Investigating Convergence of Markov Chain Monte Carlo Methods for Bayesian Phylogenetic Inference

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

David A. Spade, B.S., M.S.

Graduate Program in Statistics

The Ohio State University

2013

Dissertation Committee:

Dr. Radu Herbei, Advisor

Dr. Laura S. Kubatko (Co-Advisor)

Dr. Steven MacEachern

Dr. Dennis Pearl

Abstract

In biology, it is commonly of interest to investigate the evolutionary pattern that gave rise to an existing group of individuals, such as species or genes. This pattern is most often represented pictorially by a phylogenetic tree. Many methods of inferring evolutionary patterns have been proposed, but as advances in computational capabilities have made Bayesian inference more approachable, it has become an increasingly popular technique for phylogenetic inference.

In Bayesian inference, it is often the case that the posterior density cannot be written out in its entirety due to the intractability of the normalizing constant. One way of working around this is to use a Markov chain Monte Carlo (MCMC) method. The idea is that after several (possibly many) iterations, the chain has approximately converged to its stationary distribution, namely, the posterior distribution. After these initial iterations, subsequent steps of the chain represent an approximate sample from the posterior distribution, thus enabling Bayesian inference.

The biggest question one faces when using MCMC methods is the question of how long the chain should be run before sampling can begin, i.e., the mixing time of the chain. Many methods exist that aim to answer this question by

using the output of the chain, but these methods can only give indications that the chain has not converged. They cannot be used to conclude that a Markov chain has converged.

In this dissertation, we first provide upper bounds on the mixing times of two distinct Markov chains. Both chains move about the space of rooted phylogenetic tree topologies. We also explore methods of bounding the mixing time for a special case of the Metropolis-Hastings algorithm for inference of the branch lengths of a phylogenetic tree given the tree topology. We first provide an upper bound on the mixing time through analytical methods. When this provides results that do not give a helpful upper bound on the mixing time, we present a Monte Carlo method. The Monte Carlo method of bounding the mixing time also gives results that do not lead to a helpful upper bound, but it does provide a substantial improvement over the analytical methods. This represents a step forward in the pursuit of an upper bound on the mixing time of a specific MCMC algorithm for Bayesian inference of the branch lengths of a phylogenetic tree.

I dedicate this dissertation to my wife, Marie S. Spade.

Acknowledgments

This dissertation is the culmination of many years of education. One does not arrive at this point without the help and support of others, and I would like to take this opportunity to acknowledge those who played a role in my reaching this stage of my career as a student.

First and foremost, I owe an enormous debt of gratitude to my advisors, Dr. Radu Herbei and Dr. Laura Kubatko. I have learned a great deal from each of them. Some of the things I have learned from them are research-related, but the others are life lessons that I will carry with me after I leave Ohio State. They have taught me that making a mistake is permissible, but that making the same mistake more than once is not. They have also taught me that there is a time for discussion, and there is a time for doing what those who have more knowledge and experience say to do. The ability to make this distinction can mean the difference between success and failure. From them, I have learned that the key to one's success in research is that he or she never stops asking questions. I believe these lessons will serve me well both in my academic career and in my life outside of my career. I would also like to express my appreciation to my dissertation committee members, Drs. Steven MacEachern and Dennis Pearl, for taking the time to discuss some matters related to my research with

me, to read my dissertation, and to sit on my dissertation defense. I would also like to thank Dr. MacEachern for the milk he brought to class for me on the final instructional day of STAT 620.

Graduate school is not easy, and for me, a stable and supportive family life has been essential to my reaching this point. I would like to thank my wife, Marie, who has, on several occasions, made sacrifices in order to facilitate my success in this undertaking. In general, I am grateful to her for agreeing to be mixed up with me for the rest of her life. I would also like to express my gratitude to the following family members and friends for their continuous support and continual encouragement: my father, Gale Spade, Jr., my sister, Jacki Volkman, my parents-in-law, David and Joann Young, my siblings-in-law, Cameron Volkman, Colleen Young, and Aubrey Young, and my dear friends Gary and Adele Griffin. As a graduate of the University of Michigan, Gary is always willing to engage in lively discussions about Michigan and Ohio State athletics. These conversations have provided a welcome distraction from the rigors of graduate school.

Throughout my career as a student, I have been fortunate to have encountered many great educators. However, there are a few that stand out for their significant roles in my choosing this path. I thank Kathleen Tucci, my eighth-grade mathematics instructor, whose unique teaching style first piqued my interest in the mathematical sciences. I would like to thank my high school calculus instructor, Carol Obermann, who not only saw great potential in me

and worked very hard to ensure that I saw it as well, but also played a major role in my decision to pursue post-secondary education in a mathematical field. I also thank my undergraduate advising group, comprised of Robert Fliess, Arden Welsh, and the late Charles Baker. They convinced me that I had the talent that is necessary to successfully complete a doctoral program in statistics.

During my six years at The Ohio State University, I have been fortunate enough to have made some great friends. I would like to thank the following people for making the grind of graduate school a bit more tolerable: Stephen Bamattre, Tayler Blake, Jon Bradley, Sara Conroy, Casey Davis, John Lewis, Grant Schneider, Michael Sonksen, and Katie Thompson.

Finally, I would like to thank Urban Meyer and Thad Matta for making Ohio State football and basketball fun to watch. It is much easier for me to use sports as a distraction when my favorite teams are winning games.

Vita

October 1, 1983	Born - Tunkhannock, PA
2006	B.S. Mathematics, West Liberty State College
2010	M.S. Statistics, The Ohio State University
2007-present	Graduate Teaching Associate, The Ohio State University.

Fields of Study

Major Field: Statistics

Table of Contents

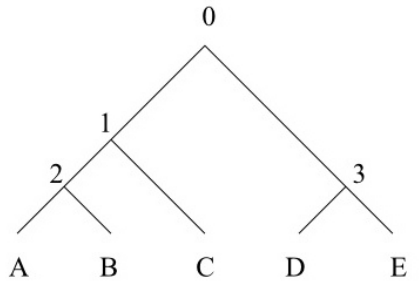
List of Tables

List of Figures

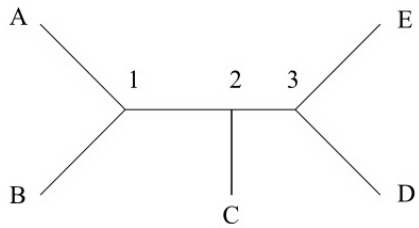
Chapter 1: Introduction, Background, and Literature Review

The focus of this work is on inference of the evolutionary pattern among a group of genes, organisms, or species. We refer to a member of the group whose evolutionary pattern we aim to infer as a *taxon*. Inference of the evolutionary pattern among a group of taxa is a common question in biology, and as a beginning to finding an answer, biologists often use a branching diagram known as a *phylogenetic tree* to represent the evolutionary relationships among the taxa. For a collection of n taxa, such a diagram can be viewed as an acyclic graph with n external vertices termed *leaves* or *tips*, where each tip represents one taxon, and $n-2$ internal nodes of degree 3. A phylogenetic tree may be rooted or unrooted. If the tree is rooted, then the graph has one internal node, termed the *root*, that has degree 2 and identifies the most recent common ancestor (MRCA) among the taxa. An illustration of a rooted and an unrooted trees can be seen in Figure 1.1.

A phylogenetic tree consists of two parts. The first part is the tree topology, which represents the branching pattern for the tree but gives no information about the time evolution has taken between the nodes. If there are taxa at



(a) A rooted tree topology, where the leaves are labelled with the letters A, B, C, D, and E. The root, which is the only node with degree 2, has been labelled 0, and the internal, non-root nodes have been labelled with the numbers $1, 2$, and 3. Each of these three nodes has degree 3.



(b) An unrooted tree topology, where the leaves are labelled with the letters A, B, C, D, and E. The other three nodes, labelled with the numbers $1, 2$, and 3, are internal nodes, and each of them has degree 3.

Figure 1.1: Two types of phylogenetic trees

the tips, the topology is *labelled*. Otherwise, the topology is *unlabelled*. For the rest of this dissertation, we restrict our attention to labelled topologies, so that any tree topology to which we refer is taken to be labelled. The branch lengths are the other component of a phylogenetic tree, and they represent the amount of evolutionary time between an internal node and its immediate descendant.

We focus on Markov chain Monte Carlo (MCMC) methods of inferring each of the two parts of a phylogenetic tree. First, we explore the time to convergence of two Markov chains on T_n, the space of n-leaf rooted tree topologies. Next, we describe methods that can be used to bound the time to convergence of an MCMC algorithm for Bayesian inference of the branch lengths of the tree

given the topology and a set of DNA sequence data.

The importance of this topic lies in the frequency with which phylogenetic trees are used to represent evolutionary relationships. In immunology, knowledge of the ancestry of a particular microorganism can be helpful not only in determining what adaptations the microbe has developed to current medications, but also in developing new medications to fight it. Phylogenetic analysis is also used in forensic epidemiology. For instance, in 1987, three patients of a particular dentist in Florida became infected with the human immunodeficiency virus (HIV) after visiting the dentist. The dentist was HIV-positive, and had performed an invasive procedure on each of the three patients. A study that included phylogenetic analysis (CDC, 1991) suggested that the three patients were infected by the dentist.

The work detailed in this dissertation finds application in many other areas, and the widespread employment of phylogenetic analysis highlights the importance of methods such as the ones detailed below. We begin by providing information required for a thorough understanding of the work we present. We first give relevant background on phylogenetic trees, and we follow that with a brief review of the literature.

1.1 Phylogenetic Trees

This section begins with a description of data used in phylogenetic inference. This is followed by a discussion of probability models for the evolutionary process by which one nucleotide base changes to another. We then describe the

role these models play in likelihood calculation. We close this section by defining two tree moves that are frequently used in MCMC algorithms for Bayesian phylogenetic inference.

1.1.1 Data for Use in Phylogenetic Inference

Suppose that we have available a set of present-day taxa, and that the evolutionary pattern among these taxa can be modelled by a phylogenetic tree. We observe genetic data from these taxa, which are represented by the tips of the tree, but in general we cannot observe such data at any of the ancestral nodes. Several types of genetic data exist, including deoxyribonucleic acid (DNA) sequences, ribonucleic acid (RNA) sequences, and protein sequences, but here we focus on DNA sequences. A DNA sequence is a large molecule that carries genetic information for a specific taxon. DNA is composed of long strands of bases called *nucleotides*. There are two types of nucleotides, and there are two bases of each type. *Purines* consist of the bases adenine and guanine, noted A and G, respectively. The *pyrimidines* are composed of the other two bases, cytosine (C) and thymine (T). Each place in the sequence where a base appears is called a *site*.

Over evolutionary time, the information contained within the same gene in different taxa changes. These changes may be due to insertions, where during the DNA replication process, extra nucleotide bases are added to the sequence, or they may be due to deletions, which are removals of nucleotide bases during replication. Changes in nucleotide bases may also occur during the replication process. Point mutations result in the replacement of a nucleotide base

with another. A point mutation that changes a purine to the other purine or a pyrimidine to the other pyrimidine is a *transition*. A point mutation that changes a purine to a pyrimidine or a pyrimidine to a purine is a *transversion*. Each of these three evolutionary events alters genetic material, but to perform phylogenetic inference, it is necessary to compare portions of DNA sequences that are the same in a common ancestor. Thus, it is necessary to infer which portions of the genomes in the taxa appear in the genome of a common ancestor. This process is called *alignment*, and it must be performed on the DNA sequences before phylogenetic analysis can begin. For a brief survey of alignment algorithms, see Li and Homer (2010).

Following alignment, the DNA sequence data can be represented by a table such as Table 1.1. This table represents a set of simple DNA sequences with four sites apiece. A tree that has at its leaves the data from Table 1.1 can be seen in Figure 1.2.

Table 1.1: Table of DNA Sequences for 5 Taxa

Taxon	Site 1	Site 2	Site 3	Site 4
Taxon 1	A	C	C	A
Taxon 2	A	T	C	T
Taxon 3	G	T	T	T
Taxon 4	G	A	T	C
Taxon 5	C	C	C	C

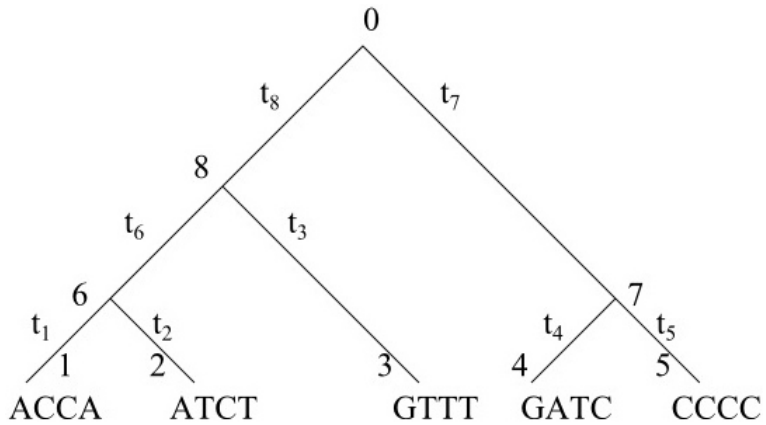


Figure 1.2: A rooted tree with DNA sequence data at the leaves. The internal nodes are denoted by the integers $0, 6, 7$ and 8. The leaves are labelled $1, 2, 3, 4,$ and 5. The branch lengths are represented by $t_1, t_2, \ldots, t_8$, where the branch length t_i corresponds to the length of the branch terminating at node i, $i = 1, 2, \ldots, 8$.

One data point $\mathbf{D}_i$ is a vector of the nucleotide bases in the DNA sequence for each taxon at the ith site. For example, $\mathbf{D}_1 =$ (A,A,G,G,C). In other words, $\mathbf{D}_i$ corresponds to the ith column of the table. Biologically, $\mathbf{D}_i$ represents a set of nucleotides, one from each DNA sequence, that are inferred to have descended from the ith site in the DNA sequence corresponding to a common ancestor. Each site is assumed to have evolved independently of all the others. At each of the tips of the tree is one of the DNA sequences in the table, and these are represented by the rows. Each time there is a split in the tree, an internal node is present. An internal node represents the most recent common ancestor among the taxa in the lineage descending from that node. In our

illustration there are four of these nodes. Between any two nodes, or between a tip and the node immediately ancestral to that tip, there is a branch having length that represents the amount of time taken by the evolution from the most recent ancestral node to the descendant node or tip, where time is taken to represent the expected number of nucleotide base substitutions per site.

1.1.2 Models of Nucleotide Substitution

Recall that we observe DNA sequences only at the tips of the tree. Our goal is to infer how, over time, evolution has given rise to the DNA sequences at the tips, and in order to do so, we need a probabilistic model that describes the evolutionary process by which one nucleotide base changes to another over time. Such a model is defined by two parameters. The first of these parameters is a matrix $\mathbf{Q}$, whose (i, j) entry represents the rate at which the nucleotide base i changes to the nucleotide base j, where $(i, j) \in \{A, G, C, T\}^2$. Let $\mathbf{P}(v)$ denote a matrix in which the (i, j) entry is the probability that base i changes to base j over v units of evolutionary time. To see the role of $\mathbf{Q}$ in the model, note that the first derivative of $\mathbf{P}(v)$ with respect to v is defined by

$$\mathbf{P}'(v) = \lim_{h \to 0^+} \frac{\mathbf{P}(v+h) - \mathbf{P}(v)}{h}. \tag{1.1}$$

Before going any further, we note that $\mathbf{P}(v)$ does not depend on the point in time at which the v units of evolutionary time began. This property is known as *time homogeneity*, and all of the models we mention here have this property. The *Chapman-Kolmogorov equation* says that if a model is time homogeneous, then for $v \geq 0$ and $u \geq 0$, $\mathbf{P}(v+u) = \mathbf{P}(v)\mathbf{P}(u)$. Applying this result to (1.1),

we see that

$$\begin{aligned}\mathbf{P}'(v) &= \lim_{h\to 0^+} \frac{\mathbf{P}(v)\mathbf{P}(h) - \mathbf{P}(v)}{h} \\ &= \mathbf{P}(v) \lim_{h\to 0^+} \frac{\mathbf{P}(h) - \mathbf{I}}{h}.\end{aligned}$$

Letting $\mathbf{Q} = \lim_{h\to 0^+} \frac{\mathbf{P}(h)-\mathbf{I}}{h}$, we obtain transition probabilities by solving the differential equation $\mathbf{P}'(v) = \mathbf{P}(v)\mathbf{Q}$, with initial condition $\mathbf{P}(0) = \mathbf{I}$. The solution to this differential equation is $\mathbf{P}(v) = e^{\mathbf{Q}v}$. The exponentiation of $\mathbf{Q}$ is not generally available by analytical methods, so numerical methods are usually required to find $\mathbf{P}$.

The second parameter of the model is a row vector π of *equilibrium probabilities* of the nucleotide bases. For a nucleotide base j and for all nucleotide bases i,

$$\pi_j = \lim_{v\to\infty} \mathbf{P}_{ij}(v).$$

In other words, π_j represents the limiting probability as $v \to +\infty$ of observing base j in a particular site given base i was in that site v time units ago. Note that π_j does not depend on i. As $v \to +\infty$, the limit is equal to the equilibrium probability of observing the base j, regardless of which base was in that site v time units ago. A sufficient condition for this limit to exist is that for all $i, j \in \{A, G, C, T\}$ and for all $v \geq 0$, $P_{ij}(v)$ is positive. The model we use in the rest of the work described in this dissertation satisfies this property, so the equilibrium probabilities exist. It can be seen that since this parameter is obtainable from the transition matrix, it is only necessary to specify the rate matrix in order to completely specify a nucleotide base substitution model.

However, in practice, the rate matrix and the equilibrium probabilities are given in the description of a model.

A substitution model is said to be *time reversible* if it satisfies the *detailed balance condition*, that is, for all $(i, j) \in \{A, G, C, T\}^2$,

$$\pi_i \mathbf{P}_{ij}(v) = \pi_j \mathbf{P}_{ji}(v)$$

for all $v \geq 0$. Many of the proposed nucleotide substitution models are time reversible. The *general time reversible* (GTR) model of nucleotide base substitution assumes six different substitution rates a_1, a_2, a_3, a_4, a_5, and a_6. The rate matrix is

$$\mathbf{Q} = \begin{array}{c} \\ A \\ G \\ C \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & G & C & T \end{array} \\ \begin{pmatrix} -J & a_1\pi_G & a_2\pi_C & a_3\pi_T \\ a_1\pi_A & -K & a_4\pi_C & a_5\pi_T \\ a_2\pi_A & a_4\pi_G & -L & a_6\pi_T \\ a_3\pi_A & a_5\pi_G & a_6\pi_C & -M \end{pmatrix} \end{array}, \text{ where}$$

$$\begin{aligned} J &= a_1\pi_G + a_2\pi_C + a_3\pi_T \\ K &= a_1\pi_A + a_4\pi_C + a_5\pi_T \\ L &= a_2\pi_A + a_4\pi_G + a_6\pi_T \\ M &= a_3\pi_A + a_5\pi_G + a_6\pi_C. \end{aligned}$$

From a biological perspective, time reversible models are justified because often, the direction of evolution is unknown. Therefore, it is possible to infer how long evolution has taken between two nodes, but it is not possible to infer which of the nodes came first. Mathematically, a reversible model is convenient,

as it allows for more efficient calculation of phylogenetic likelihood. In addition, a reversible model often fits real data closely enough that there is little to be gained by using a more general probability model (Huelsenbeck, 1998). The most significant caveat of using a reversible model, however, is that, unless additional assumptions are made, a reversible model does not allow inference of the placement of the root.

The Jukes-Cantor Model

The work we present relies on the Jukes-Cantor (JC69) model (Jukes and Cantor, 1969) of nucleotide base substitution. This model is the simplest of the time reversible models, as it assumes that the rate of substitution between any two distinct nucleotide bases is equal, so that $a_1 = a_2 = a_3 = a_4 = a_5 = a_6 = a$. The model also assumes that the equilibrium probability of each nucleotide base is 1/4. The rate matrix is

$$\mathbf{Q} = \begin{array}{c} \\ A \\ G \\ C \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & G & C & T \end{array} \\ \begin{pmatrix} -3 & 1 & 1 & 1 \\ 1 & -3 & 1 & 1 \\ 1 & 1 & -3 & 1 \\ 1 & 1 & 1 & -3 \end{pmatrix} \end{array} \frac{a}{4}.$$

This model is one for which the transition matrix is available analytically, and the (i, j) entry of the transition matrix is

$$\mathbf{P}_{ij}(v) = \begin{cases} \frac{1}{4} - \frac{1}{4}e^{-av}, & i \neq j \\ \frac{1}{4} + \frac{3}{4}e^{-av}, & i = j. \end{cases}$$

In the work we present, evolutionary time is rescaled so that v represents the expected number of nucleotide base substitutions per site. When time is scaled

in this way, $a = 4/3$, and

$$\mathbf{P}_{ij}(v) = \begin{cases} \frac{1}{4} - \frac{1}{4}e^{-4v/3}, & i \neq j \\ \frac{1}{4} - \frac{1}{4}e^{-4v/3}, & i = j. \end{cases} \tag{1.2}$$

Several other time reversible models have been proposed since the initial description of the Jukes-Cantor model. The Kimura Two-Parameter (K2P) model (Kimura, 1980), for instance, assumes that one rate of change for transitions and a different rate of change for transversions. The K2P model also assumes that the equilibrium probabilities of the nucleotide bases are equal. If the rates of change for transitions and transversions are equal, the K2P model reduces to the JC69 model.

Some other models include the HKY (Hasegawa et al., 1985) model and the F84 model, which is used in the PHYLIP (Felsenstein, 1989) phylogenetics software and is formally described by Kishino and Hasegawa (1989). Both of these models extend the K2P model by relaxing the assumption that the equilibrium probabilities of the nucleotide bases are equal. The Tamura-Nei (Tamura and Nei, 1993) model is more general than the K2P, F84, and HKY models, and it includes all three of these models as special cases.

A useful property of all the models mentioned above is that if a taxon has a particular DNA sequence, then the DNA sequence that evolves from it over the next v units of evolutionary time depends only on the current DNA sequence and the substitutions that occur over the next v time units. The sequences in the past from which the current DNA sequence evolved do not play a role in the determination of what occurs after evolution gives rise to the current sequence. This "memoryless" characteristic is known as the *Markov property.*

1.1.3 Likelihood Calculation

Recall that all we have been able to observe are the data at the tips of the tree. The tree topology is unobserved, as is the branch length vector $\mathbf{t}$. A common goal is to infer a tree topology and its branch lengths, and the likelihood function often plays a role in the inference procedure. In order to calculate the likelihood of a particular tree given our data, we first assume that evolution occurs independently in different sites and that evolution is independent among lineages.

The assumption of independence of evolution among sites allows the likelihood to be written as the product of the individual site likelihoods. The calculation of a site likelihood is completed by finding the joint probability of observing the nucleotides at the tips of the tree in that site. Since no DNA sequences are actually observed at the internal nodes, including the root, we marginalize them by summing over all possible combinations of nucleotide bases at the internal nodes. For a tree with n taxa, there are $n-1$ internal nodes for which the nucleotide base must be marginalized. Since there are four possible nucleotide bases in this site for each internal node, there are 4^{n-1} joint probabilities that need to be calculated and then summed. For a tree with even a moderately large number of taxa, this method of likelihood calculation is very inefficient because of the large number of summands. If the data set is large, the problem is compounded by a large number of site likelihood calculations.

Felsenstein (1981) presents a *peeling algorithm* to calculate phylogenetic

likelihood. This method is much faster than the brute-force method described in the preceding paragraph. The peeling algorithm makes use of the *conditional likelihood* $\mathcal{L}_s^{(i)}(k)$, which represents the probability of everything that occurs from node k down the tree, at site i, given the nucleotide base s is in site i at node k. This method is elegantly described recursively. Suppose node k has immediate descendants l and m at the bottom ends of branches with lengths t_l and t_m respectively. Then

$$\mathcal{L}_s^{(i)}(k) = \left(\sum_z \Pr(z|s, t_l) \mathcal{L}_z^{(i)}(l) \right) \left(\sum_y \Pr(y|s, t_m) \mathcal{L}_y^{(i)}(m) \right). \tag{1.3}$$

The intuition behind (1.3) is that the assumption of independence between lineages implies that the probability of everything at or below a node k given that k has base s at site i is equal to the product of the probabilities of the corresponding events in the two descendant lineages of k.

The peeling algorithm begins with the assignment of conditional likelihoods to the tips of the tree. Since data are observed at the tips, the conditional likelihood of a base s at a particular tip is assigned a value of 1 if s is observed at that tip and a value of 0 otherwise. Once conditional likelihoods are assigned at the tips, the calculation of the value in (1.3) is performed for all nodes that have only tips as their immediate descendants. This calculation is done successively for the nodes further up the tree, but it may not be carried out for any node until it has been completed for all descendant nodes. This succession continues all the way to the root node, say node 0, to obtain $\mathcal{L}_s^{(i)}(0)$. The computation of the site likelihood concludes by summing $\pi_s \mathcal{L}_s^{(i)}(0)$ over all of the nucleotide bases that can occur at the root in site i. If nodes 1 and 2

are the immediate descendant nodes of the root, then the ith site likelihood is written as

$$\mathcal{L}^{(i)}(\mathbf{D}|\mathbf{t}) = \sum_{s_0} \pi_{s_0} \left(\sum_{s_1} \Pr(s_1|s_0, t_1) \mathcal{L}^{(i)}_{s_1}(1) \right) \left(\sum_{s_2} \Pr(s_2|s_0, t_2) \mathcal{L}^{(i)}_{s_2}(2) \right). \tag{1.4}$$

Since for a full data set, we assume that evolution among sites is independent, the likelihood for a full data set having N sites is

$$\mathcal{L}(\mathbf{D}|\mathbf{t}) = \prod_{i=1}^{N} \mathcal{L}^{(i)}(\mathbf{D}|\mathbf{t}). \tag{1.5}$$

The Pulley Principle

A common assumption in phylogenetics is the assumption of a *molecular clock*, which means that specific DNA sequences spontaneously mutate at a constant rate. A consequence of this definition is that in a rooted tree, the amount of evolutionary time between each leaf and the root is the same. In our work, we make no such assumptions on the branch lengths. Felsenstein (1981) demonstrates that for a reversible model with the Markov property and no constraints on the branch lengths, the placement of the root is inconsequential to the calculation of the likelihood. In fact, the likelihood depends on the lengths t_1 and t_2 of the branches incident to the root only through their sum. This implies that the likelihood of a rooted tree and its branch lengths is equivalent to that of an unrooted tree with branch lengths $t_1 + t_2$, t_3, $\ldots$, t_{2n-2}. To see this, assume that the data set $\mathbf{D}$ has one site per sequence. The

expansion of the expression on the right-hand side of (1.4) yields

$$\mathcal{L}(\mathbf{D}|\mathbf{t}) = \sum_{s_0}\sum_{s_1}\sum_{s_2} \pi_{s_0}\Pr(s_1|s_0,t_1)\Pr(s_2|s_0,t_2)\mathcal{L}_{s_1}(1)\mathcal{L}_{s_2}(2).$$

Since the model is time reversible, we have

$$\pi_{s_0}\Pr(s_1|s_0,t_1) = \pi_{s_1}\Pr(s_0|s_1,t_1),$$

and we obtain the following.

$$\begin{aligned}
&\sum_{s_0}\sum_{s_1}\sum_{s_2} \pi_{s_0}\Pr(s_1|s_0,t_1)\Pr(s_2|s_0,t_2)\mathcal{L}_{s_1}(1)\mathcal{L}_{s_2}(2)\\
&= \sum_{s_1}\pi_{s_1}\sum_{s_2}\sum_{s_0} \Pr(s_0|s_1,t_1)\Pr(s_2|s_0,t_2)\mathcal{L}_{s_1}(1)\mathcal{L}_{s_2}(2).
\end{aligned}$$

An application of the Chapman-Kolmogorov Equation yields the result of the Pulley Principle:

$$\mathcal{L}(\mathbf{D}|\mathbf{t}) = \sum_{s_1}\sum_{s_2} \pi_{s_1}\Pr(s_1|s_2,t_1+t_2)\mathcal{L}_{s_1}(1)\mathcal{L}_{s_2}(2).$$

1.1.4 Local Tree Rearrangements

In Chapter 2 we use tree rearrangements to construct Markov chains on the space of rooted tree topologies. The two moves we consider here are the subtree prune and regraft (SPR) and the nearest neighbor interchange (NNI). An SPR is a tree move in which a branch of the current tree topology, $\mathbf{T}_1$, is broken. The broken branch, along with its associated subtree, is then attached to another branch to form a new tree topology $\mathbf{T}_2$. Figure 1.3 shows a typical SPR rearrangement. Since SPRs are most often performed on unrooted tree topologies, we adopt two conventions to handle the case of a rooted topology.

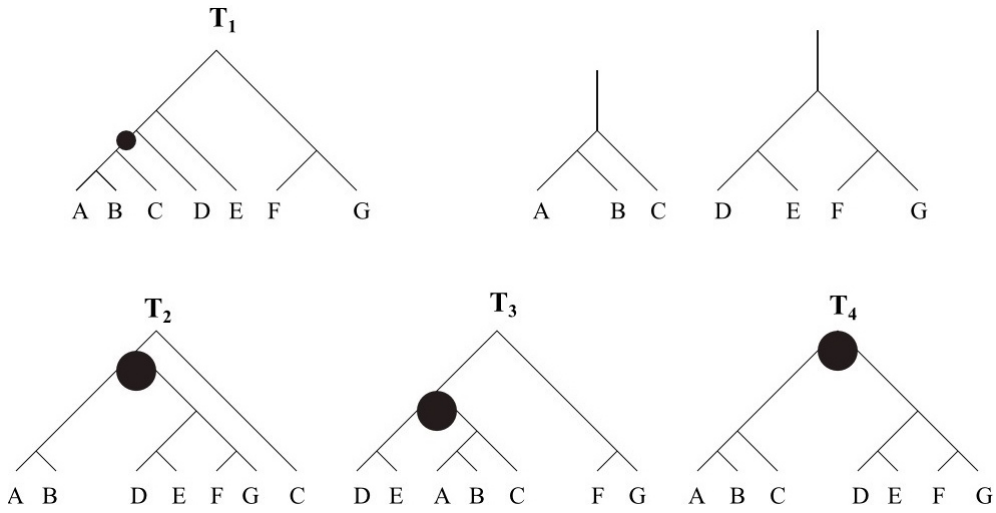


Figure 1.3: Illustrations of the three types of SPRs: $\mathbf{T}_1$ is the initial tree topology. A dot is placed on a branch of $\mathbf{T}_1$ to indicate the branch that is broken. To the right of $\mathbf{T}_1$ are the two subtrees that result from breaking the noted branch. Reattachments are indicated by a large black dot. $\mathbf{T}_2$ is the tree topology that results from attaching the subtree with leaves D, E, F, and G to the subtree with leaves A, B, and C to form the indicated node. $\mathbf{T}_3$ is the result of attaching the subtree with leaves A, B, and C to the indicated branch of the subtree that has leaves D, E, F, and G. $\mathbf{T}_4$ is the tree topology that results by connecting the two subtrees along the edges extending back from their roots.

First, extend an edge back from the root. A pruned subtree may be reattached to this edge, and the root of the new topology will be along this extended edge. An illustration of this can be seen in the forming of $\mathbf{T}_4$ in Figure 1.3. Second, it may also happen that an edge incident to the root of $\mathbf{T}_1$ is chosen to be pruned. When the reattachment occurs below the root of the remaining subtree of $\mathbf{T}_1$, the root of $\mathbf{T}_2$ is assumed to be the same as the root of the subtree of $\mathbf{T}_1$ that remained after pruning.

An NNI is performed by first choosing a non-root internal node v of $\mathbf{T}_1$ to be the *target node*. The target node has a sibling node s and two child nodes c_1 and c_2. Two nodes are chosen at random from among $s, c_1,$ and c_2, and the subtrees that descend from the two chosen nodes are interchanged to obtain $\mathbf{T}_2$. A typical NNI is illustrated in Figure 1.4.

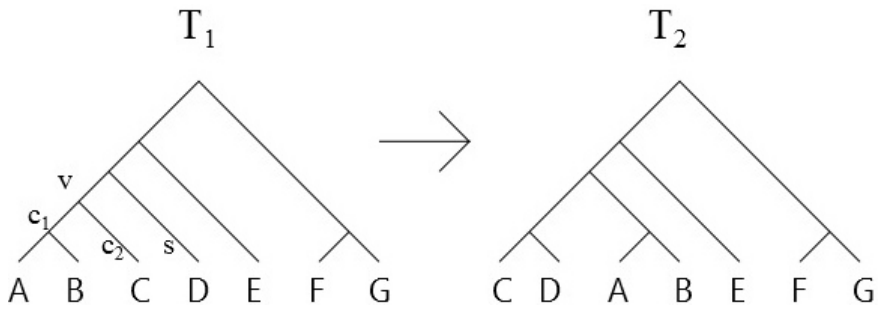


Figure 1.4: A typical NNI: The node denoted v on $\mathbf{T}_1$ has been chosen as the target. The children of v are denoted by c_1 and c_2, and the sibling node is denoted s. The descendant subtrees of c_1 and s have been interchanged to obtain $\mathbf{T}_2$.

1.2 Review of the Literature

In the literature, one finds many examples of phylogenetic analysis in practice. The information available is usually a set of data consisting of sequences of genetic material that pertains to the set of taxa about which the evolutionary history is of interest. From this information, a common task is to make an inference about not only the order of ancestry, but also the time required for one lineage to give rise to a pair of new ones. One study (Kuhnert et al.,

2000) looks at the *Pasteurella multocida* species of bacteria, which resides in the mouths of cats and dogs, and in some cases, is highly pathogenic to humans who have sustained a bite from a cat or a dog. Kuhnert et al. (2000) infer an ancestral pattern for *Pasteurella multocida* and further their analysis to identify a likely subspecies of this particular bacteria that is responsible for causing infection and illness in humans. Diezmann et al. (2004) were able to trace the origin of hemiascomycetes, which are species of yeast that are pathogenic to humans and several species of plants. They use likelihood methods as well as Bayesian methodology to infer that all species within the same taxonomic family as the hemiascomycetes likely descended from a single ancestor. Aiki-Raji et al. (2008) used phylogenetic analysis to infer the ancestral genetic characteristics that give rise to the Avian Influenza H5N1 virus in Nigeria.

The benefit of having the ability to infer ancestral patterns of particular species of organisms is clear. If it is possible to infer the evolutionary origin of a particular pathogen or virus, this knowledge can be used to gain insight into how to prevent the microbe from infecting things such as water sources, food supplies, and humans. Many other examples of such uses of phylogenetics can be found in the literature. These examples range from inferring the ancestry of certain microbial organisms such as the pathogen that leads to granuloma (Herr et al., 1999), the Influenza A virus (Chen et al., 2009), and a broad class of rapidly-evolving infectious diseases (Ross, 2011; Kuhnert et al., 2011) to inferring the genealogy of groups of larger organisms such as a class of dogs (Dowell, 2008) and a family of birds (Christensen et al., 2009).

The inference of phylogenetic trees has found a purpose in many other areas, including linguistics (Warnow et al., 2006) and network analysis (Clauset et al., 2008). In linguistics, two common assumptions are that language evolution follows a tree-like structure and that once a linguistic element changes, it changes to a state that is not yet in the tree. The second assumption represents a stark contrast to the assumptions of most models used in phylogenetic inference. However, Warnow et al. (2006) realized that the latter assumption is unnecessary. They therefore relaxed the assumption, and the result was a method of inferring the evolutionary structure of linguistic elements that very closely resembles the phylogenetic techniques that are employed in a wide variety of other areas.

Clauset et al. (2008) investigate the idea that social networks often follow a hierarchical structure. They develop a method of simultaneously inferring the topology of the network and the lengths of the paths in the networks. This is the same idea as inferring the topology and the branch lengths of a phylogenetic tree, so the model that Clauset et al. (2008) use is well-suited for phylogenetic analysis.

The ubiquity of phylogenetic inference emphasizes the need for efficient methods of inferring evolutionary patterns among a group of taxa. Some of the earliest methods of phylogenetic inference to be proposed are parsimony methods (Cavalli-Sforza and Edwards, 1963). The idea is that the preferred evolutionary tree is the one that involves the smallest net amount of evolution. At the same time, several distance methods arose, including the Unweighted Pair

Group Method with Arithmetic Mean (UPGMA) (Sokal and Sneath, 1963) and least squares (Cavalli-Sforza and Edwards, 1967). Today, maximum likelihood (Felsenstein, 1981) and Bayesian methodology are most commonly used to infer phylogenies from molecular data. Due to the complexity of phylogenetic likelihood functions and the intractability of the normalizing constant in the posterior density of a phylogenetic tree given a data set, MCMC methods have come to the forefront of phylogenetic analysis.

Our work is dedicated to establishing upper bounds on the mixing time of certain MCMC algorithms used in Bayesian inference of phylogenetic trees. We first establish upper bounds on the mixing times of two Markov chains on T_n. There are several instances in the literature in which this problem has been addressed. Randall and Tetali (2000) establish an upper bound of $\mathcal{O}(n^5 \log n)$ on the mixing time of a Markov chain that is similar to one of the chains we describe by way of a method known as Markov chain comparison. Diaconis and Holmes (2002) bound by $\mathcal{O}(n \log n)$ the mixing time of a Markov chain on rooted phylogenetic tree topologies by way of a random walk on perfect matchings. A perfect matching on a set of size $2n$ is a partition of the set into n two-element subsets. They argue that a random walk on the set of perfect matchings is isomorphic to a random walk on the set of n-leaf rooted tree topologies. A common method of bounding the time to convergence of a Markov chain on a discrete state space is to bound a closely related and more approachable quantity known as the relaxation time. Schweinsberg (2002) bounds by $\mathcal{O}(n^2)$ the relaxation time of a Markov chain on the space

of n-leaf unrooted tree topologies via the method of distinguished paths. This method involves finding a path that connects any two tree topologies and then bounding the length of the path. His work improves upon the $\mathcal{O}(n^3)$ upper bound of Aldous (2000).

In our work, the first of the two chains moves about the space T_n of rooted tree topologies through tree rearrangements that are less restrictive than those in the work of Aldous (2000) and Schweinsberg (2002). Our chain travels T_n via the SPR, while the chain studied by Aldous (2000) and Schweinsberg (2002) explores the space of unrooted n-leaf tree topologies via a special case of the SPR in which a transition is completed by the removal of one leaf from the current tree. This leaf is then attached to another edge of the tree to complete the transition. The second of the chains in our work traverses T_n via NNI transitions. We establish a $\mathcal{O}(n^{\frac{5}{2}})$ upper bound on the relaxation time of the first chain. In the process, we develop a simple path that may be useful in deciding how to explore the space of rooted tree topologies in problems where both the tree topology and the branch lengths are unknown. We also establish an upper bound of $\mathcal{O}(n^4)$ on the relaxation time of the second chain on rooted tree topologies. We develop a lower bound of $\mathcal{O}(n)$ on the relaxation time of each of the two chains.

The question of the usefulness of Markov chains that move about the space of rooted tree topologies, with no regard for branch lengths or any type of genetic data, is certainly a valid one. After all, one cannot be expected to

infer any type of evolutionary pattern when no data are available. The answer to this question is that the two chains we study are widely incorporated into algorithms designed to infer phylogenies from genetic data on the leaves. For instance, one class of algorithms employs MCMC methods to estimate the posterior distribution of trees given either DNA sequence data (Li et al., 2000; Huelsenbeck and Ronquist, 2001) or protein sequence data (Beiko et al., 2006). Another frequently used phylogenetic inference framework is maximum likelihood, which requires a search through T_n for the tree(s) that maximize the likelihood function. Standard methods for searching T_n (e.g. PHYLIP (Felsenstein, 1989), PAUP* (Swofford, 2002), RAxML (Stamatakis, 2006)) use moves similar to those defined above. Such tree moves have also been used by methods that carry out stochastic searches (e.g. SSA (Salter and Pearl, 2001) and GARLI (Zwickl, 2008)). Other applications of chains similar to ours can be seen in Yang and Rannala (1997), Guindon and Gascuel (2003), and TCS (Clement et al., 2000), a software package that is often used in population genetics. The frequency with which Markov chains on rooted tree topologies are employed highlights the need for an understanding of their rates of convergence.

The second part of this dissertation focuses on methods for bounding the mixing time of a particular MCMC algorithm used for inference of the branch lengths of a given rooted phylogenetic tree, when the tree topology is known and we have DNA sequence data available. We verify that our chain is geometrically ergodic by establishing that our chain satisfies a sufficient set of

three conditions outlined in the work of Fort et al. (2003). We then establish a minorization condition through analytical methods, and we describe Monte Carlo methods for establishing minorization and drift conditions. In many cases, drift and minorization conditions are extremely difficult, or even impossible, to verify. However, if they can be established, then doing so provides the best hope of finding an upper bound on the mixing time.

The literature shows no lack of approaches to assessing convergence by way of output-based methods. One technique that was once widely used is the thick-pen technique (Gelfand and Smith, 1990). This approach to convergence assessment relies on obtaining density estimates from various sections of the output of the chain, and these estimates come from portions of the chain which are spread far enough apart to be considered roughly independent. If these estimates differ graphically by less than the width of a thick felt-tip pen, the felt-tip pen technique does not indicate a lack of convergence. Gelman and Rubin (1992) propose a variance ratio technique that requires analysis of m independent chains $(\mathsf{X}_t^i)_{t=0}^{\infty}$, $i = 1, 2, \ldots, m$ to form an estimate of the distribution for a chosen summary statistic $\theta(\mathsf{X})$, whose value is based on the output of the chain. This gives a basis for an estimate of how close the process is to convergence by making use of the posterior variance of the means of the observations from the m sequences. Brooks and Gelman (1998) generalize this to estimation of multiple parameters through analysis of the posterior variance-covariance matrix of the m sequence mean vectors. Other procedures include methods that rely on the spectral density (Geweke, 1992), diagnostics based

on the L^2 norm (Liu et al., 1993; Roberts, 1994), a diagnostic based on the L^1 norm (Yu, 1995), and the graphical CUSUM method (Yu and Mykland, 1998). For surveys of convergence diagnostics, see Cowles and Carlin (1996) and Brooks and Roberts (1997). For more recent graphical methods of assessing Markov chain convergence for phylogenetic inference, see Li et al. (2000).

The above convergence diagnostics suffer from several drawbacks. The primary shortcoming of these methods is that none of them can actually provide an answer to the question of whether or not the chain has converged to its stationary distribution. They can only detect features that indicate a lack of convergence. Many of them are only useful for assessing the convergence of a small subset of the MCMC algorithms. If a group of convergence diagnostics are helpful for a particular MCMC algorithm, the estimates of the mixing time may vary greatly among the different convergence diagnostics. If an MCMC algorithm mixes slowly, the chosen diagnostic may suggest convergence prematurely due to the fact that, since the stationary distribution is unknown, the diagnostic must measure the distance between the sampled distributions at two different iterations instead of the distance between either of the sampled distributions and the stationary distribution. In some instances, the convergence diagnostics proposed in the literature require running the chain for thousands, or even millions, of iterations before diagnosing convergence. One may find it more helpful to have an idea of how long to run the chain *before* actually running it. The work in this dissertation represents significant progress in providing this.

