7-2013

Robust Estimation of Latent Tree Graphical Models: Inferring Hidden States With Inexact Parameters

Elchanan Mossel
University of Pennsylvania

Sébastien Roch

Allan Sly

Robust Estimation of Latent Tree Graphical Models: Inferring Hidden States With Inexact Parameters

Abstract
Latent tree graphical models are widely used in computational biology, signal and image processing, and network tomography. Here, we design a new efficient, estimation procedure for latent tree models, including Gaussian and discrete, reversible models, that significantly improves on previous sample requirement bounds. Our techniques are based on a new hidden state estimator that is robust to inaccuracies in estimated parameters. More precisely, we prove that latent tree models can be estimated with high probability in the so-called Kesten-Stigum regime with $O(\log^2 n)$ samples, where n is the number of nodes.

Keywords
Gaussian processes, spectral analysis, state estimation, trees (mathematics), Gaussian models, Kesten-Stigum regime, computational biology, discrete models, hidden state estimator, image processing, latent tree graphical models, network tomography, reversible models, robust estimation, signal processing, biological system modeling, eigenvalues and eigenfunctions, estimation, graphical models, Markov processes, measurement, vegetation, Gaussian graphical models on trees, Kesten-Stigum (KS) reconstruction bound, Markov random fields on trees, phase transitions

Disciplines
Biology | Computer Sciences | Statistics and Probability

Robust estimation of latent tree graphical models: Inferring hidden states with inexact parameters

Elchanan Mossel* Sébastien Roch† Allan Sly‡

September 23, 2011

Abstract

Latent tree graphical models are widely used in computational biology, signal and image processing, and network tomography. Here we design a new efficient, estimation procedure for latent tree models, including Gaussian and discrete, reversible models, that significantly improves on previous sample requirement bounds. Our techniques are based on a new hidden state estimator which is robust to inaccuracies in estimated parameters. More precisely, we prove that latent tree models can be estimated with high probability in the so-called Kesten-Stigum regime with $O(\log^2 n)$ samples.

Keywords: Gaussian graphical models on trees, Markov random fields on trees, phase transitions, Kesten-Stigum reconstruction bound

*Weizmann Institute and UC Berkeley.
†UCLA. Work supported by NSF grant DMS-1007144.
‡UC Berkeley.

1 Introduction

Background Latent tree graphical models and other related models have been widely studied in mathematical statistics, machine learning, signal and image processing, network tomography, computational biology, and statistical physics. See e.g. [And58, KF09, Wil02, CCL$^+$04, SS03, EKPS00] and references therein. For instance, in phylogenetics [Fel04], one seeks to reconstruct the evolutionary history of living organisms from molecular data extracted from modern species. The assumption is that molecular data consists of aligned sequences and that each position in the sequences evolves independently according to a Markov random field on a tree, where the key parameters are (see Section 1.1 for formal definitions):

- *Tree.* An evolutionary tree T, where the leaves are the modern species and each branching represents a past speciation event.
- *Rate matrix.* A $q \times q$ mutation rate matrix Q, where q is the alphabet size. A typical alphabet arising in biology would be $\{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$. Without loss of generality, here we denote the alphabet by $[q] = \{1, \ldots, q\}$. The (i, j)'th entry of Q encodes the rate at which state i mutates into state j. We normalize the matrix Q so that its spectral gap is 1.
- *Edge weights.* For each edge e, we have a scalar branch length τ_e which measures the total amount of evolution along edge e. (We use edge or branch interchangeably.) Roughly speaking, τ_e is the time elapsed between the end points of e. (In fact the time is multiplied by an edge-dependent overall mutation rate because of our normalization of Q.) We also think of τ_e as the "evolutionary distance" between the end points of e.

Other applications, including those involving Gaussian models (see Section 1.1), are similarly defined. Two statistical problems naturally arise in this context:

- *Tree Model Estimation (TME).* Given k samples of the above process at the observed nodes, that is, at the leaves of the tree, estimate the topology of the tree as well as the edge weights.
- *Hidden State Inference (HSI).* Given a fully specified tree model and a single sample at the observed nodes, infer the state at the (unobserved) root of the tree.

In recent years, a convergence of techniques from statistical physics and theoretical computer science has provided fruitful new insights on the deep connections between these two problems, starting with [Mos04].

Steel's Conjecture A crucial parameter in the second problem above is $\tau^+(T) = \max_e \tau_e$, the maximal edge weight in the tree. For instance, for the two-state symmetric Q also known as the Ising model, it is known that there exists a critical parameter $g^\star_{\mathrm{KS}} = \ln \sqrt{2}$ such that, if $\tau^+(T) < g^\star_{\mathrm{KS}}$, then it is possible to perform HSI (better than random; see the Section 2.5 for additional details). In contrast, if $\tau^+(T) \geq g^\star_{\mathrm{KS}}$, there exist trees for which HSI is impossible, that is, the correlation between the best root estimate and its true value decays exponentially in the depth of the tree. The regime $\tau^+(T) < g^\star_{\mathrm{KS}}$ is known as the Kesten-Stigum (KS) regime [KS66].

A striking and insightful conjecture of Steel postulates a deep connection between TME and HSI [Ste01]. More specifically the conjecture states that for the Ising model, in the KS regime, high-probability TME may be achieved with a number of samples $k = O(\log n)$. Since the number of trees on n labelled leaves is $2^{\Theta(n \log n)}$, this is an optimal sample requirement up to constant factors. The proof of Steel's conjecture was established in [Mos04] for the Ising model on balanced trees and in [DMR11a] for rate matrices on trees with discrete edge lengths. Furthermore, results of Mossel [Mos03, Mos04] show that for $\tau^+(T) \geq g^\star_{\mathrm{KS}}$ a polynomial sample requirement is needed for correct TME, a requirement achieved by several estimation algorithms [ESSW99a, Mos04, Mos07, GMS08, DMR11b]. The previous results have been extended to general reversible Q on alphabets of size $q \geq 2$ [Roc10, MRS11]. (Note that in that case a more general threshold $g^\star_Q$ may be defined, although little rigorous work has been dedicated to its study. See [Mos01, Sly09, MRS11]. In this paper we consider only the KS regime.)

Our contributions Prior results for general trees and general rate matrix Q, when $\tau^+(T) < g^\star_{\mathrm{KS}}$, have assumed that edge weights are discretized. This assumption is required to avoid dealing with the sensitivity of root-state inference to inexact (that is, estimated) parameters. Here we design a new HSI procedure in the KS regime which is provably robust to inaccuracies in the parameters (and, in particular, does not rely on the discretization assumption). More precisely, we prove that $O(\log^2 n)$ samples suffice to solve the TME and HSI problems in the KS regime without discretization. We consider two models in detail: discrete, reversible Markov random fields (also known as GTR models in evolutionary biology), and Gaussian models. As far as we know, Gaussian models have not previously been studied in the context of the HSI phase transition. (We derive the critical threshold for Gaussian models in Section 2.5.) Formal statements of our results can be found in Section 1.2. Section 1.3 provides a sketch of the proof.

Further related work For further related work on sample requirements in tree graphical model estimation, see [ESSW99b, MR06, TAW10, TATW11, CTAW11, TAW11, BRR10].

1.1 Definitions

Trees and metrics. Let $T = (V, E)$ be a tree with leaf set $[n]$, where $[n] = \{1, \ldots, n\}$. For two leaves $a, b \in [n]$, we denote by $\mathrm{P}(a, b)$ the set of edges on the unique path between a and b. For a node $v \in V$, let $\mathrm{N}(v)$ be the neighbors of v.

Definition 1 (Tree Metric) *A* tree metric *on $[n]$ is a positive function $\mathrm{D} : [n] \times [n] \to (0, +\infty)$ such that there exists a tree $T = (V, E)$ with leaf set $[n]$ and an edge weight function $w : E \to (0, +\infty)$ satisfying the following: for all leaves $a, b \in [n]$*

$$\mathrm{D}(a, b) = \sum_{e \in \mathrm{P}(a,b)} w_e.$$

In this work, we consider dyadic trees. Our techniques can be extended to complete trees of higher degree. We discuss general trees in the concluding remarks.

Definition 2 (Balanced tree) *A* balanced tree *is a rooted, edge-weighted, leaf-labeled h-level dyadic tree $\mathcal{T} = (V, E, [n], r; \tau)$ where: $h \geq 0$ is an integer; V is the set of vertices; E is the set of edges; $L = [n] = \{1, \ldots, n\}$ is the set of leaves with $n = 2^h$; r is the root; $\tau : E \to (0, +\infty)$ is a positive edge weight function. We denote by $(\tau(a, b))_{a,b \in [n]}$ the tree metric corresponding to the balanced tree $\mathcal{T} = (V, E, [n], r; \tau)$. We extend $\tau(u, v)$ to all vertices $u, v \in V$. We let $\mathbb{BY}_n$ be the set of all such balanced trees on n leaves and we let $\mathbb{BY} = \{\mathbb{BY}_{2^h}\}_{h \geq 0}$.*

Markov random fields on trees. We consider Markov models on trees where only the leaf variables are observed. The following discrete-state model is standard in evolutionary biology. See e.g. [SS03]. Let $q \geq 2$. Let $[q]$ be a state set and π be a distribution on $[q]$ satisfying $\pi_x > 0$ for all $x \in [q]$. The $q \times q$ matrix Q is a *rate matrix* if $Q_{xy} > 0$ for all $x \neq y$ and $\sum_{y \in [q]} Q_{xy} = 0$, for all $x \in [q]$. The rate matrix Q is *reversible with respect to* π if $\pi_x Q_{xy} = \pi_y Q_{yx}$, for all $x, y \in [q]$. By reversibility, Q has q real eigenvalues $0 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_q$. We normalize Q by fixing $\lambda_2 = -1$. We denote by $\mathbb{Q}_q$ the set of all such rate matrices.

Definition 3 (General Time-Reversible (GTR) Model) *For $n \geq 1$, let*

$$\mathcal{T} = (V, E, [n], r; \tau)$$

be a balanced tree. Let Q be a $q \times q$ rate matrix reversible with respect to π. Define the transition matrices $M^e = e^{\tau_e Q}$, for all $e \in E$. The GTR model on $\mathcal{T}$ with rate matrix Q associates a state Z_v in $[q]$ to each vertex v in V as follows: pick a state for the root r according to π; moving away from the root, choose a state for each vertex v independently according to the distribution $(M^e_{Z_u,j})_{j\in[q]}$, with $e = (u, v)$ where u is the parent of v. We let $\mathbb{GTR}_{n,q}$ be the set of all q-state GTR models on n leaves. We denote $\mathbb{GTR}_q = \left\{\mathbb{GTR}_{2^h,q}\right\}_{h\geq 0}$. We denote by Z_W the vector of states on the vertices $W \subseteq V$. In particular, $Z_{[n]}$ are the states at the leaves. We denote by $\mathcal{D}_{\mathcal{T},Q}$ the distribution of $Z_{[n]}$.

GTR models encompass several special cases such as the Cavender-Farris-Neyman (CFN) model and the Jukes-Cantor (JC) model.

Example 1 (q-state Symmetric Model) *The* q-state Symmetric model *(also called q-state Potts model) is the GTR model with $q \geq 2$ states, $\pi = (1/q, \ldots, 1/q)$, and $Q = Q^{q-\mathrm{POTTS}}$ where*

$$Q^{q-\mathrm{POTTS}}_{ij} = \begin{cases} -\frac{q-1}{q} & \text{if } i = j \\ \frac{1}{q} & \text{o.w.} \end{cases}$$

Note that $\lambda_2(Q) = -1$. The special cases $q = 2$ and $q = 4$ are called respectively the CFN and JC models in the biology literature. We denote their rate matrices by $Q^{\mathrm{CFN}}, Q^{\mathrm{JC}}$.

A natural generalization of the CFN model which is also included in the GTR framework is the Binary Asymmetric Channel.

Example 2 (Binary Asymmetric Channel) *Letting $q = 2$ and $\pi = (\pi_1, \pi_2)$, with $\pi_1, \pi_2 > 0$, we can take*

$$Q = \begin{pmatrix} -\pi_2 & \pi_2 \\ \pi_1 & -\pi_1 \end{pmatrix}.$$

The following transformation will be useful [MP03]. Let ν be a right eigenvector of the GTR matrix Q corresponding to the eigenvalue -1. Map the state space to the real line by defining $X_x = \nu_{Z_x}$ for all $x \in [n]$.

We also consider Gaussian Markov Random Fields on Trees (GMRFT). Gaussian graphical models, including Gaussian tree models, are common in statistics, machine learning as well as signal and image processing. See e.g. [And58, Wil02].

