RECONSTRUCTION ON TREES: EXPONENTIAL MOMENT BOUNDS FOR LINEAR ESTIMATORS

YUVAL PERES
Theory Group, Microsoft Research, Redmond, WA.
email: peres@microsoft.com

SEBASTIEN ROCH[1]
Department of Mathematics, University of California-Los Angeles, Los Angeles, CA.
email: roch@math.ucla.edu

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 $[k] = \{1, \dots, k\}$. 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 $v \neq 0$. Letting $\sigma_v = v_{\xi_v}$, we consider the following root-state estimator, which was introduced by Mossel and Peres (2003) in the context of the "reconstruction 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.

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

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

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 transition matrix M to each edge independently. Denote by $(\xi_v)_{v \in V}$ the state assignment so obtained and let

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

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 $n \geq 0$ and $x \in L_n$,

$$\mathbb{E}[\sigma_x \mid \xi_\rho] = \lambda^n \sigma_\rho, \tag{2}$$

and, hence,

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

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 $n \geq 0$

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

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

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

and

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

We prove the following. See below for motivation arising from computational biology.

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*

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

Note that $v_i = \mathbb{E}[S_n \mid \xi_\rho = i]$.

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.

Motivation. Our main theorem was recently used to solve an important mathematical biology problem which we now briefly discuss. As explained above, the quantity S_n arises naturally as a "linear" estimator of the root state of the Markov chain [EKPS00, MP03]. In the past few years, deep connections have been established between this "root" 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 (typically) modern species and the branchings correspond to past speciation events. To reconstruct phylogenies, biologists extract DNA sequences from extant species. It is standard in evolutionary biology to model such collections of "aligned" sequences as ℓ *i.i.d. samples from the leaves of a Markov chain on a finite tree*

$$\mathbb{S} = \{(\sigma_x^i)_{x \in L_n}\}_{i=1}^{\ell}, \tag{3}$$

where ℓ is the sequence length. In words we assume that the DNA of the ancestral species represented by the root of the tree is inherited by descendant species, up to random substitutions occuring along the branches of the tree. In this model, we ignore population-level variation and consider only a "reference" genome. Each site or position of this reference genome is identified with a sample of the Markov chain. That is, the number of samples corresponds to the length of the DNA sequence. The independence assumption is made for convenience. We also ignore other mutational events, such as insertions and deletions, which are dealt with separately through a pre-processing phase of multiple sequence alignment. The goal of the phylogenetic reconstruction problem 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. Note that, in general, the tree (which is deterministic but unknown) may not be complete and the Markov transition matrix may differ across edges.

Classical approaches for reconstructing phylogenies are typically computationally intractable [GF82, DS86, Day87, CT06, Roc06] or they require impractical sequence lengths [Att99, LC06, SS99, SS02]. Over the past two decades, however, much progress has been made in the design of computationally efficient reconstruction techniques which require shorter sequence lengths, starting with the seminal work of Erdös et al. [ESSW99].

The algorithm in [ESSW99], often dubbed the Short Quartet Method (SQM), is based on well-known distance-matrix techniques which rely on the "evolutionary distance" between each pair of species, that is, roughly an estimate of the time elapsed since their most recent common ancestor. In our notation, suppose $x, y \in L_n$ are separated by $2m$ edges and have lowest common ancestor $z \in L_{n-m}$. Then, using (2) and conditioning on the state at z, we see that

$$\mathbb{E}[\sigma_x^1 \sigma_y^1] = \lambda^{2m},$$

and, hence,

$$\hat{d}(x, y) = -\ln \frac{1}{\ell} \sum_{i=1}^{\ell} \sigma_x^i \sigma_y^i \to [\ln \lambda^{-2}]m, \tag{4}$$

almost surely, as $\ell \to +\infty$. In other words, $\hat{d}(x, y)$ is a consistent estimator of the number of edges separating x and y from z up to a multiplicative constant. From all such pairwise distances, it is a simple matter to reconstruct the complete binary tree—iteratively merge all closest pairs. Algorithms have been designed to deal efficiently with more general trees.

Unlike "naive" distance-matrix methods, however, the key behind SQM is that it discards "long" evolutionary distances whose estimates from sequence comparisons as in (4) are known to be statistically unreliable. For instance, in the two-state symmetric case, that is, when $k = 2$ and M symmetric, we have $v = (+1, -1)$ and the signal-to-noise ratio for x and y as above is given by

$$\frac{\mathbb{E}[\sigma_x^1 \sigma_y^1]}{\sqrt{\text{Var}[\sigma_x^1 \sigma_y^1]}} = \frac{\lambda^{2m}}{\sqrt{1 - \lambda^{4m}}} \to 0, \tag{5}$$

as $m \to +\infty$. The SQM algorithm works by first building subtrees of small diameter and, in a second stage, glueing the pieces back together. In fact this step is not needed for the complete binary tree, but it is crucial for more general trees. The algorithm is then guaranteed to return the correct topology with high probability in polynomial time from sequences that grow *polynomially* with the number of leaves of the tree.

