Electronic Journal of Probability

Electron. J. Probab. **23** (2018), no. 47, 1–24.
ISSN: 1083-6489 https://doi.org/10.1214/18-EJP165

Necessary and sufficient conditions for consistent root reconstruction in Markov models on trees

Wai-Tong (Louis) Fan* Sebastien Roch†

Abstract

We establish necessary and sufficient conditions for consistent root reconstruction in continuous-time Markov models with countable state space on bounded-height trees. Here a root state estimator is said to be consistent if the probability that it returns to the true root state converges to 1 as the number of leaves tends to infinity. We also derive quantitative bounds on the error of reconstruction. Our results answer a question of Gascuel and Steel [GS10] and have implications for ancestral sequence reconstruction in a classical evolutionary model of nucleotide insertion and deletion [TKF91].

Keywords: Markov models on trees; reconstruction problem; concentration inequalities; consistent estimation; information-theoretic bounds; applications to phylogenetics.
AMS MSC 2010: 60J25; 60J80; 62B10; 62M05.
Submitted to EJP on July 16, 2017, final version accepted on April 4, 2018.

1 Introduction

Background In biology, the inferred evolutionary history of organisms and their relationships is depicted diagrammatically as a phylogenetic tree, that is, a rooted tree whose leaves represent living species and branchings indicate past speciation events [Fel04]. The evolution of species features, such as protein sequences, linear arrangements of genes on a chromosome or the number of horns of a lizard, is commonly assumed to follow Markovian dynamics along this tree [Ste16]. That is, on each edge of the tree, the state of the feature changes according to a continuous-time Markov process; at bifurcations, two independent copies of the feature evolve along the outgoing edges starting from the state at the branching point. The length of an edge is a measure of the expected amount of change along it. See Section 1.1 for a formal definition.

In this paper, we are concerned with the problem of inferring an ancestral state from observations at the leaves of a given tree under known Markovian dynamics. We

*Department of Mathematics, UW–Madison. Work supported by NSF grant DMS–1149312 (CAREER) to SR.
†Departments of Mathematics, UW–Madison. E-mail: roch@math.wisc.edu. Work supported by NSF grants DMS-1149312 (CAREER) and DMS-1614242.

refer to this problem, which has important applications in biology [Tho04, Lib07], as the **root reconstruction problem**. Many rigorous results have been obtained in the finite state space case, although much remains to be understood; see, e.g., [KS66, BRZ95, Iof96, EKPS00, Mos01, MP03, BCMR06, Sly09, BST10, BVVW11, Sly11] for a partial list. Typically, one seeks an estimator of the root state which is strictly superior to random guessing—uniformly in the depth of the tree—under a uniform prior on the root [Mos01]. Whether such an estimator exists has been shown to hinge on a trade-off between the mixing rate of the Markov process (i.e., the speed at which information is lost) and the growth rate of the sequence of trees considered (i.e., the speed at which information is duplicated). In some cases, for instance two-state symmetric Markov chains on d-ary trees [KS66, Iof96], sharp thresholds have been established.

Main results Here, we study the root reconstruction problem in an alternative setting where estimators with stronger properties can be derived. We consider sequences of nested trees with *uniformly bounded depths*. This is motivated by contemporary applications in evolutionary biology where the rapidly increasing availability of data from ever-growing numbers of organisms, particularly genome sequencing data, has allowed dense sampling of species within the same family or genus. This is sometimes referred to as the **taxon-rich setting** and has been considered in a number of recent theoretical studies [GS10, HA13, FR]. As shown in [GS10], a key difference with the traditional setting described above is that, in the taxon-rich setting, **consistent root state estimation** is possible. In this context, a consistent estimator is one whose probability of success tends to 1 as the number of leaves goes to infinity. See Section 1.1 for a formal definition. In particular, for general finite-state-space Markov processes on ultrametric trees, i.e., trees whose leaves are equidistant from the root, Gascuel and Steel [GS10] give sufficient conditions for the existence of consistent root state estimators by introducing a notion of "well-spread trees."

Building on this work, we give both necessary and sufficient conditions for consistent root reconstruction for general trees and general Markov processes on countable state spaces, a question left open in [GS10]. On an intuitive level, the greater the number of leaves, the more information we have about the root state. However, the leaves do not provide *independent* information due to the correlation arising from the partial overlap of the paths from the root to the leaves. In particular we cannot appeal, for instance, to the consistency of maximum likelihood estimation for independent samples [LR05]. We show however that, under a certain "root density" assumption we refer to as the big bang condition, one can identify a subset of leaves that are "sufficiently independent." We also derive quantitative bounds on the error of reconstruction in terms of natural properties of the tree sequence and Markov process.

One applied motivation for our results, especially our consideration of countable state spaces, is ancestral sequence reconstruction in DNA evolution models accounting for nucleotide insertion and deletion. Our main theorem immediately gives necessary and sufficient conditions for the existence of consistent root estimators for a classical such model known as the TKF91 process [TKF91]. This is detailed in Section B. In this context, our work is also related to trace reconstruction, which corresponds roughly to the star tree case under simplified analogues to the TKF91 process. See, e.g., [Mit09] for a survey. See also [ADHR12] for related work in the phylogenetic setting.

Organization Definitions and main results are stated in Sections 1.1 and 1.2. The connection between our big bang condition and the well-spread trees of [GS10] is established in Section 2. Our impossibility result is proved in Section 3, while our consistency result and error bound are detailed respectively in Sections 4 and 5.

1.1 Basic definitions

Markov chains on trees We consider the following class of latent tree models arising in phylogenetics. The model has two main components:

- The first component is a tree. More precisely, throughout, by a tree we mean a finite, edge-weighted, rooted tree $T = (V, E, \rho, \ell)$, where V is the set of vertices, E is the set of edges oriented away from the root ρ, and $\ell : E \to (0, +\infty)$ is a positive edge-weighting function. We denote by ∂T the leaf set of T. No assumption is made on the degree of the vertices. We think of T as a continuous object, where each edge e is a line segment of length ℓ_e and whose elements we refer to as points. We let Γ_T be the set of points of T.
- The second component is a time-homogeneous, continuous-time Markov process taking values in a countable state space $\mathcal{S}$. Without loss of generality, we let $\mathcal{S} = \{1, \ldots, |\mathcal{S}|\}$ in the finite case and $\mathcal{S} = \{1, 2, \ldots\}$ in the infinite case. We denote by $\mathbf{P}_t = (p_{ij}(t) : i, j \in \mathcal{S})$ the transition matrix at time $t \in [0, \infty)$, that is, $p_{ij}(t)$ is the probability that the state at time t is j given that it was i at time 0. We also let

$$\mathbf{p}^i(t) = (p_{i1}(t), p_{i2}(t), \ldots), \tag{1.1}$$

be the i-th row in the transition matrix. We assume that $(\mathbf{P}_t)_t$ admits a Q-matrix $Q = (q_{ij} : i, j \in \mathcal{S})$ which is stable and conservative, that is,

$$q_{ij} := \left. \frac{d}{dt} p_{ij}(t) \right|_{t=0} \in [0, \infty) \qquad \forall i \neq j,$$

and

$$q_i := -q_{ii} = \sum_{j \neq i} q_{ij} \in [0, \infty) \qquad \forall i. \tag{1.2}$$

See, e.g., [Lig10, Chapter 2] or [And91] for more background on continuous-time Markov chains.

We consider the following stochastic process indexed by the points of T. The root is assigned a state $X_\rho \in \mathcal{S}$, which is drawn from a probability distribution on $\mathcal{S}$. This state is then propagated down the tree according to the following recursive process. Moving away from the root, along each edge $e = (u, v) \in E$, conditionally on the state X_u, we run the Markov process $\mathbf{P}_t$ started at X_u for an amount of time $\ell_{(u,v)}$. We denote by X_γ the resulting state at $\gamma \in e$. We call the process $\mathcal{X} = (X_\gamma)_{\gamma \in \Gamma_T}$ a $\mathbf{P}_t$**-chain on** T. For $i \in \mathcal{S}$, we let $\mathbb{P}^i$ be the probability law when the root state X_ρ is i. If X_ρ is chosen according to a distribution π, then we denote the probability law by $\mathbb{P}^\pi$. Note that the leaf distribution conditioned on the root state is given by

$$\mathcal{L}_T^i\left((x_u)_{u \in \partial T}\right) := \mathbb{P}^i\left[(X_u)_{u \in \partial T} = (x_u)_{u \in \partial T}\right] = \sum_{\substack{(x'_u)_{u \in V} : \\ (x'_u)_{u \in \partial T} = (x_u)_{u \in \partial T}, \\ x'_\rho = i}} \prod_{e=(u,v) \in E} p_{x'_u, x'_v}(\ell_e), \tag{1.3}$$

for all $(x_u)_{u \in \partial T} \in \mathcal{S}^{\partial T}$.

Root reconstruction In the **root reconstruction problem**, we seek a good estimator of the root state X_ρ based on the leaf states $X_{\partial T}$. More formally, let $\{T^k = (V^k, E^k, \rho^k, \ell^k)\}_{k \geq 1}$ be a sequence of trees with $|\partial T^k| \to +\infty$ and let $\mathcal{X}^k = (X_\gamma^k)_{\gamma \in \Gamma_{T^k}}$ be a $\mathbf{P}_t$-chain on T^k with root state distribution π.

Definition 1.1 (Consistent root reconstruction). *A sequence of root estimators*

$$F_k : \mathcal{S}^{\partial T^k} \to \mathcal{S},$$

is said to be consistent *for* $\{T^k\}_k$*,* $(\mathbf{P}_t)_t$ *and* π *if*

$$\liminf_{k\to+\infty} \mathbb{P}^\pi \left[F_k \left(X^k_{\partial T^k} \right) = X^k_{\rho^k} \right] = 1.$$

The basic question we address is the following.

Question 1.2. Under what conditions on $\{T^k\}_k$, $(\mathbf{P}_t)_t$, and π does there exist a sequence of consistent root estimators?

Before stating our main theorems, we make some assumptions and introduce further notation.

Basic setup For concreteness, we let $\{T^k\}_k$ be a nested sequence of trees with common root ρ. That is, for all $k > 1$, T^{k-1} is a restriction of T^k, as defined next.

Definition 1.3 (Restriction). *Let* $T = (V, E, \rho, \ell)$ *be a tree. For a subset of leaves* $L \subset \partial T$*, the* restriction of T to L *is the tree obtained from* T *by keeping only those points on a path between the root* ρ *and a leaf* $u \in L$*.*

Observe that a restriction of T is always rooted at ρ. Without loss of generality, we assume that $|\partial T^k| = k$, so that T^k is obtained by adding a leaf edge to T^{k-1}. (More general sequences can be obtained as subsequences.) In a slight abuse of notation, we denote by ℓ the edge-weight function for all k. For $\gamma \in \Gamma_T$, we denote by ℓ_γ the length of the unique path from the root ρ to γ. We refer to ℓ_γ as the distance from γ to the root. Our standing assumptions throughout this paper are as follows.

(i) *(Uniformly bounded height)* The sequence of trees $\{T^k\}_k$ has uniformly bounded height. Denote by $h^k := \max\{\ell_x : x \in \partial T^k\}$ the height of T^k. Then the bounded height assumption says that

$$h^* := \sup_k h^k < +\infty.$$

(ii) *(Initial-state identifiability)* The Markov process $(\mathbf{P}_t)_t$ is initial-state identifiable, that is, all rows of the transition matrix $\mathbf{P}_t$ are distinct for all $t \in [0, \infty)$. In other words, given the distribution at time t, the initial state of the chain is uniquely determined.