1.3 Overview of Dissertation

The remainder of this dissertation is organized into four chapters. Chapter 2 is dedicated to establishing upper and lower bounds on the relaxation times of two Markov chains on rooted tree topologies. In Chapter 3, we give a description of a Markov chain that is used to approximate the posterior density of the branch lengths of a rooted phylogenetic tree given a tree topology and a data set. We then derive the posterior density, up to a normalizing constant, of the branch lengths given the tree topology and a set of DNA sequence data at the leaves, and we show that our chain satisfies the set of conditions given by Fort et al. (2003), thus ensuring it is geometrically ergodic. Chapter 4 focuses on methods of obtaining an upper bound on the mixing time of our chain. We begin by establishing a minorization condition, and then we propose methods for obtaining a lower bound on the minorization coefficient ϵ and upper bounds on the drift coefficients λ and b. The estimates of the drift and minorization coefficients provide a key step toward obtaining a useful upper bound on the mixing time of the chain. In an illustrative example, we apply the methods described to a specified 10-taxon tree. We close Chapter 4 with a discussion of how the behavior of our MCMC algorithm compares to our expectations. Chapter 5 provides a discussion of the results outlined in this work, as well as a brief description of plans for future work.

Chapter 2: Relaxation Times of Two Markov Chains on Rooted Phylogenetic Tree Topologies

In this chapter, we describe two Markov chains on the space T_n of n-leaf rooted phylogenetic tree topologies. The first of these chains moves about T_n via SPR moves, while the second explores T_n via NNI transitions. We demonstrate that the relaxation time of the SPR chain is bounded above by $\mathcal{O}(n^{5/2})$ and that the relaxation time of the NNI chain is bounded above by $\mathcal{O}(n^4)$. The chapter concludes with a derivation of a lower bound on the relaxation times of each of the two chains as well as a description of the link between the relaxation time and the mixing time of a given Markov chain.

2.1 Preliminaries

This section gives the background that is essential to understanding the rest of the work presented in this chapter. We begin by providing information about finite-state Markov chains. This is followed by a statement and proof of a result that is useful in deriving upper bounds on the convergence rates of each of the two chains we describe later in the chapter.

Let $(\mathsf{X}_t)_{t=0}^{\infty}$ be a sequence of random variables, and let Ω be a finite set.

Then $(\mathsf{X}_t)_{t=0}^{\infty}$ is a *finite-state Markov chain* if for all $x, y \in \Omega$, all integers $t \geq 1$, and all events $H_{t-1} = \bigcap_{s=0}^{t-1} \{\mathsf{X}_s = x_s\}$,

$$\Pr\left(\mathsf{X}_{t+1} = y | \{\mathsf{X}_t = x\} \cap H_{t-1}\right) = \Pr(\mathsf{X}_{t+1} = y | \mathsf{X}_t = x). \tag{2.1}$$

Let $\mathbf{P}$ be a $|\Omega| \times |\Omega|$ square matrix, where $|\Omega|$ denotes the number of elements in Ω. Then $\mathbf{P}$ is the *transition matrix* for $(\mathsf{X}_t)_{t=0}^{\infty}$ if for all $x, y \in \Omega$, the (x, y) entry of $\mathbf{P}$ is

$$\mathbf{P}_{xy} = \Pr(\mathsf{X}_{t+1} = y | \mathsf{X}_t = x).$$

For an integer $k \geq 1$, the k-step transition matrix $\mathbf{P}^k$ has as its (x, y) entry

$$\mathbf{P}^k_{xy} = \Pr(\mathsf{X}_{t+k} = y | \mathsf{X}_t = x).$$

Note that since the xth row of $\mathbf{P}$ gives the conditional distribution the conditional distribution of X_{t+1} given $\mathsf{X}_t = x$ for each $x \in \Omega$, the sum of the entries in the xth row of $\mathbf{P}$ is 1. A matrix for which the sum of the entries in each row is 1 is a *stochastic matrix*.

A desirable property of a finite-state Markov chain is the ability to reach any state $y \in \Omega$ from any other state $x \in \Omega$ in a finite number of steps. This property is called *irreducibility*, and it is defined as follows.

Definition 1. A finite-state Markov chain $(\mathsf{X}_t)_{t=0}^{\infty}$ with state space Ω and transition matrix $\mathbf{P}$ is irreducible if for all $x, y \in \Omega$, there exists an integer $t \geq 1$ such that

$$\mathbf{P}^t(x, y) > 0.$$

Another useful property of a Markov chain is *aperiodicity*, which is defined here.

Definition 2. Let $\mathcal{T}(x) := \{t \geq 1 : \mathbf{P}^t(x,x) > 0\}$ be the set of times at which it is possible for the chain to return to its initial state x. The *period* of the state x is defined to be the greatest common divisor of $\mathcal{T}(x)$. If the period of all states is 1, then $(\mathsf{X}_t)_{t=0}^{\infty}$ is *aperiodic*.

Lemma 1. *(Levin et al., 2009) If $(\mathsf{X}_t)_{t=0}^{\infty}$ is an irreducible finite-state Markov chain with state space Ω, then for all $x, y \in \Omega$, gcd $\mathcal{T}(x)$ = gcd $\mathcal{T}(y)$.*

Lemma 1 says that for an irreducible, finite-state Markov chain, all states have the same period. Therefore, when the chain is irreducible, aperiodicity can be established by showing that a particular state has period 1.

Definition 3. Let π be a probability distribution on Ω. For a finite-state Markov chain $(\mathsf{X}_t)_{t=0}^{\infty}$ with state space Ω and transition matrix $\mathbf{P}$, π is a *stationary distribution* if

$$\pi\mathbf{P} = \pi.$$

The properties of irreducibility and aperiodicity lead to the following result pertaining to the stationary distribution of a finite-state Markov chain.

Lemma 2. *(Karlin and Taylor, 1975) Suppose $(\mathsf{X}_t)_{t=0}^{\infty}$ is an irreducible and aperiodic finite-state Markov chain with state space Ω and transition matrix $\mathbf{P}$. Then $(\mathsf{X}_t)_{t=0}^{\infty}$ has a unique stationary distribution π with the property that for all $x, y \in \Omega$,*

$$\lim_{k\to\infty} \mathbf{P}^k_{xy} = \pi_y,$$

where π_y is the stationary probability of state y.

Note that limit in Lemma 2 is independent of the initial state x. This means that as k gets large, the rows of $\mathbf{P}^k$ become more similar to the row vector π. Furthermore, if $\mathbf{P}$ is symmetric, then the unique stationary distribution is known.

Lemma 3. *If $(\mathsf{X}_t)_{t=0}^{\infty}$ is an irreducible finite-state Markov chain with state space Ω and symmetric transition matrix $\mathbf{P}$, then the unique stationary distribution for $(\mathsf{X}_t)_{t=0}^{\infty}$ is the discrete uniform distribution on Ω.*

Proof. Since $(\mathsf{X}_t)_{t=0}^{\infty}$ is irreducible, the stationary distribution is unique by Lemma 2. Recall that the stationary probability of a state y is given by

$$\pi_y = \sum_{x\in\Omega} \pi_x \mathbf{P}_{xy}$$

Suppose $\pi_x = \frac{1}{|\Omega|}$ for all $x \in \Omega$. Then

$$\begin{aligned} \pi_y &= \frac{1}{|\Omega|} \sum_{x\in\Omega} \pi_x \mathbf{P}_{xy} \\ &= \frac{1}{|\Omega|} \sum_{x\in\Omega} \mathbf{P}_{yx} \\ &= \frac{1}{|\Omega|}. \end{aligned}$$

Therefore, the uniform distribution is stationary for $(\mathsf{X}_t)_{t=0}^{\infty}$. □

Lemma 4. *(Karlin and Taylor, 1975) For a finite-state Markov chain $(\mathsf{X}_t)_{t=0}^{\infty}$ with state space Ω and transition matrix $\mathbf{P}$, suppose that for all $x, y \in \Omega$, $(\mathsf{X}_t)_{t=0}^{\infty}$ satisfies the* detailed balance equations*:*

$$\pi_x \mathbf{P}_{xy} = \pi_y \mathbf{P}_{yx} \text{ for all } x, y \in \Omega \tag{2.2}$$

for some probability distribution π. Then π is a stationary distribution for $(\mathsf{X}_t)_{t=0}^{\infty}$.

If $(\mathsf{X}_t)_{t=0}^{\infty}$ satisfies (2.2) and has π as the initial distribution (i.e. $\mathsf{X}_0 \sim \pi$), then the distribution of $(\mathsf{X}_0, \mathsf{X}_1, \ldots, \mathsf{X}_n)$ is the same as the distribution of $(\mathsf{X}_n, \mathsf{X}_{n-1}, \ldots, \mathsf{X}_0)$. A Markov chain with this property is called *reversible*, and often, the simplest way to show that a chain is reversible is to verify that it satisfies the detailed balance equations.

A common question pertaining to Markov chains is the question of how long a chain takes to become close to its stationary distribution. The answer to this question is based on a quantity known as the total variation distance, which measures the distance of the chain from its stationary distribution after n steps by comparing π to the n-step transition matrix $\mathbf{P}^n$.

Definition 4. The total variation distance between the two probability distributions μ and ν on Ω is defined by

$$\|\mu - \nu\|_{TV} = \max_{A \subset \Omega} |\mu(A) - \nu(A)|.$$

The following lemma provides an expression for the total variation distance between two probability distributions that does not involve taking a maximum over all $2^{|\Omega|}$ subsets of Ω.

Lemma 5. *(Levin et al., 2009) Let μ and ν be two probability distributions on Ω. Then*

$$\|\mu - \nu\|_{TV} = \frac{1}{2} \sum_{x \in \Omega} |\mu(x) - \nu(x)|. \tag{2.3}$$

The time a chain takes to become close to its stationary distribution in total variation distance is the *mixing time*, and it is defined as follows.

Definition 5. For fixed $\epsilon > 0$ the mixing time of a finite-state Markov chain with state space Ω and transition matrix $\mathbf{P}$ is defined by

$$\tau_{mix}(\epsilon) := \min\left\{k : \|\mathbf{P}^k - \pi\|_{TV} \leq \epsilon\right\}.$$

For a Markov chain on a finite state space, a commonly chosen value of ϵ is 1/4 (Bayer and Diaconis, 1992; Levin et al., 2009), so that $\tau_{mix} = \tau_{mix}(1/4)$.

The form of the total variation distance given in Lemma 5 is not efficiently calculated in large state spaces. In some other cases, including one of the chains we describe, the transition matrix is difficult to write down, making calculation of the total variation distance difficult. In order to avoid this, we use another measure of the convergence rate. This measure is known as the *relaxation time*, and in certain situations, an upper bound on the relaxation time can be used to obtain an upper bound on the mixing time.

Definition 6. For a reversible, irreducible, and aperiodic finite-state Markov chain $(\mathsf{X}_t)_{t=0}^{\infty}$ with state space Ω and transition matrix $\mathbf{P}$, let $1 = \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{|\Omega|}$ be the eigenvalues of $\mathbf{P}$ in decreasing order. The relaxation time of $(\mathsf{X}_t)_{t=0}^{\infty}$ is defined to be

$$\tau_{rel}(\mathsf{X}) := \frac{1}{1-\lambda_2}.$$

The quantity $1 - \lambda_2$ is the *spectral gap* of $\mathbf{P}$.

The rest of this section consists of proving a result that is useful in the derivation of the upper bounds on the relaxation times of the chains we describe in this chapter. Let $\mathsf{L}_{\mathbf{x}}$ denote the set of leaves of a tree $\mathbf{x} \in \mathsf{T}_n$. Any two leaves $i, j \in \mathsf{L}_{\mathbf{x}}$ are connected by a unique path δ_{ij} that does not intersect itself and does not traverse any edge more than once. An example of a path of the type defined above can be seen in Figure 2.1. The diameter of $\mathbf{x}$, denoted $\mathrm{diam}(\mathbf{x})$,

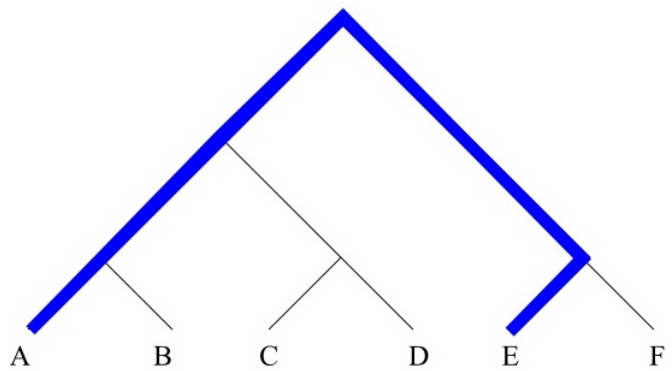


Figure 2.1: A six-leaf rooted tree topology having diameter 5. One of the paths that traverses five edges is the path from leaf A to leaf E, and this path is highlighted in blue. Similar paths connecting leaves A and F, leaves B and E, and leaves B and F also traverse five edges.

is the number of edges in the longest such path between any two leaves. Let $|\delta_{ij}|$ be the number of edges in the path δ_{ij}. Then

$$\mathrm{diam}(\mathbf{x}) := \max_{(i,j) \in \mathsf{L}_{\mathbf{x}} \times \mathsf{L}_{\mathbf{x}}} |\delta_{ij}|.$$

The following lemma states that a typical element of T_n has diameter of order no larger than $\mathcal{O}(\sqrt{n})$.

Lemma 6. *There exists a constant $C_1 < \infty$, which does not depend on n, such that if π denotes the discrete uniform probability measure on T_n, then the median diameter with respect to π is no larger than $C_1\sqrt{n}$.*

Proof. Schweinsberg (2002) shows that if π^* is the uniform distribution on the space of n-leaf unrooted tree topologies, then there exists a constant $C_2 < \infty$ that does not depend on n such that the median diameter with respect to π^* is no larger than $C_2\sqrt{n}$. Since an n-leaf rooted tree topology can be viewed as an n-leaf unrooted tree topology with a root inserted along a specific edge, it follows that rooting an unrooted tree increases the diameter by no more than one. Thus, the median diameter with respect to π is no larger than $(C_2 + 1)\sqrt{n}$. □

2.2 The SPR Chain

In this section, we provide an upper bound on the relaxation time of $(\mathsf{X}_t)_{t=0}^{\infty}$, the chain that explores T_n via SPR moves. In order to do so, we make use of the distinguished paths method. This approach has been used in many settings (see Jerrum and Sinclair (1989), Diaconis and Stroock (1991), and Schweinsberg (2002) for examples). Let V be a subset of the leaf labels $\{1, 2, \ldots, n\}$ of $\mathbf{y} \in \mathsf{U}_n$. Now consider the tree that results from removing all leaves whose labels are not in V from $\mathbf{y}$. This tree is termed the *V-tree derived from* $\mathbf{y}$. The following lemma is proven in Schweinsberg (2002).

Lemma 7. *(Schweinsberg, 2002) If V is a k-element subset of $\{1, 2, \ldots, n\}$ and $\mathbf{y}$ is a uniform random unrooted n-leaf tree, then the V-tree derived from $\mathbf{y}$ is a uniform random unrooted k-leaf tree.*

Since a rooted n-leaf tree topology can be viewed as an unrooted $(n+1)$-leaf tree topology with the $(n+1)^{st}$ leaf extending from the root, it is clear that Lemma 7 holds for rooted tree topologies.

Let $\mathbf{P}_{\mathsf{X}}$ be the transition matrix corresponding to $(\mathsf{X}_t)_{t=0}^{\infty}$. Obtaining $\mathbf{P}_{\mathsf{X}}$ exactly is difficult, but it is possible to obtain useful lower bounds on the transition probabilities. Recall that an SPR is a tree rearrangement in which a randomly selected branch is broken, leaving two subtrees. One of the subtrees is attached by the edge extending from its root to a randomly chosen branch of the other subtree. Since the number of distinct SPRs that result in the same tree is difficult to find, the exact transition probabilities for $(\mathsf{X}_t)_{t=0}^{\infty}$ are hard to obtain. However, the following lemma provides information about the transition probabilities for $(\mathsf{X}_t)_{t=0}^{\infty}$ that is useful in bounding the relaxation time.

Lemma 8. *For $\mathbf{x}, \mathbf{y} \in \mathsf{T}_n$,*

$$\begin{aligned}
\mathbf{P}_{\mathsf{X}}(\mathbf{x}, \mathbf{y}) &= \frac{2n-2}{(4n-3)(n-2)} \text{ if } \mathbf{y} = \mathbf{x} \\
\mathbf{P}_{\mathsf{X}}(\mathbf{x}, \mathbf{y}) &\geq \frac{1}{(4n-3)(n-2)} \text{ if } \mathbf{y} \sim \mathbf{x} \\
\mathbf{P}_{\mathsf{X}}(\mathbf{x}, \mathbf{y}) &= 0, \text{ otherwise.}
\end{aligned}$$

Proof. For topologies $\mathbf{x}, \mathbf{y} \in \mathsf{T}_n$, consider three cases that may arise in the completion of a rooted SPR.

(1) An external edge of $\mathbf{x}$ is removed.

(2) An internal edge of $\mathbf{x}$ that is not incident to the root is broken.

(3) An internal edge of $\mathbf{x}$ that is incident to the root is broken.

Case 1: There are n leaves from which to choose, and $2n-3$ edges to which the pruned leaf can be reattached. One of the reattachments results in $\mathbf{y} = \mathbf{x}$, so that there are $n(2n-4)$ SPRs under Case 1 that result in some topology $\mathbf{y} \sim \mathbf{x}$, where $\mathbf{y} \sim \mathbf{x}$ is taken to mean that $\mathbf{y}$ can be reached in one SPR from $\mathbf{x}$ and that $\mathbf{y} \neq \mathbf{x}$. In addition, there are n SPRs under Case 1 that result in no change to the labelled tree topology.

Case 2: Suppose an internal edge that is not incident to the root of $\mathbf{x}$ is broken. Let the subtree that has been removed be denoted $\mathbf{c}_1$ and the remaining subtree be called $\mathbf{c}_2$. Allowing $n_i,\ i = 1, 2$ to denote the number of leaves in $\mathbf{c}_i$, there are $2n_2 - 1$ edges of $\mathbf{c}_2$ along which $\mathbf{c}_1$ can be reattached. Similarly, $\mathbf{c}_2$ can be regarded as the pruned subtree, and there are $2n_1 - 1$ edges of $\mathbf{c}_1$ to which $\mathbf{c}_2$ can be reattached. Of these, n_1 are leaves and two edges are incident to the root. In $\mathbf{x}$ there are $n-4$ edges that are neither leaves nor are incident to the root. Given an edge that has been broken, there are $(2n_1 - 1) + (2n_2 - 1)$ possible SPRs. Of these, one results in $\mathbf{y} = \mathbf{x}$. In total, $(2n-2)(n-4)$ SPRs are possible in Case 2. Therefore, $(2n-3)(n-4)$ SPRs in this case result in $\mathbf{y} \sim \mathbf{x}$, while $(n-4)$ SPRs result in $\mathbf{y} = \mathbf{x}$.

Case 3: When an edge that is incident to the root is broken, the reasoning is similar to that used to establish the number of possible SPRs from Case 2. The only differences are that there are two edges that are incident to the root, and regardless of which is broken, the same two subtrees $\mathbf{c}_1$ and $\mathbf{c}_2$ result. Therefore, there are $2n-2$ possible SPRs from Case 3, two of which result in $\mathbf{y} = \mathbf{x}$. The remaining $(2n-4)$ result in $\mathbf{y} \sim \mathbf{x}$.

Summing the number of possible SPRs from $\mathbf{x}$ in each case gives a total of $(4n-3)(n-2)$ SPRs, of which $2n-2$ result in a return to $\mathbf{x}$ and the remaining SPRs yield a tree topology $\mathbf{y} \sim \mathbf{x}$. $\square$

Define $\mathsf{E} = \{e \equiv (\mathbf{x}, \mathbf{y}) : \mathbf{P}_{\mathsf{X}}(\mathbf{x}, \mathbf{y}) > 0\}$ to be the set of edges that connect two tree topologies $\mathbf{x}, \mathbf{y} \in \mathsf{T}_n$ such that $\mathbf{y} \sim \mathbf{x}$, and let $G = (\mathsf{T}_n, \mathsf{E})$ be the underlying graph corresponding to $\mathbf{P}_{\mathsf{X}}$. The vertices of this graph are the elements of T_n, and the edges connect topologies for which $\mathbf{P}_{\mathsf{X}}(\mathbf{x}, \mathbf{y}) > 0$. The result of Lemma 8 implies that $(\mathsf{X}_t)_{t=0}^{\infty}$ is aperiodic, since there is a positive probability at each transition that the state of the chain does not change. Irreducibility comes from the construction of the SPR path. That construction implies that there is a positive probability of moving from any given tree topology to any other given tree topology in no more than $n-1$ steps. This means that $(\mathsf{X}_t)_{t=0}^{\infty}$ is irreducible, so its stationary distribution is unique. In addition, the fact that the SPR is a symmetric move implies that $\mathbf{P}_{\mathsf{X}}$ is symmetric. Therefore, the unique stationary distribution for $(\mathsf{X}_t)_{t=0}^{\infty}$ is the discrete uniform distribution on T_n. Letting π denote the stationary distribution, we have that for all $\mathbf{t} \in \mathsf{T}_n$, $\pi_{\mathbf{t}} = 1/c_n$, where c_n is the cardinality of T_n. The

value of c_n is given by

$$c_n = (2n-3)!! = \prod_{i=1}^{n-1}(2i-1).$$

Given an edge $e \in \mathsf{E}$, we set $Q(e) = Q(\mathbf{x}, \mathbf{y}) = \pi_{\mathbf{x}}\mathbf{P}_{\mathsf{X}}(\mathbf{x}, \mathbf{y})$, where $\mathbf{x}$ and $\mathbf{y}$ are the vertices that are connected by e. The following lemma is central to our derivation of an upper bound on $\tau_{rel}(\mathsf{X})$.

Lemma 9. *(Schweinsberg, 2002) Let $\mathsf{B} \subset \mathsf{T}_n$. Suppose that for all $\mathbf{x} \in \mathsf{T}_n$ and $\mathbf{y} \in \mathsf{B}$, $\gamma_{\mathbf{xy}}$ is a path in G, possibly random, from $\mathbf{x}$ to $\mathbf{y}$ that has at most L edges. Then*

$$\tau_{rel}(\mathsf{X}) \leq \frac{4L}{\pi(\mathsf{B})} \max_{e \in \mathsf{E}} \left\{ \frac{1}{Q(e)} \sum_{\mathbf{x} \in \mathsf{T}_n} \sum_{\mathbf{y} \in \mathsf{B}} \pi_{\mathbf{x}} \pi_{\mathbf{y}} Pr(e \in \gamma_{\mathbf{xy}}) \right\}. \tag{2.4}$$

We are now ready to state the first of the two main results of this chapter.

Theorem 1. *There exists a finite constant M_1 such that $\tau_{rel}(\mathsf{X}) \leq M_1 n^{\frac{5}{2}}$.*

Proof. To establish Theorem 1, we take an approach that consists of analyzing each of the factors appearing on the right hand side of (2.4) individually. To begin, choose a subset $\mathsf{B} \subset \mathsf{T}_n$ to be the set of all rooted n-leaf tree topologies that have diameter no larger than $\mathcal{O}(\sqrt{n})$. Let $R = \{r_1, r_2, \ldots, r_n\}$ be a uniform random permutation of the leaf labels of $\mathbf{x}$. To obtain an upper bound on the length of a path from $\mathbf{x} \in \mathsf{T}_n$ to $\mathbf{y} \in \mathsf{B}$, we construct a random path $\gamma_{\mathbf{xy}} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m)$ such that $\mathbf{x}_1 = \mathbf{x}$ and $\mathbf{x}_m = \mathbf{y}$. At step k of the path, tree $\mathbf{x}_{k+1}$ is formed by removing a leaf from $\mathbf{x}_k$ and reattaching it in such a way that:

(1) The $\{r_1, \ldots, r_{k+1}\}$-tree derived from $\mathbf{x}_{k+1}$ is the same as the $\{r_1, \ldots, r_{k+1}\}$-tree derived from $\mathbf{y}$;

(2) The $\{r_1, r_{k+1}, \ldots, r_n\}$-tree derived from $\mathbf{x}_{k+1}$ is the same as the $\{r_1, r_{k+1}, \ldots, r_n\}$-tree derived from $\mathbf{x}$.

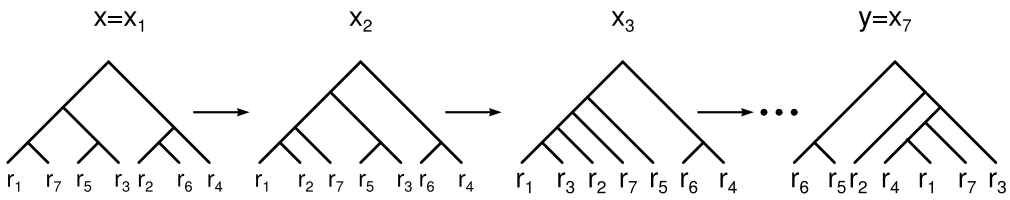


Figure 2.2: An illustration of the first two steps of the SPR path from $\mathbf{x}$ (leftmost tree) to $\mathbf{y}$ (rightmost tree). In the first step, the leaf labelled r_2 has been removed and re-attached to the branch immediately ancestral to the leaf r_1. In the second step, the leaf labelled r_3 has been removed and re-attached to the branch immediately ancestral to leaf r_1. Leaves $r_4, \ldots, r_7$ are subsequently removed and reattached in a similar fashion to obtain $\mathbf{x}_7 = \mathbf{y}$.

For some set $A \subset R$ and tree $\mathbf{x} \in \mathsf{T}_n$, let $\mathbf{x}(A)$ be the A-tree derived from $\mathbf{x}$. The path begins with $\mathbf{x}_1 = \mathbf{x}$. The first step of the path consists of removing the leaf with label r_2 and attaching it to the leaf labelled r_1. The resulting tree is denoted $\mathbf{x}_2$. The tree $\mathbf{x}_2$ contains a subtree $\mathbf{s}_2$, which we define to be the rooted subtree that contains only the leaves r_1 and r_2 and the edge extending from its root. Assume that for $k \geq 2$, we have constructed a tree $\mathbf{x}_k$ such that it contains a subtree $\mathbf{s}_k$ with the properties that

(i) the root of $\mathbf{s}_k$ is the most recent common ancestor of $\{r_1, r_2, \ldots, r_k\}$;

(ii) $\mathbf{x}_k(\{r_1, r_2, \ldots, r_k\}) = \mathbf{y}(\{r_1, r_2, \ldots, r_k\})$;

(iii) $\mathbf{s}_k$ includes an edge ascending from its root.

The tree $\mathbf{x}_{k+1}$ is obtained by removing the leaf r_{k+1} from $\mathbf{x}_k$ and attaching it to the unique edge of $\mathbf{s}_k$ that produces $\mathbf{x}_{k+1}$ with the property that $\mathbf{x}_{k+1}(\{r_1, r_2, \ldots, r_{k+1}\}) = \mathbf{y}(\{r_1, r_2, \ldots, r_{k+1}\})$. At step $n-1$, we complete the path by removing leaf r_n from $\mathbf{x}_{n-1}$ and attaching it to $\mathbf{s}_{n-1}$ in such a manner that $\mathbf{x}_n = \mathbf{s}_n = \mathbf{y}$. The first two steps of a typical path of this type are illustrated in Figure 2.2.

This approach defines a path $\gamma_{\mathbf{xy}}$ from $\mathbf{x}$ to $\mathbf{y}$. If an edge $e \in \mathsf{E}$ connects $\mathbf{x}_k$ to $\mathbf{x}_{k+1}$ in the path $\gamma_{\mathbf{xy}}$, we say $e \in \gamma_{\mathbf{xy}}$ at the kth step. Since no leaf has to be moved more than once in order to complete the path $\gamma_{\mathbf{xy}}$, it follows that $|\gamma_{\mathbf{xy}}| \leq n-1$ for all $\mathbf{x}, \mathbf{y} \in \mathsf{T}_n$. Lemma 6 gives a lower bound of $1/2$ on $\pi(\mathsf{B})$. By Lemma 9,

$$\tau_{rel}(\mathsf{X}) \leq \frac{4(n-1)}{\pi(\mathsf{B})} \max_{e \in \mathsf{E}} \left\{ \frac{1}{Q(e)} \sum_{\mathbf{x} \in \mathsf{T}_n} \sum_{\mathbf{y} \in \mathsf{B}} \pi_{\mathbf{x}} \pi_{\mathbf{y}} \Pr(e \in \gamma_{\mathbf{xy}}) \right\}.$$

If $\mathbf{P}_{\mathsf{X}}(\mathbf{x}, \mathbf{y}) > 0$, then by Lemma 8, $Q(e) \geq 1/[(4n-3)(n-2)c_n]$, which implies that

$$\begin{aligned} \tau_{rel}(\mathsf{X}) &\leq 8(n-1)c_n(4n-3)(n-2) \max_{e \in \mathsf{E}} \sum_{\mathbf{x} \in \mathsf{T}_n} \sum_{\mathbf{y} \in \mathsf{B}} \pi_{\mathbf{x}} \pi_{\mathbf{y}} \Pr(e \in \gamma_{\mathbf{xy}}) \\ &\leq 32n^3 c_n \max_{e \in \mathsf{E}} \sum_{\mathbf{x} \in \mathsf{T}_n} \sum_{\mathbf{y} \in \mathsf{B}} \pi_{\mathbf{x}} \pi_{\mathbf{y}} \Pr(e \in \gamma_{\mathbf{xy}}). \end{aligned}$$

Consider the set $K(e) = \{k : \Pr(e \in \gamma_{\mathbf{xy}} \text{ at step } k) > 0 \text{ for some } \mathbf{x} \in \mathsf{T}_n, \mathbf{y} \in \mathsf{B}\}$. Schweinsberg (2002) shows that in the case of unrooted trees, $|K(e)|=$

$\mathcal{O}(n^{\frac{1}{2}})$. For rooted trees, the argument is similar. Therefore,

$$\begin{aligned}
&\tau_{rel}(\mathsf{X}) \\
&\leq 32n^3 c_n \max_{e\in\mathsf{E}} \sum_{\mathbf{x}\in\mathsf{T}_n}\sum_{\mathbf{y}\in\mathsf{B}} \pi_{\mathbf{x}}\pi_{\mathbf{y}} \left(\sum_{k\in K(e)} \Pr(e \in \gamma_{\mathbf{xy}} \text{ at step } k) \right) \\
&\leq 32C_3 n^{\frac{7}{2}} c_n \max_{e\in\mathsf{E}} \max_{k\in K(e)} \sum_{\mathbf{x}\in\mathsf{T}_n}\sum_{\mathbf{y}\in\mathsf{T}_n} \pi_{\mathbf{x}}\pi_{\mathbf{y}} \Pr(e \in \gamma_{\mathbf{xy}} \text{ at step } k) \qquad (2.5)
\end{aligned}$$

for some positive constant $C_3 < \infty$. We derive a further upper bound on (2.5) in the following way. Assume that $\mathbf{x}, \mathbf{y} \in \mathsf{T}_n$ are independent uniform random n-leaf tree topologies and that $\{r_1, r_2, \ldots, r_n\}$ is a uniform random permutation of the leaf labels. We consider a fixed edge $e \in \mathsf{E}$ and a fixed $k \in K(e)$. Let $\mathbf{v}$ and $\mathbf{w}$ be the trees connected by edge e so that e is on the path $\gamma_{\mathbf{xy}}$ at step k. In other words, let $\mathbf{v}$ and $\mathbf{w}$ be such that $\mathbf{x}_k = \mathbf{v}$ and $\mathbf{x}_{k+1} = \mathbf{w}$. In this case, there are three independent events that must occur:

a) The subtree $\mathbf{s}_k$ contains the leaves $r_1, r_2, \ldots, r_k$ and the leaf being moved is r_{k+1}.

b) The $\{r_1, r_2, \ldots, r_{k+1}\}$-tree derived from $\mathbf{y}$ is the same as the $\{r_1, r_2, \ldots, r_{k+1}\}$-tree derived from $\mathbf{w}$.

c) The $\{r_1, r_{k+1}, r_{k+2}, \ldots, r_n\}$-tree derived from $\mathbf{x}$ is the same as the $\{r_1, r_{k+1}, r_{k+2}, \ldots, r_n\}$-tree derived from $\mathbf{v}$.

Event a) has probability $1/\binom{n}{k} \times 1/(n-k)$ because $r_1, r_2, \ldots r_n$ is a random permutation of the leaf labels. Events b) and c) have probabilities $1/c_{k+1}$ and $1/c_{n-k+1}$, respectively, by applying Lemma 7. Thus, we have

$$\Pr(e \in \gamma_{\mathbf{xy}} \text{ at step } k) \leq \frac{1}{\binom{n}{k}(n-k)c_{k+1}c_{n-k+1}}. \qquad (2.6)$$

As a result of (2.6),

$$\sum_{\mathbf{x}\in\mathsf{T}_n}\sum_{\mathbf{y}\in\mathsf{T}_n}\pi_{\mathbf{x}}\pi_{\mathbf{y}}\Pr(e\in\gamma_{\mathbf{xy}}\text{ at step }k)\leq\frac{1}{\binom{n}{k}(n-k)c_{k+1}c_{n-k+1}}. \tag{2.7}$$

Combining (2.5) and (2.7), we get

$$\tau_{rel}(\mathsf{X}) \;\leq\; 32C_3n^{\frac{7}{2}}\frac{c_n}{\binom{n}{k}(n-k)(2(n-k)-1)(2k-1)c_kc_{n-k}}.$$

Stirling's formula gives an approximate value of $c_n\approx 2^nn^{n-1}e^{-n}$, where $a\approx b$ is taken to mean that the ratio a/b is bounded away from 0 and ∞ as n varies. We also make use of the following well-known approximation to the binomial coefficient:

$$\binom{n}{k}\approx\frac{n^{n+1/2}}{k^{k+1/2}(n-k)^{(n-k)+1/2}}. \tag{2.8}$$

Applying Stirling's approximation and (2.8), we see that there exists a positive constant $C_4<\infty$ such that

$$\begin{aligned}
\tau_{rel}(\mathsf{X}) &\leq 32C_3C_4n^{\frac{7}{2}}\frac{n^{n-1}}{\binom{n}{k}(n-k)(2(n-k)-1))(2k-1)k^{k-1}(n-k)^{n-k-1}}\\
&\approx 32C_3C_4n^{\frac{7}{2}}\frac{n^{n-1}k^{k+1/2}(n-k)^{n-k+1/2}}{(2k-1)(2(n-k)-1)n^{n+1/2}k^{k-1}(n-k)^{n-k-1}}\\
&= 32C_3C_4n^{\frac{7}{2}}\frac{n^{-3/2}k^{3/2}(n-k)^{1/2}}{(2k-1)(2(n-k)-1)}\\
&\leq 32C_3C_4n^{\frac{7}{2}}n^{-3/2}k^{1/2}(n-k)^{-1/2}\\
&\leq 32C_3C_4n^{\frac{7}{2}}n^{-3/2}n^{1/2}\\
&= \mathcal{O}(n^{\frac{5}{2}}),
\end{aligned}$$

thus establishing the desired result. □

2.3 The NNI Chain

An argument similar to the one used to derive an upper bound on the relaxation time of the Markov chain $(\mathsf{X}_t)_{t=0}^{\infty}$, which explores T_n via SPR transitions, can be used to derive an upper bound on the relaxation time of $(\mathsf{Y}_t)_{t=0}^{\infty}$. In order to apply Lemma 9 to obtain an upper bound on the relaxation time we need the transition matrix of $(\mathsf{Y}_t)_{t=0}^{\infty}$. In order to ensure that $(\mathsf{Y}_t)_{t=0}^{\infty}$ is aperiodic, we specify a 1/2 probability at each transition that the chain remains in its current state and that with probability 1/2 the chain moves to another tree topology by way of an NNI that is chosen uniformly at random. The following lemma gives the transition probabilities for $(\mathsf{Y}_t)_{t=0}^{\infty}$.

Lemma 10. *Let* $\mathbf{P}_{\mathsf{Y}}$ *denote the transition matrix for* (Y_t)*. For two trees* $\mathbf{x}$ *and* $\mathbf{y}$ *in* T_n*,*

$$\mathbf{P}_{\mathsf{Y}}(\mathbf{x},\mathbf{y}) = \begin{cases} \frac{1}{2} & \textit{if } \mathbf{y} = \mathbf{x} \\ \frac{1}{4(n-2)} & \textit{if } \mathbf{y} \sim \mathbf{x} \\ 0 & \textit{otherwise}, \end{cases}$$

where $\mathbf{y} \sim \mathbf{x}$ *is taken to mean that* $\mathbf{y}$ *can be reached from* $\mathbf{x}$ *in one NNI and* $\mathbf{x} \neq \mathbf{y}$*.*

Proof. The 1/2 probability of remaining at the same tree comes from the construction of $(\mathsf{Y}_t)_{t=0}^{\infty}$. In any n-leaf rooted tree topology, there are a total of $n-2$ internal non-root nodes that may be chosen as the target. From each of these, we may interchange the sibling and either of the two children to obtain a different tree. Therefore, there are $2(n-2)$ NNIs that result in a different tree. Since $(\mathsf{Y}_t)_{t=0}^{\infty}$ moves to a different tree with probability 1/2, and it does

so by choosing an NNI uniformly at random from the $2(n-2)$ possible NNIs, it follows that $\mathbf{P}_{\mathsf{Y}}(\mathbf{x}, \mathbf{y})$ if $\mathbf{y} \sim \mathbf{x}$. $\square$

The construction of the NNI path will demonstrate that in no more than $(n-1)^2$ NNIs, it is possible to reach any tree from any other given tree. Therefore, $(\mathsf{Y}_t)_{t=0}^{\infty}$ is irreducible and has a unique stationary distribution. The transition matrix is symmetric, so the discrete uniform distribution on T_n is the stationary distribution for $(\mathsf{Y}_t)_{t=0}^{\infty}$. We are now ready to derive an upper bound on the relaxation time of $(\mathsf{Y}_t)_{t=0}^{\infty}$.

Theorem 2. *There exists a finite constant $M_2 < \infty$ such that $\tau_{rel}(\mathsf{Y}) \leq M_2 n^4$.*

Proof. We construct a path from $\mathbf{x} \in \mathsf{T}_n$ to $\mathbf{y} \in \mathsf{T}_n$, where each step of the path is completed by performing an NNI on the tree that results from the previous step. Recall that an NNI is a tree rearrangement that involves choosing an internal, non-root node as the target. Two nodes, along with their descendant subtrees, are chosen uniformly at random from among the sibling and the two children of the target, and the chosen nodes and subtrees are interchanged. The construction of the NNI path uses the SPR path described in the previous section as a skeleton. The idea is to decompose each SPR into a sequence of NNI steps, thus constructing an NNI path. We then make use of Lemma 9 to develop an upper bound on the relaxation time of $(\mathsf{Y}_t)_{t=0}^{\infty}$.

Consider an SPR that changes tree $\mathbf{v} \in \mathsf{T}_n$ to tree $\mathbf{w} \in \mathsf{T}_n$ by pruning and regrafting a leaf l. We decompose this move into a sequence of NNI transitions

in which leaf l is moved "up" one level at a time in the tree and/or moved "down" the tree one level at a time until $\mathbf{w}$ is reached. Figure 2.3 shows how any leaf can be moved up or down one level of the tree by performing a NNI. This idea enables us to construct an NNI path $\psi_{\mathbf{xy}}$ between any two trees $\mathbf{x}$ and $\mathbf{y}$ in T_n by constructing an SPR path $\gamma_{\mathbf{xy}}$ as before, and then breaking each SPR transition into a series of NNI steps.

Formally, let $\mathbf{x}, \mathbf{y} \in \mathsf{T}_n$ and let $\gamma_{\mathbf{xy}} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ be the SPR path

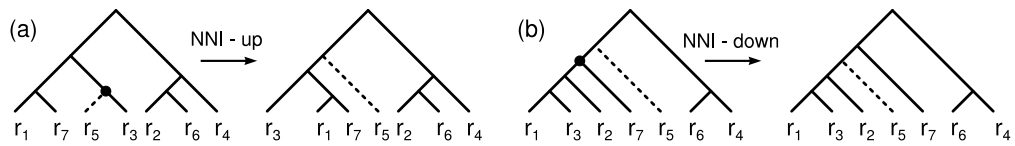


Figure 2.3: (a) An NNI that results in leaf r_5 being moved up one level of the tree. In the starting tree, the target node is identified by a large black dot. The sibling and the child r_5 are interchanged, yielding the tree to the right. (b) An NNI that results in leaf r_5 being moved down one level of the tree. The left child and sibling (r_5) of the target are interchanged to yield the tree on the right.

described in the previous subsection such that $\mathbf{x} = \mathbf{x}_1$ and $\mathbf{y} = \mathbf{x}_n$. The corresponding NNI path can be written as

$$\psi_{\mathbf{xy}} = (\mathbf{x}_1^{(1)}, \mathbf{x}_1^{(2)}, \ldots, \mathbf{x}_1^{(n_1-1)}, \mathbf{x}_2^{(1)}, \ldots, \mathbf{x}_2^{(n_2-1)}, \ldots, \mathbf{x}_{n-1}^{(1)}, \ldots, \mathbf{x}_{n-1}^{(n_{n-1})}, \mathbf{x}_n^{(1)}),$$

where n_k denotes the number of NNI moves required to change tree $\mathbf{x}_k$ into $\mathbf{x}_{k+1}$ and $\mathbf{x}_1 = \mathbf{x}_1^{(1)}$, $\mathbf{x}_2 = \mathbf{x}_2^{(1)}$, ..., $\mathbf{x}_n = \mathbf{x}_n^{(1)} = \mathbf{y}$. An n-leaf rooted tree topology has $n-1$ internal nodes, so to reconstruct one of the SPRs from

the path described in the previous section, a leaf must be moved in the way described above a maximum of $n-1$ times. Since the number of SPRs required to construct the SPR path described in the previous section is no larger than $n-1$, the length of an NNI path is no greater than $(n-1)^2$ steps. In addition, observe that each intermediary tree has the property that it contains a subtree $\mathbf{s}_k$ satisfying i), ii), and iii) as in the construction of the SPR path.

