|\mu_1 - \mu_2\|_{\mathrm{TV}} = \frac{1}{2}\sup_{f:|f|\le 1} |\mu_1(f) - \mu_2(f)| = \sup_{f:0\le f\le 1} |\mu_1(f) - \mu_2(f)|$ (where f is a measurable function). Note that $\|\mu_1 - \mu_2\|_{\mathrm{TV}}$ takes value in $[0,1]$.

Denote by $X = (X_u, u \in T)$ the hidden (stochastic) process with values in a Polish space (i.e. a complete separable metric space) $\mathcal{X}$, and by $Y = (Y_u, u \in T)$ the observed (stochastic) process with values in a Polish 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 defined on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ and initial (probability) distribution ν on $\mathcal{X}$ if for all $n \in \mathbb{N}$, we have:

$$\mathbb{P}(X_{T_n} \in \mathrm{d}x_{T_n}) = \nu(\mathrm{d}x_\partial)\, \Pi_{u \in T_n^*} Q(x_{\mathrm{p}(u)}; \mathrm{d}x_u).$$

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, ν) if:

(i) the hidden process $X = (X_u, u \in T)$ is a branching Markov process with transition kernel Q and initial distribution ν,

(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:

$$\mathbb{P}(Y_{T_n} = y_{T_n} \mid X_{T_n} = x_{T_n}) = \Pi_{u \in T_n} G(x_u; \mathrm{d}y_u).$$

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:

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

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_∂), or equivalently if its initial distribution ν 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.

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 θ 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 Θ the set of all possible vector parameters θ, which we assume to be a subset of $\mathbb{R}^d$ for some integer d. And we denote by $\theta^\star$ 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).

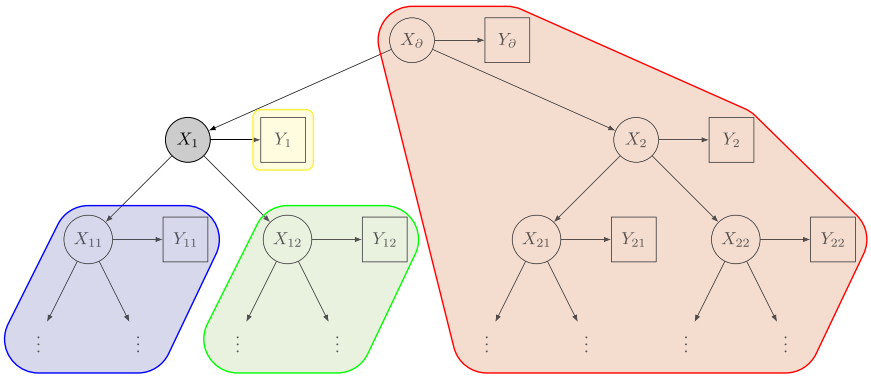

Fig 2. *Illustration of the Markov property for the HMT. Conditioning on X_1 (in grey) implies that the HMT process (X, Y) becomes independent between the four connected components of variable-dependence tree from Figure 1 where the vertex X_1 is removed, that is, Y_1, $(X_{T(11)}, Y_{T(11)})$, $(X_{T(12)},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

We denote by $\mathbb{P}_\theta$ the probability distribution under the parameter θ 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 (Cappé, Moulines and Rydén, 2005, Chapter 12) and Douc, Moulines and Rydén (2004). We first assume that the HMT model is fully dominated. For two measures λ, μ on the same space, we write $\lambda \ll \mu$ to denote that λ is absolutely continuous w.r.t. to μ.

Assumption 2 (Fully dominated model, (Cappé, Moulines and Rydén, 2005, Assumption 12.0.1))**.**

(i) There exists a probability measure λ 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, $Q_\theta(x; A) = \int_A q_\theta(x, x')\, \lambda(\mathrm{d}x')$ 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 σ-finite measure μ 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, $G_\theta(x; B) = \int_B g_\theta(x, y)\, \mu(\mathrm{d}y)$ 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_θ and g_θ.

Assumption 3 (Regularity, (Cappé, Moulines and Rydén, 2005, Assumption 12.2.1))**.**

(i) The transition density q_θ is bounded: there exist $\sigma^-, \sigma^+ \in (0, +\infty)$ such that $\forall x, x' \in \mathcal{X}$, $\forall \theta \in \Theta$, $0 < \sigma^- \le q_\theta(x, x') \le \sigma^+ < +\infty$.

(ii) For every $y \in \mathcal{Y}$, the function $\theta \mapsto \int_\mathcal{X} g_\theta(x, y)\, \lambda(\mathrm{d}x)$ is bounded away from 0 and ∞ uniformly on Θ, that is, $\sup_{\theta\in\Theta} \int_\mathcal{X} g_\theta(x, y)\, \lambda(\mathrm{d}x) < \infty$ and $\inf_{\theta\in\Theta} \int_\mathcal{X} g_\theta(x, y)\, \lambda(\mathrm{d}x) > 0$.

(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 λ is a probability measure and q_θ 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_θ has a unique invariant measure π_θ 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 ρ. Moreover, note that Assumption 3-(i) also implies that $\pi_\theta \ll \lambda$ with density $\frac{\mathrm{d}\pi_\theta}{\mathrm{d}\lambda}$ taking value in $[\sigma^-, \sigma^+]$.

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

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

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:

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

The Dobrushin coefficient gives the following coupling bound in the total variation norm. (Note that the definition of the total variation norm $\|\cdot\|_{\mathrm{TV}}$ used in (Cappé, Moulines and Rydén, 2005, Chapter 4) differs by a factor 2 from ours, see (Cappé, Moulines and Rydén, 2005, Lemma 4.3.5).)

Lemma 2.4 ((Cappé, Moulines and Rydén, 2005, Lemma 4.3.8))**.** *Let μ_1, μ_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 ((Cappé, Moulines and Rydén, 2005, 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 know define the Doeblin condition.

Definition 2.6 (Doeblin condition, (Cappé, Moulines and Rydén, 2005, 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 ν 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 ((Cappé, Moulines and Rydén, 2005, Lemma 4.3.13))**.** *Let K be transition kernel (on a metric space $\mathcal{X}$) that satisfies a Doeblin condition with (ε, ν). 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 ((Cappé, Moulines and Rydén, 2005, Theorem 4.3.16)). *Let K be a transition kernel on a metric space $\mathcal{X}$ that satisfies a Doeblin condition with (ε, ν). Then, K admits a unique invariant probability measure π. Moreover, for any probability measure ζ 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 (Douc et al., 2018, 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:

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

the conditional density w.r.t. $\mu^{\otimes T_n}$ under the parameter θ of $Y_{T_n} = y_{T_n}$ conditionally on $X_\partial = x_\partial$. Note that Assumption 3 guarantees that $p_\theta(y_{T_n} \mid X_\partial = x_\partial)$ 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:

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

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

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

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 Θ 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 $\mathbb{P}_{\theta^\star}$ that $g_{\theta^\star}(x, Y_\partial) = 0$ for some $x \in \mathcal{X}$, implying $\ell_{n,x}(\theta^\star) = -\infty$, and thus preventing the MLE to converge to $\theta^\star$. Several modifications of the log-likelihood $\ell_{n,x}(\theta)$ can be considered to get rid of Assumption 3-(iii).

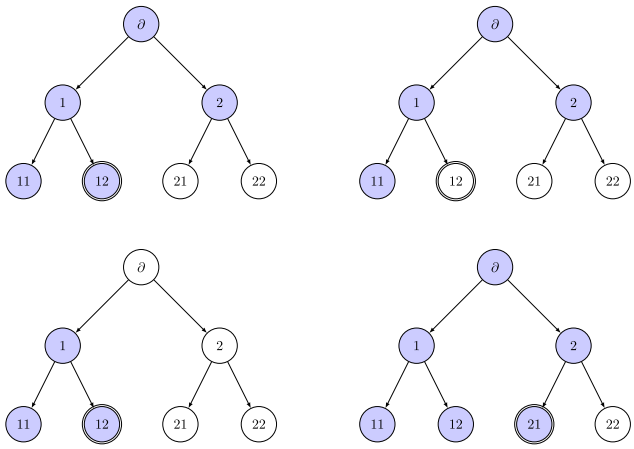

Fig 3. *Illustration of the subtrees* $\Delta(u,k)$ *and* $\Delta^*(u,k)$ *where vertices in* $\Delta(u,k)$ *and* $\Delta^*(u,k)$ *are in blue and the vertex* u *is inside a double circle. From left to right and top to bottom, using Neveu's notation, we have* $\Delta(12) = \Delta(12,2)$, $\Delta^*(12)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

A first option would be to replace $p_\theta(y_{T_n} \mid X_\partial = x)$ by $p_\theta(y_{T_n^*} \mid X_\partial = x)$ in (6). A second option would be to extend the tree T and the HMT (X,Y) to add a parent vertex $\mathrm{p}(\partial)$ for the root vertex ∂ (see Section 3.1.1), and then replace $p_\theta(y_{T_n} \mid X_\partial = x)$ by $p_\theta(y_{T_n} \mid X_{\mathrm{p}(\partial)} = x)$ in (6).

2.4. Ergodic theorems with neighborhood-dependent functions

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

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

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 $h(u) \ge k$, 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:

$$\mathcal{S}h\big(\Delta(u,k)\big) = \Delta(v_u). \tag{8}$$

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:

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

For any $k \in \mathbb{N}$, we define a collection of *neighborhood-shape-dependent* functions as a collection of functions $(f_{\mathcal{S}} : \mathcal{Z}^{\mathcal{S}} \to \mathbb{R})_{\mathcal{S}\in\mathcal{N}_k}$ 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 $f_{\mathcal{S}h(\Delta(u,k))}$. 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 $\mathcal{S}h(\Delta(u,k))$ 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 $(f_{\mathcal{S}} : \mathcal{Y}^{\mathcal{S}} \to \mathbb{R})_{\mathcal{S}\in\mathcal{N}_k}$ be a collection of neighborhood-shape-dependent Borel functions that are in $L^2(\mathbb{P}_{\theta^\star})$. Then, we have:*

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

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

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

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

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. The proof of Lemma 2.11 is postponed to the end of this section.

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 $\mathcal{S}h(\Delta(V_n,k))$ converges to the uniform distribution over $\mathcal{N}_k$ as $n \to \infty$. 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 (Second ergodic theorem for neighborhood-dependent functions)**.** *Assume that Assumptions 1–3 hold. Let $k \geq 0$. Let $(f_{\mathcal{S}} : \mathcal{Y}^{\mathcal{S}} \to \mathbb{R})_{\mathcal{S}\in\mathcal{N}_k}$ be a collection of neighborhood-shape-dependent Borel functions that are in $L^2(\mathbb{P}_{\theta^\star})$. 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:*

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

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

Proof of Lemmas 2.11 and 2.12. 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, Lemma 2.11 (resp. Lemma 2.12) follows immediately from applying the ergodic Theorems A.2 and A.4 for neighborhood-shape-dependent functions from the appendix. □

3. Strong consistency of the MLE

In this section, we first define the extended tree T^∞ 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^∞ 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^∞ that contains T as a subtree and is also rooted at ∂ the root vertex of T, and where each vertex (including ∂) has exactly one parent node and two children nodes. To construct T^∞, we start from T and add a line $L = \{u_{-j} : j \in \mathbb{N}^*\}$ of ancestors for ∂ (that is, $u_{-j} = \mathrm{p}^j(\partial)$ 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^∞ 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^∞.

Thus, we have constructed the deterministic non-plane version of the tree T^∞, and we are left to define the random plane embedding of T^∞. 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 ordered 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 $\partial^{(j+1)}$ (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^∞. (Note that $\mathcal{U}$ can be seen as the random shape of the backward spine of ∂.) See Figure 4 for an illustration of the extended random plane rooted tree T^∞. 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^∞ allows to use Neveu's notation to represent the random path between any vertex in the plane tree T^∞ 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 extended from T to T^∞ using the random plane embedding of T^∞ (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 $U_{(k:n)} \leq_{\mathrm{lex}} V_{(k:n)}$ where $U_{(k:n)}$ (resp. $V_{(k:n)}$) is

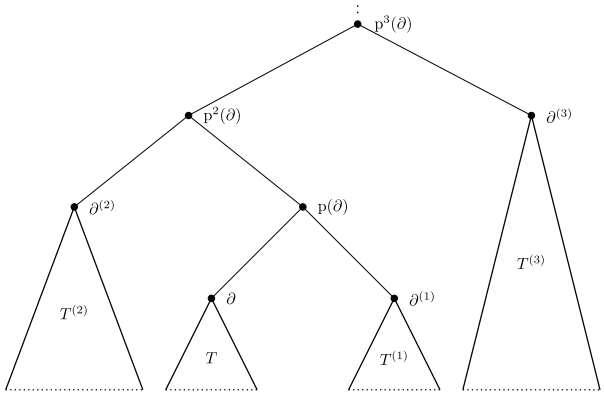

Fig 4. *Illustration of the construction of the extended random plane rooted tree T^∞ (which is rooted at ∂) for $\mathcal{U}_{(-1)} = 0$ and $\mathcal{U}_{(-2)} = 1$.*

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 $T^\infty(\mathrm{p}^k(\partial))$, and thus by Kolmogorov's extension theorem, the HMT process (X, Y) can be defined on the whole tree T^∞. Recall that Kolmogorov's extension theorem holds for general Polish spaces and not just the real line (see (Bogachev, 2007, Theorem 7.7.1 with Theorem 7.1.7)), and also recall that $\mathcal{X}$ and $\mathcal{Y}$, and thus also $\mathcal{X} \times \mathcal{Y}$, are Polish spaces. In particular, note that the stationarity assumption implies that the distribution of the HMT process (X, Y) is invariant by translation on T^∞, 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^∞.

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

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

and $\boldsymbol{\Delta}_{\mathcal{U}}(u, k) = \boldsymbol{\Delta}^*_{\mathcal{U}}(u, k) \cup \{u\}$. For simplicity, we will write instead $\boldsymbol{\Delta}^*(u, k)$ and $\boldsymbol{\Delta}(u, k)$, 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 $\boldsymbol{\Delta}(u, k)$ do indeed depend on the value of $\mathcal{U}$: for $u = \partial$ and $k = 1$, note that $\boldsymbol{\Delta}(\partial, 1)$ 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 $\boldsymbol{\Delta}(u, k)$. However, for $u \in T$ and $k \le h(u)$, we have that $\boldsymbol{\Delta}(u, k) = \Delta(u, k)$ and $\boldsymbol{\Delta}^*(u, k) = \Delta^*(u, k)$ are deterministic. Also note that we have the following inclusions:

$$T^\infty(u, k-1) \subset \boldsymbol{\Delta}(u, k) \subset T^\infty(u, k), \tag{12}$$

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 $h(u) = n \ge k$, note that $\Delta(u, k)$, up to re-rooting (i.e. up to translation), can be identified with $\boldsymbol{\Delta}(\partial, k)$ 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 $n \ge k$, 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 $\mathbf{\Delta}(\partial, k)$, $\Delta(U_n, k)$ and $\Delta(U_k)$:

$$\mathcal{S}h\big(\mathbf{\Delta}(\partial, k)\big) \overset{(d)}{=} \mathcal{S}h(\Delta(U_n, k)) \overset{(d)}{=} \Delta(U_k). \tag{13}$$

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

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

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

3.1.2. The log-likelihood as a sum of increments

For any (possibly random) subtree Δ of T^∞ with root vertex w, note that we have:

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

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

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

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

$$\mathrm{h}_{u,k,x}(\theta) = \log\big(\mathrm{H}_{u,k,x}(\theta)\big). \tag{17}$$

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

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):

$$\ell_{n,x}(\theta) = \sum_{u\in T_n} \mathrm{h}_{u,h(u),x}(\theta). \tag{18}$$

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 ν and ν' on $\mathcal{X}$, that:*

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

For simplicity, Lemma 3.2 is stated with ∂ as the initial vertex, but note that the results still holds when replacing ∂ 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 $u_0, \cdots, u_k$ with $k = h(u)$ the vertices on the path from ∂ 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:

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

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 $p_\theta(y_{T(u_j,n-j)} \mid X_{u_j} = x_{u_j})$ 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:

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

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

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

Note that the difference between the definitions of F_j and ν_j is that the term $q_\theta(x_{\mathrm{p}(u_j)}, x_{u_j})$ has disappear from both the numerator and the denominator of ν_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. □

To construct the limit of the functions $\mathrm{h}_{u,k,x}(\theta)$ we first prove the following lemma which states some uniform bound about the asymptotic behavior of those functions when $k \to \infty$. For this lemma, we need the following assumption on the density function g_θ 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, (Cappé, Moulines and Rydén, 2005, Assumption 12.3.1))**.** Assume that we have:

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

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

Lemma 3.3 (Uniform bounds for $\mathrm{h}_{u,k,x}(\theta)$)**.** *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:*

$$\sup_{\theta\in\Theta} \sup_{x,x'\in\mathcal{X}} |\mathrm{h}_{u,k,x}(\theta) - \mathrm{h}_{u,k',x'}(\theta)| \le \frac{\rho^{(k\wedge k')-1}}{1-\rho}, \tag{20}$$

$$\sup_{\theta\in\Theta} \sup_{k\in\mathbb{N}^*} \sup_{x\in\mathcal{X}} |\mathrm{h}_{u,k,x}(\theta)| \le \log b^+ \vee |\log(\sigma^- b^-(Y_u))|. \tag{21}$$

Proof. [The proof of this lemma is a straightforward adaptation of the proof of (Cappé, Moulines and Rydén, 2005, Lemma 12.3.2) using Lemma 3.2 for the coupling.] Remind the definition of $\mathrm{H}_{u,k,x}(\theta)$ in (16). Let $k' \ge k \ge 1$, and write $w = \mathrm{p}^k(u)$, $w' = \mathrm{p}^{k'}(u)$. Then, write:

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

and using the Markov property at X_w, write:

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

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

$$\begin{aligned}|\mathrm{H}_{u,k,x}(\theta) - \mathrm{H}_{u,k',x'}(\theta)| &\le \rho^{k-1} \sup_{x_{\mathrm{p}(u)}\in\mathcal{X}} \int_{\mathcal{X}} g_\theta(x_u, Y_u) q_\theta(x_{\mathrm{p}(u)}, x_u)\lambda(\mathrm{d}x_u) \\ &\le \rho^{k-1}\sigma^+ \int_{\mathcal{X}} g_\theta(x_u, Y_u)\lambda(\mathrm{d}x_u).\end{aligned} \tag{24}$$

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

$$\mathrm{H}_{u,k,x}(\theta) \ge \sigma^- \int_{\mathcal{X}} g_\theta(x_u, Y_u)\lambda(\mathrm{d}x_u), \tag{25}$$

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

$$|\mathrm{h}_{u,k,x}(\theta) - \mathrm{h}_{u,k',x'}(\theta)| \le \frac{\sigma^+}{\sigma^-}\rho^{k-1} = \frac{\rho^{k-1}}{1-\rho}.$$

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

We are now ready to construct the limit of the functions $\mathrm{h}_{u,k,x}(\theta)$ 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^∞.