Whether the last assumption holds in general for countable-space, continuous-time Markov processes (that are stable and conservative) seems to be open. We show in the appendix that it holds for two broad classes of chains: reversible chains and uniform chains, including finite state spaces. (Observe, on the other hand, that in the discrete-time case it is easy to construct a transition matrix which does not satisfy initial-state identifiability.) We use the notation $a \wedge b := \min\{a, b\}$ and $a \vee b := \max\{a, b\}$. For two probability measures μ_1, μ_2 on $\mathcal{S}$, let

$$\|\mu_1 - \mu_2\|_{\mathrm{TV}} = \frac{1}{2} \sum_{\sigma\in\mathcal{S}} |\mu_1(\sigma) - \mu_2(\sigma)| = \sup_{\mathcal{A}\subseteq\mathcal{S}} |\mu_1(\mathcal{A}) - \mu_2(\mathcal{A})| = 1 - \sum_{\sigma\in\mathcal{S}} \mu_1(\sigma) \wedge \mu_2(\sigma), \quad (1.4)$$

be the total variation distance between μ_1 and μ_2. (The last equality follows from noticing that $\|\mu_1 - \mu_2\|_{\mathrm{TV}} = \frac{1}{2}\sum_{\sigma\in\mathcal{S}}[\mu_1(\sigma) \vee \mu_2(\sigma) - \mu_1(\sigma) \wedge \mu_2(\sigma)]$ and $1 = \frac{1}{2}\sum_{\sigma\in\mathcal{S}}[\mu_1(\sigma) \vee \mu_2(\sigma) + \mu_1(\sigma) \wedge \mu_2(\sigma)]$.) Then initial-state identifiability is equivalent to

$$\|\mathbf{p}^i(t) - \mathbf{p}^j(t)\|_{\mathrm{TV}} > 0, \qquad \forall i \neq j \in \mathcal{S}, t \in (0, \infty), \quad (1.5)$$

where recall that $\mathbf{p}^i(t)$ was defined in (1.1).

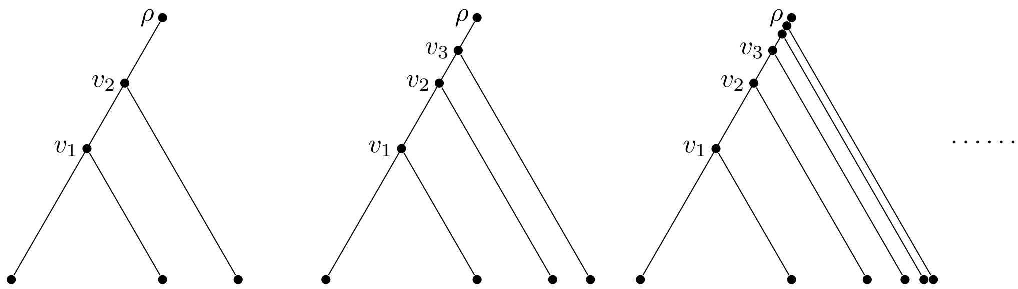

Figure 1: A sequence of trees $\{T^k\}_k$ (from left to right) satisfying the big bang condition. The distance from v_k to the root is 2^{-k}.

Big bang condition Our combinatorial condition for consistency says roughly that the T^ks are arbitrarily dense around the root.

Definition 1.4 (Truncation)**.** *For a tree $T = (V, E, \rho, \ell)$, let*

$$T(s) = \{\gamma \in \Gamma_T : \ \ell_\gamma \leq s\},$$

denote the tree obtained by truncating T at distance s from the root. We refer to $T(s)$ as a truncation *of T.*

See the left-hand side of Figure 3 for an illustration. Note that, if s is greater than the height of T, then $T(s) = T$.

Definition 1.5 (Big bang condition)**.** *We say that a sequence of trees $\{T^k\}_k$ satisfies the* big bang condition *if: for all $s \in (0, +\infty)$, we have $|\partial T^k(s)| \to +\infty$ as $k \to +\infty$.*

See Figure 1 for an illustration. For $i \in \mathcal{S}$, let $\mathcal{D}_i$ be the set of states reachable from i, that is, the states j for which $p_{ij}(t) > 0$ for some $t > 0$ (and, therefore, for all $t > 0$; see e.g. [Lig10, Chapter 2]).

1.2 Statements of main results

Our main result is the following.

Theorem 1.6 (Consistent root reconstruction: necessary and sufficient conditions)**.** *Let $\{T^k\}_k$ and $(\mathbf{P}_t)_t$ satisfy our standing assumptions (i) and (ii), and let π be a probability distribution on $\mathcal{S}$. Then there exists a sequence of root estimators that is consistent for $\{T^k\}_k$, $(\mathbf{P}_t)_t$ and π if and only if at least one of the following conditions hold:*

(a) (Downstream disjointness) *For all $i \neq j$ such that $\pi(i) \wedge \pi(j) > 0$, the reachable sets $\mathcal{D}_i$ and $\mathcal{D}_j$ are disjoint.*

(b) (Big bang) *The sequence of trees $\{T^k\}_k$ satisfies the big bang condition.*

An application to DNA evolution by nucleotide insertion and deletion is detailed in Section B. We also derive error bounds under the big bang condition. For $\epsilon > 0$, let $n_\epsilon < \infty$ be the smallest integer such that $\sum_{i > n_\epsilon} \pi(i) < \epsilon$ and $\Lambda_\epsilon = \{i \in \mathcal{S} \ : \ i \leq n_\epsilon\}$. Define also

$$q_\epsilon^* = \max_{i \in \Lambda_\epsilon} (q_i \vee 1),$$

and

$$\Delta_\epsilon = \min_{i_1 \neq i_2 \in \Lambda_\epsilon} \|\mathbf{p}^{i_1}(h^*) - \mathbf{p}^{i_2}(h^*)\|_{\mathrm{TV}},$$

which is positive under initial-state identifiability.

Theorem 1.7 (Root reconstruction: error bounds)**.** *Let $\{T^k\}_k$ and $(\mathbf{P}_t)_t$ satisfy our standing assumptions (i) and (ii) as well as the big bang condition, and let π be a probability distribution on $\mathcal{S}$. Fix $\epsilon > 0$ and $k \geq 1$. Then there exist universal constants $C_0, C_1 > 0$ and an estimator F_k such that for all $s > 0$,*

$$\mathbb{P}^\pi\left[F_k(X^k_{\partial T^k}) \neq X^k_\rho\right] < \epsilon + C_0\,\Delta_\epsilon^{-2}\,q^*_\epsilon\,s + n_\epsilon \exp\left(-C_1\,\Delta_\epsilon^2\,|\partial T^k(s)|\right). \tag{1.6}$$

Further, if the chain is uniform, that is, if $q^* = \sup_{i\in\mathcal{S}}(q_i \vee 1) < +\infty$, then there exist universal constants $C_0^U, C_1^U, C_2^U > 0$ and an estimator F_k^U such that for all $s > 0$ and all i

$$\mathbb{P}^i\left[F_k^U(X^k_{\partial T^k}) \neq X^k_\rho\right] < C_0^U\,f_*^{-4}\,q^*\,s + C_2^U f_*^{-1}\exp\left(-C_1^U\,f_*^4\,|\partial T^k(s)|\right), \tag{1.7}$$

where $f_* = e^{-q^* h^*}$.

The following example gives some intuition for the terms in (1.6) and (1.7).

Example 1.8 (Two-state chain on a pinched star)**.** Consider the following tree T. The root ρ is adjacent to a single vertex $\widetilde{\rho}$ through an edge of length $s > 0$. The vertex $\widetilde{\rho}$ is also adjacent to m vertices $x_1, \ldots, x_m$ through edges of length $h - s > 0$, where m is an odd integer. Consider the $(\mathbf{P}_t)_t$-chain on T with state space $\mathcal{S} = \{1, 2\}$, Q-matrix

$$Q = \begin{pmatrix} -q & q \\ q & -q \end{pmatrix},$$

and uniform root distribution π. It can be shown (see e.g. [SS03]) that under this chain

$$p_{11}(t) = \frac{1 + e^{-2qt}}{2} \quad \text{and} \quad p_{12}(t) = \frac{1 - e^{-2qt}}{2}. \tag{1.8}$$

Let N_1 be the number of leaves in state 1, let $\alpha = p_{11}(s) \in (1/2, 1)$ and let $\beta = p_{12}(h-s) \in (0, 1/2)$. The estimator that maximizes the probability of correct reconstruction is the maximum a posteriori estimate (see Lemma 3.2), which in this case boils down to setting $F(N_1) = 1$ if

$$\begin{aligned}
&\frac{1}{2}\alpha\binom{m}{N_1}(1-\beta)^{N_1}\beta^{m-N_1} + \frac{1}{2}(1-\alpha)\binom{m}{N_1}\beta^{N_1}(1-\beta)^{m-N_1} \\
&\qquad > \frac{1}{2}\alpha\binom{m}{N_1}\beta^{N_1}(1-\beta)^{m-N_1} + \frac{1}{2}(1-\alpha)\binom{m}{N_1}(1-\beta)^{N_1}\beta^{m-N_1},
\end{aligned}$$

and $F(N_1) = 2$ otherwise. Observing that

$$\alpha x + (1-\alpha)y > \alpha y + (1-\alpha)x \iff (2\alpha - 1)(x - y) > 0 \iff x > y,$$

where we used that $\alpha > 1/2$, we get that $F(N_1) = 1$ if and only if $N_1 > m/2$. Hence by symmetry, for $i = 1, 2$,

$$\begin{aligned}
\mathbb{P}^\pi[F(N_1) \neq X_\rho] &= \mathbb{P}^i[F(N_1) \neq i] \\
&= \sum_{n < m/2}\left\{\alpha\binom{m}{n}(1-\beta)^n\beta^{m-n} + (1-\alpha)\binom{m}{n}\beta^n(1-\beta)^{m-n}\right\} \\
&\leq (1-\alpha) + \alpha\,\mathbb{P}[N_1 < m/2 \,|\, X_{\widetilde{\rho}} = 1] \\
&\leq (1-\alpha) + \alpha\,\exp\left(-2m\left\{\frac{1}{2} - \beta\right\}^2\right),
\end{aligned}$$

by Hoeffding's inequality [Hoe63]. By (1.8), as $s \to 0$,

$$\mathbb{P}^i[F(N_1) \neq i] \leq 2qs\,(1 + o(1)) + \exp\left(-\frac{1}{2}mf^4\,(1 + o(1))\right),$$

where $f = e^{-qh}$.

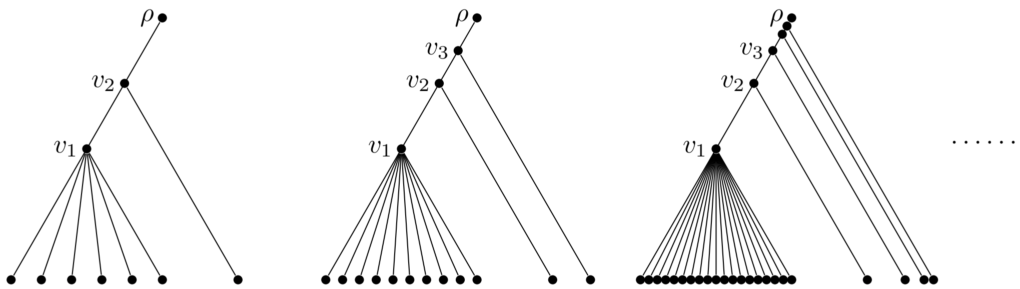

Figure 2: A (sub-)sequence of trees $\{T^k\}_k$ (from left to right) satisfying the big bang condition, but such that $\mathrm{Spr}(T^k)$ does not tend to 0.

2 Spread

We begin the proof by relating the big bang condition to a notion of spread introduced in [GS10]. This connection captures the basic combinatorial insights behind the proof of Theorem 1.6.

Let $T = (V, E, \rho, \ell)$ be a tree. We let ℓ_{xy} be the length of the shared path from the root ρ to the leaves x and y. That is, if $\mathrm{P}(u, v)$ denotes the set of edges on the unique path between vertices u and v, then we have

$$\ell_{xy} := \sum_{e \in \mathrm{P}(\rho,x) \cap \mathrm{P}(\rho,y)} \ell_e.$$

Roughly speaking, a tree is "well-spread" if the average value of ℓ_{xy} over all pairs (x, y) is small. The formal definition is as follows.

Definition 2.1 (Spread)**.** *The* spread *of a tree T is defined as*

$$\mathrm{Spr}(T) := \frac{\sum_{x,y}(\ell_{xy} \wedge 1)}{|\partial T|(|\partial T| - 1)},$$

where the summation is over all ordered pairs of distinct leaves $x \neq y$. For $\beta \in (0, \infty)$, we say that T is $(1-\beta)$-spread if $\mathrm{Spr}(T) \leq \beta$. For a sequence of trees $\{T^k\}_k$, we say that $\{T^k\}_k$ has vanishing spread *if*

$$\limsup_{k\to\infty} \mathrm{Spr}(T^k) = 0.$$

We show below that, if $\{T^k\}_k$ has vanishing spread, then the big bang condition holds. The converse is false as illustrated in Figure 2, where the root is arbitrarily dense but the spread is dominated by a subtree away from the root. We show however that, if the big bang condition holds, then one can find a sequence of arbitrarily large restrictions with vanishing spread. (Restrictions were introduced in Definition 1.3.) Our main result of this section is the following lemma.

