FAST CONVERGENCE OF MARKOV CHAIN MONTE CARLO ALGORITHMS FOR PHYLOGENETIC RECONSTRUCTION WITH HOMOGENEOUS DATA ON CLOSELY RELATED SPECIES*

DANIEL ŠTEFANKOVIČ AND ERIC VIGODA

Abstract. This paper studies a Markov chain for phylogenetic reconstruction which uses a popular transition between tree topologies known as subtree pruning-and-regrafting (SPR). We analyze the Markov chain in the simpler setting where the generating tree consists of very short edge lengths, short enough so that each sample from the generating tree (or character in phylogenetic terminology) is likely to have only one mutation, and where there are enough samples so that the data looks like the generating distribution. We prove in this setting that the Markov chain is rapidly mixing, i.e., it quickly converges to its stationary distribution, which is the posterior distribution over tree topologies. Our proofs use that the leading term of the maximum likelihood function of a tree T is the maximum parsimony score, which is the size of the minimum cut in T needed to realize single edge cuts of the generating tree. Our main contribution is a combinatorial proof that, in our simplified setting, SPR moves are guaranteed to converge quickly to the maximum parsimony tree. Our results are in contrast to recent works showing examples with heterogeneous data (namely, the data is generated from a mixture distribution) where many natural Markov chains are exponentially slow to converge to the stationary distribution.

1. Introduction. We study Markov chain Monte Carlo (MCMC) methods for Bayesian inference of phylogeny. We begin by presenting the relevant background material by defining phylogenetic trees, evolutionary models (in section 1.1), and the associated MCMC methods (in section 1.2). We refer the interested reader to Semple and Steel [15] for a more comprehensive introduction to the mathematics of phylogeny. Finally, we present our results and discuss related work in section 1.3.

A phylogenetic tree is an unrooted tree T on n leaves (called taxa, corresponding to n species) where internal vertices have degree three. Let $E(T)$ denote the edges of T and $V(T)$ denote the vertices. In the phylogenetic reconstruction problem, we observe a collection of labelings of the leaves of T from a set $\mathbf{\Omega}$, and our goal is to infer the tree T from which they were generated. For example, if $\mathbf{\Omega} = \{A, C, G, T\}$, then we are given (aligned) DNA sequences for n species, and we are trying to determine the tree describing the evolutionary history of the present-day species.

1.1. Evolutionary models and maximum likelihood. The labelings on the leaves of T are the projection of labelings on all vertices of T, and these labelings of V are generated in the following manner. There is a stochastic process along edges of T (e.g., modeling the evolutionary process of DNA substitutions) which is defined by a continuous-time Markov chain. Thus, for each edge $e \in T$ there is an $|\mathbf{\Omega}| \times |\mathbf{\Omega}|$ rate matrix Q and a time $\mathbf{t}_e > 0$, which is called the branch length of e. In this paper,

as is typical in the phylogenetic setting, we assume there is a single rate matrix Q that is common to all edges. The rate matrix is assumed to be reversible with respect to some distribution π on Ω. Hence, fix π as the stationary vector for Q (i.e., $\pi Q = 0$). (The matrix Q is usually scaled so that we expect one "substitution" (i.e., change) per unit of time.) The rate matrix Q defines a continuous-time Markov chain, and together with $\mathbf{t}_e$ defines a transition matrix on edge e:

$$P_e = \exp(\mathbf{t}_e Q) = I + \mathbf{t}_e Q + \mathbf{t}_e^2 Q^2/2! + \mathbf{t}_e^3 Q^3/3! + \dots . \tag{1.1}$$

The matrix P_e is a stochastic matrix of size $|\Omega| \times |\Omega|$ and thus defines a discrete-time Markov chain, which is time-reversible with stationary distribution π, i.e., $\pi P_e = \pi$, and $\pi_i (P_e)_{ij} = \pi_j (P_e)_{ji}$ (for all i, $j \in \Omega$).

The simplest four-state (i.e., $|\Omega| = 4$) evolutionary model has a single parameter for the off-diagonal entries of the rate matrix Q; this model is known as the Jukes–Cantor model. The most general reversible four-state model is the general time-reversible model. For $|\Omega| = 2$ (often studied for mathematical interest), the model is binary and the rate matrix has a single parameter; this model is known as the Cavender–Farris–Neyman model. See Felsenstein [6] or Yang [22] for an introduction to these evolutionary models.

Given T, the rate matrix Q, and the branch lengths $\mathbf{t} = (\mathbf{t}_e)_{e \in E(T)}$, we then define the following distribution on labelings of the vertices of T. Let $P_e = \exp(\mathbf{t}_e Q)$ for $e \in E(T)$. We first orient the edges of T away from an arbitrarily chosen root r of the tree. (We can choose the root arbitrarily since each P_e is reversible with respect to π.) Then the probability of a labeling $\ell \colon V(T) \to \Omega$ is

$$\mu'_{T,Q,\mathbf{t}}(\ell) \coloneqq \pi(\ell(r)) \prod_{\overrightarrow{uv} \in E(T)} P_{uv}(\ell(u), \ell(v)). \tag{1.2}$$

The distribution $\mu'_{T,Q,\mathbf{t}}$ can be generated in an equivalent algorithmic manner. Choose $\ell(r)$ from π. Then for each edge $e = (u, v) \in E(T)$, given an assignment for exactly one of the endpoints, say, $\ell(u)$, choose $\ell(v)$ from the distribution defined by the row of P_e corresponding to the label $\ell(u)$.

Let $\mu_{T,Q,\mathbf{t}}$ be the marginal distribution of $\mu'_{T,Q,\mathbf{t}}$ on the labelings of the leaves of T (thus $\mu_{T,Q,\mathbf{t}}$ is a distribution on Ω^n, where n is the number of leaves of T). Fix T^* with parameters Q^* and $\mathbf{t}^*$ as the generating tree. The goal of phylogeny reconstruction is to reconstruct T^* (and possibly Q^* and $\mathbf{t}^*$) from $\mu_{T^*,Q^*,\mathbf{t}^*}$ (more precisely, from independent samples from $\mu_{T^*,Q^*,\mathbf{t}^*}$).

Let $\mathcal{Q}$ denote a set of rate matrices with nonzero entries where $Q^* \in \mathcal{Q}$. The set $\mathcal{Q}$ is the set of possible rate matrices. The set can be arbitrary; usually it is determined by the model considered (e.g., for the Jukes–Cantor model $\mathcal{Q}$ would contain rate matrices whose off-diagonal entries are the same). One often assumes that the rate matrix Q^* is known. In this case we would set $\mathcal{Q} = \{Q^*\}$. On the other hand, our results also apply if one sets $\mathcal{Q}$ to be the set of all rate matrices with nonzero entries.

We consider the likelihood of a tree T as the maximum over rate matrices $Q \in \mathcal{Q}$ and over assignments of nonzero branch lengths $\mathbf{t}$ to the edges of T of the probability that the tree $(T, Q, \mathbf{t})$ generated μ. More formally, the maximum expected log-likelihood of tree T for distribution μ^* is defined by

$$\mathcal{L}_T(\mu^*) = \sup_{Q \in \mathcal{Q}} \sup_{\mathbf{t}} \mathcal{L}_{T,Q,\mathbf{t}}(\mu^*), \tag{1.3}$$

where

$$\mathcal{L}_{T,Q,\mathbf{t}}(\mu^*) = \sum_{y \in \Omega^n} \mu^*(y) \ln(\mu_{T,Q,\mathbf{t}}(y)). \tag{1.4}$$

For a set of characters $\mathbf{D} = (D_1, \ldots, D_N)$ where $D_i \in \Omega^n$, define the log-likelihood of a tree T as

$$\mathcal{L}_T(\mathbf{D}) = \sup_{Q \in \mathcal{Q}} \sup_{\mathbf{t}} \ln\left(\mu_{T,Q,\mathbf{t}}(\mathbf{D})\right) \quad = \sup_{Q \in \mathcal{Q}} \sup_{\mathbf{t}} \sum_{i=1}^{N} \ln\left(\mu_{T,Q,\mathbf{t}}(D_i)\right).$$

Our goal is to sample from the distribution on the set of phylogenetic trees with n leaves where the weight of a tree is $\mathcal{L}_T(\mathbf{D})$. In section 3 we will look at the straightforward extension to the setting where we are given a prior on trees and parameters Q, $\mathbf{t}$, and our goal is to sample from the posterior distribution.

1.2. Subtree pruning-and-regrafting Markov chain. We analyze a Markov chain using transitions made by subtree pruning-and-regrafting (SPR). SPR transitions are a natural combinatorial transition, which is also popular in practice. In section 4 we discuss several other well-studied choices for the transitions. Here we consider trees weighted by their maximum likelihood. In section 3 we discuss how the Markov chain definition and our main result extends to sampling the posterior distribution.