Now let A denote the edge set of the underlying graph corresponding to the transition matrix $\mathbf{P}_{\mathsf{Y}}$. Define the subpath $\psi_{\mathbf{xy}}^{(k)} = (\mathbf{x}_k^{(1)}, \mathbf{x}_k^{(2)}, \ldots, \mathbf{x}_k^{(n_k-1)}, \mathbf{x}_{k+1}^{(1)})$ and note that this is a NNI path from $\mathbf{x}_k$ to $\mathbf{x}_{k+1}$. Using Lemma 9, the relaxation time for the NNI chain is bounded above by

$$\begin{aligned}
\tau_{rel}(\mathsf{Y}) &\leq D_1 n^3 c_n \max_{a\in\mathsf{A}} \sum_{\mathbf{x}\in\mathsf{T}_n} \sum_{\mathbf{y}\in\mathsf{B}} \pi_{\mathbf{x}}\pi_{\mathbf{y}} \Pr(a \in \psi_{\mathbf{xy}}) \\
&= \leq D_1 n^3 c_n \max_{a\in\mathsf{A}} \sum_{\mathbf{x}\in\mathsf{T}_n} \sum_{\mathbf{y}\in\mathsf{B}} \pi_{\mathbf{x}}\pi_{\mathbf{y}} \left(\sum_k \Pr(a \in \psi_{\mathbf{xy}}^{(k)}) \right)
\end{aligned}$$

for some positive constant $D_1 < \infty$. We claim that the rightmost sum above contains no more than $\mathcal{O}(n^{1/2})$ non-zero terms (as in the SPR case). To see this, note that the event "$a \in \psi_{\mathbf{xy}}^{(k)}$" can only occur if edge a is on the NNI path that produced edge $e = (\mathbf{x}_k, \mathbf{x}_{k+1})$ in the SPR path at step k. In other words, the events "$a \in \psi_{\mathbf{xy}}^{(k)}$" and "$e \in \gamma_{\mathbf{xy}}$ at step k" are equivalent. As previously noted, there are at most $\mathcal{O}(n^{1/2})$ possible values for the integer k such that $\Pr(e \in \gamma_{\mathbf{xy}} \text{ at step } k) > 0$. Therefore, it follows that

$$\tau_{rel}(\mathsf{Y}) \leq D_1 n^{7/2} c_n \max_{a\in\mathsf{A}} \max_k \sum_{\mathbf{x}\in\mathsf{T}_n} \sum_{\mathbf{y}\in\mathsf{T}_n} \pi_{\mathbf{x}}\pi_{\mathbf{y}} \Pr(a \in \psi_{\mathbf{xy}}^{(k)}).$$

Also, observe that if $a \in \psi_{\mathbf{xy}}^{(k)}$, then $a = (\mathbf{v}, \mathbf{w})$ for some $(\mathbf{v}, \mathbf{w}) \in \mathsf{T}_n \times \mathsf{T}_n$, and the following three events must occur:

a') The subtree $\mathbf{s}_k$ contains the leaves $r_1, r_2, \ldots, r_k$, and the leaf being moved is r_{k+1};

b') The $\{r_1, r_2, \ldots, r_k\}$-tree derived from $\mathbf{y}$ is the same as the $\{r_1, r_2, \ldots, r_k\}$-tree derived from $\mathbf{w}$;

c') The $\{r_1, r_{k+2}, r_{k+3}, \ldots, r_n\}$-tree derived from $\mathbf{x}$ is the same as the $\{r_1, r_{k+2}, r_{k+3}, \ldots, r_n\}$-tree derived from $\mathbf{v}$.

Event a') has probability $1/\binom{n}{k} \times 1/(n-k)$, since a random permutation of the leaves is used to form the SPR path. Conditionally on a'), events b') and c') are independent and have probabilities $1/c_k$ and $1/c_{n-k}$, respectively, by Lemma 7. Therefore,

$$\begin{aligned}
\tau_{rel}(\mathsf{Y}) &\leq D_1 n^{7/2} c_n \sum_{\mathbf{x}\in\mathsf{T}_n} \sum_{\mathbf{y}\in\mathsf{T}_n} \pi_{\mathbf{x}}\pi_{\mathbf{y}} \Pr(a \in \psi^{(k)}_{\mathbf{xy}}) \\
&\leq \frac{1}{\binom{n}{k}(n-k)c_k c_{n-k}} \\
&\approx D_1 n^{7/2} 2^n n^{n-1} e^{-n} \frac{k^{k+1/2}(n-k)^{n-k+1/2}}{n^{n+1/2}(n-k)^{n-k} 2^n k^{k-1} e^{-n}} \\
&= D_1 n^{7/2} k^{3/2} (n-k)^{1/2} n^{-3/2} \\
&\leq D_1 n^4. \qquad (2.9)
\end{aligned}$$

□

This establishes an upper bound of $\mathcal{O}(n^4)$ on the relaxation time of $(\mathsf{Y}_t)_{t=0}^{\infty}$.

2.3.1 Remarks

In the preceding two subsections, we established upper bounds on the relaxation times of two Markov chains on rooted phylogenetic trees. While the

bounds established here are the best ones known to date, it is unlikely that they are sharp. For example, Aldous (2012) recently conjectured that the relaxation time for a chain similar to one of those discussed above is $\mathcal{O}(n^{3/2})$. Simulation studies presented by Herbei and Kubatko (2013) show similar results. This is an indication that it is important to continue investigating the mixing times of these types of chains in hopes of designing more efficient MCMC algorithms for Bayesian inference of phylogenetic trees given data at the leaves. We now provide a lower bound on the relaxation time of each of the two chains.

2.4 Lower Bounds on the Relaxation Times of the SPR and NNI Chains

In order to give lower bounds on the relaxation times of the SPR and NNI chains, we must first provide information about the distribution of the number of cherries on a randomly selected tree topology $\mathbf{t} \in \mathsf{T}_n$, where a cherry is a pair of leaves that are adjacent to a common ancestral node (McKenzie and Steel, 2000). We first give an upper bound on the net change in the number of cherries under an SPR transition. We then derive the expected value and the variance of the number of cherries on a randomly selected tree under the stationary distribution.

2.4.1 Upper Bound on the Net Change in the Number of Cherries Under an SPR

The following lemma gives an upper bound on the net change in the number of cherries under an SPR transition.

Lemma 11. *Let* $f : \mathsf{T}_n \mapsto \left\{1, 2, \ldots, \left\lfloor \frac{n}{2} \right\rfloor\right\}$ *be a function that gives the number of cherries on a For any pair* $(\mathbf{t}_1, \mathbf{t}_2) \in \mathsf{T}_n \times \mathsf{T}_n$ *such that* $\mathbf{P}_{\mathsf{X}}(\mathbf{t}_1, \mathbf{t}_2) > 0$,

$$|f(\mathbf{t}_1) - f(\mathbf{t}_2)| \leq 1.$$

Proof. Let $\mathbf{p}$ be the subtree that is pruned from $\mathbf{t}_1$, and let e denote the edge to which $\mathbf{p}$ is reattached. Note that since $\mathbf{t}_1$ is a binary tree, it follows that if $\mathbf{p}$ has more than one leaf, $\mathbf{p}$ must contain at least one cherry. Four cases exist in performing an SPR to move from $\mathbf{t}_1$ to $\mathbf{t}_2$.

1. The subtree $\mathbf{p}$ consists of one leaf, the edge e is a leaf, and together, $\mathbf{p}$ and e form a cherry.

2. The subtree $\mathbf{p}$ consists of one leaf, and that leaf is a part of a cherry. The edge e is not a part of a cherry.

3. The subtree $\mathbf{p}$ is not a part of a cherry, but the edge e is a leaf that is a part of a cherry.

4. Neither $\mathbf{p}$ nor e is a part of a cherry.

Case 1

In pruning the one-leaf subtree $\mathbf{p}$, the cherry from which $\mathbf{p}$ descends is destroyed. This results in either the formation of a new cherry at a node ancestral to the parent node of $\mathbf{p}$, or in the destruction of this cherry without the formation of a new one. Regardless of which is the case, attaching $\mathbf{p}$ to e

results in simultaneously destroying the cherry ancestral to e and creating a new cherry. Thus, if a new cherry is formed in the pruning of $\mathbf{p}$, $f(\mathbf{t}_2) = f(\mathbf{t}_1)$. If the pruning of $\mathbf{p}$ does not result in the creation of a new cherry, then the move results in a net loss of one cherry, so that $f(\mathbf{t}_2) = f(\mathbf{t}_1) - 1$.

Case 2

The pruning of $\mathbf{p}$ results in either the simultaneous creation and destruction of a cherry or the destruction of a cherry without the formation of a new one. Suppose that pruning $\mathbf{p}$ results in the simultaneous destruction of a cherry and the formation of a new one. If e is a leaf, then the attachment of $\mathbf{p}$ to e results in the creation of a new cherry, and $f(\mathbf{t}_2) = f(\mathbf{t}_1) + 1$. If e is not a leaf, then attaching $\mathbf{p}$ to e neither creates nor destroys a cherry, so that $f(\mathbf{t}_2) = f(\mathbf{t}_1)$. Now suppose that pruning $\mathbf{p}$ destroys a cherry without creating a new one. Then if e is a leaf, $f(\mathbf{t}_2) = f(\mathbf{t}_1)$. Otherwise, $f(\mathbf{t}_2) = f(\mathbf{t}_1) - 1$.

Case 3

If pruning $\mathbf{p}$ results in the creation of a cherry, then it is not possible for $\mathbf{p}$ to be a one-leaf subtree. Assume for now that pruning $\mathbf{p}$ results in the creation of a cherry. Then attaching $\mathbf{p}$ to e results in the destruction of a cherry, so that $f(\mathbf{t}_2) = f(\mathbf{t}_1)$. Now assume that pruning $\mathbf{p}$ creates no cherries. Then since $\mathbf{p}$ is not a part of a cherry, pruning it does not destroy any cherries. If $\mathbf{p}$ is a one-leaf subtree, then attaching it to e results in a net gain of zero cherries, so that $f(\mathbf{t}_2) = f(\mathbf{t}_1)$. If $\mathbf{p}$ is not a one-leaf subtree, then attaching it to e destroys a cherry, and $f(\mathbf{t}_2) = f(\mathbf{t}_1) - 1$.

Case 4

If $\mathbf{p}$ is a one-leaf subtree and e is a leaf, then pruning $\mathbf{p}$ neither creates nor destroys a cherry. However, attachment of $\mathbf{p}$ to e creates a cherry. Therefore, $f(\mathbf{t}_2) = f(\mathbf{t}_1) + 1$. If $\mathbf{p}$ is not a one-leaf subtree and e is a leaf, then it is possible to create a cherry by pruning $\mathbf{p}$, but it is not possible to destroy one. The reattachment neither creates nor destroys any cherries, so that either $f(\mathbf{t}_2) = f(\mathbf{t}_1)$ or $f(\mathbf{t}_2) = f(\mathbf{t}_1) + 1$. If $\mathbf{p}$ is a one-leaf subtree and e is not a leaf, then pruning $\mathbf{p}$ can neither create nor destroy a cherry. Since e is not a leaf, attaching $\mathbf{p}$ to e cannot create a cherry. Since e is not a descendant of a cherry, the reattachment cannot destroy one. Therefore, $f(\mathbf{t}_2) = f(\mathbf{t}_1)$. If $\mathbf{p}$ is not a one-leaf subtree and e is not a leaf, the pruning of $\mathbf{p}$ can create a cherry, but it cannot destroy one. The attachment of $\mathbf{p}$ to e neither creates nor destroys a cherry. This yields that either $f(\mathbf{t}_2) = f(\mathbf{t}_1)$ or $f(\mathbf{t}_2) = f(\mathbf{t}_1) + 1$. This establishes Lemma 11. □

Since the NNI move is a special case of the SPR, it follows that Lemma 11 holds for the NNI transition as well.

2.4.2 Distribution of the Number of Cherries

Hendy and Penny (1982) provide a result that is useful in deriving the expectation and the variance of the number of cherries for unrooted trees under the stationary distribution π^*, where π^* is the discrete uniform distribution on U_n, the space of n-leaf unrooted tree topologies. Let $N^*_{n,r}$ denote the number

of n-leaf unrooted tree topologies with r cherries. Then for $n \geq 4$,

$$N^*_{n,r} = \frac{n!\,(n-4)!}{(n-2r)!\,r!\,(r-2)!\,2^{2(r-1)}},\ 2 \leq r \leq \left\lfloor \frac{n}{2} \right\rfloor.$$

We modify this result to derive the stationary variance of the number of cherries for rooted trees. Let $N_{n,r}$ represent the number of rooted tree topologies with n taxa and r cherries. Let C_n be a random variable that corresponds to the number of cherries on a randomly selected topology in T_n. The support of C_n is $\left\{1, 2, \ldots, \left\lfloor \frac{n}{2} \right\rfloor\right\}$, where $\lfloor x \rfloor$ denotes the greatest integer that is no larger than x. Let $\mathbb{E}_\pi[\cdot]$ denote expected value with respect to the measure π, where π is the discrete uniform measure on T_n, and let $\mathbb{E}_{\pi^*}[\cdot]$ be the expected value with respect to π^*. Let $\mathbb{V}_\pi[\cdot]$ and $\mathbb{V}_{\pi^*}[\cdot]$ represent variances with respect to π and π^*, respectively. Let C^*_n be a random variable that corresponds to the number of cherries on a randomly selected unrooted tree topology. The support of C^*_n is the same as that of C_n. We are now ready to present the following result.

Lemma 12. *For $n \geq 6$,*

$$N_{n,r} = \begin{cases} \frac{n!}{2}, & \text{if } r = 1 \\ \frac{n!(n-4)![2(n-r)-3]}{(n-2r)!r!(r-2)!2^{2(r-1)}} + & \\ \frac{n!(n-4)!2(r+1)}{(n-2(r+1))!(r+1)!(r-1)!2^{2r}}, & \text{if } 2 \leq r \leq \left\lfloor \frac{n}{2} \right\rfloor - 1 \\ \frac{n!(n-4)![2(n-r)-3]}{(n-2r)!r!(r-2)!2^{2(r-1)}}, & \text{if } r = \left\lfloor \frac{n}{2} \right\rfloor. \end{cases} \tag{2.10}$$

For $n = 4$ or 5,

$$N_{n,r} = \begin{cases} \frac{n!}{2} & \text{if } r = 1 \\ \frac{n!(2n-7)}{8} & \text{if } r = 2. \end{cases} \tag{2.11}$$

For $n = 2$ or 3, $N_{n,1} = (2n-3)!!$. Under the stationary distribution, for $r = 1, 2, \ldots, \lfloor \frac{n}{2} \rfloor$,

$$Pr(C_n = r) = \frac{N_{n,r}}{(2n-3)!!}. \tag{2.12}$$

Proof. For an unrooted tree topology, there are $2n-3$ edges, and hence, $2n-3$ places to insert a root. When we insert a root, we either insert it along one of the descendant branches of a cherry, or we insert the root along a branch that is not a part of a cherry. There are $2r$ ways to put a root on a branch of a cherry, each of which results in a different rooted tree topology. This leaves $2(n-r)-3$ ways to place a root elsewhere. Each placement of a root results in a distinct tree topology.

Placing a root on a branch of a cherry destroys the cherry, while placing a root elsewhere neither creates nor destroys a cherry. Therefore, rooting the tree either decreases the number of cherries by one, or it leaves the number of cherries unchanged. To obtain a rooted n-leaf tree topology with one cherry, we must root an unrooted tree topology with two cherries along a branch of a cherry. Therefore,

$$N_{n,1} = N^*_{n,2} 2(2) = 4N^*_{n,2} = \frac{n!}{2}.$$

There are two ways to obtain a rooted topology with n taxa and r cherries from an unrooted topology with n taxa, $2 \leq r \leq \lfloor n/2 \rfloor - 1$. One way is to insert a root along a branch that is not part of a cherry on an unrooted topology with r cherries. The other is to insert a root along a branch that is part of a cherry on an unrooted tree topology with $r+1$ cherries. There are $2(n-r)-3$ ways to perform the first, with each resulting in a different topology. There

are $2(r+1)$ ways to perform the second, and each of these results in a different topology. Therefore,

$$N_{n,r} = \frac{n!\,(n-4)!\,[2(n-r)-3]}{(n-2r)!\,r!\,(r-2)!\,2^{2r-2}} + \frac{n!\,(n-4)!\,2(r+1)}{[n-2(r+1)]\,!\,(r+1)!\,(r-1)!\,2^{2r}},$$

where $2 \leq r \leq \lfloor n/2 \rfloor - 1$. When $r = \lfloor n/2 \rfloor$, there is only one way to get a rooted topology with r cherries from an unrooted topology with r cherries. We must insert a root along a branch that is not part of a cherry. There are $2(n-r)-3$ ways to do this, each resulting in a different topology. As a result,

$$\begin{aligned} N_{n,r} &= [2(n-r)-3]\,N^*_{n,r} \\ &= \frac{n!\,(n-4)!\,[2(n-r)-3]}{(n-2r)!\,r!\,(r-2)!\,2^{2(r-1)}}, \end{aligned}$$

where $r = \lfloor n/2 \rfloor$.

When $n = 4$ or $n = 5$, r must be either 1 or 2. Thus, we apply the same arguments we used for the case where $n \geq 6$ and $r = 1$ and $r = \lfloor n/2 \rfloor$ to obtain

$$N_{n,r} = \begin{cases} \frac{n!}{2}, & r = 1 \\ \frac{n!(2n-7)}{8}, & r = 2. \end{cases}$$

When $n = 2$ or $n = 3$, r must be 1. Thus, we apply the argument for the case where $n \geq 6$ and $r = 1$ to obtain

$$N_{n,1} = \frac{n!}{2}.$$

Since the stationary distribution for the chain we describe is the uniform distribution over T_n, for $n \geq 6$, (2.12) follows immediately. Similarly, for $n = 4$ or $n = 5$, we obtain (2.11). In addition, it is clear that a 2-taxon or 3-taxon tree must have exactly one cherry, so that $\Pr(C_n = 1) = 1$ for $n = 2$ or $n = 3$. $\square$

In order to obtain the stationary variance of the number of cherries, we must first find the expected number of cherries on a randomly selected rooted tree topology with respect to the stationary distribution.

Lemma 13. *For* $n \geq 2$, $\mathbb{E}_\pi [C_n] = \frac{n(n-1)}{2(2n-3)}$.

Proof. McKenzie and Steel (2000) showed that

$$\mathbb{E}_{\pi^*} [C_n^*] = \frac{n(n-1)}{2(2n-5)}.$$

We see that the expected number of cherries on a rooted topology is given by

$$\begin{aligned}
&\mathbb{E}_\pi [C_n] \\
&= \frac{1}{(2n-3)!!} \left[\frac{n!}{2} + \sum_{r=2}^{\lfloor \frac{n}{2} \rfloor} r \frac{n!\,(n-4)!\,[2(n-r)-3]}{(n-2r)!\,r!\,(r-2)!\,2^{2r-2}} \right] \\
&+ \frac{1}{(2n-3)!!} \sum_{r=2}^{\lfloor \frac{n}{2} \rfloor - 1} r \frac{n!\,(n-4)!\,2(r+1)}{(n-2(r+1))!\,(r+1)!\,(r-1)!\,2^{2r}} \\
&= \frac{n!}{(2n-3)!!} + \frac{1}{2n-3} \left[(2n-3)\mathbb{E}_{\pi^*} [C_n^*] - 2\mathbb{E}_{\pi^*} \left[(C_n^*)^2 \right] \right] \\
&+ \frac{2}{(2n-3)!!} \sum_{r=3}^{\lfloor \frac{n}{2} \rfloor} (r-1) \frac{n!\,(n-4)!\,r}{(n-2r)!\,r!\,(r-2)!\,2^{2r-2}}. \qquad (2.13)
\end{aligned}$$

The term involving the summation in (2.13) can also be written in terms of expectations with respect to π^* as

$$\frac{2}{2n-3} \left[\mathbb{E}_{\pi^*} \left[(C_n^*)^2 \right] - \mathbb{E}_{\pi^*} [C_n^*] \right] - \frac{n!}{2(2n-3)!!}. \qquad (2.14)$$

Adding to (2.14) the part of (2.13) that is written in terms of expectations with respect to π^*, we get

$$\begin{aligned}\mathbb{E}_\pi\left[C_n\right] &= \frac{2n-5}{2n-3}\mathbb{E}_{\pi^*}\left[C_n^*\right] \\ &= \frac{(2n-5)n(n-1)}{2(2n-3)(2n-5)} \\ &= \frac{n(n-1)}{2(2n-3)}.\end{aligned}$$

If $n = 4$,

$$\mathbb{E}_\pi\left[C_n\right] = \frac{4!}{2(2n-3)!\,!} + 2\frac{n!\,(2n-7)}{2(2n-3)!\,!} = \frac{4(3)}{2(5)}.$$

If $n = 5$,

$$\mathbb{E}_\pi\left[C_n\right] = \frac{5!}{2(2n-3)!\,!} + 2\frac{n!\,(2n-7)}{2!\,2^2(2n-3)!\,!} = \frac{5(4)}{2(7)},$$

so that in both cases, $\mathbb{E}_\pi\left[C_n\right] = \frac{n(n-1)}{2(2n-3)}$.

If $n = 2$ or $n = 3$, $\mathbb{E}_\pi\left[C_n\right] = 1$, which is equal to $\frac{n(n-1)}{2(2n-3)}$ in either case. Therefore, for $n \geq 2$,

$$\mathbb{E}_\pi\left[C_n\right] = \frac{n(n-1)}{2(2n-3)}.$$

□

Now that we have the distribution and the expectation of C_n, we can find the variance. In order to do so, we need to find $\mathbb{E}_\pi\left[(C_n)^2\right]$. Using a similar approach to that which was employed to find $\mathbb{E}_\pi\left[C_n\right]$, we find that for $n \geq 6$,

$$\begin{aligned}\mathbb{E}_\pi\left[(C_n)^2\right] &= \frac{2n-7}{2n-3}\left[\mathbb{V}_{\pi^*}\left[C_n^*\right] + \left(\mathbb{E}_{\pi^*}\left[C_n^*\right]\right)^2\right] \\ &+ \frac{2}{2n-3}\mathbb{E}_{\pi^*}\left[C_n^*\right].\end{aligned}$$

McKenzie and Steel (2000) showed that

$$\mathbb{V}_{\pi^*}\left[C_n^*\right] = \frac{n(n-1)(n-4)(n-5)}{2(2n-5)^2(2n-7)}. \tag{2.15}$$

Therefore, for $n \geq 6$,

$$\begin{aligned}\mathbb{E}_\pi\left[(C_n)^2\right] &= \frac{2n-7}{2n-3}\left[\frac{n(n-1)(n-4)(n-5)}{2(2n-5)^2(2n-7)} + \frac{n^2(n-1)^2}{4(2n-5)^2}\right] \\ &+ \frac{n(n-1)}{(2n-3)(2n-5)},\end{aligned}$$

which implies

$$\begin{aligned}\mathbb{V}_\pi\left[C_n\right] &= \frac{2n-7}{2n-3}\left[\frac{n(n-1)(n-4)(n-5)}{2(2n-5)^2(2n-7)} + \frac{n^2(n-1)^2}{4(2n-5)^2}\right] \\ &+ \frac{n(n-1)}{(2n-3)(2n-5)} - \frac{n^2(n-1)^2}{4(2n-3)^2}. \qquad (2.16)\end{aligned}$$

If $n = 4$ or $n = 5$,

$$\begin{aligned}\mathbb{E}_\pi\left[(C_n)^2\right] &= \sum_{r=1}^{2} \Pr(C_n = r) r^2 \\ &= \frac{n!}{2(2n-3)!!} + \frac{4n!\,(2n-7)}{8(2n-3)!!} \\ &= \frac{(2n-6)n!}{2(2n-3)!!}.\end{aligned}$$

Therefore,

$$\mathbb{V}_\pi[C_n] = \frac{(2n-6)n!}{2(2n-3)!!} - \frac{n^2(n-1)^2}{4(2n-3)^2}.$$

If $n = 2$ or $n = 3$, $\mathbb{V}_\pi[C_n] = 0$ since $C_n = 1$ with probability 1. It is straightforward to show that the value on the right hand side of (2.16) is approximately $n/16$. We are now ready to give a lower bound on the relaxation time of both chains.

2.4.3 Lower Bound on the Relaxation Time

In order to derive our lower bounds on the relaxation time of the SPR and NNI chains, we need the following theorem.

Theorem 3. *(Diaconis and Stroock, 1991) Suppose* $\mathbf{t}_1$ *and* $\mathbf{t}_2$ *are two tree topologies in* T_n*. Then for a Markov chain* $(\mathsf{Z}_t)_{t=0}^{\infty}$ *having state space* T_n*,*

$$\tau_{rel}(\mathsf{Z}) := \sup_{f:\mathsf{T}_n \mapsto \mathbb{R}} \frac{2\,Var_\pi f}{\sum_{\mathbf{t}_1}\sum_{\mathbf{t}_2} \pi_{\mathbf{t}_1}\mathbf{P}_{\mathsf{Z}}(\mathbf{t}_1,\mathbf{t}_2)(f(\mathbf{t}_1)-f(\mathbf{t}_2))^2}, \tag{2.17}$$

where $Var_\pi f$ *is the variance of the function* f *with respect to the stationary distribution* π*, and* $\mathbf{P}_{\mathsf{Z}}$ *is the transition matrix for* $(\mathsf{Z}_t)_{t=0}^{\infty}$*.*

In order to use Theorem 3 to establish a lower bound on the relaxation times of the two chains we described in Section 2.2, it is necessary to select a function $f : \mathsf{T}_n \mapsto \mathbb{R}$ for which we can bound the variance with respect to the stationary distribution. The work in the previous section shows that the number of cherries meets this requirement, so we let $f(\mathbf{t})$ be the number of cherries on a rooted tree topology in T_n.

Recall the result of Lemma 11. This, combined with the following, implies that the denominator of (2.17) is bounded above by 1:

$$\begin{aligned}
&\sum_{\mathbf{t}_1\in\mathsf{T}_n}\sum_{\mathbf{t}_2\in\mathsf{T}_n} \pi_{\mathbf{t}_1}\mathbf{P}_{\mathsf{X}}(\mathbf{t}_1,\mathbf{t}_2)\\
&= \sum_{\mathbf{t}_1\in\mathsf{T}_n}\sum_{\mathbf{t}_2\in\mathsf{T}_n} \pi_{\mathbf{t}_1}\Pr(\mathbf{t}_2|\mathbf{t}_1)\\
&= 1.
\end{aligned}$$

This establishes that $\tau_{rel}(\mathsf{X}) \geq \mathcal{O}(n)$ for $n \geq 6$. For $n = 4$ or $n = 5$, we can substitute the exact values of the variance of C_n in each case and use that to obtain a lower bound on the relaxation time. When $n < 4$, this method of obtaining a lower bound on the relaxation time is not very helpful, since $\mathbb{V}_\pi[C_n] = 0$.

For the NNI chain $(\mathsf{Y}_t)_{t=0}^{\infty}$, the moves that we use to construct the NNI

path are special cases of those used to construct the SPR path. In addition, the stationary distributions for the two chains are the same, thus implying that $\tau_{rel}(\mathsf{Y}) \geq \mathcal{O}(n)$.

2.4.4 Bounds on the Mixing Times of the SPR and NNI Chains

The following result relates the upper and lower bounds on the relaxation times of $(\mathsf{X}_t)_{t=0}^{\infty}$ and $(\mathsf{Y}_t)_{t=0}^{\infty}$ to the mixing times of each chain. Let $\tau_{mix}(\mathsf{X})$ and $\tau_{mix}(\mathsf{Y})$ denote the mixing times of $(\mathsf{X}_t)_{t=0}^{\infty}$ and $(\mathsf{Y}_t)_{t=0}^{\infty}$, respectively. Let ϵ be the total variation threshold for mixing.

Theorem 4. *(Levin et al., 2009) Let $\mathbf{P}_{\mathsf{Z}}$ be the transition matrix of a reversible, irreducible Markov chain $(\mathsf{Z}_t)_{t=0}^{\infty}$ with state space T_n, and let $\pi_{min} := \min_{x \in \mathsf{T}_n} \pi_x$, where π is the stationary measure for $(\mathsf{Z}_t)_{t=0}^{\infty}$. Then*

$$(\tau_{rel}(\mathsf{Z}) - 1) \log\left(\frac{1}{2\epsilon}\right) \leq \tau_{mix}(\mathsf{Z}) \leq \log\left(\frac{1}{\epsilon \pi_{min}}\right) \tau_{rel}.$$

Recall that $\mathcal{O}(n) \leq \tau_{rel}(\mathsf{X}) \leq \mathcal{O}(n^{\frac{5}{2}})$. As a result, we have the following lower bound on the mixing time. For some finite positive constant M,

$$\begin{aligned} \tau_{mix}(\mathsf{X}) &\geq (\tau_{rel}(\mathsf{X}) - 1) \log\left(\frac{1}{2\epsilon}\right) \\ &\geq (Mn - 1) \log\left(\frac{1}{2\epsilon}\right) \\ &\geq \mathcal{O}(n). \end{aligned}$$

In addition, for some positive finite constant N, we obtain the following upper bound on $\tau_{mix}(\mathsf{X})$.

$$\begin{aligned}
\tau_{mix}(\mathsf{X}) &\leq \log\left(\frac{1}{\epsilon\pi_{min}}\right) \\
&\leq \log\left(\frac{c_n}{\epsilon}\right) Nn^{5/2} \\
&= \left[\left(\sum_{i=1}^{n-1}\log(2i-1)\right) - \log(\epsilon)\right] Nn^{5/2} \\
&\leq \mathcal{O}(n^{7/2}\log(n)).
\end{aligned}$$

Thus, we have established that $\mathcal{O}(n) \leq \tau_{mix}(\mathsf{X}) \leq \mathcal{O}(n^{7/2}\log(n))$. Similar reasoning can be used to find that $\mathcal{O}(n) \leq \tau_{mix}(\mathsf{Y}) \leq \mathcal{O}(n^5\log(n))$. It is important to recall that the upper bound on the relaxation time, and thus on the mixing time, of $(\mathsf{X}_t)_{t=0}^{\infty}$ is likely not sharp. However, Randall and Tetali (2000) showed that the mixing time of a chain on rooted tree topologies whose transitions consist of tree rearrangements that are similar to NNIs is exactly $\mathcal{O}(n^5\log(n))$. This matches the upper bound on the mixing time that we obtained by finding an upper bound on the relaxation time.

Chapter 3: Geometric Ergodicity of a Markov Chain Monte Carlo Method for Inference of Phylogenetic Branch Lengths

The goal of this chapter is to describe an MCMC algorithm for inferring the branch lengths of a phylogenetic tree when the tree topology is known and there are DNA sequence data at the leaves. During the first several iterations of the algorithm, the chain explores the state space until it is approximately stationary. Subsequent states can be regarded as an approximate sample from the target density. In a Bayesian setting, the target density is a posterior density, and the ability to sample from the approximate posterior density allows Bayesian inference.

The chapter opens with a description of a general method of verifying geometric ergodicity of a Markov chain. This is followed by an introduction to a specific MCMC algorithm, known as the random scan Metropolis (RSM) algorithm, that we employ in the rest of this dissertation. In the literature, one can find several methods of establishing geometric ergodicity for the RSM sampler. We mention some of these methods briefly and then provide a description of the method that is most similar to the sampler we use to approximate the posterior

distribution of the branch lengths given the tree topology and a DNA sequence data set. The chapter concludes with a presentation of a version of the RSM sampler for approximating the posterior density of the branch lengths, as well as a verification of its geometric ergodicity.

3.1 Preliminaries

The work in this chapter verifies that the Markov chain we use to approximate the posterior distribution of the branch lengths converges to its stationary distribution at a geometric rate. Later, we take a step toward providing an honest (Jones and Hobert, 2001) upper bound on the mixing time of our sampler, meaning that the upper bound is obtained prior to running the chain and thus, not determined by the output of the sampler. In order to explain how such bounds are obtained, we must first provide some background on continuous-state Markov chains. These concepts are required in order to develop an understanding of the work we provide in the rest of this dissertation.

Definition 7. A *transition kernel* is a function $K : \mathbb{R}^m \times \mathcal{B}(\mathbb{R}^m) \mapsto [0, 1]$, where $m \geq 1$ is an integer and $\mathcal{B}(\mathbb{R}^m)$ is the σ-field of Borel subsets of $\mathbb{R}^m$, such that

1. For all $\mathbf{x} \in \mathbb{R}^m$, $K(\mathbf{x}, \cdot)$ is a probability measure, and

2. For all $A \in \mathcal{B}(\mathbb{R}^m)$, $K(\cdot, A)$ is measurable.

Definition 8. A sequence $(\mathsf{X}_t)_{t=0}^{\infty}$ of random variables is a *Markov chain*, defined by a transition kernel $K(\cdot, \cdot)$, if for any t, the conditional distribution of X_t

given X_{t-1}, ..., X_0 is the same as the distribution of X_t given X_{t-1}; i.e.

$$\begin{aligned}\Pr(\mathsf{X}_t \in A|\mathsf{X}_0, \ldots, \mathsf{X}_{t-1}) &= \Pr(\mathsf{X}_t \in A|\mathsf{X}_{t-1}) \\ &= \int_A K(\mathsf{X}_{t-1}, \mathrm{d}\mathsf{X}_t),\end{aligned}$$

where $A \in \mathcal{B}(\mathbb{R}^m)$.

Definition 9. If π is a σ-finite probability measure, then π is a *stationary measure* for the transition kernel $K(\cdot,\cdot)$, and for the associated chain $(\mathsf{X}_t)_{t=0}^{\infty}$ if

$$\pi(A) = \int_{\mathbb{R}^m} K(\mathbf{x}, A)\pi(d\mathbf{x}), \text{ for all } A \in \mathcal{B}(\mathbb{R}^m).$$

The definitions of the transition kernel and a stationary distribution enable us to define reversibility, which is a desirable property of an MCMC algorithm.

Definition 10. A stationary Markov chain $(\mathsf{X}_t)_{t=0}^{\infty}$ is *reversible* if the conditional distribution of X_{t+1} given $\mathsf{X}_{t+2} = \mathbf{x}$ is the same as the distribution of X_{t+1} given $\mathsf{X}_t = \mathbf{x}$ for all $t \geq 0$. A sufficient condition for reversibility is satisfaction of the *detailed balance condition*:

$$k(\mathbf{x}|\mathbf{y})\pi(\mathbf{y}) = k(\mathbf{y}|\mathbf{x})\pi(\mathbf{x}),$$

for every $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^m \times \mathbb{R}^m$ where $k(\cdot|\mathbf{x})$ is the density of $K(\mathbf{x}, \cdot)$ with respect to Lebesgue measure.

Definition 11. Let $(\Omega, \mathcal{F}, \mu)$ and $(\Omega, \mathcal{F}, \nu)$ be two probability spaces. The *total variation distance* between μ and ν is given by

$$\|\mu(\cdot) - \nu(\cdot)\|_{TV} := \sup_{A \in \mathcal{F}} |\mu(A) - \nu(A)|$$

Let $K^k(\cdot,\cdot)$ denote the k-step transition kernel for $(\mathsf{X}_t)_{t=0}^{\infty}$, so that

$$K^k(\mathbf{x}, A) = \Pr(\mathsf{X}_{t+k} \in A | \mathsf{X}_t = \mathbf{x}).$$

where k and t are non-negative integers.

Definition 12. The *mixing time* τ_{mix} of $(\mathsf{X}_t)_{t=0}^{\infty}$ is given by

$$\tau_{mix} := \min\left\{k : \max_{\mathbf{x}\in\mathbb{R}^m} \|K^k(\mathbf{x},\cdot) - \pi(\cdot)\|_{TV} \leq \epsilon\right\}$$

for $\epsilon < 1$ chosen by the researcher.

A commonly used method of obtaining an upper bound on the mixing time of a continuous-state Markov chain $(\mathsf{X}_t)_{t=0}^{\infty}$ is the coupling method, which involves the construction of a Markov chain $(\mathsf{Y}_t)_{t=0}^{\infty}$ that is a copy of $(\mathsf{X}_t)_{t=0}^{\infty}$ such that $\mathsf{X}_0 \neq \mathsf{Y}_0$. One can then bound the mixing time by bounding the time required for $(\mathsf{X}_t)_{t=0}^{\infty}$ and $(\mathsf{Y}_t)_{t=0}^{\infty}$ to become the same chain (i.e $(\mathsf{X}_t)_{t=0}^{\infty}$ and $(\mathsf{Y}_t)_{t=0}^{\infty}$ couple). The method assumes that $(\mathsf{X}_t)_{t=0}^{\infty}$ and $(\mathsf{Y}_t)_{t=0}^{\infty}$ are two Markov chains on $\mathbb{R}^m$ such that there exists a time τ_C, called the *coupling time*, which is defined in the following way:

Definition 13. Suppose that for $(\mathsf{X}_t)_{t=0}^{\infty}$ and $(\mathsf{Y}_t)_{t=0}^{\infty}$, there exists a set S such that for each $s \in S$, for all $r \geq s$, $\mathsf{X}_r = \mathsf{Y}_r$. Then the *coupling time* is defined as

$$\tau_C := \inf\{s : s \in S\}.$$

A bound on the mixing time is found through the *coupling inequality*:

$$\|K^k(\mathbf{y},\cdot) - \pi(\cdot)\|_{TV} \leq \Pr(\tau_C > k).$$

One can find many examples of the use of the coupling method in the literature. Propp and Wilson (1996) provide a description of a technique referred to as coupling from the past, which allows perfect sampling from continuous distributions. Pinto and Neal (2001) present a procedure in which a chain is coupled with an approximation to an exact copy of the original chain. For a brief survey of the uses and variations of the coupling method, see the work of Breyer and Roberts (2000). Coupling is not the only useful approach for bounding the mixing time of a Markov chain. For a more comprehensive treatment of this topic, one may consult Meyn and Tweedie (2009).

In many cases, we desire an MCMC algorithm that converges to its stationary distribution at at least a geometric rate. We say that $(\mathsf{X}_t)_{t=0}^{\infty}$ is *geometrically ergodic* if there exist a positive finite constant $R(\mathbf{x})$, which may depend on the inital state of the chain, and a constant $r < 1$ such that for all $n \geq 1$,

$$\|K^n(\mathbf{x}, \cdot) - \pi(\cdot)\|_{TV} \leq R(\mathbf{x})r^n.$$

Geometric ergodicity does not ensure rapid convergence of an MCMC algorithm. However, geometric ergodicity guarantees that, under certain conditions, we have a Central Limit Theorem for ergodic averages. In order to discuss this, we need to define several properties that are desirable for a well-behaved Markov chain.

Definition 14. Given a probability measure ϕ, the Markov chain $(\mathsf{X}_t)_{t=0}^{\infty}$ with transition kernel $K(\cdot, \cdot)$ is *ϕ-irreducible* if for every $A \in \mathcal{B}(\mathbb{R}^m)$ with $\phi(A) > 0$, there exists n such that $K^n(\mathbf{x}, A) > 0$ for all $\mathbf{x} \in \mathbb{R}^m$.

Irreducibility ensures that every set $A \in \mathcal{B}(\mathbb{R}^m)$ will be visited by the chain, but it does not guarantee that the chain will enter A frequently enough to ensure convergence at a geometric rate. To guarantee this, we need the property of recurrence.

Definition 15. A Markov chain $(\mathsf{X}_t)_{t=0}^{\infty}$ is *recurrent* if

1. there exists a probability measure ϕ such that $(\mathsf{X}_t)_{t=0}^{\infty}$ is ϕ-irreducible, and

2. for every $A \in \mathcal{B}(\mathbb{R}^m)$ such that $\phi(A) > 0$, $\mathbb{E}_{\mathbf{x}}[\eta_A] = \infty$ for every $\mathbf{x} \in A$,

where η_A denotes the number of times the chain returns to the set A, so that

$$\eta_A = \sum_{t=1}^{\infty} \mathbb{I}_A(\mathsf{X}_t)$$

and

$$\mathbb{I}_A(\mathsf{X}_t) = \begin{cases} 1 & \text{if } \mathsf{X}_t \in A \\ 0 & \text{otherwise} \end{cases}.$$

In order to take advantage of Central Limit theorems for ergodic averages, we need a property that is stronger than recurrence.

Definition 16. A set $A \in \mathcal{B}(\mathbb{R}^m)$ is *Harris recurrent* if $\Pr_{\mathbf{x}}(\eta_A = \infty) = 1$ for all $\mathbf{x} \in A$. The chain $(\mathsf{X}_t)_{t=0}^{\infty}$ is Harris recurrent if there exists a measure ϕ such that $(\mathsf{X}_t)_{t=0}^{\infty}$ is ϕ- irreducible, and for every set A such that $\phi(A) > 0$, A is Harris recurrent.

Definition 17. A set $C \subset \mathbb{R}^m$ is a *small set* if there exist an integer $n > 0$ and a probability measure ν such that for all $\mathbf{x} \in C$ and for all $A \in \mathcal{B}(\mathbb{R}^m)$,

$$K^n(\mathbf{x}, A) \geq \nu(A).$$

Definition 18. A ϕ-irreducible Markov chain $(\mathsf{X}_t)_{t=0}^{\infty}$ has a *cycle of length* d if there exists a small set C, an associated integer M, and a probability distribution ν_M such that d is the gcd of

$$\mathcal{D}(C) := \{m \geq 1 : \text{There exists } \delta_m > 0 \text{ such that } C \text{ is small for } \nu_m \geq \delta_m \nu_M\}.$$

The number d is termed the *period* of the chain, and $(\mathsf{X}_t)_{t=0}^{\infty}$ is *aperiodic* if $d = 1$. This definition says that for each $m \in \mathcal{D}(C)$, the set C satisfies the requirements of a small set for some probability measure ν_m, so that if the chain starts in C, then it has a positive probability of returning to C in m steps.

3.1.1 The Metropolis Hastings Algorithm

The literature contains descriptions of a wide variety of MCMC algorithms. We describe two of the most common MCMC algorithms here. The Metropolis Hastings algorithm (Metropolis et al., 1953; Hastings, 1970) is a general way of obtaining an approximate sample from π. Suppose X_t is the current state of a Markov chain $(\mathsf{X}_t)_{t=0}^{\infty}$. We obtain X_{t+1} as follows:

1. Draw X^* from a proposal distribution $q(\cdot|\mathsf{X}_t)$.

2. Calculate the acceptance ratio

$$\alpha(\mathsf{X}_t, \mathsf{X}^*) = \frac{\pi(\mathsf{X}^*)q(\mathsf{X}_t|\mathsf{X}^*)}{\pi(\mathsf{X}_t)q(\mathsf{X}^*|\mathsf{X}_t)}.$$

3. Set $\mathsf{X}_{t+1} = \mathsf{X}^*$ with probability $\min\{\alpha(\mathsf{X}_t, \mathsf{X}^*), 1\}$, and X_t with probability 1-$\min\{\alpha(\mathsf{X}_t, \mathsf{X}^*), 1\}$.

Provided that the chain generated by the Metropolis Hastings algorithm is irreducible and aperiodic, it has a unique stationary distribution π.

3.1.2 Gibbs Sampler

The Gibbs sampler (Gelfand and Smith, 1990) is a special case of the Metropolis Hastings algorithm that produces a Markov chain with target density π when samples from the set of full conditional densities are available. The sampling algorithm proceeds through a one-at-a-time updating scheme in the following way:

1. Choose a starting value $\mathsf{X}_0 \in \mathbb{R}^m$, $m \geq 1$.

2. Draw X_{t+1} with dimension m from the full conditional densities

$$\begin{aligned}
\mathsf{X}_{t+1}^{(1)} &\sim \pi(\cdot|\mathsf{X}_t^{(2)}, \ldots, \mathsf{X}_t^{(m)}) \\
\mathsf{X}_{t+1}^{(2)} &\sim \pi(\cdot|\mathsf{X}_{t+1}^{(1)}, \mathsf{X}_t^{(3)}, \ldots, \mathsf{X}_t^{(m)}) \\
&\vdots \\
\mathsf{X}_{t+1}^{(m-1)} &\sim \pi(\cdot|\mathsf{X}_{t+1}^{(1)}, \ldots, \mathsf{X}_{t+1}^{(m-2)}, \mathsf{X}_t^{(m)}) \\
\mathsf{X}_{t+1}^{(m)} &\sim \pi(\cdot|\mathsf{X}_{t+1}^{(1)}, \ldots, \mathsf{X}_{t+1}^{(m-1)}).
\end{aligned}$$

In this chapter and the next, we provide convergence results for a particular hybrid sampler. This algorithm was extensively studied in Fort et al. (2003), and we use it to approximate the posterior density of a set of branch lengths given a data set and a tree topology.