Lemma 2.2 (Big bang and spread)**.** *Let $\{T^k\}_k$ be a sequence of trees satisfying our standing assumptions (i) and (ii). The big bang condition holds if and only if there exists a nested sequence of restrictions $\widetilde{T}^k$ of T^k such that $|\partial \widetilde{T}^k| \to \infty$ and $\{\widetilde{T}^k\}_k$ has vanishing spread.*

Proof. For the *if* part, we argue by contradiction. Assume the big bang condition fails and let $\{\widetilde{T}^k\}_k$ be a nested sequence of restrictions of $\{T^k\}_k$ with vanishing spread such that $|\partial \widetilde{T}^k| \to \infty$. Then there exist $s_0 \in (0, 1)$, $m_0 \geq 1$ and $k_0 \geq 1$ such that

$$|\partial T^k(s_0)| = m_0, \qquad \forall k \geq k_0.$$

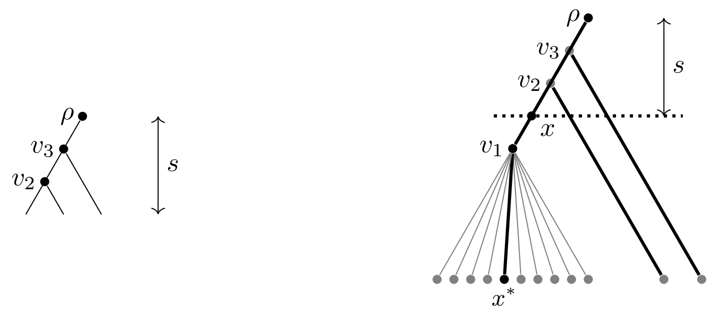

Figure 3: Consider again the second tree in Figure 2. On the left side, $T^k(s)$ is shown where $k = 3$. On the right side, the subtree $\widetilde{T}^{k,s}$ is highlighted.

Also, by the nested property, the truncation $T^k(s_0)$ remains the same for all $k \geq k_0$. We show that at least one of the subtrees of $\widetilde{T}^k$ rooted at a point in $\partial T^k(s_0)$ makes a large contribution to the spread. For $k \geq k_0$ and $z \in \partial T^k(s_0)$, let $\partial \widetilde{T}^k_{[z]}$ be the leaves of $\widetilde{T}^k$ below z. Then, since

$$\sum_{z \in \partial T^k(s_0)} \left| \partial \widetilde{T}^k_{[z]} \right| = |\partial \widetilde{T}^k|,$$

there is a $z_k \in \partial T^k(s_0)$ such that

$$\left| \partial \widetilde{T}^k_{[z_k]} \right| \geq \left\lceil \frac{|\partial \widetilde{T}^k|}{m_0} \right\rceil. \tag{2.1}$$

Observe that, for all distinct x, y in $\partial \widetilde{T}^k_{[z_k]}$, it holds that $\ell_{xy} \geq s_0$ because the paths to x and y share at least the path to z_k. Then, counting only the contribution from $\partial \widetilde{T}^k_{[z_k]}$, we get the following bound on the spread of $\widetilde{T}^k$

$$\mathrm{Spr}(\widetilde{T}^k) \geq \frac{\left| \partial \widetilde{T}^k_{[z_k]} \right| \left(\left| \partial \widetilde{T}^k_{[z_k]} \right| - 1 \right) s_0}{|\partial \widetilde{T}^k| (|\partial \widetilde{T}^k| - 1)}.$$

By (2.1),

$$\liminf_{k \to \infty} \mathrm{Spr}(\widetilde{T}^k) \geq \frac{s_0}{m_0^2} > 0.$$

As a result, $\{\widetilde{T}^k\}_k$ does not have vanishing spread.

For the *only if* part, assume the big bang condition holds. For every $k \geq 1$ and $s \in (0, 1)$, we extract a $(1 - s)$-spread restriction $\widetilde{T}^{k,s}$ of T^k as follows. See Figure 3 for an illustration. Let $\partial T^k(s) = \{z_1, \ldots, z_m\}$. For each z_i, $i = 1, \ldots, m$, pick an arbitrary leaf $x_i \in \partial T^k_{[z_i]}$ in the subtree $T^k_{[z_i]}$ of T^k rooted at z_i. We let $\widetilde{T}^{k,s}$ be the restriction of T^k to $\{x_1, \ldots, x_m\}$. Observe that $\widetilde{T}^{k,s}$ is $(1 - s)$-spread because the paths to each pair of leaves in $\partial \widetilde{T}^{k,s}$ diverge within $T^k(s)$. To construct a sequence of restrictions with vanishing spread, we take a sequence of positive reals $(s_i)_{i \geq 1}$ with $s_i \downarrow 0$ and proceed as follows:

- Let $k_1 \geq 1$ be such that $|\partial T^k(s_2)| \geq 2$ for all $k > k_1$. The value k_1 exists under the big bang condition. For all $k \leq k_1$, let $\widetilde{T}^k = \widetilde{T}^{k,s_1}$.
- Let $k_2 > k_1$ be such that $|\partial T^k(s_3)| \geq 3$ for all $k > k_2$. The value k_2 exists under the big bang condition. For all $k_1 < k \leq k_2$, let $\widetilde{T}^k = \widetilde{T}^{k,s_2}$.
- And so forth.

By construction, for all $k_{j-1} < k \le k_j$, it holds that $\mathrm{Spr}(\widetilde{T}^k) \le s_j$ and $|\partial\widetilde{T}^k| = |\partial T^k(s_j)| \ge j$. Thus,

$$\limsup_{k\to\infty} \mathrm{Spr}(\widetilde{T}^k) = 0$$

and

$$\lim_k |\partial\widetilde{T}^k| = +\infty$$

as required. □

3 Impossibility of reconstruction

The goal of this section is to show that, in the absence of downstream disjointness, the big bang condition is necessary for consistent root reconstruction. The following proposition implies the *only if* part of Theorem 1.6.

Proposition 3.1 (Impossibility of reconstruction without the big bang condition)**.** *Let $\{T^k\}_k$ and $(\mathbf{P}_t)_t$ satisfy our standing assumptions (i) and (ii), and let π be a probability distribution on $\mathcal{S}$. Assume that neither downstream disjointness nor the big bang condition hold. Then consistent reconstruction of the root state is impossible, in the sense that there exists an $\epsilon > 0$ such that for all $k \ge 1$*

$$\sup_{F_k} \mathbb{P}^\pi \left[F_k(X^k_{\partial T^k}) = X^k_\rho\right] \le 1 - \epsilon, \tag{3.1}$$

where the supremum is over all root estimators $F_k : \mathcal{S}^{\partial T^k} \to \mathcal{S}$.

3.1 Information-theoretic bounds

To prove Proposition 3.1, we need some information-theoretic bounds that relate the best achievable reconstruction probability to the total variation distance between the conditional distributions of pairs of initial states. Our first bound says roughly that the reconstruction probability is only as good as the worst total variation distance. Our second bound shows that a good reconstruction probability can be obtained from selecting a subset of initial states with high prior probability whose corresponding conditional distributions have “little overlap.” See e.g. [CT06, Chapter 2] and [SS99, SS02] for some related results.

Lemma 3.2 (Information-theoretic bounds)**.** *Let Y_0 and Y_1 be random variables taking values in the countable spaces $\mathcal{Y}_0$ and $\mathcal{Y}_1$ respectively. Let μ_0 denote the distribution of Y_0 and let μ_1^i denote the distribution of Y_1 conditioned on $\{Y_0 = i\}$.*

1. (Reconstruction upper bound) *It holds that*

$$\sup_{f:\mathcal{Y}_1\to\mathcal{Y}_0} \mathbb{P}[f(Y_1) = Y_0] \le 1 - \sup_{i_1\neq i_2\in\mathcal{Y}_0} \left\{\mu_0(i_1) \wedge \mu_0(i_2)\left[1 - \|\mu_1^{i_1} - \mu_1^{i_2}\|_{\mathrm{TV}}\right]\right\}. \tag{3.2}$$

2. (Reconstruction lower bound) *For any $\Lambda \subseteq \mathcal{Y}_0$, it holds that*

$$\sup_{f:\mathcal{Y}_1\to\mathcal{Y}_0} \mathbb{P}[f(Y_1) = Y_0] \ge \sum_{i\in\Lambda} \mu_0(i) - \sum_{i_1\neq i_2\in\Lambda} \left\{\mu_0(i_1) \vee \mu_0(i_2)\left[1 - \|\mu_1^{i_1} - \mu_1^{i_2}\|_{\mathrm{TV}}\right]\right\}. \tag{3.3}$$

Proof. For both bounds, our starting point is the formula

$$\mathbb{P}[f(Y_1) = Y_0] = \sum_{i\in\mathcal{Y}_0} \sum_{j\in\mathcal{Y}_1} \mathbf{1}_{\{f(j)=i\}}\, \mu_1^i(j)\, \mu_0(i). \tag{3.4}$$

We will also need the following alternative expression for total variation distance

$$\|\mu_1^{i_1} - \mu_1^{i_2}\|_{\mathrm{TV}} = 1 - \sum_{j\in\mathcal{Y}_1} \mu_1^{i_1}(j) \wedge \mu_1^{i_2}(j), \tag{3.5}$$

which follows from the last equality in (1.4).

To derive (3.2), observe first that by (3.4) for any f

$$\mathbb{P}[f(Y_1) = Y_0] \leq \sum_{j\in\mathcal{Y}_1} \sup_{i\in\mathcal{Y}_0} \mu_1^i(j)\,\mu_0(i) = \mathbb{P}[f^*(Y_1) = Y_0],$$

where f^* is a maximum a posteriori estimate

$$f^*(j) \in \arg\max_{i\in\mathcal{Y}_0} \left\{\mu_1^i(j)\,\mu_0(i)\right\}. \tag{3.6}$$

This is well-known; see e.g. [LC98, Chapter 4]. The maximum above is indeed attained because $\mu_1^i(j)\,\mu_0(i)$ is summable over i, and therefore must converge to 0. Then, by (3.4) applied to $f = f^*$, for any $i_1 \neq i_2 \in \mathcal{Y}_0$

$$\begin{aligned}\mathbb{P}[f^*(Y_1) = Y_0] &\leq \sum_{i\in\mathcal{Y}_0}\sum_{j\in\mathcal{Y}_1} \mu_1^i(j)\,\mu_0(i) - \sum_{j\in\mathcal{Y}_1} [\mu_1^{i_1}(j)\,\mu_0(i_1)] \wedge [\mu_1^{i_2}(j)\,\mu_0(i_2)] \\ &\leq 1 - (\mu_0(i_1)\wedge\mu_0(i_2)) \sum_{j\in\mathcal{Y}_1} \mu_1^{i_1}(j)\wedge\mu_1^{i_2}(j).\end{aligned}$$

Bound (3.2) then follows from (3.5) and taking a supremum over $i_1 \neq i_2$.

