Tropical Machine Learning Models and Applications to Phylogenetic Trees

Georgios Aliatimis, BA (Hons), MSc, MRes

Submitted for the degree of Doctor of Philosophy at Lancaster University.

November 2025

STOR-i
excellence with impact

Abstract

Classification of gene trees is an important task both in the analysis of multi-locus phylogenetic data, and assessment of the convergence of Markov Chain Monte Carlo (MCMC) analyses used in Bayesian phylogenetic tree reconstruction. The logistic regression model is one of the most popular classification models in statistical learning, thanks to its computational speed and interpretability. However, it is not appropriate to directly apply the standard logistic regression model to a set of phylogenetic trees, as the space of phylogenetic trees is non-Euclidean and thus contradicts the standard assumptions on covariates. It is well-known in tropical geometry and phylogenetics that the space of phylogenetic trees is a tropical linear space in terms of the max-plus algebra. Therefore, in this thesis, we propose an analogue approach of the logistic regression model in the setting of tropical geometry. Our proposed method outperforms classical logistic regression in terms of Area under the ROC Curve in numerical examples, including with data generated by the multi-species coalescent model. Theoretical properties such as statistical consistency are proved and generalization error rates are derived. Finally, our classification algorithm is proposed as an MCMC convergence criterion for Mr Bayes. Unlike the convergence metric used by Mr Bayes which is only dependent on tree topologies, our method is sensitive to branch lengths and therefore provides a more robust metric for convergence. In a test case, it is illustrated that the tropical logistic regression can differentiate between two independently run MCMC chains, even when the standard metric cannot.

Building upon this tropical geometric foundation, deep neural networks show great success when input vectors are in an Euclidean space. However, those classical neural networks show a poor performance when inputs are phylogenetic trees, which can be written as vectors in the tropical projective torus. Here we propose tropical embedding to transform a vector in the tropical projective torus to a vector in the Euclidean space via the tropical metric. We introduce a tropical neural network where the first layer is a tropical embedding layer and the following layers are the same as the classical ones. We prove that a tropical neural network is a universal approximator and we derive a backpropagation rule for deep tropical neural networks. Then we provide TensorFlow 2 codes for implementing a tropical neural network in the same fashion as the classical one, where the weights initialization problem is considered according to the extreme value statistics. We apply our method to empirical data including sequences of hemagglutinin for influenza virus from New York. Finally we show that a tropical neural network can be interpreted as a generalization of a tropical logistic regression.

While the first two parts focus on statistical learning within tree space geometry, the final part of this thesis addresses the underlying statistical noise in phylogenomic data. In phylogenomics, species-tree methods must contend with two major sources of noise: stochastic gene-tree variation under the multispecies coalescent model (MSC) and finite-sequence substitutional noise. Fast agglomerative methods such as GLASS, STEAC, and METAL combine multi-locus information via distance-based clustering. We derive the exact covariance matrix of these pairwise distance estimates under a joint MSC-plus-substitution model and leverage it for reliable confidence estimation, and we algebraically decompose it into components attributable to coalescent variation versus sequence-level stochasticity. Our theory identifies parameter regimes where one source of variance greatly exceeds the other. For both very low and very high mu-

tation rates, substitutional noise dominates, while coalescent variance is the primary contributor at intermediate mutation rates. Moreover, the interval over which coalescent variance dominates becomes narrower as the species-tree height increases. These results imply that in some settings one may legitimately ignore the weaker noise source when designing methods or collecting data. In particular, when gene-tree variance is dominant, adding more loci is most beneficial, while when substitution noise dominates, longer sequences or imputation are needed. Finally, leveraging the derived covariance matrix, we implement a Gaussian-sampling procedure to generate split support values for METAL trees and demonstrate empirically that this approach yields more reliable confidence estimates than traditional bootstrapping.

Acknowledgements

I would first like to express my deepest gratitude to Professor Ruriko (Rudy) Yoshida, who proposed this project and guided me throughout my PhD. Her extensive knowledge of tropical geometry and phylogenetics, exceptional research intuition, and innovative ideas have been truly inspiring. I have greatly benefited from her wide network of research collaborations, high standards, and strong work ethic.

I am also sincerely thankful to James Grant, my supervisor at Lancaster, for his continuous support and valuable guidance throughout my PhD. His insightful suggestions, detailed feedback, and actionable plans for improvement have significantly strengthened my work. I would also like to thank Burak Boyaci, my co-supervisor at Lancaster, for his support and encouragement during the often bumpy journey of pursuing a PhD.

This thesis would not have been possible without the contributions of my external collaborators, Keiji Miura and David Barnhill, with whom I had the pleasure of co-authoring research papers. Their expertise and collaboration were instrumental to the success of this work.

I gratefully acknowledge the STOR-i Centre for Doctoral Training and EPSRC for funding my research. I am also thankful to our academic and strategic partner, the Postgraduate Naval School in California, which I had the opportunity to visit twice during my PhD. These visits provided invaluable opportunities to collaborate more closely with Rudy and David.

On a personal note, I owe my deepest thanks to my family and friends for their

unwavering support. To my parents, my sister, Maria-Christina, and my friends Rafael, Theo and Fan Wang I cannot thank you enough for your constant encouragement, belief in me, and support in every possible way. This achievement would not have been possible without you.

Declaration

I declare that the work in this thesis has been done by myself and has not been submitted elsewhere for the award of any other degree.

A version of Chapter 3 has been published as Aliatimis, G., Yoshida, R., Boyacı, B. and Grant, J.A., 2024. Tropical logistic regression model on space of phylogenetic trees. Bulletin of Mathematical Biology, Volume 86, article number 99.

A version of Chapter 4 has been published as Yoshida, R., Aliatimis, G. and Miura, K., 2024, June. Tropical neural networks and its applications to classifying phylogenetic trees. In 2024 International Joint Conference on Neural Networks (IJCNN) (pp. 1-9). IEEE.

A version of Chapter 5 has been submitted for publication as Aliatimis, G., Yoshida, R., Boyaci, B. and Grant, J.A., 2025. Covariance Decomposition for Distance Based Species Tree Estimation.

Georgios Aliatimis

Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation

Over the past few decades, advances in sequencing technologies have transformed the field of genetics. The rapid decline in sequencing costs and the rise of large scale genomic projects (such as the human genome project (Hood and Rowen, 2013) and the Darwin Tree of Life Project (Darwin Tree of Life Project Consortium, 2022)) have led to an accumulation of molecular data. Analysing these vast datasets requires statistically robust but also computationally efficient methods capable of inferring *phylogenetic trees* that provide a graphical representation of evolutionary relationships. This demand for *genetic* and *genomic* data (where a *genome* refers to the complete set of an organism's genes, genetic refers to information at the individual gene level and genomic relates to studies that consider the genome as a whole) has led to the development of new statistical models designed to accommodate their nature, scale and complexity. Phylogenetic tree reconstruction is central to answering key biological questions, such as understanding how species are related, dating species divergence events, identifying how traits evolve and even how pathogens diversify and spread (Grenfell et al., 2004).

Arising from this proliferation of data the emerging field of *phylogenomics* has been born. While phylogenetics focuses on tree reconstruction from a single or a small number of genes, phylogenomic inference involves combining information across hundreds or thousands of genes/loci. Individual gene trees often differ due to variation in the evolutionary history across genes, e.g. ancestral genetic variation can persist across speciation events, leading to *incomplete lineage sorting*. Inferring a species tree from discordant gene trees requires models that explicitly acknowledge and model this variation.

In most of classical statistical inference the goal is to estimate scalar or vector parameters, for example the mean height of primary school pupils or the coefficients in linear regression. In phylogenetics and phylogenomics the inferred objects are more complex. Trees reside in non-Euclidean spaces with combinatorial and continuous properties. Standard statistical techniques often fail when directly applied in this context, since they do not respect the intricate tree space geometry which leads to poor out-of-sample performance in supervised learning tasks.

Many aspects of the tree space can be described elegantly using the framework of *tropical geometry*. In tropical geometry, arithmetic operations are different, with addition becoming equivalent to taking the maximum, and multiplication being classical addition. In other words tropical geometry operates on the max-plus algebra. Under this arithmetic, the space of phylogenetic trees can be represented as a tropical linear space, allowing for the definition of metrics, convexity, and regression analogues that respect the tree space geometry. This perspective provides a way to extend classical statistical models such as regression, PCA, and neural networks to their tropical equivalents which conform to the geometry of trees. A key application is the assessment of convergence in Bayesian phylogenetic analysis, where current MCMC diagnostics are

often insensitive to critical differences in tree branch lengths (Aliatimis et al., 2024). Furthermore, the framework extends to machine learning, where another application considers convolutional tropical neural networks which are proved to be robust against adversarial attacks (Pasque et al., 2024).

1.2 Overview of Thesis

The remainder of this thesis is organised into five main chapters.

Chapter 2 provides a comprehensive literature review that introduces the necessary background for the work that follows in the remainder of the thesis. It begins with phylogenetic inference based on distance (i.e., measures of dissimilarity between sequences of different species), parsimony, likelihood and Bayesian posterior sampling methods, before discussing phylogenomic inference where information from multiple parts of the DNA sequence (loci) are analysed under the multispecies coalescent (MSC) model, one of the most widely estabished models of how gene lineages trace back for organisms belonging to different species. The chapter concludes with an introduction to tropical geometry and the theoretical justification for modelling phylogenetic trees as elements in a tropical space.

Chapter 3 introduces the *tropical logistic regression* (TLR) model, which adapts the classical logistic regression to the tropical setting. Standard logistic regression assumes Euclidean covariates, which is invalid for data represented by trees. Theoretical properties such as generalization bounds are derived, and the TLR method is applied to both simulated and empirical datasets. Moreover, the method is proposed as a convergence diagnostic for Bayesian phylogenetic inference in software such as `MrBayes`, where it can distinguish between independent chains that appear converged under `MrBayes`'s

convergence metric. A practical implementation of TLR is available in the `TML` CRAN package (Barnhill et al., 2024).

Chapter 4 generalizes the TLR model to generalized linear models and deep neural networks. The chapter introduces a tropical embedding that maps trees, represented in the tropical projective torus, to Euclidean space. This mapping constitutes a "tropical" layer. A *tropical neural network* (TNN) combines an initial tropical layer with subsequent classical layers. The chapter's theoretical contributions include universal approximation results for functions on the tree space and the derivation of a backpropagation rule adapted to the tropical setting. Practical implementations in TensorFlow 2 are provided, as well as empirical applications to both simulated and real datasets, such as influenza hemagglutinin sequences. A disclaimer is made that this chapter includes collaborative work with Keiji Miura and Ruriko Yoshida.

Chapter 5 shifts the focus to statistical modelling in the phylogenomic setting of multilocus sequence data. It examines how the covariance structure of pairwise distances between taxa (groups of organisms, e.g. species) can be decomposed into two components; variance arising from the multispecies coalescent process and variance arising from substitution noise in the sequences. The chapter provides derivations under a multispecies coalescent with Jukes-Cantor substitution (MSC+JC) model and identifies parameter regimes where one source of randomness dominates. These theoretical results can inform the design of more efficient inference methods as well as provide fast and robust confidence estimation procedures such as estimating support values in clades (groups of organisms descended from a common ancestor) for inferred species trees.

Chapter 6 summarizes the main contributions of my work and outlines directions

for further research.

Chapter 2

Literature Review

This chapter provides an overview of the foundations that underpin this thesis. It is divided into three main parts, reflecting the progression from classical phylogenetic inference based on single-gene data to modern phylogenomic frameworks and, finally, to tropical geometry and its application in phylogenetics.

Section 2.1 reviews the core principles of *phylogenetics*, focusing on how evolutionary relationships among taxa can be inferred from molecular sequence data. We examine the major classes of inference methods; distance-based, parsimony, likelihood, and Bayesian approaches. We highlight their assumptions, statistical properties, and computational comparisons. Section 2.2 extends this discussion to the *phylogenomics* era, where multiple independent loci across the genome are analyzed jointly. This section introduces the coalescent model as the theoretical foundation for gene genealogies, explains how incomplete lineage sorting leads to discordance among gene trees, and reviews both full-likelihood and approximate inference methods under the multispecies coalescent (MSC). Finally, Section 2.3 introduces the emerging field of *tropical geometry in phylogeny*, which provides novel algebraic and geometric tools for representing and analyzing the tree space.

2.1 Phylogenetics

The history of phylogenetics is, in many ways, the history of evolutionary biology itself. When Charles Darwin published *On the Origin of Species* in 1859, he did not use the term phylogenetics, nor did he present a detailed algorithm for reconstructing evolutionary relationships. What he offered instead was a powerful metaphor; the *tree of life*. Darwin proposed that all living things are connected by descent from common ancestors, and that the history of life is best imagined as a branching tree process. This simple image set in motion an enduring scientific challenge: if life truly is a branching tree, how can we accurately reconstruct it?

Early evolutionary biologists approached this problem using the information most readily available to them, namely morphological characters. From Darwin's conception of a "tree of life" linking all species by common descent (Darwin, 1859) to Hennig's formalization of phylogenetic systematics based on shared derived traits (Hennig, 1999), morphology provided the first systematic framework for reconstructing evolutionary relationships. For much of the late nineteenth and early twentieth centuries, phylogenetic inference was synonymous with the careful cataloguing and comparison of physical traits—the number of limbs, the structure of bones, the arrangement of floral parts. These morphological characters were encoded, compared, and arranged into hypothesized relationships among species. Yet morphology, while rich, presented significant difficulties: how to distinguish traits that are *homologous* (shared by descent) from those that are *analogous* (similar by convergent adaptation) (Patterson, 1982). For example, the wings of bats and birds perform the same function and share a superficially similar form, but they evolved independently from very different ancestral structures, and they are thus analogous rather than homologous traits. By contrast, the forelimbs of whales, bats, and humans, though adapted to swimming, flying, and grasping respectively, are homologous structures derived from the same ancestral tetrapod limb.

Similarity due to shared ancestry is hard to differentiate from similarity due to function.

The mid-century brought the rise of more quantitative approaches: numerical taxonomy and cladistics. Numerical taxonomy, or phenetics, emphasized the use of quantitative characters and objective algorithms to group organisms by overall similarity, without necessarily invoking evolutionary hypotheses (Sokal and Sneath, 1963). In contrast, cladistics, developed by Willi Hennig, proposed that evolutionary relationships should be reconstructed by grouping species according to shared derived characters, defined as synapomorphies, not simply overall similarity (Hennig, 1999). This was, in essence, an appeal to *parsimony*—a preference for explanations that minimize the number of ad hoc assumptions. It echoes *Occam's razor*; do not multiply assumptions beyond necessity. Cladistics provided a conceptual framework that still underlies phylogenetics today, even as the data and methods have changed.

The 60s and 70s witnessed a "molecular turn". The advent of protein and DNA sequencing technologies opened entirely new possibilities for phylogenetics. For the first time, scientists could compare organisms at the level of their genetic material, examining thousands of characters at once rather than relying on ambiguous features of morphology. This shift from morphological to molecular data fundamentally changed the scope and accuracy of phylogenetic inference. Sequencing provided discrete, abundant, and universally comparable data across the tree of life.

It was then that the first formal statistical models of sequence evolution were developed. The pioneering work of Jukes and Cantor (Jukes, 1969) introduced a simple Markov process for nucleotide substitution, assuming equal base frequencies and equal mutation rates among all nucleotides. Although highly simplified, the Jukes–Cantor model was revolutionary in demonstrating that molecular evolution could be described

with mathematical precision. It remains a pedagogical cornerstone today, serving both as a teaching tool and as a tractable framework for testing theoretical ideas about sequence evolution.

Building on these ideas, Felsenstein (1981) introduced the maximum likelihood (ML) framework for phylgenetic infererence, showing how substitution models could be used to evaluate the probability of sequence data under competing tree hypotheses. This work marked the beginning of the statistical era in phylogenetics, in which probabilistic models became central to tree reconstruction. These models progressively relaxed the simplifying assumptions of equal base frequencies and symmetric substitution rates, allowing patterns that more closely reflect observed molecular evolution—namely, the T92 (Tamura, 1992) and T93 (Tamura and Nei, 1993) models, and the General Time Reversible (GTR) model (Tavaré, 1986), the latter of which contains all the aforementioned models as special cases.

Likelihood methods allowed researchers to explicitly model nucleotide substitution as a stochastic process, making inference more precise and robust. However, a limitation of the ML approaches was that they naturally lead to a single best tree, without an obvious way to quantify uncertainty about the tree topology. Although standard likelihood theory provides confidence intervals for numerical parameters such as branch lengths or substitution rates, it does not extend neatly to the discrete and combinatorially vast space of tree topologies. To address this, Felsenstein (1985) introduced the bootstrap as a frequentist means of assessing support for clades, that is, sets of taxa that share a common ancestor and include all its descendants. The method involves resampling alignment columns and re-estimating trees across replicates. While bootstrap support values became widely used, they are not strictly interpretable as probabilities. This gap in the frequentist framework helped motivate the adoption of Bayesian ap-

proaches (Rannala and Yang, 1996), which treat trees themselves as random variables and summarize their posterior distribution over the space of trees. In Bayesian inference of phylogeny, uncertainty is naturally expressed in the form of posterior probabilities on clades, providing a more direct measure of confidence in inferred relationships.

In the remainder of this section, I describe the four major classes of methods used in modern phylogenetic inference. To illustrate their application, we consider a toy dataset consisting of 10-nucleotide sequences for four species, as shown in the following matrix,

$$\begin{array}{l} \text{Species A} \\ \text{Species B} \\ \text{Species C} \\ \text{Species D} \end{array} \begin{bmatrix} A & C & T & G & G & G & T & C & G & A \\ A & C & T & G & A & G & T & C & G & A \\ A & G & T & G & T & G & A & A & G & T \\ G & C & T & G & G & A & T & C & G & A \end{bmatrix} \tag{2.1.1}$$

2.1.1 Distance-based methods

Distance-based methods are among the earliest computational approaches for reconstructing phylogenetic trees. The central idea is intuitive: rather than directly searching the combinatorial space of tree topologies, one first summarizes the differences among sequences as a set of pairwise distances, and then builds a tree whose branch lengths best reproduce those distances. This approach transforms the complex problem of tree inference into a simpler problem of metric fitting.

For DNA sequences, the simplest measure of distance is the *Hamming distance*, which counts the number of sites at which two sequences differ and expresses it as a fraction of the total sequence length. Although straightforward to compute, the raw Hamming distance underestimates the true number of substitutions over evolutionary time, because multiple substitutions at the same site can mask earlier changes. To account for this, one can apply a substitution model such as the Jukes–Cantor model

(Jukes, 1969). Under this model, if the observed fraction of differences between two sequences is p, the expected number of substitutions per site d is estimated as

$$d = -\frac{3}{4} \ln \left(1 - \frac{4}{3} p \right), \tag{2.1.2}$$

which is commonly referred to as the *Jukes-Cantor distance*.

For small values of p, the first order approximation is $d = p + \mathcal{O}(p^2)$. As the values of p increase, this correction provides a better reflection of evolutionary divergence than the raw fraction p. Note that the relationship between p and d is nonlinear. If the Hamming distance between two sequences doubles, the estimated number of substitutions per site increases by more than a factor of two, because some sites will have experienced multiple substitutions that cancel each other out. Moreover, consider the extreme case of $p = 3/4$: two sequences agree at only $1/4$ of sites, which is what we would expect purely by chance if the sequences were completely unrelated. Under the Jukes–Cantor model, this corresponds to $d = \infty$, reflecting the fact that the true evolutionary separation is effectively unbounded. These properties highlight why the Jukes–Cantor correction is essential for interpreting observed differences in terms of underlying evolutionary change.

Suppose the task is to infer the phylogeny of m taxa. There are $\binom{m}{2}$ pairwise distances to be computed between all possible pairs of taxa. All these components form the distance matrix $D \in \mathbb{R}^{m \times m}$. Once the distance matrix has been computed, several algorithms can be used to reconstruct the phylogenetic tree.

One of the most classical approaches is the *Unweighted Pair Group Method with Arithmetic Mean* (UPGMA), which is a hierarchical clustering method that iteratively joins the closest clusters and updates the distance matrix (Sokal and Michener, 1958).

UPGMA assumes a constant rate of evolution across lineages (following the molecular clock hypothesis), which can be restrictive for some datasets. Another widely used approach is the *Neighbour-Joining* method (Saitou and Nei, 1987a), which constructs an unrooted tree by iteratively finding pairs of taxa that minimize the total branch length of trees. Neighbour-Joining is a computationally efficient and scalable method, and is not restricted to the molecular clock assumption.

Distance-based inferred trees, like all phylogenetic estimates, are subject to sampling error: the sequences we observe are finite realizations of an underlying stochastic process, and any given dataset may support some clades more strongly than others. To assess the robustness of inferred clades, *bootstrapping* is often applied. Introduced by Felsenstein (1985), bootstrapping involves resampling the aligned columns replacement to generate pseudo-replicate datasets. For each replicate, the distance matrix is recalculated, a tree is inferred, and the presence or absence of a specific clade is recorded. By repeating this process many times, one obtains a bootstrap *support value* for each clade, representing the proportion of replicates in which the clade appears. This procedure provides a measure of confidence for each branch.

To illustrate the use of distance-based methods, we return to the toy dataset given in (2.1.1). We first compute the pairwise Hamming distances between the four species, which correspond to the proportion of differing nucleotide sites between each pair of sequences. Then, we transform those components using Jukes-Cantor distances

from (2.1.2). This results in the following distance matrices

$$D_{\text{Hamming}} = \begin{bmatrix} 0.0 & 0.1 & 0.2 & 0.5 \\ 0.1 & 0.0 & 0.3 & 0.5 \\ 0.2 & 0.3 & 0.0 & 0.7 \\ 0.5 & 0.5 & 0.7 & 0.0 \end{bmatrix}, \quad D_{\text{JC}} = \begin{bmatrix} 0.00 & 0.11 & 0.23 & 0.82 \\ 0.11 & 0.00 & 0.38 & 0.82 \\ 0.23 & 0.38 & 0.00 & 2.03 \\ 0.82 & 0.82 & 2.03 & -0.00 \end{bmatrix}.$$

Comparing the two matrices highlights how the Jukes–Cantor correction preserves small distances (e.g. $p = 0.1$ becomes $d = 0.11$), while penalizing larger distances more strongly (e.g. $p = 0.7$ becomes $d = 2.03$). From this distance matrix we would suspect that species D is the most distant and that species A and B are probably sisters since they have the shortest distance from each other. Note however that this matrix does not naturally produce an equidistant tree, i.e. one where the distances from each leaf to the root are equal. Since D is the most distant species, we would require $D_{\text{JC}}(A, D) = D_{\text{JC}}(B, D) = D_{\text{JC}}(C, D)$, which is not the case.