3.2 A General Method of Establishing Geometric Ergodicity

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and let $(\mathsf{X}_t)_{t=0}^{\infty}$ be a Markov chain on $\mathbb{R}^m, m \geq 1$, that has transition kernel $K(\cdot,\cdot)$. A sufficient condition for geometric ergodicity of $(\mathsf{X}_t)_{t=0}^{\infty}$ is that it satisfies a *minorization condition* and an associated *drift condition*. We define these two conditions here.

Definition 19. A Markov chain $(\mathsf{X}_t)_{t=0}^{\infty}$ satisfies a *minorization condition* if there exist $\epsilon > 0$, a small set $C \subset \mathbb{R}^m$, and a probability measure ν such that for all $\mathbf{x} \in C$,

$$K(\mathbf{x}, A) \geq \epsilon\nu(A) \ \forall A \in \mathcal{B}(\mathbb{R}^m).$$

Definition 20. A Markov chain $(\mathsf{X}_t)_{t=0}^{\infty}$ satisfies a *drift condition* if there exist a function $V : \mathbb{R}^m \mapsto [1, \infty)$, constants $\lambda \in (0, 1)$ and $b < \infty$, and a small set $C \subset \mathbb{R}^m$ such that for all $\mathbf{x} \in \mathbb{R}^m$,

$$\mathbb{E}\left[V(\mathsf{X}_{t+1})|\mathsf{X}_t = \mathbf{x}\right] \leq \lambda V(\mathbf{x}) + b\mathbb{I}_C(\mathbf{x}).$$

Equivalently, $(\mathsf{X}_t)_{t=0}^{\infty}$ satisfies a drift condition if there exist a function $V : \mathbb{R}^m \mapsto [0, \infty)$ and constants $\lambda \in (0, 1)$ and $b < \infty$ such that for all $\mathbf{x} \in \mathbb{R}^m$,

$$\mathbb{E}\left[V(\mathsf{X}_{t+1})|\mathsf{X}_t = \mathbf{x}\right] \leq \lambda V(\mathbf{x}) + b.$$

Two natural questions arise here. The first is the question of what is meant by an "associated" drift condition. The second question is how satisfaction of

a minorization condition and an associated drift condition implies geometric ergodicity. We resolve the second question, and in the process, we provide an answer to the first.

3.2.1 Role of the Minorization Condition

Suppose $K(\cdot,\cdot)$ admits a transition density $k(\cdot|\cdot)$ with respect to the m-dimensional Lebesgue measure λ_m, and suppose that $(\mathsf{X}_t)_{t=0}^{\infty}$ satisfies a minorization condition on a small set $C \subset \mathbb{R}^m$, with minorizing constant $\epsilon < 1$ and minorizing density $q(\cdot)$ so that for any $\mathbf{x} \in C$,

$$k(\mathbf{y}|\mathbf{x}) \geq \epsilon q(\mathbf{y}) \text{ for all } \mathbf{y} \in \mathbb{R}^m.$$

Let $r(\cdot|\cdot)$ be a probability density with respect to λ_m, termed the *residual density*, on $\mathbb{R}^m$. The residual density is given by

$$r(\mathbf{y}|\mathbf{x}) = \frac{k(\mathbf{y}|\mathbf{x}) - \epsilon q(\mathbf{y})}{1-\epsilon}$$

for each fixed $\mathbf{x} \in C$. This representation of the residual density allows the transition density to be written as a mixture of the residual and minorizing densities so that for any $\mathbf{x} \in C$,

$$k(\mathbf{y}|\mathbf{x}) = \epsilon q(\mathbf{y}) + (1-\epsilon)\, r(\mathbf{y}|\mathbf{x}).$$

This representation of the transition density corresponds to the following description of $(\mathsf{X}_t)_{t=0}^{\infty}$. If $\mathsf{X}_t \in C$, generate X_{t+1} as follows. Generate a $0-1$ coin flip from a Bernoulli(ϵ) distribution. If the outcome is 1, generate X_{t+1} from $q(\cdot)$. Otherwise, generate X_{t+1} from $r(\cdot|\mathsf{X}_t)$.

Now construct a second Markov chain $(\mathsf{Y}_t)_{t=0}^{\infty}$. This chain is a copy of $(\mathsf{X}_t)_{t=0}^{\infty}$, and it is constructed in such a way that $(\mathsf{X}_t)_{t=0}^{\infty}$ and $(\mathsf{Y}_t)_{t=0}^{\infty}$ eventually become the same chain (i.e. $(\mathsf{X}_t)_{t=0}^{\infty}$ and $(\mathsf{Y}_t)_{t=0}^{\infty}$ couple). Let X_0 be an arbitrary initial value of $(\mathsf{X}_t)_{t=0}^{\infty}$ and draw Y_0 from the stationary distribution π. Given $(\mathsf{X}_t, \mathsf{Y}_t)$, the simulation of $(\mathsf{X}_{t+1}, \mathsf{Y}_{t+1})$ occurs in one of two ways, the choice of which depends on whether $(\mathsf{X}_t, \mathsf{Y}_t) \in C \times C$.

If $(\mathsf{X}_t, \mathsf{Y}_t)$ is not in $C \times C$, draw X_{t+1} according to $r(\cdot|\mathsf{X}_t)$, and independently draw Y_{t+1} according to $r(\cdot|\mathsf{Y}_t)$. Otherwise, generate a $0-1$ coin flip according to a Bernoulli(ϵ) distribution. If the outcome is 0, draw X_{t+1} according to $r(\cdot|\mathsf{X}_t)$ and independently draw Y_{t+1} according to $r(\cdot|\mathsf{Y}_t)$. If the outcome is 1, generate $\mathsf{X}_{t+1} = \mathsf{Y}_{t+1}$ according to $q(\cdot)$. The first value of t for which $\mathsf{X}_t = \mathsf{Y}_t$ is the coupling time, denoted τ_C. For all $t \geq \tau_C$, draw $(\mathsf{X}_t, \mathsf{Y}_t)$ so that $\mathsf{X}_t = \mathsf{Y}_t$.

Recall that for a Markov chain $(\mathsf{X}_t)_{t=0}^{\infty}$ with transition kernel K and stationary distribution π,

$$\|K^k(\mathbf{x}, \cdot) - \pi(\cdot)\|_{TV} \leq \Pr(\tau_C > k).$$

Let $r_1 = \inf\{m : (\mathsf{X}_m, \mathsf{Y}_m) \in C \times C\}$ represent the first time $(\mathsf{X}_t, \mathsf{Y}_t)_{t=0}^{\infty}$ returns to $C \times C$, and for $i = 2, 3, \ldots,$ let $r_i = \inf\{m > r_{i-1} : (\mathsf{X}_m, \mathsf{Y}_m) \in C \times C\}$. The r_i are known as the *return times* to $C \times C$. Let $N_k = \max\{i : r_i < k\}$ be the number of times $(\mathsf{X}_t, \mathsf{Y}_t)_{t=0}^{\infty}$ returns to $C \times C$ prior to time k. Whenever $(\mathsf{X}_t, \mathsf{Y}_t)_{t=0}^{\infty} \in C \times C$, the probability that the chains couple is ϵ. To bound

$\Pr(\tau_C > k)$, note that

$$\begin{aligned}\Pr(\tau_C > k) &= \Pr(\tau_C > k \text{ and } N_k \geq j) + \Pr(\tau_C > k \text{ and } N_k < j) \\ &\leq \Pr(\tau_C > k \text{ and } N_k \geq j) + \Pr(N_k < j), \qquad (3.1)\end{aligned}$$

where $j \in \mathbb{Z}^+$.

The first term on the right-hand side of (3.1) gives the probability that the pair of chains has entered $C \times C$ at least j times prior to time k, and at none of these times have the two chains coupled. This quantity is bounded above by $(1-\epsilon)^j$. Therefore,

$$\|K^k(\mathbf{x}, \cdot) - \pi(\cdot)\|_{TV} \leq (1-\epsilon)^j + \Pr(N_k < j).$$

In some cases, it is not required that we go any further than this to obtain an upper bound on the total variation distance between K^k and π. The following theorem (Meyn and Tweedie, 2009) gives a scenario in which an upper bound on the total variation distance can be obtained by verifying only a minorization condition.

Theorem 5. *(Meyn and Tweedie, 2009) Suppose a Markov chain $(\mathsf{X}_t)_{t=0}^{\infty}$ with state space $\mathbb{R}^m$ has a stationary distribution π and is π-irreducible, aperiodic, and Harris recurrent. If $(\mathsf{X}_t)_{t=0}^{\infty}$ satisfies a minorization condition with small set $C = \mathbb{R}^m$, then*

$$\|K^k(\mathsf{X}_0, \cdot) - \pi(\cdot)\|_{TV} \leq (1-\epsilon)^k.$$

If the entire state space is small, then clearly $(\mathsf{X}_t, \mathsf{Y}_t) \in C \times C$ for all $t \geq 0$. Therefore, at each transition, the chains couple with probability ϵ, so that

$\Pr(\tau_C > k) = (1-\epsilon)^k$. In addition, for all $j < k$, $\Pr(N_k < j) = 0$ since $(\mathsf{X}_t, \mathsf{Y}_t)_{t=0}^{\infty}$ returns to $C \times C$ at every step.

The conditions of Theorem 5 are very strong, and instances of MCMC algorithms that satisfy all of them are very rare. Therefore, an upper bound on $\Pr(N_k < j)$ is usually necessary. The verification of a drift condition takes care of this.

3.2.2 Role of the Drift Condition

In order to obtain an upper bound on $\Pr(N_k < j)$, note first that if $N_k < j$, then $r_j > k$. Therefore,

$$\begin{aligned}\Pr(N_k < j) &= \Pr(r_j > k) \\ &= \Pr\left(r_1 + \sum_{i=2}^{j}(r_i - r_{i-1}) > k\right).\end{aligned}$$

This implies that for any $\alpha > 1$,

$$\begin{aligned}\Pr\left(r_1 + \sum_{i=2}^{j}(r_i - r_{i-1}) > k\right) &= \Pr\left(\alpha^{r_1 + \sum_{i=2}^{j}(r_i - r_{i-1})} > \alpha^k\right) \\ &\leq \frac{1}{\alpha^k}\mathbb{E}\left[\alpha^{r_1}\prod_{i=2}^{j}\alpha^{r_i - r_{i-1}}\right],\end{aligned}$$

by the Markov Inequality.

The goal here is to bound the exponential moments of the times the pair of chains spends outside of $C \times C$. This is generally difficult, but the following lemma (Rosenthal, 1995) provides the ability to obtain an upper bound on the total variation distance between K^k and π without having to find the exponential moments of the time between the returns of $(\mathsf{X}_t, \mathsf{Y}_t)_{t=0}^{\infty}$ to $C \times C$.

Lemma 14. *(Rosenthal, 1995) Let* $(\mathsf{X}_t)_{t=0}^{\infty}$, $(\mathsf{Y}_t)_{t=0}^{\infty}$, *and* $r_i, i = 1, 2, \ldots$ *be defined as above. Suppose there exists a constant* $\alpha > 1$ *and a function* $h : \mathbb{R}^m \times \mathbb{R}^m \mapsto [1, \infty)$ *such that*

$$\mathbb{E}\left[h(\mathsf{X}_1, \mathsf{Y}_1)|\mathsf{X}_0 = \mathbf{x}, \mathsf{Y}_0 = \mathbf{y}\right] \leq \alpha^{-1} h(\mathbf{x}, \mathbf{y}) \qquad \forall (\mathbf{x}, \mathbf{y}) \in C \times C.$$

Then for any integer $i > 1$ *and any choice of* $r_1, \ldots, r_{i-1}$ *such that* $r_1 < r_2 < \ldots < r_{i-1}$,

1) $\mathbb{E}\left[\alpha^{r_1}\right] \leq \mathbb{E}_{\psi \times \pi}\left[h(\mathsf{X}_0, \mathsf{Y}_0)\right]$ *and for* $i > 1$ *and any choice of* $r_1, \ldots, r_{i-1}$,

2) $\mathbb{E}\left[\alpha^{r_i - r_{i-1}} | r_1, \ldots, r_{i-1}\right] \leq \sup_{(\mathbf{x},\mathbf{y}) \in C \times C} \mathbb{E}\left[h(\mathsf{X}_1, \mathsf{Y}_1)|\mathsf{X}_0 = \mathbf{x}, \mathsf{Y}_0 = \mathbf{y}\right].$

Suppose $(\mathsf{X}_t)_{t=0}^{\infty}$ satisfies a minorization condition on a small set C for some $\epsilon < 1$ and density $q(\cdot)$, and suppose that $(\mathsf{X}_t)_{t=0}^{\infty}$ also satisfies the conditions of Lemma 14 for some function $h : \mathbb{R}^m \times \mathbb{R}^m \mapsto [1, \infty)$ and a fixed $\alpha > 1$. Set $A = \sup_{(\mathbf{x},\mathbf{y}) \in C \times C} \mathbb{E}\left[h(\mathsf{X}_1, \mathsf{Y}_1)|\mathsf{X}_0 = \mathbf{x}, \mathsf{Y}_0 = \mathbf{y}\right]$. The result of Lemma 14 implies that

$$\begin{aligned}
&\Pr(N_k < j) \\
&\leq \alpha^{-k} \mathbb{E}\left[\alpha^{r_1} \prod_{i=2}^{j} \alpha^{r_i - r_{i-1}}\right] \\
&\leq \alpha^{-k} \left(\prod_{i=2}^{j} \mathbb{E}\left[\alpha^{r_i - r_{i-1}} | r_1, \ldots, r_{i-1}\right]\right) \mathbb{E}\left[\alpha^{r_1}\right] \\
&= \alpha^{-k+j-1} A^{j-1} \mathbb{E}_{\psi \times \pi}\left[h(\mathsf{X}_0, \mathsf{Y}_0)\right],
\end{aligned}$$

where ψ denotes the distribution from which X_0 is drawn. The above work brings us closer to geometric ergodicity, but since π is not known in general, it is difficult to bound $\mathbb{E}_{\psi \times \pi}\left[h(\mathsf{X}_0, \mathsf{Y}_0)\right]$. This problem can be circumvented if

it can be shown that $(\mathsf{X}_t)_{t=0}^{\infty}$ satisfies a drift condition.

Suppose $(\mathsf{X}_t)_{t=0}^{\infty}$ satisfies a drift condition, so that there exists a function $V : \mathbb{R}^m \mapsto [1, \infty)$ and constants $\lambda \in (0, 1)$ and $b < \infty$ such that for all $\mathbf{x} \in C$,

$$\mathbb{E}\left[V(\mathsf{X}_1)|\mathsf{X}_0 = \mathbf{x}\right] \leq \lambda V(\mathbf{x}) + b\mathbb{I}_C(\mathbf{x}),$$

and let $h(\mathbf{x}, \mathbf{y}) = 1 + V(\mathbf{x}) + V(\mathbf{y})$. Meyn and Tweedie (2009) show that

$$\mathbb{E}_{\psi\times\pi}\left[h(\mathsf{X}_0, \mathsf{Y}_0)\right] \leq 1 + \mathbb{E}_{\psi}\left[V(\mathsf{X}_0)\right] + \frac{b}{1-\lambda}.$$

The result of Meyn and Tweedie (2009) yields the following upper bound on $\Pr(N_k < j)$:

$$\Pr(N_k < j) \leq \alpha^{-k}\alpha^{j-1}A^{j-1}\left(1 + \mathbb{E}_{\psi}\left[V(\mathsf{X}_0)\right] + \frac{b}{1-\lambda}\right).$$

Since $j < k$, there exists $r \in (0, 1)$ such that $j = rk + 1$, and the total variation distance between K^k and π is bounded in the following way.

$$\begin{aligned}
\|K^k(\mathbf{x}, \cdot) - \pi(\cdot)\|_{TV} &\leq (1-\epsilon)^{rk} + \alpha^{-k+rk}A^{rk}\left(1 + \mathbb{E}_{\psi}\left[V(\mathsf{X}_0)\right] + \frac{b}{1-\lambda}\right) \\
&= (1-\epsilon)^{rk} + \left(\alpha^{-(1-r)}A^r\right)^k\left(1 + \mathbb{E}_{\psi}\left[V(\mathsf{X}_0)\right] + \frac{b}{1-\lambda}\right).
\end{aligned}$$

Choose r so that $\alpha^{-(1-r)}A^r < 1$, let

$$\kappa = \max\left\{(1-\epsilon)^r, \alpha^{-(1-r)}A^r\right\},$$

and let $R(\mathbf{x}) = 1 + \mathbb{E}_{\psi}\left[V(\mathsf{X}_0)\right] + \frac{b}{1-\lambda}$. These representations provide the following result:

$$\|K^k(\mathbf{x}, \cdot) - \pi(\cdot)\|_{TV} \leq R(\mathbf{x})\kappa^k.$$

Therefore, if $(\mathsf{X}_t)_{t=0}^{\infty}$ satisfies a minorization condition and an associated drift condition, then $(\mathsf{X}_t)_{t=0}^{\infty}$ is geometrically ergodic. The association between

the minorization and drift conditions is in the small set. Note that the minorization condition and the drift condition each depend on C. The roles of each condition demonstrate why this is necessary. Returning to the coupling $(\mathsf{X}_t, \mathsf{Y}_t)_{t=0}^{\infty}$, recall that each time the pair of chains enters $C \times C$, there is a chance for the chains to couple. The minorization condition ensures that the chain requires only a geometric number of such opportunities in order to guarantee convergence. The drift condition guarantees that the distribution of the times between returns to $C \times C$ has tails that are thin enough to ensure that these times have finite exponential moments.

Rosenthal (1995) takes the use of the drift and minorization conditions a step further by showing that they can be used not only to verify geometric ergodicity, but that the values of ϵ, λ, and b can be used to provide an explicit upper bound on the mixing time.

Theorem 6. *(Rosenthal, 1995) Suppose that for a function $V : \mathbb{R}^m \mapsto [1, \infty)$ and constants $\lambda \in (0, 1)$ and $b < \infty$,$(\mathsf{X}_t)_{t=0}^{\infty}$ satisfies*

$$\mathbb{E}\left[V(\mathsf{X}_1)|\mathsf{X}_0 = \mathbf{x}\right] = \lambda V(\mathbf{x}) + b\mathbb{I}_C(\mathbf{x})$$

for all $\mathbb{R}^m$, where $C = \{\mathbf{x} : V(\mathbf{x}) \leq d\}$ and $d > \frac{2b}{1-\lambda} - 1$. Suppose also that for some $\epsilon > 0$ and some probability measure $Q(\cdot)$ on $\mathcal{B}(\mathbb{R}^m)$,

$$K(\mathbf{x}, A) \geq \epsilon Q(A)$$

for all $A \in \mathcal{B}(\mathbb{R}^m)$ and for all $\mathbf{x} \in C$. Then for any $r \in (0, 1)$ with $(\mathsf{X}_t)_{t=0}^{\infty}$ beginning in the initial distribution ψ,

$$\|K^k(\mathsf{X}_0, \cdot) - \pi(\cdot)\|_{TV} \leq (1-\epsilon)^{rk} + \left(\alpha^{-(1-r)} A^r\right)^k \left(1 + \frac{b}{1-\lambda} + \mathbb{E}_{\psi}\left[V(\mathsf{X}_0)\right]\right),$$

where $\alpha^{-1} = \lambda + [b + (1-\lambda)] / \left[1 + \frac{d-1}{2}\right]$ *and* $A = 1 + (\lambda d + b)$.

All of the work presented in this section has been dependent on a one-step minorization condition. Verification of a more general minorization condition can be used to guarantee geometric ergodicity.

A Markov chain $(\mathsf{X}_t)_{t=0}^{\infty}$ on $\mathbb{R}^m$ with transition kernel $K(\mathbf{x}, \cdot)$ satisfies an m_0-step minorization condition if there exists a probability measure $Q(\cdot)$, a small set $C \subset \mathbb{R}^m$, a positive integer m_0 and $\epsilon \in (0,1)$ such that

$$K^{m_0}(\mathbf{x}, A) \geq \epsilon Q(A) \text{ for all } \mathbf{x} \in C, \tag{3.2}$$

for all $A \in \mathcal{B}(\mathbb{R}^m)$. This more general minorization condition is very important to the work presented later in this dissertation. The establishment of geometric ergodicity using this minorization condition is a straightforward extension of the work above.

3.3 The Random-Scan Metropolis Algorithm

The RSM algorithm is a version of the Metropolis-Hastings algorithm that updates one parameter at a time. Let $(\mathsf{X}_t)_{t=0}^{\infty}$ be a Markov chain on $\mathbb{R}^m$, where X_0 is drawn from some initial distribution ψ whose density has the same support as does the target distribution. Given X_t, obtain X_{t+1} in the following way. First choose one of the components of X_t uniformly at random. The chosen entry, say $\mathsf{X}_t^{(i)}$, $i \in \{1, 2, \ldots m\}$, is updated as follows. From a symmetric increment density $q_i(\cdot)$, draw an increment value y. The increment is added to $\mathsf{X}_t^{(i)}$, so that the proposed value is $\mathsf{X}^* = \mathsf{X}_t + y\mathbf{e}_i$, where $\mathbf{e}_i$ is the unit vector in the ith direction.

Let $p(\cdot)$ denote the target density for $(\mathsf{X}_t)_{t=0}^{\infty}$. Then $\mathsf{X}_{t+1} = \mathsf{X}_t + y\mathbf{e}_i$ with probability

$$\alpha_i(\mathsf{X}_t, \mathsf{X}^*) = \min\left\{1, \frac{p(\mathsf{X}^*)}{p(\mathsf{X}_t)}\right\}.$$

Given that the ith entry has been selected for updating, the conditional transition density is

$$k_i(\mathsf{X}_{t+1}|\mathsf{X}_t) = \alpha_i(\mathsf{X}_t, \mathsf{X}^*)q_i(y)\prod_{\substack{j=1\\ j\neq i}}^{m}\delta_{\mathsf{X}_t^{(j)}}(\mathsf{X}_{t+1}^{(j)}) +$$
$$\left(\int_{\mathbb{R}}(1-\alpha_i(\mathsf{X}_t, \mathsf{X}^*)q_i(y))\,\mathrm{d}y\right)\delta_{\mathsf{X}_t}(\mathsf{X}_{t+1}),$$

where $\delta_{\mathbf{x}}(\cdot)$ is the Dirac mass measure at $\mathbf{x}$. Since the parameter to be updated is chosen uniformly at random, obtaining the full transition density amounts to simply averaging the m conditional transition densities, so that

$$k(\mathsf{X}_{t+1}|\mathsf{X}_t) = \frac{1}{m}\sum_{i=1}^{m} k_i(\mathsf{X}_{t+1}|\mathsf{X}_t). \tag{3.3}$$

3.4 Geometric Ergodicity of the Random Scan Metropolis Sampler

For an RSM algorithm on $\mathbb{R}^m$, there are several methods of verifying geometric ergodicity that work by providing conditions that are sufficient to ensure satisfaction of a drift and an associated minorization condition. These methods are useful because for the RSM algorithm, a drift function is often difficult to find analytically, and this makes direct verification of a drift and a minorization condition difficult. These methods provide a way to deal with this challenge.

Roberts and Tweedie (1996) provide a set of four conditions on the curvature of the contour manifold $\{\mathbf{y} : p(\mathbf{y}) = p(\mathbf{x})\}$ as $\|\mathbf{x}\| \to \infty$ of the target density. These four conditions together ensure geometric ergodicity of the RSM sampler. Verification of these conditions is not easy for many applications of the RSM sampler, including the one that is described later in this chapter. Jarner and Hansen (2000) provide a set of conditions that ensure geometric ergodicity of the RSM algorithm under the condition that the target density is *super-exponential*. This means that

$$\lim_{\|\mathbf{x}\| \to \infty} n(\mathbf{x}) \nabla \log p(\mathbf{x}) = -\infty,$$

where $n(\mathbf{x})$ is the unit vector $\mathbf{x}/\|\mathbf{x}\|$ and ∇ denotes the gradient. This method requires taking partial derivatives of the log target density. While this method and the method of Roberts and Tweedie (1996) ensure the satisfaction of a drift and a minorization condition, neither of them provide a drift function. Fort et al. (2003) describes a method that takes care of this.

Fort et al. (2003) provide a sufficient set of three assumptions on p that ensure geometric ergodicity of the RSM sampler. Their work establishes geometric ergodicity of the RSM sampler with essentially no conditions on the curvature of the contour manifold of p. These assumptions are easier to verify than those given by Roberts and Tweedie (1996) and Jarner and Hansen (2000) for our RSM sampler, so we choose to verify the conditions of Fort et al. (2003) in order to establish geometric ergodicity of an RSM sampler for estimating the posterior density of the branch lengths of a phylogenetic tree.

The assumptions of Fort et al. (2003) ensure that the target density is positive, continuous, and bounded on $\mathbb{R}^m$, that the tails of the target density are decreasing, and that the target density is bounded away from 0 on compact sets. The formal statement of these assumptions is as follows:

Assumption 1 The stationary distribution π is absolutely continuous with respect to λ_m, with positive and continuous density $p(\cdot)$ on $\mathbb{R}^m$.

Assumption 2 Let $\{q_i\}_{i=1}^m$ be a family of symmetric increment densities with respect to λ_1, the one-dimensional Lebesgue measure. There exist constants $\eta_i > 0$ and $\delta_i < \infty$ for all $i \in \{1, 2, \ldots, m\}$ such that $q_i(y) \geq \eta_i$ whenever $|y| \leq \delta_i$.

Assumption 3 There exist constants δ and Δ with $0 \leq \delta < \Delta \leq +\infty$ such that

$$\xi := \inf_{1 \leq i \leq m} \int_{\delta}^{\Delta} q_i(y)\lambda_1(dy) > 0. \tag{3.4}$$

In addition, for any sequence $\mathbf{x} = \{\mathbf{x}_n\}$ with $\lim_{n\to\infty}\|\mathbf{x}_n\| = +\infty$, it is possible to extract a subsequence $\tilde{\mathbf{x}} = \{\tilde{\mathbf{x}}_n\}$ with the property that, for some $i \in \{1, 2, \ldots, m\}$ and for all $y \in [\delta, \Delta]$,

$$\lim_{n\to\infty} \frac{p(\tilde{\mathbf{x}}_n)}{p(\tilde{\mathbf{x}}_n - \text{sign}(\tilde{\mathbf{x}}_n)y\mathbf{e}_i)} = 0 \text{ and } \lim_{n\to\infty} \frac{p(\tilde{\mathbf{x}}_n + \text{sign}(\tilde{\mathbf{x}}_n)y\mathbf{e}_i)}{p(\tilde{\mathbf{x}}_n)} = 0, \tag{3.5}$$

where $p(\cdot)$ is the target density for the RSM algorithm on $\mathbb{R}^m$.

The following theorem (Fort et al., 2003) provides a drift function if Assumptions $1, 2,$ and 3 hold.

Theorem 7. *Assume that Assumptions* $1, 2,$ *and* 3 *hold, and let* $s \in (0,1)$ *such that*

$$s(1-s)^{\frac{1}{s}-1} < \frac{\xi}{m-2\xi}, \tag{3.6}$$

where ξ *is defined as in* (3.4)*. Let* $V_s(\mathbf{x}) := [p(\mathbf{x})]^{-s}$*. Then there exist constants* $\lambda \in (0,1)$ *and* $b < \infty$ *as well as a small set* $C \in \mathcal{B}(\mathbb{R}^m)$ *such that*

$$\mathbb{E}\left[V_s(\mathsf{X}_{t+1})|\mathsf{X}_t = \mathbf{x}\right] \leq \lambda V_s(\mathbf{x}) + b\mathbb{I}_C(\mathbf{x}), \tag{3.7}$$

for all $\mathbf{x} \in \mathbb{R}^m$*.*

3.5 An RSM Algorithm for Bayesian Inference of the Branch Lengths

Theorem 7 implies that verification of Assumptions 1, 2, and 3 for any RSM sampler is sufficient to ensure that the sampler is geometrically ergodic. Our approach to approximating the posterior density of the branch lengths relies on a version of the RSM algorithm. In this description, $\mathbf{D}$ shall denote the set of DNA sequence data for the individuals at the leaves of the tree. We assume the tree topology is known or has been inferred, and we assume that the tree is unrooted. Whether or not the tree is rooted is irrelevant to the RSM algorithm we describe, since by the Pulley Principle, there is no information about the likelihood in the placement of the root. The payoff of assuming unrootedness is a one-dimension reduction in the state space.

The assumption of unrootedness implies that the vector $\mathbf{t}$ of branch lengths is in $[0,\infty)^{2n-3}$, where n is the number of leaves on the tree. Let $\phi(\cdot)$ denote the prior density for the branch lengths $\mathbf{t}$, and let the increment density $q_i(\cdot)$ be

the $U[-\gamma, \gamma]$ density for each $i \in \{1, 2, \ldots, 2n-3\}$, where $\gamma > 0$ is fixed. The choice of γ is made in such a way that it ensures that the algorithm accepts a proposal at a rate that is optimal for mixing behavior. As is usual in Bayesian analysis, the target density is known up to a normalizing constant:

$$p(\mathbf{t}|\mathbf{D}) \propto \mathcal{L}(\mathbf{D}|\mathbf{t})\phi(\mathbf{t}).$$

We require that p is positive and continuous over all of $\mathbb{R}^{2n-3}$, but since the branch lengths are non-negative, $\phi(\mathbf{t}) = 0$ for all values of $\mathbf{t}$ that have at least one negative component. A solution to this problem is to re-parameterize the branch lengths. Many transformations exist that result in the satisfaction of our requirement, but the natural log transformation is a continuous, one-to-one, monotone increasing transformation, so its inverse is well-defined. This makes conversion of the log branch lengths very simple, and makes the log transformation a good choice of a transformation for the branch lengths. The nice mathematical properties of the log transformation enable easy transformation of the likelihood, so that if $\mathbf{w} = \log(\mathbf{t})$, the likelihood can be written in terms of $\mathbf{w}$:

$$\mathcal{L}(\mathbf{D}|\mathbf{t}) = \mathcal{L}(\mathbf{D}|e^{\mathbf{w}}),$$

where $e^{\mathbf{w}}$ denotes the vector that results from the replacement of each component of $\mathbf{w}$ by the exponentiated value of that component. We omit the exponentiation of $\mathbf{w}$ and denote the likelihood by $\mathcal{L}(\mathbf{D}|\mathbf{w})$.

The Markov chain $(\mathsf{W}_t)_{t=0}^{\infty}$ associated with the RSM sampler that updates the log branch lengths begins with a choice of an initial value W_0 from a density that has the same support as the target density. The prior distribution is

one of many good candidates for the initial distribution, provided it has the required support, and we use it as such for our version of the RSM algorithm.

Once an initial value W_0 is drawn from $\phi(\cdot)$, choose a component $i \in \{1, 2, \ldots, 2n-3\}$ uniformly at random, and an increment $y \sim q_i(\cdot)$. Given W_t, the proposal for W_{t+1} is $\mathsf{W}^* = \mathsf{W}_t + y\mathbf{e}_i$, and this proposal is accepted with probability

$$\alpha_i(\mathsf{W}_t, \mathsf{W}^*) = \min\left\{1, \frac{\mathcal{L}(\mathbf{D}|\mathsf{W}^*)\phi(\mathsf{W}^*)}{\mathcal{L}(\mathbf{D}|\mathsf{W}_t)\phi(\mathsf{W}_t)}\right\}.$$

Calculation of the Likelihood

To complete the derivation of the transition kernel, it is necessary to find $p(\mathbf{w}|\mathbf{D})$. Recall that for branch lengths $\mathbf{t}$, the likelihood $\mathcal{L}(\mathbf{D}|\mathbf{t})$ is given by

$$\mathcal{L}(\mathbf{D}|\mathbf{t}) = \sum_{s_1}\sum_{s_2} \pi_{s_1} \Pr(s_2|s_1, t^{(1)})\mathcal{L}_{s_1}(1)\mathcal{L}_{s_2}(2), \tag{3.8}$$

where π_{s_1} is the equilibrium probability of s_1, $s_1 \in \{A, C, G, T\}$. Recall also that under the Jukes-Cantor model, for two nucleotides r and s,

$$\Pr(r|s, w) = \begin{cases} \frac{1}{4} + \frac{3}{4}e^{-\frac{4}{3}e^w} & \text{if } r = s \\ \frac{1}{4} - \frac{1}{4}e^{-\frac{4}{3}e^w} & \text{if } r \neq s. \end{cases}$$

We define the following three terms, none of which depend on $w^{(1)}$, the first entry in the log branch length vector. An illustration that indicates the reason that these three quantities do not depend on $w^{(1)}$ is provided in Figure 3.1. Define the following quantities:

$$\begin{aligned} K_1 &= \sum_{s_1} \mathcal{L}_{s_1}(1) \\ K_2 &= \sum_{s_2} \mathcal{L}_{s_2}(2) \\ J &= \sum_{s_1} \mathcal{L}_{s_1}(1)\mathcal{L}_{s_1}(2). \end{aligned}$$

The likelihood can be written in terms of J, K_1, and K_2 as follows:

$$\begin{aligned}\mathcal{L}(\mathbf{D}|\mathbf{w}) &= \frac{1}{4}\left[\left(\frac{1}{4}+\frac{3}{4}e^{-\frac{4}{3}e^{w^{(1)}}}\right)J+\left(\frac{1}{4}-\frac{1}{4}e^{-\frac{4}{3}e^{w^{(1)}}}\right)(K_1K_2-J)\right] \\ &= \frac{1}{4}\left[\frac{1}{4}K_1K_2+\left(J-\frac{1}{4}K_1K_2\right)e^{-\frac{4}{3}e^{w^{(1)}}}\right]. \qquad (3.9)\end{aligned}$$

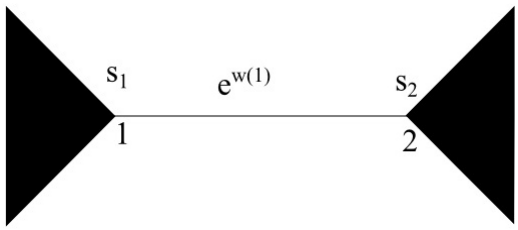


Figure 3.1: The conditional likelihood $\mathcal{L}_{s_1}(1)$ represents the likelihood of the subtree on the left side of the unrooted tree above given the nucleotide base s_1 is observed at node 1, with $\mathcal{L}_{s_2}(2)$ representing a similar quantity pertaining to node 2. Since the branch with length $e^{w^{(1)}}$ is not a part of either of the two outer subtrees, neither $\mathcal{L}_{s_1}(1)$ nor $\mathcal{L}_{s_2}(2)$ depend on $w^{(1)}$.

In general, the DNA sequences at the leaves will have some number $N > 1$ of sites apiece. Recall that in likelihood calculation, it is usually assumed that evolution is independent between sites. Therefore, we can treat the vector of bases at each site as a one-site data set and then calculate the likelihood as in (3.9). The full likelihood is then obtained by taking the product of all N one-site likelihoods.

In Bayesian phylogenetic inference, common branch length priors include the exponential, lognormal, and gamma densities. The branch length units are in expected number of substitutions per site, and this number is typically

small. This leads to data sets for which a high percentage of sites have no mutations (i.e. sites for which the nucleotide base is the same in each sequence). We refer to these as *constant sites*. The the normal density is symmetric, and therefore puts equal weight on large and small branch lengths. The exponential, lognormal, and gamma densities are all right-skewed densities, so they put more weight on small branch lengths than they do on large branch lengths. While smaller branch lengths are what is expected with a realistic data set, values that are extremely close to 0 are associated with data sets which have an unrealistically high percentage of constant sites. Of the three densities, the lognormal places the least weight on these extremely small branch lengths. In addition, the other priors mentioned have tails that are too thick to allow the methods of Fort et al. (2003) to be used to verify geometric ergodicity of our RSM algorithm for inferring branch lengths. As a result, we place a lognormal prior density on the branch lengths in such a way that the branch lengths are uncorrelated. The prior assumption that the branch lengths are uncorrelated is an unrealistic one, but it is, nonetheless, a common assumption in phylogenetic analysis (see Mar et al. (2005),Yang and Rannala (2005), and Brown et al. (2010) for examples). Given a vector $\boldsymbol{\mu}$ and a constant $\sigma^2 > 0$, for all $i \in \{1, 2, \ldots, 2n-3\}$, the ith branch length has a lognormal$(\mu^{(i)}, \sigma^2)$ distribution. The placement of this prior distribution on the branch length implies that the prior density for the log branch lengths is the $N_{2n-3}(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_{2n-3})$ density. For the case in which the DNA sequences each have one site, the target

density, up to a normalizing constant, is

$$\begin{aligned} p(\mathbf{w}|\mathbf{D}) &\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{2n-3}(w^{(i)}-\mu^{(i)})^2\right\} \\ &\times \left(\frac{1}{4}K_1K_2+\left(J-\frac{1}{4}K_1K_2\right)e^{-\frac{4}{3}w^{(1)}}\right). \end{aligned}$$

3.6 Geometric Ergodicity of the RSM Algorithm for Inference of the Branch Lengths

In this section, we establish geometric ergodicity of $(\mathsf{W}_t)_{t=0}^{\infty}$ by verifying that the three assumptions made by Fort et al. (2003) are satisfied.

Theorem 8. *The Markov chain* $(\mathsf{W}_t)_{t=0}^{\infty}$ *is geometrically ergodic.*

Proof. For the RSM algorithm described in Section 3.5, we verify the three assumptions outlined in Section 3.4.

Verification of Assumption 1

For Assumption 1, it is required to show that the target density p is positive and continuous over $\mathbb{R}^{2n-3}$. This is false only if at least one of the site likelihoods is 0. Recall from (3.8) that the site likelihood is a sum of products of three non-negative terms: one base substitution probability and two conditional likelihoods. Therefore, each of the summands in the site likelihood must be 0, which means that either the base substitution probability or one of the conditional likelihoods is 0.

Under the Jukes-Cantor model, for nucleotide bases r and s, the substitution probability can only be 0 if $r \neq s$. If $r \neq s$,

$$\Pr(r|s,t) = \frac{1}{4} - \frac{1}{4}e^{-\frac{4}{3}t}.$$

This probability is equal to 0 only if $t = 0$.

The conditional likelihood is a sum of products of substitution probabilities, so if each summand in the conditional likelihood is 0, then some of the substitution probabilities are 0. Therefore, a necessary condition for the likelihood to be 0, is that at least one branch length is 0. By our formulation, $t^{(i)} = e^{w^{(i)}}$ for all $i \in \{1, 2, \ldots, 2n-3\}$, so that $t^{(i)} > 0$ for all $i \in \{1, 2, \ldots, 2n-3\}$. This and the fact that the prior density is the $N(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_{2n-3})$ density implies that the target density p is positive for all $\mathbf{w} \in \mathbb{R}^{2n-3}$. In addition, p is continuous in $\mathbf{w}$ over $\mathbb{R}^{2n-3}$, so Assumption 1 is verified.

Verification of Assumption 2

Assumption 2 is easily verified, since the increment density is uniform on the interval $[-\gamma, \gamma]$ for each log branch length. In keeping with the notation in the statement of Assumption 2, let $\eta_i = \frac{1}{2\gamma}$ and let $\delta_i = \gamma$ for each $i \in \{1, 2, \ldots, 2n-3\}$. Then $q_i(y) = \eta_i$ whenever $|y| \leq \gamma$, thus verifying Assumption 2.

Verification of Assumption 3

Using the notation from the statement of Assumption 3, let $\delta = 0$ and $\Delta = \gamma$. Since the increment density associated with each log branch length is

the $U[-\gamma, \gamma]$ density, we obtain the following.

$$\begin{aligned}
\xi &:= \inf_{1 \le i \le 2n-3} \int_0^{\gamma} q_i(y) \mathrm{d}y \\
&= \int_0^{\gamma} \frac{1}{2\gamma} \mathrm{d}y \\
&= \frac{1}{2} > 0,
\end{aligned}$$

and this verifies the first part of Assumption 3.

Let $\{\mathbf{w}_k\}_{k=0}^{\infty}$ be a sequence of log branch lengths with the property that $\|\mathbf{w}_k\| \to \infty$ as $k \to \infty$. If $\|\mathbf{w}_k\| \to \infty$, then at least one of the log branch lengths has magnitude that tends to infinity. Therefore, there exists $j \in \{1, 2, \ldots, 2n-3\}$ and a subsequence $\{\tilde{\mathbf{w}}_k\}_{k=0}^{\infty}$ such that if $k \to \infty$, either $\tilde{w}_k^{(j)} \to -\infty$ or $\tilde{w}_k^{(j)} \to \infty$. To verify Assumption 3, it suffices to show that both limits in the statement of Assumption 3 are 0 in each of these two cases. Doing this is equivalent to verifying that both limits are 0 if $w_k^{(j)} \to -\infty$ as $k \to \infty$ and if $w_k^{(j)} \to \infty$ as $k \to \infty$. Without loss of generality, assume $|w_k^{(1)}| \to \infty$ as $k \to \infty$.

Our approach to verifying Assumption 3 consists of establishing a finite upper bound on the likelihood ratio and then showing that the prior ratio tends to 0 as $|w^{(1)}| \to +\infty$. Assume $\mathbf{D}$ is a set of DNA sequences that have one site apiece. We first show that

$$\lim_{k \to \infty} \frac{p(\mathbf{w}_k | \mathbf{D})}{p(\mathbf{w}_k - \mathrm{sign}(w_k^{(1)}) y \mathbf{e}_1 | \mathbf{D})} = 0. \tag{3.10}$$

The expansion of the limit on the left side of (3.10) gives

$$\begin{aligned}
&\lim_{k\to\infty} \frac{p(\mathbf{w}_k|\mathbf{D})}{p(\mathbf{w}_k - \text{sign}(w_k^{(1)})y\mathbf{e}_1|\mathbf{D})} \\
&= \lim_{k\to\infty} \frac{\mathcal{L}(\mathbf{D}|\mathbf{w}_k)\phi(\mathbf{w}_k)}{\mathcal{L}(\mathbf{D}|\mathbf{w}_k - \text{sign}(w_k^{(1)})y\mathbf{e}_1)\phi(\mathbf{w}_k - \text{sign}(w_k^{(1)})y\mathbf{e}_1)} \\
&= \lim_{k\to\infty} \frac{e^{-\frac{1}{2\sigma^2}(w_k^{(1)}-\mu^{(1)})^2}}{e^{-\frac{1}{2\sigma^2}(w_k^{(1)}-\mu^{(1)}-\text{sign}(w_k^{(1)})y)^2}} \\
&\times \frac{\frac{1}{4}K_k 1 K_{2k} + (J_k - \frac{1}{4}K_{1k}K_{2k})e^{-\frac{4}{3}e^{w_k^{(1)}}}}{\frac{1}{4}K_{1k}K_{2k} + (J_k - \frac{1}{4}K_{1k}K_{2k})e^{-\frac{4}{3}e^{w_k^{(1)}-\text{sign}(w_k^{(1)})y}}}.
\end{aligned}$$

First, let $w_k^{(1)} \to -\infty$. If for a fixed value of k, $J_k - \frac{1}{4}K_{1k}K_{2k} < 0$,

$$\begin{aligned}
&\frac{1}{4}K_{1k}K_{2k} + \left(J_k - \frac{1}{4}K_{1k}K_{2k}\right)e^{-\frac{4}{3}e^{w_k^{(1)}}} \leq \\
&\frac{1}{4}K_{1k}K_{2k} + \left(J_k - \frac{1}{4}K_{1k}K_{2k}\right)e^{-\frac{4}{3}e^{w_k^{(1)}+y}},
\end{aligned} \tag{3.11}$$

for all $y \geq 0$.

