Reconstruction on Trees: Exponential Moment Bounds for Linear Estimators[*]

Yuval Peres[†] Sebastien Roch[‡]

June 14, 2013

Abstract

Consider a Markov chain $(\xi_v)_{v\in V} \in [k]^V$ on the infinite b-ary tree $T = (V, E)$ with irreducible edge transition matrix M, where $b \geq 2$, $k \geq 2$ and $[k] = \{1, \ldots, k\}$. We denote by L_n the level-n vertices of T. Assume M has a real second-largest (in absolute value) eigenvalue λ with corresponding real eigenvector $\nu \neq 0$. Letting $\sigma_v = \nu_{\xi_v}$, we consider the following root-state estimator, which was introduced by Mossel and Peres (2003) in the context of the "recontruction problem" on trees:

$$S_n = (b\lambda)^{-n} \sum_{x\in L_n} \sigma_x.$$

As noted by Mossel and Peres, when $b\lambda^2 > 1$ (the so-called Kesten-Stigum reconstruction phase) the quantity S_n has uniformly bounded variance. Here, we give bounds on the moment-generating functions of S_n and S_n^2 when $b\lambda^2 > 1$. Our results have implications for the inference of evolutionary trees.

Keywords: Markov chains on trees, reconstruction problem, Kesten-Stigum bound, phylogenetic reconstruction

1 Introduction

We first state our main theorem. Related results and applications are discussed at the end of the section.

Basic setup. For $b \geq 2$, let $T = (V, E)$ be the infinite b-ary tree rooted at ρ. Denote by T_n the first $n \geq 0$ levels of T. Let $M = (M_{ij})_{i,j=1}^k$ be a $k \times k$ irreducible stochastic matrix with stationary distribution $\pi > 0$. Assume M has a

[*] Keywords: Markov models on trees, reconstruction, Kesten-Stigum bound, exponential moment

[†] Theory Group, Microsoft Research.

[‡] Department of Mathematics, UCLA.

real second-largest (in absolute value) eigenvalue λ and let $\nu \neq 0$ be a real right eigenvector corresponding to λ with

$$\sum_{i=1}^{k} \pi_i \nu_i^2 = 1.$$

Let $[k] = \{1, \ldots, k\}$. Consider the following Markov process on T: pick a root state ξ_ρ in $[k]$ according to π; moving away from the root, apply the channel M to each edge independently. Denote by $(\xi_v)_{v \in V}$ the state assignment so obtained and let

$$\sigma_v = \nu_{\xi_v},$$

for all $v \in V$

Reconstruction. In the so-called "reconstruction problem," one seeks—roughly speaking—to infer the state at the root from the states at level n, as $n \to \infty$. This problem has been studied extensively in probability theory and statistical physics. See e.g. [EKPS00] for background and references. Here, we are interested in the following root-state estimator introduced in [MP03]. For $n \geq 0$, let L_n be the vertices of T at level n. Consider the following quantity

$$S_n = \frac{1}{(b\lambda)^n} \sum_{x \in L_n} \sigma_x. \tag{1}$$

It is easy to show that for all $n \geq 0$

$$\mathbb{E}[S_n \,|\, \xi_\rho] = \sigma_\rho,$$

that is, S_n is "unbiased." Moreover, it was shown in [MP03] that in the so-called Kesten-Stigum reconstruction phase, that is, when $b\lambda^2 > 1$, it holds that for all $n \geq 0$

$$\max_i \mathbb{E}[S_n^2 \,|\, \xi_\rho = i] \leq C < +\infty,$$

where $C = C(M)$ is a constant depending only on M (not on n).

Main results. For $n \geq 0$, $i = 1, \ldots, k$, and $\zeta \in \mathbb{R}$, let

$$\Gamma_n^i(\zeta) = \mathbb{E}[e^{\zeta S_n} \,|\, \xi_\rho = i],$$

and

$$\widetilde{\Gamma}_n^i(\zeta) = \mathbb{E}[e^{\zeta S_n^2} \,|\, \xi_\rho = i].$$

We prove the following.

Theorem 1 (Exponential Moment Bound) *Assume M is such that $b\lambda^2 > 1$. Then, there is $c = c(M) < +\infty$ such that for all $n \geq 0$, $i = 1, \ldots, k$, and $\zeta \in \mathbb{R}$, it holds that*

$$\Gamma_n^i(\zeta) \leq e^{\nu_i \zeta + c\zeta^2} < +\infty.$$

Note that $\nu_i = \mathbb{E}[S_n \,|\, \xi_\rho = i]$.