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

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

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

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

3.3. Properties of the contrast function

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

$$\ell(\theta) = \mathbb{E}_\mathcal{U} \otimes \mathbb{E}_{\theta\star}\big[\mathrm{h}_{\partial,\infty}(\theta)\big], \tag{26}$$

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

We prove under the following L^2 regularity assumption the convergence of the normalized log-likelihood to the contrast function. Remind that $b^-(y) = \inf_\theta \int_\mathcal{X} g_\theta(x,y)\lambda(\mathrm{d}x)$. 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 $\mathbb{E}_{\theta\star}\big[(\log b^-(Y_\partial))^2\big] < \infty$.

Remind that the log-likelihood $\ell_{n,x}$ is defined in (7) on page 3382.

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 $\mathbb{P}_{\theta\star}$-a.s. to the contrast function $\ell(\theta)$ as $n \to \infty$.*

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

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

Note that by (21), we have that $|\mathrm{h}_{u,h(u),x}(\theta)| < \infty$ $\mathbb{P}_{\theta\star}$-a.s. for all $u \in T \setminus \{\partial\}$. For $u = \partial$, we have $\mathrm{h}_{\partial,0,x}(\theta) = \log g_\theta(x, Y_\partial)$ which is finite $\mathbb{P}_{\theta\star}$-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 3383), then we have:

$$\mathrm{h}_{u,k,x}(\theta; Y_{\Delta(u,k)} = y_{\Delta(u,k)}) = \mathrm{h}_{v_u,k,x}(\theta; Y_{\Delta(v_u)} = y_{\Delta(u,k)}). \tag{28}$$

Moreover, using (21) together with Assumption 5, we get for every $u \in T \setminus T_{k-1}$ that the random variable $\mathrm{h}_{u,k,x}(\theta; Y_{\Delta(u,k)})$ is in $L^2(\mathbb{P}_{\theta\star})$. Hence, applying Lemma 2.11 to the collection of neighborhood-shape-dependent functions $(\mathrm{h}_{v,k,x}(\theta; Y_{\Delta(v)} = \cdot))_{v \in G_k}$ (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:

$$|T_n|^{-1} \sum_{u \in T_n \setminus T_{k-1}} \mathrm{h}_{u,k,x}(\theta) \underset{n\to\infty}{\longrightarrow} \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta\star}\left[\mathrm{h}_{\partial,k,x}(\theta)\right] \qquad \mathbb{P}_{\theta\star}\text{-a.s. (and in } L^2(\mathbb{P}_{\theta\star})). \tag{29}$$

Using (20) with Proposition 3.4, we get:

$$\left|\mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta\star}\left[\mathrm{h}_{\partial,k,x}(\theta)\right] - \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta\star}\left[\mathrm{h}_{\partial,\infty}(\theta)\right]\right| \le \frac{\rho^{k-1}}{1-\rho}.$$

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

$$\limsup_{n\to\infty} \left| |T_n|^{-1} \ell_{n,x}(\theta) - \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta\star}\left[\mathrm{h}_{\partial,\infty}(\theta)\right]\right| \le 2\,\frac{\rho^{k-1}}{1-\rho}.$$

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

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

Assumption 6 (Continuity, (Cappé, Moulines and Rydén, 2005, 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 (ℓ 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 $\mathbb{P}_{\theta\star}$-a.s. continuous on Θ.*

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

$$\mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta\star}\left[\sup_{\theta' \in \Theta : \|\theta - \theta'\| \le \delta} \left|\mathrm{h}_{\partial,\infty}(\theta') - \mathrm{h}_{\partial,\infty}(\theta)\right|\right] \to 0 \quad \text{as } \delta \to 0,$$

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

Proof. This proof is a straightforward adaptation from the proof of (Cappé, Moulines and Rydén, 2005, Proposition 12.3.6).

Recall that $\mathrm{h}_{\partial,\infty}(\theta)$ is the limit of $\mathrm{h}_{\partial,k,x}(\theta)$ as $k \to \infty$. We first prove that, for every $x \in \mathcal{X}$ and $k \ge 0$, $\mathrm{h}_{\partial,k,x}(\theta)$ is a continuous function of θ, and then use this to show continuity

of the limit. Recall from (16) the second equality defining $\mathrm{H}_{u,k,x}(\theta)$, which we remind for convenience for any $u \in T$ and $x \in \mathcal{X}$:

$$\mathrm{H}_{u,k,x}(\theta) = \frac{p_\theta(Y_{\mathbf{\Delta}(u,k)} \mid X_{\mathrm{p}^k(u)} = x)}{p_\theta(Y_{\mathbf{\Delta}^*(u,k)} \mid X_{\mathrm{p}^k(u)} = x)}.$$

Recall from (15) the definition of $p_\theta(Y_\Delta \mid X_{\mathrm{p}^k(u)} = x)$ where here the possibly random subtree Δ is either $\mathbf{\Delta}(u,k)$ or $\mathbf{\Delta}^*(u,k)$. First note that the integrand in (15) is by assumption continuous w.r.t. θ and upper bounded by $(1 \vee \sigma^+ b^+)^{|\Delta|}$. Thus, dominated convergence shows that $p_\theta(Y_\Delta \mid X_{\mathrm{p}^k(u)} = x)$ is continuous w.r.t. to θ (remind that λ, defined in Assumption 2, is finite). Moreover, note that $p_\theta(Y_{\mathbf{\Delta}^*(u,k)} \mid X_{\mathrm{p}^k(u)} = x)$ is lower bounded by $\prod_{v \in \mathbf{\Delta}^*(u,k) \setminus \{\mathrm{p}^k(u)\}} \sigma^- b^-(Y_v)$ which is positive $\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^\star}$-a.s. (by Assumption 3). Thus, we get that $\mathrm{H}_{u,k,x}(\theta)$ and $\mathrm{h}_{u,k,x}(\theta) = \log \mathrm{H}_{u,k,x}(\theta)$ (remind (17)) are continuous w.r.t. θ $\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^\star}$-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. θ $\mathbb{P}_{\theta^\star}$-a.s.

Remind from Proposition 3.4 that $(\mathrm{h}_{u,k,x}(\theta))_{k \in \mathbb{N}}$ converges to $\mathrm{h}_{u,\infty}(\theta)$ uniformly in θ $\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^\star}$-a.s. Thus, the function $\theta \mapsto \mathrm{h}_{u,\infty}(\theta)$ is continuous $\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^\star}$-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:

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

This concludes the proof. □

We are now ready to state and prove that the convergence to the contrast function ℓ holds uniformly in θ.

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

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

Proof. [We mimic the proof of (Cappé, Moulines and Rydén, 2005, Proposition 12.3.7).] As Θ is compact, it is sufficient to prove that for every $\theta \in \Theta$:

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

As this claim is not proven in the proof of (Cappé, Moulines and Rydén, 2005, 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 ℓ is continuous, and thus uniformly continuous as Θ is compact. In particular, there exists $\delta > 0$ such that for all $\theta, \theta' \in \Theta$, we have that $\|\theta - \theta'\| \le \delta$ implies $|\ell(\theta) - \ell(\theta')| \le \varepsilon$. For every $\theta \in \Theta$, let $\delta_\theta \in (0, \delta)$ be such that $\limsup_{n \to \infty} \sup_{\theta' \in \Theta : \|\theta' - \theta\| \le \delta_\theta} \left| |T_n|^{-1} \ell_{n,x}(\theta') - \ell(\theta) \right| < \varepsilon$. As $\cup_{\theta \in \Theta} \{\theta' : \|\theta' - \theta\| \le \delta_\theta\}$ is an open cover of Θ and as Θ is compact, there exists a finite subset $\{\theta_j : 1 \le j \le m\}$ of Θ

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:

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

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 $\lim_{n \to \infty} |T_n|^{-1} \ell_n(\theta) = \ell(\theta)$ $\mathbb{P}_{\theta^\star}$-a.s. Using this fact, we get:

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

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

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

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

Remark 3.8 (Uniform convergence for the log-likelihood with general initial condition)**.** Let ν be a probability distribution on $\mathcal{X}$ such that $\mathbb{P}_{\theta^\star}$-a.s. $\sup_\theta |\int g_\theta(x, Y_\partial) \nu(\mathrm{d}x)|$ is finite. 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 δ_x by ν for the distribution of the root hidden variable X_∂. When ν is the stationary distribution π_θ associated to q_θ, uniform convergence holds without this extra regularity assumption by conditioning on the state of the root's parent $X_{\mathrm{p}(\partial)}$ instead (which allows to replace $\mathrm{h}_{\partial,0,x}(\theta) = g_\theta(x, Y_\partial)$ in (27) by $\mathrm{h}_{\partial,1,\nu}(\theta) := \log \int_{\mathcal{X}} \mathrm{H}_{\partial,1,x}(\theta) \, \nu(\mathrm{d}x)$ for which $\sup_\theta |\mathrm{h}_{\partial,1,\nu}(\theta)|$ 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, θ and θ' are equivalent if and only if they define the same law on every finite tree T_n, i.e. for $(Y_u, u \in T_n)$. Recall that Kolmogorov's extension theorem holds for general Polish spaces and not just the real line (see (Bogachev, 2007, Theorem 7.7.1 with Theorem 7.1.7)), and also recall that $\mathcal{X}$ and $\mathcal{Y}$, and thus also $\mathcal{X} \times \mathcal{Y}$, are Polish spaces.

The following proposition characterizes global maxima of the contrast function ℓ.

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

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

The proof of Proposition 3.10, which is postponed to the end of this section, is an adaptation of the proof of (Cappé, Moulines and Rydén, 2005, 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 $\mathbb{P}_{\theta^\star}$-a.s. under Assumptions 1–4 and 6. Thus, when we further assume that Θ 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 Θ of the log-likelihood $\ell_{n,x}$, that is as the following random variable (which depends on Y_{T_n}):

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

Note that the argmax set in (33) is not necessarily unique, in which case we select one parameter θ from the argmax set in a measurable manner (which is possible, see (Bertsekas and Shreve, 1996, 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 ℓ has a unique maximum (which is then located at $\theta^\star \in \Theta$ by Proposition 3.10) and Θ is compact. Then, for any $x \in \mathcal{X}$, the MLE $\hat{\theta}_{n,x}$ (defined in (33)) converges $\mathbb{P}_{\theta^\star}$-a.s. as $n \to \infty$ to the true parameter $\theta^\star \in \Theta$, i.e. the MLE is strongly consistent.*

Proof. [The proof is a straightforward adaptation of an argument for HMMs in (Cappé, Moulines and Rydén, 2005, Section 12.1), which itself adapts an argument that goes back to Wald (1949).]

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 ℓ has a unique maximum located at $\theta^\star$, we have that $\ell(\theta^\star) \geq \ell(\theta)$ for every $\theta \in \Theta$, and in particular, $\ell(\theta^\star) \geq \ell(\hat{\theta}_n)$ for every $n \in \mathbb{N}$. Combining those two bounds, we get that:

$$0 \leq \ell(\theta^\star) - \ell(\hat{\theta}_n)$$

$$\leq \ell(\theta^\star) - |T_n|^{-1}\ell_{n,x}(\theta^\star) + |T_n|^{-1}\ell_{n,x}(\theta^\star) - |T_n|^{-1}\ell_{n,x}(\hat{\theta}_n) + |T_n|^{-1}\ell_{n,x}(\hat{\theta}_n) - \ell(\hat{\theta}_n)$$
$$\leq 2\sup_{\theta\in\Theta}\left|\ell(\theta) - |T_n|^{-1}\ell_{n,x}(\theta)\right|,$$

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

We now prove Proposition 3.10.

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

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

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

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

where the inner expectation is on Y_∂ conditionally on $X_{\mathrm{p}^k(\partial)} = x$ 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 $\mathrm{H}_{\partial,k,x}(\theta)$ is the conditional density of Y_∂ given $Y_{\Delta^*(\partial,k)}$ and $X_{\mathrm{p}^k(\partial)} = x$, 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 $\ell(\theta^\star) - \ell(\theta)$ as $k \to \infty$ are non-negative as well, and thus $\theta^\star$ is a global maximum of ℓ.

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

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

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

We first need a variant of the convergence in Proposition 3.4 where instead of considering one vertex u as in $\mathrm{h}_{u,k,x}(\theta)$ we consider a whole subtree $T^\infty(u,m)$ for any $m \geq 1$ (this can be seen as a convergence by block). To this end, we need to define an analogue of the breadth-first-search order relation $<$ on T^∞ for subtree blocks of the form $T^\infty(u,m)$. Let $m \geq 1$ be fixed. For $u, v \in T^\infty$ with $h(u) \equiv h(v) \mod m+1$, 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

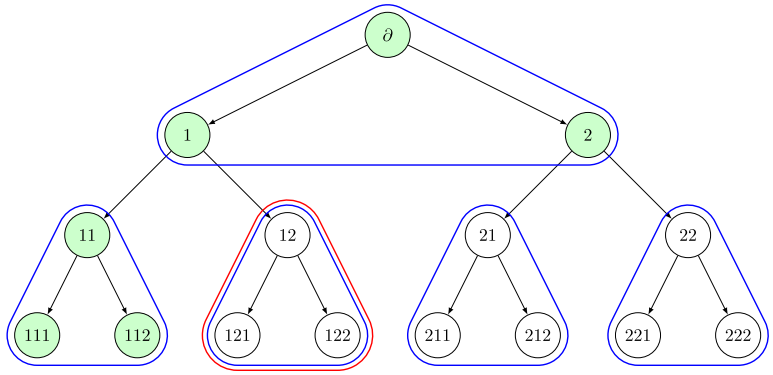

Fig 5. *Illustration of the "past" subtree* $\mathbf{\Delta}^*(T(u,m),k)$ *of the block subtree* $T(u,m)$ *for* $m=1$, $u=12$ *and* $k=1$. *The block subtrees are circled with blue lines, and the block subtree* $T(12,1)$ *is circled a second time with a red line. The vertices in green are those in* $\mathbf{\Delta}^*(T(12,1),1)$.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

is there to insure the collection of block subtrees $T^\infty(u,m)$ with $h(u) \equiv h(\partial) \mod m+1$ form a partition (i.e. a cover with non-overlapping subsets) of T^∞ (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 $\mathbf{\Delta}^*(u,k)$ for subtree blocks, that is, for all $u \in T^\infty$ and $k \in \mathbb{N}$, we define:

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

See Figure 5 for an illustration of the "past" subtree $\mathbf{\Delta}^*(T(u,m),k)$ of the block subtree $T(u,m)$. (Informally, "the subset $\mathbf{\Delta}^*(T(u,m),k)$ 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 $\mathbf{\Delta}^*(T(u,m),k)$, we only need to remember that all its vertices are upstream of the edge $(\mathrm{p}(u),u)$, 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 $m \in \mathbb{N}^*$:

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

Let $\mathcal{U}^+ = (\mathcal{U}_{(j)})_{1\le j<\infty}$ 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 ∂ is encoded by $\mathcal{U}_{(1:n)}$ in Neveu's notation. For all $n \in \mathbb{N}$, define the deterministic vertex $U_{-n} = \mathrm{p}^n(\partial)$. 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 $(\mathcal{S}h(\mathbf{\Delta}(T^\infty(U_n,m),k)))_{n\in\mathbb{Z}}$ is stationary.