An SPR transition from a tree T works by choosing an (internal or terminal) edge $e = (u, v)$. If e is an internal edge, we consider one of the two subtrees in $T \setminus e$: either the subtree S_u containing u or the subtree S_v containing v. Let S_u denote the selected subtree. If e is a terminal edge, let S_u be the endpoint of e that is a leaf. Let T' denote the tree formed by removing S_u from T; in particular, we remove S_u and edge e from T and "smooth away" the vertex v (that is, contract one of the two adjacent edges). We then choose an edge e^* in T', and we attach S onto e^* by adding a new intermediate vertex along e^*. See Figure 1.1 for an illustration. Let $SPR(T, S_u, e^*)$ denote the tree resulting from the above transition.

We analyze the following Markov chain, which chooses a random subtree S to prune and then chooses an edge to regraft S along, based on the maximum likelihood of the resulting tree. This Markov chain is analogous to heat bath chains studied in statistical physics (as opposed to Metropolis chains) (e.g., see [4]); thus we refer to the below chain

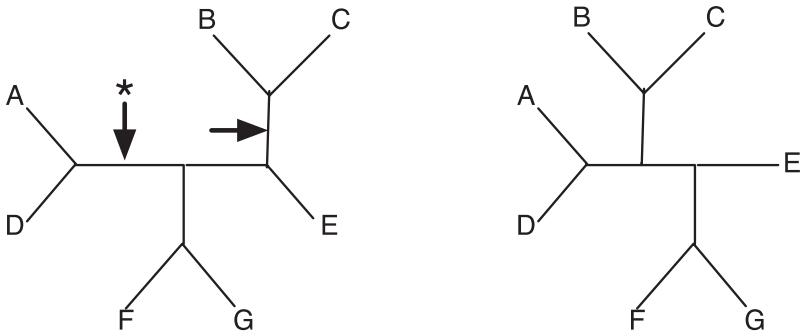

FIG. 1.1. *Illustration of an SPR transition. The randomly chosen edge e is marked by an arrow. The subtree containing B and C is pruned and then regrafted at the edge e^* marked by a starred arrow. The resulting tree is illustrated.*

as the heat bath SPR Markov chain. Here is the formal definition of the transitions $T_t \to T_{t+1}$ of the heat bath SPR Markov chain.

From a tree T_t at time t we proceed as follows:

1. Choose a random subtree S of T_t by choosing a random edge e and then choosing one of the two subtrees hanging off of e. Let T' denote the tree formed by deleting S and e from T.
2. For each edge e^* of tree T', let $w(e^*) = \mathcal{L}_{\hat{T}}(\mathbf{D})$, where $\hat{T} = SPR(T, S, e^*)$ is the tree formed by pruning S from T and regrafting S onto edge e^*. Let ω be the distribution on edges of T', where $\omega(e^*) = w(e^*)/Z$ and $Z = \sum_{e' \in E(T')} w(e')$.
3. Sample an edge e^* from the distribution ω on edges of T'.
4. Graft S onto edge e^* and move to this new tree, i.e., set $T_{t+1} = SPR(T, S, e^*)$.

We now verify that the above Markov chain is ergodic and reversible with respect to the distribution π on trees where $\pi(T) \propto \mathcal{L}_T(\mathbf{D})$, and thus π is also the unique stationary distribution. Let T_1 and T_2 be neighboring states of the Markov chain. Let S be the tree that is pruned and regrafted to obtain T_2 from T_1. Note that the same tree S can be pruned and regrafted to obtain T_1 from T_2. The transition probability from T_1 to T_2 is the probability of choosing S in step 1 times $\mathcal{L}_{T_2}(\mathbf{D})/Z$. Similarly the transition probability from T_2 to T_1 is the probability of choosing S in step 1 times $\mathcal{L}_{T_1}(\mathbf{D})/Z$ (note that Z is the same in both cases since pruning S results in the same tree T'). The detailed balance condition is satisfied for $\pi(T) \propto \mathcal{L}_T(\mathbf{D})$, and hence it is the unique stationary distribution.

Let $d_{\mathrm{TV}}(\mu, \nu)$ denote the (total) variation distance between a pair of probability distributions μ and ν defined on the same finite, discrete space, and let $P^t(T_0, \cdot)$ denote the $t$step distribution of the Markov chain from initial state T_0. The mixing time τ_{mix} is defined as

$$\tau_{\mathrm{mix}} = \max_{T_0} \min\{t : d_{\mathrm{TV}}(P^t(T_0, \cdot), \pi) \leq 1/2\mathrm{e}\},$$

which is the time to reach variation distance $\leq 1/2\mathrm{e}$ of the stationary distribution from the worst initial state. Note it is straightforward to then "boost" so that for any $\delta > 0$, after $\tau_{\mathrm{mix}} \ln(1/\delta)$ steps we are within variation distance $\leq \delta$ of π from the worst initial state (see Aldous [1]).

1.3. Results on MCMC for phylogenetic reconstruction. MCMC algorithms are an important tool for phylogenetic reconstruction. MrBayes [10] is a popular program that relies on MCMC methods for Bayesian inference of phylogeny. MrBayes uses a sophisticated variant of MCMC known as Metropolis-coupled MCMC [8].

For statistical inference problems, such as phylogenetic reconstruction, it is often easy to design appropriate MCMC algorithms, such as the above Markov chain we defined using SPR transitions, which converge in the limit over time to the desired posterior distribution. However, the computational efficiency of these methods relies on their *fast* convergence to the posterior distribution. Since theoretical results are typically lacking, heuristic methods are used to measure convergence to the desired distribution. Hence, there are often no rigorous guarantees on the scientific computations which rely on the random samples produced by the MCMC methods. Our goal is to provide some theoretical understanding of settings where MCMC methods for phylogenetic reconstruction are provably fast and hence yield accurate results and settings where the MCMC methods are slow, and consequently the samples may be misleading.

There are several works with computational experiments on the convergence rates of MCMC algorithms for phylogenetic reconstruction; e.g., see the recent works [3], [11].

There is relatively little theoretical work. Diaconis and Holmes [5] proved fast convergence of a Markov chain to the uniform distribution over phylogenetic trees. Recently, several works have shown examples of heterogeneous data where MCMC algorithms are provably slow to converge. Mossel and Vigoda [13], [14] proved slow convergence for a class of examples with data arising from a uniform mixture of a pair of 5-taxa trees (with different topologies). Štefankovič and Vigoda [19], [20] proved slow convergence for a class of mixture examples from a pair of 5-taxa trees that share the same topology but differ in their branch lengths. In these slow mixing results, the convergence time is exponential in the number of characters (i.e., sequence length).

In this paper we show fast convergence for data from a homogenous source of closely related species. In particular, for data generated from a single tree (of any size) when all the branch lengths are sufficiently short, we prove fast convergence. The requirement of sufficiently short branches is for our proof technique, but it is important to note that the slow mixing results mentioned earlier [13], [14], [19], [20] require, in an analogous manner, sufficiently short branches. If one searches for the tree with the maximum likelihood (or maximum a posteriori probability), our methods show that in our setting of very short branch lengths, the space of trees (connected by the SPR moves) has no local maxima, and hence one can find the optimal tree using hill climbing.

For simplicity, we present our results here for the case where the weight of a tree is the maximum likelihood of generating the given data D where the maximum is over a rate matrix Q (common to all edges) and a set of branch lengths $\mathbf{t}$. This is closely related to the posterior distribution when the priors are uniformly distributed. Our results extend to δ-regular priors, which are priors that are lower bounded by some $\delta > 0$; see section 3 for a discussion on the extension of our results to sampling the posterior distribution. We are interested in the mixing time τ_{mix}, defined as the number of steps until the chain is within variation distance $\leq 1/2\text{e}$ of the stationary distribution.

We prove that the heat bath SPR Markov chain converges quickly to its stationary distribution when the data is generated from a tree T^* where all of the branch lengths are sufficiently small, and there are sufficiently many samples generated from T^*. Here is the formal statement of our main result.

THEOREM 1.1. *Consider any reversible 4-state model, any phylogenetic tree T^* on n taxa, and any rate matrix Q^* with no zero entries. For all $\alpha_{\min} > 0$, there exists $\epsilon_0 > 0$ such that for all $0 < \varepsilon < \varepsilon_0$ and any choice of branch lengths $\mathbf{t}^*_e \in (\alpha_{\min}\epsilon, \epsilon)$ for $e \in E(T^*)$, there exists $N_0 > 0$ where the following holds.*

For a data set with $N > N_0$ characters, each chosen independently from the distribution $\mu_{T^*,Q^*,\mathbf{t}^*}$, then, with probability $> 1 - \exp(-\sqrt{N})$ over the data generated, the heat bath SPR Markov chain has mixing time $\leq 50n$.