For (3.3), define the approximate maximum a posteriori estimator

$$f_\Lambda^*(j) \in \arg\max_{i\in\Lambda} \left\{\mu_1^i(j)\,\mu_0(i)\right\},$$

where note that, this time, the supremum is over Λ only. Then (3.4) applied to $f = f_\Lambda^*$ implies

$$\begin{aligned}\mathbb{P}[f_\Lambda^*(Y_1) = Y_0] &= \sum_{i\in\Lambda}\sum_{j\in\mathcal{Y}_1} \mathbf{1}_{\{f_\Lambda^*(j)=i\}}\,\mu_1^i(j)\,\mu_0(i) \\ &= \sum_{i\in\Lambda}\sum_{j\in\mathcal{Y}_1} \mu_1^i(j)\,\mu_0(i) - \sum_{j\in\mathcal{Y}_1}\sum_{i\neq f_\Lambda^*(j)} \mu_1^i(j)\,\mu_0(i) \\ &\geq \sum_{i\in\Lambda}\mu_0(i) - \sum_{j\in\mathcal{Y}_1}\sum_{i_1\neq i_2\in\Lambda} [\mu_1^{i_1}(j)\mu_0(i_1)]\wedge[\mu_1^{i_2}(j)\mu_0(i_2)] \\ &\geq \sum_{i\in\Lambda}\mu_0(i) - \sum_{i_1\neq i_2\in\Lambda}\left\{(\mu_0(i_1)\vee\mu_0(i_2))\sum_{j\in\mathcal{Y}_1}\mu_1^{i_1}(j)\wedge\mu_1^{i_2}(j)\right\}.\end{aligned}$$

By (3.5), that implies (3.3) and concludes the proof. □

3.2 Characterization of consistent root reconstruction

From Lemma 3.2, we obtain a characterization of consistent root reconstruction in terms of total variation. This characterization is key to proving both directions of Theorem 1.6. Recall that $\mathcal{L}_T^i$ was defined in (1.3) as the leaf distribution on T given root state i.

Lemma 3.3 (Consistent root reconstruction: characterization). *Let $\{T^k\}_k$ and $(\mathbf{P}_t)_t$ satisfy our standing assumptions (i) and (ii), and let π be a probability distribution on $\mathcal{S}$. Then there exists a sequence of root estimators that is consistent for $\{T^k\}_k$, $(\mathbf{P}_t)_t$ and π if and only if for all $i \neq j \in \mathcal{S}$ such that $\pi(i)\wedge\pi(j) > 0$*

$$\liminf_{k\to\infty} \|\mathcal{L}_{T^k}^i - \mathcal{L}_{T^k}^j\|_{\mathrm{TV}} = 1. \tag{3.7}$$

Proof. For the *only if* part, assume by contradiction that there is $i_1 \neq i_2 \in \mathcal{S}$ with $\pi(i_1) \wedge \pi(i_2) > 0$, $\epsilon > 0$ and $k_0 \geq 1$ such that

$$\|\mathcal{L}^{i_1}_{T^k} - \mathcal{L}^{i_2}_{T^k}\|_{\mathrm{TV}} \leq 1 - \epsilon,$$

for all $k \geq k_0$. By (3.2) in Lemma 3.2, for all $k \geq k_0$ and any root estimator F_k

$$\mathbb{P}^\pi[F_k(X^k_{\partial T^k}) = X^k_\rho] \leq 1 - [\pi(i_1) \wedge \pi(i_2)]\,\epsilon < 1.$$

That proves that consistent root estimation is not possible.

For the *if* part, assume (3.7) holds. Fix $\epsilon > 0$ and let $1 \leq n_\epsilon < +\infty$ be the smallest integer such that

$$\sum_{i \leq n_\epsilon} \pi(i) > 1 - \epsilon, \tag{3.8}$$

and let $\Lambda_\epsilon = \{i \,:\, i \leq n_\epsilon\}$. Applying (3.3) in Lemma 3.2 with $\Lambda = \Lambda_\epsilon$, we get by (3.7) and (3.8)

$$\begin{aligned}\sup_{F_k} \mathbb{P}^\pi[F_k(X^k_{\partial T^k}) = X^k_\rho] &\geq 1 - \epsilon - \sum_{i_1 \neq i_2 \in \Lambda_\epsilon} \left\{\pi(i_1) \vee \pi(i_2)\left[1 - \|\mathcal{L}^{i_1}_{T^k} - \mathcal{L}^{i_2}_{T^k}\|_{\mathrm{TV}}\right]\right\}\\ &\to 1 - \epsilon,\end{aligned}$$

as $k \to \infty$. Because ϵ is arbitrary, we have shown that a sequence of maximum posteriori estimates is consistent for $\{T^k\}_k$, $(\mathbf{P}_t)_t$ and π. □

3.3 Proof of Proposition 3.1

We now prove our main result of this section.

Proof of Proposition 3.1. Let $\{T^k\}_k$ and $(\mathbf{P}_t)_t$ satisfy our standing assumptions (i) and (ii), and let π be a probability distribution on $\mathcal{S}$. Assume that $\{T^k\}_k$ satisfies neither downstream disjointness nor the big bang condition. Then, as we argued in the proof of Lemma 2.2, there exist $s_0 \in (0, \infty)$ and $k_0 \geq 1$ such that the truncation $T^k(s_0)$ remains unchanged for all $k \geq k_0$. Since downstream disjointness fails and $\ell_u > 0$ for all $u \in \partial T^k$ (by the positivity assumption on ℓ), there are $i_1 \neq i_2$ with $\pi(i_1) > 0$ and $\pi(i_2) > 0$ such that the supports of $\mathbb{P}^{i_1}[X^k_{\partial T^k(s_0)} \in \cdot\,]$ and $\mathbb{P}^{i_2}[X^k_{\partial T^k(s_0)} \in \cdot\,]$ have a non-empty intersection. This holds for all k and implies that

$$\|\mathbb{P}^{i_1}[X^k_{\partial T^k(s_0)} \in \cdot\,] - \mathbb{P}^{i_2}[X^k_{\partial T^k(s_0)} \in \cdot\,]\|_{\mathrm{TV}} < 1.$$

Because $T^k(s_0)$ is unchanged after k_0, it follows that

$$\liminf_{k \to \infty} \|\mathbb{P}^{i_1}[X^k_{\partial T^k(s_0)} \in \cdot\,] - \mathbb{P}^{i_2}[X^k_{\partial T^k(s_0)} \in \cdot\,]\|_{\mathrm{TV}} < 1. \tag{3.9}$$

Finally we observe that, by the triangle inequality and the conditional independence of X^k_ρ and $X^k_{\partial T^k}$ given $X^k_{\partial T^k(s_0)}$, we get

$$\begin{aligned}\|\mathcal{L}^{i_1}_{T^k} - \mathcal{L}^{i_2}_{T^k}\|_{\mathrm{TV}} &= \frac{1}{2}\sum_{\sigma \in \mathcal{S}^{\partial T^k}} \left|\mathcal{L}^{i_1}_{T^k}(\sigma) - \mathcal{L}^{i_2}_{T^k}(\sigma)\right|\\ &= \frac{1}{2}\sum_{\sigma \in \mathcal{S}^{\partial T^k}} \Bigg| \sum_{\tau \in \mathcal{S}^{\partial T^k(s_0)}} \mathbb{P}[X^k_{\partial T^k} = \sigma \,|\, X^k_{\partial T^k(s_0)} = \tau]\\ &\qquad \left[\mathbb{P}^{i_1}[X^k_{\partial T^k(s_0)} = \tau] - \mathbb{P}^{i_2}[X^k_{\partial T^k(s_0)} = \tau]\right]\Bigg|\\ &\leq \|\mathbb{P}^{i_1}[X^k_{\partial T^k(s_0)} \in \cdot\,] - \mathbb{P}^{i_2}[X^k_{\partial T^k(s_0)} \in \cdot\,]\|_{\mathrm{TV}}.\end{aligned} \tag{3.10}$$

Combining this inequality with (3.9) shows by Lemma 3.3 that consistent root estimation is not possible in this case. That concludes the proof. □

4 Consistent root reconstruction

In this section, we prove the *if* part of Theorem 1.6. Observe first that, under downstream disjointness, the result is immediate. Let $u \in \partial T^1$ and $I = \{i : \pi(i) > 0\}$. Note that, by the nested property, $u \in \partial T^k$ for all k. Then, let $F_k(X^k_{\partial T^k})$ be the state in I from which X^k_u is reachable. Downstream disjointness ensures that such a state exists and is unique. We then have $\mathbb{P}^\pi[F_k(X^k_{\partial T^k}) = X^k_\rho] = 1$, proving consistency in that case.

Here we show that the big bang condition also suffices for consistent root reconstruction. We use the characterization in Lemma 3.3 to reduce the problem to pairs of initial states. Our strategy is then to extract a "well-spread" subtree of T^k, as we did in the proof of Lemma 2.2, and generalize results of [GS10] on root reconstruction for well-spread trees. Formally we prove the following proposition, which together with Lemma 3.3 and the argument above in the downstream disjointness case, implies the *if* part of Theorem 1.6.

Proposition 4.1 (Reconstruction under the big bang condition)**.** *Let $\{T^k\}_k$ and $(\mathbf{P}_t)_t$ satisfy our standing assumptions (i) and (ii), and let π be a probability distribution on $\mathcal{S}$. Assume that $\{T^k\}_k$ satisfies the big bang condition. Then for all $i \neq j \in \mathcal{S}$ such that $\pi(i) \wedge \pi(j) > 0$*

$$\liminf_{k\to\infty} \|\mathcal{L}^i_{T^k} - \mathcal{L}^j_{T^k}\|_{\mathrm{TV}} = 1. \qquad (4.1)$$

4.1 Well-spread restriction

We will use the following construction. We extract a well-spread restriction of T^k and stretch the leaf edges to enforce that all leaves are at the same distance from the root. Fix $k \geq 1$ and $s > 0$. Recall that h^* is a (uniform) bound on the height of the trees.

- **Step 1: Well-spread restriction.** By Lemma 2.2, there exists a a nested sequence of restrictions with vanishing spread. Let $\widetilde{T}^{k,s}$ be the restriction of T^k constructed in the proof of Lemma 2.2. Recall that $\widetilde{T}^{k,s}$ is $(1-s)$-spread and has $|\partial T^k(s)|$ leaves.
- **Step 2: Stretching.** We then modify $\widetilde{T}^{k,s}$ to make all leaves be at distance h^* from the root as follows. For each leaf $x \in \partial\widetilde{T}^{k,s}$, we extend the corresponding leaf edge by $h^* - \ell_x$ and run the $\mathbf{P}_t$-chain started at X^k_x for time $h^* - \ell_x$. We then let $\widehat{T}^{k,s}$ be the resulting tree and assign the states generated above along the extensions. Observe that $\widehat{T}^{k,s}$, like $\widetilde{T}^{k,s}$, is $(1-s)$-spread and has $|\partial T^k(s)|$ leaves.

Let $N^{(k)}_j$ be the number of leaves of the stretched restriction $\widehat{T}^{k,s}$ that are in state $j \in \mathcal{S}$ and let $\mathbf{N}^{k,s} = (N^{k,s}_1, N^{k,s}_2, \cdots)$. Denote by $\mathcal{M}^i_{\widehat{T}^{k,s}}$ the law of $\mathbf{N}^{k,s}$ when the root state is i. By a computation similar to (3.10), by the conditional independence of $\mathbf{N}^{k,s}$ and X^k_ρ given $X^k_{\partial T^k}$, we have that

$$\|\mathcal{M}^i_{\widehat{T}^{k,s}} - \mathcal{M}^j_{\widehat{T}^{k,s}}\|_{\mathrm{TV}} \leq \|\mathcal{L}^i_{T^k} - \mathcal{L}^j_{T^k}\|_{\mathrm{TV}}.$$

Therefore, Proposition 4.1 follows from the following lemma.

Lemma 4.2 (Separation of state frequencies on stretched restrictions)**.** *Consider the setting of Proposition 4.1 and let $\mathcal{M}^i_{\widehat{T}^{k,s}}$ be as defined above. Then,*

$$\sup_{s>0} \liminf_{k\to\infty} \|\mathcal{M}^i_{\widehat{T}^{k,s}} - \mathcal{M}^j_{\widehat{T}^{k,s}}\|_{\mathrm{TV}} = 1.$$

When all leaves of T^k are assumed to be at the same distance from the root, T^k is said to be ultrametric (see e.g. [SS03, Chapter 7]). Here we do not make this assumption on T^k. Instead we enforce it artificially through the stretching in Step 2. The reason we do this is that our proof relies on initial-state identifiability which, by (1.5), implies

$$\|\mathbf{p}^i(h^*) - \mathbf{p}^j(h^*)\|_{\mathrm{TV}} > 0, \qquad \forall i \neq j \in \mathcal{S}. \qquad (4.2)$$

In contrast, it may not be the case that the expected state frequencies at $\partial\widetilde{T}^{k,s}$, that is,

$$\frac{1}{|\partial T^k(s)|}\sum_{x\in\partial\widetilde{T}^{k,s}}\mathbf{p}^i(\ell_x)$$

uniquely characterize the root state i.