Definition 4 (Gaussian Markov Random Field on a Tree (GMRFT)) *For $n \geq 1$, let $\mathcal{T} = (V, E, [n], r; \tau)$ be a balanced tree. A GMRFT on $\mathcal{T}$ is a multivariate Gaussian vector $X_V = (X_v)_{v \in V}$ whose covariance matrix $\Sigma = (\Sigma_{uv})_{u,v \in V}$ with inverse $\Lambda = \Sigma^{-1}$ satisfies*

$$(u, v) \notin E, u \neq v \implies \Lambda_{uv} = 0.$$

We assume that only the states at the leaves $X_{[n]}$ are observed. To ensure identifiability (that is, to ensure that two different sets of parameters generate different distributions at the leaves), we assume that all internal nodes have zero mean and unit variance and that all non-leaf edges correspond to a nonnegative correlation. Indeed shifting and scaling the states at the internal nodes does not affect the leaf distribution. For convenience, we extend this assumption to leaves and leaf edges. With the choice

$$\Sigma_{uv} = \prod_{e \in \mathrm{P}(u,v)} \rho_e, \quad u, v \in V,$$

where $\rho_e = e^{-\tau_e}$, for all $e \in E$, a direct calculation shows that

$$\Lambda_{uv} = \begin{cases} 1 + \sum_{w \in \mathrm{N}(v)} \frac{\rho^2_{(v,w)}}{1-\rho^2_{(v,w)}}, & \text{if } u = v, \\ -\frac{\rho_{(u,v)}}{1-\rho^2_{(u,v)}}, & \text{if } (u, v) \in E, \\ 0, & \text{o.w.} \end{cases}$$

(Note that, in computing $(\Sigma\Lambda)_{uv}$ with $u \neq v$, the product $\prod_{e \in \mathrm{P}(u,w)} \rho_e$ factors out, where $w \in \mathrm{N}(v)$ with $(w, v) \in \mathrm{P}(u, v)$.) In particular, $\{-\log |\Sigma_{uv}|\}_{uv \in [n]}$ is a tree metric. We denote by $\mathcal{D}_{\mathcal{T},\Sigma}$ the distribution of $X_{[n]}$. We let $\mathbb{GMRFT}_n$ be the set of all GMRFT models on n leaves. We denote $\mathbb{GMRFT} = \{\mathbb{GMRFT}_{2^h}\}_{h \geq 0}$.

Remark 1 *Our techniques extend to cases where leaves and leaf edges have general means and covariances. We leave the details to the reader.*

Equivalently, in a formulation closer to that of the GTR model above, one can think of a GMRFT model as picking a root value according to a standard Gaussian distribution and running independent Ornstein-Uhlenbeck processes on the edges.

Both the GTR and GMRFT models are *globally Markov*: for all disjoint subsets A, B, C of V such that B *separates* A and C, that is, all paths between A and C go through a node in B, we have that the states at A are conditionally independent of the states at C given the states at B.

1.2 Results

Our main results are the following. We are given k i.i.d. samples from a GMRFT or GTR model and we seek to estimate the tree structure with failure probability going to 0 as the number of leaves n goes to infinity. We also estimate edge weights within constant tolerance.

Theorem 1 (Main Result: GMRFT Models) *Let $0 < f < g < +\infty$ and denote by $\mathbb{GMRFT}^{f,g}$ the set of all GMRFT models on balanced trees $\mathcal{T} = (V, E, [n], r; \tau)$ satisfying $f < \tau_e < g,\ \forall e \in E$. Then, for all $0 < f < g < g^{\star}_{\mathrm{KS}} = \ln\sqrt{2}$, the tree structure estimation problem on $\mathbb{GMRFT}^{f,g}$ can be solved with $k = \kappa \log^2 n$ samples, where $\kappa = \kappa(f, g) > 0$ is large enough. Moreover all edge weights are estimated within constant tolerance.*

This result is sharp as we prove the following negative results establishing the equivalence of the TME and HSI thresholds.

Theorem 2 *If $0 < f \leq g$ with $g > g^{\star}_{\mathrm{KS}} = \ln\sqrt{2}$, then the tree structure estimation problem on $\mathbb{GMRFT}^{f,g}$ cannot, in general, be solved without at least $k = n^{\gamma}$ samples, where $\gamma = \gamma(f, g) > 0$.*

The proof of the theorem is in Section 2.

Theorem 3 (Main Result: GTR Models) *Let $0 < f < g < +\infty$ and denote by $\mathbb{GTR}_q^{f,g}$ the set of all q-state GTR models on balanced trees $\mathcal{T} = (V, E, [n], r; \tau)$ satisfying $f < \tau_e < g,\ \forall e \in E$. Then, for all $q \geq 2$, $0 < f < g < g^{\star}_{\mathrm{KS}} = \ln\sqrt{2}$, the tree structure estimation problem on $\mathbb{GTR}_q^{f,g}$ can be solved with $k = \kappa \log^2 n$ samples, where $\kappa = \kappa(q, f, g) > 0$ is large enough. Moreover all edge weights are estimated within constant tolerance.*

The proof of this theorem is similar to that of Theorem 1. However dealing with unknown rate matrices requires some care and the full proof of the modified algorithm in that case can be found in Section 3.

Remark 2 *Our techniques extend to d-ary trees for general (constant) $d \geq 2$. In that case, the critical threshold satisfies $de^{-2\tau} = 1$. We leave the details to the reader.*

1.3 Proof Overview

We give a sketch of the proof of our main result. We discuss the case of GTR models with known Q matrix. The unknown Q matrix and Gaussian cases are similar. See Sections 2 and 3 for details. Let $(Z^i_{[n]})_{i=1}^k$ be i.i.d. samples from a GTR model on a balanced tree with n leaves. Let (Z_V) be a generic sample from the GTR model.

Boosted algorithm As a starting point, our algorithm uses the reconstruction framework of [Mos04]. This basic "boosting" approach is twofold:

- *Initial Step.* Build the first level of the tree from the samples at the leaves. This can be done easily by standard quartet-based techniques. (See Section 2.2.)
- *Main Loop.* Repeat the following two steps until the tree is built:
  1. *HSI.* Infer hidden states at the roots of the reconstructed subtrees.
  2. *One-level TME.* Use the hidden state estimates from the previous step to build the next level of the tree using quartet-based techniques.

The heart of the procedure is Step 1. Note that, assuming each level is correctly reconstructed, the HSI problem in Step 1 is performed on a known, correct topology. However the edge weights are unknown and need to be estimated from the samples at the leaves.

This leads to the key technical issue addressed in this paper. Although HSI with known topology and edge weights is well understood (at least in the so-called Kesten-Stigum (KS) regime [MP03]), little work has considered the effect of inexact parameters on hidden state estimation, with the notable exception of [Mos04] where a parameter-free estimator is developed for the Ising model. The issue was averted in prior work on GTR models by assuming that edge weights are discretized, allowing exact estimation [DMR11a, Roc10].

Quartet-based tree structure and edge weight estimation relies on the following distance estimator. It is natural to use a distance estimator involving the eigenvectors of Q. Let ν be a second right eigenvector of the GTR matrix Q corresponding to the eigenvalue -1. For $a \in V$ and $i = 1, \ldots, k$, map the samples to the real line by defining $X^i_a = \nu_{Z^i_a}$. Then define

$$\hat{\tau}(a,b) = -\ln\left(\frac{1}{k}\sum_{i=1}^{k} X^i_a X^i_b\right). \tag{1}$$

It can be shown that: For all $a, b \in V$, we have $-\ln \mathbb{E}[e^{-\hat{\tau}(a,b)}] = \tau(a,b)$. Note that, in our case, this estimate is only available for pairs of *leaves*. Moreover, it is known that the quality of this estimate degrades quickly as $\tau(a,b)$ increases [ESSW99a, Att99]. To obtain accuracy ε on a τ distance with inverse polynomial failure probability requires

$$k \geq C_1 \varepsilon^{-2} e^{C_2 \tau} \log n \tag{2}$$

samples, where C_1, C_2 are constants. We use HSI to replace the X's in (1) with approximations of hidden states in order to improve the accuracy of the distance estimator between *internal* nodes.

Weighted majority For the symmetric CFN model with state space $\{+1, -1\}$, hidden states can be inferred using a linear combination of the states at the leaves—a type of weighted majority vote. A natural generalization of this linear estimator in the context of more general mutation matrices was studied by [MP03]. The estimator at the root r considered in [MP03] is of the form

$$S_r = \sum_{x \in [n]} \left(\frac{\Psi(x)}{e^{-\tau(r,x)}} \right) X_x, \tag{3}$$

where Ψ is a unit flow between r and $[n]$. For any such Ψ, S_r is a conditionally unbiased estimator of X_r, that is, $\mathbb{E}[S_r \mid X_r] = X_r$. Moreover, in the KS regime, that is, when $\tau^+ < g^\star_{\mathrm{KS}}$, one can choose a flow such that the variance of S_r is uniformly bounded [MP03] and, in fact, we have the following stronger moment condition

$$\mathbb{E}[\exp(\zeta S_r) | X_r] \leq \exp(\zeta X_r + c\zeta^2)$$

for all $\zeta \in \mathbb{R}$ [PR11]. In [Roc10] this estimator was used in Step 1 of the boosted algorithm. On a balanced tree with $\log n$ levels, obtaining sufficiently accurate estimates of the coefficients in (3) requires accuracy $1/\Omega(\log(n))$ on the edge weights. By (2), such accuracy requires a $O(\log^3 n)$ sequence length. Using misspecified edge weights in (3) may lead to a highly biased estimate and generally may fail to give a good reconstruction at the root. Here we achieve accurate hidden state estimation using only $O(\log^2 n)$ samples.

Recursive estimator We propose to construct an estimator of the form (3) *recursively*. For $x \in V$ with children y_1, y_2 we let

$$S_v = \omega_{y_1} S_{y_1} + \omega_{y_2} S_{y_2}, \tag{4}$$

and choose the coefficients $\omega_{y_1}, \omega_{y_2}$ to guarantee the following conditions:

- We have
$$\mathbb{E}[S_x \mid Z_x] = \mathcal{B}(x) X_x,$$
with a bias term $\mathcal{B}(x)$ close to 1.

- The estimator satisfies the exponential moment condition
$$\mathbb{E}[\exp(\zeta S_x) | Z_x] \leq \exp(\zeta X_x + c\zeta^2).$$

We show that these conditions can be guaranteed provided the model is in the KS regime. To do so, the procedure measures the bias terms $\mathcal{B}(y_1)$ and $\mathcal{B}(y_2)$ using methods similar to distance estimation. By testing the bias and, if necessary, compensating for any previously introduced error, we can adaptively choose coefficients ω_1, ω_2 so that S_x satisfies these two conditions.

Unknown rate matrix Further complications arise when the matrix Q is not given and has to be estimated from the data. We give a procedure for recovering Q and an estimate of its second right eigenvector. Problematically, any estimate $\hat{\nu}$ of ν may have a small component in the direction of the first right eigenvector of Q. Since the latter has eigenvalue 0, its component builds up over many recursions and it eventually overwhelms the signal. However, we make use of the fact that the first right eigenvector is identically 1: by subtracting from S_x its empirical mean, we show that we can cancel the effect of the first eigenvector. With a careful analysis, this improved procedure leads to an accurate estimator.

2 Gaussian Model

In this section, we prove our main theorem in the Gaussian case. The proof is based on a new hidden state estimator which is described in Section 2.1. For $n = 2^h$ with$h \geq 0$, let $\mathcal{T} = (V, E, [n], r; \tau)$ be a balanced tree. We assume that $0 \leq \tau_e < g$, $\forall e \in E$, with $0 < g < g^\star_{\mathrm{KS}} = \ln \sqrt{2}$. The significance of the threshold $g^\star_{\mathrm{KS}}$ is explained in Section 2.5 where we also prove Theorem 2. We generate k i.i.d. samples $(X^i_{[n]})_{i=1}^k$ from the GMRFT model $\mathcal{D}_{\mathcal{T},\Sigma}$ where $k = \kappa \log^2 n$.

Our construction is recursive, building the tree and estimating hidden states one level at a time. To avoid unwanted correlations, we use a fresh block of samples for each level. Let $K = \kappa \log n$ be the size of each block.

2.1 Recursive Linear Estimator

The main tool in our reconstruction algorithm is a new hidden state estimator. This estimator is recursive, that is, for a node $x \in V$ it is constructed from estimators for its children y, z. In this subsection, we let X_V be a generic sample from the GMRFT independent of everything else. We let $(X^i_{[n]})_{i=1}^K$ be a block of independent samples at the leaves. For a node $u \in V$, we let $\lfloor u \rfloor$ be the leaves below u and $X_{\lfloor u \rfloor}$, the corresponding state.

Linear estimator We build a *linear* estimator for each of the vertices recursively from the leaves. Let $x \in V - [n]$ with children (direct descendants) y_1, y_2. Assume that the topology of the tree rooted at x has been correctly reconstructed, as detailed in Section 2.2. Assume further that we have constructed linear estimators

$$S_u \equiv \mathcal{L}_u(X_{\lfloor u \rfloor})$$

of X_u, for all $u \in V$ below x. We use the convention that $\mathcal{L}_u(X_{\lfloor u \rfloor}) = X_u$ if u is a leaf. We let $\mathcal{L}_x$ be a linear combination of the form

$$S_x \equiv \mathcal{L}_x(X_{\lfloor x \rfloor}) = \omega_{y_1} \mathcal{L}_{y_1}(X_{\lfloor y_1 \rfloor}) + \omega_{y_2} \mathcal{L}_{y_2}(X_{\lfloor y_2 \rfloor}), \tag{5}$$

where—ideally—the ω's are chosen so as to satisfy the following conditions:

1. **Unbiasedness.** The estimator $S_x = \mathcal{L}_x(X_{\lfloor x \rfloor})$ is *conditionally unbiased*, that is,
$$\mathbb{E}[S_x \,|\, X_x] = X_x.$$
2. **Minimum Variance.** The estimator has minimum variance amongst all estimators of the form (5).