Since there is considerable quantification in Theorem 1.1,we will take a moment to dissect it at a high level. First, the requirement that $N > N_0$ comes from needing the data to look very much like the generating distribution $\mu^* = \mu_{T^*,Q^*,\mathbf{t}^*}$. Therefore, how much data we need depends on several quantities, such as the minimum probability of a configuration arising in μ^*, which depends on the minimum branch length and the minimum rate in Q^*. Hence, N_0 depends on $\alpha_{\min}$ and Q^* and is exponential in n. A somewhat related question has been studied by Steel and Székely [16], [17], [18] on how large N needs to be so that the maximum likelihood tree is the generating tree T^*. In their results one also needs N to be exponential in n.

Our proof uses the fact that in the setting of Theorem 1.1, where the branch lengths are sufficiently short, the leading term of the maximum likelihood function is actually maximum parsimony. Such a result is well known in the mathematical phylogeny

community and was first observed by Felsenstein [7]. We require a more detailed statement of such a result, which we present in Lemma 2.1 in section 2.2. Our main technical contribution is a combinatorial proof that, in the setting of Theorem 1.1, SPR moves can be used in a greedy manner to quickly find the maximum parsimony tree. This result is presented in section 2.3. Finally, in section 2.4 we show how Theorem 1.1 follows in a straightforward manner from these combinatorial results. In section 3 we discuss how Theorem 1.1 extends to Bayesian inference. We make some concluding remarks in section 4.

2. Proof of rapid mixing.

2.1. Overview. To prove Theorem 1.1 we will analyze, for every tree T, the maximum expected log-likelihood $\mathcal{L}_T(\mu^*)$, where $\mu^* = \mu_{T^*,Q^*,\mathbf{t}^*}$ (recall that $\mathcal{L}_T(\mu^*)$ is the maximum expected log-likelihood of T maximized over all rate matrices Q and all edge lengths $\mathbf{t}$; see (1.3)). To analyze $\mathcal{L}_T(\mu^*)$ we will consider the dominant terms of the likelihood function. We will show that

$$\mathcal{L}_T(\mu^*) = \mathcal{E}(\pi^*) + A(T)\varepsilon \ln \varepsilon + o(\varepsilon \ln \varepsilon),$$

where $\mathcal{E}(\pi^*)$ is the entropy of the stationary distribution of Q^* and thus is the same for every T. By taking ε sufficiently small, the last term $o(\varepsilon \ln \varepsilon)$ can be ignored. Therefore, the dominant term is $A(T)\varepsilon \ln \varepsilon$. We will prove that the function $A(T)$ decreases with each optimal SPR move. Hence, since $\ln \varepsilon$ is negative, we then have that T^* has the highest maximum expected log-likelihood, and as the Markov chain gets closer to T^* the maximum expected log-likelihood will increase. Theorem 1.1 will then follow in a straightforward manner.

2.2. Analyzing likelihood. Let T be a tree with leaves $\{1, \ldots, n\}$. Let R be a partition of the leaves into two parts (R_1, R_2). Note that we consider only the partition of the leaves without any regard for the internal vertices. Let $\mathrm{cut}_R(T)$ denote the size of the cut (i.e., a subset of edges) of minimum size that disconnects R_1 from R_2. Tuffley and Steel [21] showed that the quantity $\mathrm{cut}_R(T)$ is the parsimony score of the character corresponding to R (see also Semple and Steel [15, Proposition 5.1.6]), where the character corresponding to $R = (R_1, R_2)$ assigns all leaves in R_1 some $\alpha \in \mathbf{\Omega}$ and assigns all leaves in R_2 some $\beta \in \mathbf{\Omega}$, $\beta \neq \alpha$.

For an edge $e \in T$, the removal of e splits T into two components. This induces a partition of the leaves of T into two parts. We will call this partition $R(T, e)$.

The following is the main technical tool for our results. The lemma describes the high-order terms of the likelihood function as $\varepsilon \to 0$. Throughout this paper, the asymptotic notations $o()$ and $O()$ are parameterized by $\varepsilon \to 0$.

Roughly speaking, the lemma shows there is a function $A(T)$ which plays a leading role in the maximum likelihood. In words, $A(T)$ looks at each partition R of leaves in the generating tree T^* realized by cutting a single edge e^*. It then considers the minimum number of edges in T to realize this partition R times the branch length of e^* in the generating tree. As mentioned in the introduction, there are earlier results which show that the leading term of the likelihood function is the parsimony score, e.g., Felsenstein [7], and in Lemma 2.1, the function A is the leading term of the expected parsimony score. We require a more detailed statement than we found in the literature.

LEMMA 2.1. *Let T^*, T be trees with leaves $\{1, \ldots, n\}$ and Q^* be a rate matrix reversible with respect to π^*. Assume that the matrix Q^* is normalized (that is, $\sum_{i \neq j} \pi_i^* Q_{ij}^* = 1$) and that Q^* has no zero entries. Let T^* have branch lengths $\mathbf{t}_e^* = \alpha_e^* \varepsilon$, where $\alpha_e^* \in [\alpha_{\min}^*, \alpha_{\max}^*]$, for all $e \in E(T^*)$, where $\alpha_{\min} > 0$. Let*

$$A = A_{T^*,\alpha^*}(T) = \sum_{e \in E(T^*)} \alpha^*_e \mathrm{cut}_{R(T^*,e)}(T). \tag{2.1}$$

For $\mu^* = \mu_{T^*,Q^*,\mathbf{t}^*}$ *the following holds:*

$$\mathcal{L}_T(\mu^*) = \mathcal{E}(\pi^*) + A\varepsilon \ln \varepsilon + O(\varepsilon \ln \ln(1/\varepsilon)), \tag{2.2}$$

where $\mathcal{E}(\pi^*) = \sum_{i \in \Omega} \pi^*_i \ln \pi^*_i$ *is the entropy of* π^* *and the constant in the* $O(\cdot)$ *is independent of the choice of the* α^*_e *(but does depend on* $\alpha^*_{\min}$ *and* $\alpha^*_{\max}$*).*

As a consequence of Lemma 2.1, to analyze the expected log-likelihood on the tree space, when ε is sufficiently small, we simply have to consider the function $A = A(T)$. In the next subsection we will investigate the combinatorial properties of A.

Before presenting the proof of Lemma 2.1 we give a brief outline of its proof.

Let μ^* denote the probability distribution defined by Q^* and T^* on assignments of labels from Ω to the leaves. In this generating distribution μ^*, to prove (2.2) we need only to consider two types of assignments. The first type is constant assignments where no substitutions occur, and thus all leaves receive the same label $i \in \Omega$; these are denoted as σ_i. The second type is assignments obtained by a substitution along just one edge e^*. In this case, the cut obtained by deleting edge e^* plays an important role. By deleting e^* from T^*, the leaves are partitioned into two sets R_1 and R_2, denoted as $R(T^*, e^*)$. If a substitution occurs only along edge e^*, then the leaves in R_1 will receive the same label $i \in \Omega$, and the leaves in R_2 will receive another label $j \in \Omega$, $j \neq i$. We denote such an assignment by $\sigma^{e^*}_{ij}$. Any other type of assignment requires at least two substitutions and hence has probability at most $O(\varepsilon^2)$, which is dominated by the $O(\varepsilon \ln \ln(1/\varepsilon))$ term of (2.2).

For any tree T, to prove that $\mathcal{L}_T(\mu^*)$ is lower bounded by the right-hand side of (2.2), we compute the expected log-likelihood of μ^* for the rate matrix $Q = Q^*$ and the set of branch lengths $\mathbf{t}$, where $\mathbf{t}_e = \varepsilon$ for every edge e. For each edge e^* and its corresponding assignment $\sigma^{e^*}_{ij}$, the quantity $\mathrm{cut}_{R(T^*,e)}(T)$ is the minimum number of edges which require a substitution to obtain the assignment $\sigma^{e^*}_{ij}$ on T. Hence, the quantity $A = A(T)$ plays an important role when we sum over all edges e^* of T^*. In particular, by a calculation (as detailed in (2.9) below), the expected log-likelihood $\mathcal{L}_{T,Q,\mathbf{t}}(\mu^*)$ for this set of branch lengths $\mathbf{t}$ is $\sum_{i \in \Omega} \pi^*_i \ln \pi^*_i + A\varepsilon \ln \varepsilon + O(\varepsilon)$. Since $O(\varepsilon) = O(\varepsilon \ln \ln(1/\varepsilon))$, this implies the lower bound of (2.2).

To obtain the upper bound of (2.2) we consider three cases: when the rate matrix Q has a stationary distribution different from Q^*, when there is an edge e, where $\mathbf{t}_e$ is long (namely, $\geq \varepsilon(\ln(1/\varepsilon))^2$), and when all edges are short. In the first case of different stationary distributions, by considering the constant assignments, it will be easy to establish that there is a difference in the first term of the right-hand side of (2.2). When there is a long edge, the constant assignments are too unlikely to occur. Finally, if all edges are shorter than $\varepsilon(\ln(1/\varepsilon))^2$, then, by calculation, we show that the expected log-likelihood is at most the right-hand side of (2.2).