Corollary 1 *Assume M is such that $b\lambda^2 > 1$. Then, there is $\tilde{\zeta} = \tilde{\zeta}(M) \in (0, +\infty)$ and $\widetilde{C} = \widetilde{C}(M) < +\infty$ such that for all $n \geq 0$, $i = 1, \ldots, k$, and $\zeta \in (-\tilde{\zeta}, \tilde{\zeta})$, it holds that*

$$\widetilde{\Gamma}_n^i(\zeta) \leq \widetilde{C} < +\infty.$$

The proofs of Theorem 1 and Corollary 1 can be found in Section 2.

Related results. Moment-generating functions of random variables similar to (1) have been studied in the context of multi-type branching processes. In particular, Athreya and Vidyashankar [AV95] have obtained large-deviation results for quantities of the type (in our setting)

$$R_n = b^{-n} Z_n \cdot w - \pi \cdot w,$$

where $w \in \mathbb{R}^k$ and $Z_n = (Z_n^{(1)}, \ldots, Z_n^{(k)})$ is the "census" vector, that is,

$$Z_n^{(i)} = |\{x \in L_n \,:\, \xi_x = i\}|,$$

for all $i \in [k]$. However, note that we are interested in the *degenerate* case $w = \nu \perp \pi$ (see e.g. [HJ85]) and our results cannot be deduced from [AV95].

Note moreover that our bounds cannot hold when $b\lambda^2 < 1$. Indeed, in that case, a classical CLT of Kesten and Stigum [KS66] for multi-type branching processes implies that the quantity

$$Q_n \equiv (b\lambda^2)^{n/2} S_n = \frac{1}{b^{n/2}} \sum_{x \in L_n} \sigma_x,$$

converges in distribution to a centered Gaussian with a finite variance (independently of the root state). See [MP03] for more on the Kesten-Stigum CLT and its relation to the reconstruction problem.

Motivation. The motivation behind our results comes from mathematical biology. More particularly, our main theorem has recently played a role in the solution of important questions in mathematical phylogenetics, which we now briefly discuss.

As mentioned above, the quantity S_n arises naturally in the reconstruction problem as a simple "linear" estimator of the root state [EKPS00, MP03]. In the past few years, deep connections have been established between the reconstruction problem and the inference of phylogenies—a central problem in computational biology [SS03, Fel04]. A phylogeny is a tree representing the evolutionary history of a group of organisms, where the leaves are modern species and the branchings correspond to past speciation events. To reconstruct phylogenies, biologists extract (aligned) biomolecular sequences from extant species. It is standard in evolutionary biology to model such collections of sequences as *independent samples from the leaves of a Markov chain on a finite tree*

$$\mathbb{S} = \{(\sigma_x^i)_{x \in L_n}\}_{i=1}^{\ell}, \tag{2}$$

where ℓ is the sequence length. The goal of phylogenetics is to infer the *leaf-labelled* tree that generated these samples. In particular, developing reconstruction techniques that require as few samples as possible is of practical importance.

An insightful conjecture of Steel [Ste01] suggests that the reconstruction of phylogenies can be achieved from much shorter sequences when the reconstruction problem is "solvable," in particular in the Kesten-Stigum reconstruction phase. This conjecture has been established in the binary symmetric case (equivalent to the ferromagnetic Ising model), that is, the case $k = 2$ and M symmetric, by Mossel [Mos04] and Daskalakis et al. [DMR09]. The main idea behind these results is to "boost" standard tree-building techniques by inferring ancestral sequences. See [Mos04, DMR09] for details.