Observe two things from (3.11). First, we deal only with the case in which $y \geq 0$ because that is all that is required to satisfy Assumption 3. Second, the sign function does not appear in the exponent in the denominator because $w_k^{(1)} \to -\infty$. Therefore, we only deal with the tail of the sequence, and in the tail, $w_k^{(1)} < 0$. A direct consequence of (3.11) is that

$$\frac{\frac{1}{4}K_{1k}K_{2k} + \left(J_k - \frac{1}{4}K_{1k}K_{2k}\right)e^{-\frac{4}{3}e^{w_k^{(1)}}}}{\frac{1}{4}K_{1k}K_{2k} + \left(J_k - \frac{1}{4}K_{1k}K_{2k}\right)e^{-\frac{4}{3}e^{w_k^{(1)}+y}}} \leq 1$$

as $w_k^{(1)} \to \infty$, for $y \geq 0$. If $J_k - \frac{1}{4}K_{1k}K_{2k} \geq 0$ for a fixed value of k,

$$\begin{aligned}
&\frac{1}{4}K_{1k}K_{2k} + \left(J_k - \frac{1}{4}K_{1k}K_{2k}\right) e^{-\frac{4}{3}e^{w_k^{(1)}}} \\
&\leq J_k \left[1 + e^{-\frac{4}{3}e^{w_k^{(1)}}}\right] \\
&\leq 2J_k \\
&\leq 2K_{1k}K_{2k} \\
&\leq 2K_{1k}K_{2k} + 8\left(J_k - \frac{1}{4}K_{1k}K_{2k}\right) e^{-\frac{4}{3}e^{w_k^{(1)}+y}} \\
&= 8\left(\frac{1}{4}K_{1k}K_{2k} + (J_k - \frac{1}{4}K_{1k}K_{2k})e^{-\frac{4}{3}e^{w_k^{(1)}+y}}\right).
\end{aligned}$$

Since the quantity in parentheses in the last expression is equal to the likelihood given in (3.9), the quantity on the right-hand side of the equal sign is bounded above by 8. This gives an upper bound of 8 on the likelihood ratio. Therefore,

$$\begin{aligned}
&\lim_{k\to\infty} \frac{p(\mathbf{w}_k|\mathbf{D})}{p(\mathbf{w}_k - \text{sign}(w_k^{(1)})y\mathbf{e}_1|\mathbf{D})} \leq \\
&8 \lim_{k\to\infty} \exp\left\{\frac{1}{\sigma^2}(w_k^{(1)} - \mu^{(1)})y + \frac{1}{2\sigma^2}y^2\right\}. \qquad (3.12)
\end{aligned}$$

Note that the limit on the right hand side of (3.12) is 0, so the limit of the ratio of the target densities is established for the case where $w_k^{(1)} \to -\infty$.

If $w_k^{(1)} \to \infty$, then bound the likelihood ratio in the following way.

$$\begin{aligned}
\frac{\mathcal{L}(\mathbf{w}_k|\mathbf{D})}{\mathcal{L}(\mathbf{w}_k - \text{sign}(w_k^{(1)})y\mathbf{e}_1|\mathbf{D})} &= \frac{\frac{1}{4}K_{1k}K_{2k} + (J_k - \frac{1}{4}K_{1k}K_{2k})e^{-\frac{4}{3}e^{w_k^{(1)}}}}{\frac{1}{4}K_{1k}K_{2k} + (J_k - \frac{1}{4}K_{1k}K_{2k})e^{-\frac{4}{3}e^{w_k^{(1)}-y}}} \\
&\leq \frac{\frac{3}{4}K_{1k}K_{2k} + \frac{3}{4}K_{1k}K_{2k}e^{-\frac{4}{3}e^{w_k^{(1)}}}}{\frac{1}{4}K_{1k}K_{2k} - \frac{1}{4}K_{1k}K_{2k}e^{-\frac{4}{3}e^{w_k^{(1)}-y}}} \\
&= \frac{3(1 + e^{-\frac{4}{3}e^{w_k^{(1)}}})}{1 - e^{-\frac{4}{3}e^{w_k^{(1)}-y}}} \qquad (3.13)
\end{aligned}$$

The limit as $w_k^{(1)} \to \infty$ of the last expression on the right-hand side of (3.13) is 3, so

$$\lim_{k\to\infty} \frac{p(\mathbf{w}_k|\mathbf{D})}{p(\mathbf{w}_k - y\mathbf{e}_1|\mathbf{D})} \leq$$
$$3 \lim_{k\to\infty} \exp\left\{-\frac{1}{\sigma^2}(w_k^{(1)} - \mu^{(1)})y + \frac{1}{2\sigma^2}y^2\right\}. \tag{3.14}$$

It is easily verified that the limit on the right hand side of (3.14) is 0, thus establishing the second part of Assumption 3.

To verify the final part of Assumption 3, we use a similar approach to that used to establish the left side of (3.5). Observe that if $w_k^{(1)} \to -\infty$ as $k \to \infty$,

$$\lim_{k\to\infty} \frac{p(\mathbf{w}_k + \text{sign}(w_k^{(1)})y\mathbf{e}_1|\mathbf{D})}{p(\mathbf{w}_k|\mathbf{D})} =$$
$$\lim_{k\to\infty} \frac{e^{-\frac{1}{2\sigma^2}((w_k^{(1)}-\mu^{(1)})-y)^2}}{e^{-\frac{1}{2\sigma^2}(w_k^{(1)}-\mu^{(1)})^2}} \times \frac{\frac{1}{4}K_{1k}K_{2k} + (J_k - \frac{1}{4}K_{1k}K_{2k})e^{-\frac{4}{3}e^{w_k^{(1)}-y}}}{\frac{1}{4}K_{1k}K_{2k} + (J_k - \frac{1}{4}K_{1k}K_{2k})e^{-\frac{4}{3}e^{w_k^{(1)}}}}.$$

Suppose $J_k - 1/4K_{1k}K_{2k} < 0$ for some k. Then for $y \geq 0$,

$$\frac{\frac{1}{4}K_{1k}K_{2k} + (J_k - \frac{1}{4}K_{1k}K_{2k})e^{-\frac{4}{3}e^{w_k^{(1)}-y}}}{\frac{1}{4}K_{1k}K_{2k} + (J_k - \frac{1}{4}K_{1k}K_{2k})e^{-\frac{4}{3}e^{w_k^{(1)}}}} \leq 1.$$

If for a fixed value of k, $J_k - 1/4K_{1k}K_{2k} \geq 0$, we obtain an upper bound of 8 on the likelihood ratio by a parallel argument to that used to bound the likelihood ratio in the other limit specified in Assumption 3. Therefore, the likelihood ratio is bounded above by 8 in the tail of $\{\mathbf{w}_k\}_{k=0}^{\infty}$, and

$$\lim_{k\to\infty} \frac{p(\mathbf{w}_k - y\mathbf{e}_1|\mathbf{D})}{p(\mathbf{w}_k|\mathbf{D})} \leq 8 \lim_{k\to-\infty} \exp\left\{\frac{1}{\sigma^2}(w_k^{(1)} - \mu^{(1)})y - \frac{1}{2\sigma^2}y^2\right\}.$$

Observe that this limit is equal to 0. As $w_k^{(1)} \to \infty$, we obtain a limit of 0 on the posterior ratio in a manner similar to that which was used to establish

part 2 of Assumption 3.

The preceding work establishes Assumption 3 when each of the sequences at the leaves of the tree have one site. One thing to observe when the DNA sequences have one site is that in none of the situations described above does the limit of the upper bound on the likelihood ratio exceed 8. The assumption of independence of evolution among sites implies that when the DNA sequences each have $N \geq 1$ sites, the likelihood ratio is asymptotically bounded above by 8^N. Since the prior density does not depend on the data, it doesn't change when the number of sites changes from 1 to N. Therefore, the prior ratios from (3.5) do not change, and their limits are 0 as $\|\mathbf{w}_k\| \to \infty$. This completes verification of Assumption 3 and establishes geometric ergodicity of $(\mathsf{W}_t)_{t=0}^{\infty}$. Since all three of the assumptions of Fort et al. (2003) hold for $(\mathsf{W})_{t=0}^{\infty}$, it follows that $(\mathsf{W}_t)_{t=0}^{\infty}$ is geometrically ergodic. $\square$

We have verified that $(\mathsf{W}_t)_{t=0}^{\infty}$ is geometrically ergodic without directly verifying a drift and a minorization condition. However, knowing that a Markov chain is geometrically ergodic does not give a lot of insight into the mixing time of the chain. Gaining insight into this requires the verification of a drift and a minorization condition, and we take this up in the next chapter.

Chapter 4: Assessing Convergence of the Random Scan Metropolis Algorithm for Inference of the Branch Lengths

The results of Chapter 3 establish geometric ergodicity of the RSM sampler with a $N(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_{2n-3})$ prior distribution on the log branch lengths of a phylogenetic tree with a known tree topology, but they give no indication of the rate at which the chain converges to its stationary distribution. In this chapter, we take a significant step toward resolving this by providing a minorization and an associated drift condition. It is important to note that the methods of Fort et al. (2003) implicitly verify that a drift and a minorization condition hold. Therefore, Theorem 8 could be proven by verifying that a drift and an associated minorization condition are satisfied. In some situations, such as those in which it is only necessary to know a chain is geometrically ergodic (say, for use of Central Limit Theorems for ergodic averages), the methods of Fort et al. (2003) will suffice. However, in situations where the goal is to bound the mixing time, the best hope is to verify a drift and a minorization condition.

We provide a minorization condition by way of analytical methods, and

then present a Monte Carlo method of obtaining a lower bound on the minorization parameter ϵ. The Monte Carlo method provides a lower bound on ϵ that is more helpful in obtaining an upper bound on the mixing time of $(\mathsf{W}_t)_{t=0}^{\infty}$, Markov chain associated with the RSM algorithm presented in Chapter 3. Following the establishment of a minorization condition, we numerically establish an associated drift condition by way of a Monte Carlo integration method. Once a minorization and an associated drift condition are established, we describe in detail some commonly-used output-based methods of convergence assessment, and we comment on some of the benefits and caveats of such methods. The chapter closes with an illustrative example of all the convergence assessment methods described in this chapter, and we also provide a graphical description of how the chain behaves with different prior distributions and for DNA sequence data with sequences of varying lengths and percentages of constant sites.

4.1 A Minorization Condition

Assume the mean vector $\boldsymbol{\mu}$ for the prior distribution is equal to $\mu\mathbf{1}_{2n-3}$, where $\mu \in \mathbb{R}$. Recall the transition density

$$k(\mathsf{W}_{t+1}|\mathsf{W}_t) = \frac{1}{2n-3}\sum_{i=1}^{2n-3} k_i(\mathsf{W}_{t+1}|\mathsf{W}_t),$$

where

$$k_i(\mathsf{W}_{t+1}|\mathsf{W}_t) = \alpha_i(\mathsf{W}_t, \mathsf{W}^*)q_i(y)\prod_{\substack{j=1\\ j\neq i}}^{2n-3} \delta_{\mathsf{W}_t^{(j)}}(\mathsf{W}_{t+1}^{(j)}) +$$

$$\left(\int_{\mathbb{R}} (1-\alpha_i(\mathsf{W}_t, \mathsf{W}^*)q_i(y))\,\mathrm{d}y\right)\delta_{\mathsf{W}_t}(\mathsf{W}_{t+1}).$$

The Dirac masses in the transition density pose a significant challenge in verifying a minorization condition. One way to deal with this is to verify a $(2n-3)$-step minorization condition. To do this, consider the $(2n-3)$-step transition density

$$k(\mathsf{W}_{t+2n-3}|\mathsf{W}_t) = \bigcirc_{j=1}^{2n-3} k(\mathsf{W}_{t+j}|\mathsf{W}_{t+j-1}),$$

where $\bigcirc$ denotes the kernel composition operator. The full expansion of this composition is necessary to verify a $(2n-3)$-step minorization condition. To begin the expansion, write $k(\mathsf{W}_{t+2n-3}|\mathsf{W}_t)$ in the following way:

$$\begin{aligned}
&\bigcirc_{j=1}^{2n-3} k(\mathsf{W}_{t+j}|\mathsf{W}_{t+j-1}) \\
&= \int_{\mathbb{R}^{2n-3}} \cdots \int_{\mathbb{R}^{2n-3}} \prod_{j=1}^{2n-3} k(\mathsf{W}_{t+j}|\mathsf{W}_{t+j-1}) \mathrm{d}\mathsf{W}_{t+1} \ldots \mathrm{d}\mathsf{W}_{t+2n-4} \\
&= \int_{\mathbb{R}^{2n-3}} \cdots \int_{\mathbb{R}^{2n-3}} \prod_{j=1}^{2n-3} \left(\frac{1}{2n-3} \sum_{i=1}^{2n-3} k_i(\mathsf{W}_{t+j}|\mathsf{W}_{t+j-1}) \right) \\
&\mathrm{d}\mathsf{W}_{t+1} \ldots \mathrm{d}\mathsf{W}_{t+2n-4}. \qquad (4.1)
\end{aligned}$$

Let S be the set of all possible samples of size $2n-3$, drawn with replacement, from $\{1, 2, \ldots, 2n-3\}$. Note that the expression in (4.1) is a sum of compositions of conditional transition densities given the log branch length that is chosen to be incremented. This expression can be written more concisely as a sum over all samples $\mathbf{s}$, with replacement, from $\{1, 2, \ldots, 2n-3\}$:

$$k(\mathsf{W}_{t+2n-3}|\mathsf{W}_t) = \frac{1}{(2n-3)^{2n-3}} \sum_{\mathbf{s} \in S} \bigcirc_{j=1}^{2n-3} k_{s^{(j)}}(\mathsf{W}_{t+j}|\mathsf{W}_{t+j-1}),$$

where $s^{(j)}$ is the jth element of the sample $\mathbf{s}$. If R is the set of all possible permutations of the elements of $\{1, 2, \ldots, 2n-3\}$, then since $R \subset S$,

$$k(\mathsf{W}_{t+2n-3}|\mathsf{W}_t) \geq \frac{1}{(2n-3)^{2n-3}} \sum_{\mathbf{r} \in R} \bigcirc_{j=1}^{2n-3} k_{r^{(j)}}(\mathsf{W}_{t+j}|\mathsf{W}_{t+j-1}). \qquad (4.2)$$

Omitting the second term in the sum in the transition density for the RSM algorithm provides the next step in obtaining a lower bound on the $(2n-3)$-step transition density:

$$\begin{aligned}
&\bigcirc_{j=1}^{2n-3} k_j(\mathsf{W}_{t+j}|\mathsf{W}_{t+j-1}) \\
&\geq \int_{\mathbb{R}^{2n-3}} \cdots \int_{\mathbb{R}^{2n-3}} \prod_{j=1}^{2n-3} \left(\prod_{\substack{i=1 \\ i\neq j}}^{2n-3} \delta_{\mathsf{W}_{t+j-1}^{(i)}}(\mathsf{W}_{t+j}^{(i)}) \right) \\
&\times \alpha_j(\mathsf{W}_{t+j-1}, \mathsf{W}_{t+j}) q_j(\mathsf{W}_{t+j}^{(j)}|\mathsf{W}_{t+j-1}^{(j)}) \mathrm{d}\mathsf{W}_{t+1} \ldots \mathrm{d}\mathsf{W}_{t+2n-4}. \qquad (4.3)
\end{aligned}$$

Now consider the set $C = [\mu - 3\gamma/2, \mu + 3\gamma/2]^{2n-3}$, where γ is the maximum increment for a log branch length, and recall that the proposal density given the log branch length that is chosen to be incremented is

$$q_j(\mathsf{W}_{t+j}^{(j)}|\mathsf{W}_{t+j-1}^{(j)}) = \frac{1}{2\gamma} \mathbb{I}_{\left[\mathsf{W}_{t+j-1}^{(j)} - \gamma, \mathsf{W}_{t+j-1}^{(j)} + \gamma\right]}(\mathsf{W}_{t+j}^{(j)}).$$

Then if W_t is in C, each log branch length is between $\mu - 3\gamma/2$ and $\mu + 3\gamma/2$. This implies that

$$q_j(\mathsf{W}_{t+j}^{(j)}|\mathsf{W}_{t+j-1}^{(j)}) \geq \frac{1}{2\gamma} \mathbb{I}_{[\mu-\gamma/2, \mu+\gamma/2]}(\mathsf{W}_{t+j}^{(j)}). \qquad (4.4)$$

To see the result of (4.4), note that the composition in (4.3) updates the log branch lengths in order. Therefore, if W_t is in C, then $\mathsf{W}_{t+j-1}^{(j)}$ is between $\mu - 3\gamma/2$ and $\mu + 3\gamma/2$ for each $j \in \{1, 2, \ldots, 2n-3\}$. Since γ is the largest possible change in $\mathsf{W}_{t+j-1}^{(j)}$ in either direction, it follows that $\mathsf{W}_{t+j}^{(j)}$ cannot lie between $\mu - \gamma/2$ and $\mu + \gamma/2$ unless $\mathsf{W}_{t+j-1}^{(j)}$ is between $\mu - 3\gamma/2$ and $\mu + 3\gamma/2$.

In the expression in (4.3), the integration of the α_j terms presents a daunting mathematical challenge due to the complexity of the likelihood function.

One way to deal with this problem is to derive a lower bound on the product of the α_j terms in such a way that the parts that are difficult to integrate are bounded by a constant value. To begin, expand the product in (4.3) in the following way:

$$\prod_{j=1}^{2n-3} \alpha_j(\mathsf{W}_{t+j-1}, \mathsf{W}_{t+j}) =$$
$$\prod_{j=1}^{2n-3} \min\left\{1, \frac{\mathcal{L}(\mathbf{D}|\mathsf{W}_{t+j})e^{-\frac{1}{2\sigma^2}(\mathsf{W}_{t+j}^{(j)}-\mu)^2}}{\mathcal{L}(\mathbf{D}|\mathsf{W}_{t+j-1})e^{-\frac{1}{2\sigma^2}(\mathsf{W}_{t+j-1}^{(j)}-\mu)^2}}\right\}. \tag{4.5}$$

Recall the form of the likelihood:

$$\mathcal{L}(\mathbf{D}|\mathbf{w}) = \left[\frac{1}{4}\sum_{s_1}\sum_{s_2} \Pr(s_2|s_1, e^{w^{(1)}})\mathcal{L}_{s_1}(1)\mathcal{L}_{s_2}(2)\right]^N.$$

The conditional likelihoods do not depend on $w^{(1)}$, and under the Jukes-Cantor Model, we have the following relation:

$$\frac{1}{4} - \frac{1}{4}e^{-\frac{4}{3}e^{w^{(1)}}} \leq \Pr(s_2|s_1, e^{w^{(1)}}) \leq \frac{1}{4} + \frac{3}{4}e^{-\frac{4}{3}e^{w^{(1)}}}.$$

Observe that, regardless of which log branch length is chosen for updating, the likelihood can be written as a sum of products of a transition probability, which depends on the chosen log branch length, and the conditional likelihoods at the nodes at each end of the chosen branch, which do not depend on the chosen log branch length. Therefore, if the jth log branch length is chosen to be incremented, the likelihood ratio is bounded below in the following way:

$$\frac{\mathcal{L}(\mathbf{D}|\mathsf{W}_{t+j})}{\mathcal{L}(\mathbf{D}|\mathsf{W}_{t+j-1})} \geq \left[\frac{\frac{1}{4} - \frac{1}{4}e^{-\frac{4}{3}e^{\mathsf{W}_{t+j}^{(j)}}}}{\frac{1}{4} + \frac{3}{4}e^{-\frac{4}{3}e^{\mathsf{W}_{t+j-1}^{(j)}}}}\right]^N. \tag{4.6}$$

The numerator in the quantity on the right hand side of (4.6) increases in $\mathsf{W}_{t+j}^{(j)}$, while the denominator decreases in $\mathsf{W}_{t+j-1}^{(j)}$. Therefore, to obtain a lower

bound on the likelihood ratio, it is necessary to find the smallest supported values of $\mathsf{W}_{t+j-1}^{(j)}$ and $\mathsf{W}_{t+j}^{(j)}$. Recall that if the log branch lengths are updated in order, then for each $j \in \{1, 2, \ldots, 2n-3\}$, $\mathsf{W}_{t+j-1}^{(j)} = \mathsf{W}_t^{(j)}$. Therefore, $\mathsf{W}_{t+j-1}^{(j)}$ is no smaller than $\mu - 3\gamma/2$ if $\mathsf{W}_t^{(j)}$ is in C. As a result, $\mathsf{W}_{t+j}^{(j)}$ can be no smaller than $\mu - 5\gamma/2$, so that

$$\frac{L(\mathbf{D}|\mathsf{W}_{t+j})}{\mathcal{L}(\mathbf{D}|\mathsf{W}_{t+j-1})} \geq \left[\frac{\frac{1}{4} - \frac{1}{4}e^{-\frac{4}{3}e^{\mu-5\gamma/2}}}{\frac{1}{4} + \frac{3}{4}e^{-\frac{4}{3}e^{\mu-3\gamma/2}}}\right]^N.$$

This result provides what is needed to complete the derivation of an upper bound on the expression in (4.5).

$$\begin{aligned}
&\prod_{j=1}^{2n-3} \alpha_j(\mathsf{W}_{t+j-1}, \mathsf{W}_{t+j}) \\
&\geq \prod_{j=1}^{2n-3} \min\left\{1, \left[\frac{\frac{1}{4} - \frac{1}{4}e^{-\frac{4}{3}e^{\mu-5\gamma/2}}}{\frac{1}{4} + \frac{3}{4}e^{-\frac{4}{3}e^{\mu-3\gamma/2}}}\right]^N \frac{e^{-\frac{1}{2\sigma^2}(\mathsf{W}_{t+j}^{(j)}-\mu)^2}}{e^{-\frac{1}{2\sigma^2}(\mathsf{W}_{t+j-1}^{(j)}-\mu)^2}}\right\}. \qquad (4.7)
\end{aligned}$$

Since the denominator in the second fraction in the product in (4.7) is smaller than 1, and since in the composition in (4.3) each log branch length is updated exactly once in the $2n-3$ steps, a lower bound on the product in (4.7) is

$$\begin{aligned}
&\left[\frac{\frac{1}{4} - \frac{1}{4}e^{-\frac{4}{3}e^{\mu-5\gamma/2}}}{\frac{1}{4} + \frac{3}{4}e^{-\frac{4}{3}e^{\mu-3\gamma/2}}}\right]^{(2n-3)N} \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{2n-3}(\mathsf{W}_{t+2n-3}^{(j)} - \mu)^2\right\} \\
&\times \prod_{j=1}^{2n-3} \mathbb{I}_{[\mu-\gamma/2,\mu+\gamma/2]}(\mathsf{W}_{t+2n-3}^{(j)}).
\end{aligned}$$

To complete verification of a minorization condition, it remains to bound the expression

$$\begin{aligned}
&\int_{\mathbb{R}^{2n-3}} \cdots \int_{\mathbb{R}^{2n-3}} \prod_{j=1}^{2n-3} \left(\prod_{\substack{i=1 \\ i\neq j}}^{2n-3} \delta_{\mathsf{W}_{t+j-1}^{(j)}}(\mathsf{W}_{t+j}^{(i)})\right) \\
&\qquad \times q_j(\mathsf{W}_{t+j}^{(j)}|\mathsf{W}_{t+j-1}^{(j)}) \mathrm{d}\mathsf{W}_{t+1} \ldots \mathrm{d}\mathsf{W}_{t+2n-4}. \qquad (4.8)
\end{aligned}$$

However, all of the Dirac masses integrate to 1, as do the q_j terms when $j < 2n-4$. Therefore, what remains after all of these things are integrated out is

$$\int_{\mathbb{R}^{2n-3}} \frac{1}{2\gamma} \mathbb{I}_{\left[\mathsf{W}^{(2n-3)}_{t+2n-4}-\gamma, \mathsf{W}^{(2n-3)}_{t+2n-4}+\gamma\right]}(\mathsf{W}^{(2n-3)}_{t+2n-3}) \mathrm{d}\mathsf{W}^{(2n-3)}_{t+2n-4}$$

To deal with this term, observe that

$$\mathbb{I}_{\left[\mathsf{W}^{(2n-3)}_{t+2n-4}-\gamma, \mathsf{W}^{(2n-3)}_{t+2n-4}+\gamma\right]}(\mathsf{W}^{(2n-3)}_{t+2n-3}) = \mathbb{I}_{\left[\mathsf{W}^{(2n-3)}_{t+2n-3}-\gamma, \mathsf{W}^{(2n-3)}_{t+2n-3}+\gamma\right]}(\mathsf{W}^{(2n-3)}_{t+2n-4}).$$

This result implies that the expression in (4.8) is equal to 1, and that a lower bound on the product of the alpha terms is

$$\left[\frac{\frac{1}{4} - \frac{1}{4}e^{-\frac{4}{3}e^{\mu-5\gamma/2}}}{\frac{1}{4} + \frac{3}{4}e^{-\frac{4}{3}e^{\mu-3\gamma/2}}}\right]^{(2n-3)N} \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{2n-3}(\mathsf{W}^{(j)}_{t+2n-3} - \mu)^2\right\}$$
$$\times \prod_{j=1}^{2n-3} \mathbb{I}_{[\mu-\gamma/2, \mu+\gamma/2]}(\mathsf{W}^{(j)}_{t+2n-3}).$$

One thing of note from the above work is that due to the dependence among the log branch lengths in the likelihood function, the transition kernel may change based on the order in which the log branch lengths are chosen. However, the methods used to obtain a lower bound provide a lower bound on the likelihood ratio that does not depend on the order in which the log branch lengths are updated. In addition, the prior and the increment densities do not exhibit dependence among the log branch lengths. Thus, provided each of the log branch lengths are updated in the $2n-3$ steps, the order in which they are updated is of no consequence to the derivation of a lower bound on the expression on the right-hand side of (4.1). Since the sum on the right hand side of (4.2) is over the set of all permutations of $\{1, 2, \ldots, 2n-3\}$, it follows

that there are $(2n-3)!$ summands. We normalize all of the factors of the lower bound obtained above to show that $(\mathsf{W}_t)_{t=0}^{\infty}$ satisfies a minorization condition over the set $C = [\mu - 3\gamma/2, \mu + 3\gamma/2]$, and

$$\begin{aligned}\epsilon &= \frac{(2n-3)!}{(2n-3)^{2n-3}} \left[\frac{\frac{1}{4} - \frac{1}{4}e^{-\frac{4}{3}e^{\mu-5\gamma/2}}}{\frac{1}{4} + \frac{3}{4}e^{-\frac{4}{3}e^{\mu-3\gamma/2}}}\right]^{(2n-3)N} \\ &\times \left[\left(\Phi\left(\frac{\gamma}{2\sigma}\right) - \Phi\left(\frac{-\gamma}{2\sigma}\right)\right)\sqrt{2\pi\sigma^2}\right]^{2n-3}.\end{aligned}$$

The minorizing density is the $N(\mu\mathbf{1}_{2n-3}, \sigma^2\mathbf{I}_{2n-3})$ density, where the support is truncated to $[\mu - \gamma/2, \mu + \gamma/2]^{2n-3}$.

Minorization via Monte Carlo Simulations

It is clear that as n and/or N increase, the value of ϵ decreases very quickly. This property results in an upper bound on the mixing time that increases rapidly in both n and N. In an effort to deal with this problem, we propose a Monte Carlo method to estimate a lower bound on ϵ. The goal is to obtain a tighter lower bound on the product

$$\prod_{j=1}^{2n-3} \min\left\{1, \frac{\mathcal{L}(\mathbf{D}|\mathsf{W}_{t+j})}{\mathcal{L}(\mathbf{D}|\mathsf{W}_{t+j-1})}\right\}.$$

The algorithm begins with the selection of M_1 initial values of W_t from inside the small set obtained in the previous section. From each starting point, run M_2 $(2n-3)$-step chains, and at step j, update the jth log branch length. For the jth step of the lth chain, compute

$$\min\left\{1, \frac{\mathcal{L}(\mathbf{D}|\mathsf{W}^l_{t+j})}{\mathcal{L}(\mathbf{D}|\mathsf{W}^l_{t+j-1})}\right\},$$

and then decide whether or not to accept the proposed value of W^l_{t+j} by way of the updating mechanism provided in the description of $(\mathsf{W}_t)_{t=0}^{\infty}$. For each of

the M_2 chains, calculate

$$\prod_{j=1}^{2n-3} \min\left\{1, \frac{\mathcal{L}(\mathbf{D}|\mathsf{W}_{t+j}^{l})}{\mathcal{L}(\mathbf{D}|\mathsf{W}_{t+j-1}^{l})}\right\},$$

and calculate the minimum value of these products over the M_2 chains. The minimum value over the M_2 chains provides an intermediate estimate of a lower bound on ϵ given the starting point from which each of the M_2 chains are initialized. The next step in this method is to obtain an estimated lower bound $\tilde{\epsilon}$ by finding the minimum value of the intermediate lower bounds over the M_1 initial values.

Note that since this method is intended to estimate a minimum, it is inherently biased so that it is larger than the true minimum. In an effort to mitigate this concern, we shrink $\tilde{\epsilon}$ by a factor of L, where L is determined by the variation in the intermediate estimates. This shrinking provides the final estimated lower bound $\hat{\epsilon}$ of ϵ:

$$\hat{\epsilon} = \frac{\tilde{\epsilon}}{L}.$$

4.2 A Drift Condition

Recall that for a Markov chain $(\mathsf{X}_t)_{t=0}^{\infty}$, if there exists a small set C, constants $\lambda < 1$ and $b < \infty$, and a function $V : \mathbb{R}^m \mapsto [1, \infty)$ such that

$$\mathbb{E}\left[V(\mathsf{X}_{t+1})|\mathsf{X}_t = \mathbf{x}\right] \leq \lambda V(\mathbf{x}) + b\mathbb{I}_C(\mathbf{x}),$$

then $(\mathsf{X}_t)_{t=0}^{\infty}$ satisfies a drift condition. Fort et al. (2003) shows that if $(\mathsf{X}_t)_{t=0}^{\infty}$ satisfies the three sufficient conditions for geometric ergodicity outlined in

Chapter 3, then the chain satisfies a drift condition with

$$V(\mathbf{x}) = [p(\mathbf{x})]^{-s},$$

where p is the target density and $s \in (0,1)$ is such that

$$s(1-s)^{\frac{1}{s}-1} < \frac{\xi}{m-2\xi},$$

where

$$\xi := \inf_{1 \le i \le 2n-3} \int_0^{\gamma} q_i(y)\lambda_1(\mathrm{d}y)$$

as in Theorem 7.

For our chain and choice of increment density, $\xi = 1/2$, so that $s(1-s)^{\frac{1}{s}-1} < \frac{1}{4n-8}$. Table 4.1 provides useful values of s for different numbers of leaves.

Table 4.1: Values of s for Varying Numbers of Leaves

Number of Leaves	s
3	< 0.5
4	< 0.2892
5	< 0.2030
6	< 0.1564
8	< 0.1071
10	< 0.0814
15	< 0.0509
20	< 0.0370
50	< 0.0140
100	< 0.0070
500	< 0.00014

Let $PV(\mathbf{x}) = \mathbb{E}\left[V(\mathsf{X}_{t+1})|\mathsf{X}_t = \mathbf{x}\right]$, and let $P_iV(\mathbf{x}) = \mathbb{E}\left[V(\mathsf{X}_{t+1})|\mathsf{X}_t = \mathbf{x}, i\right]$, the expected drift given that the ith log branch length is chosen to be updated.

Since the log branch length to be updated is chosen uniformly at random,

$$PV(\mathbf{x}) = \frac{1}{2n-3}\sum_{i=1}^{2n-3} P_i V(\mathbf{x}).$$

Let $\mathcal{A}(y,i) = \left\{y : \frac{p(\mathbf{x}+y\mathbf{e}_i)}{p(\mathbf{x})} \geq 1\right\}$, and let $\mathcal{R}(y,i) = \left\{y : \frac{p(\mathbf{x}+y\mathbf{e}_i)}{p(\mathbf{x})} < 1\right\}$. Partitioning the support of the increment densities into one region of certain acceptance and one region in which the proposal is potentially rejected allows the following representation of $P_i V(\mathbf{x})$:

$$\begin{aligned} P_i V(\mathbf{x}) &= \int_{\mathcal{A}(y,i)} V(\mathbf{x}+y\mathbf{e}_i) q_i(y) \mathrm{d}y \\ &+ \int_{\mathcal{R}(y,i)} V(\mathbf{x}+y\mathbf{e}_i) \frac{p(\mathbf{x}+y\mathbf{e}_i)}{p(\mathbf{x})} q_i(y) \mathrm{d}y \\ &+ V(\mathbf{x}) \int_{\mathcal{R}(y,i)} \left(1 - \frac{p(\mathbf{x}+y\mathbf{e}_i)}{p(\mathbf{x})}\right) q_i(y) \mathrm{d}y. \end{aligned} \tag{4.9}$$

Our method of establishing a drift condition for $(\mathsf{W}_t)_{t=0}^{\infty}$ is based on the representation in (4.9). When (4.9) and the drift function provided by Fort et al. (2003) are applied to $(\mathsf{W}_t)_{t=0}^{\infty}$, we get

$$\begin{aligned} P_i V(\mathbf{w}) &= \frac{1}{2\gamma} \int_{\mathcal{A}(y,i)} V(\mathbf{w}+y\mathbf{e}_i) \mathrm{d}y \\ &+ \frac{1}{2\gamma} \int_{\mathcal{R}(y,i)} V(\mathbf{w}+y\mathbf{e}_i) \frac{p(\mathbf{w}+y\mathbf{e}_i|\mathbf{D})}{p(\mathbf{w}|\mathbf{D})} \mathrm{d}y \\ &+ \frac{1}{2\gamma} V(\mathbf{w}) \int_{\mathcal{R}(y,i)} \left(1 - \frac{p(\mathbf{w}+y\mathbf{e}_i|\mathbf{D})}{p(\mathbf{w}|\mathbf{D})}\right) \mathrm{d}y. \end{aligned}$$

Due to the complexity of the likelihood function, this integral is mathematically intractable. Because of this, we present a Monte Carlo integration method to estimate λ.

Since for $\mathbf{w} \notin C$,

$$\frac{PV(\mathbf{w})}{V(\mathbf{w})} \leq \lambda,$$

our method of estimating λ begins with the selection of M_3 starting points $\mathbf{w}^l$ that are outside C, $l = 1, 2, \ldots, M_3$. Divide $[-\gamma, \gamma]$ into intervals of width v. Given $\mathbf{w}^l$, for each $i \in \{1, 2, \ldots, 2n-3\}$, find $\mathcal{A}(y, i)$ and $\mathcal{R}(y, i)$ as follows.

For $j \in \{0, 1, \ldots, 2\gamma/v\}$, let $y_j = -\gamma + vj$, and calculate the acceptance probability $\alpha = \min\left\{1, \frac{p(\mathbf{w}^l + y_j\mathbf{e}_i|\mathbf{D})}{p(\mathbf{w}^l|\mathbf{D})}\right\}$. If $\alpha = 1$, then $y_j \in \mathcal{A}(y, i)$. Otherwise, $y_j \in \mathcal{R}(y, i)$. Next, for each starting point, calculate

$$\begin{aligned}
\frac{\hat{P}_i V(\mathbf{w}^l)}{V(\mathbf{w}^l)} &= \frac{1}{2\gamma} v \sum_{\{j: y_j \in \mathcal{A}(y,i)\}} \frac{V(\mathbf{w}^l + y_j\mathbf{e}_i)}{V(\mathbf{w}^l)} \\
&+ \frac{1}{2\gamma} v \sum_{\{j: y_j \in \mathcal{R}(y,i)\}} \frac{V(\mathbf{w}^l + y_j\mathbf{e}_i)}{V(\mathbf{w}^l)} \frac{p(\mathbf{w}^l + y_j\mathbf{e}_i|\mathbf{D})}{p(\mathbf{w}^l|\mathbf{D})} \\
&+ \frac{1}{2\gamma} v \sum_{\{j: y_j \in \mathcal{R}(y,i)\}} \left(1 - \frac{p(\mathbf{w}^l + y_j\mathbf{e}_i|\mathbf{D})}{p(\mathbf{w}^l|\mathbf{D})}\right).
\end{aligned}$$

Once the expected values of the drift given the chosen component of $\mathbf{w}^l$ are available, we calculate the average value of these conditional expectations to obtain an estimated value of λ based on $\mathbf{w}^l$:

$$\frac{\hat{P}V(\mathbf{w}^l)}{V(\mathbf{w}^l)} = \frac{1}{2n-3} \sum_{i=1}^{2n-3} \frac{\hat{P}_i V(\mathbf{w}^l)}{V(\mathbf{w}^l)}.$$

Once this process is complete for all $l \in \{1, 2, \ldots, M_3\}$, we calculate the standard error $s_{\hat{\lambda}}$ of the $\frac{\hat{P}V(\mathbf{w}^l)}{V(\mathbf{w}^l)}$. Our estimate of λ is

$$\hat{\lambda} = \max_{l \in \{1,2,\ldots,M_3\}} \frac{\hat{P}V(\mathbf{w}^l)}{V(\mathbf{w}^l)} + s_{\hat{\lambda}}.$$

We add the standard error because estimation of a maximum inherently provides estimates that are biased low. The standard error is added in an effort

to be conservative in our estimation of the maximum.

In the estimation of λ, we have no need to deal with the marginal probability $m(\mathbf{D})$ of the data set. In the estimation of b, this is not the case. While the marginal probability of the data set is difficult to obtain, it is possible obtain an upper bound on $m(\mathbf{D})$. This upper bound will be used to obtain a Monte Carlo upper bound on b. In order to bound $m(\mathbf{D})$, we view the joint probability density $p(\mathbf{w}, \mathbf{D})$ as the expected likelihood with respect to the prior density:

$$\begin{aligned} m(\mathbf{D}) &= \int_{\mathbb{R}^{2n-3}} \mathcal{L}(\mathbf{D}|\mathbf{w})\phi(\mathbf{w})\mathrm{d}\mathbf{w} \\ &= \mathbb{E}_{\phi}\left[\mathcal{L}(\mathbf{D}|\mathbf{w})\right]. \end{aligned}$$

This representation of $m(\mathbf{D})$ makes it clear that an upper bound on $m(\mathbf{D})$ is given by $\sup_{\mathbf{w}\in\mathbb{R}^m} \mathcal{L}(\mathbf{D}|\mathbf{w})$, and we use this as an upper bound on $m(\mathbf{D})$ in the estimation of b.

We are now ready to present the Monte Carlo method of bounding b. To begin, choose M_5 starting points inside C. For each starting point $\mathbf{w}^l, l = 1, 2, \ldots, M_5$, and for each $i \in \{1, 2, \ldots 2n-3\}$, we estimate an upper bound on $\hat{P}_i V(\mathbf{w}^l)$ in the following way.

Let $\hat{p}(\mathbf{w}|\mathbf{D})$ be defined as follows:

$$\hat{p}(\mathbf{w}|\mathbf{D}) := \frac{\phi(\mathbf{w})\mathcal{L}(\mathbf{D}|\mathbf{w})}{\sup_{\mathbf{w}\in\mathbb{R}^m} \mathcal{L}(\mathbf{D}|\mathbf{w})}.$$

For each starting point, the acceptance and the potential rejection regions are obtained in the same way as prescribed in the estimation of λ. Then, letting

$\hat{V}(\mathbf{w}) = [\hat{p}(\mathbf{w}|\mathbf{D})]^{-s}$, we calculate

$$\begin{aligned}
\hat{P}_i\hat{V}(\mathbf{w}^l) &= \frac{1}{2\gamma}v \sum_{\{j:y_j\in\mathcal{A}(y,i)\}} \hat{V}(\mathbf{w}^l + y_j\mathbf{e}_i) \\
&+ \frac{1}{2\gamma}v \sum_{\{j:y_j\in\mathcal{R}(y,i)\}} \hat{V}(\mathbf{w}^l + y_j\mathbf{e}_i)\frac{p(\mathbf{w}^l + y_j\mathbf{e}_i|\mathbf{D})}{p(\mathbf{w}^l|\mathbf{D})} \\
&+ \frac{1}{2\gamma}v \sum_{\{j:y_j\in\mathcal{R}(y,i)\}} \hat{V}(\mathbf{w}^l)\left(1 - \frac{p(\mathbf{w}^l + y_j\mathbf{e}_i|\mathbf{D})}{p(\mathbf{w}^l|\mathbf{D})}\right).
\end{aligned}$$

For each $l \in \{1, 2, \ldots, M_5\}$, calculate $\hat{P}\hat{V}(\mathbf{w}^l)$ in a manner similar to that used in the estimation of λ, so that

$$\hat{P}\hat{V}(\mathbf{w}^l) = \frac{1}{2n-3}\sum_{i=1}^{2n-3} \hat{P}_i\hat{V}(\mathbf{w}^l).$$

Next, calculate an intermediate upper bound $\hat{b}_l$ on b for each starting point as follows:

$$\hat{b}_l = \hat{P}_i\hat{V}(\mathbf{w}^l) - \hat{\lambda}\hat{V}(\mathbf{w}^l),$$

and then calculate the standard error $s_{\hat{b}}$ of the $\hat{b}_l$. The estimation of an upper bound on b is completed by obtaining the maximum of the $\hat{b}_l$ values and then adding $s_{\hat{b}}$ to this maximum, so that

$$\hat{b} = \max_{l\in\{1,2,\ldots,M_5\}} \hat{b}_l + s_{\hat{b}}.$$

Since this process is estimating an upper bound, the standard deviation of the $\hat{b}_l$ is added in an effort to obtain a conservative upper bound.

The procedures above provide Monte Carlo estimates of λ and a Monte Carlo upper bound on b that, along with the Monte Carlo lower bound on ϵ, can be useful in obtaining an upper bound on the mixing time of the RSM

algorithm. We will provide an illustrative example of these methods, and what we will observe is that the estimated lower bound on ϵ and the estimated upper bound on b are not very helpful in providing a useful upper bound on the mixing time of the Markov chain associated with the version of the RSM algorithm that we describe. However, the ability to obtain these values represents a step toward finding an honest upper bound on the mixing time of the RSM algorithm for Bayesian estimation of the branch lengths of a phylogenetic tree. This brings us closer to no longer having to rely on ad-hoc convergence assessment methods, some of which are outlined in the next section.

4.3 Output-Based Methods of Convergence Assessment

In this section, we describe four methods of convergence assessment, each of which is based on output from the RSM algorithm. The first, and the simplest, of these methods involves examining trace plots, which are plots of the parameter value at each iteration against the iteration numbers, and the acceptance rate of the chain. The second, proposed by Yu and Mykland (1998), involves looking at a different type of plot. We also present a modification of the method of Yu and Mykland (1998) that was proposed by Brooks (1996) and relies on the calculation of a statistic used to measure the "hairiness" of the plot. The third method (Geweke, 1992) is based on a hypothesis test for a difference in the mean value of some scalar function of the parameter values over different regions of the sequence. The final method (Gelman and Rubin, 1992) is based on the output of several independent chains, and the

procedure depends on estimation of the factor by which the posterior density is expected to shrink as the number of steps becomes large. We close this section by describing the drawbacks of output-based convergence assessment. For a survey of output-based convergence diagnostics, see Cowles and Carlin (1996) or Brooks and Roberts (1997).