We now present the formal proof of Lemma 2.1.

Proof of Lemma 2.1. We first make some observations about the distribution μ^*. Let P^* denote the transition matrix for Q^*, as defined in (1.1).

Note, for any e, any i, $j \in \Omega$, where $i \neq j$, we have

$$(P^*_e)_{ij} = Q^*_{ij} \alpha^*_e \varepsilon + O(\varepsilon^2). \tag{2.3}$$

For any $i \in \Omega$ we have

$$(P^*_e)_{ii} = 1 - \sum_{j \neq i} (P^*_e)_{ij} = 1 + O(\varepsilon).$$

For $i \in \Omega$, let $\sigma_i \in \Omega^n$ denote the constant assignment $\sigma_i(v) = i$ for all leaves v. Note to achieve σ_i in μ^* we assign label i to the root and then we have no substitutions, or we have at least two edges with substitutions. Thus,

$$\mu^*(\sigma_i) = \pi^*_i \prod_{e \in E(T^*)} (P^*_e)_{ii} + O(\varepsilon^2) = \pi^*_i + O(\varepsilon). \tag{2.4}$$

For an edge $e \in E(T^*)$ and i, $j \in \Omega$, where $i \neq j$, let $\sigma^e_{ij} \in \Omega^n$ denote the assignment of label i to all leaves in one of the partitions of $R(T^*, e)$ and label j to all leaves in the other partition of $R(T^*, e)$. In this case we have

$$\mu^*(\sigma^e_{ij}) = \pi^*_i Q^*_{ij} \alpha^*_e \varepsilon + O(\varepsilon^2). \tag{2.5}$$

(To see why (2.5) is correct, w.l.o.g., assume that the root is a leaf in the first partition of $R(T^*, e)$, and hence to achieve σ_{ij} we need to label the root by i and have a substitution on e or at least two edges with substitutions.)

Now we compute $\mathcal{L}_{T,Q,\mathbf{t}}(\mu^*)$, where $\mathbf{t}_e = \varepsilon$ for each edge e of T and $Q = Q^*$. Again we will make some observations about $\mu = \mu_{T,Q,\mathbf{t}}$. By the same reasoning as we used for (2.4), we obtain

$$\mu(\sigma_i) = \pi^*_i + O(\varepsilon). \tag{2.6}$$

We can obtain assignment σ^e_{ij} on T using a substitution on $\mathrm{cut}_{R(T^*,e)}(T)$ edges, and we cannot obtain this assignment with fewer substitutions. Hence,

$$\mu(\sigma^e_{ij}) = \Theta(\varepsilon^{\mathrm{cut}_{R(T^*,e)}(T)}). \tag{2.7}$$

Therefore,

$$\ln \mu(\sigma^e_{ij}) = \Theta(1) + \mathrm{cut}_{R(T^*,e)}(T) \ln \varepsilon. \tag{2.8}$$

In order to compute the high-order terms of $\mathcal{L}_{T,Q,\mathbf{t}}(\mu^*)$, we do not need to consider labelings other than σ_i and σ^e_{ij} (the other labelings have probability $O(\varepsilon^2)$ in μ^*).

Combining (2.5), (2.4), (2.6), and (2.8) we obtain

$$\begin{aligned}
\mathcal{L}_{T,Q,\mathbf{t}}(\mu^*) = {} & O(\varepsilon^2 \ln \varepsilon) + \sum_{i \in \Omega} (\pi^*_i + O(\varepsilon)) \ln(\pi^*_i + O(\varepsilon)) \\
& + \sum_{e \in E(T^*)} \sum_{i \neq j} (\pi^*_i Q^*_{ij} \alpha^*_e \varepsilon + O(\varepsilon^2))(\Theta(1) + \mathrm{cut}_{R(T^*,e)}(T) \ln \varepsilon) \\
= {} & O(\varepsilon) + \sum_{i \in \Omega} \pi^*_i \ln \pi^*_i + A\varepsilon \ln \varepsilon,
\end{aligned} \tag{2.9}$$

where in the last inequality we used the fact that Q^* is normalized. This proves the lower bound in (2.2).

It remains to prove the upper bound in (2.2). We will show that no rate matrix and no assignment of branch lengths can do better than the bound established in (2.9). Let Q be a rate matrix with stationary distribution π. If $\pi \neq \pi^*$, then we bound $\mathcal{L}_{T,Q,\mathbf{t}}(\mu^*)$

as follows. First, note that the terms in the sum (1.4) are negative, and hence to obtain an upper bound we will consider only the constant assignments. Second, the probability of constant assignment σ_i in μ^* is $\mu^*(\sigma_i) \leq \pi_i^*$ and similarly $\mu(\sigma_i) \leq \pi$. Thus

$$\mathcal{L}_{T,Q,\mathbf{t}}(\mu^*) \leq \sum_{i\in\Omega} \pi_i^* \ln \pi_i = \sum_{i\in\Omega} \pi_i^* \ln \pi_i^* - D_{KL}(\pi^* \| \pi),$$

where $D_{KL}(\pi^* || \pi) \coloneqq \sum_{i\in\Omega} \pi_i^* (\ln(\pi_i^*/\pi_i))$ is the KL-divergence of π from π^*. Since, by the Gibbs' inequality, the KL-divergence is positive when $\pi \neq \pi^*$, we have established the upper bound in (2.2) for the case $\pi \neq \pi^*$.

Now we assume $\pi = \pi^*$. Let $\mathbf{t}$ be an assignment of branch lengths to T. Let $\mu = \mu_{T,Q,\mathbf{t}}$. Suppose that there exists an edge $f \in E(T)$ with branch length $\mathbf{t}_f > \varepsilon(\ln(1/\varepsilon))^2$. We are going to show that such a $\mathbf{t}$ has a tiny log-likelihood because of the constant leaf labelings (i.e., σ_i, $i \in \Omega$). By (1.1), we have $(P_f)_{ii} \leq 1 - q_{\min}\varepsilon(\ln(1/\varepsilon))^2 + O(\varepsilon^2(\ln(1/\varepsilon))^4)$, where $q_{\min} = \min_{i,j\in\Omega}|Q(i,j)|$. Hence,

$$\mu(\sigma_i) \leq \pi_i(1 - q_{\min}\varepsilon(\ln(1/\varepsilon))^2 + O(\varepsilon^2(\ln(1/\varepsilon))^4)).$$

Thus

$$\begin{aligned}\mathcal{L}_{T,Q,\mathbf{t}}(\mu^*) &\leq O(\varepsilon) + \sum_{i\in\Omega} \pi_i^*(\ln(\pi_i) - q_{\min}\varepsilon(\ln(1/\varepsilon))^2 + O(\varepsilon^2(\ln(1/\varepsilon))^4)) \\ &\leq \mathcal{E}(\pi^*) - q_{\min}\varepsilon(\ln(1/\varepsilon))^2 + O(\varepsilon). \qquad (2.10)\end{aligned}$$

As $\varepsilon \to 0$, (2.10) is smaller than the right-hand side of (2.2), and we are done.

We are now left with the case in which all edges $f \in E(T)$ have branch lengths $\mathbf{t}_f \leq \varepsilon(\ln(1/\varepsilon))^2$. Since we can generate the leaf labelings starting from any vertex, then by starting at a leaf, we see that

$$\ln \mu(\sigma_i) \leq \ln \pi_i. \qquad (2.11)$$

Moreover, for $e \in E(T^*)$, to generate σ_{ij}^e, we need to have substitutions across all edges in a cut that realizes $R(T^*, e)$. Since the edges are short, this happens with probability $\leq(\varepsilon(\ln(1/\varepsilon))^2)^k$, where k is the size of the cut. Since there are at most 2^n such cuts and each has size at least $\mathrm{cut}_{R(T^*,e)}(T)$, we have that

$$\ln \mu(\sigma_{ij}^e) = \mathrm{cut}_{R(T^*,e)}(T)(O(\ln \ln(1/\varepsilon)) + \ln \varepsilon). \qquad (2.12)$$

Hence,

$$\begin{aligned}\mathcal{L}_{T,Q,\mathbf{t}}(\mu^*) &\leq O(\varepsilon^2 \ln \varepsilon) + \mathcal{E}(\pi) \\ &\quad + \sum_{e\in E(T^*)} \sum_{i\neq j} (\pi_i^* Q_{ij}^* \alpha_e^* \varepsilon + O(\varepsilon^2)) \mathrm{cut}_{R(T^*,e)}(T)(O(\ln \ln(1/\varepsilon)) + \ln \varepsilon) \\ &= O(\varepsilon \ln \ln(1/\varepsilon)) + \mathcal{E}(\pi) + A\varepsilon \ln \varepsilon. \qquad \square\end{aligned}$$

2.3. Analyzing the cut distance $A(T)$. In light of Lemma 2.1, we need to analyze how $A(T)$ changes with SPR moves. By taking N sufficiently large, for each subtree S, we will only need to analyze the effect of the optimal SPR move for S (optimal in terms of minimizing $A(T')$).