Establishing Steel's conjecture under more realistic models of sequence evolution (i.e., more general transition matrices M) is a major open problem in mathematical phylogenetics. Roughly, to reconstruct a phylogeny from samples at level n one iteratively joins the most correlated pairs of nodes, starting from level n and moving towards the root. To estimate the correlation between *internal* nodes u and v on level $m < n$ using only (2) it is natural to consider quantities such as

$$\widehat{\mathrm{Cov}}[u, v] = \frac{1}{\ell} \sum_{i=1}^{\ell} \left((b\lambda)^{-(n-m)} \sum_{x \in L_n^u} \sigma_x^i \right) \left((b\lambda)^{-(n-m)} \sum_{x \in L_n^v} \sigma_x^i \right), \tag{3}$$

where L_n^u is the set of nodes on level n below u. In words, we estimate the correlation between the *reconstructed* states at u and v. Proving concentration of such quantities necessitates uniform bounds on the moment-generating functions of S_n and S_n^2—our main result. We note in particular that our main theorem was recently used by Roch [Roc09], building on [Roc08], to prove Steel's conjecture for general k and reversible transition matrices of the form $M = e^{tQ}$ in the Kesten-Stigum phase. Moreover, this result was established using a surprisingly simple algorithm known in phylogenetics as a "distance-based method," thereby contradicting a conjecture regarding the weakness of this widely used class of methods. See [Roc08] for background.

Organization. The proof of our results can be found in Section 2.

2 Proof

We first prove our main theorem in a neighbourhood around zero.

Lemma 1 *Assume M is such that $b\lambda^2 > 1$. Then, there is $c' = c'(M) < +\infty$ and $\zeta_0 \in (0, +\infty)$ such that for all $n \geq 0$, $i = 1, \ldots, k$, and $|\zeta| < \zeta_0$, it holds that*

$$\Gamma_n^i(\zeta) \leq e^{\nu_i \zeta + c'\zeta^2}.$$

Proof: We prove the result by induction on n. For $n = 0$, note that

$$\Gamma_0^i(\zeta) = e^{\nu_i \zeta},$$

so the first step of the induction holds for all $c' > 0$ and all $\zeta \in \mathbb{R}$.

Now assume the result holds for $n > 0$ with c' and ζ_0 to be determined later. For $n \geq 0$, $i = 1, \ldots, k$, and $\zeta \in \mathbb{R}$, let

$$\gamma_n^i(\zeta) = \ln \Gamma_n^i(\zeta).$$

Let $\alpha_1, \ldots, \alpha_b$ be the children of ρ and, for $\omega = 1, \ldots, b$, denote by L_{n+1}^ω the descendants of α_ω on the $n+1$'st level. For $\omega = 1, \ldots, b$, let

$$S_{n+1}^\omega = \frac{1}{(b\lambda)^n} \sum_{x \in L_{n+1}^\omega} \sigma_x.$$

Note that conditioned on ξ_ρ, the random vectors

$$(\xi_x)_{x \in L_{n+1}^1}, \ldots, (\xi_x)_{x \in L_{n+1}^b},$$

are independent and identically distributed. Hence, the variables

$$S_{n+1}^1, \ldots, S_{n+1}^b,$$

are also conditionally independent and identically distributed. Applying the channel to the first level of the tree and using the induction hypothesis, we have for $\zeta \in (-\zeta_0, \zeta_0)$

$$\begin{aligned}
\gamma^i_{n+1}(\zeta) &= \ln \mathbb{E}[e^{\zeta S_{n+1}} \,|\, \xi_\rho = i] \\
&= \ln \mathbb{E}\left[\exp\left(\frac{\zeta}{b\lambda}\sum_{\omega=1}^{b} S^\omega_{n+1}\right) \,\Big|\, \xi_\rho = i\right] \\
&= b \ln \mathbb{E}\left[\exp\left(\frac{\zeta}{b\lambda} S^1_{n+1}\right) \,\Big|\, \xi_\rho = i\right] \\
&= b \ln \left(\sum_{j=1}^{k} M_{ij} \mathbb{E}\left[\exp\left(\frac{\zeta}{b\lambda} S^1_{n+1}\right) \,\Big|\, \xi_{\alpha_1} = j\right]\right) \\
&= b \ln \left(\sum_{j=1}^{k} M_{ij} \Gamma^j_n\left(\frac{\zeta}{b\lambda}\right)\right) \\
&\leq b \ln \left(\sum_{j=1}^{k} M_{ij} e^{\nu_j(\frac{\zeta}{b\lambda}) + c'(\frac{\zeta}{b\lambda})^2}\right),
\end{aligned}$$

where we used that by assumption

$$|b\lambda| \geq \frac{1}{|\lambda|} \geq 1,$$

so that $\zeta/(b\lambda) \in (-\zeta_0, \zeta_0)$. By a Taylor expansion, as ζ_0 goes to zero (in particular $\zeta_0 < 1$), we have