4.3.1 Trace Plots and Acceptance Rates

The simplest output-based method of convergence assessment is one that involves looking at trace plots of the individual parameters, or of some function $\theta : \mathbb{R}^m \mapsto \mathbb{R}$ of the parameters. If the chain is mixing well, we should see regular oscillation around a central value. In other words, we should see "hairiness" in the plot, and this "hairiness" should be centered around a particular value. The next four figures show examples of trace plots for chains that exhibit varying degrees of mixing behavior.

We must also consider the acceptance rate of an MCMC algorithm in our evaluation of mixing behavior. In our RSM algorithm, we have a univariate proposal density. Roberts and Rosenthal (2001) show that the optimal acceptance rate for any type of Metropolis-Hastings algorithm with a univariate proposal density is 0.44. By optimal, we mean in the sense of mixing behavior. We need a proposal density that proposes moves that are not so large that they are rarely accepted and thus result in slow exploration of the state space, but not so small as to be accepted too often. The frequent acceptance of the small moves results in slow exploration of the state space. The proposal density must avoid these two extremes, and a typical way to do it is to choose

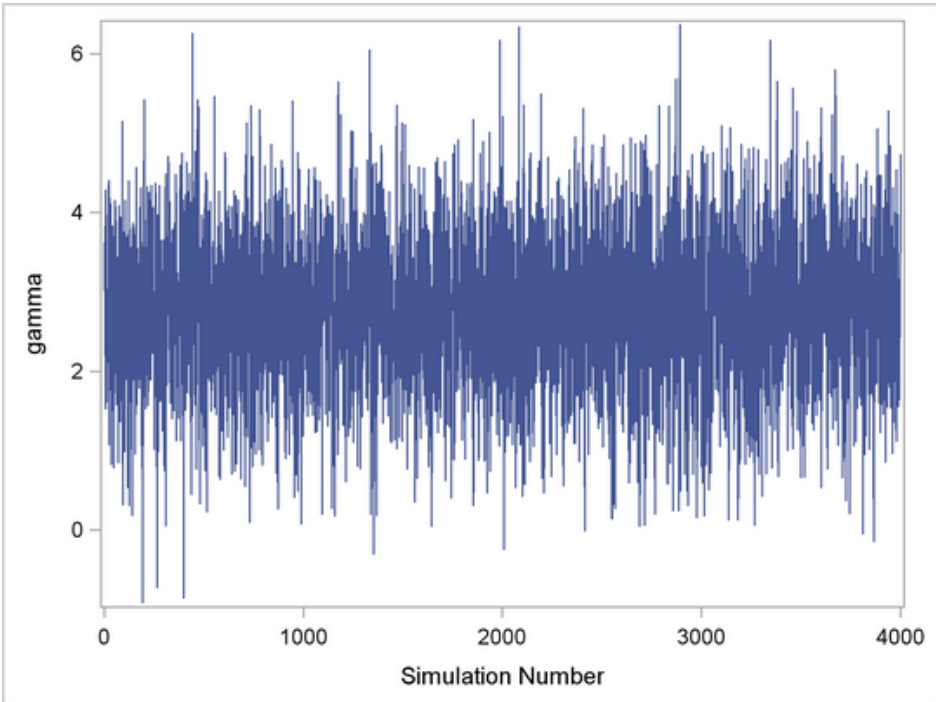


Figure 4.1: A trace plot for a Markov chain that exhibits good mixing. The chain appears to begin roughly in its target distribution. The plot shows regular oscillation around 3. Though the chain is approximately stationary, it should still be thinned in order to obtain roughly independent samples from the target distribution.

a proposal density based on the percentage of the time proposals from it are accepted. This procedure is known as *optimal scaling*. As the dimensionality of the proposal density gets large, the optimal acceptance rate tends to 0.234 (Roberts and Rosenthal, 2001).

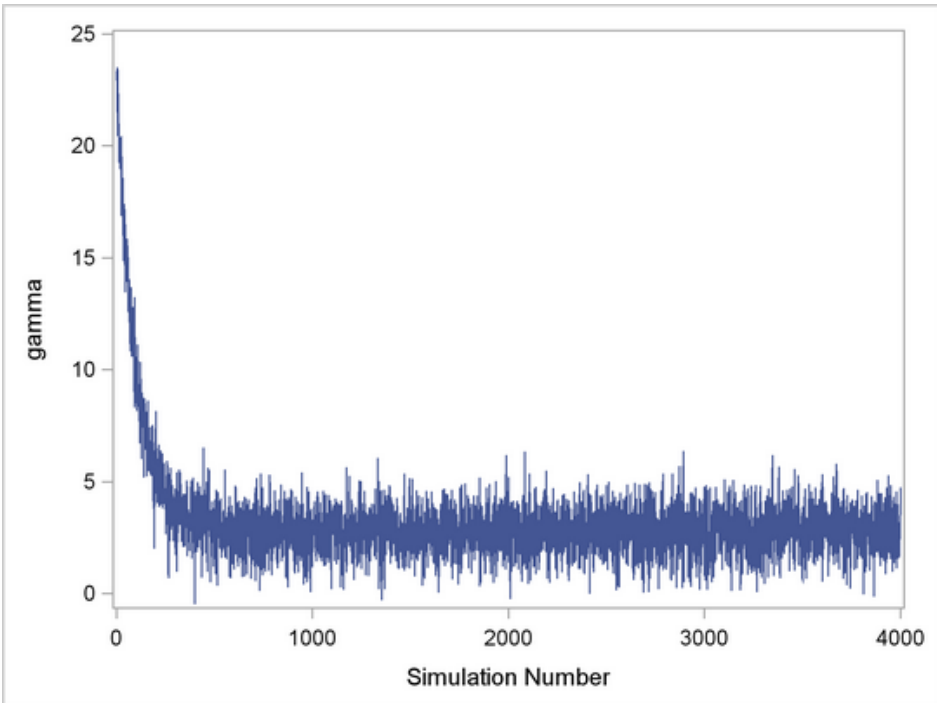


Figure 4.2: A trace plot for a chain that appears to reach its target distribution in approximately 500 steps. These initial output values should be discarded. After the first 500 steps, the plot shows regular oscillation around 3.

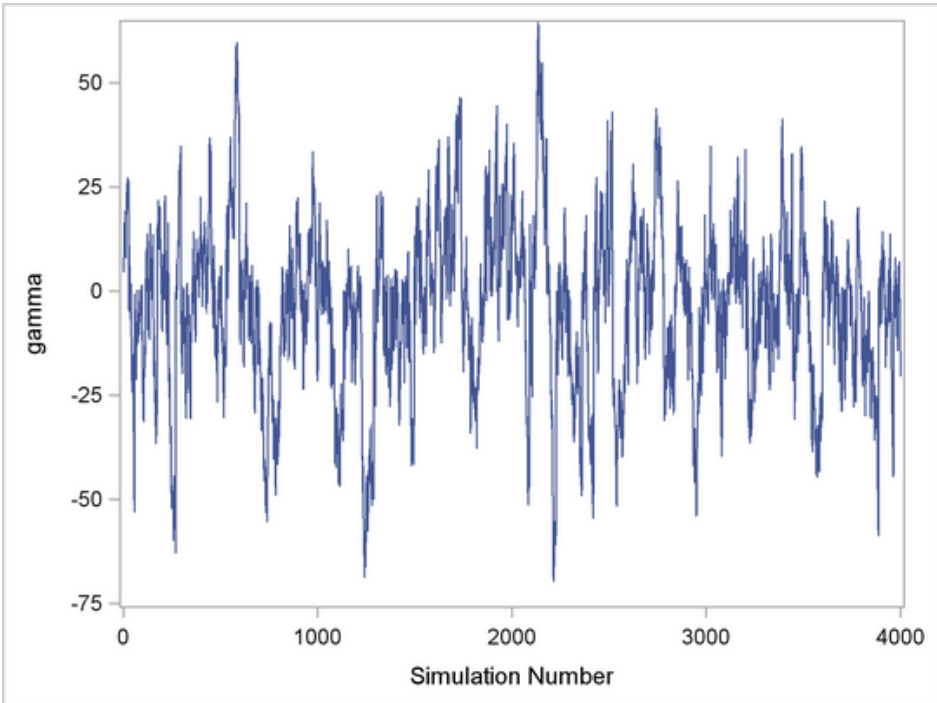


Figure 4.3: A trace plot for a chain that appears to take small steps, so that it does not explore the target distribution quickly. This is an indication of high correlation among the samples, so in order to obtain independent samples the chain must be run for a larger number of steps in order to accommodate thinning the output by a larger factor.

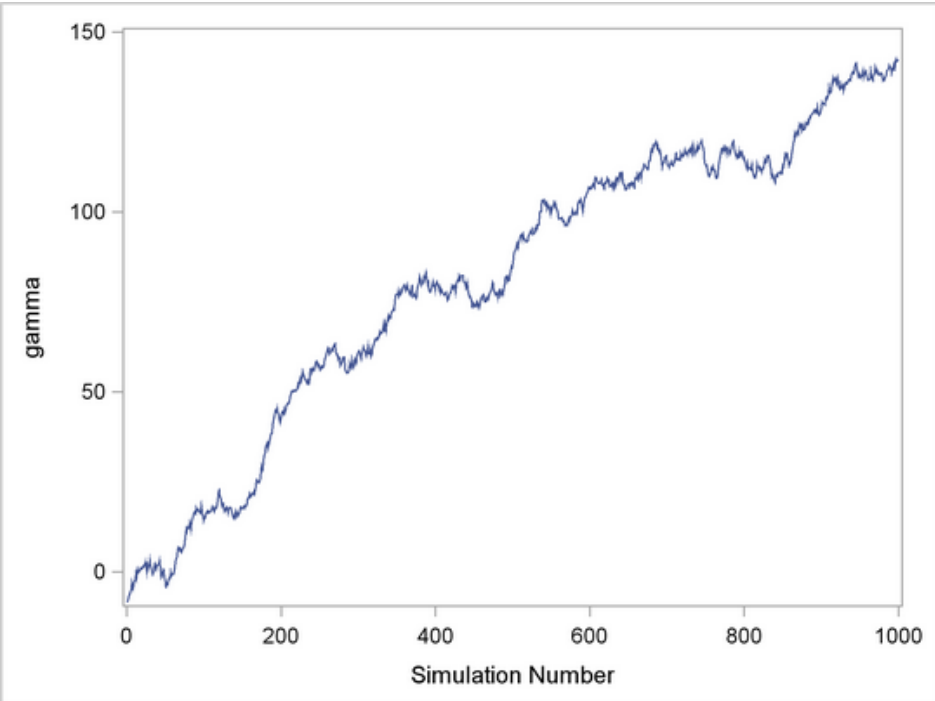


Figure 4.4: A trace plot that indicates a chain that is not mixing well. The chain is exploring the target distribution extremely slowly. This chain is not suitable for parameter inference.

4.3.2 Yu and Mykland's CUSUM Plot

The CUSUM plot is a way of monitoring convergence that can be applied to any sampler. It is presented by Yu and Mykland (1998), and its primary appeal is that it can be implemented through generic code that is independent of the problem. For a sampler $(\mathsf{X}_t)_{t=0}^{\infty}$, we have output $\{\mathbf{x}^1, \ldots, \mathbf{x}^n\}$ after the chain has run for n steps. The initial iterations, say the first n_0 of them, are discarded. Let $\theta : \mathbb{R}^m \mapsto \mathbb{R}$ be a scalar function of the parameters. Some common functions of this type are $\theta(\mathbf{x}) = \bar{\mathbf{x}}$ and $\theta(\mathbf{x}) = x^{(i)}$, $i \in \{1, 2, \ldots, m\}$. The construction of the CUSUM plot proceeds in the following way. First, calculate

$$\hat{\mu} = \frac{1}{n - n_0} \sum_{t=n_0+1}^{n} \theta(\mathbf{x}^t),$$

the mean of the values of θ that are not discarded. Next, calculate the partial sum, termed the CUSUM or cumulative sum

$$\hat{S}_T = \sum_{t=n_0+1}^{T} \left[\theta(\mathbf{x}^t) - \hat{\mu}\right], \text{ for each } T = n_0 + 1 \ldots, n.$$

Note that this is not the same as the CUSUM one would come across in the quality control literature (Bissell, 1969). The process ends by plotting $\left\{\hat{S}_T\right\}$ against T for each $T \in \{n_0 + 1, \ldots, n\}$, and connecting successive points with line segments.

In the CUSUM plot, poor mixing is indicated by smoothness. If a chain is mixing well, we will see a lot of "hairiness" in the plot, but the plot will not necessarily show oscillation around a central value, as we see in the trace plot. This is a highly subjective way to assess convergence, and Brooks (1996)

presents a more objective method that relies on the CUSUM. The idea is that a perfectly hairy plot will consist of line segments with alternating positive and negative slopes. Brooks (1996) suggests the following as a measure of hairiness. Define

$$d_T = \begin{cases} 1 & \text{if } S_{T-1} < S_T \text{ and } S_{T+1} < S_T \text{ or } S_{T-1} > S_T \text{ and } S_{T+1} > S_T \\ 0 & \text{otherwise,} \end{cases}$$

for all $T = n_0 + 1, \ldots, n - 1$. Then

$$D_{n_0,n} = \frac{1}{n - n_0} \sum_{T=n_0+1}^{n-1} d_T$$

is a measure of the "hairiness" of the plot. Brooks (1996) shows that $D_{n_0,n}$ has an asymptotic normal distribution with mean 1/2 and variance $1/[4(n - n_0)]$. A lack of convergence is indicated if $D_{n_0,n}$ falls outside of the bounds

$$\frac{1}{2} \pm Z_{\alpha/2} \sqrt{\frac{1}{4(n - n_0)}}.$$

Note that if a large number of samples are taken after the first n_0 are discarded, this interval is very small, and this method will frequently indicate a lack of convergence.

4.3.3 Geweke's Spectral Density Diagnostic

The spectral density diagnostic of Geweke (1992) relies on a hypothesis test for a difference in the mean values of a scalar function $\theta : \mathbb{R}^m \mapsto \mathbb{R}$ over different regions of the sequence of values from $(\mathsf{X}_t)_{t=0}^{\infty}$ that are not discarded. Let $\theta^t = \theta(\mathbf{x}^{t+n_0})$, where n_0 is the number of initial iterations that are discarded. Consider two subsequences $\{\theta^t : t = 1, \ldots, n_A\}$ and $\{\theta^t : t = n^*, \ldots, n\}$, where

$1 < n_A < n^* < n$ and $n_B = n - n^* + 1$. Next, define

$$\bar{\theta}_A = \frac{1}{n_A} \sum_{t=1}^{n_A} \theta^t \text{ and } \bar{\theta}_B = \frac{1}{n_B} \sum_{t=n^*}^{n} \theta^t,$$

the means of the θ^t values in each of the two regions of the sequence of $n - n_0$ steps of the chain that are not discarded. In addition, let $\hat{S}_\theta^A(0)$ and $\hat{S}_\theta^B(0)$ be estimates of the spectral density at frequency 0 for $\{\theta^t : t = 1, \ldots, n_A\}$ and $\{\theta^t : t = n^*, \ldots, n_B\}$. Geweke (1992) argues that if the ratio $\frac{n_A + n_B}{n} < 1$ is fixed and if the sequence $\{\theta^t\}$ is stationary, then

$$Z_n = \frac{\bar{\theta}_A - \bar{\theta}_B}{\sqrt{\frac{1}{n_A}\hat{S}_\theta^A(0) + \frac{1}{n_B}\hat{S}_\theta^B(0)}} \xrightarrow{D} N(0, 1),$$

as $n \to \infty$. Therefore, if the sampler is approximately in its stationary distribution, we will not reject the null hypothesis that there is no difference between the means of the θ values in the two regions of the chain.

4.3.4 Gelman and Rubin Potential Scale Reduction Factor

Gelman and Rubin (1992) propose a method of convergence assessment that relies on analyzing m independent chains. The process provides the basis for an estimate of how close the chain is to the target distribution as well as an estimate of how much we expect this estimate to improve with further simulations. For integers $m \geq 2$ and $n > 0$, independently simulate $m \geq 2$ sequences of length $2n$, each of which are started at different initial values that are over-dispersed with respect to the stationary distribution. For a scalar

summary statistic $\theta(\cdot)$, we calculate

$$B = \frac{1}{m-1}\sum_{i=1}^{m}(\bar{\theta}_{i\cdot} - \bar{\theta}_{\cdot\cdot})^2,$$

where $\bar{\theta}_{i\cdot}$ is the mean value of θ in the ith sequence and $\bar{\theta}_{\cdot\cdot}$ is the mean of the m within-sequence averages. Next, calculate the mean W of the within-sequence variances s_i^2, and then calculate $\hat{V}$, the estimated variance of the stationary distribution, as a weighted average of W and B:

$$\hat{V} = \frac{n-1}{n}W + \frac{1}{n}B.$$

The diagnostic relies on the calculation of a potential scale reduction factor (PSRF), which is a measure of the scale by which the posterior distribution of θ will shrink as n gets large. This factor is given by

$$\hat{R} = \frac{\hat{V}}{W}\ . \tag{4.10}$$

A value of $\hat{R}$ that is far from 1 is an indication of a lack of convergence.

4.3.5 Caveats of Output-Based Convergence Assessment

Despite the benefits of simplicity and ease of implementation of the ad hoc methods described above, such convergence assessment methods suffer from several drawbacks. Perhaps the most important of these is that they cannot determine that a chain is close to its stationary distribution. Instead, they can only provide evidence of a lack of convergence. In addition, these methods may require the researcher to run an MCMC algorithm for a long time just to determine how many of the first iterations should be discarded. Some of the

methods described above, such as the method of observing trace plots or the CUSUM plot method, are very subjective. They require only visual inspection of plots. These caveats of output-based convergence assessment highlight the need for honest (Jones and Hobert, 2001) upper bounds on the mixing time, where the word "honest" is taken to mean that an upper bound is developed in a mathematically rigorous manner that is not based on the output of the chain. The ability to obtain such upper bounds through verification of a drift and a minorization condition provides a sense of how long to run the chain before actually running it. Since this method relies on the Markov chain theory, there is no inherent subjectivity, thus addressing the concerns described above.

In some cases, however, it may be impossible to verify a drift and/or a minorization condition. In other cases, the drift and minorization conditions may lead to an upper bound on the mixing time, but this bound may not be very useful. In these cases, the output-based convergence assessments are the only way to obtain helpful information about the convergence behavior of our chain. In cases where honest upper bounds are both available and useful, the output-based diagnostics can be used to provide a check on the work in deriving an honest upper bound on the mixing time.

4.4 Illustrative Example

In this section, we provide an example that illustrates everything we discussed earlier in this chapter. We begin with a description of the specific RSM algorithm, including the tree topology, the data set, and the prior and proposal

densities. We then detail the establishment of the minorization condition, first by way of the analytical method and then via the Monte Carlo method of bounding ϵ. We also present the results of the Monte Carlo algorithms for estimating λ and bounding b, the parameters from the drift condition, and then we provide the output from the convergence diagnostics described previously for different summary statistics.

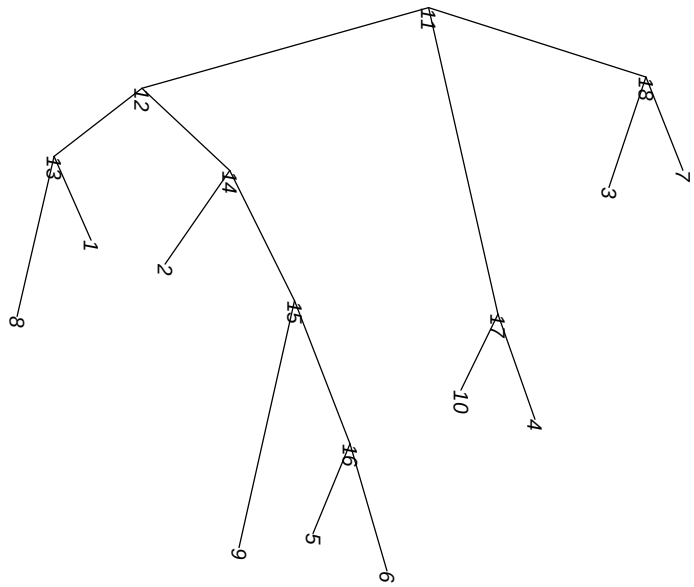


Figure 4.5: The unrooted tree topology used in this example. The tips are labelled 1 through 10, and the internal nodes are labelled 11 through 18. In this example, we look at branch 1, which connects nodes 11 and 12, branch 6, which connects nodes 2 and 14, and branch 16, which connects nodes 3 and 18.

In this example, we use an unrooted tree topology with 10 leaves. This tree topology is pictured in Figure 4.5. Therefore, our chain $(\mathsf{W}_t)_{t=0}^{\infty}$ moves on $\mathbb{R}^{2n-3} = \mathbb{R}^{17}$. We use a fictitious set of DNA sequence data that has 10,000 sites per sequence, 83.03% of which are constant. This is representative of a typical data set, as in a normal set of DNA sequence data, between 60% and 90% of the

sites are constant. To generate the data set, we first generate a set of log branch lengths $\mathbf{w}$ from the prior density, which is the $N_{17}(-4.5\mathbf{1}_{17}, 0.25\mathbf{I}_{17})$ density. This corresponds to a distribution of the branch length where between each node of the tree, we expect 0.0126 substitutions per site on average, and the expected number of substitutions per site has standard deviation 0.0063. The increment density is the $U[-0.33, 0.33]$ density, and we chose this increment density because it gives an acceptance rate near 44%.

4.4.1 The Minorization Condition

Using the analytical method, we find that

$$\begin{aligned} \epsilon &\geq \left(\frac{1 - e^{-\frac{4}{3}e^{-5.325}}}{1 + 3e^{-\frac{4}{3}e^{-4.995}}} \right)^{170,000} \\ &= 3.839998 \times 10^{-474,026}. \end{aligned} \tag{4.11}$$

The minorization condition is satisfied on $C = [-4.995, -4.005]^{17}$, with the $N_{17}(-4.5\mathbf{1}_{17}, 0.25\mathbf{I}_{17})$ density, truncated to the set $[-4.665, -4.335]^{17}$, as the density admitted by the minorizing measure. The lower bound on ϵ is very small, so we estimate a lower bound on ϵ using the Monte Carlo method.

To implement the Monte Carlo method, choose $5,000$ starting points inside C. From each of these $5,000$ starting points, run $1,000$ 17-step chains. The choices of $5,000$ and $1,000$ were made after experimentation with different values and the discovery that larger numbers of starting points and larger numbers of chains from each starting point gave only a small change in the estimate of ϵ, while greatly increasing the computational cost. Proceeding as described above, we get a Monte Carlo lower bound $\hat{\epsilon}$ equal to 8.09×10^{-284}.

This algorithm was repeated 50 times in order to obtain an idea of a suitable value of L. Once all the runs of the algorithm were complete, the smallest of the 50 values of $\hat{\epsilon}$ was 2.73×10^{-287}. This led to the conclusion that $L = 10^{10}$ is sufficiently conservative, and we obtain lower bound of $\hat{\epsilon} = 8.09 \times 10^{-294}$.

4.4.2 The Drift Condition

To formulate the estimates of the drift parameters, we use the drift function

$$V(\mathbf{w}) = [p(\mathbf{w}|\mathbf{D})]^{-0.05}. \tag{4.12}$$

To estimate λ, begin by choosing $20,000$ points $\mathbf{w}^k, k = 1, \ldots, 20,000$ outside of C. For each of these points, numerically calculate $PV(\mathbf{w}^k)/V(\mathbf{w}^k)$. In the numerical integration, we use interval width $v = 0.02$. Once $PV(\mathbf{w}^k)/V(\mathbf{w}^k)$ is obtained for each of the chosen points, we maximize over them and add the standard error. This process yields $\hat{\lambda} = 0.990777$ as an estimate of λ.

Let $\hat{V}(\mathbf{w})$ be defined as

$$[\hat{p}(\mathbf{w}|\mathbf{D})]^{-0.05}.$$

For the estimation of b, choose $5,000$ points $\mathbf{w}^l, l = 1, \ldots, 5,000$ inside C, and from each point, we numerically calculate $\hat{P}\hat{V}(\mathbf{w}^l) - \hat{\lambda}\hat{V}(\mathbf{w}^l)$ as described above for each $\mathbf{w}^l, l = 1, \ldots, 5000$. We maximize over the selected points and add a standard error to obtain an estimated upper bound on b of $\hat{b} = 3.6443 \times 10^{40}$.

4.4.3 Results of the Output-Based Methods of Convergence Assessment

We first present the trace plots for four summary statistics: $w^{(1)}$, $w^{(6)}$, $w^{(16)}$, and $\bar{\mathbf{w}}$, the mean log branch length. We then present the CUSUM plots for the

same four summary statistics. Finally, we present the Brooks (1996) CUSUM diagnostic, the Geweke (1992) diagnostic, and the Gelman and Rubin (1992) diagnostic results for each of the four summary statistics, as well as a brief discussion. For each diagnostic except for the trace plots, the first $170,000$ iterations are discarded, and the output is thinned by a factor of 1,000, which means that in the trace plot, only every 1,000th observation is plotted. This is done because the RSM algorithm updates the log branch lengths one-at-a-time, and each time the log branch length that is chosen to be updated is chosen uniformly at random. This means that it is possible for a log branch length to go a long time without being updated. Observing only every $1,000^{th}$ step not only takes care of this concern, but also reduces autocorrelation among the samples.

Trace Plots and Acceptance Rate

Figure 4.6 shows the trace plots for the four summary statistics listed above over the first 1.7 million steps of the chain. The chain has been thinned by a factor of $1,000$ because of the one-at-a-time updating scheme and the dependence among the log branch lengths. We observe that the plots indicate that the chain reaches its target distribution rather quickly. In each plot, there is rapid oscillation around a central value, and this does not indicate any problems with convergence. The acceptance rate is near the optimal rate, at 44.3%.

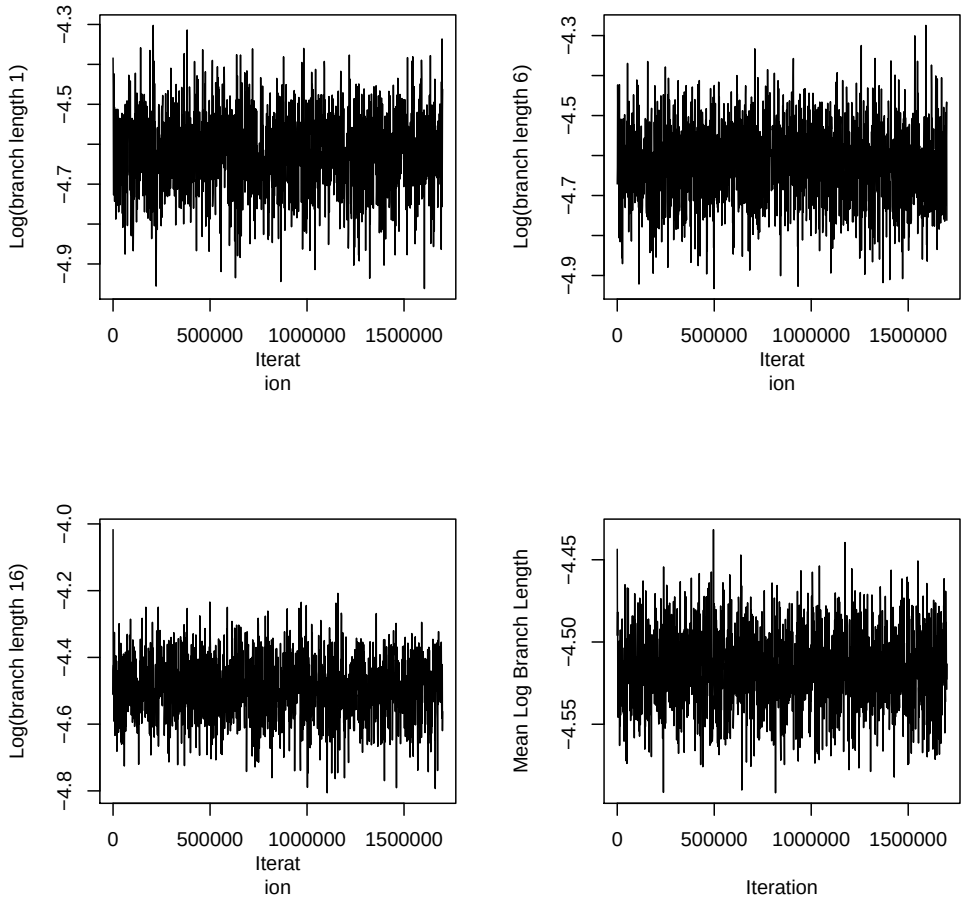


Figure 4.6: Trace plots of different summary statistics over the first 1.7 million steps of the chain. The summary statistics are $w^{(1)}$, $w^{(6)}$, $w^{(16)}$, and $\bar{\mathbf{w}}$. The trace plots do not indicate any problems with convergence, as in each of them, we see regular oscillation around a center value.

CUSUM Plots

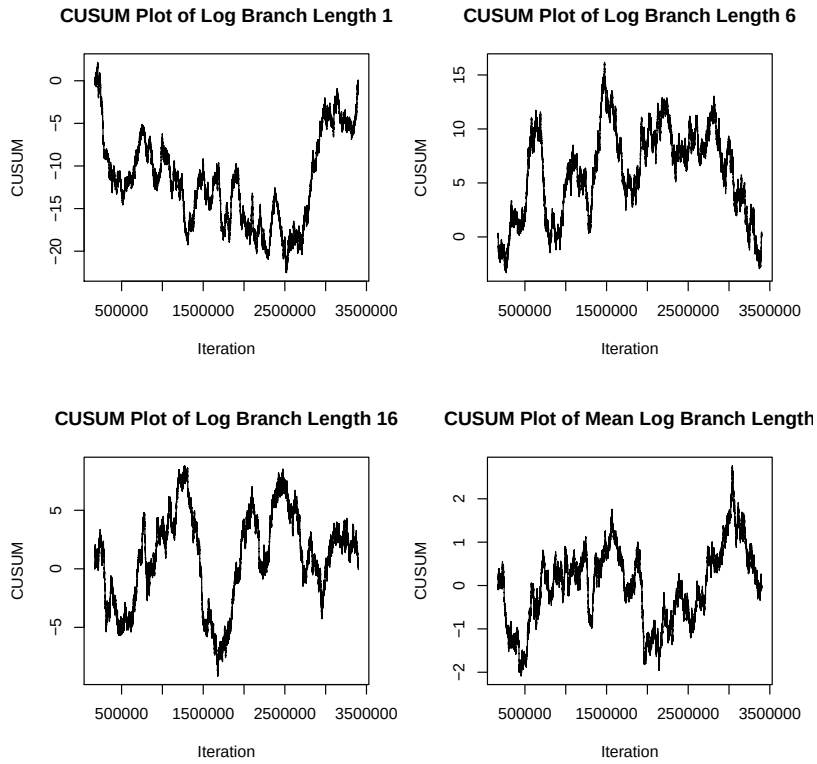


Figure 4.7: CUSUM plots for the summary statistics $w^{(1)}$, $w^{(6)}$, $w^{(16)}$, and $\bar{\mathbf{w}}$. The first 3.4 million steps of the chain are shown, with the first 170,000 steps discarded. The CUSUM plots show a lot of oscillation, and this does not indicate any problems with convergence.

Figure 4.7 shows the CUSUM plots for $w^{(1)}$, $w^{(6)}$, $w^{(16)}$, and $\bar{\mathbf{w}}$. The chain is run for 3.4 million steps, and the first 170,000 steps are discarded. The CUSUM plots show a lot of hairiness, and this is not indicative of any concerns with convergence.

Brooks, Geweke, and Gelman and Rubin Diagnostics

Table 4.2: Results of Brooks, Geweke, and Gelman and Rubin Diagnostics

Summary Statistic	Brooks CUSUM	Geweke Z_n	Gelman/Rubin PSRF
$w^{(1)}$	0.46183	-1.50531	1.00133
$w^{(6)}$	0.46025	0.211402	1.000239
$w^{(16)}$	0.45789	-0.46622	1.00129
$\bar{\mathbf{w}}$	0.47585	-0.06847	1.00026

Table 4.2 shows the results of the Brooks (1996) CUSUM diagnostic, the Geweke (1992) spectral density diagnostic, and the Gelman and Rubin (1992) diagnostic. Each of the three diagnostics comes from running the chain an additional 170,000 steps after the burn-in. For the Brooks (1996) diagnostic, all the values are between 0.45789 and 0.47585. These are reasonably close to 1/2, but by the Brooks (1996) standard, they indicate a lack of convergence. The bounds on the $D_{n_0,n}$ values for any summary statistic are 0.4976 and 0.5024, based on a 95% confidence interval. If the value of $D_{n_0,n}$ is between these values, no lack of convergence is indicated. This highlights a major concern with this convergence diagnostic. The chain may have, in fact, run for an adequately large number of steps to allow approximate sampling from the target density. However, if the number of steps the chain runs after the first n_0 is large, this method of convergence assessment will indicate a lack of convergence where there may be none.

For the Geweke (1992) diagnostic, all of the values of Z_n fall between -1.96

and 1.96, and at a 5% significance level, we would not reject the hypothesis that the mean values of the chosen summary statistic are the same in both parts of the output sequence. This does not indicate a lack of convergence after $170,000$ steps. With this diagnostic, we always have the concern that we have not run the chain long enough. It may be the case that the chain stays in one region for a long time, especially in cases where the posterior density has several local maxima. This can can result in a premature indication that there is no evidence of a lack of convergence.

For the Gelman and Rubin (1992) method, we ran 4 independent chains from different starting points. What we see in the table is that for each of the summary statistics, the PSRF is very close to 1. This indicates that as n gets large, we do not expect the estimated posterior density to shrink very much, and this indicates that the chain is close to its stationary distribution.

4.4.4 Discussion

The drift and minorization conditions provide values of λ, b, and ϵ that indicate that the chain will take a very long time to become close to its stationary distribution. However, most of the diagnostics tell a different story. The major issue we see in the theory-based approach to bounding the mixing time is with ϵ. The value of ϵ is much too small to be useful in providing an upper bound on the mixing time. However, there is hope that the lower bound on ϵ can be improved. One approach is to find a larger small set, so that the chain will visit the small set more frequently. To do this, it is necessary to increase the size of the set of values to which the chain can move

in one step, and this should result in faster mixing. Despite this concern, a clear argument for the theory-based approach emerges from the results of the convergence diagnostics. The fact that most of the convergence diagnostics fail to indicate a lack of convergence does not mean that the chain has converged. For instance, it is possible that the chain explores the state space slowly, so it may stay in one region for long periods of time. If this is the case, it would come as no surprise that the difference in the summary statistics in different parts of the chain is small enough that the diagnostics do not find it significant.

4.5 The Behavior of the RSM Algorithm

With any problem in which MCMC methods are employed, it is useful to perform simulations in a variety of situations in order to determine whether the chain behaves as one expects. In this section, we explore the behavior of the RSM algorithm under several scenarios. We begin with a look at how the choice of a prior distribution impacts the behavior of the RSM algorithm. In order to do this, we use four different prior distributions that have different amounts of variability. We then investigate how the behavior of the sampler differs with varying percentages of constant sites in data set. Next, we provide a description of how the number of sites per sequence in the data set affects how the algorithm behaves, and then we provide a brief discussion that ties together the other three parts of this section and that provides insight into how the behavior of the RSM algorithm in the illustrative example of Section 4.4

compares with the behavior of similar chains that use different data sets and prior distributions.

To carry out the investigations mentioned above, we examine the marginal behavior of four summary statistics pertaining to the log branch lengths $\mathbf{w}$. We look at $w^{(1)}$, $w^{(6)}$, $w^{(16)}$, and $\bar{\mathbf{w}}$, the same summaries we examined in the illustrative example. We observe trace plots and histograms for each summary, and we mark the maximum likelihood estimates (MLEs) of each of those summaries. In each simulation, the algorithm is run for 1.7×10^6 steps, and the first 1.7×10^5 are discarded. We choose to discard the first 1.7×10^5 values because after this number of iterations, the convergence diagnostics above did not indicate any problems with convergence. We also thin the chain by a factor of 340, so that the plots we provide show every 340th value of each summary, and we do this because after 340 iterations, each component is expected to have been updated 20 times. This helps to reduce the autocorrelation among the states of the chain.

4.5.1 The Effect of the Prior Distribution on the Behavior of the RSM Algorithm

Here, we compare the behavior of four RSM algorithms for approximating the posterior density of the branch lengths of a phylogenetic tree. The chains we investigate in this subsection use the following prior densities: the 17-dimensional double exponential density, where each log branch length has a $DE(-4.5, 0.25)$ prior distribution and and there is no correlation among the log branch lengths, the $N_{17}(-4.5\mathbf{1}_{17}, 0.25\mathbf{I}_{17})$ density, the $N(-4.5\mathbf{1}_{17}, 4\mathbf{I}_{17})$ density,

and an improper prior that has density 1 over all of $\mathbb{R}^{17}$. The working data set in each of these four situations is a DNA sequence data set with 10,000 sites per sequence, and for which 83.03% of the sites are constant.

Double Exponential Prior Distribution

The trace plots in Figure 4.8 show the output from a version of the RSM algorithm that utilizes the $DE_{17}(-4.5\mathbf{1}_{17}, 0.25\mathbf{1}_{17})$ prior distribution, and they do not indicate any problems with convergence. The acceptance rate for the RSM algorithm using the double exponential prior distribution is 44.94%, which is not indicative of problems with mixing. The increment density for each of the log branch lengths is the Uniform(-0.32, 0.32) density. In the plots, we see the hairiness we desire, and the oscillation is centered near the MLEs for each summary. The value of $w^{(1)}$ moves from approximately -4.3 to -3.5, with centering around -3.9. The values of $w^{(6)}$ are centered around -4.2, and the values of $w^{(16)}$ are oscillating around -4.45, which is closer to the MLE for that summary than what we see in the plots of $w^{(1)}$ and $w^{(6)}$. The mean, as expected, is more stable than the individual branch lengths, taking most of its values between -4.6 and -4.55.

The histograms in Figure 4.9 show a clear shift from the prior distribution, indicating that the chain is pulling the values of these log branch lengths toward regions of higher likelihood. This indicates that the likelihood is controlling the posterior density, while the prior density has little effect.

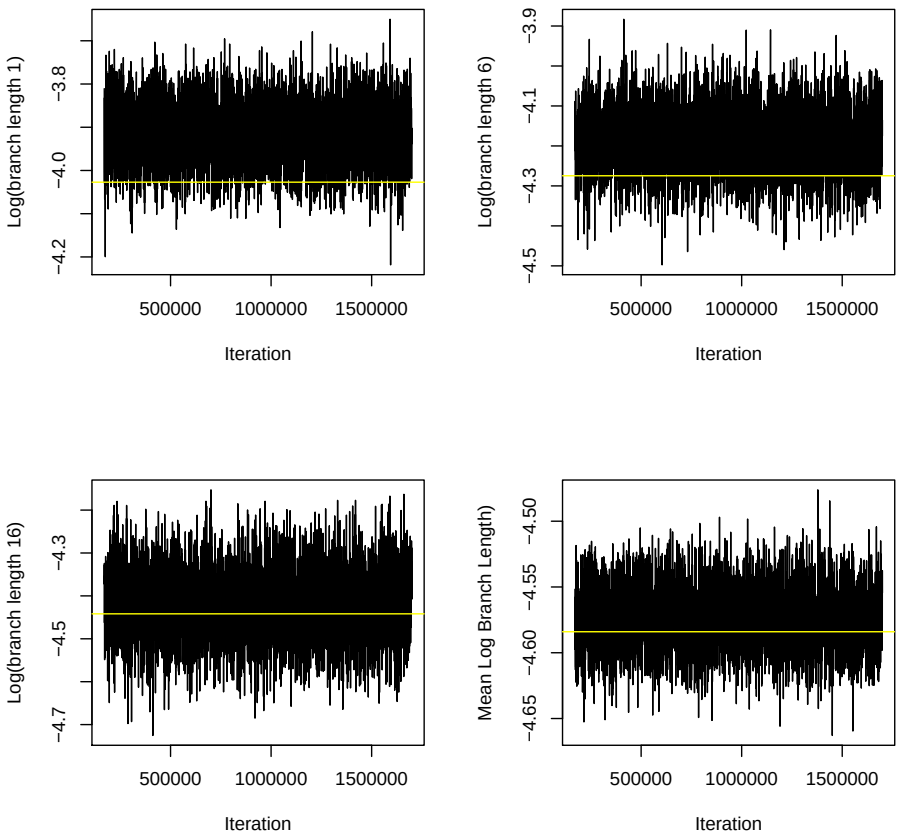


Figure 4.8: Trace plots of $w^{(1)}, w^{(6)}, w^{(16)}$, and $\bar{\mathbf{w}}$ over 1.7×10^6 steps, with the first 1.7×10^5 discarded. The prior density for this RSM algorithm is the $DE_{17}(-4.5\mathbf{1}_{17}, 0.25\mathbf{1}_{17})$ density. 83.03% of the sites in the data set are constant. A yellow line is drawn in each trace plot to indicate the MLE of the corresponding summary for the log branch length vector.

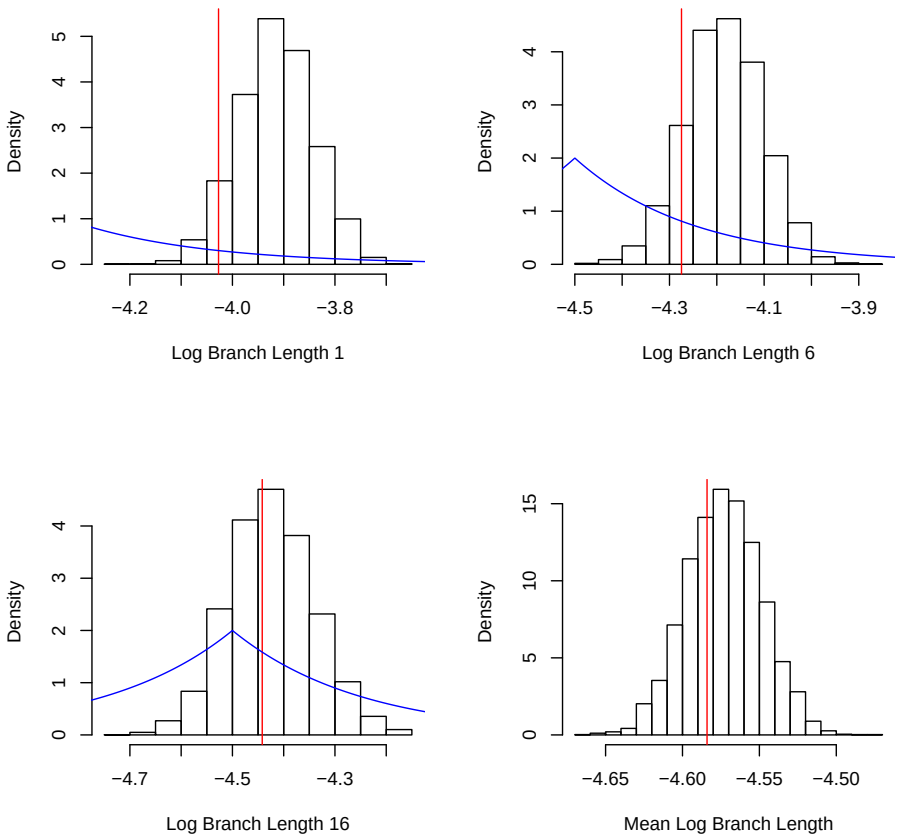


Figure 4.9: Histograms of $w^{(1)}, w^{(6)}, w^{(16)}$, and $\bar{\mathbf{w}}$ over 1.7×10^6 steps, where the first 1.7×10^5 have been discarded. The prior density for an individual log branch length is the $DE(-4.5, 0.25)$ density, and it has been overlaid in blue on the histogram of each of the individual log branch lengths. 83.03% of the sites in the data set are constant. A red vertical line is drawn in each histogram to indicate the MLE of the corresponding summary for the log branch length vector.