Now, pick $\theta \in \Theta$ such that $\ell(\theta) = \ell(\theta^\star)$. Thus for any positive integer $n < m$, we have (where for simplicity we write $w_k = \mathrm{p}^{k(m+1)}(\partial)$):

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

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 $(\mathcal{S}h(\mathbf{\Delta}(T^\infty(U_n, m), k)))_{n\in\mathbb{Z}}$. 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 $m, k \in \mathbb{N}^*$:

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

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 π_θ after Lemma 2.3. From (36), we deduce that (where m in (37) and (38) below corresponds to $m - n$ in (36)):

$$\forall m \in \mathbb{N}^*, \qquad \lim_{k\to\infty} \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star} \left[W_{m,k}(\theta^\star) - W_{m,k}(\theta) \right] = 0. \qquad (38)$$

Hence, we have managed to insert a gap between the variables $(Y_v, v \in T_n)$ whose density we examine and the variables $(Y_v, v \in \mathbf{\Delta}^*(T(U_{-m}, m+n), k))$ and $X_{\mathrm{p}^{k(m+n+1)}(U_{-m})}$ that appear in the conditioning. Remark that the following fact illustrates the gap between the variables: if $u \in T_n$ and $v \in \mathbf{\Delta}^*(T(U_{-m}, m+n), k)$, 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:

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

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

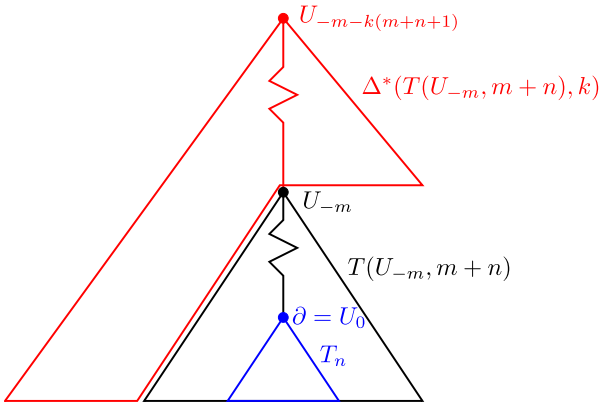

Fig 6. *Illustration of the gap in* (38) *between the variables* $(Y_v, v \in T_n)$ *(bottom triangle in blue) and the variables* $(Y_v, v \in \mathbf{\Delta}^*(T(U_{-m}, m+n), k))$ *and* $X_{\mathrm{p}^{k(m+n+1)}(U_{-m})}$ *that appear in the conditioning (top partial triangle in red). Note that the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

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

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

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

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

and

$$\begin{aligned}
\exp(W(\theta)) &= p_\theta(Y_{T_n}) \\
&= \int_{\mathcal{X}\times\mathcal{X}} p_\theta(Y_{T_n} \mid X_{\mathrm{p}(\partial)} = x_{\mathrm{p}(\partial)})\, Q_\theta^{m-1}(x_{U_{-m}}; \mathrm{d}x_{\mathrm{p}(\partial)}) \pi_\theta(\mathrm{d}x_{U_{-m}}),
\end{aligned}$$

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

$$p_\theta(Y_{T_n} \mid X_{\mathrm{p}(\partial)} = x_{\mathrm{p}(\partial)}) = \int_{\mathcal{X}^{T_n}} \prod_{u\in T_n} q_\theta(x_{\mathrm{p}(u)}, x_u) g_\theta(x_u, Y_u)\, \lambda(\mathrm{d}x_u) \le (b^+)^{|T_n|}. \tag{41}$$

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_θ, we obtain $\mathbb{P}_{\theta^\star}$-a.s. :

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

Moreover, as we have the lower bound:

$$p_\theta(Y_{T_n} \mid X_{\mathrm{p}(\partial)} = x_{\mathrm{p}(\partial)}) = \int_{\mathcal{X}^{T_n}} \prod_{u\in T_n} q_\theta(x_{\mathrm{p}(u)}, x_u) g_\theta(x_u, Y_u)\lambda(\mathrm{d}x_u) \geq \prod_{u\in T_n} \sigma^- b^-(Y_u), \quad (43)$$

this implies that $p_\theta(Y_{T_n} \mid Y_{\Delta^*(T(U_{-m},m+n),k)}, X_{\mathrm{p}^{k(m+n+1)}(U_{-m})} = x)$ 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:

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

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

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

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

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 λ and μ are defined in Assumption 2. We denote by ∇_θ and ∇^2_θ, 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, (Cappé, Moulines and Rydén, 2005, Assumption 12.5.1))**.** There exists an open (for the trace topology on $\Theta \subset \mathbb{R}^d$) neighborhood $O = \{\theta \in \Theta : \|\theta - \theta^\star\| < \delta_0\}$ of $\theta^\star$ 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 O.

(ii) We have:

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

(iii) We have:

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

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

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

(v) For λ-almost all $x \in \mathcal{X}$, there exist functions $f^1_x : \mathcal{X} \to \mathbb{R}_+$ and $f^2_x : \mathcal{X} \to \mathbb{R}_+$ in $L^1(\mu)$ such that $\sup_{\theta\in O}\|\nabla^i_\theta g_\theta(x,y)\| \le f^i_x(y)$ for $i \in \{1,2\}$.

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* $-\nabla^2_\theta \ell_{n,x}(\theta)$ exist and are in $L^2(\mathbb{P}_{\theta^\star})$ and $L^1(\mathbb{P}_{\theta^\star})$, respectively.

4.1. Asymptotic normality of the score

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

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

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 $k \le h(u)$ (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 ∂ 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:

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

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 (Cappé, Moulines and Rydén, 2005, Proposition 10.1.6)), we get:

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

where

$$\phi_\theta(x', x, y) = \nabla_\theta \log\big[q_\theta(x', x) g_\theta(x, y)\big]. \tag{46}$$

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

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

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 $\dot{\mathrm{h}}_{u,k,x}(\theta)$ by $\dot{\mathrm{h}}_{u,0,x}(\theta) = \nabla_\theta \log g_\theta(x, Y_u)$ if $k = 0$, and otherwise by:

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

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

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

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

4.1.2. Construction of score increments with infinite past

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

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

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 3386 that $\mathbf{\Delta}(u,k)$ is a random subtree of the deterministic subtree $T^\infty(\mathrm{p}^k(u),k)$.

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 $v \in T^\infty(\mathrm{p}^k(u),k) \setminus \{u\}$. Then, we have:*

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

The proof of Lemma 4.1, which is postponed to Appendix B, relies on a "backward in time" bound from $\mathrm{p}(u)$ to $u \wedge v$ and then a "forward in time" bound from $u \wedge v$ to v, and using the initial distributions $\mathbb{P}_\theta(X_{\mathrm{p}(u)} \in \cdot \,|\, Y_\Delta, X_{\mathrm{p}^k(u)} = x)$ with Δ equal to $\mathbf{\Delta}(u,k)$ and $\mathbf{\Delta}^*(u,k)$, respectively. Note that this proof is similar to the proofs for Lemma 3.2 and (Cappé, Moulines and Rydén, 2005, Proposition 12.5.4).

The following lemma gives an L^2-bound on the difference between $\dot{\mathrm{h}}_{u,k,x}(\theta)$ and $\dot{\mathrm{h}}_{u,k',x'}(\theta)$ 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 ρ 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 Θ_0 be a closed ball in Θ, and let ψ 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 Θ_0. Furthermore, assume that there exists $b \in [1, +\infty)$ such that:

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

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

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

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 $L^b(\mathbb{P}_\mathcal{U} \otimes \mathbb{P}_{\theta^\star})$ 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 $\dot{\mathrm{h}}_{u,\infty}(\theta)$ the limit function of the sequence $(\dot{\mathrm{h}}_{u,k,x}(\theta))_{n\in\mathbb{N}}$ (for all $x \in \mathcal{X}$) which is in $L^2(\mathbb{P}_\mathcal{U} \otimes \mathbb{P}_{\theta^\star})$.

As an immediate corollary of Lemma 4.2, there exists a finite constant $C' < \infty$ such that for all $\theta \in \Theta_0$, $x \in \mathcal{X}$, $u \in T$ and $k \geq 1$, we have that:

$$\left(\mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}\left[\sup_{x\in\mathcal{X}} \|\xi_{u,k,x}(\theta)\|^b\right]\right)^{1/b} \leq \left(\mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}\left[\sup_{x\in\mathcal{X}} \|\xi_{u,1,x}(\theta)\|^b\right]\right)^{1/b} + C' < \infty,$$

where the expectation $\mathbb{E}_{\mathcal{U}}$ in the right hand side is only used for $u = \partial$ (for $u \in T^*$, $\mathbf{\Delta}(u,1) \subset T$ is deterministic). Also note that by stationarity, for $u \in T^*$, we have that $\mathbb{E}_{\theta^\star}[\sup_{x\in\mathcal{X}} \|\xi_{u,1,x}(\theta)\|^b]) = \mathbb{E}_{\theta^\star}[\sup_{x\in\mathcal{X}} \|\xi_{v,1,x}(\theta)\|^b])$ where v is the only children of the root ∂ such that $\Delta(u,1) = \Delta(v,1)$. Hence, we get:

$$\sup_{\theta\in\Theta_0} \sup_{u\in T} \sup_{k\geq 1} \left(\mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}\left[\sup_{x\in\mathcal{X}} \|\xi_{u,k,x}(\theta)\|^b\right]\right)^{1/b} < \infty. \tag{49}$$

Proof. We mimic the scheme of the proof of (Cappé, Moulines and Rydén, 2005, 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,

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

   and

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

   and similarly for the corresponding terms with $\mathbf{\Delta}(u,k)$ and $\mathbf{\Delta}(u,k')$ replaced by $\mathbf{\Delta}^*(u,k)$ and $\mathbf{\Delta}^*(u,k')$, respectively;
2. For v far from u,

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

   and

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

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

Remind from (12) on page 3386 that $\mathbf{\Delta}(u,k) \subset T^\infty(\mathrm{p}^k(u),k)$ and that the subtree $\mathbf{\Delta}(u,k)$ is random while the subtree $T^\infty(\mathrm{p}^k(u),k)$ is deterministic. Let $(x,x') \in \mathcal{X}\times\mathcal{X}$ and let $v \in T^\infty(\mathrm{p}^k(u),k) \setminus \{\mathrm{p}^k(u)\}$, which implies that $\mathrm{p}(v) \in \Delta(u,k)$.

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

$$\begin{aligned}
&\left\|\mathbb{E}_\theta[\psi_\theta(X_{\mathrm{p}(v)}, X_v, Y_v) \,|\, Y_{\mathbf{\Delta}(u,k)}, X_{\mathrm{p}^k(u)} = x] - \mathbb{E}_\theta[\psi_\theta(X_{\mathrm{p}(v)}, X_v, Y_v) \,|\, Y_{\mathbf{\Delta}(u,k')}, X_{\mathrm{p}^{k'}(u)} = x']\right\| \\
&\quad = \left\| \int_{\mathcal{X}^3} \psi_\theta(x_{\mathrm{p}(v)}, x_v, Y_v)\, \mathbb{P}_\theta(X_v \in \mathrm{d}x_v \,|\, Y_{\mathbf{\Delta}(u,k)\cap T(v)}, X_{\mathrm{p}(v)} = x_{\mathrm{p}(v)}) \right. \\
&\qquad\qquad \times \mathbb{P}_\theta(X_{\mathrm{p}(v)} \in \mathrm{d}x_{\mathrm{p}(v)} \,|\, Y_{\mathbf{\Delta}(u,k)}, X_{\mathrm{p}^k(u)} = x_{\mathrm{p}^k(u)}) \\
&\qquad\qquad \left. \times \left[\delta_x(\mathrm{d}x_{\mathrm{p}^k(u)}) - \mathbb{P}_\theta(X_{\mathrm{p}^k(u)} \in \mathrm{d}x_{\mathrm{p}^k(u)} \,|\, Y_{\mathbf{\Delta}(u,k')}, X_{\mathrm{p}^{k'}(u)} = x')\right] \right\|
\end{aligned}$$

$$\leq 2 \sup_{x_1,x_2\in\mathcal{X}} \|\psi_\theta(x_1,x_2,Y_v)\|\,\rho^{d(v,\mathrm{p}^k(u))-1}, \tag{50}$$

where the inequality is obtained using Lemma 3.2 (note that $d(\mathrm{p}(v),\mathrm{p}^k(u)) = d(v,\mathrm{p}^k(u))-1$). Note that this upper bound is a.s. finite as we have that $\sup_{x_1,x_2\in\mathcal{X}}\|\psi_\theta(x_1,x_2,Y_v)\|$ is in $L^b(\mathbb{P}_\mathcal{U}\otimes\mathbb{P}_{\theta^\star})$ 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 $\mathbf{\Delta}(u,k)$ and $\mathbf{\Delta}(u,k')$ are replaced by $\mathbf{\Delta}^*(u,k)$ and $\mathbf{\Delta}^*(u,k')$, respectively. Obviously, this bound is small if v is far away from $\mathrm{p}^k(u)$ (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 $d(u,v) = d(u,\mathrm{p}(v))+1$), using the Markov property (remind (2)) and Lemma 4.1, we get:

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

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

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

In both cases, we get:

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

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

Remind from (12) on page 3386 that $\mathbf{\Delta}(u,k)\subset T^\infty(\mathrm{p}^k(u),k)$ and as $k'\geq k$ note that:

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

For a vertex $v\in T^\infty(\mathrm{p}^k(u),k)\setminus\{\mathrm{p}^k(u)\}$ (note that $u\wedge v\in T^\infty(\mathrm{p}^k(u))$), note that the term $\rho^{d(v,\mathrm{p}^k(u))-1}$ is smaller than $\rho^{d(u,v)-2}$ whenever $d(v,\mathrm{p}^k(u)) > d(u,v)-1$, that is when $d(u\wedge v,\mathrm{p}^k(u))\geq d(u\wedge v,u)$, 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 $(\mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}[\sup_{\theta\in\Theta_0} \sup_{x,x'\in\mathcal{X}} \|\xi_{u,k,x}(\theta) - \xi_{u,k',x'}(\theta)\|^b])^{1/b}$ is upper bounded by:

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

up to the factor $\left(\mathbb{E}_{\theta^\star}\left[\sup_{\theta\in\Theta_0} \sup_{x_1,x_2\in\mathcal{X}} \|\psi_\theta(x_1,x_2,Y_v)\|^b\right]\right)^{1/b}$ (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 $q \le j$. Note that if $q > 0$, then the first vertex after $u\wedge v = \mathrm{p}^j(u)$ on the path from u to v cannot be $\mathrm{p}^{j-1}(u)$ and must be the other children of $\mathrm{p}^j(u)$. Thus, there are 2^{q-1} choices of v with the same coding (j,q) with $0 < q \le j$. Hence, we get:

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

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

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

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

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

4.1.3. Asymptotic normality of the score

Define the limiting Fisher information as:

$$\mathcal{I}(\theta^\star) = \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}\left[\dot{\mathrm{h}}_{\partial,\infty}(\theta^\star)\dot{\mathrm{h}}_{\partial,\infty}(\theta^\star)^t\right], \tag{54}$$

where we see $\dot{\mathrm{h}}_{\partial,\infty}(\theta^\star)$ 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:

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

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 ρ 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 ρ.

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

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

where $\mathcal{N}(0, M)$ denotes the centered Gaussian distribution with covariance matrix M, and $\mathcal{I}(\theta^\star)$ 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 $\pi_{\theta^\star}$ denotes the invariant distribution for the hidden process X associated with $Q_{\theta^\star}$. Define the stationary score $\nabla_\theta \ell_{n,\pi_{\theta^\star}}(\theta)$ as:

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

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:

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

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

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

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

$$\lim_{n\to\infty} \frac{1}{|T_n|^{1/2}} \Big(\nabla_\theta \ell_{n,x}(\theta^\star) - \nabla_\theta \ell_{n,\pi_{\theta^\star}}(\theta^\star)\Big) = 0 \qquad \text{in } L^2(\mathbb{P}_{\theta^\star}). \tag{55}$$

In particular, to prove asymptotic normality for the score $\nabla_\theta \ell_{n,x}(\theta^\star)$ for any $x \in \mathcal{X}$, it is enough to prove asymptotic normality for the stationary score $\nabla_\theta \ell_{n,\pi_{\theta^\star}}(\theta^\star)$ (see for instance (Billingsley, 1999, Theorem 3.1)).