The quantity $A(T)$ looks at cuts obtained by single edges of T^*. For a tree T, we classify the edges of T^* as good or bad if their corresponding cut in T^* is realizable in T by cutting a single edge. More precisely, let

$$\text{GOOD}_{T^*}(T) = \{ e^* \in E(T^*) : \text{there exists } e \in E(T) \text{ where } R(T, e) = R(T^*, e^*)\}$$

be the set of good edges for T. Let $\text{BAD}_{T^*}(T) = E(T^*) \setminus \text{GOOD}_{T^*}(T)$.

Lemma 2.2 says that for every tree $\tilde{T}$ obtained from T by an SPR move using S, if $\tilde{T}$ has more bad edges than T, then this was not the optimal SPR move using S. Namely, there is a tree T' which is also obtained from T by an SPR move using S, and T' is such that $A(T') < A(\tilde{T})$. (More precisely, each term in $A(T')$ is less than or equal to the corresponding term in $A(\tilde{T})$, and there is a term in $A(T')$ which is strictly smaller than the corresponding term in $A(\tilde{T})$). Our proof has some similarity to those of Bruen and Bryant [2] which connect the parsimony score of a character to the minimum number of SPR transitions needed to obtain the character.

LEMMA 2.2. *For every generating tree T^* and all trees T, $\tilde{T}$, where T and $\tilde{T}$ differ by a prune-and-regraft of a subtree S and such that there exists $f^* \in \text{BAD}_{T^*}(\tilde{T}) \setminus \text{BAD}_{T^*}(T)$, the following holds. There exists a tree T' which differs from T by a prune-and-regraft of S and such that $\text{cut}_R(T') \leq \text{cut}_R(\tilde{T})$ for every partition R realized by single edges in T^* and $\text{cut}_R(T') < \text{cut}_R(\tilde{T})$ for partition R realized by f^* in T^*.*

Proof. Suppose an edge $f^* \in E(T^*)$ is good for T and is bad for $\tilde{T}$. Let L_1, L_2 be the partition of the leaves induced by f^* in T^*. Thus, in T, there is an edge $f = (v_1, v_2)$ which partitions the leaves into L_1 and L_2. See Figure 2.1 for an illustration of the setup.

Let S_1 denote the subtree "hanging off" of v_1 in T. More precisely, after deleting f from T, let S_1 be the subtree containing v_1. Let L_1 denote the leaves in S_1. Similarly, let L_2 denote the leaves and S_2 denote the subtree hanging off of v_2. Let v denote the root of the subtree S.

First we claim that $f \notin S$. Suppose $f \in S$ and w.l.o.g. suppose $S_1 \subset S$. See Figure 2.2 for an illustration of this case. Thus we must be grafting S into an edge of $S_2 \setminus S$.

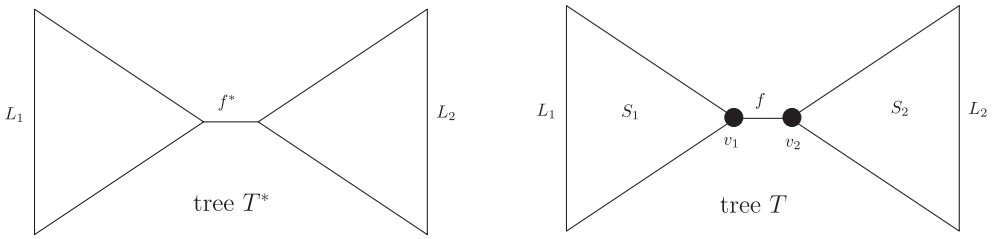

FIG. 2.1. *Edge f^* is good for T.*

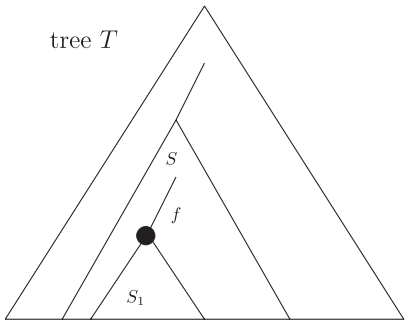

FIG. 2.2. *Case when $f \in S$; this scenario cannot occur.*

After such a move, the edge f still separates L_1 and L_2, and thus f^* is still good. Therefore, $f \notin S$.

From now on, we assume, w.l.o.g., that $S \subset S_1$, where $S \neq S_1$; see Figure 2.3. We construct the tree T' by taking T, pruning S, and then regrafting S along edge f; see Figure 2.4.

Note that $\tilde{T}$ is obtained from T by regrafting S onto an edge in S_2 (otherwise f^* would be good for $\tilde{T}$); see Figure 2.5.

The following claim says that the tree T' satisfies the conclusion of the lemma.

Claim 2.3. For every partition $R = (R_1, R_2)$ of leaves realized by edges of T^*, it holds that

$$\mathrm{cut}_R(\tilde{T}) \geq \mathrm{cut}_R(T').$$

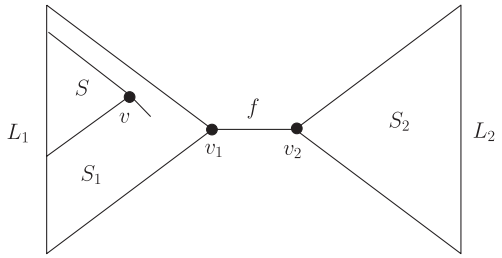

FIG. 2.3. *Case when $f \notin S$; this must be the scenario.*

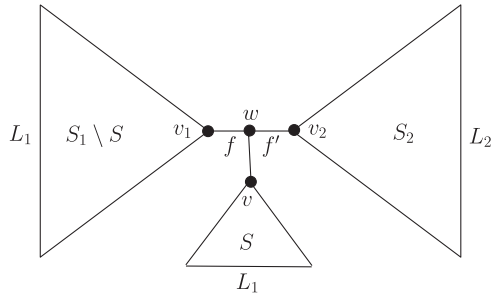

FIG. 2.4. *Construction of the tree T'.*

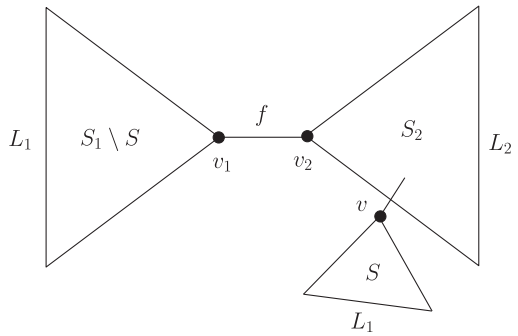

FIG. 2.5. *In tree $\tilde{T}$, S is regrafted into S_2.*

Moreover, for the partition $R^* = (L_1, L_2)$ (corresponding to f^*), we have that

$$\text{cut}_{R^*}(\tilde{T}) > \text{cut}_{R^*}(T').$$

The proof of the claim proceeds by constructing a cut in T' realizing (R_1, R_2) by a small modification of a cut in $\tilde{T}$ realizing (R_1, R_2). Assuming the claim, the proof of the lemma is now complete. □

We now prove Claim 2.3.

Proof of Claim 2.3. We continue using the setup and notation from the proof of Lemma 2.2 in section 2.3.

Recall that the claim says that for every partition $R = (R_1, R_2)$ of leaves realized by edges of T^* that $\text{cut}_R(\tilde{T}) \geq \text{cut}_R(T')$ and for the partition $R^* = (L_1, L_2)$, $\text{cut}_{R^*}(\tilde{T}) > \text{cut}_{R^*}(T')$.

First we argue that $\text{cut}_{R^*}(\tilde{T}) > \text{cut}_{R^*}(T')$. Note that $\text{cut}_{R^*}(\tilde{T}) \geq 2$ since f^* is bad for $\tilde{T}$. On the other hand, $\text{cut}_{R^*}(T') = 1$ since cutting f'' separates L_1 and L_2. Now we just need to argue that $\text{cut}_R(\tilde{T}) \geq \text{cut}_R(T')$.

Let $g^* \in E(T^*)$ be an edge in T^*. Let $R = (R_1, R_2)$ be the corresponding partition in T^*. Note that if g^* is in the subtree with leaves L_1, then

$$R_1 \subseteq L_1 \quad \text{and} \quad R_2 \supseteq L_2. \tag{2.13}$$

On the other hand, if g^* is in the subtree with leaves L_2, then

$$R_2 \subseteq L_2 \quad \text{and} \quad R_1 \supseteq L_1. \tag{2.14}$$

Consider a minimum cut $C \subset E(\tilde{T})$ that realizes (R_1, R_2) in $\tilde{T}$ and among these minimum cuts is the one with the fewest number of edges in subtrees $S_1 \setminus S$ and S.