An estimator with these properties can be constructed given exact knowledge of the edge parameters, see Section 2.5. However, since the edge parameters can only be estimated with constant accuracy given the samples, we need a procedure that satisfies these conditions only approximately. We achieve this by 1) recursively minimizing the variance at each level and 2) at the same time measuring the bias and adjusting for any deviation that may have accumulated from previously estimated branch lengths.

Setup We describe the basic recursive step of our construction. As above, let $x \in V - [n]$ with children y_1, y_2 and corresponding edges $e_1 = (x, y_1), e_2 = (x, y_2)$. Let $0 < \delta < 1$ (small) and $c > 1$ (big) be constants to be defined later. Assume that we have the following:

- Estimated edge weights $\hat{\tau}_e$ for all edges e below x such that there is $\varepsilon > 0$ with
$$|\hat{\tau}_e - \tau_e| < \varepsilon. \tag{6}$$
The choice of ε and the procedure to obtain these estimates are described in Section 2.3. We let $\hat{\rho}_e = e^{-\hat{\tau}_e}$.

- Linear estimators $\mathcal{L}_u$ for all $u \in V$ below x such that with
$$\mathbb{E}[S_u \mid X_u] = \mathcal{B}(u) X_u, \tag{7}$$
where $S_u \equiv \mathcal{L}_u(X_{\lfloor u \rfloor})$, for some $\mathcal{B}(u) > 0$ with $|\mathcal{B}(u) - 1| < \delta$ and
$$\mathcal{V}(u) \equiv \mathrm{Var}[S_u] \leq c. \tag{8}$$
Note that these conditions are satisfied at the leaves. Indeed, for $u \in [n]$ one has $S_u = X_u$ and therefore $\mathbb{E}[S_u \mid X_u] = X_u$ and $\mathcal{V}(u) = \mathrm{Var}[X_u] = 1$. We denote $\beta(u) = -\ln \mathcal{B}(u)$.

We now seek to construct S_x so that it in turn satisfies the same conditions.

Remark 3 *In this subsection, we are treating the estimated edge weights and linear estimator coefficients as deterministic. In fact, they are random variables depending on sample blocks used on prior recurrence levels—and in particular they are independent of X_V and of the block of samples used on the current level.*

Procedure Given the previous setup, we choose the weights ω_{y_α}, $\alpha = 1, 2$, as follows. For $u, v \in V$ below x and $\ell = 1, \ldots, K$ let
$$S_u^\ell \equiv \mathcal{L}_u(X_{\lfloor u \rfloor}^\ell),$$
and define
$$\ddot{\tau}(u, v) = -\ln \left(\frac{1}{K} \sum_{\ell=1}^{K} S_u^\ell S_v^\ell \right),$$
the estimated path length between u and v including bias. We let $\beta(u) = -\ln \mathcal{B}(u)$.

1. **Estimating the Biases.** If y_1, y_2 are leaves, we let $\widehat{\beta}(y_\alpha) = 0$, $\alpha = 1, 2$. Otherwise, let z_{21}, z_{22} be the children of y_2. We then compute

$$\widehat{\beta}(y_1) = \frac{1}{2}(\ddot{\tau}(y_1, z_{21}) + \ddot{\tau}(y_1, z_{22}) - \ddot{\tau}(z_{21}, z_{22}) - 2\hat{\tau}_{e_1} - 2\hat{\tau}e_2),$$

and similarly for y_2. Let $\widehat{\mathcal{B}}(y_\alpha) = e^{-\widehat{\beta}(y_\alpha)}$, $\alpha = 1, 2$.

2. **Minimizing the Variance.** For $\alpha = 1, 2$ we set $\omega_{y_1}, \omega_{y_2}$ as

$$\omega_{y_\alpha} = \frac{\widehat{\mathcal{B}}(y_\alpha)\hat{\rho}_{e_\alpha}}{\widehat{\mathcal{B}}(y_1)^2\hat{\rho}_{e_1}^2 + \widehat{\mathcal{B}}(y_2)^2\hat{\rho}_{e_2}^2}, \tag{9}$$

which corresponds to the solution of the following optimization problem:

$$\min\{\omega_{y_1}^2 + \omega_{y_2}^2 \ : \ \omega_{y_1}\widehat{\mathcal{B}}(y_1)\hat{\rho}_{e_1} + \omega_{y_2}\widehat{\mathcal{B}}(y_2)\hat{\rho}_{e_2} = 1, \ \omega_{y_1}, \omega_{y_2} > 0\}. \tag{10}$$

The constraint in the optimization above is meant to ensure that the bias condition (7) is satisfied. We set

$$\mathcal{L}_x(X_{\lfloor x \rfloor}) = \omega_{y_1}\mathcal{L}_{y_1}(X_{\lfloor y_1 \rfloor}) + \omega_{y_2}\mathcal{L}_{y_2}(X_{\lfloor y_2 \rfloor}).$$

Bias and Variance We now prove (7) and (8) recursively assuming (6) is satisfied. This follows from the following propositions.

Proposition 1 (Concentration of Internal Distance Estimates) *For all $\varepsilon > 0$, $\gamma > 0$, $0 < \delta < 1$ and $c > 0$, there is $\kappa = \kappa(\varepsilon, \gamma, \delta, c) > 0$ such that, with probability at least $1 - O(n^{-\gamma})$, we have*

$$|\ddot{\tau}(u, v) - (\tau(u, v) + \beta(u) + \beta(v))| < \varepsilon,$$

for all $u, v \in \{y_1, y_2, z_{11}, z_{12}, z_{21}, z_{22}\}$ where $z_{\alpha 1}, z_{\alpha 2}$ are the children of y_α.

Proof: First note that

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{K}\sum_{\ell=1}^{K} S_u^\ell S_v^\ell\right] &= \mathbb{E}\left[S_u S_v\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[S_u S_v | X_u, X_v\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[S_u | X_u\right]\mathbb{E}\left[S_v | X_v\right]\right] \\
&= \mathbb{E}\left[\mathcal{B}(u)\mathcal{B}(v)X_u X_v\right] \\
&= \mathcal{B}(u)\mathcal{B}(v)\Sigma_{uv},
\end{aligned}$$

where we used the Markov property on the third line, so that

$$-\ln\left(\mathbb{E}\left[\frac{1}{K}\sum_{\ell=1}^{K} S_u^\ell S_v^\ell\right]\right) = \tau(u,v) + \beta(u) + \beta(v).$$

Moreover, by assumption, S_u is Gaussian with

$$\mathbb{E}[S_u] = 0, \quad \mathrm{Var}[S_u] = \mathcal{V}(u) \le c,$$

and similarly for u. It is well-known that in the Gaussian case empirical covariance estimates as above have χ^2-type distributions [And58]. Explicitly, note that from

$$S_u S_v = \frac{1}{2}[(S_u + S_v)^2 - S_u^2 - S_v^2],$$

it suffices to consider the concentration of S_u^2, S_v^2, and $(S_u + S_v)^2$. Note that

$$\mathrm{Var}[S_u + S_v] = \mathcal{V}(u) + \mathcal{V}(v) + 2\mathcal{B}(u)\mathcal{B}(v)\Sigma_{uv} \le 2c + 2(1+\delta)^2 < +\infty,$$

independently of n. We argue about S_u^2, the other terms being similar. By definition, $S_u^2/\mathcal{V}(u)$ has a χ_1^2 distribution so that

$$\mathbb{E}\left[e^{\zeta S_u^2}\right] = \frac{1}{\sqrt{1 - 2\zeta\mathcal{V}(u)}} < +\infty, \tag{11}$$

for $|\zeta|$ small enough, independently of n. The proposition then follows from standard large-deviation bounds [Dur96]. ■

Proposition 2 (Recursive Linear Estimator: Bias) *For all $\delta > 0$, there is $\varepsilon > 0$ small enough so that, assuming that Proposition 1 holds,*

$$\mathbb{E}[S_x \mid X_x] = \mathcal{B}(x) X_x,$$

for some $\mathcal{B}(x) > 0$ with $|\mathcal{B}(x) - 1| < \delta$.

Proof: We first show that the conditional biases at y_1, y_2 are accurately estimated. From Proposition 1, we have

$$|\ddot{\tau}(z_{21}, z_{22}) - (\tau(z_{21}, z_{22}) + \beta(z_{21}) + \beta(z_{22}))| < \varepsilon,$$

and similarly for $\ddot{\tau}(y_1, z_{21})$ and $\ddot{\tau}(y_1, z_{22})$. Then from (6), we get

$$\begin{aligned}
2\widehat{\beta}(y_1) &= \ddot{\tau}(y_1, z_{21}) + \ddot{\tau}(y_1, z_{22}) - \ddot{\tau}(z_{21}, z_{22}) - 2\hat{\tau}_{e_1} - 2\hat{\tau}_{e_2} \\
&\leq (\tau(y_1, z_{21}) + \beta(y_1) + \beta(z_{21})) + (\tau(y_1, z_{22}) + \beta(y_1) + \beta(z_{22})) \\
&\qquad -(\tau(z_{21}, z_{22}) + \beta(z_{21}) + \beta(z_{22})) - 2\tau_{e_1} - 2\tau_{e_2} + 7\varepsilon \\
&= 2\beta(y_1) + (\tau(y_1, z_{21}) + \tau(y_1, z_{22}) - \tau(z_{21}, z_{22})) - 2(\tau_{e_1} + \tau_{e_2}) + 7\varepsilon \\
&= 2\beta(y_1) + ([\tau(y_1, y_2) + \tau(y_2, z_{21})] + [\tau(y_1, y_2) + \tau(y_2, z_{22})] \\
&\qquad -[\tau(z_{21}, y_2) + \tau(y_2, z_{22})]) - 2\tau(y_1, y_2) + 7\varepsilon \\
&= 2\beta(y_1) + 7\varepsilon,
\end{aligned}$$

where we used the additivity of τ on line 4. And similarly for the other direction so that

$$|\widehat{\beta}(y_1) - \beta(y_1)| \leq \frac{7}{2}\varepsilon.$$

The same inequality holds for y_2.

Given $\omega_{y_1}, \omega_{y_2}$, the bias at x is

$$\begin{aligned}
\mathbb{E}[S_x \,|\, X_x] &= \mathbb{E}[\omega_{y_1} S_{y_1} + \omega_{y_2} S_{y_2} \,|\, X_x] \\
&= \sum_{\alpha=1,2} \omega_{y_\alpha} \mathbb{E}[\mathbb{E}[S_{y_\alpha} \,|\, X_{y_\alpha}, X_x] | X_x] \\
&= \sum_{\alpha=1,2} \omega_{y_\alpha} \mathbb{E}[\mathbb{E}[S_{y_\alpha} \,|\, X_{y_\alpha}] | X_x] \\
&= \sum_{\alpha=1,2} \omega_{y_\alpha} \mathbb{E}[\mathcal{B}(y_\alpha) X_{y_\alpha} | X_x] \\
&= (\omega_{y_1} \mathcal{B}(y_1) \rho_{e_1} + \omega_{y_2} \mathcal{B}(y_2) \rho_{e_2}) X_x \\
&\equiv \mathcal{B}(x) X_x,
\end{aligned}$$

where we used the Markov property on line 2 and the fact that X_V is Gaussian on line 5. The last line is a definition. Note that by the inequality above we have

$$\begin{aligned}
\mathcal{B}(x) &= \omega_{y_1} \mathcal{B}(y_1) \rho_{e_1} + \omega_{y_2} \mathcal{B}(y_2) \rho_{e_2} \\
&= \omega_{y_1} e^{-\beta(y_1)} \rho_{e_1} + \omega_{y_2} e^{-\beta(y_2)} \rho_{e_2} \\
&\leq \omega_{y_1} e^{-\widehat{\beta}(y_1) + 7/2\varepsilon} (\hat{\rho}_{e_1} + \varepsilon) + \omega_{y_2} e^{-\widehat{\beta}(y_2) + 7/2\varepsilon} (\hat{\rho}_{e_2} + \varepsilon) \\
&= (\omega_{y_1} \widehat{\mathcal{B}}(y_1) \hat{\rho}_{e_1} + \omega_{y_2} \widehat{\mathcal{B}}(y_2) \hat{\rho}_{e_2}) + \max\{\omega_{y_1}, \omega_{y_2}\} O(\varepsilon) \\
&= 1 + \max\{\omega_{y_1}, \omega_{y_2}\} O(\varepsilon),
\end{aligned}$$

where the last line follows from the definition of ω_{y_α}. Taking ε, δ small enough, from our previous bounds and equation (9), we can derive that $\omega_{y_\alpha} = O(1)$, $\alpha = 1, 2$. In particular, $\mathcal{B}(x) = 1 + O(\varepsilon)$ and, choosing ε small enough, it satisfies $|\mathcal{B}(x) - 1| < \delta$. ■

Proposition 3 (Recursive Linear Estimator: Variance) *There exists $c > 0$ large enough and $\varepsilon, \delta > 0$ small enough such that, assuming that Proposition 1 holds, we have*

$$\mathcal{V}(x) \equiv \mathrm{Var}[S_x] \leq c.$$

Proof: From (9),

$$\begin{aligned}
\omega_{y_1}^2 + \omega_{y_2}^2 &= \left(\frac{\rho_{e_1}^2}{(\rho_{e_1}^2 + \rho_{e_2}^2)^2} + \frac{\rho_{e_2}^2}{(\rho_{e_1}^2 + \rho_{e_2}^2)^2} \right) (1 + O(\varepsilon + \delta)) \\
&= \left(\frac{1}{\rho_{e_1}^2 + \rho_{e_2}^2} \right) (1 + O(\varepsilon + \delta)) \\
&\leq \frac{1}{2(\rho^*)^2} (1 + O(\varepsilon + \delta)) < 1,
\end{aligned}$$

for $\varepsilon, \delta > 0$ small enough, where $\rho^* = e^{-g}$ so that $2(\rho^*)^2 > 1$. Moreover,

$$\begin{aligned}
\mathrm{Var}[S_x] &= \mathrm{Var}[\omega_{y_1} S_{y_1} + \omega_{y_2} S_{y_2}] \\
&= \omega_{y_1}^2 \mathrm{Var}[S_{y_1}] + \omega_{y_2}^2 \mathrm{Var}[S_{y_2}] + \omega_{y_1} \omega_{y_2} \mathbb{E}[S_{y_1} S_{y_2}] \\
&\leq (\omega_{y_1}^2 + \omega_{y_2}^2) c + \omega_{y_1} \omega_{y_2} \mathcal{B}(y_1) \mathcal{B}(y_2) \Sigma_{uv} \\
&\leq (\omega_{y_1}^2 + \omega_{y_2}^2) c + \omega_{y_1} \omega_{y_2} (1 + \delta)^2 \\
&< c,
\end{aligned}$$

taking c large enough. ■

2.2 Topology reconstruction

Propositions 2 and 3 rely on the knowing the topology below x. In this section, we show how this is performed inductively. That is, we assume the topology is known up to level $0 \leq h' < h$ and that hidden state estimators have been derived up to that level. We then construct the next level of the tree.