Table 4.3: Posterior Mean and Standard Deviation of $w^{(1)}, w^{(6)}, w^{(16)}$, and $\bar{\mathbf{w}}$ with Double Exponential Prior

	Posterior Mean	Posterior Std. Dev.	MLE
$w^{(1)}$	-3.91445	0.07232	-4.02708
$w^{(6)}$	-4.18433	0.08142	-4.27461
$w^{(16)}$	-4.42344	0.08379	-4.44184
$\bar{\mathbf{w}}$	-4.5730	0.02429	-4.58403

The posterior mean, posterior standard deviation, and the MLE of each of the summaries are given in Table 4.3. From this, we see the same things that are illustrated in Figure 4.9. Note that since we placed $DE(-4.5, 0.25)$ prior distribution on each log branch length, each log branch length has prior variance 0.125, which is much larger than the variances we see in Table 4.3. We also observe that, despite the appearance of the histograms and the trace plots, none of the mean values are more than two standard deviations away from the MLE of the log branch length.

$N_{17}(-4.5\mathbf{1}_{17}, 0.25\mathbf{I}_{17})$ Prior Distribution

Figure 4.10 shows trace plots of $w^{(1)}, w^{(6)}, w^{(16)}$, and $\bar{\mathbf{w}}$ for a version of the RSM algorithm that uses the $N(-4.5\mathbf{1}_{17}, 0.25\mathbf{I}_{17})$ density as the prior density. The trace plots do not indicate any problems with convergence, and the acceptance rate for this algorithm is 44.39%. The increment density is the Uniform(-0.34, 0.34) density. The trace plots show that the values of $w^{(1)}$ oscillate around -3.9, the values of $w^{(6)}$ are centered around -4.2, and the values of $w^{(16)}$ oscillate around -4.4. The values of the mean log branch length

are concentrated between -4.60 and -4.55. The individual log branch lengths appear to center at values that are near the MLEs of the chosen log branch lengths.

The histograms that correspond to the individual log branch lengths in Figure 4.11 are each centered at values that are near the MLEs. The histograms show a shift from the prior density, and there is a great deal less variation in the histogram than in the prior density. This indicates that the chain is moving toward a region of high likelihood and staying there.

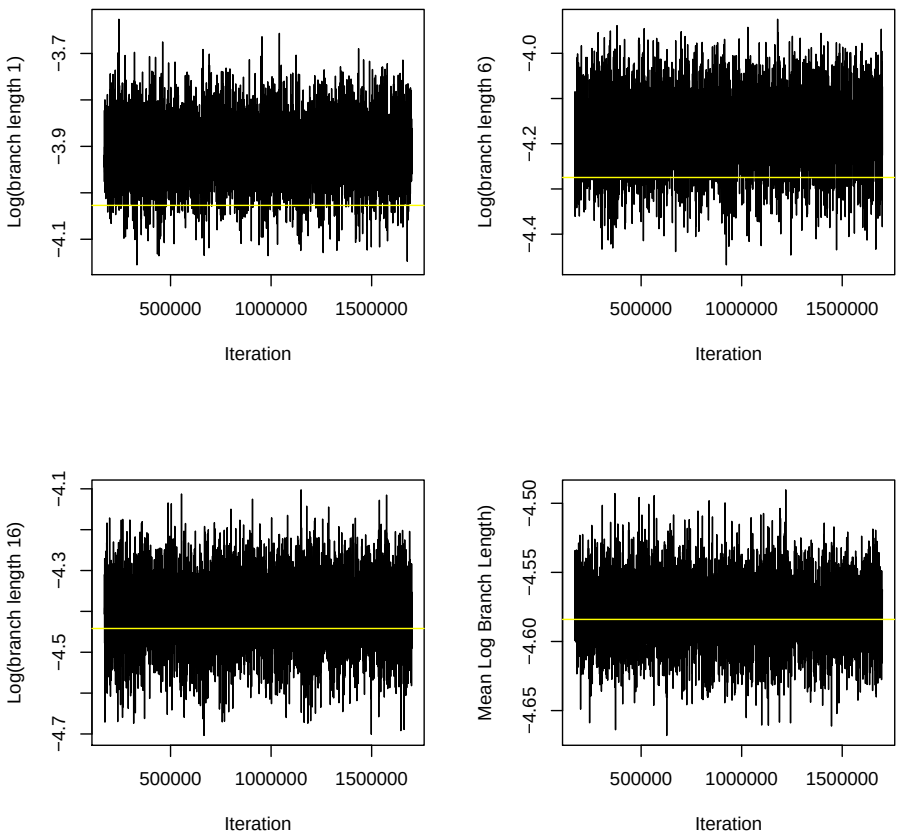


Figure 4.10: Trace plots of $w^{(1)}, w^{(6)}, w^{(16)}$, and $\bar{\mathbf{w}}$ over 1.7×10^6 steps, with the first 1.7×10^5 discarded. The prior density for this version of the RSM algorithm is the $N_{17}(-4.5\mathbf{1}_{17}, 0.25\mathbf{I}_{17})$ density. 83.03% of the sites in the data set are constant. A yellow line is drawn in each trace plot to indicate the MLE of the corresponding summary for the log branch length vector.

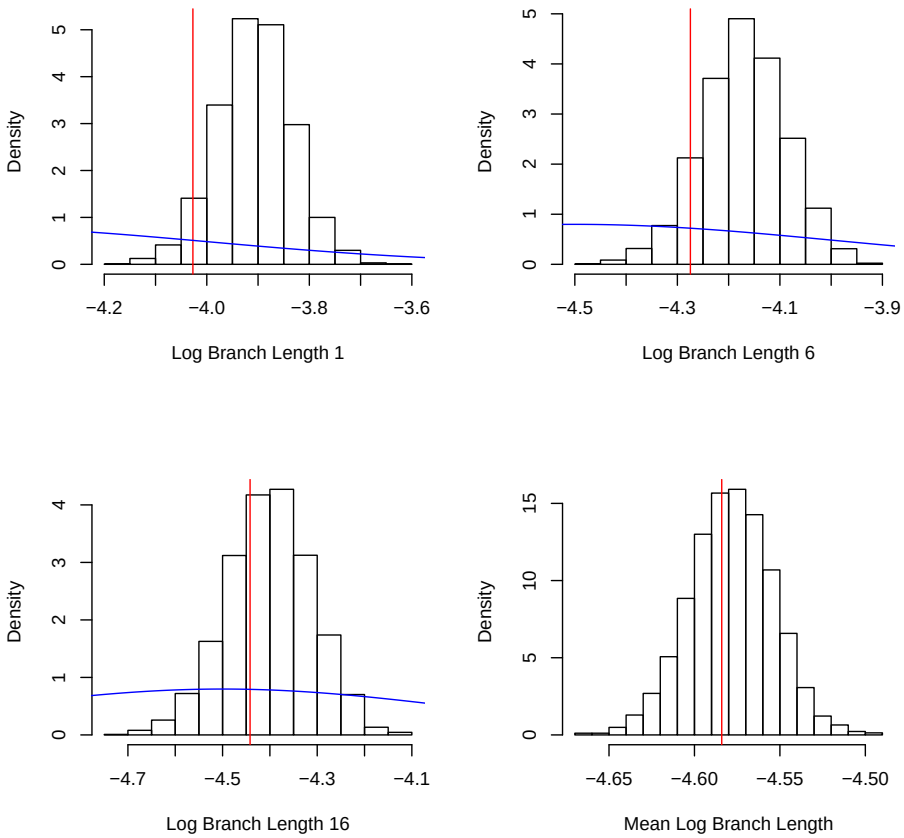


Figure 4.11: Histograms of $w^{(1)}, w^{(6)}, w^{(16)}$, and $\bar{\mathbf{w}}$ over 1.7×10^6 steps, where the first 1.7×10^5 have been discarded. The prior density for an individual log branch length is the $N(-4.5, 0.25)$ density, and it has been overlaid in blue on the histogram of each of the individual log branch lengths. 83.03% of the sites in the data set are constant. A red vertical line is drawn in each histogram to indicate the MLE of the corresponding summary for the log branch length vector.

Table 4.4: Posterior Mean and Standard Deviation of $w^{(1)}, w^{(6)}, w^{(16)}$, and $\bar{\mathbf{w}}$ with $N_{17}(-4.5\mathbf{1}_{17}, 0.25\mathbf{I}_{17})$ Prior

	Posterior Mean	Posterior Std. Dev.	MLE
$w^{(1)}$	-3.90680	0.07280	-4.02708
$w^{(6)}$	-4.17039	0.08246	-4.27461
$w^{(16)}$	-4.40147	0.09017	-4.44184
$\bar{\mathbf{w}}$	-4.57853	0.02449	-4.584033

Table 4.4 gives the posterior mean and standard deviation of the four summaries described Figures 4.10 and 4.11. We see similar things here to that which we observed when the prior distribution was the double exponential. We see that the posterior distribution is much less diffuse than the prior, and that the empirical posterior distributions for each summary are centered no more than two standard deviations away from the MLEs of the corresponding log branch lengths.

$N_{17}(-4.5\mathbf{1}_{17}, 4\mathbf{I}_{17})$ Prior Density

Figure 4.12 shows trace plots of $w^{(1)}$, $w^{(6)}$, $w^{(16)}$, and $\bar{\mathbf{w}}$ from the version of the RSM algorithm that uses the $N_{17}(-4.5\mathbf{1}_{17}, 4\mathbf{I}_{17})$ density as the prior density. There appears to be the hairiness we require in order to conclude that there is no evidence of a lack of convergence, and the acceptance rate of 42.35% does not indicate a problem with mixing. The increment density in this situation for each log branch length is the Uniform(-0.37, 0.37) density. The values of $w^{(1)}$ fall primarily between -3.95 and -3.8; for $w^{(6)}$ the values from the chain fall mostly between -4.25 and -4.1. The values of $w^{(16)}$ lie mostly between

-4.5 and -4.3. The trace plots indicate slightly less variability in the values of the three individual log branch lengths we chose than there appeared to be in the values of the same summaries from the version of the RSM algorithm that uses the $DE_{17}(-4.5\mathbf{1}_{17}, 0.25\mathbf{1}_{17})$ prior and the RSM algorithm which has the $N_{17}(-4.5\mathbf{1}_{17}, 0.25\mathbf{I}_{17})$ density as the prior density. The values of the mean log branch lengths fall primarily between -4.62 and -4.57, and the values of the mean log branch length appear to be centered near the MLE of the mean log branch length.

The histograms show that the individual log branch lengths are centered at values that differ somewhat from the MLEs of the log branch lengths, while the posterior distribution of the mean is centered near the MLE of the mean log branch length. There is much less variation in the distributions of each of the log branch lengths than there is in the prior density, and the distribution of the log branch lengths shows a large shift from the prior density. This is an indication that the likelihood is playing a dominant role in the determination of the posterior density.

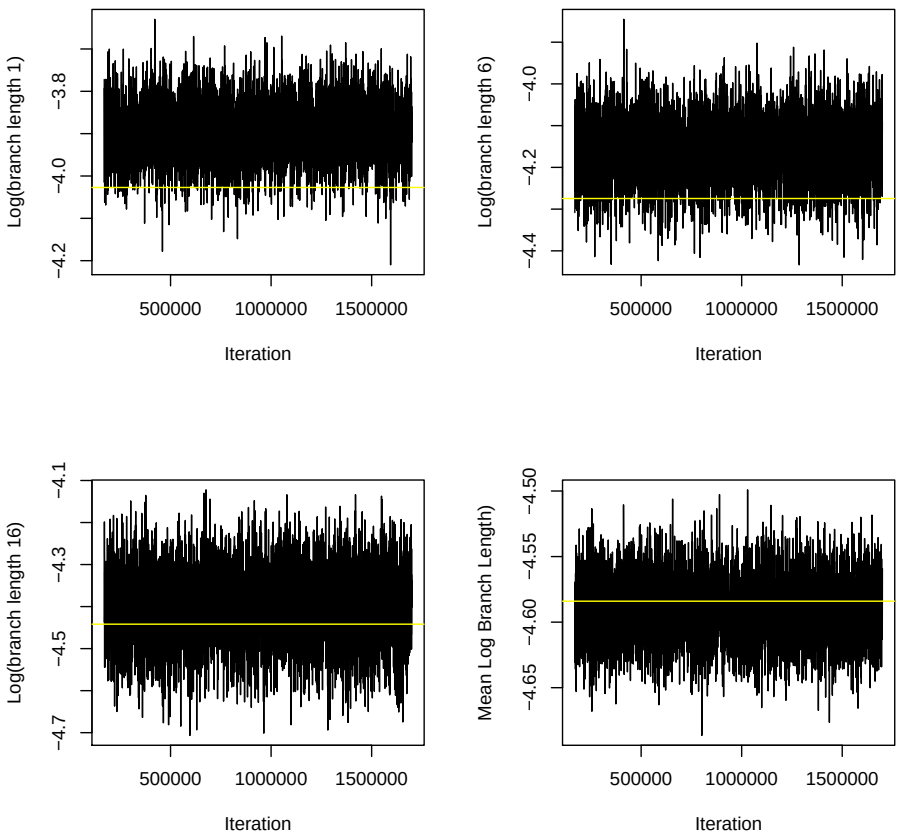


Figure 4.12: Trace plots of $w^{(1)}, w^{(6)}, w^{(16)}$, and $\bar{\mathbf{w}}$ over 1.7×10^6 steps, with the first 1.7×10^5 discarded. The prior density for this RSM algorithm is the $N_{17}(-4.5\mathbf{1}_{17}, 4\mathbf{I}_{17})$ density. 83.03% of the sites in the data set are constant. A yellow line is drawn in each trace plot to indicate the MLE of the corresponding summary for the log branch length vector.

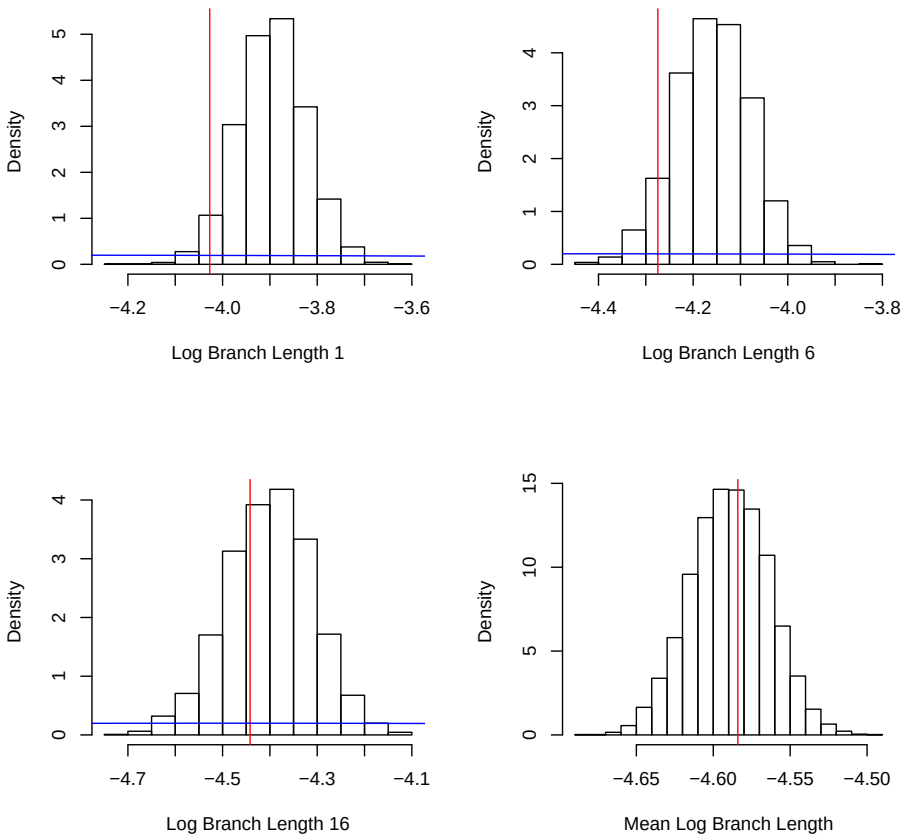


Figure 4.13: Histograms of $w^{(1)}, w^{(6)}, w^{(16)}$, and $\bar{\mathbf{w}}$ over 1.7×10^6 steps, where the first 1.7×10^5 have been discarded. The prior density for an individual log branch length is the $N(-4.5, 4)$ density, and it has been overlaid in blue on the histogram of each of the individual log branch lengths. 83.03% of the sites in the data set are constant. A red vertical line is drawn in each histogram to indicate the MLE of the corresponding summary for the log branch length vector.

Table 4.5: Posterior Mean and Standard Deviation of $w^{(1)}, w^{(6)}, w^{(16)}$, and $\bar{\mathbf{w}}$ with $N_{17}(-4.5\mathbf{1}_{17}, 4\mathbf{I}_{17})$ Prior

	Posterior Mean	Posterior Std. Dev.	MLE
$w^{(1)}$	-3.89473	0.07148	-4.02708
$w^{(6)}$	-4.15857	0.07925	-4.27461
$w^{(16)}$	-4.40101	0.09138	-4.44184
$\bar{\mathbf{w}}$	-4.58926	0.02550	-4.58403

Table 4.5 shows the posterior mean and standard deviation of each of $w^{(1)}$, $w^{(6)}$, $w^{(16)}$, and $\bar{\mathbf{w}}$. We see that, despite the fact that the variance of the prior distribution is much larger than the variance of each of the other two prior distributions used in this section, the empirical posterior mean and variance of each summary is very similar to the posterior means and variances in each of the previous two situations. This is an indication that the prior distribution has very little effect on the posterior density.

Improper Prior With Density 1 Over $\mathbb{R}$

The trace plots in Figure 4.14 show the values of $w^{(1)}$, $w^{(6)}$, $w^{(16)}$, and $\bar{\mathbf{w}}$ from the RSM algorithm which uses the constant, improper prior distribution having density 1 over $\mathbb{R}^{17}$. The increment density for each of the log branch lengths is the Uniform(-0.35, 0.35) density. The trace plots show rapid oscillation around a central value, and this does not indicate any problems with convergence. This algorithm has an acceptance rate of 44.25%. The values of $w^{(1)}$ lie mostly between -4.0 and -3.8, the values of $w^{(6)}$ are concentrated between -4.3 and -4.1, and the values of $w^{(16)}$ fall primarily between -4.5 and

-4.3. The values of the mean log branch length lie mostly between -4.62 and -4.57. The log branch lengths and the mean log branch length are each centered around values that differ slightly from the MLEs of the log branch lengths and the mean log branch length, respectively.

Figure 4.15 shows the marginal empirical distributions of each of the log branch lengths and the mean log branch lengths. In this algorithm, the prior is constant, so the likelihood is equivalent to the posterior density up to a normalizing constant. There are small discrepancies between the actual values of each of the summaries and where the histograms are centered, but these discrepancies are not large enough to cause alarm.

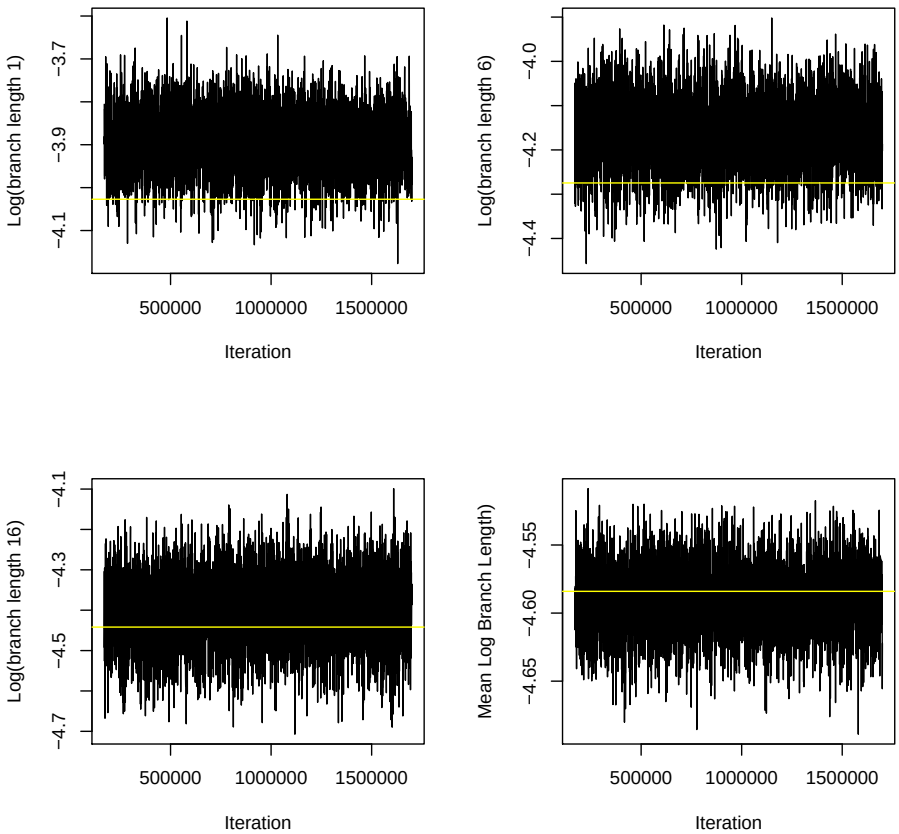


Figure 4.14: Trace plots of $w^{(1)}, w^{(6)}, w^{(16)}$, and $\bar{\mathbf{w}}$ over 1.7×10^6 steps, with the first 1.7×10^5 discarded. The prior density for this RSM algorithm is the constant, improper prior which has density 1 over all of $\mathbb{R}$. 83.03% of the sites in the data set are constant. A yellow line is drawn in each trace plot to indicate the MLE of the corresponding summary for the log branch length vector.

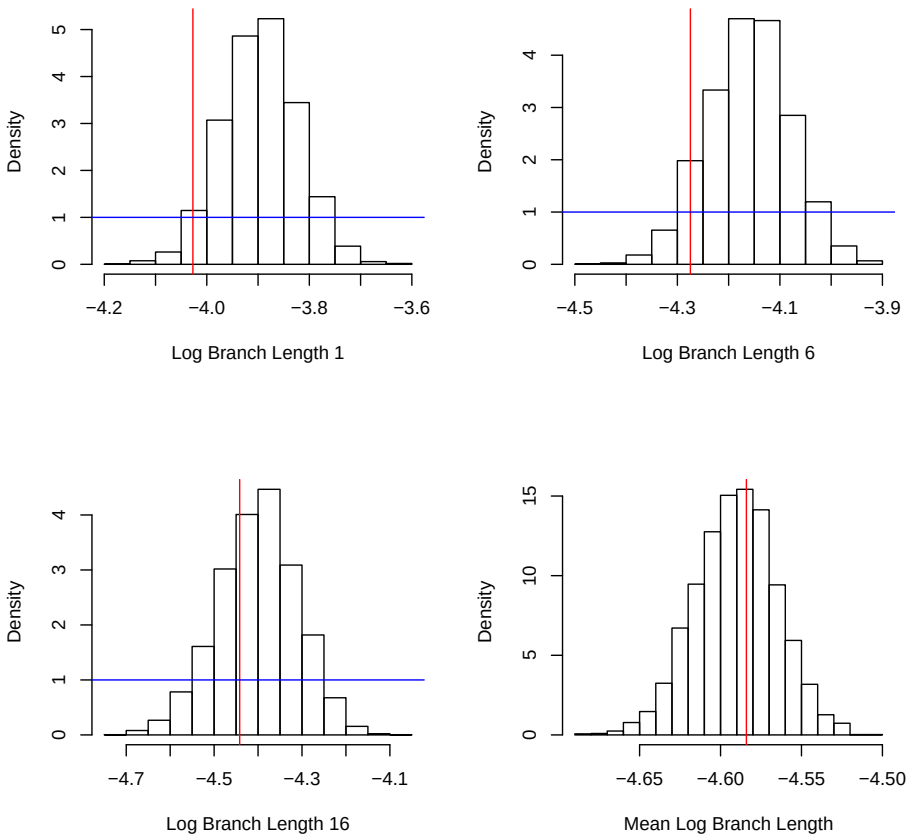


Figure 4.15: Histograms of $w^{(1)}, w^{(6)}, w^{(16)}$, and $\bar{\mathbf{w}}$ over 1.7×10^6 steps, where the first 1.7×10^5 have been discarded. The prior density for each log branch length is constant and improper with "density" 1 over $\mathbb{R}$, and it has been overlaid in blue on the histogram of each of the individual log branch lengths. 83.03% of the sites in the data set are constant. A red vertical line is drawn in each histogram to indicate the MLE of the corresponding summary for the log branch length vector.

Table 4.6: Posterior Mean and Standard Deviation of $w^{(1)}, w^{(6)}, w^{(16)}$, and $\bar{\mathbf{w}}$ Improper Prior having Density 1 over $\mathbb{R}$

	Posterior Mean	Posterior Std. Dev.	MLE
$w^{(1)}$	-3.89491	0.07308	-4.02708
$w^{(6)}$	-4.16126	0.08006	-4.27461
$w^{(16)}$	-4.40085	0.0.9028	-4.44184
$\bar{\mathbf{w}}$	-4.59047	0.02530	-4.58403

Table 4.6 shows the empirical posterior mean and standard deviation of each of the $w^{(1)}$, $w^{(6)}$, $w^{(16)}$ and $\bar{\mathbf{w}}$. We see that the posterior mean and variance of each summary are similar to those under each of the previous three prior distributions. Thus, the distance between the mean and actual value of each summary is similar under each prior distribution. Since this improper prior has no effect on the posterior distribution, this serves as an illustration of how the likelihood is largely determining the posterior density in each of the four situations described in this subsection.

We have examined RSM algorithms for inference of the log branch lengths using four different prior distributions. The differences we see in the behavior among the four algorithms are very slight. We have looked at prior densities that have very little variability, and we have considered priors which have high, and even infinite variability. For each version of the RSM algorithm, the trace plots look very similar, and they center around similar values. The sets of histograms look very similar, and the increment densities that produce a near-optimal acceptance rate are also very similar among the four versions of the

algorithm. The indication here is that the choice of the prior distribution does not play a major part in determining the posterior density. The similarity in the trace plots, histograms, and near-optimal increment densities, despite the significant differences in the prior distribution in each version of the algorithm, indicates that the posterior density is almost completely determined by the likelihood. The similarity among the posterior means and variances of each of the four summaries provides further evidence that the prior does not play a significant role in the behavior of this version of the RSM algorithm. This suggests that, in the situation where the tree topology is known, there is little difference between branch length estimates that come from Bayesian inference and those that come from maximum likelihood.

4.5.2 Effect of the Percentage of Constant Sites on the Behavior of the Chain

Here, we take a look at how the behavior of the RSM algorithm changes with varying percentages of constant sites in the data set. For each simulation, we use the $N_{17}(-4.5\mathbf{1}_{17}, 0.25\mathbf{I}_{17})$ density as the prior density.

60.56% **Constant Sites**

The trace plots in Figure 4.16 show the values of $w^{(1)}$, $w^{(6)}$, $w^{(16)}$, and $\bar{\mathbf{w}}$ from the RSM algorithm with the $N_{17}(-4.5\mathbf{1}_{17}, 0.25\mathbf{I}_{17})$ density as the prior density, and where the working data set has 10,000 sites per sequence. 60.56% of these sites are constant. The increment density for each log branch length in this version of the RSM algorithm is the Uniform(-0.225, 0.225) density.

This choice of an increment density yields an acceptance rate of 43.95%. This acceptance rate, combined with the hairiness in the trace plots, gives no indication of a lack of convergence. We see that most of the values of $w^{(1)}$ from the RSM algorithm lie between -3.9 and -3.7. The chain gives values of $w^{(6)}$ that mostly fall between -3.95 and -3.75, and the values of $w^{(16)}$ lie primarily between -3.7 and -3.55. The mean log branch length is more stable than the individual log branch lengths, and a large majority of the values of the mean log branch length fall between -3.64 and -3.61. In each of the trace plots, we see that the chain settles fairly close to the MLEs of the log branch lengths and the mean log branch length.

The histograms in Figure 4.17 show marginal distributions of each of the summaries. All of them differ greatly from the prior distribution in variability and location. The estimated marginal distributions of each of the summary statistics are centered at values that are near the MLE of each of the summary statistics. This provides evidence that the chain is moving toward values with higher likelihood. This is expected, since in the previous subsection, we observed the prominent role the likelihood plays in the determination of the posterior density. With a higher percentage of non-constant sites, the expectation is that the likelihood will play an even larger role in determining the posterior density, sine there are more distinct site patterns.

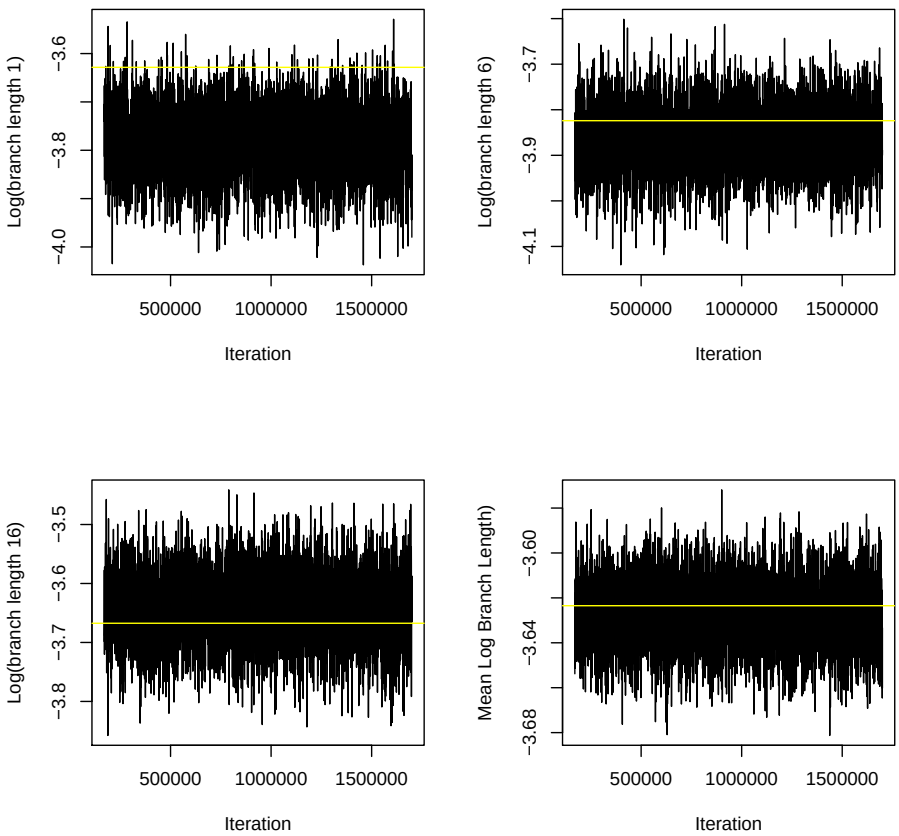


Figure 4.16: Trace plots of $w^{(1)}, w^{(6)}, w^{(16)}$, and $\bar{\mathbf{w}}$ over 1.7×10^6 steps, with the first 1.7×10^5 discarded. The prior density for this RSM algorithm the $N(-4.5\mathbf{1}_{17}, 0.25\mathbf{I}_{17})$ density. 60.56% of the sites in the data set are constant. A yellow line is drawn in each trace plot to indicate the MLE of the corresponding summary for the log branch length vector.

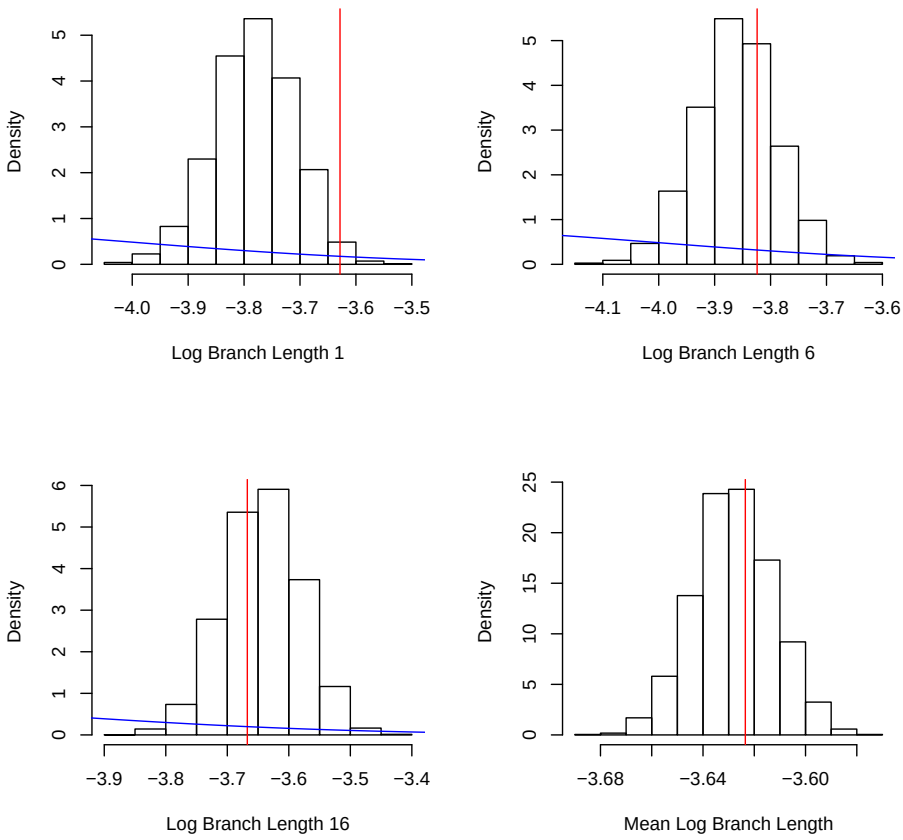


Figure 4.17: Histograms of $w^{(1)}, w^{(6)}, w^{(16)}$, and $\bar{\mathbf{w}}$ over 1.7×10^6 steps, where the first 1.7×10^5 have been discarded. The prior density for each log branch length is the $N(-4.5, 0.25)$ density, and it has been overlaid in blue on the histogram of each of the individual log branch lengths. 60.56% of the sites in the data set are constant. A red vertical line is drawn in each histogram to indicate the MLE of the corresponding summary for the log branch length vector.

Table 4.7: Posterior Mean and Standard Deviation of $w^{(1)}, w^{(6)}, w^{(16)}$, and $\bar{\mathbf{w}}$ 60.56% of Sites are Constant

	Posterior Mean	Posterior Std. Dev.	MLE
$w^{(1)}$	-3.78154	0.07946	-3.62847
$w^{(6)}$	-3.86111	0.07225	-3.82404
$w^{(16)}$	-3.64256	0.06213	-3.66747
$\bar{\mathbf{w}}$	-3.62783	0.01549	-3.62349

Table 4.7 shows the empirical posterior mean and standard deviation of $w^{(1)}$, $w^{(6)}$, $w^{(16)}$, and $\bar{\mathbf{w}}$ when 60.56% of the sites in the data set are constant and the prior distribution is the $N_{17}(-4.5\mathbf{1}_{17}, 0.25\mathbf{I}_{17})$ distribution. We see that the posterior means are larger than those in the four situations in the previous subsection. This makes sense since there are more sites that are not constant. Thus, we expect more substitutions per site. We see that each of the summaries has a mean value that is roughly within two standard deviations of the MLE of the log branch length, so although the empirical distributions of each of these summaries is centered at a value that appears far from the MLE of the summary, the difference is still not overly large in terms of the standard deviation.

91.61% **Constant Sites**

Figure 4.18 shows trace plots of the values of $w^{(1)}$, $w^{(6)}$, $w^{(16)}$, and $\bar{\mathbf{w}}$ for the version of the RSM algorithm that uses the $N_{17}(-4.5\mathbf{1}_{17}, 0.25\mathbf{I}_{17})$ prior density and a data set with 10,000 sites per sequence, where 91.61% of the sites in the data set are constant. For this version of the RSM algorithm, the increment

density for each of the log branch lengths is the Uniform(-0.475, 0.475) density, and the acceptance rate is 44.18%. The trace plots show that the values of $w^{(1)}$ from the chain are mostly between -5.4 and -5.2, the values of $w^{(6)}$ mainly lie between -5.6 and -5.3, and that the values of $w^{(16)}$ are concentrated between -5.5 and -5.3. The mean log branch length has values that lie mostly between -5.32 and -5.28. We observe that the values of $w^{(1)}, w^{(6)}$ and $w^{(16)}$ are centered very close to the MLEs of the respective log branch lengths, while the mean log branch length is centered at a value that differs somewhat from the MLE of the mean log branch length.

The histograms in Figure 4.19 show that the estimated marginal distributions of each of the summary statistics are far less diffuse than the prior distribution. For $\bar{\mathbf{w}}$, the histogram is centered around a value that differs from the MLE of the mean log branch length, while the histograms representing the estimated marginal distributions of $w^{(1)}, w^{(6)}$ and $w^{(16)}$ are centered near the respective MLEs of those summaries.

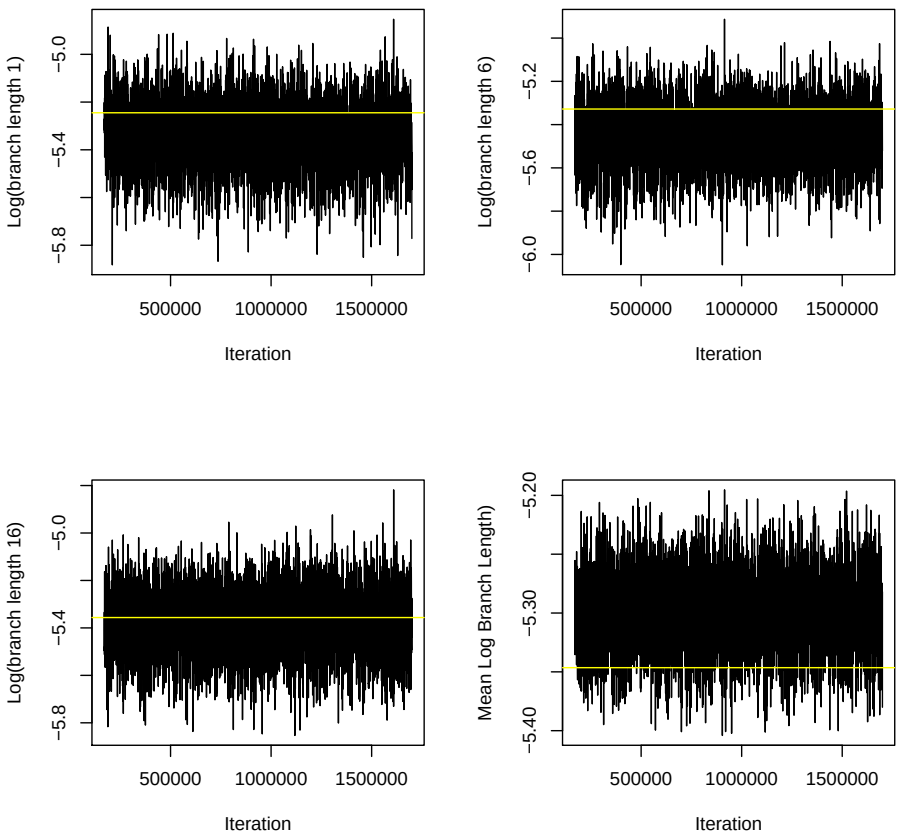


Figure 4.18: Trace plots of $w^{(1)}, w^{(6)}, w^{(16)}$, and $\bar{\mathbf{w}}$ over 1.7×10^6 steps, with the first 1.7×10^5 discarded. The prior density for this RSM algorithm the $N(-4.5\mathbf{1}_{17}, 0.25\mathbf{I}_{17})$ density. 91.61% of the sites in the data set are constant. A yellow line is drawn in each trace plot to indicate the MLE of the corresponding summary for the log branch length vector.

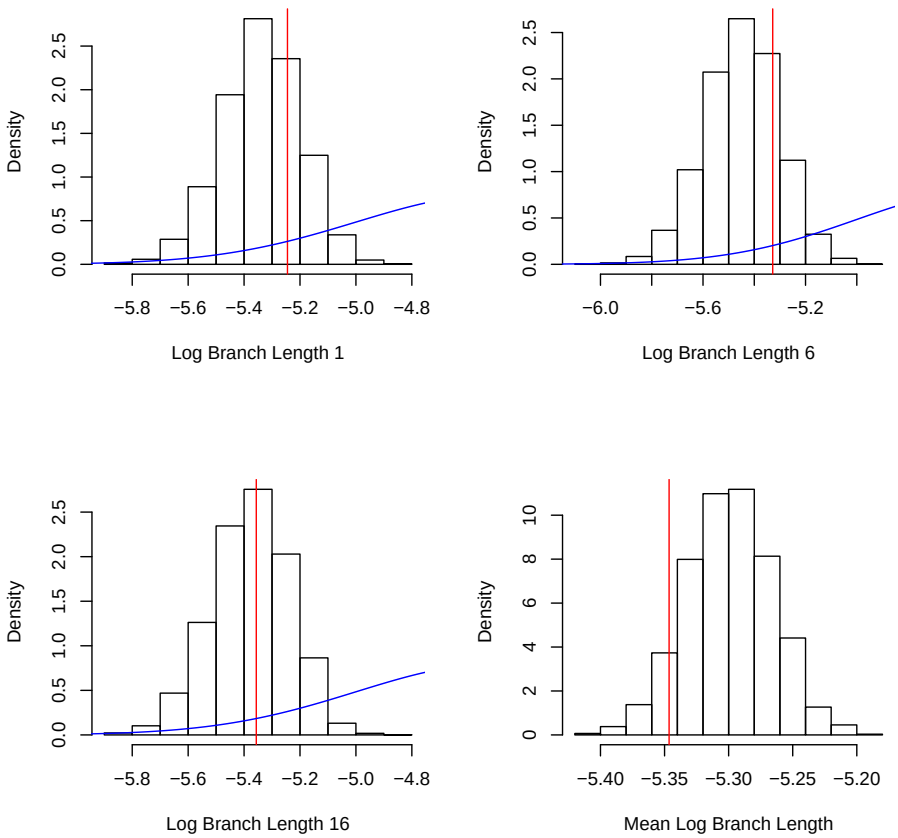


Figure 4.19: Histograms of $w^{(1)}, w^{(6)}, w^{(16)}$, and $\bar{\mathbf{w}}$ over 1.7×10^6 steps, where the first 1.7×10^5 have been discarded. The prior density for each log branch length is the $N(-4.5, 0.25)$ density, and it has been overlaid in blue on the histogram of each of the individual log branch lengths. 91.61% of the sites in the data set are constant. A red vertical line is drawn in each histogram to indicate the MLE of the corresponding summary for the log branch length vector.