Applying UPGMA yields a modified distance matrix in which the pairwise distances correspond to twice the path lengths from the species to the root:

$$D_{\text{UPGMA}} = \begin{bmatrix} 0.00 & 0.11 & 0.31 & 1.43 \\ 0.11 & 0.00 & 0.31 & 1.43 \\ 0.31 & 0.31 & 0.00 & 1.43 \\ 1.43 & 1.43 & 1.43 & 0.00 \end{bmatrix}$$

The corresponding tree, inferred under UPGMA using JC distances, is shown in Figure 2.1.1. To assess the robustness of the clades, we perform nonparametric bootstrapping, resampling sites and recomputing Hamming distances instead of JC distances to avoid infinities. Two clades are consistently recovered: the sister relationship of species A and B supported with 82% of bootstrap replicates, and the larger clade (A,B,C),

supported with 58% of replicates. Together these values suggest moderate support for the placement of A and B as closest relatives, and weaker support for grouping C with them, while species D emerges as the most distant lineage and does not form a cherry with any other taxon.

In summary, distance-based methods are computationally efficient, conceptually simple, and robust to sampling error, which has ensured the continued relevance of distance-based methods in phylogenetics.

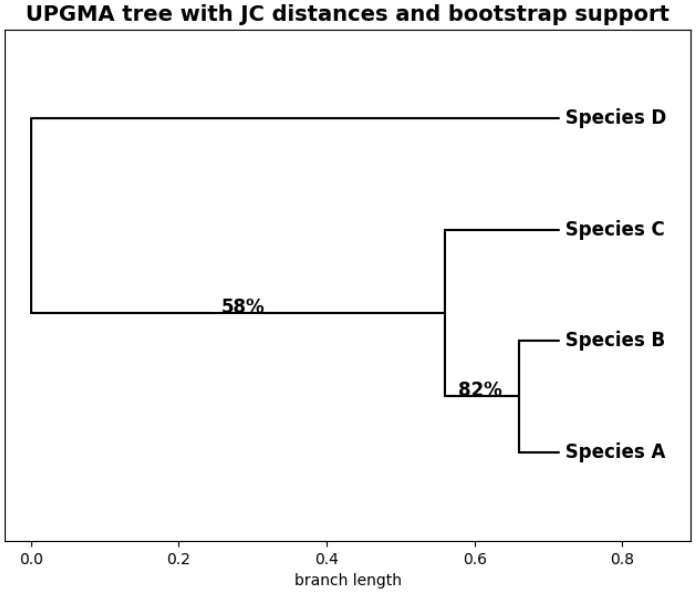


Figure 2.1.1: UPGMA tree inferred from Jukes–Cantor distances from the toy gene sequence. Bootstrap support values based on 500 replicates are shown above the branches. Species A and B are recovered as sisters with strong support, while the grouping of A,B, and C has weaker support. Species D is placed as the most divergent lineage.

2.1.2 Parsimony methods

Parsimony methods form one of the earliest systematic approaches to phylogenetic inference. The guiding principle is *Occam's razor*: the preferred phylogeny is the one that requires the fewest evolutionary changes to explain observed data. In this frame-

work, each column of a sequence alignment is treated as a discrete character analogous to morphological traits in earlier systematics. Given a candidate tree topology, one can determine the minimum number of substitutions required at that site to reconcile the character states observed at the leaves (extant species). Summing these minimal counts across all sites of the gene yields the *parsimony score* of the tree, and the most parsimonious tree is the one with the lowest total score, as first formalized in Camin and Sokal (1965).

An efficient algorithm for calculating site-wise minimal changes was presented by Fitch (Fitch, 1971). Fitch's algorithm is a bottom-up approach that assigns sets of possible ancestral states to internal tree nodes and computes the minimum number of changes required at each position. It remains foundational in parsimony-based software such as the Phylogenetic Analysis Using Parsimony (PAUP*) software (Swofford, 1998).

Subsequent developments led to heuristic searches of the tree space, since the number of possible tree topologies grows faster than exponentially with the number of taxa. Exact evaluation is impossible beyond small datasets, so practical programs rely on stepwise addition of taxa, branch-swapping procedures (such as nearest-neighbour interchange, subtree pruning, and tree bisection-reconnection), and other heuristic algorithms to explore the tree space (Felsenstein, 2004; Goloboff et al., 1999). These strategies allow parsimony methods to remain competitive even on datasets with a large number of taxa.

In applying Fitch's parsimony algorithm to our four–taxon dataset (2.1.1), we found that all 15 possible rooted trees yield the same parsimony score of 8. This outcome follows from the fact that the alignment contains no *parsimony-informative sites*. For four taxa, a site is informative only when at least two different nucleotides each occur

in at least two species, i.e. a 2–2 split. Such sites allow parsimony to favor one topology over the others. In our alignment, however, every column is either constant, exhibits a 3–1 pattern, or a 2–1–1 pattern.

Constant sites (*xxxx*) incur no changes, so the Fitch's score is 0 regardless of topology. Similarly, a 3–1 site (*xxxy*) requires one substitution, i.e. a score of 1, and a 2–1–1 site (*xxyz*) requires two substitution and has a score of 2, both scores being indendepent of tree topology. Since none of the columns match the informative 2–2 pattern, every possible topology receives the same cumulative score. Specifically, our 10 sites break down as six 3–1 sites (1 change), one 2–1–1 site (2 changes), and three constant sites (0 changes), for a total score of $6 * 1 + 1 * 2 = 8$ for all trees. Table 2.1.1 summarizes the possible patterns for 4 taxa.

Pattern type	**Example**
Constant	A A A A
3–1 split	A A A G
2–1–1 split	A A G T
2–2 split (AB | CD)	A A G G
2–2 split (AC | BD)	A G A G
2–2 split (AD | BC)	A G G A

Table 2.1.1: Examples of site patterns for four taxa. Only 2–2 splits are parsimony–informative.

Overall, parsimony remains conceptually important today as a baseline methods against which more complex frequentist and Bayesian approaches can be contrasted. However, it is also known to suffer from biases such as *long-branch attraction*, in which rapidly evolving lineages are incorrectly inferred to be closely related and are thus positively misleading (Felsenstein, 1978).

2.1.3 Frequentist inference of phylogeny

Thus far, we have discussed distance-based and parsimony methods, which are computationally efficient but limited in statistical rigor. ML methods, introduced to phylogenetics by Felsenstein (1981), place tree inference on a firm probabilistic footing. Let $\chi = (\chi^{(i)} : i \in [K])$ denote the aligned sequence data across K sites, where each site $\chi^{(i)} \in \{A, C, G, T\}^n$ represents the nucleotides observed for the n taxa at position i. Given a substitution model of sequence evolution (e.g. Jukes–Cantor), the likelihood of a branch length weighted tree T (including branch lengths) and substitution rate μ is

$$L(T, \mu) \;=\; P(\chi \mid T, \mu) \;=\; \prod_{i=1}^{K} P(\chi^{(i)} \mid T, \mu), \tag{2.1.3}$$

where P denotes the probability mass function. This product form reflects the *site independence assumption*, namely that all sites evolve identically and independently (i.i.d.) according to the same substitution process. Of course, in reality sequence sites are not strictly independent: for example, codons encode 20 amino acids from 64 possible triplets, and functional or structural constraints may induce correlations across sites. Nevertheless, the i.i.d. assumption is widely adopted in phylogenetic likelihood models, both because it renders the computations tractable and because it often provides a reasonable approximation for the statistical signal in real alignments. The maximum likelihood estimate (MLE) is then the tree topology, branch lengths and the mutation rate that maximize $L(T, \mu)$.

Under the Jukes–Cantor (Jukes, 1969) model we assume equal equilibrium base frequencies and that every ordered substitution $i \to j$ $(i \neq j)$ occurs at the same

instantaneous rate. Writing μ for the overall substitution rate, the rate matrix is

$$Q = \begin{array}{c} \\ \mathrm{A} \\ \mathrm{C} \\ \mathrm{G} \\ \mathrm{T} \end{array} \begin{array}{c} \begin{array}{cccc} \mathrm{A} & \mathrm{C} & \mathrm{G} & \mathrm{T} \end{array} \\ \begin{bmatrix} -3/4\mu & \mu/4 & \mu/4 & \mu/4 \\ \mu/4 & -3/4\mu & \mu/4 & \mu/4 \\ \mu/4 & \mu/4 & -3/4\mu & \mu/4 \\ \mu/4 & \mu/4 & \mu/4 & -3/4\mu \end{bmatrix} \end{array}.$$

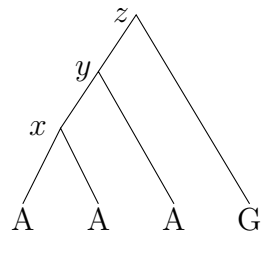
The transition probability matrix $P(t)$ giving the probability that a site in state i at time 0 is in state j after time t has entries

$$P_{ij}(t) \;=\; \begin{cases} \frac{1}{4} + \frac{3}{4}\, e^{-\mu t}, & i = j, \\ \\ \frac{1}{4}\left(1 - e^{-\mu t}\right), & i \neq j. \end{cases}$$

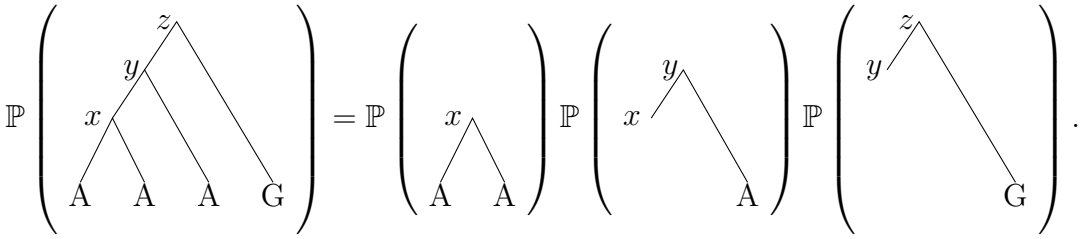
Equivalently, for the expected observed fraction of differences p between two sequences, the Jukes–Cantor correction gives the expected number of substitutions per site using equation (2.1.2), where the Jukes–Cantor distance $d = \frac{3}{4}\mu t$ is the expected number of changes per site and $p = P_{ij}, i \neq j$.

To compute the likelihood of a site across the entire tree, one sums over all possible assignments of nucleotides to internal nodes. For example, the likelihood of the first site of (2.1.1) is

$$f(\chi^{(1)}|T) = \sum_{x,y,z \in \{A,C,T,G\}} \mathbb{P}\left(\right., \tag{2.1.4}$$



where



All three probabilities of the equation above are obtained from the same continuous-time Markov chain framework. For any three nodes i, j, k with j the most recent common ancestor of i and k, let τ_{ij} and τ_{jk} denote the branch lengths from j to i and from j to k, respectively. The process evolves according to the rate matrix Q, so that if we set $X_0 = \chi_i$ and $X_{\tau_{ij}+\tau_{jk}} = \chi_k$, where χ_i and χ_k are the observed nucleotides at the leaves, then the distribution of the intermediate state $X_{\tau_{ij}}$ gives the probability of nucleotide assignments at the internal node j. This computation corresponds to the three factors in the equation above. An even simpler way of computing the term is above is as a product of the six tree edges.

Direct enumeration of all internal nodes is exponential in the number of taxa, but Felsenstein's pruning algorithm (Felsenstein, 1981) allows efficient computation: starting from the leaves, one recursively computes partial likelihoods at each internal node by combining child likelihoods weighted by the transition probabilities along the branches. The site likelihood is obtained at the root, and the total likelihood is the product across all sites in the alignment (or sum of log-likelihoods).

Modern ML analyses are typically performed with specialized software packages that implement pruning and numerical optimization of branch lengths and substitution model parameters. Widely used tools include `IQ-TREE` (Minh et al., 2020), `RAxML` (Stamatakis, 2014), and `PhyML` (Guindon and Gascuel, 2003). These programs explore tree space heuristically (since the number of possible trees grows super-exponentially with

the number of taxa) and return the tree with the highest likelihood under the specified model. Bootstrapping or approximate likelihood ratio tests can then be applied to assess support for clades. ML trees are among the most widely used approaches in modern phylogenetics, being both statistically efficient under realistic models and relatively computationally tractable with these algorithms.

For our toy dataset (2.1.1), the ML analysis under the Jukes–Cantor model yields the unrooted tree BC|AD, (meaning that B, C is split from A, D) with an internal branch length close to zero and bootstrap support of only 36%. The remaining 64% of bootstrap replicates were split between the other two possible bipartitions, indicating that none of the three alternative unrooted topologies is strongly favoured by the data. This indicates that the data provide very limited evidence for resolving the relationship among the four species. Indeed, this is consistent with our earlier parsimony analysis, which found that all fifteen rooted trees had the same parsimony score. As explained above, this occurs because the alignment contains no parsimony-informative sites. In such cases, both parsimony and likelihood methods agree in reporting little to no resolution: parsimony ties all trees, and likelihood produces a tree with extremely low internal support. Thus, while ML provides a statistical framework and formal measures of uncertainty (e.g. bootstrap values), the lack of phylogenetic signal in the data fundamentally limits inference under either approach.

2.1.4 Bayesian inference of phylogeny

Bayesian phylogenetics treats trees and evolutionary parameters as random variables and seeks to characterize their joint posterior distribution given the observed sequence alignment. By Bayes' theorem, the posterior probability of a tree T (topology and

branch lengths) and substitution rate μ is

$$P(T, \mu \mid \chi) \;\propto\; P(\chi \mid T, \mu)\, P(T)\, P(\mu),$$

where $P(\chi \mid T, \mu)$ is the likelihood of the data under the chosen substitution model, and $P(T), P(\mu)$ are prior distributions on the tree and substitution rate, respectively (Mau et al., 1999; Huelsenbeck et al., 2001). Unlike ML, which returns a single "best" tree, Bayesian inference produces a posterior distribution over trees. This allows direct probability statements about clades, in contrast to frequentist support values such as bootstraps.

Because the posterior distribution over trees and model parameters is analytically intractable, Bayesian phylogenetics relies on Markov chain Monte Carlo (MCMC) methods to approximate the posterior distribution by sampling (Metropolis et al., 1953; Hastings, 1970). After a burn-in period, the MCMC chain produces an ensemble of trees and parameter values proportional to their posterior probability. These samples are typically summarized as a maximum clade credibility tree or a consensus tree annotated with posterior clade probabilities (Drummond and Rambaut, 2007; Ronquist et al., 2012).

Bayesian approaches have several advantages: they incorporate model uncertainty explicitly through priors, yield posterior probabilities that are directly interpretable, and allow joint estimation of phylogenies with other evolutionary parameters (e.g. divergence times, population sizes). However, they are also computationally demanding and sensitive to prior specification, requiring careful diagnostics to ensure adequate mixing and convergence of the MCMC chain. Despite these challenges, Bayesian inference has become indispensable in modern phylogenetics, implemented in widely used software such as `MrBayes` (Ronquist et al., 2012), `BEAST` (Drummond and Rambaut,

2007), `RevBayes` (Höhna et al., 2016), `PhyloBayes` (Lartillot et al., 2009), and `BPP` (Yang, 2015).

Assessing convergence in Bayesian phylogenetic inference requires determining whether an MCMC chain has reached its stationary distribution and, subsequently, whether it has adequately sampled from the posterior distribution. Unlike classical MCMC applied to simple numeric or vector-valued parameters, phylogenetic MCMC samples complex, structured objects—trees with both discrete topologies and continuous branch lengths. Similar challenges arise in other domains that involve sampling over combinatorial or graph-structured spaces, such as Bayesian network inference in molecular biology (Friedman et al., 2000; Madigan et al., 1995) and MCMC-based network modelling in the social and biological sciences (Snijders, 2002), although the phylogenetic literature has largely developed its own domain-specific move sets and convergence diagnostics. In classical MCMC, convergence diagnostics such as the effective sample size (ESS), autocorrelation, or the Gelman-Rubin statistic (Gelman and Rubin, 1992) can be applied directly to numeric parameters, providing clear, quantitative criteria. In phylogenetics, however, the posterior includes both discrete topologies and continuous parameters, and these components can converge at different rates. As a result, convergence assessment in Bayesian phylogenetics requires multiple, often complementary diagnostics, because no single classical MCMC criterion can simultaneously capture convergence of both tree topology and branch lengths.

`MrBayes` primarily relies on the Average Standard Deviation of Split Frequencies (ASDSF), which compares the frequencies of bipartitions (splits) between independent chains. Low ASDSF values indicate that chains have converged to similar posterior distributions of tree topologies. While ASDSF provides a quantitative measure for topological convergence, it does not assess convergence of continuous parameters such

as branch lengths or substitution rates. Other software such as `RevBayes` offers similar split-frequency-based checks for topology, but continuous parameters must still be evaluated separately using trace plots or effective sample size (ESS) calculations.

`BEAST`, in contrast, provides a more flexible but user-guided approach. Convergence is typically assessed using the companion program Tracer (Rambaut et al., 2018), which allows users to visualize trace plots of continuous parameters for stationarity and calculate effective sample sizes (ESS) to evaluate the number of effectively independent samples. The process therefore depends on the user integrating multiple diagnostics to judge whether the chain has reached stationarity.

Taken together, these examples highlight a common limitation: current Bayesian phylogenetic software packages assess convergence of tree topology and continuous parameters separately, and no single criterion evaluates whether entire trees—including both their topologies and branch lengths—have converged. This division leaves convergence assessment fragmented and, to some extent, dependent on subjective user judgment. Our work (Section 3.5.3 of Chapter 3) addresses this gap by introducing a classification-based method that evaluates convergence through the stability of tree pairwise distances, thereby integrating both topology and branch lengths into a single diagnostic.

Bayesian inference implemented in `MrBayes` was run on the alignment (2.1.1) under the JC69 model with equal base frequencies, using 1000000 MCMC generations sampled every 100 generations, four Metropolis-coupled chains, a temperature of 0.2, and a burn-in of 2,500 samples. The posterior distribution of trees identified the split AD|BC as the maximum a posteriori (MAP) topology, with 45% posterior probability. Alternative splits AC|BD and AB|CD were also present, with posterior probabilities of

32% and 23%, respectively, and no star trees (trees where all pair leaf distances are the same) were sampled. Posterior probabilities are generally higher than bootstrap values because they quantify the probability of a split given the model and data, rather than the stability of a split under resampling. Across all methods, the results highlight the limited phylogenetic signal in this dataset, with only weak to moderate support for any particular resolution among the four taxa.

For comparison, the ML tree inferred by IQ-TREE2 under the JC69 model recovered the same AD|BC topology as the MAP tree, but with weaker ultrafast bootstrap support (36%) compared to the 45% posterior probability from MrBayes. This difference arises because bootstrap proportions measure the stability of a split under resampling of the alignment (i.e., how often the split appears across pseudo-replicates).

2.2 Phylogenomics

Phylogenomic datasets differ from single-gene studies because they sample many independent genomic regions (loci) across the genome. The key challenge for species tree inference is that, as we will see in this subsection, different loci can have conflicting evolutionary histories. Here, we discuss why the problem is different, how we can model phylogenomic inference and review the different approaches suggested in the literature.

By *gene* or *locus* we mean a contiguous genomic region that is treated as a unit for phylogenetic inference. Genes are the coding regions of the genome that contain the information required to produce functional molecules, typically proteins. In contrast, the remaining part of the genome comprises non-coding regions that are often discarded.

Each gene can exist in multiple alternative versions, known as *alleles*, which differ in their nucleotide sequences due to mutation. These allelic variants may be shared among populations or species as a result of ancestral polymorphism—that is, genetic

variation that was already present in a common ancestor and persisted through successive speciation events. Alleles sampled from extant species trace back through time along a *gene tree* that describes the branching relationships among copies of the gene and the times at which they coalesce to a common ancestor.

2.2.1 The Coalescent Model for Genealogical Lineages

In a coalescent framework, lineages are traced backward in time from sampled alleles to their common ancestry. Each allele in the present generation was inherited from one of the two parental gene copies in the previous generation, and this process is repeated as we move further back in time. At some point, two lineages may trace back to the same ancestral gene copy carried by a single individual in an ancestral population. *Coalescence* represents the point in the past where both alleles share a single common ancestral copy of the gene. In a gene tree drawn forward in time, this corresponds to a branching event, whereas when viewed backward in time, it is a merging of lineages. Thus, the coalescent process provides a stochastic model for how allelic lineages from different individuals or species ultimately converge on their most recent common ancestor.

Tracing lineages backward in time, each coalescent event reduces the number of distinct ancestral lineages by one, until eventually all sampled alleles converge on a single common ancestor of that gene. The times at which these coalescent events occur depend on factors such as *effective population size* [1]: in larger populations, coalescence tends to occur deeper in time because ancestral alleles persist for longer before coalescing. Thus, the topology and branch lengths of a gene tree encapsulate the stochastic history of how alleles from different species or populations trace back to their shared ancestry.

[1]The effective population size mostly reflects the mating population and is therefore lower than the total census population. It is often smaller than the breeding population due to factors such as unequal reproductive success, skewed sex ratios, and population size fluctuations that reduce genetic diversity.

Formally, the *coalescent rate* between any pair of lineages is inversely proportional to the effective population size. In a diploid population (where each individual carries two gene copies, one inherited from each parent) with effective size N_e, there are $2N_e$ gene copies, and the probability that two randomly chosen alleles share a common parent copy in the previous generation is $\lambda \coloneqq 1/(2N_e)$. Consequently, the number of generations required for two contemporary alleles to coalesce is geometrically distributed with probability of success λ. For sufficiently large population sizes, this discrete distribution can be approximated by a continuous exponential distribution with rate λ. In both cases, the expected time to the most recent common ancestor of two alleles is $2N_e$ generations. The *Kingman coalescent* (Kingman, 1982) arises as the continuous-time limit of this process as $N_e \to \infty$, providing a stochastic model in which coalescent events occur at exponentially distributed waiting times, and only pairwise mergers of lineages are allowed.