Quartet Reconstruction Let $L_{h'}$ be the set of vertices in V at level h' from the leaves and let $\mathcal{Q} = \{a, b, c, d\} \subseteq L_{h'}$ be a 4-tuple on level h'. The topology of T restricted to $\mathcal{Q}$ is completely characterized by a bipartition or *quartet split* q of the form: $ab|cd$, $ac|bd$ or $ad|bc$. The most basic operation in quartet-based reconstruction algorithms is the inference of such quartet splits. This is done by performing a *four-point test*: letting

$$\mathcal{F}(ab|cd) = \frac{1}{2}[\tau(a,c) + \tau(b,d) - \tau(a,b) - \tau(c,d)],$$

we have

$$q = \begin{cases} ab|cd & \text{if } \mathcal{F}(a,b|c,d) > 0 \\ ac|bd & \text{if } \mathcal{F}(a,b|c,d) < 0 \\ ad|bc & \text{o.w.} \end{cases}$$

Note however that we cannot estimate directly the values $\tau(a, c)$, $\tau(b, d)$, $\tau(a, b)$, and $\tau(c, d)$ for internal nodes, that is, when $h' > 0$. Instead we use the internal estimates described in Proposition 1.

Deep Four-Point Test Let $D > 0$. We let

$$\widehat{\mathcal{F}}(ab|cd) = \frac{1}{2}[\ddot{\tau}(a,c) + \ddot{\tau}(b,d) - \ddot{\tau}(a,b) - \ddot{\tau}(c,d)],$$

and

$$\widehat{\mathbb{SD}}(\mathcal{S}) = \mathbb{1}\{\ddot{\tau}(x,y) \leq D,\ \forall x, y \in \mathcal{S}\}.$$

We define the *deep four-point test*

$$\widehat{\mathbb{FP}}(a,b|c,d) = \widehat{\mathbb{SD}}(\{a,b,c,d\})\mathbb{1}\{\widehat{\mathcal{F}}(ab|cd) > f/2\}.$$

Algorithm. Fix $\gamma > 2$, $0 < \varepsilon < f/4$, $0 < \delta < 1$ and $D = 4g + 2\ln(1+\delta) + \varepsilon$. Choose c, κ so as to satisfy Proposition 1. Let $\mathcal{Z}_0$ be the set of leaves. The algorithm is detailed in Figure 1.

2.3 Estimating the Edge Weights

Propositions 2 and 3 also rely on edge-length estimates. In this section, we show how this estimation is performed, assuming the tree topology is known below $x' \in L_{h'+1}$ and edges estimates are known below level h'. In Figure 1, this procedure is used as a subroutine in the tree-building algorithm.

Algorithm
Input: Samples $(X^i_{[n]})_{i=1}^k$;
Output: Tree;

- For $h' = 0, \ldots, h-1$,

  1. **Deep Four-Point Test.** Let

  $$\mathcal{R}_{h'} = \{q = ab|cd \ : \ \forall a, b, c, d \in \mathcal{Z}_{h'} \text{ distinct such that } \widehat{\mathbb{FP}}(q) = 1\}.$$

  2. **Cherries.** Identify the cherries in $\mathcal{R}_{h'}$, that is, those pairs of vertices that only appear on the same side of the quartet splits in $\mathcal{R}_{h'}$. Let

  $$\mathcal{Z}_{h'+1} = \{x_1^{(h'+1)}, \ldots, x_{2^{h-(h'+1)}}^{(h'+1)}\},$$

  be the parents of the cherries in $\mathcal{Z}_{h'}$

  3. **Edge Weights.** For all $x' \in \mathcal{Z}_{h'+1}$,

     (a) Let y'_1, y'_2 be the children of x'. Let z'_1, z'_2 be the children of y'_1. Let w' be any other vertex in $\mathcal{Z}_{h'}$ with $\widehat{\mathbb{SD}}(\{z'_1, z'_2, y'_2, w'\}) = 1$.

     (b) Let e'_1 be the edge between y'_1 and x'. Set

     $$\hat{\tau}_{e'_1} = \widehat{\mathbb{O}}(z'_1, z'_2; y'_2, w').$$

     (c) Repeat interchanging the role of y'_1 and y'_2.

Figure 1: Tree-building algorithm. In the deep four-point test, internal distance estimates are used as described in Section 2.1.

Let y'_1, y'_2 be the children of x' and let e'_1, e'_2 be the corresponding edges. Let w' in $L_{h'}$ be a vertex not descended from x'. (One should think of w' as being on the same level as on a neighboring subtree.) Our goal is to estimate the weight of e'_1. Denote by z'_1, z'_2 the children of y'_1. (Simply set $z'_1 = z'_2 = y'_1$ if y'_1 is a leaf.) Note that the internal edge of the quartet formed by z'_1, z'_2, y'_2, w' is e'_1. Hence, we use the standard four-point formula to compute the length of e'_1:

$$\hat{\tau}_{e'_1} \equiv \widehat{\mathbb{O}}(z'_1, z'_2; y'_2, w') = \frac{1}{2}(\ddot{\tau}(z'_1, y'_2) + \ddot{\tau}(z'_2, w') - \ddot{\tau}(z'_1, z'_2) - \ddot{\tau}(y'_2, w')),$$

and $\hat{\rho}_{e'_1} = e^{-\hat{\tau}_{e'_1}}$. Note that, with this approach, the biases at z'_1, z'_2, y'_2, w' cancel each other. This technique was used in [DMR11a].

Proposition 4 (Edge-Weight Estimation) *Consider the setup above. Assume that for all* $a, b \in \{z_1', z_2', y_2', w'\}$ *we have*

$$|\ddot{\tau}(a,b) - (\tau(a,b) + \beta(a) + \beta(b))| < \varepsilon/2,$$

for some $\varepsilon > 0$*. Then,* $|\hat{\tau}_{e_1'} - \tau_{e_1'}| < \varepsilon$*.*

This result follows from a calculation similar to the proof of Proposition 2.

2.4 Proof of Theorem 1

We are now ready to prove Theorem 1.

Proof:(Theorem 1) All steps of the algorithm are completed in polynomial time in n and k.

We argue about the correctness by induction on the levels. Fix $\gamma > 2$. Take $\delta > 0$, $0 < \varepsilon < f/4$ small enough and c, κ large enough so that Propositions 1, 2, 3, 4 hold. We divide the $\kappa \log^2 n$ samples into $\log n$ blocks.

Assume that, using the first h' sample blocks, the topology of the model has been correctly reconstructed and that we have edge estimates satisfying (6) up to level h'. Assume further that we have hidden state estimators satisfying (7) and (8) up to level $h' - 1$ (if $h' \geq 1$).

We now use the next block of samples which is independent of everything used until this level. When $h' = 0$, we can use the samples directly in the Deep Four-Point Test. Otherwise, we construct a linear hidden-state estimator for all vertices on level h'. Propositions 2 and 3 ensure that conditions (7) and (8) hold for the new estimators. By Proposition 1 applied to the new estimators and our choice of $D = 4g + 2\ln(1+\delta) + \varepsilon$, all cherries on level h' appear in at least one quartet and the appropriate quartet splits are reconstructed. Note that the second and third terms in D account for the bias and sampling error respectively. Once the cherries on level h' are reconstructed, Proposition 4 ensures that the edge weight are estimated so as to satisfy (6).

That concludes the induction. ■

2.5 Kesten-Stigum regime: Gaussian case

In this section, we derive the critical threshold for HSI in Gaussian tree models. The section culminates with a proof of Theorem 2 stating that TME cannot in general be achieved outside the KS regime without at least polynomially many samples.

2.5.1 Definitions

Recall that the *mutual information* between two random vectors $\mathbf{Y}_1$ and $\mathbf{Y}_2$ is defined as

$$I(\mathbf{Y}_1; \mathbf{Y}_2) = H(\mathbf{Y}_1) + H(\mathbf{Y}_2) - H(\mathbf{Y}_1, \mathbf{Y}_2),$$

where H is the *entropy*, that is,

$$H(\mathbf{Y}_1) = -\int f_1(\mathbf{y}_1) \log f_1(\mathbf{y}_1) d\mathbf{y}_1,$$

assuming $\mathbf{Y}_1$ has density f_1. See e.g. [CT91]. In the Gaussian case, if $\mathbf{Y}_1$ has covariance matrix Σ_1, then

$$H(\mathbf{Y}_1) = \frac{1}{2} \log(2\pi e)^{n_1} |\Sigma_1|,$$

where $|\Sigma_1|$ is the determinant of the $n_1 \times n_1$ matrix Σ_1.

Definition 5 (Solvability) *Let $X_V^{(h)}$ be a GMRFT on balanced tree*

$$\mathcal{T}^{(h)} = (V^{(h)}, E^{(h)}, [n^{(h)}], r^{(h)}; \tau^{(h)}),$$

where $n^{(h)} = 2^h$ and $\tau_e^{(h)} = \tau > 0$ for all $e \in E^{(h)}$. For convenience we denote the root by 0. We say that the GMRFT root state reconstruction problem with τ is solvable *if*

$$\liminf_{h \to \infty} I\left(X_0^{(h)}; X_{[n^{(h)}]}^{(h)}\right) > 0,$$

that is, if the mutual information between the root state and leaf states remains bounded away from 0 as the tree size goes to $+\infty$.

2.5.2 Threshold

Our main result in this section is the following.

Theorem 4 (Gaussian Solvability) *The GMRFT reconstruction problem is solvable if and only if*

$$2e^{-2\tau} > 1.$$

When $2e^{-2\tau} < 1$ then

$$I\left(X_0^{(h)}; X_{[n^{(h)}]}^{(h)}\right) = \left[2e^{-2\tau}\right]^h \cdot \frac{1 - 2e^{-2\tau} + o(1)}{2 - 2e^{-2\tau}}, \tag{12}$$

as $h \to \infty$.