We claim that v_1 is reachable from a leaf of $S_1 \setminus S$ in $\tilde{T} \setminus C$ and that v is reachable from a leaf of S in $\tilde{T} \setminus C$. Suppose that v_1 is not reachable from a leaf of $S_1 \setminus S$. Let e' be the edge in $C \cap (S_1 \setminus S)$ closest to v_1. We claim that $C' = (C \setminus \{e'\}) \cup \{f\}$ realizes (R_1, R_2). If there were a pair of leaves in $S_1 \setminus S$ each in different R_i that are connected in $\tilde{T} \setminus C'$, then by the choice of e' one of those leaves would be connected to v_1 in $\tilde{T} \setminus C$, a contradiction with the assumption that v_1 is not reachable from a leaf of $S_1 \setminus S$ in $\tilde{T} \setminus C$. Thus R_1 and R_2 are still separated in $S_1 \setminus S$ in $\tilde{T} \setminus C'$; R_1 and R_2 are still separated by C' in $S_2 \cup S$ since $C = C'$ in this subtree; and $f \in C'$ ensures that pairs across f are separated. Note that $|C'| \leq |C|$ and C' has fewer edges in $S_1 \setminus S$ and S, a contradiction with the choice of C. Thus v_1 is reachable from some leaf of $S_1 \setminus S$ in $\tilde{T} \setminus C$. The argument for S and v is the same.

Since a leaf of S is reachable from v in $\tilde{T} \setminus C$, then in other words a (nonempty) subset of R_1 and/or R_2 is reachable from v. Moreover, since C realizes the partition (R_1, R_2), then a subset of only one of the R_i is reachable from v in $\tilde{T} \setminus C$; we will say v is of type R_i to signify the R_i reachable from v. Analogously, we say v_i is of type R_i for the set reachable from v_1.

If v and v_1 are of the same type R_i, let

$$C' = \begin{cases} C & \text{if } f \notin C, \\ (C \setminus \{f\}) \cup \{f'\} & \text{if } f \in C. \end{cases}$$

We claim C' realizes (R_1, R_2) in T'. To see this, note that if a path (between a pair of leaves) exists in T' and does not exist in $\tilde{T}$, then it must include w, which is the new vertex in T' where S is regrafted; see Figure 2.4 for an illustration. Now we argue that

such a path cannot connect a leaf in R_1 with a leaf in R_2 in $T' \setminus C'$. Note that only a subset of R_i is reachable from w in $T' \setminus C'$, since w can reach the same set of vertices (outside of S) in $T' \setminus C'$ as v_1 does in $\tilde{T} \setminus C$, and only a subset of R_i is reachable from v in $S \setminus C$. Finally, since $|C'| = |C|$, we have that $\mathrm{cut}_R(T') \leq \mathrm{cut}_R(\tilde{T})$, which completes the proof in this case.

Now suppose v_1 is of type R_1 and v is of type R_2. This means a leaf of S is in R_2, and since $S \subset S_1$, it is also in L_1, and we are in case (2.13); thus $R_2 \supseteq L_2$. Note C has to separate v_1 from L_2 by some set of edges $Q \subseteq C$. Let $C' = (C \setminus Q) \cup \{f\}$. The new pairs of leaves that are connected in $T' \setminus C'$ (but not in $\tilde{T} \setminus C$) either are both from L_2 and hence R_2 or are connected by a path that exists in T' and does not exist in $\tilde{T}$. As in the previous case, if a path (between a pair of leaves) exists in T' and does not exist in $\tilde{T}$, then it must include w, which is the new vertex in T' where S is regrafted. Note that w is disconnected from $S_1 \setminus S$ in $T' \setminus C'$ (since $f \in C'$). The leaves of S_2 are from R_2, and the leaves of S reachable from v in $S \setminus C$ are also from R_2. Therefore the new paths do not connect leaves of R_1 and R_2. This completes this case since $|C'| \leq |C|$.

Finally, suppose v_1 is of type R_2 and v is of type R_1. In this case a leaf of $S_1 \setminus S$ is in R_2 and is also in L_1, and therefore we are again in case (2.13); thus $R_2 \supseteq L_2$. Note C has to separate v from L_2 by some set of edges $Q \subseteq C$. Let $C' = (C \setminus Q) \cup \{w, v\}$. Once again, the new pairs of leaves that are connected in $T' \setminus C'$ (but not in $\tilde{T} \setminus C$) either are both from L_2 and hence R_2 or are connected by a path that exists in T' and does not exist in $\tilde{T}$. Note that w is disconnected from the leaves of S in $T' \setminus C'$. The leaves of S_2 are from R_2, and the leaves of $S_1 \setminus S$ reachable from v_1 are also from R_2. Therefore the new paths in $T' \setminus C'$ do not connect leaves of R_1 and R_2. This completes this case since $|C'| \leq |C|$.

This completes the proof of the claim. □

Using Lemma 2.2, we will prove that for every subtree S the optimal SPR move using S does not increase the number of bad edges, and there is a subtree S where the optimal SPR move using S decreases the number of bad edges. It will then be straightforward to prove rapid mixing by analyzing the time until the number of bad edges is zero, and hence we have reached T^*.

LEMMA 2.4. *For all trees T^*, every choice of parameters $\alpha : E(T^*) \to \mathbb{R}^+$, and for all trees $T \neq T^*$ the following holds, where $A = A_{T^*,\alpha}$ is defined in* (2.1)*.:*

1. *For any subtree S of T the following holds. Let $T_{\min}$ be any tree which minimizes $A(T_{\min})$ among the SPR neighbors of T which differ by a prune-and-regraft of S. Then*

$$\mathrm{BAD}_{T^*}(T_{\min}) \subseteq \mathrm{BAD}_{T^*}(T). \tag{2.15}$$

2. *There exists a subtree S of T where the following holds. Let $T_{\min}$ be any tree which minimizes $A(T_{\min})$ among the SPR neighbors of T which differ by a prune-and-regraft of S. Then*

$$\mathrm{BAD}_{T^*}(T_{\min}) \nsubseteq \mathrm{BAD}_{T^*}(T). \tag{2.16}$$

Part 1 of Lemma 2.4 follows immediately from Lemma 2.2. To prove part 2 we choose a particular "minimal" subtree S. Roughly speaking, we consider the bad edge f^* that is closest to the leaves in T^* and take the subtree S hanging off of f^*.

Proof of Lemma 2.4. If (2.15) is violated, then there exists $f^* \in \mathrm{BAD}_{T^*}(T_{\min}) \setminus \mathrm{BAD}_{T^*}(T)$, and hence by Lemma 2.2, there exists T' (which differs from T and $T_{\min}$ by a prune-and-regraft of S) such that no cuts increased in size and the cut

corresponding to f^* is smaller. Therefore, $A(T') < A(T_{\min})$, contradicting the choice of $T_{\min}$. Therefore, part 1 holds.

We now prove part 2. We first claim that there is an SPR move that decreases the number of bad edges.

Claim 2.5. For every tree T, there is an SPR move resulting in a tree T', where

$$\text{BAD}_{T^*}(T') \nsubseteq \text{BAD}_{T^*}(T). \tag{2.17}$$

Now we argue that part 2 of Lemma 2.4 follows from Claim 2.5 and part 1. We then go back to prove the claim.

Consider a subtree S of T. Let $N_S(T)$ denote those trees obtainable from T by a prune-and-regraft of S. Note that for any $T' \in N_S(T)$, we have that $N_S(T') = N_S(T)$, since when we prune S from T and T', we have the same subtree remaining.

Let T' denote the neighboring tree from Claim 2.5 with fewer bad edges, and let S denote the subtree where $T' \in N_S(T)$. Let $T_{\min}$ denote the tree in $N_S(T)$ which minimizes $A(T_{\min})$. As noted above, we must have that $N_S(T') = N_S(T)$. Thus, $T_{\min}$ is also the neighbor of T' that minimizes $A(T_{\min})$. Therefore, we can apply part 1 of Lemma 2.4 for tree T' and subtree S, and we conclude that $\text{BAD}_{T^*}(T_{\min}) \subseteq \text{BAD}_{T^*}(T')$. Combined with (2.17) we then have that

$$\text{BAD}_{T^*}(T_{\min}) \nsubseteq \text{BAD}_{T^*}(T),$$

which proves part 2 of Lemma 2.4. □

We now prove Claim 2.5.