Another series of theoretical improvements was triggered by an insightful conjecture of Steel claiming that the reconstruction of phylogenies should be feasible from *much shorter sequences* when the corresponding root-state reconstruction problem is "solvable," in particular, in the Kesten-Stigum reconstruction phase $b\lambda^2 > 1$ considered here. Intuitively, the fact that more information about internal sequences "diffuses" to the leaves should make phylogeny reconstruction easier. The conjecture was established in the two-state symmetric case by Mossel [Mos04] and Daskalakis et al. [DMR11]. More precisely it was shown that, when the Markov transition matrix on each branch satisfies the Kesten-Stigum bound, the phylogeny can be reconstructed in polynomial time from sequences that grow only *logarithmically* (instead of polynomially) with the number of leaves. The main idea of the proof is the "boosting" of standard tree-building techniques through the inference of ancestral sequences. In the case of the complete binary tree, one could proceed as follows:

1. Merge all closest pairs of leaves.

2. Infer ancestral sequences at the parents of the merged pairs.

3. Use those ancestral sequences to merge all closest pairs of parents.

4. And so on.

To be more precise, in the third step we estimate the evolutionary distance between *internal* nodes u and v on level $n' < n$ using

$$
\begin{aligned}
\hat{\hat{d}}(u, v) &= -\ln \frac{1}{\ell} \sum_{i=1}^{\ell} \left((b\lambda)^{-(n-n')} \sum_{x \in L_n^u} \sigma_x^i \right) \left((b\lambda)^{-(n-n')} \sum_{x \in L_n^v} \sigma_x^i \right) \\
&\equiv -\ln \frac{1}{\ell} \sum_{i=1}^{\ell} S_{n-n'}^{(u),i} S_{n-n'}^{(v),i},
\end{aligned} \tag{6}
$$

where L_n^u is the set of nodes on level n below u—assuming for simplicity that λ is known. In words, we estimate the correlation between the reconstructed sequences at u and v. The key is that, as long as the ancestral sequences are sufficiently well correlated with the true sequences,

one only needs to estimate evolutionary distances between pairs of nodes (possibly, internal) with high signal-to-noise ratio in (5). A more complex algorithm was developed in [DMR11] to deal with more general trees.

This "boosted" reconstruction algorithm performs a polynomial number of evolutionary distance estimation, each of which therefore has to be accurate with inverse polynomial probability (in the number of leaves). In order to achieve such success with a logarithmic number of samples, one needs exponential concentration on the quantity (6). Proving such concentration necessitates uniform bounds on the moment-generating functions of S_n and S_n^2—our main result. Indeed, using the notation introduced in (6), by our main result

$$
\begin{aligned}
\mathbb{E}\left[\exp(\zeta S_{n-n'}^{(u),1} S_{n-n'}^{(v),1})\right] &= \mathbb{E}\left[\mathbb{E}\left[\exp(\zeta S_{n-n'}^{(u),1} S_{n-n'}^{(v),1})\,|\,\sigma_u^1, \sigma_v^1\right]\right] \\
&\leq \mathbb{E}\left[\mathbb{E}\left[\exp(\sigma_v^1\{\zeta S_{n-n'}^{(u),1}\} + c\{\zeta S_{n-n'}^{(u),1}\}^2)\,|\,\sigma_u^1, \sigma_v^1\right]\right],
\end{aligned}
$$

which, by the Cauchy-Schwarz inequality and our corollary, is finite for small enough ζ. We can therefore apply classical large deviations techniques to obtain exponential concentration on the estimator $\hat{d}$. Our main theorem is proved by induction on the number of levels and the particular form of upper bound we consider allows the recursion to go through.

We note that our main result was recently used by one of us [Roc10], 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.

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 $Z_n = (Z_n^{(1)}, \dots, Z_n^{(k)})$ 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 $w = v \perp \pi$ (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.

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 $n \geq 0$, $i = 1, \dots, k$, and $|\zeta| < \zeta_0$, it holds that*

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

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

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

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 $n \geq 0$, $i = 1, \dots, k$, and $\zeta \in \mathbb{R}$, let

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

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

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

are independent and identically distributed. Hence, the variables

$$S_{n+1}^{(1)}, \dots, S_{n+1}^{(b)},$$

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

$$
\begin{aligned}
\gamma_{n+1}^i(\zeta) &= \ln \mathbb{E}[e^{\zeta S_{n+1}} \mid \xi_\rho = i] \\
&= \ln \mathbb{E}\left[\exp\left(\frac{\zeta}{b\lambda} \sum_{\omega=1}^b S_{n+1}^{(\omega)} \right) \Big| \xi_\rho = i \right] \\
&= b \ln \mathbb{E}\left[\exp\left(\frac{\zeta}{b\lambda} S_{n+1}^{(1)} \right) \Big| \xi_\rho = i \right] \\
&= b \ln \left(\sum_{j=1}^k M_{ij} \mathbb{E}\left[\exp\left(\frac{\zeta}{b\lambda} S_{n+1}^{(1)} \right) \Big| \xi_{\alpha_1} = j \right] \right) \\
&= b \ln \left(\sum_{j=1}^k M_{ij} \Gamma_n^j \left(\frac{\zeta}{b\lambda} \right) \right) \\
&\leq b \ln \left(\sum_{j=1}^k M_{ij} e^{v_j \left(\frac{\zeta}{b\lambda} \right) + c' \left(\frac{\zeta}{b\lambda} \right)^2} \right),
\end{aligned}
$$