Proof: Fix $h \geq 0$ and let $n = n^{(h)}$,

$$I_h = I\left(X_0^{(h)}; X_{[n]}^{(h)}\right),$$

$[[n]] = \{0, \ldots, n\}$, and $\rho = e^{-\tau}$. Assume $2\rho^2 \neq 1$. (The case $2\rho^2 = 1$ follows by a similar argument which we omit.) Denote by $\Sigma_{[n]}^{(h)}$ and $\Sigma_{[[n]]}^{(h)}$ the covariance matrices of $X_{[n]}^{(h)}$ and $(X_0^{(h)}, X_{[n]}^{(h)})$ respectively. Then

$$I_h = \frac{1}{2}\log\left(\frac{|\Sigma_{[n]}^{(h)}|}{|\Sigma_{[[n]]}^{(h)}|}\right).$$

Let $\mathbf{e}_n$ be the all-one vector with n elements. To compute the determinants above, we note that each eigenvector $\mathbf{v} \perp \mathbf{e}_n$ of $\Sigma_{[n]}^{(h)}$ gives an eigenvector $(0, \mathbf{v})$ of $\Sigma_{[[n]]}^{(h)}$ with the same eigenvalue. There are $2^h - 1$ such eigenvectors. Further $\mathbf{e}_n$ is an eigenvector of $\Sigma_{[n]}^{(h)}$ with positive eigenvalue corresponding to the sum of all pairwise correlation between a leaf and all other leaves (including itself), that is,

$$R_h = 1 + \sum_{l=1}^{h} \rho^{2l} 2^{l-1} = 1 + \rho^2\left(\frac{(2\rho^2)^h - 1}{2\rho^2 - 1}\right).$$

(The other eigenvectors are obtained inductively by noticing that each eigenvector $\mathbf{v}$ for size 2^{h-1} gives eigenvectors $(\mathbf{v}, \mathbf{v})$ and $(\mathbf{v}, -\mathbf{v})$ for size 2^h.) Similarly the remaining two eigenvectors of $\Sigma_{[[n]]}^{(h)}$ are of the form $(1, \beta\mathbf{e}_n)$ with

$$\Sigma_{[[n]]}^{(h)}(1, \beta\mathbf{e}_n)' = (1 + \beta 2^h \rho^h, (\rho^h + \beta R_h)\mathbf{e}_n)' = \lambda(1, \beta\mathbf{e}_n)',$$

whose solution is

$$\beta_h^{\pm} = \frac{(R_h - 1) \pm \sqrt{(R_h - 1)^2 + 4\rho^{2h}2^h}}{2\rho^h 2^h},$$

and

$$\lambda_h^{\pm} = 1 + \beta_h^{\pm} 2^h \rho^h.$$

Moreover note that

$$\begin{aligned} \lambda_h^+ \lambda_h^- &= 1 + (\beta_h^+ + \beta_h^-)2^h\rho^h + \beta_h^+\beta_h^- 2^{2h}\rho^{2h} \\ &= 1 + (R_h - 1) - \rho^{2h}2^h] \\ &= R_h - (2\rho^2)^h. \end{aligned}$$

Hence

$$\begin{aligned} I_h &= \frac{1}{2}\log\left(\frac{|\Sigma^{(h)}_{[n]}|}{|\Sigma^{(h)}_{[[n]]}|}\right) \\ &= \frac{1}{2}\log\left(\frac{R_h}{\lambda^+_h\lambda^-_h}\right) \\ &= -\frac{1}{2}\log\left(1-\frac{(2\rho^2)^h}{R_h}\right). \end{aligned}$$

Finally,

$$I_h \to \begin{cases} 0, & \text{if } 2\rho^2 < 1, \\ -\frac{1}{2}\log\left(\frac{1}{\rho^2}-1\right), & \text{if } 2\rho^2 > 1, \end{cases}$$

as $h \to +\infty$ with equation (12) established by a Taylor series expansion in the limit. ■

2.5.3 Hidden state reconstruction

We make precise the connection between solvability and hidden state estimation. We are interested in deriving good estimates of $X^{(h)}_0$ given $X^{(h)}_{[n]}$. Recall that the conditional expectation $\mathbb{E}[X^{(h)}_0|X^{(h)}_{[n]}]$ minimizes the mean squared error (MSE) [And58]. Let $\Lambda^{(h)}_{[n]} = (\Sigma^{(h)}_{[n]})^{-1}$. Under the Gaussian distribution, conditional on $X^{(h)}_{[n]}$, the distribution of $X^{(h)}_0$ is Gaussian with mean

$$\rho^h \mathbf{e}_n \Lambda^{(h)}_{[n]} X^{(h)}_{[n]} = \frac{\rho^h}{R_h}\mathbf{e}_n X^{(h)}_{[n]}, \tag{13}$$

and covariance

$$1 - \rho^{2h}\mathbf{e}_n\Lambda^{(h)}_{[n]}\mathbf{e}'_n = 1 - \frac{(2\rho^2)^h}{R_h} = e^{-2I_h}. \tag{14}$$

The MSE is then given by

$$\mathbb{E}[(X^{(h)}_0 - \mathbb{E}[X^{(h)}_0|X^{(h)}_{[n]}])^2] = \mathbb{E}[\mathrm{Var}[X^{(h)}_0|X^{(h)}_{[n]}]] = e^{-2I_h}.$$

Theorem 5 (Linear root-state estimation) *The linear root-state estimator*

$$\frac{\rho^h}{R_h}\mathbf{e}_n X^{(h)}_{[n]}$$

has asymptotic MSE < 1 *as* $h \to +\infty$ *if and only if* $2e^{-2\tau} > 1$. *(Note that achieving an MSE of* 1 *is trivial with the estimator identically zero.)*

The following observation explains why the proof of our main theorem centers on the derivation of an unbiased estimator with finite variance. Let $\widehat{X}_0^{(h)}$ be a random variable measurable with respect to the σ-field generated by $X_{[n]}^{(h)}$. Assume that $\mathbb{E}[\widehat{X}_0^{(h)}|X_0^{(h)}] = X_0^{(h)}$, that is, $\widehat{X}_0^{(h)}$ is a conditionally unbiased estimator of $X_0^{(h)}$. In particular $\mathbb{E}[\widehat{X}_0^{(h)}] = 0$. Then

$$\begin{aligned}
\mathbb{E}[(X_0^{(h)} - \alpha\widehat{X}_0^{(h)})^2] &= \mathbb{E}[\mathbb{E}[(X_0^{(h)} - \alpha\widehat{X}_0^{(h)})^2|X_0^{(h)}]] \\
&= 1 - 2\alpha\mathbb{E}[\mathbb{E}[X_0^{(h)}\widehat{X}_0^{(h)}|X_0^{(h)}]] + \alpha^2\mathrm{Var}[\widehat{X}_0^{(h)}] \\
&= 1 - 2\alpha + \alpha^2\mathrm{Var}[\widehat{X}_0^{(h)}],
\end{aligned}$$

which is minimized for $\alpha = 1/\mathrm{Var}[\widehat{X}_0^{(h)}]$. The minimum MSE is then $1 - 1/\mathrm{Var}[\widehat{X}_0^{(h)}]$. Therefore:

Theorem 6 (Unbiased root-state estimator) *There exists a root-state estimator with MSE* < 1 *if and only if there exists a conditionally unbiased root-state estimator with finite variance.*

2.5.4 Proof of Theorem 2

Finally in this section we establish that when $2e^{-2\tau} < 1$ the number of samples needed for TME grows like n^γ proving Theorem 2.

Proof:(Theorem 2) The proof follows the broad approach laid out in [Mos03, Mos04] for establishing sample size lower bounds for phylogenetic reconstruction. Let $\mathcal{T}$ and $\tilde{\mathcal{T}}$ be h-level balanced trees with common edge weight τ and the same vertex set differing only in the quartet split between the four vertices at graph distance 2 from the root $U = \{u_1, \ldots, u_4\}$ (that is, the grand-children of the root). Let $\{X_V^i\}_{i=1}^k$ and $\{\tilde{X}_V^i\}_{i=1}^k$ be k i.i.d. samples from the corresponding GMRFT.

Suppose that we are given the topology of the trees below level two from the root so that all that needs to be reconstructed is the top quartet split, that is, how U splits. By the Markov property and the properties of the multivariate Gaussian distribution, $\{Y_u^i\}_{u\in U, i\in\{1,\ldots,k\}}$ with $Y_u^i = \mathbb{E}[X_u^i \mid X_{\lfloor u \rfloor}^i]$ is a sufficient statistic for the topology of the top quartet, that is, it contains all the information given by the leaf states. Indeed, the conditional distribution of the states at U depends on the leaf states only through the condition expectations. To prove the impossibility

of TME with high probability, we will bound the total variation distance between $\underline{Y} = \{Y_u\}_{u\in U}$ and $\underline{\tilde{Y}} = \{\tilde{Y}_u\}_{u\in U}$. We have that $\underline{Y}$ is a mean 0 Gaussian vector and using equations (13) and (14) its covariance matrix Σ^*_U is given by

$$(\Sigma^*_U)_{uu} = \mathrm{Var}[Y_u] = e^{-2I_{h-2}} = 1 - O((2\rho^2)^h),$$

and

$$\begin{aligned}(\Sigma^*_U)_{uu'} &= \mathrm{Cov}[Y_u, X_u]\mathrm{Cov}[X_u, X_{u'}]\mathrm{Cov}[X_{u'}, Y_{u'}] \\ &= \frac{(2\rho^2)^{2(h-2)}}{R^2_{h-2}}(\Sigma_U)_{uu'} \\ &= O((2\rho^2)^{2h}).\end{aligned}$$

where Σ_U is the covariance matrix of X_U. The covariance matrix of $\underline{\tilde{Y}}$ is defined similarly. Let Λ^*_U (resp. $\tilde{\Lambda}^*_U$) denote the inverse covariance matrix $(\Sigma^*_U)^{-1}$ (resp. $(\tilde{\Sigma}^*_U)^{-1}$). We note that Σ^*_U and $\tilde{\Sigma}^*_U$ are close to the identity matrix and, hence, so are their inverses [HJ85]. Indeed, with I_U the 4×4-identity matrix, the elements of $\Sigma^*_U - I_U$ are all $O((2\rho^2)^h)$ and, similarly for $\tilde{\Sigma}^*_U$, which implies that

$$\sup_{u,u'} |\Lambda^*_{uu'} - \tilde{\Lambda}^*_{uu'}| = O((2\rho^2)^h). \tag{15}$$

We let $d_{\mathrm{TV}}(\cdot,\cdot)$ denote the total variation distance of two random vectors. Note that by symmetry $|\det \Lambda^*_U| = |\det \tilde{\Lambda}^*_U|$ and so, with $f_{\underline{Y}}(y)$ the density function of $\underline{Y}$, the total variation distance satisfies

$$\begin{aligned}d_{\mathrm{TV}}(\underline{Y}, \underline{\tilde{Y}}) &= \frac{1}{2}\int_{\mathbb{R}^4} \left|\frac{f_{\underline{\tilde{Y}}}(\underline{y})}{f_{\underline{Y}}(\underline{y})} - 1\right| f_{\underline{Y}}(\underline{y})d\underline{y} \\ &= \frac{1}{2}\int_{\mathbb{R}^4} \left|\exp\left[-\frac{1}{2}\underline{y}^T\tilde{\Lambda}^*_U\underline{y} + \frac{1}{2}\underline{y}^T\Lambda^*_U\underline{y}\right] - 1\right| f_{\underline{Y}}(\underline{y})d\underline{y} \\ &\le \frac{1}{2}\int_{\mathbb{R}^4} \left(\exp\left[O((2\rho^2)^h\sum_{j=1}^4 y_j^2)\right] - 1\right) f_{\underline{Y}}(\underline{y})d\underline{y} \\ &\le \frac{1}{2}\int_{\mathbb{R}^4} \left(\exp\left[O((2\rho^2)^h y_1^2)\right] - 1\right) f_{\underline{Y}}(\underline{y})d\underline{y} \\ &= \frac{1}{2}\left(\mathbb{E}\exp\left[O((2\rho^2)^h Y^2_{u_1})\right] - 1\right) \\ &= O((2\rho^2)^h),\end{aligned}$$

where the first inequality follows from equation (15) while the second follows from an application of the AM-GM inequality and fact that the Y_{u_i} are identically distributed. The final equality follows from an expansion of equation (11).

It follows that when $k = o((2\rho^2)^{-h})$ we can couple $\{Y_u^i\}_{u\in U, i\in\{1,\ldots,k\}}$ and $\{\tilde{Y}_u^i\}_{u\in U, i\in\{1,\ldots,k\}}$ with probability $(1 - O((2\rho^2)^h))^k$ which tends to 1. Since they form a sufficient statistic for the top quartet, this top structure of the graph cannot be recovered with probability approaching 1. Recalling that $n = 2^h$, $\rho = e^{-\tau}$ and that if $\gamma < (2\tau)/\log 2 - 1$ then $\mathbb{GMMRFT}^{f,g}$ is not solvable with $k = n^\gamma = o((2\rho^2)^{-h})$ samples. ■

3 GTR Model with Unknown Rate Matrix

In this section, we prove our reconstruction in the GTR case. We only describe the hidden-state estimator as the other steps are the same. We use notation similar to Section 2. We denote the tree by $T = (V, E)$ with root r. The number of leaves is denoted by n. Let $q \geq 2$, $0 < f < g < +\infty$, and $\mathcal{T} = (V, E, [n], r; \tau) \in \mathbb{BY}^{f,g}$. Fix $Q \in \mathbb{Q}_q$. We assume that $0 < g < g^\star_{\mathrm{KS}} = \ln\sqrt{2}$. We generate k i.i.d. samples $(Z_V^i)_{i=1}^k$ from the GTR model $(\mathcal{T}, Q)$ with state space $[q]$. Let ν^2 be a second right eigenvector of Q, that is, an eigenvector with eigenvalue -1. We will use the notation $X_u^i = \nu^2_{Z_u^i}$, for all $u \in V$ and $i = 1, \ldots, k$. We shall denote the leaves of T by $[n]$.

3.1 Estimating Rate and Frequency Parameters

We discuss in this section the issues involved in estimating Q and its eigenvectors using data at the leaves. For the purposes of our algorithm we need only estimate the first left eigenvector and the second right eigenvector. Let π be the stationary distribution of Q (first left eigenvector) and denote $\Pi = \mathrm{diag}(\pi)$. Let

$$\nu^1, \nu^2, \ldots, \nu^q,$$

be the right eigenvectors of Q corresponding respectively to eigenvalues

$$0 = \lambda_1 > \lambda_2 \geq \ldots \geq \lambda_q.$$

Because of the reversibility assumption, we can choose the eigenvectors to be orthonormal with respect to the inner product,

$$\langle \nu, \nu' \rangle_\pi = \sum_{i\in[q]} \pi_i \nu_i \nu'_i$$

In the case of multiplicity of eigenvalues this description may not be unique.