Specifically, when there are n ancestral lineages present, any pair of them may coalesce next. Under the Kingman coalescent, each pair of lineages coalesces independently at rate $\lambda = 1/(2N_e)$, so the overall rate of the next coalescent event is the sum of the pairwise rates, $\Lambda_n \coloneqq \binom{n}{2}\lambda$.[2] Accordingly, the waiting time T_n until the next coalescent event is exponentially distributed with rate Λ_n. Because all lineages are exchangeable under the assumptions of random mating and neutrality, each of the $\binom{n}{2}$ possible pairs is equally likely to be the one that coalesces next, with probability $1/\binom{n}{2}$. This property of exchangeability and the exponential waiting-time distribution together fully characterize the Kingman coalescent process.

After this coalescence event, there are $n-1$ lineages, and the process continues until only a single lineage remains, representing the most recent common ancestor (MRCA) of the sample. The total time to the MRCA, or *height* of the gene tree, can be expressed

[2]The minimum of independent Exponential distributions with rates $\lambda_i : i \in [N]$ is Exp $\left(\sum_{i=1}^{N} \lambda_i\right)$.

as $\sum_{i=2}^{n} T_i$ whose mean asymptotically converges (from below) to $4N_e$, since

$$E[T_{\text{MRCA}}] = \sum_{k=2}^{n} \mathbb{E}(T_k) = \sum_{k=2}^{n} \frac{1}{\Lambda_k} = \sum_{k=2}^{n} \frac{2N_e}{\binom{k}{2}} = 4N_e \left(1 - \frac{1}{n}\right) \to 4N_e \quad \text{as } n \to \infty,$$

or twice the expected length of the last coalescent interval. This illustrates that the most ancient branches of the gene tree near the root tend to be much longer than the more recent branches, reflecting the slower coalescent rate when fewer lineages remain.

2.2.2 Species Trees and Incomplete Lineage Sorting

The Kingman coalescent provides a stochastic description of how alleles sampled from a *single, panmictic population* (one in which all individuals mate randomly) trace back to their common ancestor. However, it cannot be directly applied to samples drawn from multiple distinct species, because lineages from different species are not free to coalesce at arbitrary times. If two organisms belong to different species, their lineages can only coalesce *after* their species share a common ancestral population—that is, after the corresponding *speciation event*. Formally, if g_{xy} denotes the time to coalescence between alleles sampled from species x and y, and S_{xy} the divergence time between those species, then $g_{xy} = S_{xy} + \text{Exp}(\lambda)$, where the exponential waiting time with rate $\lambda = 1/(2N_e)$ reflects the coalescent process within the ancestral population. Hence, genealogical relationships among alleles are constrained by the temporal structure of the species tree, with coalescence events only permitted within the ancestral populations that existed prior to the corresponding speciation events.

A *species tree* describes the branching pattern of species divergences through time. Each branch of the species tree represents a population that evolves for some duration and may contain multiple ancestral lineages of sampled alleles. Within each branch, lineages evolve according to a standard Kingman coalescent process, potentially coalescing with one another before reaching the ancestral population at the next higher

node of the species tree. At speciation events, lineages are partitioned into descendant populations; when tracing backward in time, those lineages enter the ancestral population at the corresponding divergence time, where they may eventually coalesce. Since coalescent events within different loci are independent given the species tree, different genes can yield distinct genealogical topologies (gene tree discordance) even under a model of neutral evolution and without gene flow. Such discordance is a natural consequence of the stochastic nature of lineage sorting in ancestral populations, a process known as *incomplete lineage sorting* (ILS).

In Figure 2.2.1, the species tree has topology $((A, B), C)$, but the embedded gene genealogies differ from the species relationships. The red gene tree reflects the true species topology, while the blue gene tree represents a discordant genealogy in which alleles from species B and C coalesce first. This discordance occurs because ancestral allelic lineages failed to coalesce before the divergence τ_{AB} of species A and B, leaving the ancestral polymorphism that persisted across speciation events. As a result, the order of coalescent events among lineages does not always match the order of speciation events, producing a gene tree that conflicts with the species tree even under neutral evolution and without gene flow.

Moving backward in time, the lineages sampled from species A and B enter their common ancestral population at time τ_{AB} and may coalesce before reaching the next ancestral population at time τ_{ABC}. Since the waiting time until this coalescence event is exponentially distributed, the probability that the two lineages fail to coalesce before the deeper speciation event at τ_{ABC} is e^{-T}, where $T = (\tau_{ABC} - \tau_{AB})/(2N_e)$ is the length of the internal branch of the species tree measured in coalescent units. Conditioned on this event of no prior coalescence, all three possible rooted gene tree topologies for the taxa A, B, C are equally probable. It follows that the total probability of discordance is $\frac{2}{3}e^{-T}$, which can be substantial when the internal branch of the species tree is short. Nonetheless, the concordant topology remains the most probable overall.

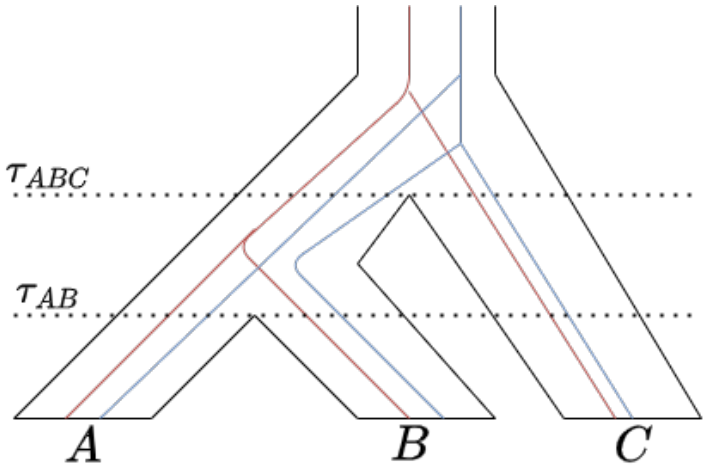


Figure 2.2.1: Illustration of gene tree discordance. The black lines depict the species tree with topology $((A, B), C)$, where τ_{AB} and τ_{ABC} denote the species divergence times and the width of the species tree represents the effective population size of the ancestral populations. Two gene genealogies are shown within the species tree: the red gene tree matches the species topology, whereas the blue tree with topology $(A, (B, C))$ is discordant, grouping B with C. This discordance occurs because ancestral allelic lineages failed to coalesce before the divergence τ_{AB} of species A and B, leaving the ancestral polymorphism that persisted across speciation events. Note that $(B, (A, C))$ is another possible gene tree topology, equiprobable to $(A, (B, C))$, but which would be good for visualization here.

2.2.3 Computational challenges and naive approaches

Because different loci are independent realizations of the coalescent process, a phylogenomic dataset produces a mixture of gene trees: some loci will have genealogies that match the species tree (red in the figure), while others will be discordant (blue in the figure). The task of phylogenomics is therefore to infer the underlying species tree S from a set of observed (or estimated) gene trees $G_1, \ldots, G_m$ or from the sequence data that generated them. This task is made harder when internal branches are short (small T), since the fraction of discordant loci increases and the signal for the species branching order is correspondingly weakened. The difficulty increases rapidly with the number of taxa because the space of possible rooted bifurcating topologies grows super-exponentially. The number of distinct rooted, fully resolved (binary) trees for n labeled taxa is $(2n-3)!! := 1 \cdot 3 \cdot 5 \cdots (2n-3)$, so even moderate values of n give

astronomically many candidate species trees to consider. This combinatorial explosion exacerbates both the statistical and computational challenges of species-tree inference from genome-scale data.

A straightforward strategy adopted in many early phylogenomic studies was to concatenate alignments from multiple loci into a single "supermatrix" and then infer a single tree using maximum-likelihood or Bayesian inference under standard substitution models. This approach is appealing because it leverages all available characters in a single analysis, permits the use of mature ML/Bayesian software, and often yields apparently well-supported topologies with high bootstrap or posterior probabilities. However, concatenation implicitly assumes that all loci share a single underlying genealogy. When gene-tree discordance is present, this assumption is violated: rather than averaging over heterogeneous histories, the concatenation procedure forces the data to fit a single tree, potentially leading to strongly supported but incorrect inferences. Indeed, under incomplete lineage sorting (ILS), concatenation has been shown to be statistically inconsistent—that is, it can converge on the wrong species tree with increasing confidence as more loci are added (Roch and Steel, 2015; Roch et al., 2019; Mendes and Hahn, 2018).

Another simplistic approach is to estimate a species tree by majority rule, selecting the topology that occurs most frequently among inferred gene trees (the *plurality* or *majority-vote* method). This strategy appears intuitive: as shown in the three-species case under constant effective population size, the concordant gene-tree topology is indeed the most probable. However, this property does not generalize. There exist regions of parameter space—known as the *anomaly zone*—where the most frequent gene-tree topology differs from the true species topology (Degnan and Rosenberg, 2006). In such cases, the majority-vote estimate will converge to the wrong tree as the number of loci increases.

2.2.4 The Multispecies Coalescent Model

To overcome the limitations of these simplistic approaches, more robust strategies explicitly model the heterogeneity among loci rather than forcing them into a single genealogy. The *multispecies coalescent* (MSC; Rannala and Yang (2003), Degnan and Rosenberg (2009)) provides such a framework by describing how gene trees are probabilistically generated within the branches of a species tree. As illustrated in Figure 2.2.2, the MSC formalizes phylogenomic inference as a hierarchical process: the species tree S specifies the topology, divergence times, and effective population sizes of ancestral populations, and is assumed to be an equidistant (ultrametric) tree in which branch lengths correspond to time rather than expected number of substitutions. Meanwhile, each gene tree $G_1, G_2, \ldots, G_m$ represents a latent genealogical history for one of m independent loci. Conditional on the species tree, each gene tree is realized through nested Kingman coalescent processes acting within the ancestral populations defined by S, giving rise to the probability distribution $P(G_i \mid S, \Theta)$, where Θ denotes the population parameters, such as the effective population size and the mutation rate. The observed sequence alignments $\chi_1, \chi_2, \ldots, \chi_m$ are then modeled as evolving along their respective gene trees according to standard substitution processes, with likelihood $P(\chi_i \mid G_i)$. In conclusion, this hierarchical structure encapsulates two levels of randomness: (1) the substitution process generating sequence data on a fixed gene tree; and (2) the coalescent process generating gene trees within the species tree.

At the lowest level of the hierarchy, the observed molecular data $\chi_i \in \{A, C, T, G\}^{n \times K_i}$ for locus i, consisting of K_i aligned nucleotide sites, are assumed to evolve along the corresponding gene tree G_i under a standard substitution model, such as JC69 or GTR. Earlier in Equations (2.1.3) and (2.1.4) of Section 2.1.3 we have shown how to compute the lower level of randomness $P(\chi_i \mid G_i)$. The overall likelihood of the data under the

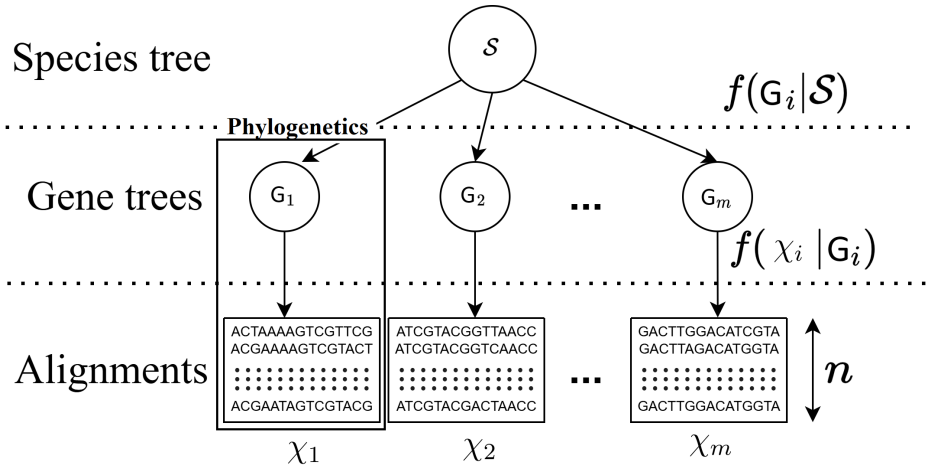


Figure 2.2.2: In phylogenetics, the goal is to reconstruct gene trees G_i for $i \in [m]$ (intermediate level) from m gene alignments (bottom level) through a nucleotide substitution model $f(\chi|G)$. In phylogenomics, the aim is to infer the species tree $\mathcal{S}$ (top level) from which these gene trees arose through the MSC model $f(G|\mathcal{S})$. This is a slightly adapted version of Figure 5.1.1 in Section 5.1.

MSC model thus integrates over the uncertainty in the unobserved gene trees:

$$L(S, \Theta) = P(\chi_1, \ldots, \chi_m \mid S, \Theta) = \prod_{i=1}^{m} \int P(\chi_i \mid G_i)\, P(G_i \mid S, \Theta)\, dG_i. \tag{2.2.1}$$

In a Bayesian framework, inference proceeds by placing priors on the species-tree parameters S and Θ (and potentially on substitution-model parameters), and computing the joint posterior distribution given the observed data,

$$P(S, \Theta, G_1, \ldots, G_m \mid \chi_1, \ldots, \chi_m) \propto \left[\prod_{i=1}^{m} P(\chi_i \mid G_i)\, P(G_i \mid S, \Theta)\right] P(S)\, P(\Theta). \tag{2.2.2}$$

Integration over the latent gene trees yields the marginal posterior of the species tree:

$$P(S, \Theta \mid \chi_1, \ldots, \chi_m) \propto P(S)\, P(\Theta) L(S, \Theta). \tag{2.2.3}$$

Evaluating the likelihood (2.2.1) or the posterior (2.2.3) requires integrating over the space of all possible gene trees G_i.

This integration is particularly challenging because the space of gene trees is both *discrete*—encompassing all possible tree topologies—and *continuous*—involving the branch-length and coalescent-time parameters associated with each topology. As mentioned earlier, the dimensionality of this space grows super-exponentially with the number of taxa, making exact analytical integration intractable even for moderately sized datasets. Numerical evaluation is therefore extremely difficult: one must, in principle, integrate a high-dimensional, multimodal function over a complex discrete–continuous domain. Consequently, practical inference under the multispecies coalescent relies on approximation or sampling-based strategies that effectively explore this tree space.

2.2.5 Inference methods

Among methods that implement the multispecies coalescent, the most statistically rigorous but computationally demanding are full-likelihood Bayesian approaches such as `StarBEAST` and its successor `StarBEAST2` (Heled and Drummond, 2009; Bouckaert et al., 2014). These methods perform joint Bayesian inference on the entire hierarchical model, sampling from the joint posterior distribution (2.2.2) using a Metropolis–Hastings Markov chain Monte Carlo (MCMC) algorithm. By explicitly modelling the probability of each gene tree given the species tree $P(G_i \mid S, \Theta)$ and the sequence likelihood $P(\chi_i \mid G_i)$, proposing changes, and accepting or rejecting according to the standard Metropolis–Hastings acceptance criterion, `StarBEAST2` simultaneously estimates gene trees, species-tree topology, divergence times, and population sizes. This full Bayesian treatment properly integrates over the enormous discrete–continuous tree space, but it is computationally expensive and scales poorly with the number of loci and taxa. Specifically, while no formal complexity bound is provided, simulation experiments for `StarBEAST` indicate that the computational cost scales roughly as $O\left(m^{2.81}\right)$ where m is the number of loci (Ogilvie et al., 2016).

A related approach, `BEST` (Bayesian Estimation of Species Trees; Liu and Pearl (2007)), extends the `MrBayes` framework to species-tree inference by incorporating the multispecies coalescent as a prior on gene trees. Because full integration over gene-tree space is infeasible, `BEST` employs several heuristics to improve efficiency—such as conditional updating of gene trees and shared tree-topology proposals across loci—while still performing approximate MCMC sampling of the joint posterior. Although somewhat faster than `StarBEAST`, it remains limited to relatively small datasets.

The following methods can be viewed as *summary-coalescent* approximations that avoid sampling over the joint gene-tree space by first estimating individual gene trees and then summarizing them into a species tree. In other words, the input to these methods can be viewed as a set of *estimated gene trees*, each obtained from independent phylogenetic inference on the corresponding locus alignment. `MP-EST` (Maximum Pseudo-likelihood Estimation of Species Trees; Liu et al. (2010)) infers the species tree that maximizes a pseudo-likelihood function derived from the distribution of rooted triplets in the input gene trees, effectively providing a maximum-likelihood analogue under the coalescent. `ASTRAL` (Mirarab et al., 2014) instead relies on quartets—subtrees of four taxa—and identifies the species tree that maximizes the number of induced gene-tree quartets that are concordant with it. Because quartet relationships can be computed efficiently, `ASTRAL` achieves a worst-case running time of $O\big((n\,m)^{2.726}\big)$, where n is the number of species and m the number of genes (Zhang et al., 2018a). Empirically, on typical data sets the running time grows roughly as $O\big((n\,m)^{2}\big)$, making it feasible for genome-scale analyses.

Analyzed in greater detail in Chapter 5, simpler and faster still are methods such as `STEAC` (Species Tree Estimation using Average Coalescence times; Liu et al. 2009) and `GLASS` (Global LAteSt Split; Mossel and Roch 2008), which summarize information across loci using pairwise coalescent times. `STEAC` averages estimated coalescent times between taxa across loci, whereas `GLASS` uses the minimum coalescent time as

an estimator of the species divergence time. These approaches are computationally trivial compared to Bayesian or pseudo-likelihood methods and are therefore particularly useful when the number of loci is extremely large, rendering full-likelihood or even summary-likelihood inference computationally infeasible. Computationally, calculating the pairwise coalescence summaries across m genes for n species requires $O(m \cdot n^2)$ operations, and constructing the species tree from the resulting distance matrix via optimized UPGMA adds only $O(n^2)$ operations, giving an overall complexity of $O(m \cdot n^2)$.

Both `STEAC` and `GLASS` operate on estimated gene trees, treating them as input to produce an inferred species tree. In contrast, the `METAL` method (Dasarathy et al., 2014) dispenses with intermediate gene-tree estimation and instead infers the species tree directly using distance–based estimates (Section 2.1.1) computed on the concatenated gene sequences. Importantly, while concatenation of loci can lead to statistical inconsistency in likelihood-based phylogenetic inference under the multispecies coalescent, it does not introduce such inconsistency for distance-based methods. `METAL` has been formally proven to be a statistically consistent estimator of the true species tree topology under both the molecular clock assumption (Theorem 2) and without it (Theorem 4) (Dasarathy et al., 2014). Computationally, `METAL` shares the same complexity as `GLASS` and `STEAC`.

Although conceptually straightforward, these distance-based estimators exhibit distinct performance profiles depending on the level of stochasticity in the data. `GLASS`, which emphasizes the earliest (minimum) coalescent times, can recover the correct species topology efficiently when gene-tree estimates are highly reliable and substitutional noise is low. In contrast, methods that average across loci, such as `STEAC` and `METAL`, tend to be more robust under higher gene-tree estimation variance or when additional sources of heterogeneity beyond incomplete lineage sorting (ILS) are present. While `GLASS` may converge more rapidly in low-noise regimes, its accuracy can deteriorate sharply when sequence alignments are short or mutation rates are high, conditions

under which averaging approaches become preferable.

From a theoretical standpoint, these three methods exhibit distinct asymptotic properties. `METAL` requires only $\mathcal{O}(f^{-2})$ total sites to recover the correct species tree topology with high probability, where f denotes the shortest internal branch length in the species tree. In comparison, `GLASS` and `STEAC` require $\mathcal{O}(f^{-3})$ and $\mathcal{O}(f^{-4})$ total sites, respectively (Dasarathy et al., 2014). Nevertheless, `GLASS` can outperform `METAL` in regimes with long sequences and low substitutional noise, owing to its minimum-distance criterion, which becomes increasingly accurate as mutational variance decreases.

Despite their computational efficiency and conceptual simplicity, the behavior of distance-based methods under different sources of stochasticity remains only partially understood. Previous work has identified limitations of early approaches such as `GLASS` and proposed refinements (e.g., `iGLASS`; Jewett and Rosenberg (2012)) to mitigate sensitivity to mutational noise. Other studies have examined the relative effects of gene-tree variance and substitutional error through simulation-based analyses (Huang et al., 2010). However, what remains unclear is how these two sources of uncertainty—coalescent variation and substitutional noise—interact across parameter regimes, and under which conditions different distance-based estimators (`STEAC`, `METAL`, `GLASS`) will be most reliable. Because the relative magnitudes of these error components depend on mutation rate, sequence length, and tree height, a unified analytical framework is needed to quantify how total uncertainty propagates through distance-based estimators.

This gap is addressed in Chapter 5, which develops a quantitative framework for decomposing total variance in distance-based species-tree reconstruction into coalescent and substitutional components. By deriving the covariance matrices corresponding to each source of randomness, the analysis provides insight into how uncertainty propagates across hierarchical levels of the inference process. Our results show that `METAL` outperforms `GLASS` when substitutional variance dominates, and conversely, that `GLASS` becomes preferable when coalescent variance is the major contributor to total

uncertainty. This decomposition further enables informed study-design decisions: when coalescent variance dominates, increasing the number of loci is most effective, whereas longer alignments (potentially through imputation or deeper sequencing) are preferable when substitutional noise is the limiting factor.

In summary, phylogenomics extends classical phylogenetics by embedding gene-tree variation within the broader stochastic framework of the multispecies coalescent. This hierarchical perspective clarifies why gene trees may conflict, how species trees can be statistically identified despite this heterogeneity, and why different inference strategies—ranging from full-likelihood Bayesian methods to lightweight distance-based estimators—represent trade-offs between realism and computational feasibility. Yet, despite major methodological progress, the interplay between coalescent and substitutional variance remains a key source of uncertainty in phylogenomic inference. Chapter 5 addresses this issue by developing a formal variance decomposition framework that quantifies how these two stochastic processes jointly shape the accuracy of distance-based species-tree estimators.