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

$$\dot{\mathrm{h}}_{u,k,\pi_{\theta^\star}}(\theta) := \int_{\mathcal{X}} \dot{\mathrm{h}}_{u,k,x}(\theta)\, \pi_{\theta^\star}(\mathrm{d}x). \tag{56}$$

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

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

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

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

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:

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

Also note that Assumption 7 (on page 3399) implies that:

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

Thus, we have:

$$\begin{aligned}&\mathbb{E}_{\theta^\star}\Big[\mathbb{E}_{\theta^\star}[\phi_{\theta^\star}(X_{\mathrm{p}(u_j)}, X_{u_j}, Y_{u_j}) \,|\, Y_{\Delta(u_j,k)}] \;\Big|\; \mathcal{F}_{j-1}\Big] \\ &\qquad = \mathbb{E}_{\theta^\star}[\phi_{\theta^\star}(X_{\mathrm{p}(u_j)}, X_{u_j}, Y_{u_j}) \,|\, Y_{\Delta^*(u_j,k)}] \\ &\qquad = \mathbb{E}_{\theta^\star}\Big[\mathbb{E}_{\theta^\star}[\phi_{\theta^\star}(X_{\mathrm{p}(u_j)}, X_{u_j}, Y_{u_j}) \,|\, Y_{\Delta^*(u_j,k)}, X_{\mathrm{p}(u_j)}] \;\Big|\; Y_{\mathbf{\Delta}^*(u_j,k)}\Big] \\ &\qquad = \mathbb{E}_{\theta^\star}\Big[\mathbb{E}_{\theta^\star}[\phi_{\theta^\star}(X_{\mathrm{p}(u_j)}, X_{u_j}, Y_{u_j}) \,|\, X_{\mathrm{p}(u_j)}] \;\Big|\; Y_{\Delta^*(u_j,k)}\Big] \\ &\qquad = 0,\end{aligned}$$

where we used the Markov property for the inner expectation in the third equality. Moreover, it is immediate that $\dot{\mathrm{h}}_{u_j,k,\pi_{\theta^\star}}(\theta^\star)$ is $\mathcal{F}_j$-measurable for all $j \in \mathbb{N}^*$ and $1 \leq k \leq h(j)$. Hence, we get that the sequence $\big(\dot{\mathrm{h}}_{u_j,h(u_j),\pi_{\theta^\star}}(\theta^\star)\big)_{j \in \mathbb{N}^*}$ is a $\mathbb{P}_{\theta^\star}$-martingale increment sequence

adapted to the filtration $\mathcal{F} = (\mathcal{F}_j)_{j\in\mathbb{N}}$ in $L^2(\mathbb{P}_{\theta^\star})$ (thanks to Assumption 7). We are going to apply a central limit theorem for martingales (see (Duflo, 2011, Corollary 2.1.10)). For all $n \in \mathbb{N}$, define $M_n = \sum_{j=0}^{n} \dot{\mathrm{h}}_{u_j,h(u_j),\pi_{\theta^\star}}(\theta^\star)$. Note that the first term $M_0 = \dot{\mathrm{h}}_{\partial,0,\pi_{\theta^\star}}(\theta^\star) = \int_{\mathcal{X}} \nabla_\theta \log g_{\theta^\star}(x, Y_\partial)\, \pi_{\theta^\star}(\mathrm{d}x)$ is in $L^2(\mathbb{P}_{\theta^\star})$ by Assumption 7. Hence, the sequence $(M_n)_{n\in\mathbb{N}}$ is a $\mathbb{P}_{\theta^\star}$-martingale sequence adapted to the filtration $\mathcal{F} = (\mathcal{F}_j)_{j\in\mathbb{N}}$ in $L^2(\mathbb{P}_{\theta^\star})$, and whose quadratic variation is:

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

where, as in (54), we see $\dot{\mathrm{h}}_{u_j,h(u_j),\pi_{\theta^\star}}(\theta^\star)$ 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 the convergence $\lim_{n\to\infty} n^{-1}\langle M\rangle_n = \mathcal{I}(\theta^\star)$ holds in $\mathbb{P}_{\theta^\star}$-probability. Indeed, we will prove that this convergence holds in $L^2(\mathbb{P}_{\theta^\star})$. 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 $\sup_{x\in\mathcal{X}} \dot{\mathrm{h}}_{u,k,x}(\theta^\star, Y_{\Delta(u,k)})$ is in $L^4(\mathbb{P}_{\theta^\star})$, and thus the random variable $\sup_{x\in\mathcal{X}} \dot{\mathrm{h}}_{u,k,x}(\theta^\star, Y_{\Delta(u,k)})\dot{\mathrm{h}}_{u,k,x}(\theta^\star, Y_{\Delta(u,k)})^t$ is in $L^2(\mathbb{P}_{\theta^\star})$ 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)):

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

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 $L^2(\mathbb{P}_{\theta^\star})$ as $n \to \infty$, 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:

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

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 3383), 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 $\dot{\mathrm{h}}_{u,k,x}(\theta^\star, Y_{\Delta(u,k)})$ is in $L^4(\mathbb{P}_{\theta^\star})$, and thus the random variable $\Psi_{u,k,x}(Y_{\Delta^*(u,k)})$ is in $L^2(\mathbb{P}_{\theta^\star})$ 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 $L^2(\mathbb{P}_{\theta^\star})$ to $\mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}\left[\dot{\mathrm{h}}_{\partial,k,x}(\theta^\star)\dot{\mathrm{h}}_{\partial,k,x}(\theta^\star)^t\right]$ as $n \to \infty$. Using Lemma 4.2, we have that $\lim_{k\to\infty} \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}\left[\dot{\mathrm{h}}_{\partial,k,x}(\theta^\star)\dot{\mathrm{h}}_{\partial,k,x}(\theta^\star)^t\right] = \mathcal{I}(\theta^\star)$. Combining those facts with (58), we get that $\lim_{n\to\infty} n^{-1}\langle M\rangle_n = \mathcal{I}(\theta^\star)$ in $L^2(\mathbb{P}_{\theta^\star})$.

Step 4: Lindeberg's condition holds. We now need to verify that Lindeberg's condition holds (see (Duflo, 2011, Corollary 2.1.10)), that is, to prove for all $\varepsilon > 0$ that $\lim_{n\to\infty} F_n(\varepsilon\sqrt{n}) = 0$ in $\mathbb{P}_{\theta^\star}$-probability where for all $n \in \mathbb{N}^*$ and $A \in \mathbb{R}_+$:

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

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

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

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

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

Let $\varepsilon > 0$. Setting $A_n = \varepsilon\sqrt{n}$ for all $n \in \mathbb{N}^*$, we get that the convergence $\lim_{n\to\infty} F_n(\varepsilon\sqrt{n}) = 0$ holds in $L^1(\mathbb{P}_{\theta^\star})$, and thus also in $\mathbb{P}_{\theta^\star}$-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 (Duflo, 2011, Corollary 2.1.10)), which gives us that $\mathbb{P}_{\theta^\star}$-a.s. $\lim_{n\to\infty} n^{-1}M_n = 0$ and that the sequence $(n^{-1/2}M_n)_{n\in\mathbb{N}^*}$ converges in $\mathbb{P}_{\theta^\star}$-distribution towards a centered Gaussian distribution $\mathcal{N}(0, \mathcal{I}(\theta^\star))$ whose covariance matrix is $\mathcal{I}(\theta^\star)$. In particular, using (55), we get that $\mathbb{P}_{\theta^\star}$-a.s. $\lim_{n\to\infty} |T_n|^{-1}\nabla_\theta \ell_{n,x}(\theta^\star) = 0$ and that:

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

This concludes the proof of the theorem. □

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 $\mathbb{P}_{\theta^\star}$-a.s. $\lim_{n\to\infty} \theta_n = \theta^\star$, then the normalized observed information $-n^{-1}\nabla^2_\theta \ell_{n,x}(\theta_n)$ converges $\mathbb{P}_{\theta^\star}$-a.s. as $n \to \infty$ to the limiting Fisher information matrix $\mathcal{I}(\theta^\star)$ which is defined in (54).

Remind the definition of the log-likelihood $\ell_{n,x}(\theta)$ in (7) on page 3382. 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 (Cappé, Moulines and Rydén, 2005, Proposition 10.1.6)), we get for all $\theta \in O$ and $x \in \mathcal{X}$:

$$\nabla_\theta^2 \ell_{n,x}(\theta) = \nabla_\theta^2 \log(g_\theta(X_\partial, Y_\partial)) + \mathbb{E}_\theta \left[\sum_{u \in T_n^*} \varphi_\theta(X_{\mathrm{p}(u)}, X_u, Y_u) \middle| Y_{T_n}, X_\partial = x \right] + \mathrm{Var}_\theta \left[\sum_{u \in T_n^*} \phi_\theta(X_{\mathrm{p}(u)}, X_u, Y_u) \middle| Y_{T_n}, X_\partial = x \right],$$

where remind that ϕ_θ is defined in (46) on page 3401, and φ_θ is defined as:

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

Note that similarly to the case of ϕ_θ, the random variale $\varphi_\theta(X_{\mathrm{p}(u)}, X_u, Y_u)$ 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 $\mathbb{P}_{\theta^\star}$-a.s. finite by Assumption 7-(iii).

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

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

and:

$$\Gamma_{u,k,x}(\theta) = \mathrm{Var}_\theta \left[\sum_{v \in \mathbf{\Delta}(u,k) \setminus \{\mathrm{p}^k(u)\}} \phi_\theta(X_{\mathrm{p}(v)}, X_v, Y_v) \middle| Y_{\mathbf{\Delta}(u,k)}, X_\partial = x_\partial \right] - \mathrm{Var}_\theta \left[\sum_{v \in \mathbf{\Delta}^*(u,k) \setminus \{\mathrm{p}^k(u)\}} \phi_\theta(X_{\mathrm{p}(v)}, X_v, Y_v) \middle| Y_{\mathbf{\Delta}^*(u,k)}, X_\partial = x_\partial \right], \tag{63}$$

where Var_θ (resp. Cov_θ) 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 $Y_{\mathbf{\Delta}(u,k)}$ 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 $\theta \in O$ and $x \in \mathcal{X}$ as:

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

To prove the convergence of the two sums in the right hand side of (64), and thus 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_θ of the HMT.

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

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

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 $\mathbb{P}_{\theta^\star}$-a.s. and in $L^2(\mathbb{P}_{\theta^\star})$ when $k \to \infty$. 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 $\mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}[\Lambda_{\partial,\infty}(\theta)]$ and $\mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}[\Gamma_{\partial,\infty}(\theta)]$, respectively, with some uniformity in θ near $\theta^\star$.

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 ρ 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 ρ.

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

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

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

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

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 ρ 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 ρ.

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

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

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

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

With Propositions 4.4 and 4.5, we are now ready to prove the following theorem which states that the normalized observed information $-|T_n|^{-1}\nabla^2_\theta \ell_{n,x}(\theta_n)$ converges $\mathbb{P}_{\theta^\star}$-a.s. locally uniformly to the limiting Fisher information $\mathcal{I}(\theta^\star)$ (which is defined in (54)). Note that the condition $\rho < 1/2$ on the mixing rate ρ 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 ρ.

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 Θ is compact. Then, for all $x \in \mathcal{X}$, we have:*

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

As an immediate corollary, for any possibly random sequence $(\theta_n)_{n \in \mathbb{N}}$ such that $\mathbb{P}_{\theta^\star}$-a.s. $\lim_{n \to \infty} \theta_n = \theta^\star$ and for any $x \in \mathcal{X}$, we get that:

$$\mathbb{P}_{\theta^\star}\text{-a.s.} \qquad \lim_{n \to \infty} -|T_n|^{-1} \nabla_\theta^2 \ell_{n,x}(\theta_n) = \mathcal{I}(\theta^\star).$$

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 3394), 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 $\mathcal{I}(\theta^\star)$.

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

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 (Rasch and Schott, 2018, p.21) or (Cappé, Moulines and Rydén, 2005, p.355)), we get for all $\theta \in O$ and $x \in \mathcal{X}$:

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

Setting $\theta = \theta^\star$ and taking the expectation over X_∂, we get:

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

Using (64) on page 3410, Propositions 4.4 and 4.5 give us that the right hand side of (68) converges as $n \longrightarrow \infty$ to $-\mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}[\Lambda_{\partial,\infty}(\theta^\star) + \Gamma_{\partial,\infty}(\theta^\star)]$.

Remind, using (48) along with Assumption 8, that $\dot{\mathrm{h}}_{u,k,x}(\theta^\star, Y_{\Delta(u,k)})$ is in $L^4(\mathbb{P}_{\theta^\star})$, and thus the random variable $\dot{\mathrm{h}}_{u,k,x}(\theta^\star, Y_{\Delta(u,k)}) \dot{\mathrm{h}}_{u,k,x}(\theta^\star, Y_{\Delta(u,k)})^t$ is in $L^2(\mathbb{P}_{\theta^\star})$ for every $u \in T \backslash T_{k-1}$. 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:

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

where remind that we see $\dot{\mathrm{h}}_{u,k,x}(\theta^\star)$ 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:

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

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

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

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 $\mathcal{I}(\theta^\star)^{-1}$ where the Fisher information matrix $\mathcal{I}(\theta^\star)$ is defined in (54) on page 3405. Recall that the contrast function ℓ is defined in (26) on page 3390, that the MLE $\hat{\theta}_{n,x}$ is defined in (33) on page 3394, and that the mixing rate ρ of the HMT process (X, Y) is defined after Assumption 3 on page 3380.

Note that the condition $\rho < 1/2$ on the mixing rate ρ 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 ρ.

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 ℓ has a unique maximum (which is then located at $\theta^\star \in \Theta$ by Proposition 3.10) and that Θ is compact, $\theta^\star$ is an interior point of Θ, and the limiting Fisher information matrix $\mathcal{I}(\theta^\star)$ (which is defined in (54)) is non-singular. Then, for all $x \in \mathcal{X}$, we have:*

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

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 (Bickel, Ritov and Rydén, 1998, 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 $\theta^\star$, we get:

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

where we see $\hat{\theta}_{n,x}$ and $\theta^\star$ as column vectors. As $\mathcal{I}(\theta^\star)$ is non-singular (indeed definite positive), remark that Theorems 3.11 and 4.6 insure that $\mathbb{P}_{\theta^\star}$-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 $\mathbb{P}_{\theta^\star}$-a.s. for n large enough:

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

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

4.2.1. *Proof of Proposition 4.4*

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

Let Θ_0 be a compact subset of Θ, Let Θ_0 be a closed ball in Θ, and let $\varphi : \Theta_0 \times \mathcal{X}^2 \times \mathcal{Y} \to \mathbb{R}$ 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 Θ_0, and such that:

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

Let $\Lambda_{u,k,x}(\theta)$ be defined as in (62) on page 3410 and note that it is in $L^2(\mathbb{P}_\mathcal{U} \otimes \mathbb{P}_{\theta^\star})$.

The proof of Proposition 4.4 is decomposed into several lemmas.

We start with the following lemma, stating a uniform $L^2(\mathbb{P}_\mathcal{U} \otimes \mathbb{P}_{\theta^\star})$ 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 $L^2(\mathbb{P}_\mathcal{U} \otimes \mathbb{P}_{\theta^\star})$ such that for all $k \in \mathbb{N}^*$, we have:*

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

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 $L^2(\mathbb{P}_\mathcal{U} \otimes \mathbb{P}_{\theta^\star})$ to $\Lambda_{u,\infty}(\theta)$ which does not depend on x.