Proposition 5 *There exists $\kappa(\epsilon, \varrho, Q)$ such that given $\kappa \log n$ samples there exist estimators $\hat{\pi}$ and $\hat{\nu}^2$ such that*

$$\| \pi - \hat{\pi} \| \leq \epsilon, \tag{16}$$

and

$$\hat{\nu}^2 = \sum_{l=1}^{q} \alpha_l \nu^l, \tag{17}$$

where $|\alpha_2 - 1| \leq \varepsilon$ and $|\frac{\alpha_l}{\alpha_2}| < \varrho$ for $l \geq 3$, (for some choice of ν^l if the second eigenvalue has multiplicity greater than 1).

Estimates Let $\widehat{F}$ denote the empirical joint distribution at leaves a and b as a $q \times q$ matrix. (We use an extra sample block for this estimation.) To estimate π and ν^2, our first task is to find two leaves that are sufficiently close to allow accurate estimation. Let $a^*, b^* \in [n]$ be two leaves with minimum log-det distance

$$(a^*, b^*) \in \arg\min \left\{ -\log\det \widehat{F}^{ab} \ : \ (a, b) \in [n] \times [n] \right\}.$$

Let

$$F = F^{a^* b^*},$$

and consider the symmetrized correlation matrix

$$\widehat{F}^{\dagger} = \frac{1}{2}(\widehat{F}^{a^* b^*} + (\widehat{F}^{a^* b^*})^{\top}).$$

Then we estimate π from

$$\hat{\pi}_{\upsilon} = \sum_{\upsilon' \in [q]} \widehat{F}^{\dagger}_{\upsilon \upsilon'},$$

for all $\upsilon \in [q]$. Denote $\widehat{\Pi} = \mathrm{diag}(\hat{\pi})$. By construction, $\hat{\pi}$ is a probability distribution. Let $\varphi = \tau(a^*, b^*)$ and define G to be the symmetric matrix

$$G = \Pi^{-1/2} F \Pi^{-1/2} = \Pi^{-1/2}(\Pi e^{\varphi Q})\Pi^{-1/2} = \Pi^{1/2} e^{\varphi Q} \Pi^{-1/2}.$$

Then denote the right eigenvectors of G as

$$\mu^1 = \Pi^{1/2}\nu^1, \mu^2 = \Pi^{1/2}\nu^2, \ldots, \mu^q = \Pi^{1/2}\nu^q,$$

with corresponding eigenvalues

$$1 = \theta^{(1)}_{(a^*,b^*)} = e^{\varphi\lambda_1} > \theta^{(2)}_{(a^*,b^*)} = e^{\varphi\lambda_2} \geq \cdots \geq \theta^{(q)}_{(a^*,b^*)} = e^{\varphi\lambda_q},$$

orthonormal with respect to the Euclidean inner product. Note that $\theta^{(2)}_{(a^*,b^*)} < e^{-f}$ and that ν^1 is the all-one vector. Assuming $\hat{\pi} > 0$, define

$$\widehat{G} = \widehat{\Pi}^{-1/2}\widehat{F}^{\dagger}\widehat{\Pi}^{-1/2}.$$

which we use to estimate the eigenvectors and eigenvalues of Q. Since $\widehat{G}$ is real symmetric, it has q real eigenvalues $\hat{\theta}^{(1)} > \hat{\theta}^{(2)} \geq \ldots \geq \hat{\theta}^{(q)}$. with a corresponding orthonormal basis $\hat{\mu}^1, \hat{\mu}^2, \ldots, \hat{\mu}^q$. It can be checked that, provided $\widehat{G} > 0$, we have $1 = \hat{\theta}^{(1)} > \hat{\theta}^{(2)}$. We use

$$\hat{\nu}^2 = \widehat{\Pi}^{-1/2}\hat{\mu}^2.$$

as our estimate of the "second eigenvector" and $\hat{\theta}^{(2)}$ as our estimate of the second eigenvalue of the channel.

Discussion The sensitivity of eigenvectors is somewhat delicate [HJ85]. With sufficiently many samples ($k = \kappa \log n$ for large enough κ) the estimator $\widehat{G}$ will approximate G within any constant tolerance. When the second eigenvalue is distinct from the third one our estimate will satisfy (17) provided κ is large enough.

If there are multiple second eigenvectors the vector $\hat{\nu}^2$ may not exactly be an estimate of ν^2 since indeed the second eigenvalue is not uniquely defined: using classical results (see e.g. [GVL96]) it can be shown that $\hat{\nu}^2$ is close to a combination of eigenvectors with eigenvalues equal to $\theta^{(2)}$. Possibly after passing to a different basis of eigenvectors $\nu^1, \nu^2, \ldots, \nu^q$, we still have that equation (17) holds. By standard large deviations estimate this procedure satisfies Proposition 5 when κ is large enough.

Remark 4 *This procedure provides arbitrary accuracy as κ grows, however, for fixed κ it will not in general go to 0 as n goes to infinity as the choice of a^*, b^* may bias the result. An error of size $O(1/\sqrt{k})$ may be obtained by taking all pairs with log-det distance below some small threshold (say $4g$), randomly picking such a pair a', b' and estimating the matrix $\widehat{G}$ using a', b'.*

We could also have estimated $\hat{\pi}$ by taking the empirical distribution of the states at one of the vertices or indeed the empirical distribution over all vertices.

3.2 Recursive Linear Estimator

As in the Gaussian case, we build a recursive linear estimator. We use notation similar to Section 2. Let $K = \kappa \log n$ be the size of each block. We let Z_V be a generic sample from the GRT model independent of everything else, and we define $X_u = \hat{\nu}^2_{Z_u}$ for all $u \in V$. We let $(Z^i_{[n]})_{i=1}^K$ be a block of independent samples at the leaves, and we set $X^\ell_u = \hat{\nu}^2_{Z^\ell_u}$, for all $u \in V$ and $\ell = 1, \ldots, K$. For a node $u \in V$, we let $\lfloor u \rfloor$ be the leaves below u and $X_{\lfloor u \rfloor}$, the corresponding state. Let $0 < \delta < 1$ (small) and $c > 1$ (big) be constants to be defined later.

Linear estimator We build a linear estimator for each of the vertices recursively from the leaves. Let $x \in V - [n]$ with children (direct descendants) y_1, y_2. Assume that the topology of the tree rooted at x has been correctly reconstructed. Assume further that we have constructed linear estimators

$$S_u \equiv \mathcal{L}_u(X_{\lfloor u \rfloor})$$

of X_u, for all $u \in V$ below x. We use the convention that

$$\mathcal{L}_u(X_{\lfloor u \rfloor}) = X_u$$

if u is a leaf. We let $\mathcal{L}_x$ be a linear combination of the form

$$S_x \equiv \mathcal{L}_x(X_{\lfloor x \rfloor}) = \omega_{y_1} \mathcal{L}_{y_1}(X_{\lfloor y_1 \rfloor}) + \omega_{y_2} \mathcal{L}_{y_2}(X_{\lfloor y_2 \rfloor}), \tag{18}$$

where the ω's are chosen below.

Recursive conditions Assume that we have linear estimators $\mathcal{L}_u$ for all u below x satisfying

$$\mathbb{E}[S_u \mid Z_u] = \sum_{l=1}^{q} \mathcal{B}^l(u) \nu^l_{Z_u}, \tag{19}$$

for some $\mathcal{B}^l(u)$ such that $|\mathcal{B}^2(u) - 1| < \delta$ and $|\mathcal{B}^l(u)/\mathcal{B}^2(u)| < \varrho$ for $l = 3, \ldots, q$. Note that no condition is placed on $\mathcal{B}^1(u)$. Further for all $i \in [q]$

$$\Gamma^i_u(\zeta) \leq \zeta \mathbb{E}[S_u \mid Z_u = i] + c\zeta^2, \tag{20}$$

where as before

$$\Gamma^i_u(\zeta) \equiv \ln \mathbb{E}[\exp(\zeta S_u) \mid Z_u = i].$$

Observe that these conditions are satisfied at the leaves. Indeed, for $u \in [n]$ one has $S_u = \hat{\nu}^2_{Z_u} = \sum_{l=1}^q \alpha_l \nu^l_{Z_u}$ and therefore $\mathbb{E}[S_u \,|\, Z_u] = \sum_l^q \alpha_l \nu^l_{Z_u}$ and $\Gamma^i_u(\zeta) = \zeta \mathbb{E}[S_u \,|\, Z_u = i]$. We now seek to construct S_x so that it in turn satisfies the same conditions.

Moreover we assume we have a priori estimated edge weights $\hat{\tau}_e$ for all e below x such that for $\varepsilon > 0$ we have that

$$|\hat{\tau}_e - \tau_e| < \varepsilon. \tag{21}$$

Let $\hat{\theta}_e = e^{-\hat{\tau}_e}$.

First eigenvalue adjustment As discussed above, because we cannot estimate exactly the second eigenvector, our estimate $\hat{\nu}^2$ may contain components of other eigenvectors. While eigenvectors ν^3 through ν^q have smaller eigenvalues and will thus decay in importance as we recursively construct our estimator, the presence of a component in the direction of the first eigenvalue poses greater difficulties. However, we note that ν^1 is identically 1. So to remove the effect of the first eigenvalue from equation (19) we subtract the empirical mean of S_u,

$$\bar{S}_u = \frac{1}{K} \sum_{\ell=1}^{K} S_u^\ell.$$

As $\langle \pi, \nu^l \rangle = 0$ for $l = 2, \ldots, q$ and $\nu^1 \equiv 1$ we have that $\mathbb{E} S_u = \mathcal{B}^1(u)$ from (19) and hence the following proposition follows from standard large deviations estimates.

Proposition 6 (Concentration of Empirical Mean) *For $u \in V$, $\varepsilon' > 0$ and $\gamma > 0$, suppose that conditions* (19) *and* (20) *hold for some δ, ε and c. Then there exists $\kappa = \kappa(\varepsilon', c, \gamma, \delta, \varepsilon) > 0$ such that, when we have $K \geq \kappa \log n$ then*

$$|\bar{S}_u - \mathcal{B}^1(u)| < \varepsilon',$$

with probability at least $1 - O(n^{-\gamma})$.

Proof: Let $\varepsilon_\pi > 0$. By Chernoff's bound, of the K samples, $\widehat{K}_i$ are such that $Z^\ell_u = i$ where

$$\left| \frac{\widehat{K}_i}{K} - \pi_i \right| \leq \varepsilon_\pi,$$

except with inverse polynomial probability, given that κ is large enough. By (19) and (20), we have

$$\mathbb{E}[e^{\zeta(S_u-\mathcal{B}^1(u))}|Z_u=i] \le \zeta\mathbb{E}[(S_u-\mathcal{B}^1(u))|Z_u=i]+c\zeta^2,$$

where

$$\left|\mathbb{E}[(S_u-\mathcal{B}^1(u))|Z_u=i]\right| = \left|\sum_{l=2}^{q}\mathcal{B}^l(u)\nu_i^l\right| \le (1+\delta)(1+q\varrho)\max_j 1/\sqrt{\pi_j} \equiv \Upsilon.$$

Let $\varepsilon_\Gamma > 0$. Choosing $\zeta = \frac{\varepsilon_\Gamma}{2c}$ in Markov's inequality for $e^{\zeta(S_u-\mathcal{B}^1(u))}$ gives that the average of $S_u^\ell - \mathcal{B}^1(u)$ over the samples with $Z_u^\ell = i$ is within ε_Γ of $\sum_{l=2}^q \mathcal{B}^l(u)\nu_i^l$ except with probability at most $e^{-\varepsilon_\Gamma^2 K(\pi_i-\varepsilon_\pi)/4c} = 1/\text{poly}(n)$ for κ large enough and ε_π small enough. Therefore, in that case,

$$\left|\frac{1}{K}\sum_{\ell=1}^{K}(S_u^\ell-\mathcal{B}^1(u))\right| \;\le\; q\varepsilon_\Gamma+\varepsilon_\pi[\Upsilon+\varepsilon_\Gamma] < \varepsilon',$$

for $\varepsilon_\pi, \varepsilon_\Gamma$ small enough, where we used $\langle \pi, \nu^l\rangle = 0$ for $l = 2, \ldots, q$. ■

For $\alpha = 1, 2$, using the Markov property we have the following important conditional moment identity which we will use to relate the bias at y_α to the bias at x,

$$\begin{aligned}\mathbb{E}\left(S_{y_\alpha}^\ell - \mathcal{B}^1(y_\alpha) \mid Z_x = i\right) &= \sum_{l=2}^{q}\sum_{j=1}^{q}\mathcal{B}^l(y_\alpha)M_{ij}^{e_\alpha}\nu_j^l \\ &= \sum_{l=2}^{q}\mathcal{B}^l(y_\alpha)\theta_{e_\alpha}^{(l)}\nu_i^l, \end{aligned} \tag{22}$$

where we used the fact that the ν^l's are eigenvectors of $M_{ij}^{e_\alpha}$ with eigenvectors $\theta_e^{(l)} = \exp(-\lambda_l\tau_e)$.