2.3 Tropical Geometry in Phylogeny

Tropical geometry is a relatively new field in mathematics, more specifically in combinatorics and algebraic geometry. It is based on the max-plus algebra, a different arithmetic system where the sum of two numbers is their maximum and their product is their sum in regular arithmetic. In this tropical algebra, polynomials are piecewise-linear functions [3]. The field is based on the foundational work of Imre Simon on tropical semirings (Simon, 1994, 1988). The term 'tropical' was attributed to him by French mathematicians, who viewed his country, Brazil, as a tropical place.

Tropical geometry has found many applications in recent years. Finding solutions

[3] more specifically, polynomials are envelopes of linear functions

to systems of polynomial equations efficiently has wide-ranging applications in the natural sciences. This is one of the most standard applications of tropical geometry, but in recent years there have been many others that are more unexpected. In economics, tropical geometry underpins the theory of product-mix auctions, as established in Baldwin and Klemperer (2019). These auctions were introduced by the Bank of England after the 2008 financial crisis to provide liquidity to the UK financial system. In operations research, performance analysis of emergency call centres has been conducted using tropical polynomials (Allamigeon et al., 2015). In phylogenetics, statistical models that use tropical geometry have also been employed. Chapters 3 and 4 focus on this application.

2.3.1 Tropical Arithmetic

In tropical geometry, addition and multiplication are different than regular arithmetic. Throughout the thesis, tropical arithmetic operations are performed in the max-plus tropical semiring $(\mathbb{R}\cup\{-\infty\}, \oplus, \odot)$ as defined in Pin (1998). Tropical arithmetic, metrics, and convexity form the geometric foundation necessary for understanding Chapters 3 and 4.

Definition 2.3.1 (Tropical Arithmetic Operations)**.** *In the tropical semiring, the basic tropical arithmetic operations of addition and multiplication are defined as:*

$$a \oplus b := \max\{a, b\}, \qquad a \odot b := a + b, \qquad \textit{where } a, b \in \mathbb{R} \cup \{-\infty\}.^{4}$$

Remark 1 (Tropical subtraction)**:** Let $\ominus x$ be the additive inverse of $x \in \mathbb{R}$. Then,

[4]The element $-\infty$ ought to be included as it is the identity element of tropical addition and the absorbing element of tropical multiplication with zero being the multiplicative identity. Clearly, associative and commutative laws of addition and multiplication hold in max-plus algebra. The distributive law $x \odot (y \oplus z) = (x \odot y) \oplus (x \odot z)$ also holds, being the equivalent of the identity $x + \min(y, z) = \min(x + y, x + z)$ in regular arithmetic. It is noted that tropical multiplication takes precedence over tropical addition when they both occur in the same expression.

$x \oplus (\ominus x) = -\infty$, since $-\infty$ is the addition identity. However,

$$x \oplus (\ominus x) = \max(x, \ominus x) \geq x > -\infty.$$

It is concluded that there is no tropical subtraction.

Remark 2 (Tropical exponentiation)**:** The n^{th} power of x is denoted as

$$x^{\odot n} = \underbrace{x \odot \cdots \odot x}_{n \text{ times}} = nx.$$

This result can be generalised from $n \in \mathbb{N}$ to $n \in \mathbb{R}$; tropical exponentiation is the same as regular multiplication.

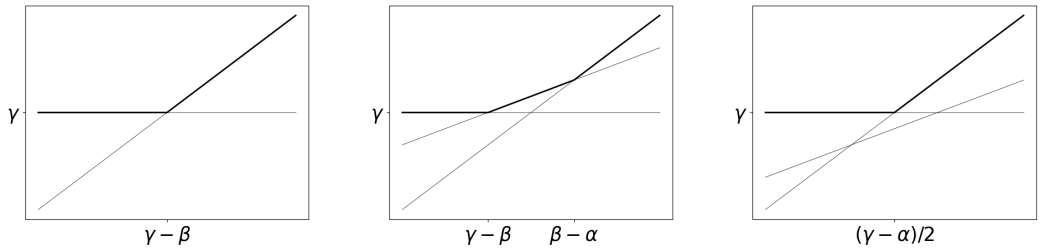
Example 1 (Tropical lines and quadratics)**:** A tropical line $\beta \odot x \oplus \gamma = \max(\beta + x, \gamma)$ is the upper envelope of lines $y = x + \beta$ and $y = \gamma$ as shown in Figure 2.3.1a. These two lines intersect at $x = \gamma - \beta$ which is called the *root* of the tropical line. Tropical lines have a unique root similar to regular lines. A tropical quadratic defined as

$$\alpha \odot x^{\odot 2} \oplus \beta \odot x \oplus \gamma = \max(\alpha + 2x, \beta + x, \gamma)$$

is the upper envelope of the three lines $y = \alpha + 2x$, $y = \beta + x$ and $y = \gamma$, whose roots are the intersection points of the three lines; $x_1 = \gamma - \beta$ and $x_2 = \beta - \alpha$ provided that $\gamma - \beta \leq \beta - \alpha$ as shown in Figure 2.3.1b. However, if $\gamma - \beta > \beta - \alpha$, the tropical quadratic has a unique root at $(\gamma - \alpha)/2$ as shown in Figure 2.3.1c.

Definition 2.3.2 (Tropical Scalar Multiplication and Vector Addition)**.** *For any scalars* $a, b \in \mathbb{R} \cup \{-\infty\}$ *and for any vectors* $v = (v_1, \ldots, v_e), w = (w_1, \ldots, w_e) \in (\mathbb{R} \cup \{-\infty\})^e$*, where* $e \in \mathbb{N}$ *is the dimension of the space, tropical scalar multiplication and tropical vector addition is defined as follows:*

$$a \odot v := (a + v_1, \ldots, a + v_e),$$



(a) (b) (c)

Figure 2.3.1: (a) Tropical line, (b),(c) Tropical quadratics with $\gamma - \beta < \beta - \alpha$ and $\gamma - \beta > \beta - \alpha$, respectively.

$$a \odot v \oplus b \odot w := (\max\{a + v_1, b + w_1\}, \ldots, \max\{a + v_e, b + w_e\}).$$

2.3.2 Distance metric

In classical Euclidean geometry, distances are induced by norms such as the ℓ_2 or ℓ_∞ norm, which measure absolute differences between points in a linear space. In tropical geometry, however, the underlying algebra is the max-plus semiring, and points $x \in \mathbb{R}^n$ are considered up to the equivalence relation $x \sim y$ if $x - y = \alpha\mathbf{1}$ (i.e. tropical scaling). A Euclidean norm would not respect this equivalence, since it changes under uniform translations. A more appropriate metric is the *Hilbert projective metric*, which measures the range of coordinate-wise differences between two vectors and is invariant under tropical scaling. It is not an analogue of the Euclidean or ℓ_1 distance specifically, but rather the unique metric that respects the fundamental tropical scaling invariance. It provides a natural way to quantify distance on the tropical projective torus $\mathbb{R}^e/\mathbb{R}\mathbf{1}$. Note that e is an integer used to indicate the dimension of the space and should not be confused with Euler's number.

Definition 2.3.3 (Generalised Hilbert Projective Metric)**.** *For any two vectors v, $w \in (\mathbb{R} \cup \{-\infty\})^e$ the* tropical distance $d_{\mathrm{tr}}(v, w)$ *between v and w is defined as:*

$$d_{\mathrm{tr}}(v, w) := \max_i\{v_i - w_i\} - \min_i\{v_i - w_i\} = \mathcal{R}(v - w),$$

where $v = (v_1, \ldots, v_e)$, $w = (w_1, \ldots, w_e)$ *and* $\mathcal{R} : (\mathbb{R} \cup \{-\infty\})^e \to \mathbb{R}_{\geq 0},\ x \mapsto \max_i\{x_i\} - \min_i\{x_i\}$ *is the "range" operarator.*

Example 2: The tropical distance between $v = (-\infty, 2, 0)$ and $w = (-\infty, 0, 1)$ is

$$d_{\text{tr}}(v, w) = \mathcal{R}(v - w) = \mathcal{R}(0, 2, -1) = 2 - (-1) = 3$$

Note that subtraction of minus infinities is assumed to be zero.

Remark 3: Consider two vectors $v = (c, \ldots, c) = c\mathbf{1} \in \mathbb{R}^e$ and $w = \mathbf{0} \in \mathbb{R}^e$. Their tropical distance is

$$d_{\text{tr}}(v, w) = \mathcal{R}(v - w) = \mathcal{R}(c, \ldots, c) = c - c = 0$$

Hence, d_{tr} is not a metric in $\mathbb{R}^e$. The space in which d_{tr} is a metric treats all points in $\{c\mathbf{1} : c \in \mathbb{R}\} = \mathbb{R}\mathbf{1}$ as the same point. The quotient space $(\mathbb{R} \cup \{-\infty\})^e/\mathbb{R}\mathbf{1}$ achieves just that.

Proposition 2.3.4. *The function* d_{tr} *is a well-defined metric on* $(\mathbb{R} \cup \{-\infty\})^e/\mathbb{R}\mathbf{1}$, *where* $\mathbf{1} \in \mathbb{R}^e$ *is the vector of all-ones.*

Proof. Clearly, $d_{\text{tr}}(u, u) = 0$. If $d_{\text{tr}}(u, v) = 0$, then $\max_i\{u_i - v_i\} = \min_i\{u_i - v_i\}$ so $u - v = c \cdot \mathbf{1}$ for some $c \in \mathbb{R}$ and since $u, v \in (\mathbb{R} \cup \{-\infty\})^e/\mathbb{R}\mathbf{1}$, the two elements are equal. The function d_{tr} can be expressed as $\max_i\{u_i - v_i\} + \max_i\{v_i - u_i\}$ which is commutative. Finally, for the triangular inequality

$$\begin{aligned} d_{\text{tr}}(u, w) &= \max\{u - w\} + \max\{w - u\} \\ &\overset{(*)}{\geq} (\max\{u - v\} + \max\{v - w\}) + (\max\{v - w\} + \max\{w - v\}) \\ &= d_{\text{tr}}(u, v) + d_{\text{tr}}(v, w), \end{aligned}$$

where $(*)$ uses the trivial inequality $\max\{x\} + \max\{y\} \geq \max\{x + y\}$ □

Example 3 (Unit circle)**:** In this example the unit circle under the tropical Hilbert metric $d_{\text{tr}}(x, \mathbf{0}) = \max\{x\} - \min\{x\}$ in $\mathbb{R}^3/\mathbb{R}\mathbf{1}$ is derived. In particular, since $\mathbb{R}^3/\mathbb{R}\mathbf{1} \cong \mathbb{R}^2$ it can be derived in two dimensions by setting $x_3 = 0$. Assuming that $x_2 > 0$,

$$d_{\text{tr}}(x, \mathbf{0}) = \max\{x_1, x_2, 0\} + \max\{-x_1, -x_2, 0\} = \begin{cases} \max\{x_1, x_2\}, & \text{if } x_1, x_2 > 0 \\ x_2 - x_1, & \text{if } x_2 > 0 > x_1 \end{cases},$$

and so the unit circle $d_{\text{tr}}(x, \mathbf{0}) = 1$ is the locus $\max\{x_1, x_2\} = 1$ in the first quadrant and the line $x_2 - x_1 = 1$ in the second quadrant. Since $d_{\text{tr}}(x, \mathbf{0}) = d_{\text{tr}}(-x, \mathbf{0})$, the other two quadrants can be drawn using reflective symmetry as shown in Figure 2.3.2. Any circle in $\mathbb{R}^3/\mathbb{R}\mathbf{1}$ centred at $x^* = (x_1^*, x_2^*, 0)$ with radius R can be expressed as $d_{\text{tr}}(x, x^*) = R$ and can be drawn by enlarging the unit circle by the factor R and translating it by the vector x^*.

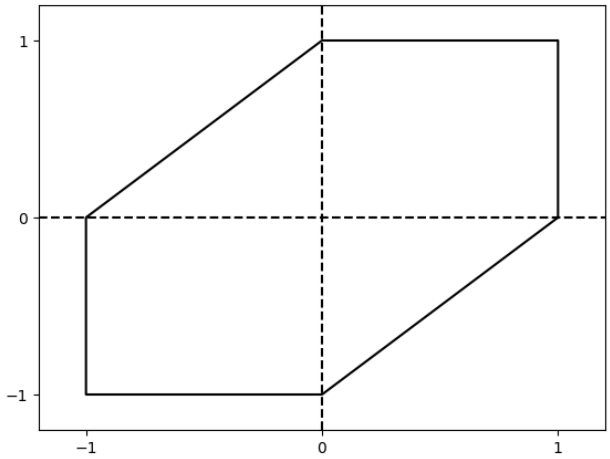


Figure 2.3.2: Unit circle $d_{\text{tr}}(x, \mathbf{0}) = 1$ in $\mathbb{R}^3/\mathbb{R}\mathbf{1}$.

Remark 4 (Tropical arithmetic operations in $\mathbb{R}^e/\mathbb{R}\mathbf{1}$)**:** Consider for example $x = (1, 2)$ and $y = (0, 3)$, whose tropical sum is $x \oplus y = (1, 3)$. Under the tropical projective torus $x = \lambda \odot x$ for all $\lambda \in \mathbb{R}$, but $1 \odot x \oplus y = (2, 3)$ which is not equal to $(1, 3)$ in $\mathbb{R}^2/\mathbb{R}\mathbf{1}$. Hence, tropical arithmetic operations are not well-defined in the tropical projective torus.

2.3.3 Tropical convexity

Convexity is a fundamental concept in Euclidean geometry. For every pair of points x, y in some convex polytope or linear space, all points $\lambda x + (1-\lambda)y$, $\lambda \in [0,1]$ that lie in a line segment connecting them are also elements of that structure. In tropical geometry the equivalent property that respects tropical arithmetic operations is *tropical convexity*. This concept is particularly relevant in statistics for modeling data with underlying tree structures, such as phylogenetic trees. As we will see in the following subsection, these trees naturally correspond to points in a tropical linear space which is a tropically convex set, making tropical convexity the essential framework for studying their geometric relationships and interpolations. In this subsection tropical convexity and tropical polytopes, such as tropical line segments and triangles, are presented. More detail regarding useful properties of tropical convexity and tropical linear spaces can be found in Hampe (2015).

Definition 2.3.5 (Tropical Convexity, Hampe (2015), Develin and Sturmfels (2004))**.** *Suppose we have $S \subset \mathbb{R}^e$. If*

$$c_1 \odot v \oplus c_2 \odot w \in S \tag{2.3.1}$$

for any $c_1, c_2 \in \mathbb{R}$ and for any points $v, w \in S$, then S is called tropically convex*. Note that classically convex sets are defined similarly under regular arithmetic and with c_1, $c_2 \in [0,1]$ and $c_2 = 1 - c_1$.*

Tropically convex sets are well-defined in the tropical projective torus, since for all $v \in S$, $c_1 \odot v \in S$ for all $c_1 \in \mathbb{R}$; this follows from the definition by allowing $c_2 - c_1$ to be sufficiently small. In the tropical projective torus, a set $S \subset \mathbb{R}^e/\mathbb{R}\mathbf{1}$ is tropically convex if its preimage $\{\phi^{-1}(x) : x \in \mathbb{R}^e/\mathbb{R}\mathbf{1}\}$ is also tropically convex, where here ϕ is the quotient map $\phi : \mathbb{R}^e \to \mathbb{R}^e/\mathbb{R}\mathbf{1}$. Note that in the tropical projective torus the convex

property (2.3.1) *of a tropically convex set* S *simplifies to*

$$c_1 \odot v \oplus c_2 \odot w = (c_1 \odot v \oplus c_2 \odot w) \odot (-c_2) = \lambda \odot v \oplus w \in S \tag{2.3.2}$$

for all $\lambda \coloneqq c_1 - c_2 \in \mathbb{R}$ *and any* $v, w \in S$.

In what follows, we demonstrate that geometric objects such as tropical circles, tropical lines, and tropical triangles are tropically convex. This result confirms that tropical convexity successfully preserves the fundamental geometric properties of its classical Euclidean counterpart.

Remark 5 (Tropical geodesics)**:** The tropical metric admits a natural geometric meaning: it turns $\mathbb{R}^e/\mathbb{R}\mathbf{1}$ into a metric space whose geodesics are tropical line segments. In particular, the tropical segment joining two points $v, w \in \mathbb{R}^e/\mathbb{R}\mathbf{1}$, that is, their tropical convex hull $\mathrm{tconv}(\{v, w\})$, realizes a shortest path between them with respect to d_{tr} (see Maclagan and Sturmfels (2021)). However, as observed by Lee et al. (2022), the tropical Hilbert metric admits infinitely many geodesics between any two points. Related geometric constructions of tree spaces equipped with geodesic metrics have been proposed in other contexts, such as the Wald space of Garba et al. (2021).

Lemma 2.3.6. *Tropical circles are tropically convex.*

Proof. Tropical convexity is preserved under classical scaling and translation: if $S \subset \mathbb{R}^e$ is tropically convex, then so is $\alpha S + c = \{\alpha x + c : x \in S\}$ for any $\alpha \in \mathbb{R}$ and $c \in \mathbb{R}^e$. Hence it suffices to prove that the unit tropical circle

$$S = \{x \in \mathbb{R}^e : d_{\mathrm{tr}}(x, \mathbf{0}) \leq 1\}$$

is tropically convex.

By definition, $x \in S$ if and only if $x_i - x_j \leq 1$ for all i, j. Let $x, y \in S$ and $\lambda \in \mathbb{R}$,

and set $z = \lambda \odot x \oplus y$, so $z_k = \max(\lambda + x_k,\, y_k)$. Then, for any i, j,

$$z_i - z_j = \max(\lambda + x_i,\, y_i) - \max(\lambda + x_j,\, y_j) \leq \max(x_i - x_j,\, y_i - y_j) \leq 1.$$

Thus $z \in S$, and S is tropically convex. Consequently, every tropical circle $\alpha S + c$, with centre c and radius α, is tropically convex as well. □

Proposition 2.3.7 (Tropical polytope, first defined in Develin and Sturmfels (2004))**.** *Suppose* $V = \{v_1, \ldots, v_s\} \subset \mathbb{R}^e/\mathbb{R}\mathbf{1}$. *The smallest tropically-convex subset containing* V *is called the* tropical convex hull *or* tropical polytope *of* V. *It can be expressed as the set of all tropical linear combinations of vectors in* V*,*

$$\operatorname{tconv}(V) = \left\{ \bigoplus_{i=1}^{s} a_i \odot v_i : a \in \mathbb{R}^s \right\}$$

Classical polytopes are defined similarly under regular arithmetic with $a \in [0, 1]^s$ and $\sum_i a_i = 1$.

Example 4: A *tropical line segment* $\Gamma_{u,v}$ between two points $u,\, v \in \mathbb{R}^e/\mathbb{R}\mathbf{1}$ is a tropical polytope of a set of two points $\{u,\, v\} \subset \mathbb{R}^e/\mathbb{R}\mathbf{1}$. In other words, using the convexity property (2.3.2) the tropical line segment is defined as $\Gamma_{u,v} = \{\lambda \odot u \oplus v : \lambda \in \mathbb{R}\}$.

Without loss of generality, it may be assumed that

$$v_e - u_e \geq v_{e-1} - u_{e-1} \geq \ldots v_1 - u_1 = 0$$

after permuting the coordinates of $v - u$ and adding multiples of $\mathbf{1}$. Then the tropical

line $\Gamma_{u,v}$ contains the following points corresponding to $\lambda = u_i - v_i$ for $i \in [e]$,

$$\begin{aligned}
w_1 &\coloneqq (u_1 - v_1) \odot v \oplus u = v = (u_1, u_1 - v_1 + v_2, u_1 - v_1 + v_3, \ldots, u_1 - v_1 + v_e) \\
w_2 &\coloneqq (u_2 - v_2) \odot v \oplus u = (u_1, u_2, u_2 - v_2 + v_3, \ldots, u_2 - v_2 + v_e) \\
w_3 &\coloneqq (u_3 - v_3) \odot v \oplus u = (u_1, u_2, u_3, u_3 - v_3 + v_4, \ldots, u_3 - v_3 + v_e) \\
&\ldots\ldots \\
w_{e-1} &\coloneqq (u_{e-1} - v_{e-1}) \odot v \oplus u = (u_1, \ldots, u_{e-1}, u_{e-1} - v_{e-1} + v_e) \\
w_e &\coloneqq (u_e - v_e) \odot v \oplus u = u
\end{aligned}$$

For $\lambda \in [u_{i+1} - v_{i+1}, u_i - v_i]$

$$\lambda \odot v \oplus u = (u_1, \ldots, u_i, \lambda + v_{i+1}, \lambda + v_{i+2}, \ldots, \lambda + v_e)$$

which corresponds to a regular line segment connecting w_i and w_{i+1}. Hence, the tropical line segment is the union of regular line segments $\Gamma_{u,v} = \bigcup_{i=1}^{e-1} \gamma_{w_i, w_{i+1}}$, where $\gamma_{x,y}$ denotes a regular line segment between x and y.

Example 5 (Tropical line segment)**:** In this example the task is to draw the tropical line segment connecting two points $u = (0, 0, 0)$ and $v = (0, 2, 5)$. The coordinates do not need to be permuted since $v_3 - u_3 \geq v_2 - u_2 \geq v_1 - u_1 = 0$. The tropical line segment $\Gamma_{u,v}$ contains the following points

$$\begin{aligned}
0 \odot v \oplus u &= (0, 2, 5) = v \\
-2 \odot v \oplus u &= (0, 0, 3) \\
-5 \odot v \oplus u &= (0, 0, 0) = u
\end{aligned}$$

Hence, the tropical line segment consists of two regular line segments connecting u and v to $(0, 0, 3)$ as shown in Figure 2.3.3.

Example 6 (Tropical triangle)**:** A tropical triangle is a tropical polytope with three vertices that do not all lie in the same tropical line. As with regular triangles, the boundary of a tropical triangle consists of its three tropical line segments connecting the vertices pairwise. The tropical triangle $\text{tconv}(\{v_1, v_2, v_3\})$ with vertices $v_1 = (0, 0, 0), v_2 = (0, 3, 1), v_3 = (0, 2, 5)$ is the area enclosed by the three tropical edges (tropical line segments connecting the vertices) as shown in Figure 2.3.3. In Euclidean space this tropical triangle appears as a hexagon. Interestingly, the tropical circle shown in Figure 2.3.2 has the same Euclidean shape: for $e = 3$, the tropical circle coincides with the tropical triangle generated by the standard basis vectors

$$e_1 = (1, 0, 0) = (0, -1, -1), \quad e_2 = (0, 1, 0), \quad e_3 = (0, 0, 1).$$

The following proposition generalises this observation.