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

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

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:

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

$$\begin{aligned}
&\leq \mathbb{E}_{\theta^\star}\left[\sup_{x,x'\in\mathcal{X}}\left|\frac{1}{|T_n|}\sum_{u\in T^*_{k-1}}\Lambda_{u,h(u),x}(\theta)\right|^2\right]^{1/2}\\
&\quad+\mathbb{E}_{\theta^\star}\left[\sup_{x,x'\in\mathcal{X}}\left|\frac{1}{|T_n|}\sum_{u\in T_n\setminus T_{k-1}}\Lambda_{u,h(u),x}(\theta)-\Lambda_{u,k,x'}(\theta)\right|^2\right]^{1/2}\\
&\quad+\mathbb{E}_{\theta^\star}\left[\left|\frac{1}{|T_n|}\sum_{u\in T_n\setminus T_{k-1}}\Lambda_{u,k,x'}(\theta)-\mathbb{E}_{\mathcal{U}}\otimes\mathbb{E}_{\theta^\star}\left[\Lambda_{\partial,k,x'}(\theta)\right]\right|^2\right]^{1/2}\\
&\quad+\mathbb{E}_{\mathcal{U}}\otimes\mathbb{E}_{\theta^\star}\left[|\Lambda_{\partial,k,x'}(\theta)-\Lambda_{\partial,\infty}(\theta)|^2\right]^{1/2}.
\end{aligned}\tag{70}$$

Using Lemma 4.8 together with (49) on page 3403 (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 3383), 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 φ_θ 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 $L^2(\mathbb{P}_{\theta^\star})$ 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. □

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

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

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

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

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

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 $\mathbb{P}_{\mathcal{U}}\otimes\mathbb{P}_{\theta^\star}$-a.s. continuous on Θ_0. Moreover, for all $\theta \in \Theta_0$, we have:*

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

Proof. We mimic the proof of (Douc, Moulines and Rydén, 2004, 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 $L^2(\mathbb{P}_{\theta^\star})$. Thus, for all $v \in T^\infty$, the random variable $\|\varphi^v\|_\infty$ is in $L^2(\mathbb{P}_{\theta^\star})$. Remind from (12) on page 3386 that $\mathbf{\Lambda}(\partial, k)$ is a random subtree of the deterministic subtree $T^\infty(\mathrm{p}^k(u), k)$. Then, note that we have:

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

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

$$\lim_{\delta\to 0}\sup_{\theta'\in\Theta_0:\|\theta'-\theta\|\le\delta}\Big|\mathbb{E}_{\theta'}[\varphi_{\theta'}(X_{\mathrm{p}(v)}, X_v, Y_v)\,|\,Y_{\mathbf{\Lambda}(\partial,k)}, X_{\mathrm{p}^k(\partial)} = x] - \mathbb{E}_{\theta}[\varphi_{\theta}(X_{\mathrm{p}(v)}, X_v, Y_v)\,|\,Y_{\mathbf{\Lambda}(\partial,k)}, X_{\mathrm{p}^k(\partial)} = x]\Big| = 0, \qquad \mathbb{P}_{\mathcal{U}}\otimes\mathbb{P}_{\theta^\star}\text{-a.s.}$$

Denote $x_{\mathrm{p}^k(\partial)} = x$ and $\mathbf{\Lambda}^{\mathrm{o}}(\partial, k) = \mathbf{\Lambda}(\partial, k) \setminus \{\mathrm{p}^k(\partial)\}$, and write:

$$\mathbb{E}_{\theta}[\varphi_{\theta}(X_{\mathrm{p}(v)}, X_v, Y_v)\,|\,Y_{\mathbf{\Lambda}(\partial,k)}, X_{\mathrm{p}^k(\partial)} = x] = \frac{\int_{\mathcal{X}^{\mathbf{\Lambda}^{\mathrm{o}}(\partial,k)}}\varphi_{\theta}(x_{\mathrm{p}(v)}, x_v, Y_v)\prod_{w\in\mathbf{\Lambda}^{\mathrm{o}}(\partial,k)}q_\theta(x_{\mathrm{p}(w)}, x_w)g_\theta(x_w, Y_w)\lambda(\mathrm{d}x_w)}{\int_{\mathcal{X}^{\mathbf{\Lambda}^{\mathrm{o}}(\partial,k)}}\prod_{w\in\mathbf{\Lambda}^{\mathrm{o}}(\partial,k)}q_\theta(x_{\mathrm{p}(w)}, x_w)g_\theta(x_w, Y_w)\lambda(\mathrm{d}x_w)}. \tag{72}$$

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. θ and is upper bounded by the random variable $\|\varphi^v\|_\infty(\sigma^+b^+)^{|T^\infty(\mathrm{p}^k(u),k)|-1}$ (remind that $\sigma^+ \ge 1$ and

$b^+ \geq 1$). And similarly, the denominator is continuous w.r.t. θ, and, using Assumption 3-(ii), is lower bounded by the random variable:

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

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

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

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

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

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

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

Using Lemma 4.11, we get:

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

and taking $k \to \infty$, the upper bound vanishes. This concludes the proof. □

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:*

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

Proof. First, write:

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

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:

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

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

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

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

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

Proof of Proposition 4.4. Applying Lemma 4.8, for all $u \in T$, we get that the sequence $(\Lambda_{u,k,x}(\theta))_{k \in \mathbb{N}^*}$ is a Cauchy sequence uniformly w.r.t. $\theta \in \Theta_0$ in $L^2(\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta\star})$ that converges to some limit $\Lambda_{u,\infty}(\theta)$ (that does not depend on x). By Corollary 4.10, we have that $\mathbb{P}_{\theta\star}$-a.s. the convergence for the the average of the quantities $\Lambda_{u,h(u),x}(\theta^\star)$ 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 $\theta \mapsto \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta\star}[\Lambda_{\partial,\infty}(\theta)]$ is continuous on Θ_0. Finally, the last part of the proposition is given by Lemma 4.13. □

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 ϕ_θ used in (63) to define $\Gamma_{u,k,x}(\theta)$ are replaced by scalar-valued functions, still denoted by φ_θ, under more general assumptions. The extension to matrix-valued functions is then straightforward by applying the result coordinate-wise.

Let Θ_0 be a compact subset of Θ, Let Θ_0 be a closed ball in Θ, and let $\phi : \Theta_0 \times \mathcal{X}^2 \times \mathcal{Y} \to \mathbb{R}$ 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 Θ_0, and such that:

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

Let $\Gamma_{u,k,x}(\theta)$ be defined as in (63) on page 3410 and note that it is in $L^2(\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta\star})$.

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 ν and ν' on $\mathcal{X}$, we have:*

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

For simplicity, Lemma 4.14 is stated with ∂ as the initial vertex, but note that the results still holds when replacing ∂ 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. We denote by $((X'_w, Y'_w), w \in T)$ the process started from the distribution ν' (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 ∂, 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 ∂ 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. □

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 $v, w \in T^\infty(\mathrm{p}^k(u), k) \setminus \{u\}$. Then, we have:*

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

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 $w_0 \in T^\infty(\mathrm{p}^k(u), k) \setminus \{u\}$. Note that w_0 is on the path from $\mathrm{p}(u)$ to v, and thus $d(\mathrm{p}(u), v) = d(\mathrm{p}(u), w_0) + d(w_0, v)$, and similarly when replacing v by w. On the path from $\mathrm{p}(u)$ to w_0, we use the coupling provided by the "backward in time" coupling bound from Lemma 4.1 with successful probability $1 - \rho^{d(\mathrm{p}(u),w_0)}$. 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 $\mathrm{p}(u)$ 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 $2\,\rho^{\min(d(u,v),d(u,w))-1}$. □

For brevity, for all $u \in T$ we will denote $\phi_{\theta,u} = \phi_\theta(X_{\mathrm{p}(u)}, X_u, Y_u)$ 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 3386 that $\boldsymbol{\Delta}(\partial, k)$ is a random subtree of the deterministic subtree $T^\infty(\mathrm{p}^k(u), k)$.

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 ρ 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 $v, w \in T^\infty(\mathrm{p}^k(u), k) \setminus \{\mathrm{p}^k(u)\}$, we have:*

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

and,

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

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

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

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 $\mathrm{p}(v)$), then using the Markov property of the HMT process (X, Y), we get:

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

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 $\mathrm{p}(u)$ replaced by w and using the initial distributions $\mathbb{P}((X_{\mathrm{p}(w)}, X_w) \in \cdot \,|\, Y_{\boldsymbol{\Delta}(u,k)}, X_w \in B_2, X_{\mathrm{p}^k(u)} = x)$ with $B' = B$ and $\mathcal{X}$ respectively).

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

$$\begin{aligned}&\Big|\mathbb{P}_\theta(X_{\{\mathrm{p}(v),v\}} \in A \,|\, Y_{\boldsymbol{\Delta}(u,k)}, X_{\mathrm{p}^k(u)} = x)\mathbb{P}_\theta(X_{\{\mathrm{p}(w),w\}} \in B \,|\, Y_{\boldsymbol{\Delta}(u,k)}, X_{\mathrm{p}^k(u)} = x) \\ &\quad -\mathbb{P}_\theta(X_{\{\mathrm{p}(v),v\}} \in A, X_{\{\mathrm{p}(w),w\}} \in B \,|\, Y_{\boldsymbol{\Delta}(u,k)}, X_{\mathrm{p}^k(u)} = x)\Big|\end{aligned}$$

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

Thus, in both case, we get:

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

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:

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

and then use the joint coupling Lemma 4.14 with $v' = \mathrm{p}(v)$ and $w' = \mathrm{p}(w)$ for the first term in the upper bound, and use the coupling Lemma 3.2 for the other two terms with $\mathrm{p}(v)$ and $\mathrm{p}(w)$, 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 $m = \min\{d(\mathrm{p}(u), w_0) : w_0 \in \{\mathrm{p}(v), v, \mathrm{p}(w), w\}\}$. Noting $m \leq \min(d(u,v), d(u,w)) - 2$, we get that (77) holds. This concludes the proof of the lemma. □

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 $L^2(\mathbb{P}_{\theta^\star})$ approximation bound.

Note that the condition $\rho < 1/2$ on the mixing rate ρ 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:*

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

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 θ is implicit). Similarly to the proof of (Douc, Moulines and Rydén, 2004, 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 $Y_{\mathbf{\Delta}(u,k')}$ and implicitly on $\mathcal{U}$, and are define as:

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

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 $\mathbb{E}_{\theta^\star}[\|\phi_v\|_\infty^2 \|\phi_w\|_\infty^2] \leq \mathbb{E}_{\theta^\star}[\|\phi_\partial\|_\infty^4] < \infty$ for all $v, w \in T^\infty$.

Remind from (12) on page 3386 that $\Delta(u,k)$ is a random subtree of the deterministic subtree $T^\infty(\mathrm{p}^k(u), k)$.

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

$$2 \sum_{v,w \in T^\infty(\mathrm{p}^k(u),k)} \big(2 \times 8\rho^{\min(d(v,u),d(w,u))-2} \wedge 2 \times 8\rho^{\min(d(v,\mathrm{p}^k(u)),d(w,\mathrm{p}^k(u)))-2} \wedge 4 \times 2\rho^{d(v,w)-2}\big)$$

$$\leq \frac{32}{\rho^2} \sum_{v,w \in T^\infty(\mathrm{p}^k(u),k)} \left(\rho^{\min(d(v,u),d(w,u))} \wedge \rho^{\min(d(v,\mathrm{p}^k(u)),d(w,\mathrm{p}^k(u)))} \wedge \rho^{d(v,w)}\right). \quad (78)$$

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

For all $j \in \mathbb{N}$, denote $u_j = \mathrm{p}^j(u)$. 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:

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

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

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

Note that for v, w in the third or fourth case, either $\min(d(v, \mathrm{p}^k(u)), d(w, \mathrm{p}^k(u)))$ 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 $\mathbb{E}_{\theta^\star}[|A|^2]^{1/2}$ is upper bounded by an expression as in the lemma.

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

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

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 ρ).

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

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

As annonce 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 $Q_{\theta^\star}$ and $G_{\theta^\star}$ that are parametrized by some $\theta^\star \in \Theta$ as before, and the hidden variable X'_∂ of the root vertex ∂ has distribution ζ. This initial distribution ζ is unknown to us, may depend on $\theta^\star$, and in general is different from the invariant distribution $\pi_{\theta^\star}$. As before, we will denote by $(X, Y) = ((X_u, Y_u) : u \in T)$ a stationary process distributed according to the same parameter $\theta^\star$. 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 3382, define its log-likelihood for all $n \in \mathbb{N}$ and $x \in \mathcal{X}$ as:

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

where $\ell_{n,x}(\theta;\cdot)$ is defined in (6) on page 3382. Moreover, when Assumptions 1–4 and 6 hold and Θ is compact, similarly to the stationary case in (33) on page 3394, we define the MLE $\hat{\theta}'_{n,x}$ for the non-stationary process as:

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

Denote by $\mathbb{P}_{\theta^\star,\zeta}$ the probability distribution of the non-stationary HMT process (X', Y'), and by $\mathbb{E}_{\theta^\star,\zeta}$ 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 ℓ has a unique maximum (which is then located at $\theta^\star \in \Theta$ by Proposition 3.10) and Θ is compact. Then, the MLE is strongly consistent, that is, for all initial distributions ζ and all $x \in \mathcal{X}$, the MLE $\hat{\theta}'_{n,x}$ converges $\mathbb{P}_{\theta^\star,\zeta}$-a.s. as $n \to \infty$ to the true parameter $\theta^\star \in \Theta$.*

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:

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

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

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

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

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

Hence, we get that (81) holds.

Using (81), we get that Proposition 3.7 also holds $\mathbb{P}_{\theta^\star,\zeta}$-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 ℓ, which is the same in the stationary and non-stationary cases). This concludes the proof of the Theorem. □

Using a similar argument as for Theorem 5.1, we get, in the non-stationary case, that the normalized observed information $-|T_n|^{-1}\nabla^2_\theta \ell'_{n,x}(\theta_n)$ converges $\mathbb{P}_{\theta^\star,\zeta}$-a.s. locally uniformly to the limiting Fisher information $\mathcal{I}(\theta^\star)$ (which is defined in (54)). Note that the condition $\rho < 1/2$ on the mixing rate ρ 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 ρ.

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 Θ is compact. Then, for all initial distributions ζ and all $x \in \mathcal{X}$, we have:*

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

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

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 $\sigma^- \ge 1/2$. 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 $n \ge N$.*

We will denote by $\mathbb{P}_{\theta^\star\bowtie\zeta}$ the probability distribution that realizes this coupling.

Note that $\rho \le 1/2$ implies that $\sigma^- \ge \sigma^+/2 \ge 1/2$ (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_∂ and X'_∂, which are distributed according to $\pi_{\theta^\star}$ and ζ respectively (note that it is also possible to use a perfect coupling with probability error $\|\pi_{\theta^\star} - \zeta\|_{\mathrm{TV}}$). 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 $\mathrm{p}(u)$, that is $X_{\mathrm{p}(u)} = X'_{\mathrm{p}(u)}$, then we choose the new value $X_u = X'_u$ according to the transition kernel $Q_{\theta^\star}$. Otherwise, if $X_{\mathrm{p}(u)} \ne X'_{\mathrm{p}(u)}$, then using the uniform geometric ergodicity (remind Assumption 3) of the transition kernel $Q_{\theta^\star}$, we know that $\sup_{x,x'\in\mathcal{X}} \|Q_{\theta^\star}(x;\cdot) - Q_{\theta^\star}(x';\cdot)\|_{\mathrm{TV}} \le 1-\sigma^-$, and thus we can construct a coupling of X_u and X'_u conditionally on $X_{\mathrm{p}(u)} \ne X'_{\mathrm{p}(u)}$ 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's extension theorem, there exists a coupling measure for the whole tree T whose finite dimensional marginals are the ones given above. Recall that Kolmogorov's extension theorem holds for general Polish spaces and not just the real line (see (Bogachev, 2007, Theorem 7.7.1 with Theorem 7.1.7)), and also recall that $\mathcal{X}$ and $\mathcal{Y}$, and thus also $\mathcal{X}\times\mathcal{Y}$, are Polish spaces.

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 $\mathbb{P}_{\theta^\star\bowtie\zeta}$ 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 $\mathbb{P}_{\theta^\star \bowtie \zeta}$-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 if finite $\mathbb{P}_{\theta^\star \bowtie \zeta}$-a.s., that is, N is finite $\mathbb{P}_{\theta^\star \bowtie \zeta}$-a.s., if and only if $2(1-\sigma^-) \leq 1$, that is, $\sigma^- \geq 1/2$. This concludes the proof. □

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 3382 (resp. just before (79) on page 3425). 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_θ denotes possibly conditional density (see (5) on page 3382).

Remind that when Assumptions 1–4 and 6 hold and Θ 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 3394 (resp. in (80) on page 3425).

To prove that $\lim_{n\to\infty} |T_n|^{1/2}(\hat{\theta}'_{n,x} - \hat{\theta}_{n,x}) = 0$ $\mathbb{P}_{\theta^\star \bowtie \zeta}$-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:

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

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 ζ and all $x \in \mathcal{X}$, we have:*

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

Proof. [The proof of the lemma is a straightforward adaptation of the proof of (Douc, Moulines and Rydén, 2004, Lemmas 11 and 12).]

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

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

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

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

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 $h(u) > N$:

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

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 $\mathbb{P}_{\theta^\star \bowtie \zeta}$-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 $\mathbb{P}_{\theta^\star \bowtie \zeta}$-a.s.

Let $\varepsilon > 0$. Using (82), there exists a random integer N_ε which $\mathbb{P}_{\theta^\star \bowtie \zeta}$-a.s. is finite and satifies (remind that $\Delta^*(u) = \Delta^*(u, h(u))$):

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

Thus, $\mathbb{P}_{\theta^\star \bowtie \zeta}$-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 $\mathbb{P}_{\theta^\star \bowtie \zeta}$-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 $\mathbb{P}_{\theta^\star \bowtie \zeta}$-a.s. to $\theta^\star$ (see Theorem 5.1), and that ε was arbitrary. □

We can now prove the asymptotic normality of the MLE $\hat\theta'_{n,x}$ in the non-stationary case. Remind that the contrast function ℓ is defined in (26) on page 3390.