Procedure We first define a procedure for estimating the path length (that is, the sum of edge weights) between a pair of vertices u_1 and u_2 including the bias. For $u_1, u_2 \in V$ with common ancestor v we define

$$\ddot{\tau}(u_1,u_2) = -\ln\left(\frac{1}{K}\sum_{\ell=1}^{K}\left(S_{u_1}^\ell-\bar{S}_{u_1}\right)\left(S_{u_2}^\ell-\bar{S}_{u_2}\right)\right).$$

This estimator differs from Section 2.1 in that we subtract the empirical means to remove the effect of the first eigenvalue. Using the fact that $\sum_{\ell=1}^{k} S_{u_1}^{\ell} - \bar{S}_{u_1} = 0$ and Proposition 6 we have that with probability at least $1 - O(n^{-\gamma})$

$$\begin{aligned}
&\frac{1}{K}\sum_{\ell=1}^{K}\left(S_{u_1}^{\ell} - \bar{S}_{u_1}\right)\left(S_{u_2}^{\ell} - \bar{S}_{u_2}\right) \\
&\quad= \frac{1}{K}\sum_{\ell=1}^{K}\Big[\left(S_{u_1}^{\ell} - \mathcal{B}^1(u_1)\right)\left(S_{u_2}^{\ell} - \mathcal{B}^1(u_2)\right) \\
&\qquad\quad + \left(\bar{S}_{u_1}^{\ell} - \mathcal{B}^1(u_1)\right)\left(\bar{S}_{u_2}^{\ell} - \mathcal{B}^1(u_2)\right)\Big] \\
&\quad\leq \frac{1}{K}\sum_{\ell=1}^{K}\left(S_{u_1}^{\ell} - \mathcal{B}^1(u_1)\right)\left(S_{u_2}^{\ell} - \mathcal{B}^1(u_2)\right) + (\varepsilon')^2,
\end{aligned}$$

and similarly the other direction so,

$$\begin{aligned}
&\Bigg|\frac{1}{K}\sum_{\ell=1}^{K}\left(S_{u_1}^{\ell} - \bar{S}_{u_1}\right)\left(S_{u_2}^{\ell} - \bar{S}_{u_2}\right) \\
&\qquad -\frac{1}{K}\sum_{\ell=1}^{K}\left(S_{u_1}^{\ell} - \mathcal{B}^1(u_1)\right)\left(S_{u_2}^{\ell} - \mathcal{B}^1(u_2)\right)\Bigg| \leq (\varepsilon')^2. \qquad (23)
\end{aligned}$$

It follows that $\ddot{\tau}(u_1, u_2)$ is an estimate of the length between u_1 and u_2 including bias since

$$\begin{aligned}
&\mathbb{E}\left[\left(S_{u_1}^{\ell} - \mathcal{B}^1(u_1)\right)\left(S_{u_2}^{\ell} - \mathcal{B}^1(u_2)\right)\right] \\
&\quad= \sum_{i\in[q]}\pi_i \mathbb{E}\left(S_{u_1}^{\ell} - \mathcal{B}^1(u_1) \mid Z_v = i\right)\mathbb{E}\left(S_{u_2}^{\ell} - \mathcal{B}^1(u_2) \mid Z_v = i\right) \\
&\quad= \sum_{i\in[q]}\pi_i\left(\sum_{l=2}^{q}\mathcal{B}^l(u_1)\theta_{(v,u_1)}^{(l)}\nu_j^l\right)\left(\sum_{l=2}^{q}\mathcal{B}^l(u_2)\theta_{(v,u_2)}^{(l)}\nu_j^l\right) \\
&\quad= \mathcal{B}^2(u_1)\theta_{(v,u_1)}^{(2)}\mathcal{B}^2(u_2)\theta_{(v,u_2)}^{(2)} + O(\varrho) \\
&\quad= \mathcal{B}^2(u_1)\mathcal{B}^2(u_2)e^{-\tau(u_1,u_2)} + O(\varrho), \qquad (24)
\end{aligned}$$

where line 2 follows from equation (22). Above we also used the recursive assumptions and the fact that $\sum_{i\in[q]}\pi_i(\nu_i^2)^2 = 1$. We will use the estimator $\ddot{\tau}(u, v)$ to estimate $\beta(u) = -\ln\mathcal{B}^2(u)$. Given the previous setup, we choose the weights ω_{y_α}, $\alpha = 1, 2$, as follows:

1. **Estimating the Biases.** If y_1, y_2 are leaves, we let $\widehat{\beta}(y_\alpha) = 0$, $\alpha = 1, 2$. Otherwise, let $z_{\alpha 1}, z_{\alpha 2}$ be the children of y_α. We then compute

$$\widehat{\beta}(y_1) = \frac{1}{2}(\ddot{\tau}(y_1, z_{21}) + \ddot{\tau}(y_1, z_{22}) - \ddot{\tau}(z_{21}, z_{22}) - 2\hat{\tau}_{e_1} - 2\hat{\tau}_{e_2}),$$

   And similarly for y_2. Let $\widehat{\mathcal{B}}^2(y_\alpha) = e^{-\widehat{\beta}(y_\alpha)}$, $\alpha = 1, 2$.

2. **Minimizing the Variance.** Set $\omega_{y_\alpha}, \alpha = 1, 2$ as

$$\omega_{y_\alpha} = \frac{\widehat{\mathcal{B}}^2(y_\alpha)\theta^{(2)}_{e_\alpha}}{(\widehat{\mathcal{B}}^2(y_1))^2(\theta^{(2)}_{e_1})^2 + (\widehat{\mathcal{B}}^2(y_2))^2(\theta^{(2)}_{e_2})^2}, \tag{25}$$

   the solution of the following optimization problem:

$$\min\{\omega^2_{y_1} + \omega^2_{y_2} \,:\, \omega_{y_1}\widehat{\mathcal{B}}^2(y_1)\theta^{(2)}_{e_1} + \omega_{y_2}\widehat{\mathcal{B}}^2(y_2)\theta^{(2)}_{e_2} = 1,\ \omega_{y_1}, \omega_{y_2} > 0\}. \tag{26}$$

   The constraint above guarantees that the bias condition (19) is satisfied when we set

$$\mathcal{L}_x(X_{\lfloor x \rfloor}) = \omega_{y_1}\mathcal{L}_{y_1}(X_{\lfloor y_1 \rfloor}) + \omega_{y_2}\mathcal{L}_{y_2}(X_{\lfloor y_2 \rfloor}).$$

Bias and Exponential Moment We now prove (19) and (20) recursively assuming (21) is satisfied. Assume the setup of the previous paragraph. We already argued that (19) and (20) are satisfied at the leaves. Assume further that they are satisfied for all descendants of x. We first show that the $\ddot{\tau}$-quantities are concentrated.

Proposition 7 (Concentration of Internal Distance Estimates) *For all $\varepsilon > 0$, $\gamma > 0$, $0 < \delta < 1$ and $c > 0$, there are $\kappa = \kappa(\varepsilon, \gamma, \delta, c) > 0$, $\varrho = \varrho(\varepsilon, \gamma, \delta, c) > 0$ such that, with probability at least $1 - O(n^{-\gamma})$, we have*

$$|\ddot{\tau}(u, v) - (\tau(u, v) + \beta(u) + \beta(v))| < \varepsilon,$$

for all $u, v \in \{y_1, y_2, z_{11}, z_{12}, z_{21}, z_{22}\}$ where $z_{\alpha 1}, z_{\alpha 2}$ are the children of y_α.

Proof: This proposition is proved similarly to Proposition 1 by establishing concentration of $\frac{1}{K}\sum_{\ell=1}^{K} \widetilde{S}^\ell_u \widetilde{S}^\ell_v$, where $\widetilde{S}^\ell_u = S^\ell_u - \mathcal{B}^1(u)$, around its mean which is approximately $e^{-\tau(u,v)-\beta(u)-\beta(v)}$ by equation (24). The only difference with Proposition 1 is that, in this non-Gaussian case, we must estimate the exponential moment directly using (20). We use an argument of [PR11, Roc10].

Let $\zeta > 0$. Let N be a standard normal. Using that $\mathbb{E}[e^{\alpha N}] = e^{\alpha^2/2}$ and applying (19) and (20),

$$\begin{aligned}\mathbb{E}[e^{\zeta\widetilde{S}_u\widetilde{S}_v}|Z_{\{u,v\}}] &\leq \mathbb{E}[e^{(\zeta\widetilde{S}_u)\mathbb{E}[\widetilde{S}_v|Z_v]+c(\zeta\widetilde{S}_u)^2}|Z_{\{u,v\}}] \\ &= \mathbb{E}[e^{\zeta\widetilde{S}_u\mathbb{E}[\widetilde{S}_v|Z_v]+\sqrt{2c}\zeta\widetilde{S}_u N}|Z_{\{u,v\}}] \\ &\leq \mathbb{E}[e^{(\zeta\mathbb{E}[\widetilde{S}_v|Z_v]+\sqrt{2c}\zeta N)\mathbb{E}[\widetilde{S}_u|Z_u]+c(\zeta\mathbb{E}[\widetilde{S}_v|Z_v]+\sqrt{2c}\zeta N)^2}|Z_{\{u,v\}}].\end{aligned}$$

We factor out the constant term and apply Cauchy-Schwarz on the linear and quadratic terms in N

$$\begin{aligned}&\mathbb{E}[e^{\zeta\widetilde{S}_u\widetilde{S}_v}|Z_{\{u,v\}}] \\ &\quad\leq e^{\zeta\mathbb{E}[\widetilde{S}_u|Z_u]\mathbb{E}[\widetilde{S}_v|Z_v]}e^{c\zeta^2\Upsilon^2}\mathbb{E}[e^{4c^2\zeta^2N^2}]^{1/2} \\ &\qquad\times\mathbb{E}\left[e^{2(\sqrt{2c}\zeta\mathbb{E}[\widetilde{S}_u|Z_u]+2c\sqrt{2c}\zeta^2\mathbb{E}[\widetilde{S}_v|Z_v])N}|Z_{\{u,v\}}\right]^{1/2} \\ &\quad\leq e^{\zeta\mathbb{E}[\widetilde{S}_u|Z_u]\mathbb{E}[\widetilde{S}_v|Z_v]}e^{c\zeta^2\Upsilon^2}\frac{1}{(1-8c^2\zeta^2)^{1/4}}e^{2c\Upsilon^2\zeta^2(1+2c\zeta)^2} \\ &\quad= 1+\zeta\mathbb{E}[\widetilde{S}_u\widetilde{S}_v|Z_{\{u,v\}}]+\Upsilon'\zeta^2+O(\zeta^3),\end{aligned}$$

as $\zeta \to 0$, where Υ was defined in the proof of Proposition 6 and $\Upsilon' > 0$ is a constant depending on Υ and c. Taking expectations and expanding

$$e^{-\zeta(\mathbb{E}[\widetilde{S}_u\widetilde{S}_v]+\varepsilon)}\mathbb{E}[e^{\zeta\widetilde{S}_u\widetilde{S}_v}] = 1-\varepsilon\zeta+\Upsilon'\zeta^2+O(\zeta^3) < 1,$$

for ζ small enough, independently of n. Applying Markov's inequality gives the result. ■

Proposition 8 (Recursive Linear Estimator: Bias) *Assuming (19), (20), and (21) hold for some $\varepsilon > 0$ that is small enough, we have*

$$\mathbb{E}[S_x \mid Z_x] = \sum_{l=1}^{q}\mathcal{B}^l(x)\nu^l_{Z_x},$$

for some $\mathcal{B}^l(x)$ such that $|\mathcal{B}^2(x)-1| < \delta$ and $|\mathcal{B}^l(x)/\mathcal{B}^2(x)| < \varrho$ for $l = 3,\ldots,q$.

Proof: We first show that the biases at y_1, y_2 are accurately estimated. Applying a similar proof to that of Proposition 2 (using Proposition 7 in place of Proposition 1) we have that

$$|\widehat{\beta}(y_1)-\beta(y_1)| \leq O(\varepsilon+\varrho).$$

The same inequality holds for y_2. Taking ε, δ small enough, our previous bounds on $\mathcal{B}$, θ and their estimates, we derive from equation (25) that $\omega_{y_\alpha} = \Theta(1)$, $\alpha = 1, 2$ with high probability. We now calculate the bias at x to be,

$$\begin{aligned}
\mathbb{E}[S_x \mid Z_x = i] &= \mathbb{E}[\omega_{y_1} S_{y_1} + \omega_{y_2} S_{y_2} \mid Z_x = i] \\
&= \sum_{\alpha=1,2} \omega_{y_\alpha} \sum_{l=1}^{q} \mathcal{B}^l(y_\alpha) \theta_{e_\alpha}^{(l)} \nu_j^l \\
&= \sum_{l=1}^{q} \left(\omega_{y_1} \mathcal{B}^l(y_1) \theta_{e_1}^{(l)} + \omega_{y_2} \mathcal{B}^l(y_2) \theta_{e_2}^{(l)} \right) \nu_j^l \\
&\equiv \sum_{l=1}^{q} \mathcal{B}^l(x) \nu_j^l
\end{aligned}$$

where we used equation (22) on line 2. Observe that since $\omega_{y_1}, \omega_{y_2}$ are positive and $0 < \theta_{e_\alpha}^{(l)} \leq \theta_{e_\alpha}^{(2)}$ for $l \geq 3$,

$$\begin{aligned}
\left| \frac{\mathcal{B}^l(x)}{\mathcal{B}^2(x)} \right| &= \left| \frac{\omega_{y_1} \mathcal{B}^l(y_1) \theta_{e_1}^{(l)} + \omega_{y_2} \mathcal{B}^l(y_2) \theta_{e_2}^{(l)}}{\omega_{y_1} \mathcal{B}^2(y_1) \theta_{e_1}^{(2)} + \omega_{y_2} \mathcal{B}^2(y_2) \theta_{e_2}^{(2)}} \right| \\
&\leq \left| \frac{\omega_{y_1} \varrho \mathcal{B}^2(y_1) \theta_{e_1}^{(2)} + \omega_{y_2} \varrho \mathcal{B}^2(y_2) \theta_{e_2}^{(2)}}{\omega_{y_1} \mathcal{B}^2(y_1) \theta_{e_1}^{(2)} + \omega_{y_2} \mathcal{B}^2(y_2) \theta_{e_2}^{(2)}} \right| \\
&= \varrho.
\end{aligned}$$