Table 4.8: Posterior Mean and Standard Deviation of $w^{(1)}, w^{(6)}, w^{(16)}$, and $\bar{\mathbf{w}}$ 91.61% of Sites are Constant

	Posterior Mean	Posterior Std. Dev.	MLE
$w^{(1)}$	-5.33702	0.14075	-5.24482
$w^{(6)}$	-5.44864	0.14612	-5.32798
$w^{(16)}$	-5.37539	0.13867	-5.35635
$\bar{\mathbf{w}}$	-5.29922	0.03347	-5.34645

Table 4.8 shows the empirical posterior mean and standard deviation of $w^{(1)}$, $w^{(6)}$, $w^{(16)}$, and $\bar{\mathbf{w}}$ where 91.61% of the sites in the data set are constant and the prior distribution is the $N_{17}(-4.5\mathbf{1}_{17}, 0.25\mathbf{I}_{17})$ distribution. The mean values of each summary are smaller than those from the version of the RSM algorithm that uses the data set with 60.56% of the sites are constant. This is expected, since with 91.61% of the sites having no mutations, the expected number of substitutions per sequence is smaller. Each of the summaries have posterior mean values that are within two standard deviations of their MLEs.

In the two situations in this subsection, we see that with a higher percentage of constant sites, we require an increment density that can propose larger changes in the value of the parameter that is chosen to be incremented. We observed in Section 4.5.1 the prior density has little effect on the posterior density. The percentage of constant sites seems to play a slightly larger role here, as is evidenced by the change in where the marginal distributions of the summaries are centered with a higher percentage of constant sites.

4.5.3 Effect of the Size of the Data Set on the Behavior of the Chain

Here, we take a look at how the RSM algorithm behaves with data sets of varying sizes. Both versions of the RSM algorithm discussed here use the $N_{17}(-4.5\mathbf{1}_{17}, 0.25\mathbf{I}_{17})$ distribution as the prior density.

DNA Sequences with 1,000 Sites Apiece

The trace plots in Figure 4.20 show the values of $w^{(1)}$, $w^{(6)}$, $w^{(16)}$, and $\bar{\mathbf{w}}$ for a version of the RSM algorithm that uses the $N_{17}(-4.5\mathbf{1}_{17}, 0.25\mathbf{I}_{17})$ prior density and a data set that consists of 10 1,000-site DNA sequences. 82.00% of the sites in this data set are constant. The increment density for this chain is the Uniform(-0.9, 0.9) density, and the acceptance rate is 44.86%. In the trace plots, there are no indications of a lack of convergence. We also observe that the values of $w^{(1)}$ that come from the RSM algorithm mostly fall between -4.75 and -4.25, the values of $w^{(6)}$ mostly lie between -4.9 and -4.4, and most of the values of $w^{(16)}$ fall between -4.4 and -3.9. The majority of the values of the mean log branch lengths are between -4.65 and -4.45. Thus, we see that with the smaller data set, the mean is still more stable than the individual log branch lengths, and all four of the summary statistics are less stable than the same summary statistics in the versions of the RSM algorithm that rely on the 10,000-site DNA sequence data set. We expect this, as variability in estimation decreases with increasing sample size. We observe also that $w^{(1)}$, $w^{(16)}$, and $\bar{\mathbf{w}}$ are centered at values that differ from the MLEs of these summary statistics, and that do so by more than do the values of the summaries for the version

of the RSM algorithm that used a 10,000-site DNA sequence data set. The values of $w^{(6)}$ from the RSM algorithm appear to be centered near the MLE of $w^{(6)}$.

Figure 4.21 shows histograms for the four summary statistics. Here, we see that, while the marginal distributions for the individual log branch lengths are still less diffuse than the prior, they appear to resemble the prior density more closely than do the marginal distributions of the same log branch lengths in the chains that use the 10,000-site DNA sequence data sets. This is an indication that the prior has a larger effect on the posterior density in this situation than in the ones we discussed previously. This comes as no surprise, as with a smaller data set, one expects the likelihood to have a smaller effect on the posterior density.

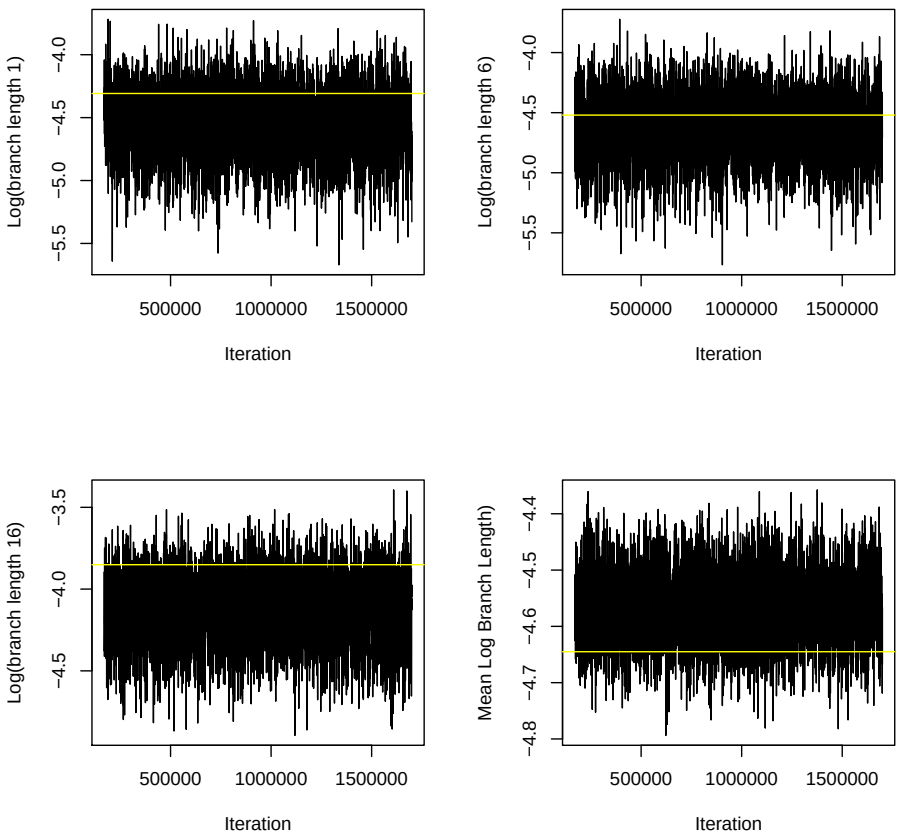


Figure 4.20: Trace plots of $w^{(1)}, w^{(6)}, w^{(16)}$, and $\bar{\mathbf{w}}$ over 1.7×10^6 steps, with the first 1.7×10^5 discarded. The prior density for this version of the RSM algorithm the $N(-4.5\mathbf{1}_{17}, 0.25\mathbf{I}_{17})$ density. 82.00% of the sites in the data set are constant. A yellow line is drawn in each trace plot to indicate the MLE of the corresponding summary for the log branch length vector.

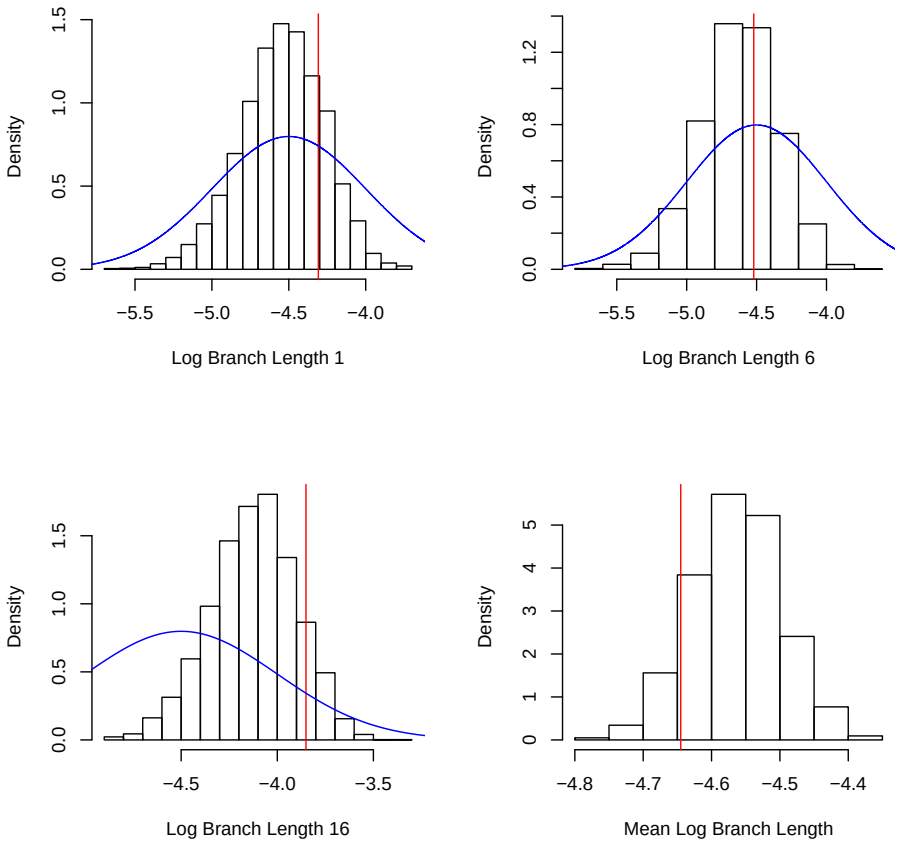


Figure 4.21: Histograms of $w^{(1)}, w^{(6)}, w^{(16)}$, and $\bar{\mathbf{w}}$ over 1.7×10^6 steps, where the first 1.7×10^5 have been discarded. The prior density for each log branch length is the $N(-4.5, 0.25)$ density, and it has been overlaid in blue on the histogram of each of the individual log branch lengths. 82.00% of the sites in the data set are constant. A red vertical line is drawn in each histogram to indicate the MLE of the corresponding summary for the log branch length vector.

Table 4.9: Posterior Mean and Standard Deviation of $w^{(1)}, w^{(6)}, w^{(16)}$, and $\bar{\mathbf{w}}$– 1,000 Sites Per Sequence

	Posterior Mean	Posterior Std. Dev.	MLE
$w^{(1)}$	-4.54331	0.27262	-4.30811
$w^{(6)}$	-4.62716	0.27776	-4.52030
$w^{(16)}$	-4.12636	0.22305	-3.85078
$\bar{\mathbf{w}}$	-4.56385	0.06550	-4.64505

Table 4.9 shows the empirical posterior mean and standard deviation of $w^{(1)}$, $w^{(6)}$, $w^{(16)}$, and $\bar{\mathbf{w}}$ for a version of the RSM algorithm that uses a set of 10 DNA sequences with 1,000 sites apiece and a $N_{17}(-4.5\mathbf{1}_{17}, 0.25\mathbf{I}_{17})$ prior distribution for the log branch lengths. We see that the posterior mean of $w^{(6)}$ is quite close to the MLE of $w^{(6)}$, but that the rest of the summaries have means that are further away from their MLEs. We also note that the standard deviation of each of the summaries are much higher than they have been in previous situations. This is expected, since the data set here only has 1,000 sites per sequence, whereas the other situations used data sets with 10,000 sites per sequence. The mean value of each summary is within two standard deviations of the MLE, so in terms of standard deviations, the distance between the mean and the MLE of each summary is similar to that in each of the previous situations.

DNA Sequences with 100,000 Sites Each

The trace plots in Figure 4.22 show the values of $w^{(1)}$, $w^{(6)}$, $w^{(16)}$ and $\bar{\mathbf{w}}$ for a version of the RSM algorithm that uses a $N_{17}(-4.5\mathbf{1}_{17}, 0.25\mathbf{I}_{17})$ prior density

and a data set that has 100,000 sites per DNA sequence. 80.934% of the sites in the data set are constant. The increment density in this version of the RSM sampler for each of the log branch lengths is the Uniform(-0.105, 0.105) density. The acceptance rate is 44.46%. We observe that the majority of the values of $w^{(1)}$ from the chain lie between -4.85 and -4.79, most of the output values of $w^{(6)}$ fall between -4.65 and -4.60, and the values of $w^{(16)}$ lie mostly between -4.56 and -4.50. The majority of the values of the mean log branch length lie between -4.49 and -4.48. The output values of each summary appear to be centered near the MLE.

The histograms in Figure 4.23 show the estimated marginal distributions of each of the four summary statistics. Each of the marginal distributions for the log branch lengths show much less variation than what is seen in the overlaid prior density. The marginal distribution of each summary is centered near the MLE. This is an indication that the likelihood is dominating the prior distribution in the determination of the posterior density. In other words, the posterior density is almost completely determined by the likelihood.

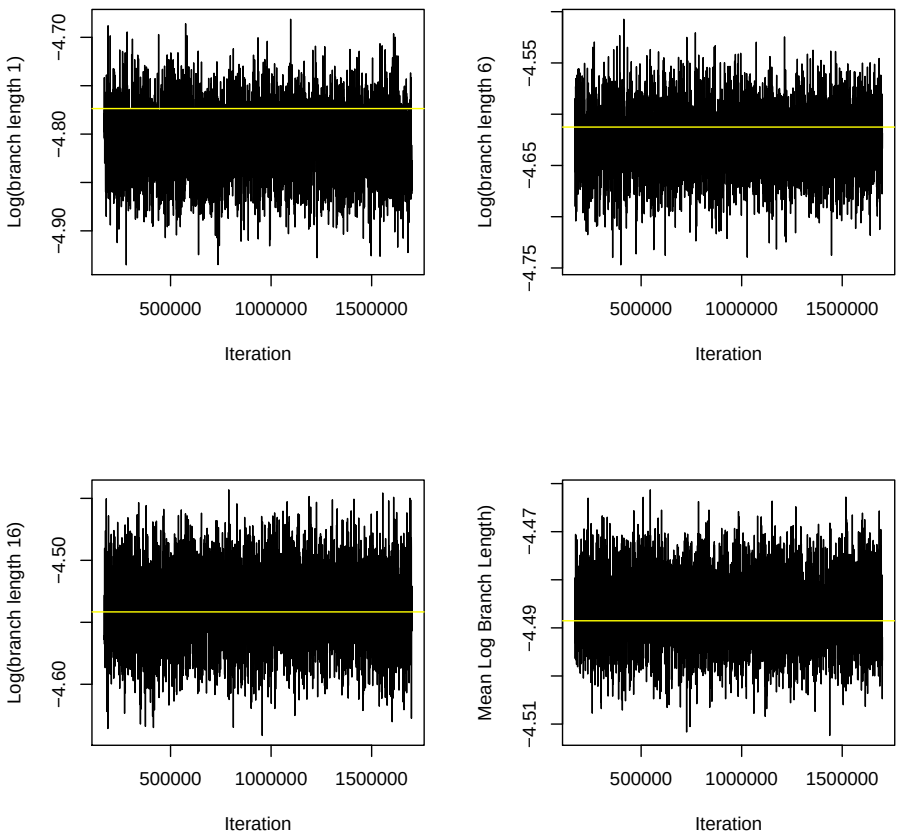


Figure 4.22: Trace plots of $w^{(1)}, w^{(6)}, w^{(16)}$, and $\bar{\mathbf{w}}$ over 1.7×10^6 steps, with the first 1.7×10^5 discarded as burn-in. The prior density for this RSM algorithm the $N(-4.5\mathbf{1}_{17}, 0.25\mathbf{I}_{17})$ density. 80.934% of the sites in the data set are constant. A yellow line is drawn in each trace plot to indicate the MLE of the corresponding summary in the log branch length vector.

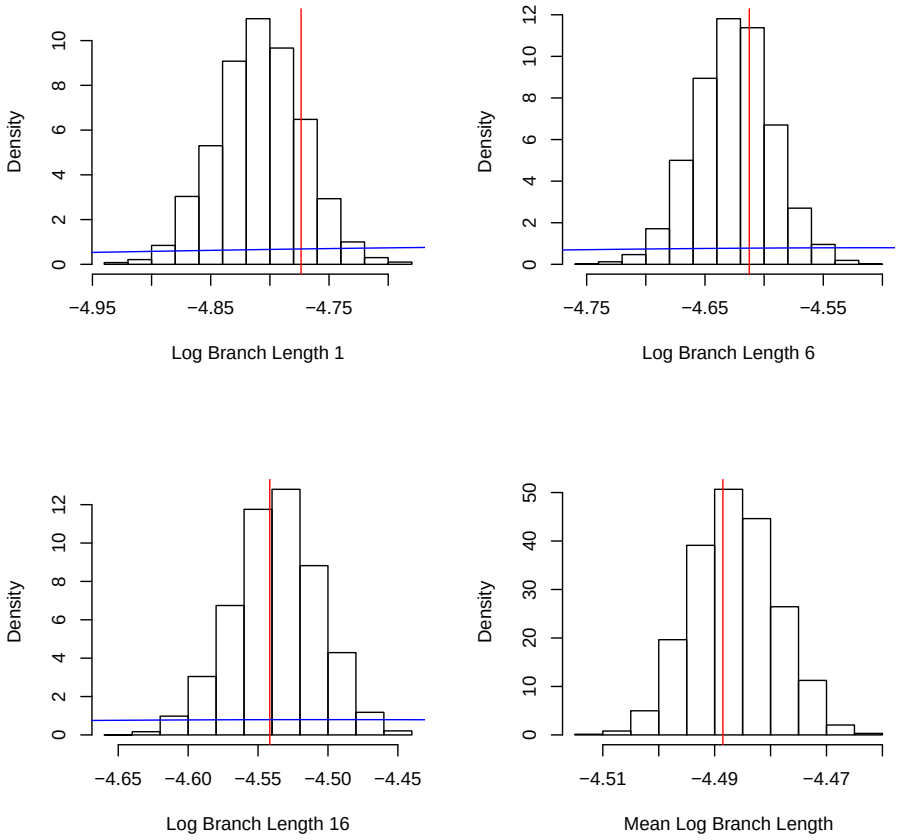


Figure 4.23: Histograms of $w^{(1)}, w^{(6)}, w^{(16)}$, and $\bar{\mathbf{w}}$ over 1.7×10^6 steps, where the first 1.7×10^5 have been discarded as burn-in. The prior density for each log branch length is the $N(-4.5, 0.25)$ density, and it has been overlaid in blue on the histogram of each of the individual log branch lengths. 80.934% of the sites in the data set are constant. A red vertical line is drawn in each histogram to indicate the MLE of the corresponding summary in the log branch length vector.

Table 4.10: Posterior Mean and Standard Deviation of $w^{(1)}, w^{(6)}, w^{(16)}$, and $\bar{\mathbf{w}}$–100,000 Sites Per Sequence

	Posterior Mean	Posterior Std. Dev.	MLE
$w^{(1)}$	-4.80847	0.03619	-4.77362
$w^{(6)}$	-4.62558	0.03256	-4.61256
$w^{(16)}$	-4.53679	0.03050	-4.54161
$\bar{\mathbf{w}}$	-4.486388	0.00765	-4.48853

Table 4.10 gives the empirical posterior mean and standard deviation of $w^{(1)}$, $w^{(6)}$, $w^{(16)}$, and $\bar{\mathbf{w}}$ for a version of the RSM algorithm that uses a data set with 100,000 sites per sequence and a $N_{17}(-4.5\mathbf{1}_{17}, 0.25\mathbf{I}_{17})$ distribution as the prior distribution of the log branch lengths. We observe that the mean values of each summary are closer to the MLEs than what we saw in previous situations. We see that the variability is much smaller in this situation than it is in any of the other situations described above. This comes as no surprise, since the data set is larger than it is in any of the other situations. Each of the log branch lengths has a mean that is within one standard deviation of its MLE, as is the mean log branch lengths. Thus, the log branch lengths appear to be closer to their MLEs in terms of standard deviations as well.

4.5.4 Discussion

In the investigation described above, we observe that as the size of the data set increases, the choice of the prior distribution becomes less important in determining posterior density. This is not unexpected. If we have more information about the taxa represented by the leaves of the tree, it should

play a bigger role in the inference of the branch lengths. From a mathematical perspective, the number of factors in the likelihood calculation is equal to the number of sites in the data set, whereas the number of factors in the prior density is equal to the number of branches. In fact, in the priors we chose, there is no dependence among the branches, so that all of the factors in the prior density corresponding to log branch lengths that are not chosen to be incremented cancel in the acceptance probability. Thus, there is only one factor that comes from the prior, and there are many (100,000 in one of our examples) factors that come from the likelihood in the calculation of the acceptance ratio. It therefore comes as no surprise that as the data set gets large, the prior density plays almost no role in determining the estimation of the posterior distribution.

We observed that as the percentage of constant sites increases, the width of the increment density must increase in order to maintain a desirable acceptance rate for the chain. This also comes as no surprise. If a higher percentage of the sites in the data set are informative, then the RSM chain should be settling in an area of high likelihood. This means that large moves should be rejected frequently, as large moves can take the chain to a state that has small likelihood. A chain that allows this often will suffer adverse effects in terms of convergence, and will therefore be of little help in Bayesian inference of the branch lengths.

The third key observation from the rest of this section is that with a larger data set, the RSM algorithm tends to settle around values that are very close

to the MLEs of the parameters that need to be inferred. This makes intuitive sense, as more data means less variation, so when the chain settles in an area of high likelihood, the empirical distributions of the summaries should be near the actual values of the summaries. We also note that for the data set with 1,000 sites, the increment density was the Uniform(-0.90, 0.90) density, for 10,000 sites it is the Uniform(-0.34, 0.34) density, and for the chain that used the data set with 100,000 sites, the increment density is the Uniform(-0.105, 0.105) density. Note the decreasing diffuseness of the increment density as the number of sites per sequence increases. This is a direct result of having more information about the individuals represented at the leaves of the tree, and the reason the increment density shrinks with more data is precisely the same as the reason it shrinks when we have a lower percentage of constant sites.

The chain we used in the illustrative example strikes a good balance between inferential and computational performance. With a large data set, the computation of the likelihood slows down, and when it needs to be done many times, the difference in computational time can be very burdensome in comparison to using a chain that performs well in terms of inference but that uses a smaller data set. The summaries for the chain we used, with a $N_{17}(-4.5\mathbf{1}_{17}, 0.25\mathbf{I}_{17})$ prior density and a 10,000-site-per-sequence data set with 83.03% of the sites constant provides output that, for the summary statistics we investigated, centers rather closely to their MLEs. In addition, there is low variability in the estimated marginal distribution. Therefore, if ergodic

averages are used to obtain Bayesian estimates of the branch lengths, the resulting estimate will have low variance and low bias. The computing time for running the chain in Section 4.4 is about one-third the computing time for the chain that uses the data set with 100,000 sites. If one had to perform this process many times, it is easy to see how the payoff of lower variability may not be worth the tripled computation time. This is especially true in this setting, since the decrease in variability is rather small, especially when viewed in terms of the branch lengths instead of the log branch lengths.

4.6 Summary

In this chapter, we presented a lower bound on the minorization parameter ϵ for the Markov chain associated with an RSM algorithm that is designed to approximate the posterior distribution of the log branch lengths of a phylogenetic tree given a DNA sequence data set and a tree topology. The analytical upper bound was extremely small, and thus would not be very helpful in providing an upper bound on the mixing time. In an effort to make the value of ϵ more useful, we presented a Monte Carlo method for estimating a lower bound on ϵ, and this gave a lower bound that, while still very small, is much tighter than the one derived analytically. We also presented Monte Carlo methods for estimating upper bounds on the drift parameters λ and b. We then illustrated the use of these methods in a situation where the tree topology is unrooted with 10 taxa, and we looked at some of the well-known convergence diagnostics and found that only one of them indicated any problems with convergence

for our chain. We closed with an investigation of how the algorithm behaves with different prior distributions and different data sets and found that in this setting, the choice of prior and the percentage of constant sites appear to have little effect on the posterior distribution. The size of the data set appears to have a significant effect on the variation in the posterior density. All plots suggest that the multivariate normal distribution provides a good approximation to the posterior distribution. In this situation, there is little difference in Bayesian inference and maximum likelihood estimation. In situations where the tree topology is not known, maximum likelihood estimation requires a search through all possible tree topologies. Bayesian inference requires a similar type of search, but because the posterior distribution is a distribution of the parameters and not of the data, we can use it to form credible sets of trees, including the tree topology. This is a benefit that we do not get from maximum likelihood estimation.

Chapter 5: Conclusion

The work detailed in the preceding pages describes methods of bounding the mixing time for Markov chain Monte Carlo methods that are used in the inference of the parameters of a phylogenetic tree. We dealt with the tree topology and the branch lengths separately. We provided useful upper bounds on the mixing time of two Markov chains that explore the space of rooted phylogenetic tree topologies. We then verified geometric ergodicity of a random scan Metropolis sampler that can be used to infer the branch lengths of a phylogenetic tree with a known tree topology. Once we verified geometric ergodicity, we provided an analytical method for establishment of a minorization condition. This gave a value of ϵ that is too small to be of any use in bounding the mixing time of the RSM sampler. We attempted to combat this problem through computational methods that are designed to estimate a lower bound on ϵ and to estimate upper bounds on λ and b, the parameters in the drift condition. We then gave a description of several of the ad hoc convergence assessment methods, and we provided an example that demonstrates the analytical minorization condition, the computational methods of establishing

minorization and drift conditions, and also shows the results of several output-based convergence assessment methods for our sampler. We closed with an investigation into the behavior of the RSM algorithm when different priors on the log branch lengths are used. In addition, we described how the behavior of the sampler changes when the percentage of constant sites in the data set changes, and when the number of sites in the data set changes while the percentage of constant sites is held steady.

5.1 Summary and Discussion of Results

In Chapter 2, we provided an upper bound of $\mathcal{O}(n^{\frac{5}{2}})$ on the relaxation time of a Markov chain that moves about T_n via SPR transitions. We also showed that a Markov chain that explores T_n via NNI moves has relaxation time that is no larger than $\mathcal{O}(n^4)$. For the RSM sampler for inference of the branch lengths, we verified a minorization and a drift condition through computational methods. While the lower bound on ϵ that came out of the computational method is a significant improvement over the value of ϵ that we obtained analytically, it is still too small to be very helpful in terms of providing a useful upper bound on the mixing time of the RSM sampler.

The establishment of a minorization condition and an associated drift condition represents a significant step forward in the pursuit of an "honest" upper bound on the mixing time of a RSM sampler for inference of the branch lengths of a known phylogenetic tree shape. To our knowledge, there has been no establishment of drift and minorization conditions for MCMC algorithms in this

setting. The methods outlined in Cowles and Rosenthal (1998) are applicable in very general settings, but for a chain that explores a state space with a large number of dimensions, the Monte Carlo method proposed by Cowles and Rosenthal (1998) for estimating ϵ is computationally infeasible. The method we presented is quite useful in situations where a minorizing measure is available, but the establishment of a lower bound on ϵ is intractable, or analytical methods fail to provide a lower bound on ϵ that is helpful in bounding the mixing time.

Despite the concerns regarding the usefulness of $\hat{\epsilon}$, $\hat{\lambda}$, and $\hat{b}$, we are optimistic that these estimates can be improved enough to aid in giving a useful upper bound on the mixing time. Though we detailed many of the criticisms of the output-based convergence assessment methods, the fact that they do not seem to indicate that there are any issues with convergence after $170,000$ steps is very encouraging. If the chain is, in fact, near convergence after $170,000$ steps, we believe that we can improve our estimates and use them to provide an upper bound on the mixing time that more closely reflects this.

In the work presented in Chapter 4, we dealt only with inference of the branch lengths, while assuming the tree topology is known. The reader may have difficulty seeing the benefit of the Bayesian methodology in this setting. We concede this point, as other methods, especially maximum likelihood, are computationally much more efficient when the tree topology is known. However, it is important to note that this work is intended to provide insight into how we can approach Bayesian inference for the setting in which the topology

is not known. It is in this situation that the computational benefit of Bayesian analysis is seen, even with large data sets and complex evolutionary models.

5.2 Future Work

Several ideas come to mind when it comes to directions for future work, but there are two that are immediately to follow the completion of the work presented here. The next step of this research is improve the bounds on the drift and the minorization parameters. One possible way to do this is to find a small set C that is larger than the one we used. With a larger small set, the chain will have to enter C more frequently to ensure geometric convergence, thus increasing the value of ϵ. Consideration of a different drift function may aid in our making a better choice of C. We would want to choose a function that is inversely proportional to the target density, but that is analytically and/or computationally tractable. Analytical tractability is a characteristic that is lacking of the drift function we chose. Another possibility is to increase the size of the set of possible moves for the chain in order to improve mixing. This may also take care of increasing the size of the small set.

Next, our goal is to provide an honest upper bound on the mixing time of a MCMC algorithm for Bayesian inference of phylogenetic trees. The idea is to find a combination of the two types of chains discussed in this dissertation to develop a reliable and efficient method of Bayesian phylogenetic inference. We have considered an RSM algorithm in which the log branch lengths are updated as before, and if the tree shape parameter is the parameter that is

chosen to be updated, we propose an SPR move in which a leaf is chosen at random, removed, and reattached to a randomly chosen branch of the tree. The removal of the assumption of a known tree shape has added a great deal of difficulty to the problem of verifying drift and minorization conditions, as the state space now has both discrete and continuous components. It is possible that in this setting, the RSM algorithm may perform poorly, so we are not committed to using it as our MCMC method for inferring phylogenetic trees. We do believe, however, that because of the high level of dependence among the branch lengths, a one-at-a-time updating scheme is preferable to an all-at-once updating method.

A third question is that of whether or not our model for likelihood calculation is based on a plausible representation of reality. Recall that we assume that evolution between sites in a set of DNA sequences is independent, and that we assume that evolution among lineages is independent. These are likely poor assumptions. It may be better to consider the idea that sites that are less distant from each other are likely to be more closely related in an evolutionary sense than are sites that are more distant. This indicates a spatial dependence among the sites. The same reasoning can be applied to the lineages. Therefore, it would be helpful to incorporate the spatial dependence into the likelihood calculation while still allowing efficient likelihood calculation. We recognize, however, that incorporating spatial dependence makes an already complex function much more complicated, and that achieving both parts of

this goal is a rather tall order.

In addition to extending the research described in this dissertation, there are opportunities for new directions. For instance, a common concern is the computational inefficiency of sampling from complex, high dimensional distributions. A distribution that comes up frequently in many settings is the m-variate truncated normal density. We want to sample a vector $\mathbf{X}$ from a normal distribution whose support is truncated to a set in which for each $i \in \{1, \ldots, m\}$, $X_i \in [-\eta_i, \eta_i]$. This problem has been addressed in many settings where the covariance matrix $\mathbf{\Sigma}$ is assumed to be of the form $\mathbf{\Sigma} = c\mathbf{I}_m$, for some constant $c > 0$. In the problem we plan to address,

$$\mathbf{\Sigma}_{ij} = c\rho^{|i-j|},$$

where $\rho \in (0, 1)$. The mean vector is an m-dimensional vector $\boldsymbol{\mu}$. An approach that shows promise is the polar slice sampler, introduced by Roberts and Rosenthal (2002). Provided the density from which sampling is to occur is log-concave, the polar slice sampler has been shown to have convergence properties that are independent of the dimension of the problem.

Bibliography

Aiki-Raji, C. O., Aguilar, P. V., Kwon, Y.-K., Goetz, S., Suarez, D. L., Jethra, A. I., Nash, O., Adeyefa, C. A., Adu, F. D., Swayne, D., and Basler, C. F. (2008). Phylogenetics and pathogenesis of early avian influenza viruses(h5n1), nigeria. *Engineering Infectious Diseases*, 14(11):1753–1755.

Aldous, D. (2000). Mixing time for a markov chain on cladograms. *Combinatorics, Probability and Computing*, 9:191–204.

Aldous, D. (2012). Mixing times for the branch-rotation chain on cladograms (or the triangulation walk). Technical report, University of California at Berkeley.

Bayer, D. and Diaconis, P. (1992). Trailing the dovetail shuffle to its lair. *Annals of Applied Probability*, 2(2):294–313.

Beiko, R. G., Keith, J. M., Harlow, T. J., and Ragan, M. A. (2006). Searching for convergence in phylogenetic markov chain monte carlo. *Systematic Biology*, 55(4):553–565.

Bissell, A. (1969). Cusum techniques for quality control. *Journal of the Royal Statistical Society, Series C*, 18:1–30.

Breyer, L. and Roberts, G. O. (2000). Some multi-step coupling constructions for markov chains. Technical report, University of Lancaster.

Brooks, S. and Gelman, A. (1998). Alternative methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455.

Brooks, S. and Roberts, G. (1997). Assessing convergence of markov chain monte carlo algorithms. *Statistics and Computing*, 8:319–335.

Brooks, S. P. (1996). Quantitative convergence diagnosis for mcmc via cusums. Technical report, University of Bristol.

Brown, J. M., Hedtke, S. M., Lemmon, A. R., and Lemmon, E. M. (2010). When trees grow too long: Investigating the causes of highly inaccurate bayesian branch-length estimates. *Systematic Biology*, 59:145–161.

Cavalli-Sforza, L. L. and Edwards, A. W. F. (1963). The reconstruction of evolution. *Ann. Hum. Genet.*, 27:105–106.

Cavalli-Sforza, L. L. and Edwards, A. W. F. (1967). Phylogenetic analysis:models and estimation procedures. *Evolution*, 21(3):550–570.

CDC (1991). Epidemiologic notes and reports update: Transmission of hiv infection during invasive dental procedures–florida. *Morbidity and Mortality Weekly Report*, 40(23):377–381.

Chen, J.-M., Sun, Y.-X., Chen, J.-W., Liu, S., Jian-Min Yu, Chao-Jian Shen, X.-D. S., and Peng, D. (2009). Panorama phylogenetic diversity and distribution of type a influenza viruses based on their six internal gene sequences. *Virol. J.*, 6:1–17.

Christensen, H., Blackall, P. J., and Bisgaard, M. (2009). Phylogenetic relationships of unclassified, satellitic pasteurellaceae obtained from different species of birds as demonstrated by 16s rrna gene sequence comparison. *Res. Microbiol.*, 160:315–321.

Clauset, A., Moore, C., and Newman, M. (2008). Hierarchical structure and the prediction of missing links in networks. *Nature*, 453:98–101.

Clement, M., Posada, D., and Crandall, K. (2000). Tcs: A computer program to estimate gene genealogies. *Molecular Ecology*, 9:1657–1659.

Cowles, M. K. and Carlin, B. P. (1996). Markov chain monte carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91(434):883–904.

Cowles, M. K. and Rosenthal, J. S. (1998). A simulation approach to convergence rates for markov chain monte carlo algorithms. *Statistical Computing*, 8:115–124.

Diaconis, P. and Holmes, S. (2002). Random walks on trees and matchings. *Electronic Journal of Probability*, 7:1–17.

Diaconis, P. and Stroock, D. (1991). Geometric bounds for eigenvalues of markov chains. *The Annals of Applied Probability*, 1:36–61.

Diezmann, S., Cox, C. J., Schonian, G., Vilgalys, R. J., , and Mitchell, T. G. (2004). Phylogeny and evolution of medical species of candida and related taxa: A multigenic analysis. *J. Clin. Microbiol.*, 42(12):5624–5635.

Dowell, K. (2008). Molecular phylogenetics. Technical report, University of Maine.

Felsenstein, J. (1981). Evolutionary trees from dna sequences: A maximum likelihood approach. *J. Mol. Evol.*, 17(6):368–376.

Felsenstein, J. (1989). Phylip- phylogenetic inference package(version 3.2). *Cladistics*, 5:164–166.

Fort, G., Moulines, E., Roberts, G. O., and Rosenthal, J. S. (2003). On the geometric ergodicity of hybrid samplers. *Journal of Applied Probability*, 40:123–146.

Gelfand, A. and Smith, A. (1990). Sampling-based approaches to calculating marginal densities. *J. Am. Stat. Assoc.*, 85:398–409.

Gelman, A. and Rubin, D. (1992). Inference from iterative simulation using multiple sequences(with discussion). *Statistical Science*, 7:457–511.

Geweke, J. (1992). *Bayesian Statistics 4*, chapter Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments, pages 169–193. Oxford University Press.

Guindon, S. and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, 52(5):696–704.

Hasegawa, M., Kishino, H., and Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial dna. *Journal of Molecular Evolution*, 22:160–174.

Hastings, W. (1970). Monte carlo sampling techniques using markov chains and their applications. *Biometrika*, 57:97–109.

Hendy, M. and Penny, D. (1982). Branch and bound algorithms to determine minimal evolutionary trees. *Mathematical Biosciences*, 59:277–290.

Herbei, R. and Kubatko, L. S. (2013). Monte carlo estimation of total variation distance of markov chains on large spaces, with applications to phylogenetics. *Statistical Applications in Genetics and Molecular Biology*, 12:39–48.

Herr, R. A., Ajello, L., Tayler, J. W., Arseculeratne, S. N., and Mendoza, L. (1999). Phylogenetic analysis of rhinosporidium seeberi 18s small- subunit ribosomal dna groups this pathogen among members of the protoctistan mesomycetozoa clade. *J. Clin. Microbiol.*, 37(9):2750–2754.

Huelsenbeck, J. and Ronquist, F. (2001). Mrbayes: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8):754–755.

Huelsenbeck, J. P. (1998). Systematic bias in phylogenetics: Is the strepsiptera problem solved? *Systematic Biology*, 43:519–537.

Jarner, S. F. and Hansen, E. (2000). Geometric ergodicity of metropolis algorithms. *Stochastic Processes and their Applications*, 85:341–361.

Jerrum, M. and Sinclair, A. (1989). Approximating the permanent. *SIAM Journal on Computing*, 18:1149–1178.

Jones, G. L. and Hobert, J. P. (2001). Honest exploration of intractable probability distributions via markov chain monte carlo. *Statistical Science*, 16(4):312–334.

Jukes, T. and Cantor, C. (1969). *Mammalian Proten Metabolism, Volume III*, chapter Evolution of Protein Molecules, pages 21–32. Academic Press, New York.

Karlin, S. and Taylor, H. (1975). *A First Course in Stochastic Processes*. Academic Press.

Kimura, M. (1980). A simple model for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16:111–120.

Kishino, H. and Hasegawa, M. (1989). Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from dna sequence data, and the branching order in hominoidea. *Journal of Molecular Evolution*, 29:170–179.

Kuhnert, D., Wu, C.-H., and Drummond, A. J. (2011). Phylogenetic and epidemic modeling of rapidly evolving infectious diseases. *Infect. Genet. Evol.*, 11:1825–1841.

Kuhnert, P., Boerlin, P., Emler, S., Krawinkler, M., and Frey, J. (2000). Phylogenetic analysis of pasteurella multocida subspecies and molecular identification of feline p multocida subspecies septica by 16s rrna gene sequencing. *Int. J. Med. Microbiol.*, 290:599–604.

Levin, D. A., Peres, Y., and Wilmer, E. L. (2009). *Markov Chains and Mixing Times*. American Mathematical Society.

Li, H. and Homer, N. (2010). A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics*, 11:473–483.

Li, S., Pearl, D. K., and Doss, H. (2000). Phylogenetic tree construction using markov chain monte carlo. *Journal of the American Statistical Association*, 95(450):493–508.

Liu, C., Liu, J., and Rubin, D. (1993). A control variable for assessment of the convergence of the gibbs sampler. In *Proceedings of the Statistical Computing Section of the American Statistical Association*, pages 74–78.

Mar, J. C., Harlow, T. J., and Ragan, M. A. (2005). Bayesian and maximum likelihood phylogenetic analyses of protein sequence data under relative branch-length differences and model violation. *BMC Evolutionary Biology*, 5:8.

McKenzie, A. and Steel, M. (2000). Distributions of cherries for two models of trees. *Mathematical Biosciences*, 164:81–92.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–1092.

Meyn, S. and Tweedie, R. L. (2009). *Markov Chains and Stochastic Stability.* Cambridge University Press.

Pinto, R. L. and Neal, R. M. (2001). Improving markov chain monte carlo estimators by coupling to an approximating chain. Technical report, University of Toronto.

Propp, J. and Wilson, D. (1996). Exact sampling with coupled chains and applications to statistical mechanics. *Random Structures and Algorithms*, 9:223–252.

Randall, D. and Tetali, P. (2000). Analyzing glauber dynamics by comparison of markov chains. *Journal of Mathematical Physics*, 41:1598–1615.

Roberts, G. (1994). *Bayesian Statistics and Econometrics: Essays in Honor of Arnold Zellner*, chapter Methods for Estimating L2 Convergence of Markov Chain Monte Carlo. North Holland.

Roberts, G. O. and Rosenthal, J. S. (2001). Optimal scaling for various metropolis-hastings algorithms. *Statistical Science*, 16:351–367.

Roberts, G. O. and Rosenthal, J. S. (2002). The polar slice sampler. *Stochastic Models*, 18:257–280.

Roberts, G. O. and Tweedie, R. (1996). Geometric convergence and central limit theorems for multidimensional hastings and metropolis algorithms. *Biometrika*, 83:95–110.

Rosenthal, J. S. (1995). Minorization conditions and convergence rates for markov chain monte carlo. *J. Am. Stat. Assoc.*, 90(430):558–566.

Ross, J. (2011). Invasion of infectious diseases in finite homogeneous populations. *J. Theor. Biol.*, 289:83–89.

Salter, L. A. and Pearl, D. K. (2001). Stochastic search strategy for estimation of maximum likelihood phylogenetic trees. *Systematic Biology*, 50(1):7–17.

Schweinsberg, J. (2002). An o(n2) bound for the relaxation time of a markov chain on cladograms. *Random Structures and Algorithms*, 20:59–70.

Sokal, R. R. and Sneath, P. H. (1963). *Numerical Taxonomy.* W.H. Freeman.

Stamatakis, A. (2006). Raxml: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–2690.

Swofford, D. L. (2002). *PAUP*–Phylogenetic Analysis Using Parsimony(*and Other Methods) Version 4.06.10.* Sinauer Associates, Inc.

Tamura, K. and Nei, M. (1993). Estimation of the umber of nucleotide substitutions in the control region of mitochondrial dna in humans and chimpanzees. *Molecular Biology and Evolution*, 10:612–526.

Warnow, T., Evans, S. N., Ringe, D., and Nakhleh, L. (2006). *Phylogenetic Methods and the Prehistory of Languages*, chapter A Stochastic Model of Language that Incorporates Homoplasy and Borrowing, pages 75–87. MacDonald Institute for Archaeological Research/University of Cambridge.

Yang, Z. and Rannala, B. (1997). Bayesian phylogenetic inference using dna sequences: a markov chain monte carlo method. *Mol. Biol. Evol.*, 14:717–724.

Yang, Z. and Rannala, B. (2005). Branch-length prior influences bayesian posterior probability of phylogeny. *Systematic Biology*, 54:455–470.

Yu, B. (1995). Estimating l1 error of kernel estimator: Monitoring convergence of markov samplers. Technical report, Department of Statistics; University of California Berkeley.

Yu, B. and Mykland, P. (1998). Looking at markov samplers through cusum path plotsa a simple diagnostic idea. *Statistics and Computing*, 8:275–286.

Zwickl, D. J. (2008). *GARLI Version 0.96 Beta*.