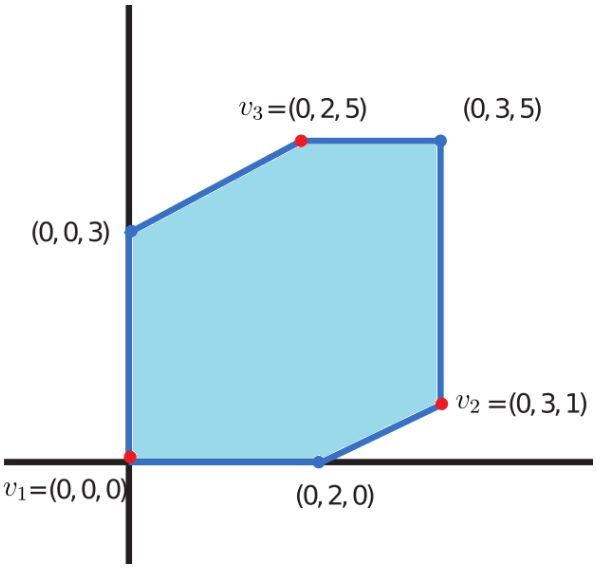


Figure 2.3.3: Tropical triangle $v_1 = (0, 0, 0), v_2 = (0, 3, 1), v_3 = (0, 2, 5)$ in $\mathbb{R}^3/\mathbb{R}\mathbf{1}$. Figure adapted from Yoshida et al. (2023c).

Proposition 2.3.8. *The tropical unit ball in* $\mathbb{R}^n/\mathbb{R}\mathbf{1}$ *is the tropical convex hull of the*

standard basis vectors:

$$\{\, x \in \mathbb{R}^n/\mathbb{R}\mathbf{1} : d_{\mathrm{tr}}(x, \mathbf{0}) \leq 1 \,\} = \mathrm{tconv}\big(\mathbf{e}_1, \ldots, \mathbf{e}_n\big),$$

where $\mathbf{e}_i$ *denotes the* i*th standard basis vector of* $\mathbb{R}^n$*. In particular, the tropical unit ball is a tropical simplex generated by* n *vertices.*

Proof. Let S denote the tropical unit ball. Since each standard basis vector $\mathbf{e}_i \in S$ and S is tropically convex (Lemma 2.3.6), we have $\mathrm{tconv}(\mathbf{e}_i : i \in [n]) \subseteq S$.

For the reverse inclusion, take any $x \in S$ and subtract its minimum coordinate so that $\min_k x_k = 0$. Then all coordinates satisfy $x_k \in [0, 1]$. Define $a_i := x_i - 1$, and note that

$$\bigoplus_{i=1}^{n} a_i \odot \mathbf{e}_i = (\max(a_k + 1, \max_{j \neq k} a_j))_k = x.$$

Hence $x \in \mathrm{tconv}(\mathbf{e}_i : i \in [n])$, and therefore $S = \mathrm{tconv}(\mathbf{e}_i : i \in [n])$. □

Tropical convexity provides a fundamental framework for studying geometric structures in a combinatorial and piecewise-linear setting. The examples of tropical line segments and tropical shapes illustrate how classical convexity is adapted under tropical arithmetic, producing geometric objects that combine discrete combinatorial structure with continuous, piecewise-linear geometry. This combination of discrete and continuous properties is particularly relevant for phylogenetic reconstruction. As we will discuss in the following subsection, phylogenetic trees can be represented as ultrametric elements (defined in Definition 2.3.9), which correspond to points in a tropical linear space. Understanding tropical convexity enables us to explore the geometry of tree spaces, bridging combinatorial and continuous perspectives in a way that classical Euclidean convexity does not.

2.3.4 Connection to Phylogenetics

A goal of this thesis is to build statistical models using tropical geometry over the *space of phylogenetic trees*. A phylogenetic tree is a weighted tree whose internal nodes do not have labels and whose external nodes/leaves, have labels $[m]$. These labels correspond to organisms. The leaf label set of phylogenetic trees is denoted as $[m]$. In particular, we are interested over a certain class of phylogenetic trees.

Definition 2.3.9. *Let T be a rooted phylogenetic tree with leaf label set $[m]$. If the distance from all leaves $i \in [m]$ to the root is the same, then T is an* equidistant tree.

To apply certain statistical and machine learning methods, a vector representation of trees is often convenient. There are many ways to map a phylogenetic tree to a point, including the BHV metric space (Billera et al., 2001). We vectorize a phylogenetic tree using dissimilarity maps. Dissimilarity maps are maps $d : [m] \times [m] \to \mathbb{R}$ such that $d(i, i) = 0$ and $d(i, j) = d(j, i)$. We consider dissimilarity maps over $[m] \times [m]$ such that $d(i, j)$ is the pairwise distance between a leaf $i \in [m]$ to a leaf $j \in [m]$ and a vector of all possible pairwise distances in T between any two leaves in $[m]$ as a vector representation of a phylogenetic tree T with $[m]$. Hence the dimension of the phylogenetic tree space is $\binom{m}{2}$.

Definition 2.3.10 (Ultrametric)**.** *Let $([m], d)$ be a metric space, where $[m] = \{1, \ldots, m\}$ and $d : [m]^2 \to \mathbb{R}_{\geq 0}$ is the distance function. The metric d is an ultrametric if, for all $i, j, k \in [m]$, it satisfies the strong triangle inequality:*

$$d(i, k) \leq \max\{d(i, j), d(j, k)\}.$$

Example 7: Suppose $m = 3$. Let d be a metric on $[m] := \{1, 2, 3\}$ such that

$$d(1, 2) = 2,\ d(1, 3) = 2,\ d(2, 3) = 1$$

Since the maximum is achieved twice, d is an ultrametric.

Theorem 2.3.11 (noted in Buneman (1974)). *Suppose we have an equidistant tree T with a leaf label set $[m]$ and suppose $d(i,j)$ for all $i,j \in [m]$ is a distance from a leaf i to a leaf j. Then, d is an ultrametric if and only if T is an equidistant tree.*

Example 8: Suppose we have $m = 5$. Then, the phylogenetic tree shown in Fig. 2.3.4 is an equidistant tree with a leaf label set $[5] := \{1, 2, 3, 4, 5\}$ and its pairwise distances are

$$u = (4, 4, 4, 4, 2, 2, 2, 1.6, 1.6, 0.6),$$

as illustrated in Figure 2.3.5. It is easy to verify that u is an ultrametric.

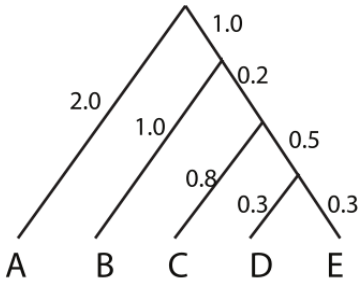


Figure 2.3.4: Example of an equidistant tree with a leaf label set [5].

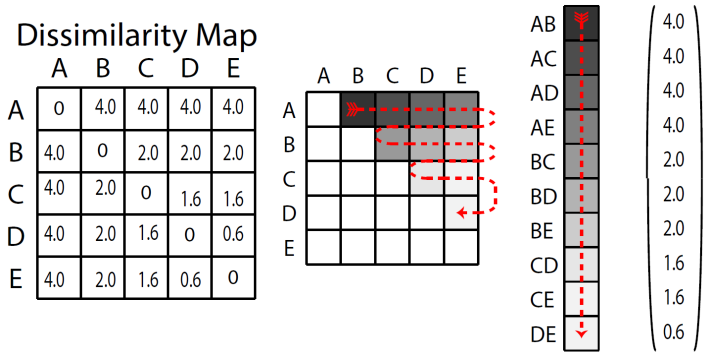


Figure 2.3.5: Vectorisation of the equidistant tree from Figure 2.3.4.

Under the strict molecular clock assumption, the rate of molecular evolution is constant across all lineages. This constancy implies that the evolutionary distance from any extant species to its Most Recent Common Ancestor (MRCA) is the same, resulting

in an equidistantphylogenetic tree (Zuckerkandl and Pauling, 1965). This concept has a direct mathematical equivalent: a distance matrix corresponds to an equidistant tree if and only if it satisfies the ultrametric property (as defined by Theorem 2.3.11). Following this equivalence (popularized by Felsenstein (2004)), the space of all possible equidistant phylogenetic trees is mathematically equivalent to the space of ultrametrics $\mathcal{U}_m$. Therefore, under the molecular clock, both species trees and gene trees can be modeled as ultrametric trees.

The following theorem describes the connection between ultrametrics and tropical spaces, providing the geometric foundation for subsequent statistical analysis.

Theorem 2.3.12 (explained in Ardila and Klivans (2006); Page et al. (2020))**.** *Suppose we have a* classical *linear subspace* $L_m \subset \mathbb{R}^e$ *, where* $e = \binom{m}{2}$*, defined by the linear equations* $x_{ij} - x_{ik} + x_{jk} = 0$ *for* $1 \leq i < j < k \leq m$*. Let* $Trop(L_m) \subseteq \mathbb{R}^e/\mathbb{R}\mathbf{1}$ *be the* tropicalization *of the linear space* $L_m \subset \mathbb{R}^e$*, that is, classical operators are replaced by tropical ones (defined in Section 3.1) in the equations defining the linear subspace* L_m*, so that all points* $(v_{12}, v_{13}, \ldots, v_{m-1,m})$ *in* $Trop(L_m)$ *satisfy the condition that*

$$\max_{i,j,k\in[m]} \{v_{ij}, v_{ik}, v_{jk}\}.$$

is attained at least twice. Then, the image of $\mathcal{U}_m$ *inside of the tropical projective torus* $\mathbb{R}^e/\mathbb{R}\mathbf{1}$ *is equal to* $Trop(L_m)$*.*

Theorem 2.3.12 establishes a critical geometric link: the non-Euclidean, discrete-looking space of ultrametric trees is precisely a *tropical linear space* when viewed through the lens of max-plus algebra. This insight is central to developing statistically rigorous methods for tree-based data. Traditional statistical models, such as linear regression and neural networks, rely on a smooth Euclidean space and ℓ_2-norms, making their direct application to the non-Euclidean space of phylogenetic trees flawed. The tropical linear space structure, however, naturally defines the appropriate distance

metric for trees, the *tropical metric* (defined in 2.3.3).

The subsequent chapters fully leverage this tropical geometry to adapt classical statistical models for tree data. Specifically, Chapter 3 introduces the Tropical Logistic Regression (TLR) model. TLR is an analogue of classical logistic regression defined within the tropical projective torus, a superset of L_m[5]. By respecting the underlying tropical geometry and utilizing the tropical metric for distance calculations, TLR outperforms its classical counterpart in phylogenetic classification tasks, providing a branch-length-sensitive metric for assessing MCMC convergence that is significantly more robust than traditional topology-only metrics used by popular software such as `MrBayes`.

Building upon this foundation, Chapter 4 extends the approach to deep learning with the Tropical Neural Network (TNN). Standard neural networks perform poorly on tree data due to their Euclidean assumptions. The TNN addresses this by incorporating a tropical embedding layer as its initial component. This layer transforms the tree vectors from the tropical space into a suitable Euclidean space representation using the tropical metric. This architecture allows the TNN to function as a **universal approximator** for tree data and is shown to be a generalisation of the Tropical Logistic Regression itself, establishing a comprehensive and geometrically sound framework for deep learning on phylogenetic tree spaces.

In conclusion, the analysis of evolutionary rates, pioneered by the molecular clock hypothesis, establishes that clock-like evolution corresponds to an equidistant tree, which is mathematically equivalent to the ultrametric distance property and, in turn, to a tropical linear space. This insight is critical because it mandates a departure from traditional Euclidean statistical methods, whose underlying assumptions are violated by tree data. The two subsequent chapters illustrate this connection through the development of statistical methods that respect this tropical geometry, yielding higher

[5]Restricting the classification problem to L_m presents computational challenges that are not resolved in the paper and require further work.

predictive power than those conforming only to Euclidean geometry.

2.3.5 Other tree spaces

Before concluding this section, we briefly mention some other tree spaces. The *The Billera–Holmes–Vogtmann* tree space is a geometric space used to compare phylogenetic trees. It was first introduced in Billera et al. (2001) and combines aspects of geometric group theory and computational biology. In particular, the BHV space is a metric space of phylogenetic trees with a fixed set of labeled n leaves. The topology is fixed within regions, while the edge lengths vary as continuous parameters. The structure of this space can be thought of as a set of glued Euclidean orthants $\mathbb{R}^{n-3}_{\geq 0}$, with each orthant representing a tree topology and the coordinates representing each of the $n-3$ edge lengths. When an edge shrinks to zero, that represents the boundary of the orthant and a transition to different topologies/orthants. The BHV space is equipped with a geodesic metric, meaning distance is defined as the length of the shortest path between two trees. If the trees have the same topologies and corresponding lengths $l_1, l_2 \in \mathbb{R}^{n-3}$, their distance is the Euclidean distance $\|l_1 - l_2\|_2$. The caveat with this discrete-continuous space is that it is not a smooth manifold and it has corners and singularities where the $(2n-5)!!$ top-dimensional orthants meet. Its main advantage is that it is a CAT(0) space, i.e. globally non-positively curved. Curvature here refers to the "fatness" of triangles, with positively curved surfaces such as a sphere having fat triangles (the sum of edge lengths exceeds those of a Euclidean triangle), a flat surface having zero curvature and Euclidean triangles, and a saddle surface having negative curvature and thin triangles. A CAT(0) space behaves like the last two. Consequently, in these spaces there is a unique geodesic (exactly one shortest path), which in turn leads to well behaved optimization (e.g. unique averages such as the Frechet mean).

Another tree space worth mentioning is the Robinson–Foulds (RF) space (Robinson and Foulds, 1981). This is a purely discrete space that considers only tree topologies

and ignores edge lengths. The distance between two trees T_1, T_2 is defined as

$$d_{\mathrm{RF}}(T_1, T_2) = |\Sigma(T_1) \triangle \Sigma(T_2)|,$$

where $\Sigma(T)$ is the set of splits/bipartitions of tree T, i.e. the number of splits that differ between the two trees. While the RF metric is simple and computationally efficient, it is relatively coarse and does not capture gradual changes in tree geometry. Unlike the BHV space, it does not admit a continuous structure or a notion of geodesics and averages.

2.4 Neural Networks

Artificial neural networks are a class of function approximators inspired by biological neural systems. A (feedforward) neural network defines a mapping $F : \mathbb{R}^d \to \mathbb{R}^k$ as a composition of layers, where each layer applies an affine transformation followed by a nonlinear activation function. More precisely, for an L-layer network, the output is computed recursively as

$$x^{(l)} = \sigma^{(l)}\big(W^{(l)} x^{(l-1)} + b^{(l)}\big), \quad l = 1, \ldots, L,$$

where $W^{(l)}$ and $b^{(l)}$ are the weights and biases, and $\sigma^{(l)}$ is a nonlinear activation function such as the ReLU or the sigmoid function. Note that if $\sigma^{(l)}$ are linear functions, the output is a linear transformation of the input.

Neural networks have been shown to possess strong approximation properties. In particular, universal approximation theorems guarantee that sufficiently large networks can approximate a wide class of functions arbitrarily well. This expressive power, combined with efficient training via stochastic gradient descent, has led to their widespread

success in machine learning. In Chapter 4, we prove that a variant of the universal approximation theorem can be applied to functions in "tropical" spaces.

In this work, specifically Chapter 4, we consider a variant of neural networks adapted to tropical geometry. While classical neural networks rely on Euclidean inner products, tropical neural networks replace these operations with tropical analogues, leading to models that are better suited for data with combinatorial or geometric structure, such as phylogenetic trees. We prove that a variant of the universal approximation theorem can be applied to functions in those spaces, and demonstrate with both simulated and real world data that it can outperform classical ReLU neural networks.

Chapter 3

Tropical Logistic Regression

3.1 Introduction

Phylogenomics is a new field that applies tools from phylogenetics to genome datasets. The multi-species coalescent model is often used to model the distribution of gene trees under a given species tree (Maddison, 2008). The first step in statistical analysis of phylogenomic data is to analyze sequence alignments to determine whether their evolutionary histories are congruent with each other. In this step, evolutionary biologists aim to identify genes with unusual evolutionary events, such as duplication, horizontal gene transfer, or hybridization (Ané et al., 2007). To accomplish this, they compare multiple sets of *gene trees*, that is, phylogenetic trees reconstructed from alignments of genes, with each gene tree characterised by the aforementioned evolutionary events. The classification of gene trees into different categories is therefore important for analyzing multi-locus phylogenetic data.

Tree classification can also help in assessing the convergence of Markov Chain Monte Carlo (MCMC) analyses for Bayesian inference on phylogenetic tree reconstruction. Often, we apply MCMC samplers to estimate the posterior distribution of a phylogenetic tree given an observed alignment. These samplers typically run multiple independent

Markov chains on the same observed dataset. The goal is to check whether these chains converge to the same distribution. This process is often done by comparing summary statistics computed from sampled trees. These statistics often only depend on the tree topologies, and so they naturally lose information about the branch lengths of the sampled trees. Alternatively, we propose the use of a classification model that classifies trees from different chains and uses statistical measures such as the Area under the ROC Curve (AUC) to indicate how distinguishable the two chains are. Consequently, high values of AUCs indicate that the chains have not converged to the equilibrium distribution. Currently, there is no classification model over the space of phylogenetic trees, the set of all possible phylogenetic trees with a fixed number of leaves. In this paper, we propose a classifier that is appropriate for the tree space and is sensitive to branch lengths, unlike the summary statistics of most MCMC convergence diagnostic tools.

In Euclidean geometry, the logistic regression model is the simplest generalized linear model for classification. It is a supervised learning method that classifies data points by modeling the log-odds of having a response variable in a particular class as a linear combination of predictors. This model is very popular in statistical learning due to its simplicity, computational speed and interpretability. However, directly applying such classical supervised models to a set of sampled trees may be misleading, since the space of phylogenetic trees does not conform to Euclidean geometry.

The space of phylogenetic trees with labeled leaves $[m]$ is a union of lower dimensional polyhedral cones with dimension $m-1$ over $\mathbb{R}^e$ where $e = \binom{m}{2}$ (Speyer and Sturmfels, 2009; Lin et al., 2017). This space is not Euclidean and even lacks convexity (Lin et al., 2017). In fact, Speyer and Sturmfels (2009) showed that the space of phylogenetic trees is a *tropicalization* of linear subspaces defined by a system of tropical linear equations (Page et al., 2020) and is therefore a tropical linear space.

Consequently, many researchers have applied tools from tropical geometry to sta-

tistical learning methods in phylogenomics, such as principal component analysis over the space of phylogenetic trees with a given set of leaves $[m]$ (Yoshida et al., 2019; Page et al., 2020), kernel density estimation (Yoshida et al., 2022d), MCMC sampling (Yoshida et al., 2022b), and support vector machines (Yoshida et al., 2023c). Recently, Akian et al. (2021) proposed a tropical linear regression over the tropical projective space as the best-fit tropical hyperplane. However, our logistic regression model is built from first principles and is not a trivial extension of the aforementioned tropical regression model.

In this paper, an analog of the logistic regression is developed over the tropical projective space, which is the quotient space $\mathbb{R}^e/\mathbb{R}\mathbf{1}$ where $\mathbf{1} := (1, 1, \ldots, 1)$. Given a sample of observations within this space, the proposed model finds the "best-fit" tree representative $\omega_Y \in \mathbb{R}^e/\mathbb{R}\mathbf{1}$ of each class $Y \in \{0, 1\}$ and the "best-fit" deviation of the gene trees. This tree representative is a statistical parameter and can be interpreted as the corresponding species tree of the gene trees. The deviation parameter is defined in terms of the variability of branch lengths of gene trees. It is established that the median tree, specifically the Fermat-Weber point, can asymptotically approximate the inferred tree representative of each class. The response variable $Y \in \{0, 1\}$ has conditional distribution $Y|X \sim \text{Bernoulli}(S(h(X)))$, where $h(x)$ is small when x is close to ω_0 and far away from ω_1 and vice versa.

In Section 3.2 an overview of tropical geometry and its connections to phylogenetics is presented. The one-species and two-species tropical logistic models are developed in Section 3.3. Theoretical results, including the optimality of the proposed method over tropically distributed predictor trees, the distance distribution of those trees from their representative, the consistency of estimators and the generalization error of each model are stated in Section 3.3 and proved in Supplement 3.A. Section 3.4 explains the benefit and suitability of using the Fermat-Weber point approximation for the inferred trees and a sufficient optimality condition is stated. Computational results are

presented in Section 3.5 where a toy example is considered for illustration purposes. Additionally, a comparison study between classical, tropical and BHV logistic regression is conducted on data generated under the coalescent model. In both the toy example and the coalescent gene trees example, our model outperforms the alternative regression models. Finally, our model is proposed as an alternative MCMC convergence criterion in Section 3.5.3. The paper concludes with a discussion in Section 3.6. The code developed and implemented for the proposed model can be found in Aliatimis (2024b).

The dataset can be found at DRYAD with DOI: 10.5061/dryad.tht76hf65.

3.2 Tropical Geometry and Phylogenetic Trees

3.2.1 Tropical Basics

This section covers the basics of tropical geometry and provides the theoretical background for the model developed in later sections. The concept of a tropical metric will be used when defining a suitable distribution for the gene trees. For more details regarding the basic concepts of tropical geometry covered in this section, readers are recommended to consult Maclagan and Sturmfels (2021).

A key tool from tropical geometry is the *tropical metric* also known as the *tropical distance* defined as follows:

Definition 3.2.1 (Tropical distance)**.** *The* tropical distance*, more formally known as the Generalized Hilbert projective metric, between two vectors* $v, w \in (\mathbb{R} \cup \{-\infty\})^e$ *is defined as*

$$d_{\text{tr}}(v, w) := \|v - w\|_{\text{tr}} = \max_i\{v_i - w_i\} - \min_i\{v_i - w_i\}, \tag{3.2.1}$$

where $v = (v_1, \ldots, v_e)$ *and* $w = (w_1, \ldots, w_e)$.