4.2 Variance bound

The proof of Lemma 4.2 relies on the following variance bound, which generalizes a result of [GS10, Proof of Lemma 3.2]. Recall the definition of q_i in (1.2).

Lemma 4.3 (Variance bound)**.** *Let $T=(V,E,\rho,\ell)$ be a tree and let $(X_\gamma)_{\gamma\in\Gamma_T}$ be a $\mathbf{P}_t$-chain on T. Let N_j be the number of leaves of T in state $j\in\mathcal{S}$. Then for all $i,j\in\mathcal{S}$,*

$$\mathrm{Var}_i(N_j)\le\frac{1}{4}|\partial T|+2(q_i\vee 1)\,\mathrm{Spr}(T)\,|\partial T|^2, \tag{4.3}$$

where we denote by Var_i the variance under $\mathbb{P}^i$.

Proof. Let θ_x^j be the indicator random variable for the event "leaf x is in state j." Then

$$N_j=\sum_{x\in\partial T}\theta_x^j,$$

and, hence,

$$\mathrm{Var}_i(N_j)=\sum_x\mathrm{Var}_i(\theta_x^j)+\sum_{x\neq y}\mathrm{Cov}_i(\theta_x^j,\ \theta_y^j). \tag{4.4}$$

Because $\theta_x^j\in\{0,1\}$, we have

$$\mathrm{Var}_i(\theta_x^j)=\mathbb{P}^i[\theta_x^j=1](1-\mathbb{P}^i[\theta_x^j=1])\le 1/4,$$

leading to the first term on the RHS of (4.3). For $x\neq y$, we have that

$$\begin{aligned}\mathrm{Cov}_i(\theta_x^j,\ \theta_y^j)&=\mathbb{E}^i\big[\big(\theta_x^j-p_{ij}(\ell_x)\big)\big(\theta_y^j-p_{ij}(\ell_y)\big)\big]\\&=\sum_{k\in\mathcal{S}}p_{ik}(\ell_{xy})\big(p_{kj}(\ell_x-\ell_{xy})-p_{ij}(\ell_x)\big)\big(p_{kj}(\ell_y-\ell_{xy})-p_{ij}(\ell_y)\big),\end{aligned}$$

which is obtained by conditioning on the state at the divergence point between the paths from the root to x and y. Splitting the sum according to whether $k=i$, we have

$$\begin{aligned}|\mathrm{Cov}_i(\theta_x^j,\ \theta_y^j)|&\le|p_{ij}(\ell_x-\ell_{xy})-p_{ij}(\ell_x)|+\sum_{k\neq i}p_{ik}(\ell_{xy}) &(4.5)\\&\le 2[(q_i\ell_{xy})\wedge 1]. &(4.6)\end{aligned}$$

To see inequality (4.6), note that the second term on the RHS of (4.5) is bounded above by the probability that the state is changed at least once along the shared path from the root to x and y, which is equal to $1-\exp(-q_i\ell_{xy})\le(q_i\ell_{xy})\wedge 1$ (see e.g. [Lig10, Chapter 2]). For the first term, the Chapman-Kolmogorov equations (see e.g. [Lig10, Chapter 2]) imply that, for all $t\ge 0$ and $\delta>0$,

$$p_{ij}(t+\delta)-p_{ij}(t)=\sum_k p_{ik}(\delta)p_{kj}(t)-p_{ij}(t)$$

so that

$$p_{ij}(t+\delta)-p_{ij}(t)\le\sum_{k\neq i}p_{ik}(\delta)=1-p_{ii}(\delta)\le 1-\exp(-q_i\delta),$$

and

$$p_{ij}(t+\delta)-p_{ij}(t)\ge-(1-p_{ii}(\delta))p_{ij}(t)\ge-(1-\exp(-q_i\delta)).$$

The proof is complete in view of (4.4) and the definition of the spread. □

4.3 Proof of Lemma 4.2

Proof of Lemma 4.2. It suffices to find a sequence of events $\mathcal{A}_k$, $k \geq 1$, depending only on $\mathbf{N}^{k,s}$ such that

$$\sup_{s>0} \liminf_{k\to\infty} \mathbb{P}^i[\mathcal{A}^{k,s}] = 1 \qquad \text{and} \qquad \inf_{s>0} \limsup_{k\to\infty} \mathbb{P}^j[\mathcal{A}^{k,s}] = 0,$$

that is, a sequence of events asymptotically likely under $\mathcal{M}^i_{\widehat{T}^{k,s}}$ but unlikely under $\mathcal{M}^j_{\widehat{T}^{k,s}}$.

Consider the norm $\|\cdot\|_*$ defined as

$$\|\mathbf{v}\|_* := \sum_{i=1}^{|\mathcal{S}|} 2^{-i}|v_i|,$$

for $\mathbf{v} = (v_1, v_2, \ldots)$. We claim that (4.2) is equivalent to

$$\Delta^*_{i,j} := \|\mathbf{p}^i(h^*) - \mathbf{p}^j(h^*)\|_* > 0. \tag{4.7}$$

Indeed, by the definition of the norms, we have $\|\cdot\|_* \leq \|\cdot\|_{\mathrm{TV}}$. For the other direction, note that, for any $\delta > 0$, there exists M such that $\sum_{k>M} 2^{-k} < \delta/2$ and so $\|\mu - \nu\|_{\mathrm{TV}} \leq \delta/2 + 2^M\|\mu - \nu\|_*$ for any probability distributions μ and ν. We consider the following events

$$\mathcal{A}^{k,s} = \left\{ \left\| \frac{\mathbf{N}^{k,s}}{|\partial T^k(s)|} - \mathbf{p}^i(h^*) \right\|_* < \frac{\Delta^*_{i,j}}{2} \right\}.$$

Because $\widehat{T}^{k,s}$ is $(1-s)$-spread, the variance bound in Lemma 4.3 implies that for $i, j \in \mathcal{S}$

$$\mathrm{Var}_i\left(N^{k,s}_j\right) \leq \frac{|\partial T^k(s)|}{4} + 2\,(q_i \vee 1)\, s\, |\partial T^k(s)|^2, \tag{4.8}$$

By the Cauchy-Schwarz inequality and (4.8),

$$\begin{aligned}
\mathbb{E}^i\left[\left\| \frac{\mathbf{N}^{k,s}}{|\partial T^k(s)|} - \mathbf{p}^i(h^*) \right\|_*^2\right] &= \mathbb{E}^i\left[\left(\sum_{j=1}^{|\mathcal{S}|} 2^{-j} \left| \frac{N^{k,s}_j}{|\partial T^k(s)|} - p_{ij}(h^*) \right|\right)^2\right] \\
&\leq \left(\sum_{j=1}^{|\mathcal{S}|} 2^{-j}\right)\left(\sum_{j=1}^{|\mathcal{S}|} 2^{-j}\mathrm{Var}_i\left(\frac{N^{k,s}_j}{|\partial T^k(s)|}\right)\right) \\
&\leq \frac{1}{4|\partial T^k(s)|} + 2(q_i \vee 1)s.
\end{aligned} \tag{4.9}$$

By Chebyshev's inequality (see e.g. [Dur10]),

$$\begin{aligned}
\mathbb{P}^i\left[\left\| \frac{\mathbf{N}^{k,s}}{|\partial T^k(s)|} - \mathbf{p}^i(h^*) \right\|_* \geq \frac{\Delta^*_{i,j}}{2}\right] &\leq \frac{4}{(\Delta^*_{i,j})^2}\, \mathbb{E}_i\left[\left\| \frac{\mathbf{N}^{k,s}}{|\partial T^k(s)|} - \mathbf{p}^i(h^*) \right\|_*^2\right] \\
&\leq \frac{4}{(\Delta^*_{i,j})^2}\left[\frac{1}{4|\partial T^k(s)|} + 2(q_i \vee 1)s\right],
\end{aligned} \tag{4.10}$$

where we used (4.9). By the big bang condition and (4.7), taking $k \to +\infty$ and then $s \to 0$, we get

$$\inf_{s>0} \limsup_{k\to\infty} \mathbb{P}^i[(\mathcal{A}^{k,s})^c] = 0.$$

Similarly, noting that by the triangle inequality and the definition of $\Delta^*_{i,j}$,

$$\mathbb{P}^j\left[\left\| \frac{\mathbf{N}^{k,s}}{|\partial T^k(s)|} - \mathbf{p}^i(h^*) \right\|_* < \frac{\Delta^*_{i,j}}{2}\right] \leq \mathbb{P}^j\left[\left\| \frac{\mathbf{N}^{k,s}}{|\partial T^k(s)|} - \mathbf{p}^j(h^*) \right\|_* \geq \frac{\Delta^*_{i,j}}{2}\right],$$

we also get that

$$\inf_{s>0} \limsup_{k\to\infty} \mathbb{P}^j[\mathcal{A}^{k,s}] = 0.$$

The proof is complete. □

5 Error bounds

The proof of Lemma 4.2 actually implies an explicit bound on the error probability (see (4.10)). That bound decays like the inverse of $|\partial T^k(s)|$. This is far from best possible: take for instance the star tree where, by conditional independence of the leaf states given the root state, one would expect an exponential inequality. Here we give an improved bound on the achievable error probability which decays exponentially in $|\partial T^k(s)|$. We also express this bound in terms of the more natural total variation distance.

Our main result is the following proposition, which implies the first part of Theorem 1.7. (The second part of the theorem is proved in Section 5.3.) For $\epsilon > 0$, recall that $n_\epsilon < \infty$ be the smallest integer such that $\sum_{i>n_\epsilon} \pi(i) < \epsilon$ and that $\Lambda_\epsilon = \{i \in \mathcal{S} \,:\, i \le n_\epsilon\}$,

$$q^*_\epsilon = \max_{i\in\Lambda_\epsilon} (q_i \vee 1),$$

and

$$\Delta_\epsilon = \min_{i_1\neq i_2\in\Lambda_\epsilon} \|\mathbf{p}^{i_1}(h^*) - \mathbf{p}^{i_2}(h^*)\|_{\mathrm{TV}}.$$

Proposition 5.1 (Achievable error bound)**.** *Fix $\epsilon > 0$ and $k \ge 1$. Then there exist universal constants $C_0, C_1 > 0$ and an estimator F_k such that the following holds. For all $s > 0$,*

$$\mathbb{P}^\pi\left[F_k(X^k_{\partial T^k}) \neq X^k_\rho\right] < \epsilon + C_0\, \Delta_\epsilon^{-2}\, q^*_\epsilon\, s + n_\epsilon \exp\left(-C_1\, \Delta_\epsilon^2\, |\partial T^k(s)|\right).$$

5.1 Deviation of frequencies

To prove Proposition 5.1, we devise a root estimator (described in details in the next subsection) based on the combinatorial construction of Section 4.1. Fix $k \ge 1$ and $s > 0$. Given the leaf states $X^k_{\partial T^k} \in \mathcal{S}^{\partial T^k}$ of the original tree T^k, we extract the subtree $\widetilde{T}^{k,s}$, run a simulation of the $\mathbf{P}_t$-chain on the extended tree $\widehat{T}^{k,s}$, and treat the leaf states of $\widehat{T}^{k,s}$ as the observed leaf states. For a subset $\mathcal{A} \subseteq \mathcal{S}$, let $N^{k,s}_{\mathcal{A}}$ be the number of leaves of $\widehat{T}^{k,s}$ whose state is in $\mathcal{A}$. The proof of Proposition 5.1 requires a bound on the deviation of $N^{k,s}_{\mathcal{A}}$. To obtain such a bound, we proceed by first controlling the number of points in $\partial T^k(s)$ whose state coincides with the root state.