Applying the bounds on ω_{y_α} and $\widehat{\beta}(y_\alpha)$ for $\alpha = 1, 2$ we have that

$$\begin{aligned}
\mathcal{B}^2(x) &= \omega_{y_1} \mathcal{B}^2(y_1) \theta_{e_1}^{(2)} + \omega_{y_2} \mathcal{B}^2(y_2) \theta_{e_2}^{(2)} \\
&= \omega_{y_1} e^{-\beta(y_1)} \theta_{e_1}^{(2)} + \omega_{y_2} e^{-\beta(y_2)} \theta_{e_2}^{(2)} \\
&\leq \omega_{y_1} e^{-\widehat{\beta}(y_1) + O(\varepsilon+\varrho)} (\hat{\theta}_{e_1}^{(2)} + O(\varepsilon + \varrho)) \\
&\quad + \omega_{y_2} e^{-\widehat{\beta}(y_2) + O(\varepsilon+\varrho)} (\hat{\theta}_{e_2}^{(2)} + O(\varepsilon + \varrho)) \\
&= (\omega_{y_1} \widehat{\mathcal{B}}^2(y_1) \hat{\theta}_{e_1}^{(2)} + \omega_{y_2} \widehat{\mathcal{B}}^2(y_2) \hat{\theta}_{e_2}^{(2)}) + O(\varepsilon + \varrho) \\
&= 1 + O(\varepsilon + \varrho).
\end{aligned}$$

Choosing ε and ρ small enough, it satisfies $|\mathcal{B}^2(x) - 1| < \delta$. ■

Proposition 9 (Recursive Linear Estimator: Exponential Bound) *There is $c > 0$ such that, assuming (19), (20), and (21) hold, we have for all $i \in [q]$*

$$\Gamma_x^i(\zeta) \leq \zeta \mathbb{E}[S_x \mid Z_x = i] + c\zeta^2.$$

Proof: We use the following lemma suitably generalized from [PR11, Roc10].

Lemma 1 (Recursion Step) *Let $M = e^{\tau Q}$ as above with eigenvectors*

$$\nu^1, \nu^2, \ldots, \nu^q,$$

with corresponding eigenvalues $1 = e^{\lambda_1} \geq \ldots \geq e^{\lambda_q}$. Let $b_2, \ldots, b_q$ we arbitrary constants with $|b_i| < 2$. Then there is $c' > 0$ depending on Q such that for all $i \in [q]$

$$F(x) \equiv \sum_{j \in [q]} M_{ij} \exp\left(x \sum_{l=2}^{q} b_l \nu_j^l \right) \leq \exp\left(x \sum_{l=2}^{q} \lambda_l b_l \nu_i^l + c' x^2 \right) \equiv G(x),$$

for all $x \in \mathbb{R}$.

We have by the Markov property and Lemma 1 above,

$$\begin{aligned}
\Gamma_x^i(\zeta) &= \ln \mathbb{E}\left[\exp\left(\zeta \sum_{\alpha=1,2} S_{y_\alpha}\omega_{y_\alpha}\right) \mid Z_x = i\right] \\
&= \sum_{\alpha=1,2} \ln \mathbb{E}\left[\exp\left(\zeta S_{y_\alpha}\omega_{y_\alpha}\right) \mid Z_x = i\right] \\
&= \sum_{\alpha=1,2} \ln\left(\sum_{j\in[q]} M_{ij}^{e_\alpha} \mathbb{E}\left[\exp\left(\zeta S_{y_\alpha}\omega_{y_\alpha}\right) \mid Z_{y_\alpha} = j\right]\right) \\
&= \sum_{\alpha=1,2} \ln\left(\sum_{j\in[q]} M_{ij}^{e_\alpha} \exp\left(\Gamma_{y_\alpha}^j\left(\zeta\omega_{y_\alpha}\right)\right)\right) \\
&\leq \sum_{\alpha=1,2} \ln\left(\sum_{j\in[q]} M_{ij}^{e_\alpha} \exp\left(\zeta\omega_{y_\alpha}\mathbb{E}[S_{y_\alpha} \mid Z_{y_\alpha} = j] + c\zeta^2\omega_{y_\alpha}^2\right)\right) \\
&= c\zeta^2 \sum_{\alpha=1,2} \omega_{y_\alpha}^2 + \sum_{\alpha=1,2} \ln\left(\sum_{j\in[q]} M_{ij}^{e_\alpha} \exp\left(\zeta\omega_{y_\alpha} \sum_{l=1}^{q} \mathcal{B}^l(y_\alpha)\nu_j^l\right)\right) \\
&= c\zeta^2 \sum_{\alpha=1,2} \omega_{y_\alpha}^2 + \zeta \sum_{\alpha=1,2} \mathcal{B}^1(y_\alpha)\omega_{y_\alpha} \\
&\qquad + \sum_{\alpha=1,2} \ln\left(\sum_{j\in[q]} M_{ij}^{e_\alpha} \exp\left(\zeta\omega_{y_\alpha} \sum_{l=2}^{q} \mathcal{B}^l(y_\alpha)\nu_j^l\right)\right) \\
&\leq c\zeta^2 \sum_{\alpha=1,2} \omega_{y_\alpha}^2 + \zeta \sum_{\alpha=1,2} \omega_{y_\alpha} \sum_{l=1}^{q} \theta_{e_\alpha}^{(l)} \mathcal{B}^l(y_\alpha)\nu_i^l + \sum_{\alpha=1,2} c'\zeta^2\omega_{y_\alpha}^2 \\
&= \zeta\mathbb{E}\left[S_x \mid Z_x = i\right] + \zeta^2\left(c + c'\right) \sum_{\alpha=1,2} \omega_{y_\alpha}^2
\end{aligned}$$

Take c large enough so that $c + c' < c(1 + \varepsilon')$ for some small $\varepsilon' > 0$. Moreover, from (25)

$$\begin{aligned}
\omega_{y_1}^2 + \omega_{y_2}^2 &= \left(\frac{\theta_{e_1}^2}{(\theta_{e_1}^2 + \theta_{e_2}^2)^2} + \frac{\theta_{e_2}^2}{(\theta_{e_1}^2 + \theta_{e_2}^2)^2} \right) (1 + O(\varepsilon + \delta + \varrho)) \\
&= \left(\frac{1}{\theta_{e_1}^2 + \theta_{e_2}^2} \right) (1 + O(\varepsilon + \delta + \varrho)) \\
&\leq \frac{1}{2(\theta^*)^2} (1 + O(\varepsilon + \delta + \varrho)) < 1,
\end{aligned}$$

where $\theta^* = e^{-g}$ so that $2(\theta^*)^2 > 1$. Hence,

$$\Gamma_x^i(\zeta) \leq \zeta \mathbb{E}[S_x \,|\, Z_x = i] + c\zeta^2.$$

■

4 Concluding remarks

We have shown how to reconstruct latent tree Gaussian and GTR models using $O(\log^2 n)$ samples in the KS regime. In contrast, a straightforward application of previous techniques $O(\log^3 n)$ samples. Several questions arise from our work:

- Can this reconstruction be done using only $O(\log n)$ samples? Indeed this is the case for the CFN model [Mos04] and it is natural to conjecture that it may be true more generally. However our current techniques are limited by our need to use fresh samples on each level of the tree to avoid unwanted correlations between coefficients and samples in the recursive conditions.

- Do our techniques extend to general trees? The boosted algorithm used here has been generalized to non-homogeneous trees using a combinatorial algorithm of [DMR11a] (where edge weights are discretized to avoid the robustness issues considered in this paper). However general trees have, in the worst case, linear diameters. To apply our results, one would need to control the depth of the subtrees used for root-state estimation in the combinatorial algorithm. We leave this extension for future work.

References

[And58] T. W. Anderson. *An introduction to multivariate statistical analysis*. Wiley Publications in Statistics. John Wiley & Sons Inc., New York, 1958.

[Att99] K. Atteson. The performance of neighbor-joining methods of phylogenetic reconstruction. *Algorithmica*, 25(2-3):251–278, 1999.

[BRR10] Shankar Bhamidi, Ram Rajagopal, and Sébastien Roch. Network delay inference from additive metrics. *Random Structures Algorithms*, 37(2):176–203, 2010.

[CCL+04] Rui Castro, Mark Coates, Gang Liang, Robert Nowak, and Bin Yu. Network tomography: recent developments. *Statist. Sci.*, 19(3):499–517, 2004.

[CT91] T. M. Cover and J. A. Thomas. *Elements of information theory*. Wiley Series in Telecommunications. John Wiley & Sons Inc., New York, 1991. A Wiley-Interscience Publication.

[CTAW11] Myung Jin Choi, Vincent Y.F. Tan, Animashree Anandkumar, and Alan S. Willsky. Learning latent tree graphical models. *Journal of Machine Learning Research*, 12:1771–1812, 2011.

[DMR11a] Constantinos Daskalakis, Elchanan Mossel, and Sébastien Roch. Evolutionary trees and the ising model on the bethe lattice: a proof of steel's conjecture. *Probability Theory and Related Fields*, 149:149–189, 2011. 10.1007/s00440-009-0246-2.

[DMR11b] Constantinos Daskalakis, Elchanan Mossel, and Sébastien Roch. Phylogenies without branch bounds: Contracting the short, pruning the deep. *SIAM J. Discrete Math.*, 25(2):872–893, 2011.

[Dur96] Richard Durrett. *Probability: theory and examples*. Duxbury Press, Belmont, CA, second edition, 1996.

[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.

[ESSW99a] P. L. Erdös, M. A. Steel, L. A. Székely, and T. A. Warnow. A few logs suffice to build (almost) all trees (part 1). *Random Struct. Algor.*, 14(2):153–184, 1999.

[ESSW99b] P. L. Erdös, M. A. Steel, L. A. Székely, and T. A. Warnow. A few logs suffice to build (almost) all trees (part 2). *Theor. Comput. Sci.*, 221:77–118, 1999.

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

[GMS08] Ilan Gronau, Shlomo Moran, and Sagi Snir. Fast and reliable reconstruction of phylogenetic trees with very short edges. In *SODA '08: Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 379–388, Philadelphia, PA, USA, 2008. Society for Industrial and Applied Mathematics.

[GVL96] Gene H. Golub and Charles F. Van Loan. *Matrix computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, third edition, 1996.

[HJ85] Roger A. Horn and Charles R. Johnson. *Matrix analysis*. Cambridge University Press, Cambridge, 1985.

[KF09] Daphne Koller and Nir Friedman. *Probabilistic graphical models*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2009. Principles and techniques.

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

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

[Mos03] E. Mossel. On the impossibility of reconstructing ancestral data and phylogenies. *J. Comput. Biol.*, 10(5):669–678, 2003.

[Mos04] E. Mossel. Phase transitions in phylogeny. *Trans. Amer. Math. Soc.*, 356(6):2379–2404, 2004.

[Mos07] E. Mossel. Distorted metrics on trees and phylogenetic forests. *IEEE/ACM Trans. Comput. Bio. Bioinform.*, 4(1):108–116, 2007.

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

[MR06] Elchanan Mossel and Sébastien Roch. Learning nonsingular phylogenies and hidden Markov models. *Ann. Appl. Probab.*, 16(2):583–614, 2006.

[MRS11] Elchanan Mossel, Sébastien Roch, and Allan Sly. On the inference of large phylogenies with long branches: How long is too long? *Bulletin of Mathematical Biology*, 73:1627–1644, 2011. 10.1007/s11538-010-9584-6.

[PR11] Yuval Peres and Sébastien Roch. Reconstruction on trees: Exponential moment bounds for linear estimators. *Electron. Comm. Probab.*, 16:251–261 (electronic), 2011.

[Roc10] Sebastien Roch. Toward extracting all phylogenetic information from matrices of evolutionary distances. *Science*, 327(5971):1376–1379, 2010.

[Sly09] Allan Sly. Reconstruction for the potts model. In *STOC*, pages 581–590, 2009.

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

[Ste01] M. Steel. My Favourite Conjecture. Preprint, 2001.

[TATW11] Vincent Y. F. Tan, Animashree Anandkumar, Lang Tong, and Alan S. Willsky. A large-deviation analysis of the maximum-likelihood learning of markov tree structures. *IEEE Transactions on Information Theory*, 57(3):1714–1735, 2011.

[TAW10] Vincent Y. F. Tan, Animashree Anandkumar, and Alan S. Willsky. Learning gaussian tree models: analysis of error exponents and extremal structures. *IEEE Transactions on Signal Processing*, 58(5):2701–2714, 2010.

[TAW11] Vincent Y.F. Tan, Animashree Anandkumar, and Alan S. Willsky. Learning high-dimensional markov forest distributions: Analysis of error rates. *Journal of Machine Learning Research*, 12:1617–1653, 2011.

[Wil02] A.S. Willsky. Multiresolution markov models for signal and image processing. *Proceedings of the IEEE*, 90(8):1396 – 1458, aug 2002.