$$\begin{aligned}
\gamma^i_{n+1}(\zeta) &\leq c' \frac{\zeta^2}{b\lambda^2} \\
&\quad + b \ln \left(\sum_{j=1}^{k} M_{ij} \left[1 + \nu_j\left(\frac{\zeta}{b\lambda}\right) + \frac{1}{2}\nu_j^2\left(\frac{\zeta}{b\lambda}\right)^2 + |\zeta|^3\right]\right) \\
&\leq c' \frac{\zeta^2}{b\lambda^2} \\
&\quad + b \ln \left(1 + \lambda\nu_i\left(\frac{\zeta}{b\lambda}\right) + \frac{1}{2}\|\nu\|^2_\infty\left(\frac{\zeta}{b\lambda}\right)^2 + |\zeta|^3\right) \\
&\leq \nu_i\zeta + \left\{c' + \frac{1}{2}\|\nu\|^2_\infty\right\}\frac{\zeta^2}{b\lambda^2} - \frac{1}{2}\frac{\nu_i^2\zeta^2}{b} + O_{\zeta_0}(|\zeta|^3) \\
&\leq \nu_i\zeta + \left\{c' + \frac{1}{2}\|\nu\|^2_\infty\right\}\frac{\zeta^2}{b\lambda^2} + O_{\zeta_0}(|\zeta|^3).
\end{aligned}$$

Choose $c' > 0$ large enough so that

$$c' > \left\{ c' + \frac{1}{2}\|\nu\|_\infty^2 \right\} \frac{1}{b\lambda^2},$$

that is,

$$c' > \frac{\|\nu\|_\infty^2}{2b\lambda^2}\left(1 - \frac{1}{b\lambda^2}\right)^{-1}.$$

Note that c' is well defined when $b\lambda^2 > 1$. Then there is $\zeta_0 \in (0, +\infty)$ such that for all $\zeta \in (-\zeta_0, \zeta_0)$

$$\gamma_{n+1}^i(\zeta) \leq \nu_i \zeta + c'\zeta^2.$$

That concludes the proof. ■

The following lemma deals with values of ζ away from zero.

Lemma 2 *Assume M is such that $b\lambda^2 > 1$. Let $\zeta_0 \in (0, +\infty)$ be as in Lemma 1. Then, there is $c'' = c''(M) < +\infty$ such that for all $n \geq 0$, $i = 1, \dots, k$, and $|\zeta| \geq \zeta_0$, it holds that*

$$\Gamma_n^i(\zeta) \leq e^{c''\zeta^2}.$$

Proof: Let c' be as in Lemma 1. Let $\zeta_1 \in (0, +\infty)$ be such that

$$\zeta_1 < \frac{\zeta_0}{|b\lambda|}. \tag{4}$$

Choose $c'' > c'$ large enough so that

$$e^{\nu_i\zeta + c'\zeta^2} \leq e^{c''\zeta^2}, \tag{5}$$

for all $|\zeta| > \zeta_1$ and for all $i = 1, \dots, k$.

Let $n \geq 0$ and ζ with $|\zeta| \geq \zeta_0$ be fixed. Note that, when we relate the exponential moment at level m to that at level $m-1$ with a recursion as in the proof of Lemma 1, the value of ζ is effectively divided by $b\lambda$. Therefore, there are two cases in the proof: either we reach the interval $(-\zeta_0, \zeta_0)$ by the time we reach $m = 0$ in the recursion; or we do not.

1. First assume that

   $$\left|\frac{\zeta}{(b\lambda)^n}\right| \geq \zeta_0, \tag{6}$$

   that is, we do not reach $(-\zeta_0, \zeta_0)$. We prove the result by induction on the level $m = 0, \dots, n$. At $m = 0$, we have

   $$\Gamma_0^i\left(\frac{\zeta}{(b\lambda)^n}\right) = e^{\nu_i(\frac{\zeta}{(b\lambda)^n})} \leq e^{c''(\frac{\zeta}{(b\lambda)^n})^2},$$

by (5) and (6) for all $i = 1, \ldots, k$. Assume for the sake of the induction that

$$\Gamma_m^i \left(\frac{\zeta}{(b\lambda)^{n-m}} \right) \leq e^{c''(\frac{\zeta}{(b\lambda)^{n-m}})^2},$$