Let i be state at the root. For any vertex v, let Z_v be 1 if the state at v is i, and let Z_v be 0 otherwise. Let $\mathcal{W}_i$ be those vertices in $\partial T^k(s)$ in state i. In particular

$$S_i = |\mathcal{W}_i| = \sum_{x\in\partial T^k(s)} Z_x.$$

Let $\widehat{N}_{\mathcal{A}}$ be the number of descendant leaves of $\mathcal{W}_i$ in $\widehat{T}^{k,s}$ whose states are in $\mathcal{A}$. We also let $m = |\partial T^k(s)|$. Then, we can bound $N^{k,s}_{\mathcal{A}}$ as follows

$$\widehat{N}_{\mathcal{A}} \le N^{k,s}_{\mathcal{A}} \le \widehat{N}_{\mathcal{A}} + m - S_i. \tag{5.1}$$

Conditioned on S_i, note that $\widehat{N}_{\mathcal{A}}$ is a binomial random variable, specifically, $\mathrm{Bin}(S_i,\, p_{i\mathcal{A}}(h^* - s))$, where $p_{i\mathcal{A}}(t)$ denotes the probability that the state is in $\mathcal{A}$ given that initially it is i. To bound the probability that $N^{k,s}_{\mathcal{A}}$ is close to its expectation, we argue in two steps. We first bound the probability that S_i itself is close to its expectation, then we apply a concentration inequality to $N^{k,s}_{\mathcal{A}}$ conditioned on that event.

Lemma 5.2 (Control of S_i). *Define the event*

$$\mathcal{E}_\delta^0 = \left\{ \left| S_i - \mathbb{E}^i[S_i] \right| > \delta m \right\},$$

where

$$\mathbb{E}^i[S_i] = p_{ii}(s)\, m. \tag{5.2}$$

Then, we have the bound

$$\mathbb{P}^i\left(\mathcal{E}_\delta^0\right) \le \frac{1 - e^{-q_i s}}{\delta^2}. \tag{5.3}$$

Proof. We use Chebyshev's inequality to control the deviation of S_i. By the Cauchy-Schwarz inequality, the variance of S_i is bounded by

$$\begin{aligned}
\mathrm{Var}_i[S_i] &= \mathrm{Var}_i\left[\sum_{x\in\partial T^k(s)} Z_x\right] \\
&= \sum_{x\in\partial T^k(s)} \sum_{y\in\partial T^k(s)} \mathbb{E}^i\left[(Z_x - p_{ii}(s))(Z_y - p_{ii}(s))\right] \\
&\le \sum_{x\in\partial T^k(s)} \sum_{y\in\partial T^k(s)} \sqrt{\mathrm{Var}_i[Z_x]\mathrm{Var}_i[Z_y]} \\
&= m^2 p_{ii}(s)(1 - p_{ii}(s)) \\
&\le m^2(1 - e^{-q_i s}),
\end{aligned}$$

where on the last line we used that the probability of being at state i at time s is at least the probability of never having left state i up to time s, i.e., $e^{-q_i s} \le p_{ii}(s) \le 1$ (see e.g. [Lig10, Chapter 2]). The result by Chebyshev's inequality. □

Lemma 5.3 ($N_{\mathcal{A}}^{k,s}$ is close to its expectation given $\mathcal{E}_\delta^0$). *Fix a subset $\mathcal{A} \subseteq \mathcal{S}$. Let $\delta > 0$. Then, the following bound holds*

$$\mathbb{P}^i\left[N_{\mathcal{A}}^{k,s} < p_{i\mathcal{A}}(h^*)\, m - [(1 - e^{-q_i s}) + 2\delta]m \,\middle|\, \mathcal{E}_\delta^0\right] \le \exp\left(-\frac{2\delta^2}{1+\delta} m\right).$$

Proof. We proceed in three steps:

1. **Conditional control of $\widehat{N}_{\mathcal{A}}$.** Condition on S_i. Define the event

$$\mathcal{E}_\delta^1 = \left\{ \widehat{N}_{\mathcal{A}} < \mathbb{E}[\widehat{N}_{\mathcal{A}} \,|\, S_i] - \delta m \right\}.$$

Here

$$\mathbb{E}\left[\widehat{N}_{\mathcal{A}} \,\middle|\, S_i\right] = p_{i\mathcal{A}}(h^* - s)\, S_i. \tag{5.4}$$

By Hoeffding's inequality [Hoe63], we then have

$$\mathbb{P}\left[\mathcal{E}_\delta^1 \,|\, S_i\right] \le \exp\left(-2\frac{\delta^2 m^2}{S_i}\right). \tag{5.5}$$

2. **Approximation of $p_{i\mathcal{A}}(h^*)$.** By (5.2) and (5.4), the expectation of $\widehat{N}_{\mathcal{A}}$ is

$$\mathbb{E}^i\left[\widehat{N}_{\mathcal{A}}\right] = p_{i\mathcal{A}}(h^* - s)\, p_{ii}(s)\, m.$$

To relate it to the expectation of $N_{\mathcal{A}}^{k,s}$, we note that

$$p_{i\mathcal{A}}(h^*) = \sum_{k\in\mathcal{S}} p_{ik}(s)\, p_{k\mathcal{A}}(h^*-s) = p_{ii}(s)\, p_{i\mathcal{A}}(h^*-s) + \sum_{k\neq i} p_{ik}(s)\, p_{k\mathcal{A}}(h^*-s),$$

so

$$\begin{aligned} |p_{i\mathcal{A}}(h^*) - p_{i\mathcal{A}}(h^*-s)\, p_{ii}(s)| &\leq \sum_{k\neq i} p_{ik}\,(s) p_{k\mathcal{A}}(h^*-s) \\ &\leq \sum_{k\neq i} p_{ik}(s) \\ &= 1-p_{ii}(s) \\ &\leq 1-e^{-q_i s}. \end{aligned} \tag{5.6}$$

3. **Overall error.** Under the event $(\mathcal{E}_\delta^0)^c \cap (\mathcal{E}_\delta^1)^c$, we have by (5.2) and (5.4) that

$$\widehat{N}_{\mathcal{A}} \geq p_{i\mathcal{A}}(h^*-s)\, p_{ii}(s)\, m - 2\delta m,$$

where we used $p_{ii}(s)\leq 1$. In turn, by (5.6) and (5.1),

$$\begin{aligned} N_{\mathcal{A}}^{k,s} - p_{i\mathcal{A}}(h^*)\, m &\geq N_{\mathcal{A}}^{k,s} - p_{i\mathcal{A}}(h^*-s)\, p_{ii}(s)\, m - (1-e^{-q_i s})m \\ &\geq N_{\mathcal{A}}^{k,s} - \widehat{N}_{\mathcal{A}} - (1-e^{-q_i s})m - 2\delta m \\ &\geq -(1-e^{-q_i s})m - 2\delta m. \end{aligned}$$

Define the event

$$\mathcal{E}_\delta^2 = \left\{ N_{\mathcal{A}}^{k,s} < p_{i\mathcal{A}}(h^*)\, m - [(1-e^{-q_i s}) + 2\delta]m \right\}.$$

Thus, by the above,

$$\begin{aligned} \mathbb{P}^i[\mathcal{E}_\delta^2] &\leq \mathbb{P}^i[\mathcal{E}_\delta^0 \cup \mathcal{E}_\delta^1] \\ &\leq \mathbb{P}^i[\mathcal{E}_\delta^0] + \mathbb{P}^i[\mathcal{E}_\delta^1 \cap (\mathcal{E}_\delta^0)^c] \\ &\leq \mathbb{P}^i[\mathcal{E}_\delta^0] + \mathbb{P}^i[\mathcal{E}_\delta^1 \,|\, (\mathcal{E}_\delta^0)^c] \\ &\leq \frac{1-e^{-q_i s}}{\delta^2} + \exp\left(-2\frac{\delta^2 m^2}{[p_{ii}(s)+\delta]m}\right) \\ &\leq \frac{1-e^{-q_i s}}{\delta^2} + \exp\left(-\frac{2\delta^2}{1+\delta} m\right) \end{aligned}$$

by (5.3) and (5.5).

That concludes the proof. □

5.2 Analysis of root estimator

We now describe our root estimator. In fact, we construct a randomized estimator (which can be made deterministic by choosing for each input the output most likely to be correct.) We restrict ourselves to a subset of root states that has high probability under π and we estimate the frequencies of events achieving the total variation distance between the leaf distributions given different root states. Fix $\epsilon > 0$ and let $\Lambda = \Lambda_\epsilon$.

Root estimator Our root estimator $G_k^\Lambda : \mathcal{S}^{\partial T^k} \to \mathcal{S}$ is defined as follows. Let $N_{\mathcal{A}}^{k,s}$ and m be defined as in the previous subsection.

- Define

$$\Delta = \inf_{i_1\neq i_2\in\Lambda} \|\mathbf{p}^{i_1}(h^*) - \mathbf{p}^{i_2}(h^*)\|_{\mathrm{TV}}.$$

- For every distinct pair of states $i_1, i_2 \in \Lambda$, let $\mathcal{A}_{i_1 \to i_2} \subseteq \mathcal{S}$ be an event achieving the total variation distance between $\mathbf{p}^{i_1}(h^*)$ and $\mathbf{p}^{i_2}(h^*)$, that is,

$$\|\mathbf{p}^{i_1}(h^*) - \mathbf{p}^{i_2}(h^*)\|_{\mathrm{TV}} = p_{i_1, \mathcal{A}_{i_1 \to i_2}}(h^*) - p_{i_2, \mathcal{A}_{i_1 \to i_2}}(h^*) > 0,$$

  where we also require that $\mathcal{A}_{i_1 \to i_2} = \mathcal{A}^c_{i_2 \to i_1}$.
- We let $G_k^{\Lambda}(X^k_{\partial T^k})$ be the state i passing the following tests

$$\frac{N^{k,s}_{\mathcal{A}_{i \to i'}}}{m} > p_{i\mathcal{A}_{i \to i'}}(h^*) - \frac{\Delta}{2}, \qquad \forall i' \neq i, \tag{5.7}$$

  if such a state exists; otherwise we let $G_k^{\Lambda}(X^k_{\partial T^k})$ be a state chosen uniformly at random in Λ.

Observe that at most one state can satisfy the condition in (5.7). Indeed, for any $i \neq i'$, if

$$\frac{N^{k,s}_{\mathcal{A}_{i \to i'}}}{m} > p_{i\mathcal{A}_{i \to i'}}(h^*) - \frac{\Delta}{2}$$

then

$$\frac{N^{k,s}_{\mathcal{A}_{i' \to i}}}{m} = 1 - \frac{N^{k,s}_{\mathcal{A}_{i \to i'}}}{m} < 1 - p_{i\mathcal{A}_{i \to i'}}(h^*) + \frac{\Delta}{2} = p_{i\mathcal{A}_{i' \to i}}(h^*) + \frac{\Delta}{2} < p_{i'\mathcal{A}_{i' \to i}}(h^*) - \frac{\Delta}{2},$$

where we used the definition of Δ and the fact that $\mathcal{A}_{i \to i'} = \mathcal{A}^c_{i' \to i}$. Observe also that G_k^{Λ} is randomized as a function of $X^k_{\partial T^k}$ since it depends on the states at the leaves of the extension $\widehat{T}^{k,s}$.

Analysis We now prove our main result of this section.

Proof of Proposition 5.1. Let $F_k = G_k^{\Lambda_\epsilon}$ be the estimator defined above, let the events $\mathcal{A}_{i \to i'}$ be as defined above and let i be the state at the root. By Lemmas 5.2 and 5.3,

$$\begin{aligned}
\mathbb{P}^i\left[F_k(X^k_{\partial T^k}) \neq i\right] &= \mathbb{P}^i\left[\exists i' \neq i,\ N^{k,s}_{\mathcal{A}_{i \to i'}} \leq p_{i\mathcal{A}_{i \to i'}}(h^*)\, m - \frac{\Delta_\epsilon}{2} m\right] \\
&\leq \mathbb{P}^i\left[\exists i' \neq i,\ N^{k,s}_{\mathcal{A}_{i \to i'}} \leq p_{i\mathcal{A}_{i \to i'}}(h^*)\, m - [(1 - e^{-q_i s}) + 2\delta] m\right] \\
&\leq \frac{1 - e^{-q_i s}}{\delta^2} + n_\epsilon \exp\left(-\frac{2\delta^2}{1+\delta} m\right),
\end{aligned}$$

provided $0 < 2\delta < \frac{\Delta_\epsilon}{2} - (1 - e^{-q^*_\epsilon s})$. Take $\delta = \frac{\Delta_\epsilon}{8}$ and s small enough that $1 - e^{-q^*_\epsilon s} \leq \Delta_\epsilon / 4$. The result follows. Note finally that, if $1 - e^{-q^*_\epsilon s} \leq \Delta_\epsilon/4$ fails, then the bound in Proposition 5.1 is trivially true as the RHS is then larger than 1. We leave that condition implicit in the statement. □

5.3 Uniform chains: minimax error bound

Here we consider chains with unformly bounded rates. We give a minimax error bound, that is, a bound uniform in the root state. We observe in Appendix A that

$$\Delta_{Q,h^*} = \inf_{i \neq j} \|\mathbf{p}^i(h^*) - \mathbf{p}^j(h^*)\|_{\mathrm{TV}} \geq \exp\left(-h^* \|Q\|\right) > 0, \tag{5.8}$$

where

$$\|Q\| = \sup_i \sum_j |q_{ij}| = \sup_i 2q_i < +\infty.$$

Let

$$q^* = \sup_{i \in \mathcal{S}} (q_i \vee 1) < +\infty,$$

and

$$f_* = e^{-q^* h^*}.$$

Note that

$$f_*^2 = e^{-2q^* h^*} \leq \Delta_{Q,h^*}. \tag{5.9}$$

We prove the following proposition, which implies the second part of Theorem 1.7.