Remark 6: Consider two vectors $v = (c, \ldots, c) = c\mathbf{1} \in \mathbb{R}^e$ and $w = \mathbf{0} \in \mathbb{R}^e$. It is easy to verify that $d_{\text{tr}}(v, w) = 0$ and as a result d_{tr} is not a metric in $\mathbb{R}^e$. The space in which

d_{tr} is a metric treats all points in $\{c\mathbf{1} : c \in \mathbb{R}\} = \mathbb{R}\mathbf{1}$ as the same point. The quotient space $(\mathbb{R} \cup \{-\infty\})^e/\mathbb{R}\mathbf{1}$ achieves just that.

Proposition 3.2.2. *The function d_{tr} is a well-defined metric on $(\mathbb{R} \cup \{-\infty\})^e/\mathbb{R}\mathbf{1}$, where $\mathbf{1} \in \mathbb{R}^e$ is the vector of all-ones.*

All proofs of this chapter can be found in the Appendix.

3.2.2 Equidistant Trees and Ultrametrics

Phylogenetic trees depict the evolutionary relationship between different taxa. For example, they may summarise the evolutionary history of certain species. The leaves of the tree correspond to the species studied, while internal nodes represent (often hypothetical) common ancestors of those species and their ancestors. In this paper, only rooted phylogenetic trees are considered, with the common ancestor of all taxa based on the root of the tree. The branch lengths of these trees are measured in evolutionary units, i.e. the amount of evolutionary change. Under the molecular clock hypothesis, the rate of genetic change between species is constant over time, which implies genetic equidistance and allows us to treat evolutionary units as proportional to time units. Consequently, phylogenetic trees of extant species are *equidistant trees*.

Definition 3.2.3 (Equidistant tree)**.** *Let T be a rooted phylogenetic tree with leaf label set $[m]$, where $m \in \mathbb{N}$ is the number of leaves. If the distance from all leaves $i \in [m]$ to the root is the same, then T is an equidistant tree.*

It is noted that the molecular clock hypothesis has limitations and the rate of genetic change can in fact vary from one species to another. However, the assumption that gene trees are equidistant is not unusual in phylogenomics; the multispecies coalescent model makes that assumption in order to conduct inference on the species tree from a sample of gene trees Maddison and Maddison (2009). The proposed classification method is not

restricted to equidistant trees, but all coalescent model gene trees produced in Section 3.5.2. are equidistant.

To apply certain statistical and machine learning methods, a vector representation of trees is often convenient. A common formulation is to use BHV metric space (Billera et al., 2001) but in this paper *distance matrices* are used instead, which are then transformed into vectors. The main reason is simplicity and computational efficiency; it is much easier to compute gradients in the tropical projective torus than in the BHV space.

Definition 3.2.4 (Distance matrix)**.** *Consider a phylogenetic tree T with leaf label set $[m]$. Its distance matrix $D \in \mathbb{R}^{m\times m}$ has components D_{ij} being the pairwise distance between a leaf $i \in [m]$ to a leaf $j \in [m]$. It follows that the matrix is symmetric with zeros on its diagonals. For equidistant trees, D_{ij} is equal to twice the difference between the current time and the latest time that the common ancestor of i and j was alive.*

To form a vector, the distance matrix D is mapped onto $\mathbb{R}^e$ by vectorizing the strictly upper triangular part of D, i.e.

$$D \mapsto (D_{12}, \dots, D_{1m}, D_{23}, \dots, D_{2m}, \dots, D_{(m-1)m}) \in \mathbb{R}^e,$$

where the dimension of the resulting vector is equal to the number of all possible pairwise combinations of leaves in T. Hence the dimension of the phylogenetic tree space is $e = \binom{m}{2}$. In what follows, the connection between the space of phylogenetic trees and tropical linear spaces is established.

Definition 3.2.5 (Ultrametric)**.** *Consider the distance matrix $D \in \mathbb{R}^{m\times m}$. Then if*

$$\max\{D_{ij}, D_{jk}, D_{ik}\}$$

is attained at least twice for any $i, j, k \in [m]$, D is an ultrametric*. Note that the distance*

map $d(i,j) = D_{ij}$ forms a metric in $[m]$, with the strong triangular inequality satisfied. The space of ultrametrics is denoted as $\mathcal{U}_m$.

Theorem 3.2.6 (noted in Buneman (1974))**.** *Suppose we have an equidistant tree T with a leaf label set $[m]$ and D as its distance matrix. Then, D is an ultrametric if and only if T is an equidistant tree.*

Using Theorem 3.2.6, if we wish to consider all possible equidistant trees, then it is equivalent to consider the space of ultrametrics as the space of phylogenetic trees on $[m]$. Here we define $\mathcal{U}_m$ as the space of ultrametrics with a set of leaf labels $[m]$. Theorem 3.B.1 (explained in Ardila and Klivans (2006); Page et al. (2020)) in Supplement 3.B establishes the connection between phylogenetic trees and tropical geometry by stating that the ultrametric space is a tropical linear space.

3.3 Method

Our logistic regression model is designed to capture the association between a binary response variable $Y \in \{0,1\}$ and an explanatory variable vector $X \in \mathbb{R}^n$, where n is the number of covariates in the model. Under the logistic model, $Y \sim \text{Bernoulli}(p(x|\omega))$ where

$$p(x|\omega) = \mathbb{P}(Y=1|x) = \frac{1}{1+\exp\left(-h_\omega(x)\right)} = \sigma\left(h_\omega(x)\right),$$

where σ is the logistic function and $\omega \in \mathbb{R}^n$ is the model parameter that needs to be estimated and h is a function that will be specified later. The log-likelihood function of logistic regression for N observation pairs $(x^{(1)}, y^{(1)}), \ldots, (x^{(N)}, y^{(N)})$ is

$$l(\omega|x,y) = \frac{1}{N}\sum_{i=1}^{N} y^{(i)} \log p_\omega^{(i)} + (1-y^{(i)}) \log(1-p_\omega^{(i)}), \tag{3.3.1}$$

where $p_\omega^{(i)} = p(x^{(i)}|\omega)$. It is the negative of the cross entropy loss. The training model seeks a statistical estimator $\hat{\omega}$ that maximizes this function.

3.3.1 Optimal Model

The framework described thus far incorporates the tropical, classical and BHV logistic regression as special cases. In this section, we show that these can be distinguished through the choice of the function h. In fact, this function h can be derived from the conditional distributions $X|Y$, as stated in Equation (3.3.2) of Lemma 1, below, by simple application of the Bayes' rule.

If $X|Y$ is a Gaussian distribution with appropriate parameters, the resulting model is the classical logistic regression. Alternatively, if $X|Y$ is a "tropical" distribution, then the resulting classification model is the "tropical" logistic regression. Examples 9 and 10 illustrate this for non-tropical and tropical distributions respectively, and Remark 7 discusses the choice of tropical distribution in more detail.

Furthermore, the function h from (3.3.2) also minimizes the expected cross-entropy loss according to Proposition 3.3.1. Therefore, the *best model* to fit data that have been generated by tropical Laplace distribution (3.3.4) is the tropical logistic regression. We conclude this section showing how the tropical metric and tropical Laplace distribution may be applied to produce two intuitive variants of tropical logistic regression, our one- and two-species models.

Lemma 1: Let $Y \sim \text{Bernoulli}(r)$ and define the random vector $X \in \mathbb{R}^n$ with conditional distribution $X|Y \sim f_Y$, where f_0, f_1 are probability density functions defined in $\mathbb{R}^n$. Then, $Y|X \sim \text{Bernoulli}(p(X))$ with $p(x) = \sigma(h(x))$, where

$$h(x) = \log\left(\frac{r f_1(x)}{(1-r) f_0(x)}\right). \tag{3.3.2}$$

Proposition 3.3.1. *Let $Y \sim$ Bernoulli(r) and define the random vector $X \in \mathbb{R}^n$ with conditional distribution $X|Y \sim f_Y$, where f_0, f_1 are probability density functions defined in $\mathbb{R}^n$. The functional p that maximises the expected log-likelihood as given by equation*

(3.3.1) *is* $p(x) = \sigma(h(x))$*, with* h *defined as in equation* (3.3.2) *of Lemma 1.*

Example 9 (Normal distribution and classical logistic regression)**:** Suppose that the two classes are equiprobable ($r = 1/2$) and that the covariate is multivariate normal

$$X|Y \sim \mathcal{N}(\omega_Y, \sigma^2 I_n),$$

where n is covariate dimension and I_n is the identity matrix. Using Lemma 1, the optimal model has

$$h(x) = -\frac{\|x-\omega_1\|^2}{2\sigma^2} + \frac{\|x-\omega_0\|^2}{2\sigma^2} = \frac{(\omega_1-\omega_0)^T}{\sigma^2}(x-\bar{\omega}), \tag{3.3.3}$$

where $\|\cdot\|$ is the Euclidean norm and $\bar{\omega} = (\omega_0+\omega_1)/2$. This model is the *c*lassical logistic regression model with translated covariate $X - \bar{\omega}$ and $\omega = \sigma^{-2}(\omega_1 - \omega_0)$.

Example 10 (Tropical Laplace distribution)**:** It may be assumed that the covariates are distributed according to the tropical version of the Laplace distribution, as presented in Yoshida et al. (2022b), with mean ω_Y and probability density functions

$$f_Y(x) = \frac{1}{\Lambda}\exp\left(-\frac{d_{\mathrm{tr}}(x,\omega_Y)}{\sigma_Y}\right), \tag{3.3.4}$$

where Λ is the normalizing constant of the distribution.

Proposition 3.3.2. *In distribution* (3.3.4)*, the normalizing factor is* $\Lambda = e!\sigma_Y^{e-1}$.

Proof. See Supplement 3.A. □

Remark 7: Consider $\mu \in \mathbb{R}^d$ and a covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$. Then the pdf of a

classical Gaussian distribution is

$$f_{\mu,\Sigma}(x) \propto \exp\left(-\frac{1}{2}(x-\mu)^t\Sigma^{-1}(x-\mu)\right) \tag{3.3.5}$$

where $x \in \mathbb{R}^d$ and y^t is the transpose of a vector $y \in \mathbb{R}^d$. When $\sigma_Y = 1$, the tropical Laplace distribution in (3.3.4) is tropicalization of the right hand side in (3.3.5) where Σ is to the tropical identity matrix

$$\begin{pmatrix} 0 & -\infty & -\infty & \dots & -\infty \\ -\infty & 0 & -\infty & \dots & -\infty \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -\infty & -\infty & -\infty & \dots & 0 \end{pmatrix}.$$

Tran (2020) nicely surveys the many different definitions of tropical Gaussian distributions. Since the space of ultrametrics is a tropical linear space (Speyer and Sturmfels, 2009), it is natural to use tropical "linear algebra" for the definition of tropical "Gaussian" distribution defined in (3.3.4) in this research. Clearly not all desirable properties of the classical Gaussian distribution are necessarily realised in a tropical space.

For example, as Tran discussed in Tran (2020), we lose some natural intuition of orthogonality of vectors. This means that we lose a nice geometric intuition of a correlation between two random vectors. Even with the loss of some nice properties of the classical Gaussian distribution, the tropical Laplace (3.3.5) is a popular choice. It has been applied to statistical analysis of phylogenetic trees: as a kernel density estimator of phylogenetic trees over the space of phylogenetic trees (Yoshida et al., 2022d), and as the Bayes estimator (Huggins et al., 2011) because this distribution is interpretable in terms of phylogenetic trees.

In particular, the tropical metric d_{tr} represents the biggest difference of divergences (speciation time and mutation rates) between two species among two trees shown in

Example 11. This is a very natural and desirable interpretation in terms of phylogenomics. The smaller difference of divergences between two species among the tree with an observed ultrametric x and the tree with the centriod has higher probability. Therefore, it is natural to apply a sample generated from the multi-species coalescent model where the species tree has the centroid as its dissimilarity map. It is worth noting that we do not know much about a well-defined distribution over the space of phylogenetic trees, despite many researchers' attempts (Woodman and Nye, 2025).

Example 11: [Tropical Metric] Suppose we have equidistant trees T_1 and T_2 with leaf

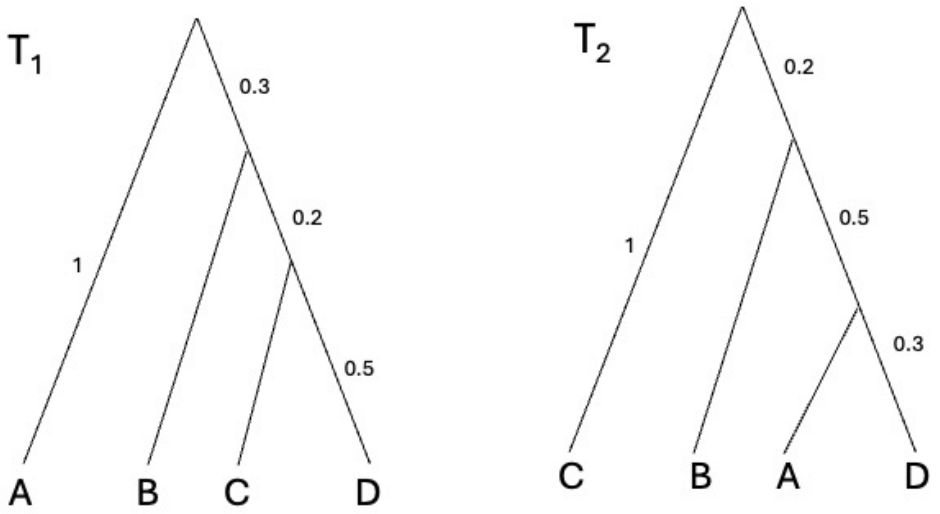


Figure 3.3.1: Visualization of trees T_1 and T_2.

labels $\{A, B, C, D\}$ shown in Fig. 3.3.1. Let u, v be dissimilarity vectors of trees T_1 and T_2 respectively,

$$\begin{aligned} u &= (2, 2, 2, 1.4, 1.4, 1), \\ v &= (1.6, 2, 0.6, 2, 1.6, 2). \end{aligned}$$

Their difference is

$$u - v = (0.4, 0, 1.4, -0.6, -0.2, -1),$$

and so their tropical distance is

$$d_{\text{tr}}(u, v) = \max_i (u - v)_i - \min_i (u - v)_i = 1.4 - (-1) = 2.4$$

Combining the result of Proposition 3.3.2 with Equations (3.3.2) and (3.3.4) yields

$$h_{\omega_0,\omega_1}(x) = \frac{d_{\text{tr}}(x, \omega_0)}{\sigma_0} - \frac{d_{\text{tr}}(x, \omega_1)}{\sigma_1} + (e - 1) \log \left(\frac{\sigma_0}{\sigma_1} \right). \tag{3.3.6}$$

In its most general form, the model parameters are $(\omega_0, \omega_1, \sigma_0, \sigma_1)$ so the parameter space is a subset of $(\mathbb{R}^e/\mathbb{R}\mathbf{1})^2 \times \mathbb{R}^2_+$ with dimension $2e$. Two instances of this general model are particularly practically useful and interpretable. We call these the one-species and two-species models and they will be our focus for tropical logistic regression in the rest of the paper.

For the *one-species model*, it is assumed that $\omega_0 = \omega_1$ and $\sigma_0 \neq \sigma_1$. If, without loss of generality, $\sigma_1 > \sigma_0$, equation (3.3.6) becomes

$$h_\omega(x) = \lambda \left(d_{\text{tr}}(x, \omega) - c \right), \tag{3.3.7}$$

where $\lambda = (\sigma_0^{-1} - \sigma_1^{-1})$ and $\lambda c = \log(\sigma_1/\sigma_0)$. Symbolically, the expression in equation (3.3.7) can be considered to be a scaled tropical inner product, whose direct analogue in classical logistic regression is the classical inner product $h_\omega(x) = \omega^T x$. See Section 3.C in the supplement for more details. The classifier is $C(x) = \mathbb{I}(d_{\text{tr}}(x, \hat{\omega}) > c)$, where $\hat{\omega}$ is the inferred estimator of ω^*, the true theoretical parameter. Note that the classification threshold and the probability contours $(p(x))$ are tropical circles, illustrated in Figure 3.5.1.

For the *two-species-tree model*, it is assumed that $\sigma_0 = \sigma_1$, and $\omega_0 \neq \omega_1$. Equation

(3.3.6) reduces to

$$h_{\omega_0,\omega_1}(x) = \sigma^{-1}(d_{\mathrm{tr}}(x,\omega_0) - d_{\mathrm{tr}}(x,\omega_1)), \tag{3.3.8}$$

with a classifier $C(x) = \mathbb{I}(d_{\mathrm{tr}}(x,\hat{\omega}_0) > d_{\mathrm{tr}}(x,\hat{\omega}_1))$, where $\hat{\omega}_y$ is the inferred tree for class $y \in \{0,1\}$. It is noted here that $\omega_0, \omega_1, \hat{\omega}_0, \hat{\omega}_1$ are not necessarily tree-like, but in the setting of phylogenetics they could be interpretted as such. The classification boundary is the tropical bisector which is extensively studied in Criado et al. (2021) between the estimators $\hat{\omega}_0$ and $\hat{\omega}_1$ and the probability contours are tropical hyperbolae with $\hat{\omega}_0$ and $\hat{\omega}_1$ as foci, as shown in Figure 3.5.3(right).

The one-species model is appropriate when the gene trees of both classes are concentrated around the same species tree ω with potentially different concentration rates. When the gene trees of each class come from distributions centered at different species trees the two-species model is preferred.

3.3.2 Model selection

In the previous subsection, we established the correspondence between the covariate conditional distribution and the function h which defines the logistic regression model. According to Proposition 3.3.1, the best regression model follows from the distribution that fits the data. The family of distributions that best fits the training data of a given class can indicate which regression model to use. The question that naturally arises is how to assess which family of conditional distributions has the best fit.

One issue is that the random covariates are multivariate and so the Kolmogorov–Smirnov test can not be readily applied. Moreover, the four families considered, namely the classical and tropical Laplace and Gaussian distributions, are not nested. Nonetheless, it is observed that for all these families the distances of the covariates from their centres are Gamma distributed. This is stated in Corollary 3.3.4 which is based on Proposition 3.3.3. Note that the distance metric corresponds to the geometry of the covariates. However, the arguments used in the proof of Corollary 3.3.4 do not work for distri-

butions defined on the space of ultrametric trees $\mathcal{U}_m$, because these spaces are not translation invariant. For a similar reason, the corollary does not apply to the BHV metric.

Proposition 3.3.3. *Consider a function $d : \mathbb{R}^n \to \mathbb{R}$ with $\alpha d(x) = d(\alpha x)$, for all $\alpha \geq 0$. If $X \sim f$ with $f(x) \propto \exp(-d^i(x)/(i\sigma^i))$ being a valid probability density function, for some $i \in \mathbb{N}$, $\sigma > 0$. Then, $d^i(X) \sim i\sigma^i \text{Gamma}(n/i)$, where $\theta\text{Gamma}(\alpha) \equiv \text{Gamma}(\alpha, \theta)$ is the Gamma distribution with shape α and scale θ.*

Remark 8: Throughout this chapter, $\theta\text{Gamma}(\alpha)$ and $\text{Gamma}(\alpha, \theta)$ are used interchangeably.

Corollary 3.3.4. *If $X \in \mathbb{R}^e$ with $X \sim f \propto \exp\left(-d^i(x, \omega^*)/(i\sigma^i)\right)$, where d is the Euclidean metric, then $d^i(X, \omega^*) \sim i\sigma^i \text{Gamma}(e/i)$. If $X \in \mathbb{R}^e/\mathbb{R}\mathbf{1}$ with $X \sim f \propto \exp\left(-d^i_{\text{tr}}(x, \omega^*)/(i\sigma^i)\right)$, where d_{tr} is the tropical metric, then $d^i_{\text{tr}}(X, \omega^*) \sim i\sigma^i \text{Gamma}((e-1)/i)$.*

In Section 3.5.3, the suitability of the tropical against the classical logistic regression is assessed for the coalescent model and the Mr Bayes trees, by visually comparing the fits of the theoretical Gamma distributions to Euclidean and tropical distances of the gene trees to the species tree.

3.3.3 Consistency and Generalization Error

In this subsection, the consistency of the statistical estimators (in Theorem 3.3.5) and of the tropical logistic regression as a learning algorithm (in Propositions 3.3.6 and 3.3.7) are established. Finally, the generalization error (probability of misclassification for unseen data) for the one-species model is derived and an upper bound is found for the generalization error of the two-species model. In both cases the error bounds are

getting better as the estimation error ϵ shrinks to zero. It is worth mentioning that in the case of exact estimation, the generalization error of the one-species model can be computed explicitly by equation (3.3.9). Moreover, there is a higher misclassification rate from the more dispersed class (inequality (3.3.10)).

Theorem 3.3.5 (Consistency)**.** *The estimator* $(\hat{\omega}, \hat{\sigma}) = (\hat{\omega}_0, \hat{\omega}_1, \hat{\sigma}_0, \hat{\sigma}_1) \in \Omega^2 \times \Sigma^2$ *of the parameter* $(\omega^*, \sigma^*) = (\omega_0^*, \omega_1^*, \sigma_0^*, \sigma_1^*) \in \Omega^2 \times \Sigma^2$ *is defined as the maximizer of the logistic likelihood function, where* $\Omega \subset \mathbb{R}^e/\mathbb{R}\mathbf{1}$ *and* $\Sigma \subset \mathbb{R}_+$ *are compact sets. Moreover, it is assumed that the covariate-response pairs* $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ *are independent and identically distributed with* $X_i \in \mathbb{R}^e/\mathbb{R}\mathbf{1}$*,* $d_{\mathrm{tr}}(X, \omega_Y)$ *being integrable and square-integrable and* $Y_i \sim$ *Bernoulli*$(S(h(X_i, (\omega^*, \sigma^*))))$*, where* S *is the sigmoid function. Then,*

$$(\hat{\omega}, \hat{\sigma}) \xrightarrow{p} (\omega^*, \sigma^*) \text{ as } n \to \infty.$$

In other words, the model parameter estimator is consistent.