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 ℓ has a unique maximum (which is then located at $\theta^\star \in \Theta$ by Proposition 3.10) and that Θ is compact, $\theta^\star$ is an interior point of Θ, and the limiting Fisher information matrix $\mathcal{I}(\theta^\star)$ (which is defined in (54)) is non-singular. Then, for all initial distributions ζ and for all $x \in \mathcal{X}$, we have:*

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

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

Proof. [The proof of the theorem is a straightforward adaptation of the proof of (Douc, Moulines and Rydén, 2004, 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 $\lim_{n\to\infty} \varepsilon_n = 0$ $\mathbb{P}_{\theta^\star \bowtie \zeta}$-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:

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

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

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

As $\mathcal{I}(\theta^\star)$ is positive definite, there exits $M > 0$ such that on a set with $\mathbb{P}_{\theta^\star \bowtie \zeta}$-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. □

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

In this appendix, we prove generalization of the ergodic theorems in Guyon (2007) and in Weibel (2025), 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 $\mathrm{h}_{u,k,x}(\theta; Y_{\Delta(u,k)})$ (defined in (17) on page 3387) 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 π 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 $\|Q^n(x;\cdot) - \pi\|_{\mathrm{TV}} \leq C\rho^n$. 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 π. 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 ν 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 νQ for the image probability measure $(\nu Q)(\cdot) = \int_{\mathcal{X}} Q(x;\cdot)\,\nu(\mathrm{d}x)$, and Qf for the Borel function $(Qf)(x) = \int_{\mathcal{Z}} f(z)\,Q(x;\mathrm{d}z)$. For a probability measure ν on $\mathcal{X}$ and a Borel integrable function f on $\mathcal{X}$, we will write $\langle \nu, f\rangle = \nu f = \int_{\mathcal{X}} f\,\mathrm{d}\nu$.

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 π, 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:*

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

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 π-almost everywhere and are π-integrable.

Proof. Using (Douc et al., 2018, 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 $\pi (Q^n f - \langle \pi, f \rangle)^2 = \| Q^n (f - \langle \pi, f \rangle) \|^2_{L^2(\pi)} \leq M \alpha^{2n} \| f - \langle \pi, f \rangle \|^2_{L^2(\pi)}$ for all $n \in \mathbb{N}$. In particular, we get that $\sup_{n \in \mathbb{N}} \pi Q^n f^2 < \infty$. □

Let $k \in \mathbb{N}$ be fixed. Remind from Section 2.4 the definitions of the subtrees $\Delta(u,k)$, of their shapes $\mathcal{S}h\big(\Delta(u,k)\big)$ (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 $\mathcal{S}h\big(\Delta(u,k)\big)$. Also remind from Section 2.4 the definition of a collection of neighborhood-shape-dependent functions $(f_{\mathcal{S}} : \mathcal{X}^{\mathcal{S}} \to \mathbb{R})_{\mathcal{S} \in \mathcal{N}_k}$. 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 $\mathrm{f} = (f_{\mathcal{S}} : \mathcal{X}^{\mathcal{S}} \to \mathbb{R})_{\mathcal{S} \in \mathcal{N}_k}$, define the empirical average of f over a finite subset $A \subset T \setminus T_{k-1}$ as:

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

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:

$$Q^{\mathcal{S}} f(x) = \mathbb{E}_Q \Big[f(X_{\Delta(u)}) \;\Big|\; X_\partial = x \Big]. \tag{84}$$

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:

$$Q^{\mathcal{S}_u} f(x) = \mathbb{E}_Q \Big[f(X_{\Delta(u,k)}) \;\Big|\; X_{\mathrm{p}^k(u)} = x \Big]. \tag{85}$$

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 $n \to \infty$.

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 $\bar{M}_{A_n}(\mathrm{f})$ 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 π. Let $X = (X_u, u \in T)$ be a branching Markov chain with transition kernel Q and initial distribution π.

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

$$\bar{M}_{A_n}(\mathrm{f}) \underset{n\to\infty}{\longrightarrow} \mathbb{E}_{U_k} \otimes \mathbb{E}_Q\left[f_{\mathcal{S}_{U_k}}(X_{\Delta(U_k)})\right] \qquad \text{in } L^2(\pi) = L^2(X), \tag{86}$$

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) O_u of the the k-neighborhood $B_T(u, k) := \{v \in T : d(u, v) \leq k\}$ of u such that O_u contains u. In that case, we must assume that the distribution of the shape (i.e. when seen up to translation) $\mathcal{S}h(O_{V_n})$ converges to some limit distribution. Also note that we could allow more general choices as in Weibel (2025) 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:

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

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

General case: We adapt the proof of (Weibel, 2025, Theorem 2.2) where the single function is replaced by a family 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:

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

Using Lemma A.1, as the transition kernel Q is uniformly geometrically ergodic and has unique invariant probability measure π, 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:

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

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

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

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 $h(u) \ge h(v)$. 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:

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

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 $d/2 \le \tilde{d} \le d$.

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

$$\forall m \in \mathbb{N}, \qquad \sup_{j\in\mathbb{N}} \pi Q^j \left(Q^m Q^{\mathcal{S}} f_{\mathcal{S}}\right)^2 \le M^2 \alpha^{2m}. \tag{91}$$

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

$$\mathbb{E}_Q\left[\bar{M}_{A_n}(\mathrm{f})^2\right]$$

$$\leq C\,\mathbb{P}_{U_n,V_n}(\tilde{d}(U_n,V_n)\leq 2K)$$
$$+ C^{1/2}\,|A_n|^{-2}\sum_{u,v\in A_n:\tilde{d}(u,v)>2K}\left(\max_{\mathcal{S}\in\mathcal{N}_k}\pi Q^{h(u\wedge v)}\left(Q^{\tilde{d}(u,v)-K}Q^{\mathcal{S}}f_{\mathcal{S}}\right)^2\right)^{1/2}$$
$$\leq C\,\mathbb{P}_{U_n,V_n}(d(U_n,V_n)\leq 4K)$$
$$+ C^{1/2}\,|A_n|^{-2}\sum_{u,v\in A_n:\tilde{d}(u,v)>2K} M\alpha^{\tilde{d}(u,v)-K} \tag{92}$$
$$\leq C\,\mathbb{P}_{U_n,V_n}(d(U_n,V_n)\leq 4K)+C^{1/2}M\alpha^K. \tag{93}$$

Let $\varepsilon>0$. Let $K\geq k$ be such that $C^{1/2}M\alpha^K<\varepsilon$. Using (Weibel, 2025, Lemma 3.1), we get that the first term in the right hand side of (93) goes to zero as $n\to\infty$. Thus, for n large enough, the right hand side of (93) is upper bounded by 2ε. This being true for all $\varepsilon>0$, we get that $\lim_{n\to\infty}\mathbb{E}_Q\left[\bar{M}_{A_n}(\mathrm{f})^2\right]=0$. This concludes the proof. □

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 π. Let $X=(X_u,u\in T)$ be a branching Markov chain with transition kernel Q and initial distribution π.*

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 $\mathrm{f}=(f_{\mathcal{S}}:\mathcal{X}^{\mathcal{S}}\to\mathbb{R})_{\mathcal{S}\in\mathcal{N}_k}$ be a collection of neighborhood-shape-dependent Borel functions that are in $L^2(X)$.

Then, we have:

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

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

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

Proof. After using (92), the proof is an easy adaptation of the proof of (Guyon, 2007, Theorem 14).

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

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

Thus, it it enough to treat the case of $\bar{M}_{G_n}(\mathrm{f})$. In the case where all functions $f_{\mathcal{S}}$ for $\mathcal{S}\in\mathcal{N}_k$ are constant, writing $\bar{M}_{G_n}(\mathrm{f})$ as in (87), and using the convergence in distribution of $(\mathcal{S}_{U_n})_{n\in\mathbb{N}}$ the uniform distribution over $\mathcal{N}_k$, we get that the sought convergence holds a.s. for $\bar{M}_{G_n}(\mathrm{f})$. 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 $\sum_{n\geq k}\mathbb{E}[\bar{M}_{G_n}(\mathrm{f})^2]<\infty$, 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:

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

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

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

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

and in particular remark that we have $U_0 = u \notin \boldsymbol{\Delta}^-(u,k,j)$ and $U_k = \mathrm{p}^k(u) \notin \boldsymbol{\Delta}^-(u,k,j)$.

Case 1: v is an ancestor of u. We mimic the proof of (Cappé, Moulines and Rydén, 2005, Proposition 12.5.4). The proof of the first case relies on the observation that conditioned on $X_{\mathrm{p}^k(u)}$ and $Y_{\boldsymbol{\Delta}(u,k)}$, the backward ancestral process X from $U_0 = u$ to $U_k = \mathrm{p}^k(u)$ is a non-homogeneous Markov chain satisfying a uniform mixing condition. The fact that $(X_{U_j})_{0\le j\le k}$ is a Markov chain comes from the Markov property of the HMT (X,Y) (remind the discussion around (2) on page 3379) which gives for all $j \in \{1, \cdots, k\}$:

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

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

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

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

$$\mathrm{B}_{x_{U_k},j}[y_{\boldsymbol{\Delta}(u,k)}](x_{U_{j-1}}; f)$$

$$= \mathbb{E}_\theta\left[f(X_{U_j}) \;\middle|\; Y_{\mathbf{\Delta}^-(u,k,j)} = y_{\mathbf{\Delta}^-(u,k,j)}, X_{U_k} = x_{U_k}, X_{U_{j-1}} = x_{U_{j-1}}\right]$$
$$= \frac{\int_{\mathcal{X}} f(x_{U_j})\, p_\theta(y_{\mathbf{\Delta}^-(u,k,j)}, x_{U_j} \mid X_{U_k} = x_{U_k}) q_\theta(x_{U_j}, x_{U_{j-1}})\, \lambda(\mathrm{d}x_{U_j})}{\int_{\mathcal{X}} p_\theta(y_{\mathbf{\Delta}^-(u,k,j)}, x_{U_j} \mid X_{U_k} = x_{U_k}) q_\theta(x_{U_j}, x_{U_{j-1}})\, \lambda(\mathrm{d}x_{U_j})}, \tag{96}$$

where:

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

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

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

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):

$$\frac{\sigma^-}{\sigma^+} \nu_{x_{U_k},j}[y_{\mathbf{\Delta}(u,k)}](f) \le \mathrm{B}_{x_{U_k},j}[y_{\mathbf{\Delta}(u,k)}](x_{U_{j-1}}; f),$$

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

$$\nu_{x_{U_k},j}[y_{\mathbf{\Delta}(u,k)}](f) = \mathbb{E}_\theta\left[f(X_{U_j}) \;\middle|\; Y_{\mathbf{\Delta}^-(u,k,j)} = y_{\mathbf{\Delta}^-(u,k,j)}, X_{U_k} = x_{U_k}\right]$$
$$= \frac{\int_{\mathcal{X}} f(x_{U_j})\, p_\theta(y_{\mathbf{\Delta}^-(u,k,j)}, x_{U_j} \mid X_{U_k} = x_{U_k})\, \lambda(\mathrm{d}x_{U_j})}{\int_{\mathcal{X}} p_\theta(y_{\mathbf{\Delta}^-(u,k,j)}, x_{U_j} \mid X_{U_k} = x_{U_k})\, \lambda(\mathrm{d}x_{U_j})},$$

where note that the only difference with the definition of $\mathrm{B}_{x_{U_k},j}[y_{\mathbf{\Delta}(u,k)}](x_{U_{j-1}}; f)$ is that the term $q_\theta(x_{U_j}, x_{U_{j-1}})$ has disappeared in both the numerator and the denominator of $\nu_{x_{U_k},j}[y_{\mathbf{\Delta}(u,k)}](f)$. 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:

$$\mathcal{L}(X_{U_1} \mid Y_{\mathbf{\Delta}(u,k)}, X_{U_k}, X_{U_0}) = \mathcal{L}(X_{U_1} \mid Y_{\mathbf{\Delta}^-(u,k,j)}, X_{U_k}, X_{U_0})$$
$$= \mathcal{L}(X_{U_1} \mid Y_{\Delta^*(u,k)}, X_{U_k}, X_{U_0}). \tag{97}$$

Finally, if we write:

$$\mathbb{P}_\theta(X_v \in \cdot \mid Y_{\mathbf{\Delta}(u,k)} = y_{\mathbf{\Delta}(u,k)}, X_{U_k} = x_{U_k})$$
$$= \int \mathbb{P}_\theta(X_v \in \cdot \mid Y_{\mathbf{\Delta}^-(u,k,1)} = y_{\mathbf{\Delta}^-(u,k,1)}, X_{U_k} = x_{U_k}, X_{U_1} = x_{U_1}) \times \mathbb{P}_\theta(X_{U_1} \in \mathrm{d}x_{U_1} \mid Y_{\mathbf{\Delta}(u,k)} = y_{\mathbf{\Delta}(u,k)}, X_{U_k} = x_{U_k}), \tag{98}$$

and we also write (using (97)):

$$\mathbb{P}_\theta(X_v \in \cdot \mid Y_{\mathbf{\Delta}^*(u,k)} = y_{\mathbf{\Delta}^*(u,k)}, X_{U_k} = x_{U_k})$$

$$= \int \mathbb{P}_\theta(X_v \in \cdot \mid Y_{\mathbf{\Delta}^-(u,k,1)} = y_{\mathbf{\Delta}^-(u,k,1)}, X_{U_k} = x_{U_k}, X_{U_1} = x_{U_1}) \times \mathbb{P}_\theta(X_{U_1} \in \mathrm{d}x_{U_1} \mid Y_{\mathbf{\Delta}^*(u,k)} = y_{\mathbf{\Delta}^*(u,k)}, X_{U_k} = x_{U_k}), \tag{99}$$

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 $X_{\mathrm{p}^k(u)}$ and $Y_{\mathbf{\Delta}(u,k)}$, if we consider the process X backward from u to $u \wedge v$ (remind that $v \in \mathbf{\Delta}^*(u,k)$) and then forward from $u \wedge v$ to v, we get a non-homogeneous Markov chain satisfying uniform mixing rate ρ. Note that as $v \in T^\infty(\mathrm{p}^k(u),k) \setminus \{u\}$, we have that $u \wedge v \in \{U_1, \cdots, U_k\}$. 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. □

Appendix 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 $h(u) \equiv h(v) \mod m+1$, 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 $h(u) \le h(v) + m$ 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 $\mathcal{U}$:

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

and $\mathbf{\Delta}(T(u,m),k) = \mathbf{\Delta}^*(T(u,m),k) \cup T(u,m)$. When $h(u) \ge k(m+1)$, then those subtrees do not depend on $\mathcal{U}$, and we simply write $\Delta^*(T(u,m),k) = \mathbf{\Delta}^*(T(u,m),k)$ and $\Delta(T(u,m),k) = \mathbf{\Delta}(T(u,m),k)$ to indicate it. See Figure 5 on page 3396 for an illustration of the "past" subtree $\mathbf{\Delta}^*(T(u,m),k)$ 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:

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

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:

$$\begin{aligned} \mathrm{H}_{T(u,m),k,x}(\theta) &= \frac{p_\theta(Y_{\mathbf{\Lambda}(T(u,m),k)} \mid X_{\mathrm{p}^k(u)} = x)}{p_\theta(Y_{\mathbf{\Lambda}^*(T(u,m),k)} \mid X_{\mathrm{p}^k(u)} = x)} \\ &= \int p_\theta(Y_{T(u,m)} \mid X_u = x_u)\, \mathbb{P}_\theta(X_u \in \mathrm{d}x_u \mid Y_{\mathbf{\Lambda}^*(T(u,m),k)}, X_{\mathrm{p}^{(k-1)(m+1)}(u)} = x), \end{aligned} \tag{100}$$

where:

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

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:

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

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:

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

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 $\mathrm{h}_{T(u,m),k,x}(\theta)$ we first prove the following lemma which states some uniform bound about the asymptotic behavior of those functions when $k \to \infty$.

Lemma C.1 (Uniform bounds for $\mathrm{h}_{T(u,m),k,x}(\theta)$)**.** *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:*

$$\sup_{\theta\in\Theta} \sup_{x,x'\in\mathcal{X}} |\mathrm{h}_{T(u,m),k,x}(\theta) - \mathrm{h}_{T(u,m),k',x'}(\theta)| \le \frac{\rho^{(k\wedge k')(m+1)-1}}{(1-\rho)^{|T_m|}}, \tag{102}$$

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

Proof. [The proof of the lemma is a straightforward adaptation of the proof of (Cappé, Moulines and Rydén, 2005, Lemma 12.3.2) using Lemma 3.2 for the coupling.] Let $k' \ge k \ge 1$, and write $v = \mathrm{p}^{k(m+1)}(u)$, $v' = \mathrm{p}^{k'(m+1)}(u)$. Then, write:

$$\mathrm{H}_{T(u,m),k,x}(\theta) = \int_{\mathcal{X}^2} \left[\int_{\mathcal{X}^{T(u,m)}} \prod_{w\in T(u,m)} g_\theta(x_w, Y_w) q_\theta(x_{\mathrm{p}(w)}, x_w) \lambda(\mathrm{d}x_w) \right] \tag{104}$$

$$\times \mathbb{P}_\theta(X_{\mathrm{p}(u)} \in \mathrm{d}x_{\mathrm{p}(u)} \,|\, Y_{\mathbf{\Delta}^*(T(u,m),k)}, X_v = x_v) \times \delta_x(\mathrm{d}x_v),$$

and using the Markov property at X_v, write:

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

Applying Lemma 3.2, we get (note that the integrands in (104) and (105) are non-negative):

$$\begin{aligned} &|\mathrm{H}_{T(u,m),k,x}(\theta) - \mathrm{H}_{T(u,m),k',x'}(\theta)| \\ &\quad \le \rho^{k(m+1)-1} \sup_{x_{\mathrm{p}(u)} \in \mathcal{X}} \int \prod_{w \in T(u,m)} g_\theta(x_w, Y_w) q_\theta(x_{\mathrm{p}(w)}, x_w) \lambda(\mathrm{d}x_w) \\ &\quad \le \rho^{k(m+1)-1} (\sigma^+)^{|T_m|} \prod_{w \in T(u,m)} \int g_\theta(x_w, Y_w) \lambda(\mathrm{d}x_w). \end{aligned} \tag{106}$$

The integral in (104) can be lower bounded giving us:

$$\mathrm{H}_{u,k,x}(\theta) \ge (\sigma^-)^{|T_m|} \prod_{w \in T(u,m)} \int g_\theta(x_w, Y_w) \lambda(\mathrm{d}x_w), \tag{107}$$

where the right hand side is positive by Assumption 3-(ii); and similarly for (105). Combining (106) with (107), and with the inequality $|\log x - \log y| \le |x - y|/(x \wedge y)$, we get the first assertion of the lemma:

$$|\mathrm{h}_{T(u,m),k,x}(\theta) - \mathrm{h}_{T(u,m),k',x'}(\theta)| \le \left(\frac{\sigma^+}{\sigma^-}\right)^{|T_m|} \rho^{k(m+1)-1} = \frac{\rho^{k(m+1)-1}}{(1-\rho)^{|T_m|}}.$$

Combining (100) and (107), we get:

$$\prod_{w \in T(u,m)} \sigma^- b^-(Y_w) \le \mathrm{H}_{T(u,m),k,x}(\theta) \le (b^+)^{|T_m|},$$

which yields the second assertion of the lemma (remind that $b^-(Y_w) > 0$ for all $w \in T^\infty$ by Assumption 3-(ii)). □

We are now ready to construct the limit of the functions $\mathrm{h}_{T(u,m),k,x}(\theta)$ 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^∞.