Proposition 5.4 (Minimax error bound for uniform chains)**.** *Fix $k \geq 1$. There exist universal constants $C_0^U, C_1^U, C_2^U > 0$ and an estimator F_k^U such that the following holds. For all $s > 0$ and all i,*

$$\mathbb{P}^i \left[F_k^U(X_{\partial T^k}^k) \neq X_\rho^k \right] < C_0^U \, f_*^{-4} \, q^* \, s + C_2^U f_*^{-1} \exp\left(-C_1^U \, f_*^4 \, |\partial T^k(s)| \right).$$

Root estimator We modify the root estimator from Section 5.2. We use the same estimator G_k^Λ, but we choose a set Λ depending on the leaf states of the extended restriction. More precisely, fix $k \geq 1$ and $s > 0$. Recall the definitions of $\widehat{T}^{k,s}$ and $N_{\mathcal{A}}^{k,s}$ from Section 5.1. When $\mathcal{A} = \{j\}$, we write $N_j^{k,s}$ for $N_{\{j\}}^{k,s}$. Our modified estimator is defined as follows. We let

$$\widehat{\Lambda} = \left\{ j \in \mathcal{S} \, : \, \frac{N_j^{k,s}}{|\partial T^k(s)|} \geq \frac{1}{2} f_* \right\},$$

and we set $F_k^U = G_k^{\widehat{\Lambda}}$.

Analysis Let i be the state at the root. Recall the definitions of S_i and $\mathcal{E}_\delta^0$ from Section 5.1. We show first that, conditioned on $\mathcal{E}_\delta^0$, the set $\widehat{\Lambda}$ is highly likely to contain i, but highly unlikely to contain any state with low enough probability at the leaves. For $\alpha \in [0,1]$, define

$$\mathcal{J}_{i,\alpha} = \{ j \in \mathcal{S} \, : \, p_{ij}(h^*) \leq \alpha \}.$$

We write $\mathcal{J}_{i,\alpha}^c$ for $\mathcal{S} \setminus \mathcal{J}_{i,\alpha}$. Let $m = |\partial T^k(s)|$.

Lemma 5.5 (Properties of $\widehat{\Lambda}$)**.** *We have*

$$\mathbb{P}^i \left[i \notin \widehat{\Lambda} \, \middle| \, \mathcal{E}_\delta^0 \right] \leq \exp\left(-\frac{f_*^2}{64} m \right), \tag{5.10}$$

and

$$\mathbb{P}^i \left[\mathcal{J}_{i,f_*/3} \cap \widehat{\Lambda} \neq \emptyset \, \middle| \, \mathcal{E}_\delta^0 \right] \leq \left(6 f_*^{-1} + 1 \right) \, \exp\left(-\frac{f_*^2}{64} m \right), \tag{5.11}$$

provided $1 - e^{-q^* s} \leq f_*/4$ *and* $\delta \leq f_*/8$.

Proof. For (5.10), let

$$\mathcal{A} = \{i\},$$

and note that

$$p_{i\mathcal{A}}(h^*) \geq e^{-q^* h^*} = f_*.$$

By Lemma 5.3, provided $1 - e^{-q^* s} \leq f_*/4$ and $\delta \leq f_*/8$, we get

$$\mathbb{P}^i \left[N_i^{k,s} < \frac{f_*}{2} m \, \middle| \, \mathcal{E}_\delta^0 \right] \leq \exp\left(-\frac{f_*^2}{64} m \right).$$

For (5.11), consider a partition $\bigsqcup_{r=1}^{R} \mathcal{H}_{i,r}$ of $\mathcal{J}_{i,f_*/3}$ into the *smallest* number of subsets with

$$p_{i\mathcal{H}_{i,r}}(h^*) \le f_*/3.$$

Observe that $R \le 6/f_* + 1$. Indeed $\sum_{r=1}^{R} p_{i\mathcal{H}_{i,r}}(h^*) \le 1$ and, if two sets in the partition have $p_{i\mathcal{H}_{i,r}}(h^*) \le f_*/6$, then they can be combined into one. By Lemma 5.3, provided $1 - e^{-q^*s} \le f_*/4$ and $\delta \le f_*/8$, we get

$$\mathbb{P}^i\left[\exists r, N^{k,s}_{\mathcal{H}_{i,r}} \ge \frac{f_*}{2} m \,\middle|\, \mathcal{E}^0_\delta\right] \le R \exp\left(-\frac{f_*^2}{64} m\right).$$

Noting that $N^{k,s}_{\mathcal{H}_{i,r}} < f_*/2$ implies $N^{k,s}_j < f_*/2$ for all $j \in \mathcal{H}_{i,r}$ concludes the proof. □

Recall from Section 5.2 the definition of the events $\mathcal{A}_{i\to i'}$. By the previous lemma, the set $\widehat{\Lambda}$ is likely to contain only elements from $\mathcal{J}^c_{i,f_*/3}$—not necessarily all of them, but at least the root state i. We show next that under $\mathcal{E}^0_\delta$ the state i is likely to be chosen against all other states in $\mathcal{J}^c_{i,f_*/3}$ in the tests performed under $G_k^{\mathcal{J}^c_{i,f_*/3}}$.

Lemma 5.6 (Full set of potential tests)**.** *We have*

$$\mathbb{P}^i\left[\exists i' \in \mathcal{J}^c_{i,f_*/3} \setminus \{i\},\ N^{k,s}_{\mathcal{A}_{i\to i'}} \le p_{i\mathcal{A}_{i\to i'}}(h^*)\, m - \frac{\Delta_{Q,h^*}}{2} m \,\middle|\, \mathcal{E}^0_\delta\right] \le 3 f_*^{-1} \exp\left(-\frac{\Delta^2_{Q,h^*}}{64} m\right), \tag{5.12}$$

provided $1 - e^{-q^*s} \le \Delta_{Q,h^*}/4$ *and* $\delta \le \Delta_{Q,h^*}/8$*.*

Proof. We use an argument similar to that in the proof of Proposition 5.1. Observe first that

$$\left|\mathcal{J}^c_{i,f_*/3}\right| \le \frac{3}{f_*}.$$

By Lemmas 5.2 and 5.3,

$$\begin{aligned}
&\mathbb{P}^i\left[\exists i' \in \mathcal{J}^c_{i,f_*/3} \setminus \{i\},\ N^{k,s}_{\mathcal{A}_{i\to i'}} \le p_{i\mathcal{A}_{i\to i'}}(h^*)\, m - \frac{\Delta_{Q,h^*}}{2} m \,\middle|\, \mathcal{E}^0_\delta\right] \\
&\le \mathbb{P}^i\left[\exists i' \in \mathcal{J}^c_{i,f_*/3} \setminus \{i\},\ N^{k,s}_{\mathcal{A}_{i\to i'}} \le p_{i\mathcal{A}_{i\to i'}}(h^*)\, m - [(1 - e^{-q_i s}) + 2\delta] m \,\middle|\, \mathcal{E}^0_\delta\right] \\
&\le \frac{3}{f_*} \exp\left(-\frac{2\delta^2}{1+\delta} m\right),
\end{aligned}$$

provided $0 < 2\delta < \frac{\Delta_{Q,h^*}}{2} - (1 - e^{-q^*s})$. Take $\delta = \frac{\Delta_{Q,h^*}}{8}$ and s small enough that $1 - e^{-q^*s} \le \Delta_{Q,h^*}/4$. The result follows. □

Finally, we prove Proposition 5.4.

Proof of Proposition 5.4. Set $F_k = G_k^{\widehat{\Lambda}}$, and let i be the root state. We let $\mathcal{E}^3$ and $\mathcal{E}^4$ be the events

$$\mathcal{E}^3 = \left\{i \in \widehat{\Lambda}\right\} \cap \left\{\mathcal{J}_{i,f_*/3} \cap \widehat{\Lambda} = \emptyset\right\},$$

and

$$\mathcal{E}^4 = \left\{\forall i' \in \mathcal{J}^c_{i,f_*/3} \setminus \{i\},\ N^{k,s}_{\mathcal{A}_{i\to i'}} > p_{i\mathcal{A}_{i\to i'}}(h^*)\, m - \frac{\Delta_{Q,h^*}}{2} m\right\}.$$

Under $\mathcal{E}^3 \cap \mathcal{E}^4$, it holds that $F_k(X^k_{\partial T^k}) = i$. Thus, by Lemmas 5.2, 5.5 and 5.6,

$$\begin{aligned}
\mathbb{P}^i[F_k(X^k_{\partial T^k}) \ne X^k_\rho] &\le \mathbb{P}^i[(\mathcal{E}^0_\delta)^c] + \mathbb{P}^i\left[(\mathcal{E}^3)^c \,\middle|\, \mathcal{E}^0_\delta\right] + \mathbb{P}^i\left[(\mathcal{E}^4)^c \,\middle|\, \mathcal{E}^0_\delta\right] \\
&\le \frac{1 - e^{-q_i s}}{\delta^2} + \left(6 f_*^{-1} + 2\right) \exp\left(-\frac{f_*^2}{64} m\right) + 3 f_*^{-1} \exp\left(-\frac{\Delta^2_{Q,h^*}}{64} m\right) \\
&\le \frac{1 - e^{-q^*s}}{\delta^2} + 11 f_*^{-1} \exp\left(-\frac{(f_* \wedge \Delta_{Q,h^*})^2}{64} m\right),
\end{aligned}$$

provided $\delta \le (f_* \wedge \Delta_{Q,h^*})/8$ and $1 - e^{-q^* s} \le (f_* \wedge \Delta_{Q,h^*})/4$. As we did in Proposition 5.1, the latter condition is implicit. Using (5.9) concludes the proof. □

Acknowledgments. We thank Tom Kurtz, Ramon van Handel, and Mykhaylo Shkolnikov for helpful discussions.

References

[And91] William J. Anderson. *Continuous-time Markov chains*. Springer Series in Statistics: Probability and its Applications. Springer-Verlag, New York, 1991. An applications-oriented approach. MR-1118840

[ADHR12] Alexandr Andoni, Constantinos Daskalakis, Avinatan Hassidim, and Sebastien Roch. Global alignment of molecular sequences via ancestral state reconstruction. *Stochastic Processes and their Applications*, 122(12):3852–3874, 2012. MR-2971717

[BST10] Nayantara Bhatnagar, Allan Sly, and Prasad Tetali. Reconstruction threshold for the hardcore model. In *APPROX-RANDOM*, pages 434–447. Springer, 2010. MR-2755854

[BVVW11] Nayantara Bhatnagar, Juan Vera, Eric Vigoda, and Dror Weitz. Reconstruction for colorings on trees. *SIAM Journal on Discrete Mathematics*, 25(2):809–826, 2011. MR-2817532

[BRZ95] P. M. Bleher, J. Ruiz, and V. A. Zagrebnov. On the purity of the limiting Gibbs state for the Ising model on the Bethe lattice. *J. Statist. Phys.*, 79(1–2):473–482, 1995. MR-1325591

[BCMR06] Christian Borgs, Jennifer Chayes, Elchanan Mossel, and Sébastien Roch. The Kesten-Stigum reconstruction bound is tight for roughly symmetric binary channels. In *FOCS'06—Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, pages 518–530, 2006.