Proposition 3.3.6 (One-species generalization error)**.** *Consider the one-species model where* $\omega = \omega_0 = \omega_1 \in \mathbb{R}^e/\mathbb{R}\mathbf{1}$ *and without loss of generality* $\sigma_0 < \sigma_1$*. The classifier is* $C(x) = \mathbb{I}(h_{\hat{\omega}}(x) \geq 0)$*, where* h *is defined in equation (3.3.7) and* $\hat{\omega}$ *is the estimate for* $\omega^\star$*. Define the covariate-response joint random variable* (X, Y) *with* $Z = \sigma_Y^{-1} d_{\mathrm{tr}}(X, \omega_Y^*)$ *drawn from the same distribution with cumulative density function* F*. Then,*

$$\begin{aligned}
&\mathbb{P}(C(X) = 1 | Y = 0) \in [1 - F(\sigma_1(\alpha + \epsilon)), 1 - F(\sigma_1(\alpha - \epsilon))], \\
&\mathbb{P}(C(X) = 0 | Y = 1) \in [F(\sigma_0(\alpha - \epsilon)), F(\sigma_0(\alpha + \epsilon))], \text{ where} \\
&\alpha = \frac{\log \frac{\sigma_1}{\sigma_0}}{\sigma_1 - \sigma_0}, \text{ and } \epsilon = (e-1)\frac{d_{\mathrm{tr}}(\hat{\omega}, \omega^*)}{\sigma_1 \sigma_0}.
\end{aligned}$$

The generalization error defined as $\mathbb{P}(C(X) \neq Y)$ *lies in the average of the two intervals above. In particular, note that if* $\hat{\omega} = \omega^*$*, then* $\epsilon = 0$ *and the intervals shrink to a single*

point, so the misclassification probabilities and generalization error can be computed explicitly.

$$\mathbb{P}(C(X) \neq Y) = \frac{1}{2}\left(1 - F(\sigma_1\alpha) + F(\sigma_0\alpha)\right) \tag{3.3.9}$$

Moreover, if $\hat{\omega} = \omega_*$ *and* $Z \sim \text{Gamma}(e-1, 1)$*, then*

$$\mathbb{P}(C(X) = 1|Y = 0) < \mathbb{P}(C(X) = 0|Y = 1). \tag{3.3.10}$$

Proposition 3.3.7 (Two-species generalization error)**.** *Consider the random vector* $X \in \mathbb{R}^e/\mathbb{R}\mathbf{1}$ *with response* $Y \in \{0, 1\}$ *and the random variable* $Z = d_{\text{tr}}(X, \omega_Y^*)$*. Assuming that the probability density function is* $f_X(x) \propto f_Z(d_{\text{tr}}(x, \omega_Y^*))$*, the generalization error satisfies the following upper bound*

$$\mathbb{P}\left(C(X) \neq Y\right) \leq \frac{1}{2}F_Z^C(\Delta_\epsilon) + h(\epsilon),$$

where $\epsilon = d_{\text{tr}}(\hat{\omega}_1, \omega_1^*) + d_{\text{tr}}(\hat{\omega}_0, \omega_0^*)$*,* $2\Delta_\epsilon = (d_{\text{tr}}(\omega_1^*, \omega_0^*) - \epsilon)$*,* F_Z^C *is the complementary cumulative distribution of* Z*, and* $h(\epsilon)$ *is an increasing function of* ϵ *with* $2h(\epsilon) \leq F_Z^C(\Delta_\epsilon)$ *and* $h(0) = 0$ *assuming that* $\mathbb{P}(d_{\text{tr}}(X, \omega_1^*)) = d_{\text{tr}}(X, \omega_{-1}^*)) = 0$*. Moreover, under the conditions of Theorem 3.3.5, our proposed learning algorithm is consistent.*

Observe that the upper bound is a strictly increasing function of ϵ.

Example 12: The complementary cumulative distribution of $\text{Gamma}(n, \sigma)$ is $F^C(x) = \Gamma(n, x/\sigma)/\Gamma(n, 0)$, where Γ is the upper incomplete gamma function and $\Gamma(n, 0) = \Gamma(n)$ is the regular Gamma function. Therefore, the tropical distribution given in equation (3.3.4) yields the following upper bound for the generalization error

$$\frac{\Gamma\left(e-1, \frac{d_{\text{tr}}(\omega_0^*, \omega_1^*)}{2\sigma}\right)}{2\Gamma(e-1)}, \tag{3.3.11}$$

under the assumptions of Proposition 3.3.7 and assuming that the estimators coincide with the theoretical parameters. This assumption is reasonable for large sample sizes and it follows from Theorem 3.3.5.

In subsequent sections, these theoretical results will guide us in implementing our model. Bounds on the generalization error from Propositions 3.3.6 and 3.3.7 are computed and the suitability of Euclidean and tropical distributions, and as a result of classical and tropical logistic regards, is assessed using the distance distribution of Proposition 3.3.3.

3.4 Optimization

As in the classical logistic regression, the parameter vectors $(\hat{\omega}, \hat{\sigma})$ maximising the log-likelihood (3.3.1), are chosen as statistical estimators. Identifying these requires the implementation of a continuous optimization routine. While root-finding algorithms typically work well for identifying maximum likelihood estimators in the classical logistic regression where the log-likelihood is concave, they are unsuitable here. The gradients of the log-likelihood under the proposed tropical logistic models are only piecewise continuous, with the number of discontinuities increasing along with the sample size. Furthermore, even if a parameter is found, it may merely be a local optimum. In light of this, the tropical Fermat-Weber problem of Lin and Yoshida (2018a) is revisited.

3.4.1 Fermat-Weber Point

A Fermat-Weber point or geometric mean $\tilde{\omega}_n$ of the sample set $(X_1, \ldots, X_n)$ is a point that minimizes the sum of distances from to sample points, i.e.

$$\tilde{\omega}_n \in \arg\min_{\omega} \sum_{i=1}^{n} d_{\mathrm{tr}}(X_i, \omega). \tag{3.4.1}$$

This point is rarely unique for finite n, indeed there will often be an infinite set of Fermat-Weber points (Lin and Yoshida, 2018a). However, the proposition below gives conditions for asymptotic convergence.

Proposition 3.4.1. *Let $X_i \overset{\text{iid}}{\sim} f$, where where f is a distribution that is symmetric around its center $\omega^* \in \mathbb{R}^e/\mathbb{R}\mathbf{1}$ i.e. $f(\omega^* + \delta) = f(\omega^* - \delta)$ for all $\delta \in \mathbb{R}^e/\mathbb{R}\mathbf{1}$. Let $\tilde{\omega}_n$ be any Fermat-Weber point as defined in equation (3.4.1). Then, $\tilde{\omega}_n \xrightarrow{p} \omega^*$ as $n \to \infty$.*

The significance of Proposition 3.4.1 is twofold. It proves that the Fermat-Weber sets of points sampled from symmetric distributions tend to a unique point. This is a novel result and ensures that for sufficiently large sample sizes the topology of any Fermat-Weber point is fixed. Additionally, using Theorem 3.3.5 and Proposition 3.4.1, $\hat{\omega}_n - \tilde{\omega}_n \xrightarrow{p} 0$ as $n \to \infty$. Furthermore, empirical evidence in Figure 3.5.4, see the following section, suggests that $d_{\text{tr}}(\hat{\omega}_n, \omega^*) = \mathcal{O}_p(1/\sqrt{n})$ and $d_{\text{tr}}(\tilde{\omega}_n, \omega^*) = \mathcal{O}_p(1/\sqrt{n})$. These statements are left as conjectures and proofs of them are beyond the scope of this paper. Assuming they hold and applying triangular inequality, it follows that $d_{\text{tr}}(\hat{\omega}_n, \tilde{\omega}_n) = \mathcal{O}_p(1/\sqrt{n})$. As a result, for a sufficiently large sample size we may use the Fermat-Weber point as an approximation for the MLE vector. Indeed, there are benefits in doing so.

Instead of having a single optimization problem with $2e-1$ variables, three simpler problems are considered; finding the Fermat-Weber point of each of the two classes, which has $e-1$ degrees of freedom and then finding the optimal σ which is a one dimensional root finding problem. The algorithms of our implementation for both model can be found in Supplement 3.D.

There is also another another benefit of using Fermat-Weber points. Proposition 3.4.2 provides a sufficient optimality condition that the MLE lacks, since a vanishing gradient in the log likelihood function merely shows that there is a local optimum.

Proposition 3.4.2. *Let* $X_1, \ldots, X_n \in \mathbb{R}^e/\mathbb{R}\mathbf{1}$, $\omega \in \mathbb{R}^e/\mathbb{R}\mathbf{1}$ *and define the function*

$$f(\omega) = \sum_{i=1}^{n} d_{\mathrm{tr}}(X_i, \omega).$$

i. The gradient vector of f *is defined at* ω *if and only if the vectors* $\omega - X_i$ *have unique maximum and minimum components for all* $i \in [n]$.

ii. If the gradient of f *at* ω *is well-defined and zero, then* ω *is a Fermat-Weber point.*

In Lin and Yoshida (2018a), Fermat-Weber points are computed by means of linear programming, which is computationally expensive. Employing a gradient-based method is much faster, but there is no guarantee of convergence. Nevertheless, if the gradient, which is an integer vector, vanishes, then it is guaranteed, as above, that the algorithm has reached a Fermat-Weber point. This tends to happen rather frequently, but not in all cases examined in Section 3.5.

Remark 9: Our choice of Fermat-Weber points to represent centers is not the only practical option, however it is an especially desirable choice due to the interpretability of its resulting solutions.

Recently, Comăneci and Joswig studied tropical Fermat-Weber points obtained using the asymmetric tropical distance (Comaneci and Joswig, 2023). They found that if all X_i are ultrametric, then the resulting tropical Fermat-Weber points are also ultrametric, all with the same tree topology. On the other hand, Lin et al. (2017) show that a tropical Fermat-Weber point defined with d_{tr} of a sample taken from the space of ultrametrics could fall outside of the ultrametric space.

Despite this, the major drawback of using the asymmetric tropical distance, is that it would result in losing the phylogenetic interpretation of the distance or dissimilarity between two trees held by the tropical metric d_{tr} - see Remark 7.

3.5 Results

In this section, tropical logistic regression is applied in three different scenarios. The first and simplest considers datapoints generated from the tropical Laplace distribution. Secondly, gene trees sampled from a coalescent model are classified based on the species tree they have been generated from, and finally it is applied as an MCMC convergence criterion for the phylogenetic tree construction, using output from the `Mr Bayes` software. The models' performance in terms of misclassification rates and AUCs on these datasets is examined.

3.5.1 Toy Example

In this example, a set of data points is generated from the tropical normal distribution as defined in Equation (3.3.4) using rejection sampling.

The data points are defined in the tropical projective torus $\mathbb{R}^e/\mathbb{R}\mathbf{1}$, which is isomorphic to $\mathbb{R}^{e-1}$. To map $x \in \mathbb{R}^e/\mathbb{R}\mathbf{1}$ to $\mathbb{R}^{e-1}$, simply set the last component of x to 0, or in other words $x \mapsto (x_1 - x_e, x_2 - x_e, \ldots, x_{e-1} - x_e)$. For illustration purposes, it is desirable to plot points in $\mathbb{R}^2$, so we use $e = 3$ which corresponds to phylogenetic trees with 3 leaves. Both the one-species model and the two-species model are examined.

In the case of the former, $\omega = \omega_0 = \omega_1$ and $\sigma_0 \neq \sigma_1$. The classification boundary in this case is a tropical circle. If $\sigma_0 < \sigma_1$, the algorithm classifies points close to the inferred centre to class 0 and those that are more dispersed away from the centre as class 1. For simplicity, the centre is set to be the origin $\omega = (0, 0, 0)$ and no inference is performed. In Figure 3.5.1 a scatterplot of of the two classes is shown, where misclassified points are highlighted. As anticipated from Proposition 3.3.6 there are more misclassified points from the more dispersed class (class 1). Out of 100 points for each class, there are 7 and 21 misclassified points from class 0 and 1 respectively, while the theoretical probabilities calculated from equation (3.3.9) of Proposition 3.3.6 are

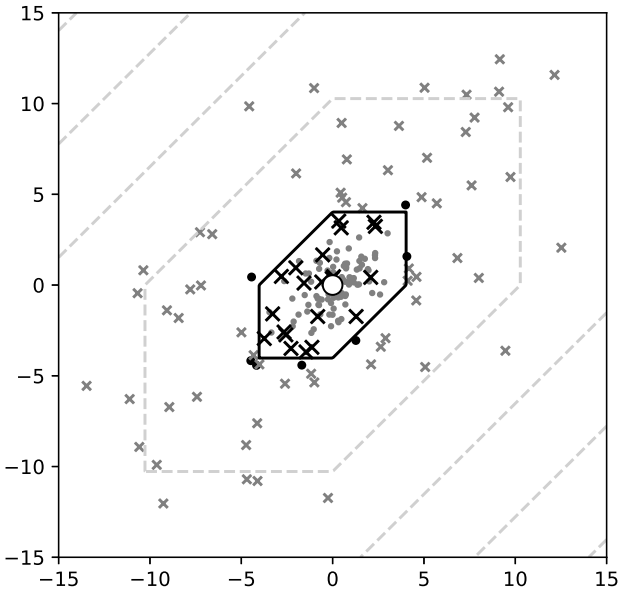


Figure 3.5.1: Scatterplot of 200 points - 100 dots for class 0 and 100 Xs for class 1, black for misclassified and grey otherwise - imposed upon a contour plot of the probability of inclusion in class 0, where the black contour is the classification threshold. The deviation parameters used in data generation were $\sigma_0 = 1, \sigma_1 = 5$ and the centre of the distribution (white-filled point) is the origin. The centres of the two distributions are $\omega_0 = \omega_1$.

9% and 19% respectively.

Varying the deviation ratio σ_1/σ_0 in the data generation process allows exploration of its effect on the generalization error in the one-species model. The closer this ratio is to unity, the higher the generalization error. For $\sigma_0 = \sigma_1$ the classes are indistinguishable and hence any model is as good as a random guess i.e. the generalization error is $1/2$. The estimate of the generalization error for every value of that ratio is the proportion of misclassified points in both classes. Assuming an inferred ω that differs from the true parameter, Fig. 3.5.2(left) verifies the bounds of Proposition 3.3.6.

For the two-species model, tropical logistic regression is directly compared to classical logistic regression. Data is generated using different centres $\omega_0 = (0, 0, 0)$, $\omega_1 = (3, 2, 0)$ but the same $\sigma = 0.5$. The classifier is $C(x) = \mathbb{I}(h(x) > 0)$ for both methods,

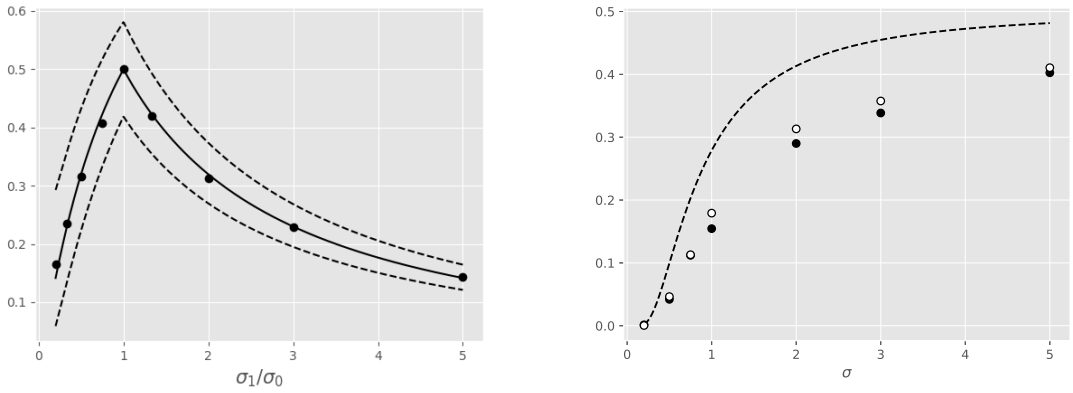


Figure 3.5.2: (left) Generalization error for 9 different deviation ratios. The estimator $\hat{\omega} = (0.3, 0, 3)$ differs from the true parameter $\omega = (0, 0)$. The upper and lower bounds of Proposition 3.3.6 are plotted in dashed lines and the generalization error for the correct estimator $\hat{\omega} = \omega^*$ plotted in solid line. The dots represent the proportion of misclassified points from a set of 2000 points in each experiment, 1000 points for each class. (right) Generalization errors for 7 different dispersion parameters with black markers for the two-species tropical logistic regression and white markers for the classical logistic regression. The upper bound (3.3.11) of Proposition 3.3.7 is plotted in dashed line.

using h as defined in equations (3.3.3) and (3.3.8) for the classical and tropical logistic regression respectively. Fig. 3.5.3 compares contours and classification thresholds of the classical (left) and tropical (right) logistic regression by overlaying them on top of the same data. Out of 100+100 points there are 5+4 and 4+3 misclassifications in classical and tropical logistic regression respectively. Fig. 3.5.2(right) visualizes the misclassification rates of the two logistic regression methods for different values of dispersion σ, showing the tropical logistic regression to have consistently lower generalization error than the classical, even in this simple toy problem.

Finally, we investigate the convergence rate of the Fermat-Weber points and of the MLEs from the two-species model as the sample size N increases. Fixing $\omega_0^* = (0, 0, 0)$ and $\omega_1^* = (3, 2, 0)$ as before, the Fermat-Weber point numerical solver and the log-likelihood optimization solver are employed to find $(\tilde{\omega}_0)_N$ and $((\hat{\omega}_0)_N, (\hat{\omega}_1)_N, \hat{\lambda}_N)$ respectively. From this, the error is computed for the two methods, which is defined

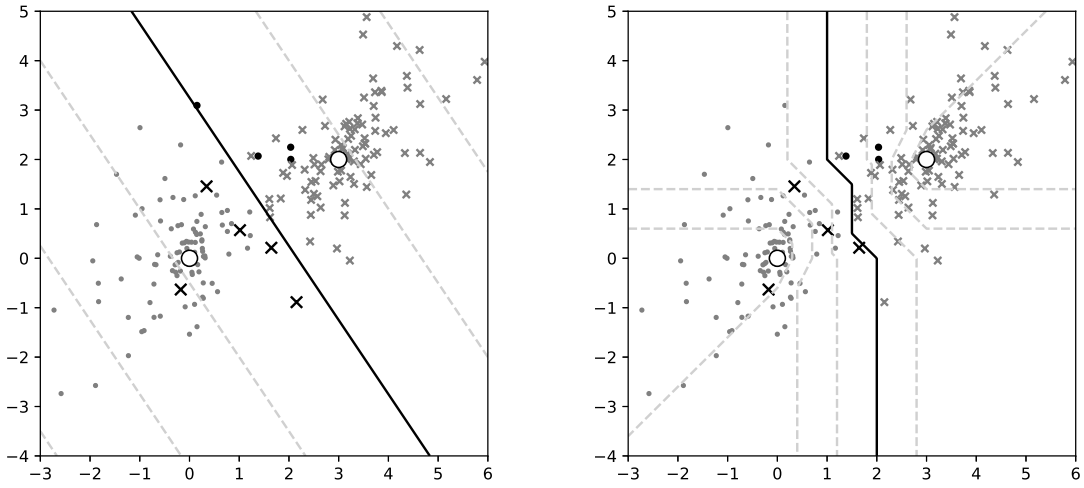


Figure 3.5.3: Scatterplot of points - dots for class 0 and X for class 1, black for misclassified according to (left) **classical logistic regression** or (right) **tropical logistic regression**, and grey otherwise - alongside a contour plot of the probabilities, where the black contour is the classification threshold. The centres, drawn as big white dots, are $\omega_0 = (0, 0, 0), \omega_1 = (3, 2, 0)$ and $\sigma = 0.5$.

as $d_N = d_{\text{tr}}((\omega_0)_N, \omega_0^*)$ for $(\omega_0)_N = (\tilde{\omega}_0)_N$ and $(\hat{\omega}_0)_N$ respectively. For each N, we repeat this procedure 100 times to get an estimate of the mean error rate $r_N = \mathbb{E}(d_N)$. Figure 3.5.4 shows that for both methods, $r_N\sqrt{N} \to C$ as $N \to \infty$, with $C_{\text{FW}} < C_{\text{MLE}}$. Since $\mathbb{E}(\sqrt{N}d_N) \to C$, it follows that $\sqrt{N}d_N = \mathcal{O}_p(1)$ as $N \to \infty$. This supports the assumption of Section 3.4 that Fermat-Weber points can be used in lieu of MLEs, since they converge to each other in probability at rate $1/\sqrt{N}$. Interestingly, the MLEs produce higher errors than FW points. This may be due to an imperfection of the MLE solver, which may be stuck at a local optimum.

3.5.2 Coalescent Model

The data that have been used in our simulations were generated under the multispecies coalescent model, using the python library `dendropy` (Sukumaran and Holder, 2010). The classification method we propose is the two-species model because two distinct species tree have been used to generate gene tree data for each class.

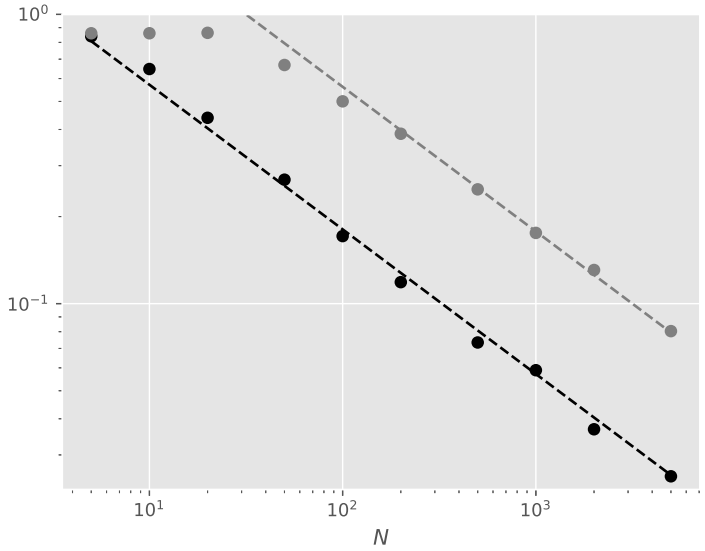


Figure 3.5.4: Expected asymptotic error for FW points $(\tilde{\omega}_0)_N$ (in black) and MLE points $(\hat{\omega}_0)_N$ (in grey) for different values of N. Error is defined as the tropical distance from the true centre ω_0^* i.e. $d_{\text{tr}}(\omega_N, \omega_0^*)$. The dashed lines are $y \propto N^{-0.5}$, so this figure illustrates that $d_{\text{tr}}((\omega_0)_N, \omega_0^*) = \mathcal{O}_p(1/\sqrt{N})$ as $N \to \infty$.