Proof of Claim 2.5. Let f^* in T^* be an edge in $\text{BAD}_{T^*}(T)$ that is "closest" to the leaves in the following precise sense. Say f^* joins subtrees S^* and Z^* in T^*, where the number of vertices in S^* is at most the number of vertices in Z^*. Then we say the distance of f^* to the leaves is the number of vertices of S^*.

Note that by the choice of f^*, S^* contains no bad edges for T. First, note that S^* must contain at least two leaves because, in any tree, any single leaf can be separated from the rest of the leaves by deleting one edge (which would contradict that f^* is bad). Let S_1^* and S_2^* denote the two subtrees of S^* hanging from the root of S^* in T^*. Both S_1^* and S_2^* must exist since S^* contains at least two leaves.

Let L_1 and L_2 denote the leaves in S_1^* and S_2^*, respectively. Since f^* is the closest bad edge to the leaves, there is a subtree S_1 in T whose leaves are L_1 and also a subtree S_2 whose leaves are L_2. Moreover, by induction, $S_1 = S_1^*$ and $S_2 = S_2^*$. In T, by pruning S_2 and then regrafting along the edge incident to S_1, we obtain a copy of S^* in T. See Figure 2.6 for an illustration. Let T' be the tree resulting from this SPR move. Note that f^* is now a good edge in T'.

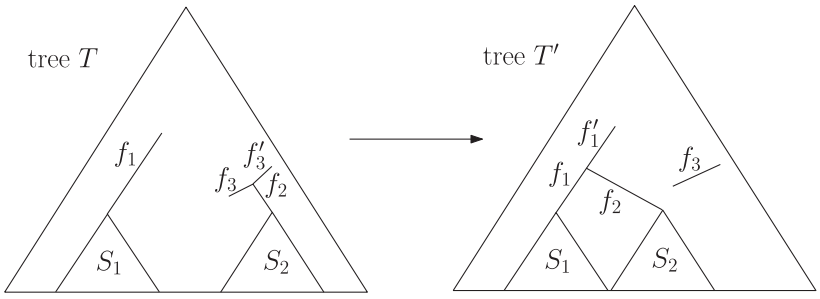

FIG. 2.6. *Construction of the tree T' with fewer bad edges.*

It remains to argue that other edges of T^* did not change from good for T to bad for T'. Note that edges in S_1^* and S_2^* remain good in T' since they are realizable in S_1 and S_2, respectively. Consider an edge e^* of T^*, where $e^* \notin S^*$. Let (R_1, R_2) be the partition of the leaves realized by e^* in T^*. Note that L_1, L_2 are in the same partition, since tree S^* is not cut by e^*. Let g be the edge in T that realizes (R_1, R_2). After pruning-and-regrafting S_2 (to form T'), g still realizes the partition (R_1, R_2) since L_1 and L_2 are in the same partition. Hence, e^* is still good for T'. Therefore, $\mathrm{GOOD}_{T^*}(T') \supseteq \mathrm{GOOD}_{T^*}(T) \cup \{f^*\}$, which completes the proof of Claim 2.5. □

Finally, we prove that when the number of bad edges increases, then $A(T)$ also increases by a significant amount. As a consequence, in our analysis of the Markov chain, by taking ε sufficiently small, we can focus on how a transition changes $A(T)$ and hence on the change in the number of bad edges.

LEMMA 2.6. *For any trees T and T' which differ by one SPR move, if $|\mathrm{BAD}_{T^*}(T')| > |\mathrm{BAD}_{T^*}(T)|$, then*

$$A(T') \geq A(T_{\min}) + \alpha_{\min}.$$

Proof. Let S be the subtree used to move between T and T'. Let $N_S(T)$ denote those trees obtainable from T by a prune-and-regraft of S. Note that $N_S(T) = N_S(T')$.

Consider $T_{\min}$, which minimizes $A(T_{\min})$ among the SPR neighbors of T which differ by a prune-and-regraft of S. Since $N_S(T') = N_S(T')$, then $T_{\min}$ is also the neighbor of T' that minimizes $A(T_{\min})$. Fix $e \in \mathrm{BAD}_{T^*}(T') \setminus \mathrm{BAD}_{T^*}(T)$. By part 1 of Lemma 2.4,

$$\mathrm{BAD}_{T^*}(T_{\min}) \subseteq \mathrm{BAD}_{T^*}(T) \cap \mathrm{BAD}_{T^*}(T'). \qquad \square$$

2.4. Proof of rapid mixing: Theorem 1.1. The proof of our main theorem now follows from a straightforward argument. We show that the heat bath SPR Markov chain behaves like a local search algorithm and then a simple coupling argument gives the mixing result.

Proof of Theorem 1.1. Let $\mathcal{T}$ denote the space of phylogenetic trees on n taxa. For a tree $T \in \mathcal{T}$ and subtree S of T, let $N_S(T)$ denote those trees obtainable from T by pruning-and-regrafting S.

Let C be the constant in the $O(\cdot)$ notation of (2.2) for the chosen $\alpha_{\min}$ and $\alpha_{\max} = 1$. By choosing ε_0 (note that ε_0 is an upper bound on ε) sufficiently small, then for every tree T, in (2.2), the $C\varepsilon \ln \ln(1/\varepsilon)$ is smaller than $|\alpha_{\min}(\varepsilon \ln \varepsilon)/10|$, and therefore

$$\mathcal{L}_T(\mu^*) = \mathcal{E}(\pi^*) + (A(T) + \delta_T)\varepsilon \ln \varepsilon \tag{2.18}$$

for some $|\delta_T| < \alpha_{\min}/10$.

Fix a tree $T \neq T^*$ and a subtree S of T. By Lemma 2.6 and (2.18), for every $T' \in N_S(T)$ where $|\mathrm{BAD}_{T^*}(T')| > |\mathrm{BAD}_{T^*}(T)|$ we have

$$\mathcal{L}_{T'}(\mu^*) < \mathcal{L}_{T_{\min}}(\mu^*) - (9/10)\alpha_{\min}\varepsilon \ln(1/\varepsilon).$$

For a character $\sigma \in \Omega^n$, let $D(\sigma) = |\{i : D_i = \sigma\}|$. A straightforward application of Hoeffding's inequality [9] and a union bound over $\sigma \in \Omega^n$ implies that, for all $\delta > 0$,

$$\Pr(\text{for all } \sigma \in \Omega^n, |D(\sigma) - \mu^*(\sigma)N| \leq \delta N) \geq 1 - 2 \cdot 4^n \exp(-2\delta^2 N).$$

Let $q_{\min} = \min_{i,j\in\Omega: i\neq j} Q_{i,j}$ denote a lower-bound on the off-diagonal entries in the rate matrix. For ε_0 sufficiently small, every labeling of the leaves has probability at least ε^{2n}; this follows from the fact that for every edge, every transition has probability $\Omega(\varepsilon)$; see (2.3) for a precise statement. Hence, by choosing ε_0 sufficiently small (relative to $\alpha_{\min}$, $q_{\min}$ and the constant in the error term of (2.3)), then for all $\sigma \in \Omega^n$, $\mu^*(\sigma) \geq \varepsilon^{2n}$. Let

$$\delta = \alpha_{\min}\varepsilon \ln(1/\varepsilon)/(20 \cdot 4^n n \ln \varepsilon).$$

Then, for $\mathbf{D} \sim \mu^*$,

$$\mathcal{L}_{T'}(\mathbf{D}) < \mathcal{L}_{T_{\min}}(\mathbf{D}) - (7/10)\alpha_{\min}\varepsilon \ln(1/\varepsilon)N$$

with probability $\geq 1 - \exp(-\sqrt{N})$ for N sufficiently large. The probability of moving from T to T' after choosing S in step 1 is at most

$$\frac{\exp(\mathcal{L}_{T'}(\mathbf{D}))}{\exp(\mathcal{L}_{T_{\min}}(\mathbf{D}))} < \exp\left(-\left(\frac{7}{10}\right)\alpha_{\min}\varepsilon \ln\left(\frac{1}{\varepsilon}\right)N\right) < \exp(-10n) \tag{2.19}$$

for N sufficiently large. Therefore, with probability $\geq 1 - 4n\exp(-10n)$, the chain will move from T to some $T_{\min}$ (where $T_{\min}$ is a tree that can be obtained from T by an SPR move and such that it minimizes $A(T_{\min})$), and thus by part 1 of Lemma 2.4 the number of bad edges will not increase. Moreover, if we choose the subtree S satisfying part 2 of Lemma 2.4, then the number of bad edges will decrease. Hence, with probability $\geq 1/(4n) - 4n\exp(-10n) \geq 1/(5n)$ the number of bad edges decreases by at least one. In expectation, after $\leq 5n$ steps of the chain, the number of bad edges will be zero, in which case we have reached T^*. By Markov's inequality, with probability $\geq 9/10$, after $50n$ steps we reach T^*. Once we reach T^* the probability of moving to a different tree within $50n$ steps is at most $50n(4n)^2\exp(-10n) < 1/100$. Hence the claimed mixing time follows by an elementary coupling argument (cf. [12] for an introduction to the coupling technique) since from any pair of initial trees, both chains (run independently) reach T^* at time $50n$ with probability $\geq 1 - 1/2\mathrm{e}$. □

3. Bayesian inference. The goal is often to randomly sample from the posterior distribution over trees. To do this, we consider a Markov chain whose stationary distribution is the posterior distribution and analyze the chain's mixing time, which is a measure of the convergence time of the chain to its stationary distribution. Let $\Phi(T, Q, \mathbf{t})$ denote a prior density, where

$$\sum_T \int_{Q\in\mathcal{Q}} \int_{\mathbf{t}} \Phi(T, Q, \mathbf{t})\mathrm{d}\mathbf{t}\mathrm{d}Q = 1.$$

Our results extend to priors that are lower bounded by some $\delta > 0$ as in Mossel and Vigoda [14]. In particular, for all trees T and all branch lengths $\mathbf{t}$, where $\mathbf{t}_e \leq t_0$ for all edges e, we require $\Phi(T, Q^*, \mathbf{t}) \geq \delta$. We refer to these priors as (δ, t_0)-regular priors.