[CT06] Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition, 2006. MR-2239987

[Dur10] Rick Durrett. *Probability: theory and examples*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, fourth edition, 2010. MR-2722836

[EKPS00] W. S. Evans, C. Kenyon, Y. Peres, and L. J. Schulman. Broadcasting on trees and the Ising model. *Ann. Appl. Probab.*, 10(2):410–433, 2000. MR-1768240

[FR] Wai-Tong (Louis) Fan and Sebastien Roch. Efficient and consistent inference of ancestral sequences in an evolutionary model with insertions and deletions under dense taxon sampling. Preprint, 2017.

[Fel04] J. Felsenstein. *Inferring Phylogenies*. Sinauer, Sunderland, MA, 2004.

[GS10] Olivier Gascuel and Mike Steel. Inferring ancestral sequences in taxon-rich phylogenies. *Math. Biosci.*, 227(2):125–135, 2010. MR-2732449

[HA13] Lam Si Tung Ho and Cécile Ané. Asymptotic theory with hierarchical autocorrelation: Ornstein?uhlenbeck tree models. *Ann. Statist.*, 41(2):957–981, 04 2013. MR-3099127

[Hoe63] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58:13–30, 1963. MR-0144363

[Iof96] D. Ioffe. On the extremality of the disordered state for the Ising model on the Bethe lattice. *Lett. Math. Phys.*, 37(2):137–143, 1996. MR-1391195

[Ken67] David G. Kendall. On Markov groups. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66), Vol. II: Contributions to Probability Theory, Part 2*, pages 165–173. Univ. California Press, Berkeley, Calif., 1967. MR-0216584

[KS66] H. Kesten and B. P. Stigum. Additional limit theorems for indecomposable multidimensional Galton-Watson processes. *Ann. Math. Statist.*, 37:1463–1481, 1966. MR-0200979

[LC98] E. L. Lehmann and George Casella. *Theory of point estimation*. Springer Texts in Statistics. Springer-Verlag, New York, second edition, 1998. MR-1639875

[LR05] E. L. Lehmann and Joseph P. Romano. *Testing statistical hypotheses*. Springer Texts in Statistics. Springer, New York, third edition, 2005. MR-2135927

[Lib07] David A Liberles. *Ancestral sequence reconstruction*. Oxford University Press on Demand, 2007.

[Lig10] Thomas M. Liggett. *Continuous time Markov processes*, volume 113 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2010. An introduction. MR-2574430

[Mit09] Michael Mitzenmacher. A survey of results for deletion channels and related synchronization channels. *Probability Surveys*, 6:1–33, 2009. MR-2525669

[Mos01] E. Mossel. Reconstruction on trees: beating the second eigenvalue. *Ann. Appl. Probab.*, 11(1):285–300, 2001. MR-1825467

[MP03] E. Mossel and Y. Peres. Information flow on trees. *Ann. Appl. Probab.*, 13(3):817–844, 2003. MR-1994038

[Roy88] H. L. Royden. *Real analysis*. Macmillan Publishing Company, New York, third edition, 1988. MR-1013117

[Rud76] Walter Rudin. *Principles of mathematical analysis*. McGraw-Hill Book Co., New York-Auckland-Düsseldorf, third edition, 1976. International Series in Pure and Applied Mathematics. MR-0385023

[SS03] C. Semple and M. Steel. *Phylogenetics*, volume 22 of *Mathematics and its Applications series*. Oxford University Press, 2003. MR-2060009

[Sly09] Allan Sly. Reconstruction of random colourings. *Communications in Mathematical Physics*, 288(3):943–961, 2009. MR-2504861

[Sly11] Allan Sly. Reconstruction for the potts model. *Ann. Probab.*, 39(4):1365–1406, 07 2011. MR-2857243

[SS99] Michael A. Steel and László A. Székely. Inverting random functions. *Ann. Comb.*, 3(1):103–113, 1999. Combinatorics and biology (Los Alamos, NM, 1998). MR-1769697

[SS02] Michael A. Steel and László A. Székely. Inverting random functions. II. Explicit bounds for discrete maximum likelihood estimation, with applications. *SIAM J. Discrete Math.*, 15(4):562–575 (electronic), 2002. MR-1935839

[Ste16] Mike Steel. *Phylogeny—discrete and random processes in evolution*, volume 89 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2016. MR-3601108

[TKF91] Jeffrey L Thorne, Hirohisa Kishino, and Joseph Felsenstein. An evolutionary model for maximum likelihood alignment of dna sequences. *Journal of Molecular Evolution*, 33(2):114–124, 1991.

[Tho04] Joseph W Thornton. Resurrecting ancient genes: experimental analysis of extinct molecules. *Nature reviews. Genetics*, 5(5):366, 2004.

[Wid41] David Vernon Widder. *The Laplace Transform*. Princeton Mathematical Series, v. 6. Princeton University Press, Princeton, N. J., 1941. MR-0005923

A Identifiability of initial state on countable state spaces

We establish initial-state identifiability in two broad classes of chains.

Uniform chains Assume the uniform bound $\sup_{i\in\mathcal{S}} q_i < \infty$, where recall that q_i was defined in (1.2). That condition implies that Q is a bounded operator on the Banach space ℓ_1 and that $\mathbf{P}_t = \exp(tQ)$ for $t \in [0,\infty)$ (see e.g. [Ken67]). It then follows that $\mathbf{P}_t$ has an inverse $\mathbf{P}_t^{-1} = \exp(-tQ)$ and that

$$\inf_{x\neq 0} \frac{\|x\mathbf{P}_t\|_1}{\|x\|_1} \geq \frac{1}{\|\exp(-tQ)\|} \geq \exp(-t\|Q\|), \tag{A.1}$$

where $\|\cdot\|_1$ is the usual ℓ_1-norm and

$$\|Q\| = \sup_i \sum_j |q_{ij}| < +\infty,$$

is the operator norm of Q. The first inequality in (A.1) follows by putting $y = x\mathbf{P}_t$ and operator $A = \exp(-tQ)$ in the inequality $\|y\,A\|_1 \leq \|y\|_1\,\|A\|$. The second inequality follows from

$$\|\exp(-tQ)\| = \sum_{k\geq 0} \frac{\|(-tQ)^k\|}{k!} \leq \sum_{k\geq 0} \frac{t^k\|Q\|^k}{k!} = \exp(t\|Q\|).$$

Since $\mathbf{p}^i(t) = \vec{e}_i\mathbf{P}_t$, we have

$$\inf_{i\neq j} \|\mathbf{p}^i(t) - \mathbf{p}^j(t)\|_{\mathrm{TV}} \geq \exp(-t\|Q\|) > 0, \tag{A.2}$$

where we used that the total variation distance is half the ℓ_1 norm. This includes the finite state-space case; see [GS10, Lemma 5.1] for another proof in that case.

We are unaware of a proof that initial-state identifiability holds more generally in the unbounded case. In particular, arguing through the inverse as above may be difficult, as it is related to the longstanding Markov group conjecture. See [Ken67]. However we argue next that, in the special case of reversible chains, initial-state identifiability does hold in general.

Reversible chains Assume now that $(\mathbf{P}_t)_t$ is reversible (or weakly symmetric) with respect to the positive measure μ on $\mathcal{S}$. By Kendall's representation (see e.g. [And91, Theorem 1.6.7]), for each pair $i, j \in \mathcal{S}$, there is a finite signed measure ϕ_{ij} on $[0,\infty)$ such that

$$p_{ij}(t) = \sqrt{\frac{\mu_j}{\mu_i}} \int_{[0,\infty)} e^{-tx} d\phi_{ij}(x), \qquad \forall t \geq 0.$$

By Jordan decomposition, ϕ_{ij} is the difference of two finite non-negative measures (see e.g. [Roy88, Chapter 11]) so that $p_{ij}(t)$ can be seen as the difference of two Laplace transforms of non-negative measures. The latter are absolutely convergent, and therefore, analytic on $(0,\infty)$ (see e.g. [Wid41, Chapter II]). Hence all $p_{ij}(t)$s are analytic. If two analytic functions agree on a set with a limit point, then they agree everywhere (see e.g. [Rud76]). Suppose there exists $t_0 > 0$ such that $p_{ij}(t_0) = p_{kj}(t_0)$ for all j. Then by the Chapman-Kolmogorov equations we have

$$p_{ij}(t) = p_{kj}(t), \qquad \forall t \geq t_0, j \in \mathcal{S}. \tag{A.3}$$

Then the same holds for all $t > 0$ and, by continuity at 0, we must have $p_{ii}(0) = p_{ki}(0)$ which implies $i = k$.

B An application: the TKF91 process

In this section, we apply Theorem 1.6 to ancestral sequence reconstruction in a DNA model accounting for nucleotide insertion and deletion known as the TKF91 process. We first describe the Markovian dynamics. Conforming with the original definition of the model [TKF91], we use an "immortal link" as a stand-in for the empty sequence.

Definition B.1 (TKF91 sequence evolution model on an edge)**.** *The* **TKF91 edge process** *is a Markov process $\mathcal{I} = (\mathcal{I}_t)_{t\geq 0}$ on the space $\mathcal{S}$ of DNA sequences together with an* **immortal link** *"•", that is,*

$$\mathcal{S} := \text{“}\bullet\text{”} \otimes \bigcup_{M\geq 0} \{A, T, C, G\}^M, \tag{B.1}$$

where the notation above indicates that all sequences begin with the immortal link (and can otherwise be empty). We also refer to the positions of a sequence (including nucleotides and the immortal link) as **sites***. Let* $(\nu, \lambda, \mu) \in (0,\infty)^3$ *with* $\lambda < \mu$ *and* $(\pi_A, \pi_T, \pi_C, \pi_G) \in [0,\infty)^4$ *with* $\pi_A + \pi_T + \pi_C + \pi_G = 1$ *be given parameters. The continuous-time Markovian dynamic is described as follows: if the current state is the sequence* $\vec{x}$*, then the following events occur independently:*

- *(Substitution) Each nucleotide (but not the immortal link) is substituted independently at rate* $\nu > 0$*. When a substitution occurs, the corresponding nucleotide is replaced by* A, T, C *and* G *with probabilities* π_A, π_T, π_C *and* π_G *respectively.*
- *(Deletion) Each nucleotide (but not the immortal link) is removed independently at rate* $\mu > 0$*.*
- *(Insertion) Each site gives birth to a new nucleotide independently at rate* $\lambda > 0$*. When a birth occurs, a nucleotide is added immediately to the right of its parent site. The newborn site has nucleotide* A, T, C *and* G *with probabilities* π_A, π_T, π_C *and* π_G *respectively.*

The **length** *of a sequence* $\vec{x} = (\bullet, x_1, x_2, \cdots, x_M)$ *is defined as the number of nucleotides in* $\vec{x}$ *and is denoted by* $|\vec{x}| = M$ *(with the immortal link alone corresponding to* $M = 0$*). When* $M \geq 1$ *we omit the immortal link for simplicity and write* $\vec{x} = (x_1, x_2, \cdots, x_M)$*.*

The TKF91 edge process is reversible [TKF91]. Suppose furthermore that

$$0 < \lambda < \mu,$$

an assumption we make throughout. Then it has an **stationary distribution** Π, given by

$$\Pi(\vec{x}) = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^M \prod_{i=1}^{M} \pi_{x_i}$$

for each $\vec{x} = (x_1, x_2, \cdots, x_M) \in \{A, T, C, G\}^M$ where $M \geq 1$, and $\Pi(\text{“}\bullet\text{”}) = \left(1 - \frac{\lambda}{\mu}\right)$. In words, under Π, the sequence length is geometrically distributed and, conditioned on the sequence length, all sites are independent with distribution $(\pi_\sigma)_{\sigma \in \{A,T,C,G\}}$. Hence, from the argument in Section A, initial-state identifiability holds for the TKF91 edge process. Theorem 1.6 gives:

Theorem B.2 (TKF91 process: consistent root estimation)**.** *Let* $\{T^k\}_k$ *satisfy assumption (i) and the big bang condition. Let* $(\mathbf{P}_t)_t$ *be the TKF91 edge process with* $\lambda < \mu$ *and let* π *be the stationary distribution of the process. Then there exists a sequence of consistent root estimators.*

In a companion paper [FR], we give an alternative consistent root estimator that is also computationally efficient and provide error bounds that are explicit in the parameters of the model.