Proposition C.2 (Properties of the limit function $\mathrm{h}_{T(u,m),\infty}(\theta)$)**.** *Assume that Assumptions 1–4 hold. For every $u \in T$ and $\theta \in \Theta$, there exists $\mathrm{h}_{T(u,m),\infty}(\theta) \in L^1(\mathbb{P}_\mathcal{U} \otimes \mathbb{P}_{\theta^\star})$ such that for all $x \in \mathcal{X}$, the sequence $(\mathrm{h}_{T(u,m),k,x}(\theta))_{k \in \mathbb{N}}$ converges $\mathbb{P}_\mathcal{U} \otimes \mathbb{P}_{\theta^\star}$-a.s. and in $L^1(\mathbb{P}_\mathcal{U} \otimes \mathbb{P}_{\theta^\star})$ to $\mathrm{h}_{T(u,m),\infty}(\theta)$.*

Furthermore, this convergence is uniform over $\theta \in \Theta$ and $x \in \mathcal{X}$, that is, we have that $\lim_{k \to \infty} \sup_{\theta \in \Theta} \sup_{x \in \mathcal{X}} |\mathrm{h}_{T(u,m),k,x}(\theta) - \mathrm{h}_{T(u,m),\infty}(\theta)| = 0$ $\mathbb{P}_\mathcal{U} \otimes \mathbb{P}_{\theta^\star}$-a.s. and in $L^1(\mathbb{P}_\mathcal{U} \otimes \mathbb{P}_{\theta^\star})$.

Note that the limit function $\mathrm{h}_{T(u,m),\infty}(\theta)$ in the proposition can be interpreted as $\log p_\theta(Y_{T(u,m)} \,|\, Y_{\Delta^*(T(u,m),\infty)})$, where $\Delta^*(T(u,m),\infty) = \{v \in T^\infty : v <_{\mathcal{U}} T(u,m)\}$ is a random subset of vertices. Note that $\mathrm{h}_{T(u,m),\infty}(\theta)$ 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 $\mathcal{U}$ trough $\Delta(T(u,m),\infty)$.

Proof. Fix some $u \in T$. Note that (102) shows that $(\mathrm{h}_{T(u,m),k,x}(\theta))_{k\in\mathbb{N}}$ is Cauchy sequence uniformly in θ and x, and thus has $\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta\star}$-almost surely a limit when $k \to \infty$ which does not depend on x; we denote this limit by $\mathrm{h}_{T(u,m),\infty}(\theta)$. Furthermore, we get from (103) that $(\mathrm{h}_{T(u,m),k,x}(\theta))_{k\in\mathbb{N}}$ is uniformly bounded in $L^1(\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta\star})$, and thus $\mathrm{h}_{T(u,m),\infty}(\theta)$ is in $L^1(\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta\star})$ and the convergence also holds in $L^1(\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta\star})$. Finally, as the bound in (102) is uniform in θ and x, we get that the convergence holds uniform over θ and x both $\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta\star}$-almost surely and in $L^1(\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta\star})$. □

C.3. Properties of the contrast function

As the functions $\mathrm{h}_{T(u,m),\infty}(\theta)$ are in $L^1(\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta\star})$ 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:

$$\ell^{(m)}(\theta) = \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta\star}\left[\mathrm{h}_{T_m,\infty}(\theta)\right],$$

where remind $\mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta\star}$ is the expectation corresponding to $\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta\star}$. 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 $\mathbb{P}_{\theta\star}$-a.s. to the contrast function $\ell^{(m)}(\theta)$ as $n \to \infty$.*

$$\lim_{n\to\infty} \frac{|T_m|}{|T_{n(m+1)-1}|}\, \ell_{n(m+1)-1,x}(\theta) = \ell^{(m)}(\theta) \qquad \mathbb{P}_{\theta\star}\text{-a.s.} \tag{108}$$

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 $j \in \{k, \cdots, n-1\}$, we get:

$$\begin{aligned}\frac{|T_m|}{|T_{n(m+1)-1}|}\left|\ell_{n(m+1)-1,x}(\theta) - \sum_{j=k}^{n-1}\sum_{u\in G_{j(m+1)}} \mathrm{h}_{T(u,m),k,x}(\theta)\right| \\ \leq \frac{\rho^{k(m+1)-1}}{(1-\rho)^{|T_m|}} + \frac{|T_m|}{|T_{n(m+1)-1}|}\sum_{j=0}^{k-1}\sum_{u\in G_{j(m+1)}} |\mathrm{h}_{T(u,m),j,x}(\theta)|.\end{aligned} \tag{109}$$

Note that by (103), we have that $|\mathrm{h}_{T(u,m),j,x}(\theta)| < \infty$ $\mathbb{P}_{\theta\star}$-a.s. for all $j \in \mathbb{N}^*$ and $u \in G_{j(m+1)}$. For $u = \partial$, we have $\mathrm{h}_{T_m,0,x}(\theta) = \log p_\theta(Y_{T_m} \,|\, X_\partial = x)$ which is finite $\mathbb{P}_{\theta\star}$-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 3383) 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:

$$\begin{aligned}\mathrm{h}_{T(u,m),k,x}\big(\theta; Y_{\Delta(T(u,m),k)} = y_{\Delta(T(u,m),k)}\big) \\ = \mathrm{h}_{T(v_u,m),k,x}\big(\theta; Y_{\Delta(T(v_u,m),k)} = y_{\Delta(T(u,m),k))}\big).\end{aligned} \tag{110}$$

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 $\mathrm{h}_{T(u,m),k,x}(\theta; Y_{\Delta(T(u,m),k)})$ is in $L^2(\mathbb{P}_{\theta^\star})$. Hence, applying a straightforward modification of Lemma 2.11 for subtree blocks $T(u,m)$ to the collection of neighborhood-shape-dependent functions $(\mathrm{h}_{T(v,m),k,x}(\theta; Y_{\Delta(T(v,m))} = \cdot))_{v \in G_{k(m+1)}}$ (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:

$$\frac{|T_m|}{|T_{n(m+1)-1}|} \sum_{j=k}^{n-1} \sum_{u \in G_{j(m+1)}} \mathrm{h}_{T(u,m),k,x}(\theta) \underset{n\to\infty}{\longrightarrow} \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}\big[\mathrm{h}_{T_m,k,x}(\theta)\big] \qquad \mathbb{P}_{\theta^\star}\text{-a.s.} \tag{111}$$

Using (102) with Proposition C.2, we get:

$$\Big|\mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}\big[\mathrm{h}_{T_m,k,x}(\theta)\big] - \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}\big[\mathrm{h}_{T_m,\infty}(\theta)\big]\Big| \leq \frac{\rho^{k(m+1)-1}}{(1-\rho)^{|T_m|}}.$$

Thus, combining this bound with (109) and (111), we get $\mathbb{P}_{\theta^\star}$-a.s. that:

$$\limsup_{n\to\infty} \left| \frac{|T_m|}{|T_{n(m+1)-1}|} \ell_{n(m+1)-1,x}(\theta) - \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}\big[\mathrm{h}_{T_m,\infty}(\theta)\big] \right| \leq 2\, \frac{\rho^{k(m+1)-1}}{(1-\rho)^{|T_m|}}.$$

As the left hand side does not depend on k, letting $k \to \infty$, 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. □

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 $\ell^{(m)}(\theta) = \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}[h_{T(\partial,m),\infty}(\theta)]$, which is equal to $|T_m|\ell(\theta)$ by Proposition C.3. □

Appendix 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 $L^2(\mathbb{P}_{\theta^\star})$ norm uniformly in $x \in \mathcal{X}$ for the the average of the quantities $\Gamma_{u,h(u),x}(\theta^\star)$ 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:*

$$\mathbb{E}_{\theta^\star}\left[\sup_{x\in\mathcal{X}}\left|\frac{1}{|T_n|}\sum_{u\in T_n^*}\Gamma_{u,h(u),x}(\theta) - \mathbb{E}_{\mathcal{U}}\otimes\mathbb{E}_{\theta^\star}\left[\Gamma_{\partial,\infty}(\theta)\right]\right|^2\right]^{1/2} \leq C\alpha^n. \tag{112}$$

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:

$$\begin{aligned}
&\mathbb{E}_{\theta^\star}\left[\sup_{x\in\mathcal{X}}\left|\frac{1}{|T_n|}\sum_{u\in T_n^*}\Gamma_{u,h(u),x}(\theta) - \mathbb{E}_{\mathcal{U}}\otimes\mathbb{E}_{\theta^\star}\left[\Gamma_{\partial,\infty}(\theta)\right]\right|^2\right]^{1/2} \\
&\qquad\leq \mathbb{E}_{\theta^\star}\left[\sup_{x,x'\in\mathcal{X}}\left|\frac{1}{|T_n|}\sum_{u\in T_{k-1}^*}\Gamma_{u,h(u),x}(\theta)\right|^2\right]^{1/2} \\
&\qquad\quad+\mathbb{E}_{\theta^\star}\left[\sup_{x,x'\in\mathcal{X}}\left|\frac{1}{|T_n|}\sum_{u\in T_n\setminus T_{k-1}}\Gamma_{u,h(u),x}(\theta) - \Gamma_{u,k,x'}(\theta)\right|^2\right]^{1/2} \\
&\qquad\quad+\mathbb{E}_{\theta^\star}\left[\left|\frac{1}{|T_n|}\sum_{u\in T_n\setminus T_{k-1}}\Gamma_{u,k,x'}(\theta) - \mathbb{E}_{\mathcal{U}}\otimes\mathbb{E}_{\theta^\star}\left[\Gamma_{\partial,k,x'}(\theta)\right]\right|^2\right]^{1/2} \\
&\qquad\quad+\mathbb{E}_{\mathcal{U}}\otimes\mathbb{E}_{\theta^\star}\left[|\Gamma_{\partial,k,x'}(\theta) - \Gamma_{\partial,\infty}(\theta)|^2\right]^{1/2}.
\end{aligned} \tag{113}$$

Using Lemma 4.17 together with (49) on page 3403 (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 3383), then we have:

$$\Gamma_{u,k,x'}(\theta; Y_{\Delta(u,k)} = y_{\Delta(u,k)}) = \Gamma_{v_u,k,x'}(\theta; Y_{\Delta(v_u)} = y_{\Delta(u,k)}). \tag{114}$$

Moreover, using the definition of $\Gamma_{u,k,x}(\theta)$ in (63) together with the assumption on ϕ_θ in Proposition 4.5, we get that the random variable $\Gamma_{u,k,x'}(\theta; Y_{\Delta(u,k)} = y_{\Delta(u,k)})$ is in $L^2(\mathbb{P}_{\theta^\star})$ 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 $(\Gamma_{v_u,k,x'}(\theta; Y_{\Delta(v)} = \cdot))_{v\in G_k}$ (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. □

Lemma D.1 implies as a corollary the convergence $\mathbb{P}_{\theta\star}$-a.s. and in $L^2(\mathbb{P}_{\theta\star})$ uniformly in $x \in \mathcal{X}$ for the the sum of the quantities $\Gamma_{u,h(u),x}(\theta^\star)$ 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:*

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

Proof. The convergence in $L^2(\mathbb{P}_{\theta\star})$ follows immediately from Lemma D.1. Moreover, using again Lemma D.1, we have:

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

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

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 $\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta\star}$-a.s. continuous on Θ_0. Moreover, for all $\theta \in \Theta_0$, we have:*

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

Proof. We mimic the proof of (Douc, Moulines and Rydén, 2004, 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 $L^4(\mathbb{P}_{\theta\star})$. Thus, for all $v \in T^\infty$, the random variable $\|\phi^v\|_\infty$ is in $L^4(\mathbb{P}_{\theta\star})$. Remind from (12) on page 3386 that $\boldsymbol{\Lambda}(\partial, k)$ is a random subtree of the deterministic subtree $T^\infty(\mathrm{p}^k(u), k)$. Then, note that we have:

$$\sup_{\theta\in\Theta_0} |\Gamma_{\partial,k,x}(\theta)| \le 4 \left(\sum_{v\in T^\infty(\mathrm{p}^k(\partial),k)} \|\phi^v\|_\infty \right)^2,$$

where the upper bound is a random variable in $L^2(\mathbb{P}_{\theta\star})$ (and thus in $L^2(\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta\star})$) which depends on $Y_{T^\infty(\mathrm{p}^k(u),k)}$ but not on $\mathcal{U}$. Hence, to prove the lemma, it suffices to prove that for all $v_1, v_2 \in T^\infty(\mathrm{p}^k(u), k) \setminus \{\mathrm{p}^k(\partial)\}$ and $\epsilon \in \{0, 1\}$, we have $\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta\star}$-a.s. :

$$\lim_{\delta\to 0}\sup_{\theta'\in\Theta_0:\|\theta'-\theta\|\le\delta}\Big|\mathbb{E}_{\theta'}[\phi^{(2,\epsilon)}_{\theta',v_1,v_2}\,|\,Y_{\boldsymbol{\Delta}(\partial,k)},X_{\mathrm{p}^k(\partial)}=x] - \mathbb{E}_{\theta}[\phi^{(2,\epsilon)}_{\theta,v_1,v_2}\,|\,Y_{\boldsymbol{\Delta}(\partial,k)},X_{\mathrm{p}^k(\partial)}=x]\Big| = 0,$$

where:

$$\phi^{(2,\epsilon)}_{\theta',v_1,v_2} := \phi_{\theta'}(X_{\mathrm{p}(v_1)},X_{v_1},Y_{v_1})\phi_{\theta'}(X_{\mathrm{p}(v_2)},X_{v_2},Y_{v_2})^{\epsilon}.$$

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

$$\mathbb{E}_{\theta}[\phi^{(2,\epsilon)}_{\theta',v_1,v_2}\,|\,Y_{\boldsymbol{\Delta}(\partial,k)},X_{\mathrm{p}^k(\partial)}=x] = \frac{\int_{\mathcal{X}^{|\boldsymbol{\Delta}(\partial,k)|-1}}\phi^{(2,\epsilon)}_{\theta'}(x_{\mathrm{p}(v_1)},x_{v_1},Y_{v_1},x_{\mathrm{p}(v_2)},x_{v_2},Y_{v_2})\,\Psi(\mathrm{d}x_{\boldsymbol{\Delta}(\partial,k)\setminus\{\mathrm{p}^k(\partial)\}})}{\int_{\mathcal{X}^{\boldsymbol{\Delta}(\partial,k)\setminus\{\mathrm{p}^k(\partial)\}}}1\,\Psi(\mathrm{d}x_{\boldsymbol{\Delta}(\partial,k)\setminus\{\mathrm{p}^k(\partial)\}})}. \quad (115)$$

where:

$$\Psi(\mathrm{d}x_{\boldsymbol{\Delta}(\partial,k)\setminus\{\mathrm{p}^k(\partial)\}}) := \prod_{w\in\boldsymbol{\Delta}(\partial,k)\setminus\{\mathrm{p}^k(\partial)\}} q_\theta(x_{\mathrm{p}(w)},x_w)g_\theta(x_w,Y_w)\lambda(\mathrm{d}x_w).$$

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. θ and is upper bounded by the random variable $\|\phi^{v_1}\|_\infty(\|\phi^{v_2}\|_\infty)^{\epsilon}(\sigma^+b^+)^{|T^\infty(\mathrm{p}^k(u),k)|-1}$ (remind that $\sigma^+\ge 1$ and $b^+\ge 1$). And similarly, the denominator is continuous w.r.t. θ, and, using Assumption 3-(ii), is lower bounded by the random variable:

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

Hence, using dominated convergence, we conclude that $\mathbb{P}_{\mathcal{U}}\otimes\mathbb{P}_{\theta^\star}$-a.s. the left hand side of (115) is continuous w.r.t. θ. This concludes the proof. □

As a corollary of Lemma D.3, we get that the function $\theta\mapsto\Gamma_{\partial,\infty}(\theta)$ is continuous in $L^2(\mathbb{P}_{\theta^\star})$.