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

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

Choose $c' > 0$ large enough so that

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

that is,

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

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

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

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 $n \geq 0$, $i = 1, \ldots, k$, 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

$$\zeta_1 < \frac{\zeta_0}{|b\lambda|}. \tag{7}$$

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

$$e^{v_i \zeta + c' \zeta^2} \leq e^{c'' \zeta^2}, \tag{8}$$

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

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

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

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

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

by (8) and (9) for all $i = 1, \ldots, k$. Assume for the sake of the induction that

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

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

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

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

$$\left| \frac{\zeta}{(b\lambda)^n} \right| < \zeta_0. \tag{10}$$

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

$$\left| \frac{\zeta}{(b\lambda)^{n-m^*}} \right| < \zeta_0. \tag{11}$$

The purpose of Assumption (7) above is to make sure that we never "jump" entirely over the subset of $(-\zeta_0, \zeta_0)$ where (8) holds. Indeed, by (7) and

$$\left| \frac{\zeta}{(b\lambda)^{n-(m^*+1)}} \right| \geq \zeta_0, \tag{12}$$

it follows that we must also have

$$\left| \frac{\zeta}{(b\lambda)^{n-m^*}} \right| > \zeta_1. \tag{13}$$

Hence, by (8) 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.

$\square$

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

$$e^{c''\zeta^2} \leq e^{v_i\zeta + c\zeta^2}, \tag{14}$$

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 defined on a joint probability space with our Markov chain, but independent from it. Using Theorem 1 and applying Fubini we have for all $n \geq 0$ and $i = 1, \ldots, k$

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

The last expectation is finite for ζ small enough. ∎

References

[Att99]   K. Atteson. The performance of neighbor-joining methods of phylogenetic reconstruction. *Algorithmica*, 25(2-3):251–278, 1999. MR1703580

[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. MR1336883

[CT06]   Benny Chor and Tamir Tuller. Finding a maximum likelihood tree is hard. *J. ACM*, 53(5):722–744, 2006. MR2263067

[Day87]   William H. E. Day. Computational complexity of inferring phylogenies from dissimilarity matrices. *Bull. Math. Biol.*, 49(4):461–467, 1987. MR0908160

[DMR11]   Constantinos Daskalakis, Elchanan Mossel, and Sébastien Roch. Evolutionaty trees and the Ising model on the Bethe lattice: a proof of Steel's conjecture. *Probability Theory and Related Fields*, 149(1-2):149–189, 2011.

[DS86]   William H. E. Day and David Sankoff. Computational complexity of inferring phylogenies by compatibility. *Syst. Zool.*, 35(2):224–229, 1986.

[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. MR1768240

[ESSW99]   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. MR1667319

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

[GF82]   R. L. Graham. and L. R. Foulds. Unlikelihood that minimal phylogenies for a realistic biological study can be constructed in reasonable computational time. *Math. Biosci.*, 60:133–142, 1982. MR0678718

[HJ85]   Roger A. Horn and Charles R. Johnson. *Matrix analysis*. Cambridge University Press, Cambridge, 1985. MR0832183

[KS66]   H. Kesten and B. P. Stigum. Additional limit theorems for indecomposable multidimensional Galton-Watson processes. *Ann. Math. Statist.*, 37:1463–1481, 1966. MR0200979

[LC06]   Michelle R. Lacey and Joseph T. Chang. A signal-to-noise analysis of phylogeny estimation by neighbor-joining: insufficiency of polynomial length sequences. *Math. Biosci.*, 199(2):188–215, 2006. MR2211625

[Mos04]   E. Mossel. Phase transitions in phylogeny. *Trans. Amer. Math. Soc.*, 356(6):2379–2404, 2004. MR2048522

[MP03]   E. Mossel and Y. Peres. Information flow on trees. *Ann. Appl. Probab.*, 13(3):817–844, 2003. MR1994038

[Roc06]   Sébastien Roch. A short proof that phylogenetic tree reconstruction by maximum likelihood is hard. *IEEE/ACM Trans. Comput. Biology Bioinform.*, 3(1):92–94, 2006.

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

[Roc10]   Sebastien Roch. Toward Extracting All Phylogenetic Information from Matrices of Evolutionary Distances. *Science*, 327(5971):1376–1379, 2010. MR2650508

[SS99]   Michael A. Steel and László A. Székely. Inverting random functions. *Ann. Comb.*, 3(1):103–113, 1999. Combinatorics and biology (Los Alamos, NM, 1998). MR1769697

[SS02]   M. A. Steel and L. A. Székely. Inverting random functions. II. Explicit bounds for discrete maximum likelihood estimation, with applications. *SIAM J. Discrete Math.*, 15(4):562–575 (electronic), 2002. MR1935839

[SS03]   C. Semple and M. Steel. *Phylogenetics*, volume 22 of *Mathematics and its Applications series*. Oxford University Press, 2003. MR2060009