Applying Bayes law we get the posterior distribution

$$\Pr(T, Q, \mathbf{t}|\mathbf{D}) = \frac{\mu_{T,Q,\mathbf{t}}(\mathbf{D})\Phi(T, Q, \mathbf{t})}{\Pr(\mathbf{D})},$$

where

$$\Pr(\mathbf{D}) = \sum_{T'} \int_{Q' \in \mathcal{Q}} \int_{\mathbf{t}'} \mu_{T',Q',\mathbf{t}'}(\mathbf{D}) \Phi(T', Q', \mathbf{t}') \mathrm{d}\mathbf{t}' \mathrm{d}Q.$$

Each tree T then has a posterior weight

$$w(T) = \int_{Q \in \mathcal{Q}} \int_{\mathbf{t}} \mu_{T,Q,\mathbf{t}}(\mathbf{D}) \Phi(T, Q, \mathbf{t}) \mathrm{d}\mathbf{t} \mathrm{d}Q. \tag{3.1}$$

Finally, the posterior distribution μ on trees is defined as $\mu(T) = w(T)/\sum_{T'} w(T')$.

3.1. Extension of Theorem 1.1 to sampling the posterior. To sample from the posterior distribution, the Markov chain is defined as in section 1.2 except that in step 2 the weight $w(e^*)$ is now set as $w(T^*)$ defined in (3.1). This ensures that the Markov chain is reversible with respect to the posterior distribution, and hence this is the unique stationary distribution.

Theorem 1.1 then extends to hold for any priors which are $(\delta, 2\varepsilon_0)$-regular. The proof easily extends to this case in the following manner.

In particular, we need to modify the statement of Lemma 2.1 so that, for any tree T, (2.2) is achieved for $Q = Q^*$ and for every set of branch lengths $\mathbf{t}$, where $\mathbf{t}_e \in (\varepsilon/2, 2\varepsilon)$ for all edges e. Then we can use the same proof as Lemma 21 in Mossel and Vigoda [14] to get an analogue of (2.19) to hold for the posterior weights defined in (3.1) in place of the maximum likelihood function $\exp(\mathcal{L}(\mathbf{D}))$, and the remainder of the proof of Theorem 1.1 remains the same.

4. Discussion. Nearest neighbor interchangetransitions. In a nearest neighbor interchange (NNI) transition, an internal edge e is chosen. Since internal vertices have degree three, there are four subtrees hanging off of e. There are three possible ways of attaching these four subtrees to e, and an NNI transition moves to one of these rearrangements. There are trees T (different from the generating tree T^*) where no NNI neighbor (strictly) improves $A(T)$; moreover, there are cases where there is also no improvement in the next term of (2.2). We are uncertain as to whether Theorem 1.1 holds for a Markov chain based on NNI transitions. It would be especially intriguing if there are cases where chains based on NNI transitions are slow to converge (so-called torpidly mixing), whereas a chain based on SPR transitions is provably fast to converge (rapidly mixing).

Possible future work. There are now several works with proofs of convergence of MCMC algorithms for phylogenetic reconstruction in certain settings—rapid mixing results in this paper and torpid mixing results in Mossel and Vigoda [13], [14] and Štefankovič and Vigoda [19], [20]). All of these results require that the branch lengths are sufficiently small so that only the dominant terms of the likelihood function need to be considered. A natural avenue for extending this paper is to allow arbitrary branch lengths on the terminal edges.

Rapid or torpid mixing for general pure distributions. The most tantalizing question to the authors is whether there exists a pure distribution (i.e., a single generating tree as in the setting of this paper) where Markov chains based on all of the natural transitions (e.g., NNI, SPR, and tree bisection reconnection transitions) are slow to converge to the stationary distribution (in other words, they are torpidly mixing). We expect simulations can be quite useful for finding such a bad example if one exists; in fact,

our previous work [19], [20] on this topic was inspired by some intriguing findings from some simple simulations.

REFERENCES

[1] D. Aldous, *Random walks on finite groups and rapidly mixing Markov chains*, in Lecture Notes Math. 986, Springer, New York, 1983, pp. 243–297.

[2] T. C. Bruen and D. Bryant, *Parsimony via consensus*, Syst. Biol., 57 (2008), pp. 251–256.

[3] R. Beiko, J. Keith, T. Harlow, and M. Ragan, *Searching for convergence in phylogenetic Markov chain Monte Carlo*, Syst. Biol., 55 (2006), pp. 553–565.

[4] B. A. Berg, *Introduction to Markov chain Monte Carlo simulations and their statistical analysis*, in Markov Chain Monte Carlo: Innovations and Applications, W. S. Kendall, F. Liang, and J.-S. Wang, eds., World Scientific, Singapore, 2005, pp. 1–52.

[5] P. Diaconis and S. P. Holmes, *Random walks on trees and matchings*, Electron. J. Probab., 7 (2002), pp. 1–17.

[6] J. Felsenstein, *Inferring Phylogenies*, Sinauer Associates, Inc., Sunderland, MA, 2004.

[7] J. Felsenstein, *A likelihood approach to character weighting and what it tells us about parsimony and compatibility*, Biol. J. Linn. Soc., 16 (1981), pp. 183–196.

[8] C. J. Geyer and E. A. Thompson, *Annealing Markov chain Monte Carlo with applications to ancestral inference*, J. Amer. Statist. Assoc., 90 (1995), pp. 909–920.

[9] W. Hoeffding, *Probability inequalities for sums of bounded random variables*, J. Amer. Statist. Assoc., 58 (1963), pp. 13–30.

[10] J. P. Huelsenbeck and F. Ronquist, *MRBAYES: Bayesian inference of phylogenetic trees*, Bioinformatics, 17 (2001), pp. 754–755.

[11] C. Lakner, P. van der Mark, J. P. Huelsenbeck, B. Larget, and F. Ronquist, *Efficiency of Markov chain Monte Carlo tree proposals in Bayesian phylogenetics*, Syst. Biol., 57 (2008), pp. 86–103.

[12] D. Levin, Y. Peres, and E. Wilmer, *Markov Chains and Mixing Times*, AMS, Providence, RI, 2008.

[13] E. Mossel and E. Vigoda, *Phylogenetic Markov chain Monte Carlo algorithms are misleading on mixtures of trees*, Science, 309 (2005), pp. 2207–2209.

[14] E. Mossel and E. Vigoda, *Limitations of Markov chain Monte Carlo algorithms for Bayesian inference of phylogeny*, Ann. Appl. Probab., 16 (2006), pp. 2215–2234.

[15] C. Semple and M. Steel, *Phylogenetics*, Oxf. Lect. Ser. Math. Appl. 24, Oxford University Press, New York, 2003.

[16] M. A. Steel and L. A. Székely, *Inverting random functions*, Ann. Comb., 3 (1999), pp. 103–113.

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

[18] M. A. Steel and L. A. Székely, *Inverting random functions* III*: Discrete MLE revisited*, Ann. Comb., 13 (2009), pp. 365–382.

[19] D. Štefankovič and E. Vigoda, *Pitfalls of heterogeneous processes for phylogenetic reconstruction*, Syst. Biol., 56 (2007), pp. 113–124.

[20] D. Štefankovič and E. Vigoda, *Phylogeny of mixture models: Robustness of maximum likelihood and non-identifiable distributions*, J. Comput. Biol., 14 (2007), pp. 144–155.

[21] C. Tuffley and M. Steel, *Links between maximum likelihood and maximum parsimony under a simple model of site substitution*, Bull. Math. Biol., 59 (1997), pp. 581–607.

[22] S. Yang, *Computational Molecular Evolution*, Oxford University Press, New York, 2007.