Corollary D.4. *Under the assumptions of Proposition 4.5, for all $\theta\in\Theta_0$, we have:*

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

In particular, the function $\theta\mapsto\mathbb{E}_{\mathcal{U}}\otimes\mathbb{E}_{\theta^\star}[\Gamma_{\partial,\infty}(\theta)]$ is continuous on Θ_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:

$$\mathbb{E}_{\mathcal{U}}\otimes\mathbb{E}_{\theta^\star}\left[\sup_{\theta'\in\Theta_0:\|\theta'-\theta\|\le\delta}|\Gamma_{\partial,\infty}(\theta')-\Gamma_{\partial,\infty}(\theta)|^2\right]^{1/2} \le 2C\beta^{k/2}+\mathbb{E}_{\mathcal{U}}\otimes\mathbb{E}_{\theta^\star}\left[\sup_{\theta'\in\Theta_0:\|\theta'-\theta\|\le\delta}|\Gamma_{\partial,k,x}(\theta')-\Gamma_{\partial,k,x}(\theta)|^2\right]^{1/2}. \quad (116)$$

Using Lemma D.3, we get:

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

and taking $k \to \infty$, the upper bound vanishes. This concludes the proof. □

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:*

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

Proof. First, write:

$$\begin{aligned}\sup_{\theta'\in\Theta_0:\|\theta'-\theta\|\le\delta} &\left|\frac{1}{|T_n|}\sum_{u\in T_n^*} \Gamma_{u,h(u),x}(\theta') - \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}[\Gamma_{\partial,\infty}(\theta)]\right| \\ &\le \frac{1}{|T_n|}\sum_{u\in T_n^*} \sup_{\theta'\in\Theta_0:\|\theta'-\theta\|\le\delta} \left|\Gamma_{u,h(u),x}(\theta') - \Gamma_{u,h(u),x}(\theta)\right| \\ &\quad + \left|\frac{1}{|T_n|}\sum_{u\in T_n^*} \Gamma_{u,h(u),x}(\theta) - \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}[\Gamma_{\partial,\infty}(\theta)]\right|. \end{aligned} \tag{117}$$

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:

$$\sup_{\theta'\in\Theta_0:\|\theta'-\theta\|\le\delta} \left|\Gamma_{u,h(u),x}(\theta') - \Gamma_{u,h(u),x}(\theta)\right|,$$

which are in $L^2(\mathbb{P}_{\theta^\star})$ using the assumptions of Proposition 4.5. This gives us that the first term in the upper bound of (117) converges $\mathbb{P}_{\theta^\star}$-a.s. as $n \to \infty$ to:

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

which, by Corollary D.4, vanishes when $\delta \to 0$. Corollary D.2 implies that the second term in the upper bound of (117) vanishes $\mathbb{P}_{\theta^\star}$-a.s. when $n \to \infty$. This concludes the proof. □

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. Applying Lemma 4.17, for all $u \in T$, we have that the sequence $(\Gamma_{u,k,x}(\theta))_{k\in\mathbb{N}^*}$ is a Cauchy sequence uniformly w.r.t. $\theta \in \Theta_0$ in $L^2(\mathbb{P}_{\mathcal{U}} \otimes \mathbb{P}_{\theta^\star})$ that converges to some limit $\Gamma_{u,\infty}(\theta)$ (that does not depend on x). By Corollary D.2, we have that $\mathbb{P}_{\theta^\star}$-a.s. the convergence for the the average of the quantities $\Gamma_{u,h(u),x}(\theta^\star)$ 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 $\theta \mapsto \mathbb{E}_{\mathcal{U}} \otimes \mathbb{E}_{\theta^\star}[\Gamma_{\partial,\infty}(\theta)]$ is continuous on Θ_0. Finally, the last part of the proposition is given by Lemma D.5. □

Acknowledgments

The author would like to thank an anonymous referee for their constructive comments that helped improve the quality of this paper.

References

ATHREYA, K. B. (1969). Limit Theorems for Multitype Continuous Time Markov Branching Processes: I. The Case of an Eigenvector Linear Functional. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **12** 320–332. https://doi.org/10.1007/BF00538753 MR0254927

BAUM, L. E. and PETRIE, T. (1966). Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *The Annals of Mathematical Statistics* **37** 1554–1563. https://doi.org/10.1214/aoms/1177699147 MR0202264

BAUM, L. E., PETRIE, T., SOULES, G. and WEISS, N. (1970). A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *The Annals of Mathematical Statistics* **41** 164–171. https://doi.org/10.1214/aoms/1177697196 MR0287613

BERCU, B., DE SAPORTA, B. and GÉGOUT-PETIT, A. (2009). Asymptotic Analysis for Bifurcating AutoRegressive Processes via a Martingale Approach. *Electronic Journal of Probability* **14**. https://doi.org/10.1214/EJP.v14-717 MR2563249

BERTSEKAS, D. P. and SHREVE, S. E. (1996). *Stochastic Optimal Control: The Discrete Time Case*. *Optimization and Neural Computation Series*. Athena Scientific, Belmont, Mass. MR4496006

BICKEL, P. J., RITOV, Y. and RYDÉN, T. (1998). Asymptotic Normality of the Maximum-Likelihood Estimator for General Hidden Markov Models. *The Annals of Statistics* **26**. https://doi.org/10.1214/aos/1024691255 MR1647705

BIESINGER, J., WANG, Y. and XIE, X. (2013). Discovering and Mapping Chromatin States Using a Tree Hidden Markov Model. *BMC Bioinformatics* **14** S4. https://doi.org/10.1186/1471-2105-14-S5-S4

BILLINGSLEY, P. (1999). *Convergence of Probability Measures*, 2nd ed ed. *Wiley Series in Probability and Statistics. Probability and Statistics Section*. Wiley, New York. MR1700749

BITSEKI PENDA, S. V. and DELMAS, J.-F. (2022a). Central Limit Theorem for Bifurcating Markov Chains under Pointwise Ergodic Conditions. *The Annals of Applied Probability* **32**. MR4497859

BITSEKI PENDA, S. V. and DELMAS, J.-F. (2022b). Central Limit Theorem for Kernel Estimator of Invariant Density in Bifurcating Markov Chains Models. *Journal of Theoretical Probability*. MR4621077

BITSEKI PENDA, S. V., DJELLOUT, H. and GUILLIN, A. (2014). Deviation Inequalities, Moderate Deviations and Some Limit Theorems for Bifurcating Markov Chains with Application. *The Annals of Applied Probability* **24**. https://doi.org/10.1214/13-AAP921 MR3161647

BOGACHEV, V. I. (2007). *Measure Theory*. **Vol. 2**. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-34514-5 MR2267655

BOUGUILA, N., FAN, W. and AMAYRI, M., eds. (2022). *Hidden Markov Models and Applications*. *Unsupervised and Semi-Supervised Learning*. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-030-99142-5 MR4442254

CAPPÉ, O., MOULINES, É. and RYDÉN, T. (2005). *Inference in Hidden Markov Models*. *Springer Series in Statistics*. Springer New York, New York, NY. https://doi.org/10.1007/0-387-28982-8 MR2159833

CHOI, H. and BARANIUK, R. G. (2001). Multiscale Image Segmentation Using Wavelet-Domain Hidden Markov Models. *IEEE Transactions on Image Processing* **10** 1309–1321. https://doi.org/10.1109/83.941855 MR1852250

CROUSE, M. S., NOWAK, R. D. and BARANIUK, R. G. (1998). Wavelet-Based Statistical Signal Processing Using Hidden Markov Models. *IEEE Transactions on Signal Processing* **46** 886–902. https://doi.org/10.1109/78.668544 MR1665651

DELMAS, J.-F. and MARSALLE, L. (2010). Detection of Cellular Aging in a Galton–Watson Process. *Stochastic Processes and their Applications* **120** 2495–2519. https://doi.org/10.1016/j.spa.2010.07.002 MR2728175

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum Likelihood from Incomplete Data Via the *EM* Algorithm. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **39** 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x MR0501537

DOUC, R. and MATIAS, C. (2001). Asymptotics of the Maximum Likelihood Estimator for General Hidden Markov Models. *Bernoulli* **7** 381. https://doi.org/10.2307/3318493 MR1836737

DOUC, R., MOULINES, É. and RYDÉN, T. (2004). Asymptotic Properties of the Maximum Likelihood Estimator in Autoregressive Models with Markov Regime. *The Annals of Statistics* **32**. https://doi.org/10.1214/009053604000000021 MR2102510

DOUC, R., ROUEFF, F. and SIM, T. (2016). The Maximizing Set of the Asymptotic Normalized Log-Likelihood for Partially Observed Markov Chains. *The Annals of Applied Probability* **26**. https://doi.org/10.1214/15-AAP1149 MR3543899

DOUC, R., MOULINES, E., OLSSON, J. and VAN HANDEL, R. (2011). Consistency of the Maximum Likelihood Estimator for General Hidden Markov Models. *The Annals of Statistics* **39**. https://doi.org/10.1214/10-AOS834 MR2797854

DOUC, R., MOULINES, É., PRIOURET, P. and SOULIER, P. (2018). *Markov Chains*. *Springer Series in Operations Research and Financial Engineering*. Springer International Publishing, Cham. MR3889011

DUARTE, M. F., WAKIN, M. B. and BARANIUK, R. G. (2008). Wavelet-Domain Compressive Signal Reconstruction Using a Hidden Markov Tree Model. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing* 5137–5140. IEEE, Las Vegas, NV, USA. https://doi.org/10.1109/ICASSP.2008.4518815

DUFLO, M. (2011). *Random Iterative Models*. Springer, Berlin. MR1485774

DURAND, J.-B., GONÇALVES, P. and GUÉDON, Y. (2004). Computational Methods for Hidden Markov Tree Models—An Application to Wavelet Trees. *IEEE Transactions on Signal Processing* **52** 2551–2560. https://doi.org/10.1109/TSP.2004.832006 MR2091804

DURAND, J.-B., GUÉDON, Y., CARAGLIO, Y. and COSTES, E. (2005). Analysis of the Plant Architecture via Tree-structured Statistical Models: The Hidden Markov Tree Models. *New*

Phytologist **166** 813–825. https://doi.org/10.1111/j.1469-8137.2005.01405.x

GENON-CATALOT, V. and LAREDO, C. (2006). Leroux's Method for General Hidden Markov Models. *Stochastic Processes and their Applications* **116** 222–243. https://doi.org/10.1016/j.spa.2005.10.005 MR2197975

GRAVE, É., OBOZINSKI, G. and BACH, F. (2013). Hidden Markov Tree Models for Semantic Class Induction. In *CoNLL* **Proceedings of the Seventeenth Conference on Computational Natural Language Learning** 94–103. Association for Computational Linguistics, Sofia, Bulgaria.

GUYON, J. (2007). Limit Theorems for Bifurcating Markov Chains. Application to the Detection of Cellular Aging. *The Annals of Applied Probability* **17** 1538–1569. https://doi.org/10.1214/105051607000000195 MR2358633

HANZOULI-BEN SALAH, H., LAPUYADE-LAHORGUE, J., BERT, J., BENOIT, D., LAMBIN, P., VAN BAARDWIJK, A., MONFRINI, E., PIECZYNSKI, W., VISVIKIS, D. and HATT, M. (2017). A Framework Based on Hidden Markov Trees for Multimodal Image Co-segmentation. *Medical Physics* **44** 5835–5848. https://doi.org/10.1002/mp.12531

HU, K., YANG, W. and GAO, X. (2017). 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. https://doi.org/10.1016/j.eswa.2017.05.062

JENSEN, J. L. and PETERSEN, N. V. (1999). Asymptotic Normality of the Maximum Likelihood Estimator in State Space Models. *The Annals of Statistics* **27**. https://doi.org/10.1214/aos/1018031205 MR1714719

KASAHARA, H. and SHIMOTSU, K. (2019). Asymptotic Properties of the Maximum Likelihood Estimator in Regime Switching Econometric Models. *Journal of Econometrics* **208** 442–467. https://doi.org/10.1016/j.jeconom.2018.09.019 MR3913246

KONDO, S., DUH, K. and MATSUMOTO, Y. (2013). Hidden Markov Tree Model for Word Alignment. In *WMT* **Proceedings of the Eighth Workshop on Statistical Machine Translation** 503–511. Association for Computational Linguistics, Sofia, Bulgaria.

KOSKI, T. (2001). *Hidden Markov Models for Bioinformatics*. *Computational Biology* **2**. Kluwer academic publ, Dordrecht [etc.]. MR1888250

LE GLAND, F. and MEVEL, L. (2000). Exponential Forgetting and Geometric Ergodicity in Hidden Markov Models. *Mathematics of Control, Signals, and Systems* **13** 63–93. https://doi.org/10.1007/PL00009861 MR1742140

LEROUX, B. G. (1992). Maximum-Likelihood Estimation for Hidden Markov Models. *Stochastic Processes and their Applications* **40** 127–143. https://doi.org/10.1016/0304-4149(92)90141-C MR1145463

MAKHIJANI, M. K., BALU, N., YAMADA, K., YUAN, C. and NAYAK, K. S. (2012). Accelerated 3D MERGE Carotid Imaging Using Compressed Sensing with a Hidden Markov Tree Model. *Journal of Magnetic Resonance Imaging* **36** 1194–1202. https://doi.org/10.1002/jmri.23755

MAMON, R. S. and ELLIOTT, R. J., eds. (2014). *Hidden Markov Models in Finance: Further Developments and Applications, Volume II. International Series in Operations Research &*

Management Science **209**. Springer US, Boston, MA. https://doi.org/10.1007/978-1-4899-7442-6 MR3013963

MANN, H. B. and WALD, A. (1943). On the Statistical Treatment of Linear Stochastic Difference Equations. *Econometrica* **11** 173. https://doi.org/10.2307/1905674 MR0009291

NAKASHIMA, S., SUGHIYAMA, Y. and KOBAYASHI, T. J. (2020). Lineage EM Algorithm for Inferring Latent States from Cellular Lineage Trees. *Bioinformatics* **36** 2829–2838. https://doi.org/10.1093/bioinformatics/btaa040

OLARIU, V., COCA, D., BILLINGS, S. A., TONGE, P., GOKHALE, P., ANDREWS, P. W. and KADIRKAMANATHAN, V. (2009). Modified Variational Bayes EM Estimation of Hidden Markov Tree Model of Cell Lineages. *Bioinformatics* **25** 2824–2830. https://doi.org/10.1093/bioinformatics/btp456

RABINER, L. R. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE* **77** 257–286. https://doi.org/10.1109/5.18626

RASCH, D. and SCHOTT, D. (2018). *Mathematical Statistics*. John Wiley & Sons, Hoboken. MR3618868

ROMBERG, J., CHOI, H., BARANIUK, R. G. and KINGSBURY, N. (2000). Multiscale Classification Using Complex Wavelets and Hidden Markov Tree Models. In *Proceedings 2000 International Conference on Image Processing (Cat. No.00CH37101)* 371–374 vol.2. IEEE, Vancouver, BC, Canada. https://doi.org/10.1109/ICIP.2000.899396

SHAHDOOSTI, H. R. and HAZAVEI, S. M. (2017). Image Denoising in Dual Contourlet Domain Using Hidden Markov Tree Models. *Digital Signal Processing* **67** 17–29. https://doi.org/10.1016/j.dsp.2017.04.011

VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* Cambridge University Press, Cambridge, UK; New York, NY, USA. https://doi.org/10.1017/CBO9780511802256 MR1652247

WALD, A. (1949). Note on the Consistency of the Maximum Likelihood Estimate. *The Annals of Mathematical Statistics* **20** 595–601. https://doi.org/10.1214/aoms/1177729952 MR0032169

WEIBEL, J. (2025). Ergodic Theorem for Branching Markov Chains Indexed by Trees with Arbitrary Shape. *Journal of Applied Probability*. https://doi.org/10.1017/jpr.2025.14

XIE, M., JIANG, Z. and SAINJU, A. M. (2018). Geographical Hidden Markov Tree for Flood Extent Mapping. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* 2545–2554. ACM, London United Kingdom. https://doi.org/10.1145/3219819.3220053

YU, D. and DENG, L. (2015). *Automatic Speech Recognition: A Deep Learning Approach*. *Signals and Communication Technology*. Springer London, London. https://doi.org/10.1007/978-1-4471-5779-3 MR3559045

ZUCCHINI, W. and MACDONALD, I. L. (2009). *Hidden Markov Models for Time Series: An Introduction Using R*, 0 ed. Chapman and Hall/CRC. https://doi.org/10.1201/9781420010893 MR2523850