for all $i = 1, \ldots, k$. Using the calculations of Lemma 1, we have

$$\begin{aligned}
\gamma_{m+1}^i \left(\frac{\zeta}{(b\lambda)^{n-(m+1)}} \right) &= b \ln \left(\sum_{j=1}^k M_{ij} \Gamma_m^j \left(\frac{1}{b\lambda} \frac{\zeta}{(b\lambda)^{n-(m+1)}} \right) \right) \\
&\leq b \ln \left(\sum_{j=1}^k M_{ij} e^{c''(\frac{\zeta}{(b\lambda)^{n-m}})^2} \right) \\
&= bc'' \left(\frac{\zeta}{(b\lambda)^{n-m}} \right)^2 \\
&= \frac{b}{b^2\lambda^2} c'' \left(\frac{\zeta}{(b\lambda)^{n-(m+1)}} \right)^2 \\
&\leq c'' \left(\frac{\zeta}{(b\lambda)^{n-(m+1)}} \right)^2,
\end{aligned}$$

where we used $b\lambda^2 > 1$ on the last line. The proof of the first case follows by induction, that is, we have

$$\Gamma_n^i(\zeta) \leq e^{c''\zeta^2},$$

for all $i = 1, \ldots, k$.

2. Assume now that

$$\left| \frac{\zeta}{(b\lambda)^n} \right| < \zeta_0. \tag{7}$$

Let m^* be the largest value in $0, \ldots, n$ such that

$$\left| \frac{\zeta}{(b\lambda)^{n-m^*}} \right| < \zeta_0. \tag{8}$$

The purpose of Assumption (4) above is to make sure that we never “jump” entirely over the subset of $(-\zeta_0, \zeta_0)$ where (5) holds. Indeed, by (4) and

$$\left| \frac{\zeta}{(b\lambda)^{n-(m^*+1)}} \right| \geq \zeta_0, \tag{9}$$

it follows that we must also have

$$\left|\frac{\zeta}{(b\lambda)^{n-m^*}}\right| > \zeta_1. \tag{10}$$

Hence, by (5) and Lemma 1, we get

$$\Gamma^i_{m^*}\left(\frac{\zeta}{(b\lambda)^{n-m^*}}\right) \leq e^{c''(\frac{\zeta}{(b\lambda)^{n-m^*}})^2},$$

for all $i = 1, \ldots, k$. The proof then follows by induction as in the first case above.

■

Proof of Theorem 1: Let ζ_0, c' and c'' be as in Lemmas 1 and 2. Choose $c > c''(> c')$ large enough so that

$$e^{c''\zeta^2} \leq e^{\nu_i\zeta + c\zeta^2}, \tag{11}$$

for all $|\zeta| \geq \zeta_0$ and for all $i = 1, \ldots, k$. The result then follows by combining Lemmas 1 and 2. ■

Proof of Corollary 1: We use a standard trick relating the exponential moment of the square to that of a Gaussian. Let X be a standard normal. Using Theorem 1 and applying Fubini we have for all $n \geq 0$ and $i = 1, \ldots, k$

$$\begin{aligned} \mathbb{E}[e^{\zeta S_n^2} \,|\, \xi_\rho = i] &= \mathbb{E}[e^{\sqrt{2\zeta} S_n X} \,|\, \xi_\rho = i] \\ &\leq \mathbb{E}[e^{\nu_i\sqrt{2\zeta} X + c2\zeta X^2} \,|\, \xi_\rho = i]. \end{aligned}$$

The last expectation is finite for ζ small enough. ■

References

[AV95] K. B. Athreya and A. N. Vidyashankar. Large deviation rates for branching processes. II. The multitype case. *Ann. Appl. Probab.*, 5(2):566–576, 1995.

[DMR09] Constantinos Daskalakis, Elchanan Mossel, and Sébastien Roch. Evolutionaty trees and the Ising model on the Bethe lattice: a proof of Steel's conjecture. Preprint, 2009.

[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.

[Fel04] J. Felsenstein. *Inferring Phylogenies*. Sinauer, New York, New York, 2004.

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

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

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

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

[Roc08] Sébastien Roch. Sequence-length requirement for distance-based phylogeny reconstruction: Breaking the polynomial barrier. In *FOCS*, pages 729–738, 2008.

[Roc09] Sébastien Roch. Phase transition in distance-based phylogeny reconstruction. Preprint, 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.