Two distinct species trees are used, which were randomly generated under a Yule process, a stochastic pure-birth model used to generate random branching trees where each species is equally likely to speciate i.e. divide in two. Then, using `dendropy`, 1000 gene trees are randomly generated for each of the two species. The trees have 10 leaves and so the number of the model variables is $\binom{10}{2} = 45$. They are labelled according to the species tree they are generated from. The tree generation is under the coalescent model for specific model parameters.

Since the species trees are known, we conduct a comparative analysis between classical, tropical and a BHV-based (Billera et al. (2001)) logistic regression. In the supplement, we show an approximation analog of our model to the BHV metric. The comparative analysis includes the distribution fitting of distances and the misclassification rates for different metrics.

In Fig. 3.5.5, the distribution of the radius $d(X, \omega)$ as given by Proposition 3.3.3, is fitted to the histograms of the Euclidean and tropical distances of gene trees to

their corresponding species tree, along with the corresponding pp-plots on the right. According to Proposition 3.3.3, for both the classical and tropical Laplace distributed covariates, $d(X, \omega^*) \sim \sigma\text{Gamma}(n)$, shown in solid lines in Fig. 3.5.5, where $n = e = 45$ and $n = e-1 = 44$ for the classical and tropical case respectively. Similarly, for normally distributed covariates, $d(X, \omega^*) \sim \sigma\sqrt{\chi^2_n}$, shown in dashed lines. It is clear that Laplace distributions produce better fits in both geometries and that the tropical Laplace fits the data best. As discussed in Section 3.3.2, the same analysis can not be applied to the BHV metric, because the condition of Proposition 3.3.3 does not hold.

Species depth SD is the time since the speciation event between the species and *effective population size* N quantifies genetic variation in the species. Datasets have been generated for a range of values $R := \text{SD}/N$ by varying species depth. For low values of R, speciation happens very recently and so the gene trees look very much alike. Hence, classification is hard for datasets with low values of R and vice versa, because the gene deviation σ_R is a decreasing function of R. We expect classification to improve in line with R. Fig. 3.H.2 and Fig. 3.H.1 in Supplement 3.H confirm that, by showing that as R increases the receiver operating characteristic (ROC) curves are improving and the Robinson-Foulds and tropical distances of inferred (Fermat-Weber point) trees are decreasing. In addition, Fig. 3.5.6 shows that as R increases, AUCs increase (left) and misclassification rates decrease (right). It also shows that tropical logistic regression produces higher AUCs than classical logistic regression and other out-of-the-box ML classifiers such as random forest classifier, neural networks with a single sigmoid output layer and support vector machines. Our model also produces lower misclassification rates than classical logistic regression. Finally, note that the generalization error upper bound as given in equation (3.3.11) is satisfied but it not very tight (dashed line in Fig. 3.5.6).

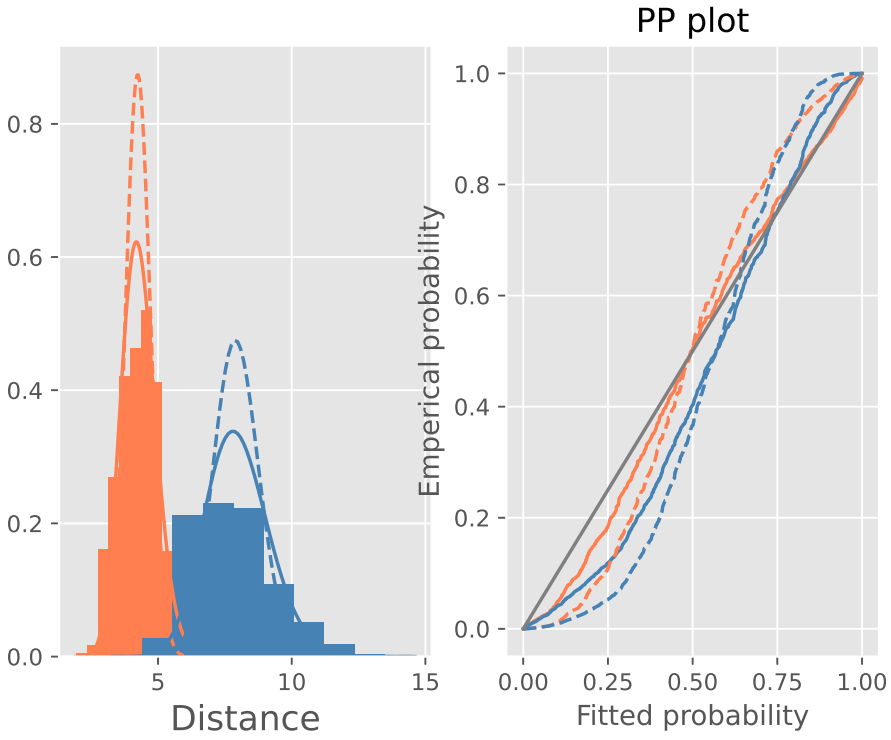


Figure 3.5.5: (Left) Histograms of the distances of 1000 gene trees from the species trees that generated them under the coalescent model with $R = 0.7$. Coral and blue corresponds to tropical and euclidean geometries respectively. The solid and dashed lines are fitted distributions $\sigma\text{Gamma}(n)$ and $\sigma\sqrt{\chi^2_n}$ respectively; σ is chosen to be the MLE, derived in the supplement. Euclidean metric has worse fit than the tropical metric. This can also be observed by the corresponding pp-plots (right).

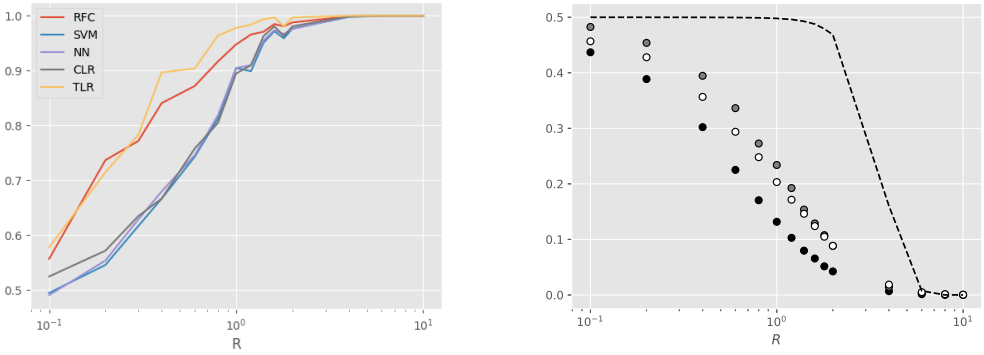


Figure 3.5.6: (left) Average AUCs against R. Five classification models which we considered are the tropical two species-tree model (TLR), random forest classifier (RFC), support vector machines (SVM), neural networks (NN) and classical logistic regression (CLR). We used default set up for TLR, SVM, NN and CLR implemented by `sklearn`.
(right) the x-axis represents the ratio R and the y-axis represents misclassification rates. Black circles represent the tropical logistic regression, white circles represent the classical logistic regression, grey points represent the logistic regression with BHV metric, and the dashed line represents the theoretical generalization error shown in Proposition 3.3.7. It is noted here that there was a flaw in the implementation of BHV regression, by erroneously assuming that the normalization constant of the two BHV distributions are equal (see more details in the Appendix).

3.5.3 Convergence of Mr Bayes

`Mr Bayes` (Huelsenbeck et al. (2001)) is a widely used software for Bayesian inference of phylogeny using MCMC to sample the target posterior distribution. An important feature of the software is the diagnostic metrics indicating whether a chain has converged to the equilibrium distribution. This is calculated at regular, specified intervals, set by the variable `diagnfreq`, using the average standard deviation of split frequencies (ASDSF introduced by Lakner et al. (2008)) between two independently run chains. The more similar the split frequencies between the two chains are, the lower the ASDSF, and the more likely it is that both chains have reached the equilibrium distribution.
Our classification model provides an alternative convergence criterion for MCMC convergence. Consider two independently run chains; the sampled trees of the two chains correspond to two classes and the AUC value is a measure of how distinguishable the two chains are. High values of AUC are associated with easily distinguishable chains,

implying that the chains have not converged to the equilibrium distribution. At every iteration that is a multiple of `diagnfreq`, the ASDSF metric is calculated and the AUC of the two chains is found by applying tropical logistic regression to the truncated chains that only keep the last 30% of the trees in each chain.

For our comparison study, the data used were the gene sequences from the `primates.nex` file. This dataset comes with the `Mr Bayes` software and it is used as an example in Ronquist et al. (2005). Figure 3.5.7 shows the two metrics at different iterations of the two independent chains ran on this dataset. According to the `Mr Bayes` manual, the convergence threshold for their metric is 10^{-2}. This is achieved at the 800-th iteration, when our method produces an AUC of 97%, which indicates that the chains may have not converged yet, contrary to the suggestion of Mr Bayes. A likely explanation for this discrepancy is the dependence of ASDSF on tree topologies instead of branch lengths. The frequencies of the tree topologies may have converged to those of the equilibrium distribution, even if the branch lengths have not. Eventually, the AUC values drop rapidly when iterations exceed $2 \cdot 10^3$, while the ASDSF metric is reduced at a much slower rate. In this second phase, the branch lengths are calibrated, while the topology frequencies do not change a lot. Finally, for iterations that exceed 10^5, neither metric can reject convergence, with ASDSF being 10 lower than the threshold and the AUC values finally dropping below 70%, which is a typical threshold for poor classification.

When our classification method is compared to other classifiers, it marginally outperforms classical logistic regression and neural networks with a single sigmoid output but underperforms support vector machines and random forest classifiers. Despite their simplicity, logistic regression models cannot capture the complexity of the chain classification problem. More advanced statistical methods that conform to tropical geometry (such as tropical support vector machines Yoshida et al. (2023d)) could be applied instead at the cost of simplicity and interpretability.

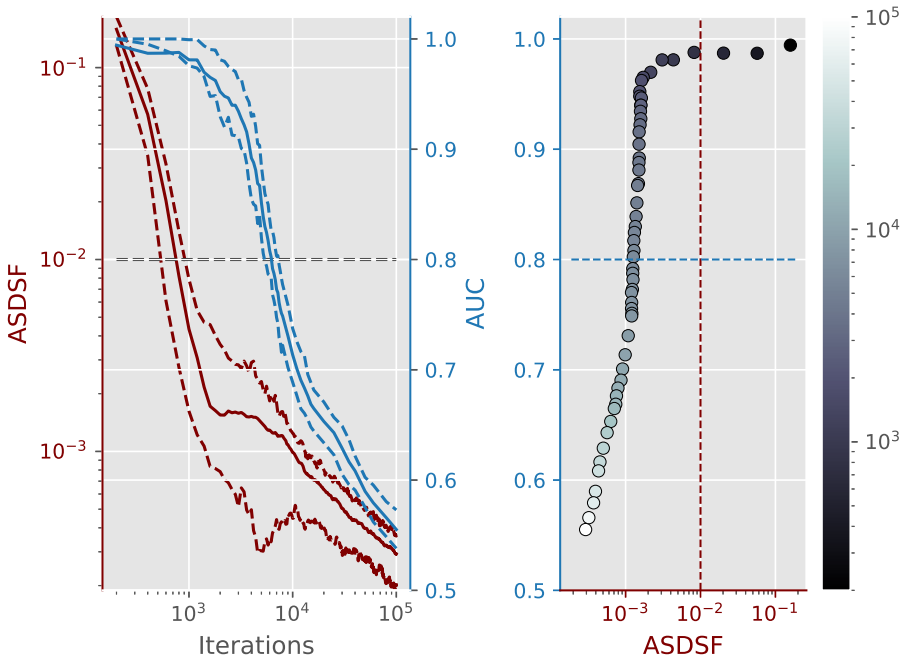


Figure 3.5.7: (Left) Average ASDSF (in red) and AUC (in blue) values plotted against the number of iterations of the MCMC chains. The coloured dashed lines correspond to the first and third quartile. The grey dashed line indicates the `Mr Bayes` threshold for ASDSF and our provisional AUC threshold of 80%. (Right) ASDSF and AUC values plotted against each other, with the iterations coloured according to the colourbar and the dashed lines corresponding to the thresholds for each metric.

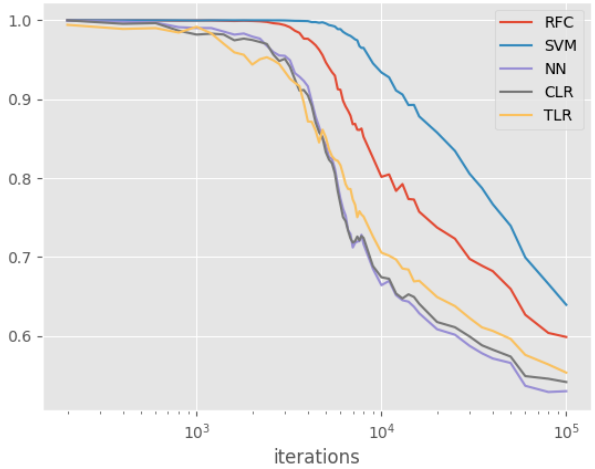


Figure 3.5.8: Average AUC values plotted against the number of MCMC iterations for the 5 supervised learning methods considered.

3.6 Discussion

In this paper we developed a tropical analog of the classical logistic regression model and considered two special cases; the one species-tree model and two species-tree model. In our empirical work the two-species model was most effective, but we anticipate both are potentially impactful tools for phylogenomic analysis. The one-species model's principal benefit is having the same number of parameters as the number of predictors, unlike the two-species model which has almost twice as many. Therefore, the one-species model more readily fits the standard definition of a generalized linear model and could generalize to a stack of GLMs to produce a "tropical" neural network, which is investigated in Yoshida et al. (2023a).

The two-species model implemented on data generated under the coalescent model outperformed classical and BHV logistic regression models in terms of misclassification rates, AUCs and fitness of the distribution of distances to their centre. It was also observed that Laplace distributions were better fitting than Gaussians, for both geometries. Empirically selecting tropical distributions over Euclidean distributions suffices for the scope of this paper, but further theoretical justification of the suitability of such distributions is needed. Moreover, further research on the generalization error for the two-species model would provide tighter bounds.

Finally, the AUC metric of our model is proposed as an alternative to the ASDSF metric for MCMC convergence checking. Our metric is more conservative and robust, taking branch lengths into account. Nonetheless, computing the ASDSF is less computationally intensive than running our method. There seems to be a tradeoff between the reliability of the convergence criterion tool and computational speed. Further research can shed light on the types of datasets where the ASDSF metric becomes unreliable. Then, the two metrics could complement each other, with our methods applied only when there is a good indication that ASDSF is unreliable.

3.A Proofs

Proof of Lemma 1. A simple application of the Bayes rule for continuous random variables yields

$$\begin{aligned} p(x) = \mathbb{P}(Y=1|X=x) &= \frac{f_1(x)\mathbb{P}(Y=1)}{f_0(x)\mathbb{P}(Y=0)+f_1(x)\mathbb{P}(Y=1)} \\ &= \frac{1}{1+\frac{f_1(x)(1-r)}{f_0(x)r}} = S(h(x)). \end{aligned}$$

□

Proof of Proposition 3.3.1. The expected log-likelihood is expressed as

$$\begin{aligned} \mathbb{E}(l) &= \mathbb{E}\left(Y\log(p(X)) + (1-Y)\log(1-p(X))\right) \\ &= \mathbb{P}(Y=1)\int_{\mathbb{R}^n} f_1(x)\log(p(x))\,dx \\ &+ \mathbb{P}(Y=0)\int_{\mathbb{R}^n} f_0(x)\log(1-p(x))\,dx \\ &= \int_{\mathbb{R}^n} L(x,p(x))\,dx, \end{aligned}$$

where $L(x,p) = rf_1(x)\log(p)+(1-r)f_0(x)\log(1-p)$ is treated as the Lagrangian. The Euler-Lagrange equation can be generalized to a several variables (in our case there are n variables). Since there are no derivatives of p, the stationary functional satisfies $\partial_p L = 0$, which yields the desired result. □

Proof of Proposition 3.3.3. The pdf of X is

$$f_\omega(x) = \frac{1}{C_\alpha}\exp\left(-\alpha^i\frac{d^i(x)}{i}\right), x \in \mathbb{R}^n$$

where $\alpha = \sigma^{-1}$ is the precision. Using the variable transformation $y = \alpha x$ with Jacobian

$1/\alpha^n$ and remembering that $\alpha d(x) = d(y)$,

$$C_\alpha = \int_{\mathbb{R}^n} \exp\left(-\alpha^i \frac{d^i(x)}{i}\right) dx = \int_{\mathbb{R}^n} \exp\left(-\frac{d^i(x)}{i}\right) \frac{dy}{\alpha^n} = \frac{C_1}{\alpha^n}.$$

The moment generating function of $d^i(X)$ is

$$\begin{aligned} M_{d^i(X)} &= \int_{\mathbb{R}^n} \exp\left(z d^i(x)\right) \frac{\exp\left(-\alpha^i \frac{d^i(x)}{i}\right)}{C_\alpha} dx \\ &= \frac{C_{\sqrt[i]{\alpha^i/i - z}}}{C_\alpha} = \frac{1}{\left(\sqrt[i]{1 - i\sigma^i z}\right)^n}, \end{aligned}$$

which coincides with the MGF of $\Gamma(n/i, i\sigma^i)$. □

Proof of Proposition 3.3.2. From the proof of Proposition 3.3.3, it was established that the normalizing constant is $C_{\sigma_Y} = C_1 \sigma_Y^{e-1}$ for the tropical projective torus, whose dimension is $n = e - 1$.

The volume of a unit tropical sphere in the tropical projective torus $\mathbb{R}^e/\mathbb{R}\mathbf{1}$ is equal to e. If the tropical radius is r, then the volume is er^{e-1} and hence the surface area is $e(e-1)r^{e-2}$. Therefore,

$$\begin{aligned} C_1 &= \int_{\mathbb{R}^e/\mathbb{R}\mathbf{1}} \exp\left(-d_{\mathrm{tr}}(x, \mathbf{0})\right) dx \\ &= \int_0^\infty e(e-1) r^{e-2} \exp\left(-r\right) dr \\ &= e(e-1)\Gamma(e-1) = e! \end{aligned}$$

It follows that the normalizing constant is $C_{\sigma_Y} = e! \sigma_Y^{e-1}$. □

Proof of Corollary 3.3.4. Suppose that X comes from the Laplace or the Normal distribution, whose pdf is proportional to $\exp\left(-d^i(x, \omega^*)/(i\sigma^i)\right)$ for $i = 1$ and 2 respectively, for all $x \in \mathbb{R}^n$ where d is the Euclidean metric. Then, $X - \omega^*$ has a distribution proportional to $\exp\left(-d^i(x, \mathbf{0})/(i\sigma^i)\right)$. Clearly, $\alpha d(x, \mathbf{0}) = d(\alpha x, \mathbf{0})$ and so from Propo-

sition 3.3.3, it follows that $d^i(X - \omega^*, \mathbf{0}) = d^i(X, \omega^*) \sim i\sigma^i \text{Gamma}(n/i)$. Note that for the normal distribution ($i = 2$), $d^i(X, \omega^*) \sim \sigma^2 \chi_{n/2}$. The same argument applies for tropical Laplace and tropical Normal distributions, where the metric is tropical ($d = d_{\text{tr}}$), the distribution is defined on $\mathbb{R}^e/\mathbb{R}\mathbf{1} \cong \mathbb{R}^{e-1}$ and the dimension is hence $n = e - 1$. □

Prerequisites for proof of Theorem 3.3.5

Theorem 3.A.1. *(Theorem 4.2.1 in Bierens (1996)) Let $(Q_n(\theta))$ be a sequence of random functions on a compact set $\Theta \subset \mathbb{R}^m$ such that for a continuous real function $Q(\theta)$ on Θ,*

$$\sup_{\theta \in \Theta} |Q_n(\theta) - Q(\theta)| \xrightarrow{p} 0 \text{ as } n \to \infty.$$

Let θ_n be any random vector in Θ satisfying $Q_n(\theta_n) = \inf_{\theta \in \Theta} Q_n(\theta)$ and let θ_0 be a unique point in Θ such that $Q(\theta_0) = \inf_{\theta \in \Theta} Q(\theta)$. Then $\theta_n \xrightarrow{p} \theta_0$.

Theorem 3.A.2. *(Lemma 2.4 in Newey and McFadden (1994)) If the data $z_1, \ldots, z_n$ are independent and identically distributed, the parameter space Θ is compact, $f(z_i, \theta)$ is continuous at each $\theta \in \Theta$ almost surely and there is $r(z) \geq |f(z, \theta)|$ for all $\theta \in \Theta$ and $\mathbb{E}(r(z)) < \infty$, then $\mathbb{E}(f(z, \theta))$ is continuous and*

$$\sup_{\theta \in \Theta} \left| n^{-1} \sum_{i=1}^{n} f(z_i, \theta) - \mathbb{E}(f(z, \theta)) \right| \xrightarrow{p} 0.$$

Lemma 3.A.3. *Consider two points $x, y \in \mathbb{R}^e/\mathbb{R}\mathbf{1}$. There exists $\eta > 0$ such that*

$$d_{\text{tr}}(x + \epsilon E_i, y) = d_{\text{tr}}(x, y) + \epsilon \phi_i(x - y),\ \forall \epsilon \in [0, \eta],\ \forall i \in [e],\ \textit{where}$$

$$\phi_i(v) = \begin{cases} 1, & \textit{if } v_i \geq v_j\ \forall j \in [e] \\ -1, & v_i < v_j\ \forall j \in [e] \backslash \{i\} \\ 0, & \textit{otherwise} \end{cases},$$

and $E_i \in \mathbb{R}^e/\mathbb{R}\mathbf{1}$ *is a vector with* 1 *in the* i*-th coordinate and* 0 *elsewhere.*

Proof. By setting $v := x - y$, $M := \max_{j\in[e]}\{v_j\}$ and $m := \min_{j\in[e]}\{v_j\}$,

$$d_{\text{tr}}(x, y) = M - m$$
$$d_{\text{tr}}(x + \epsilon E_i, y) = \max_{j\in[e]}\{v_j + \epsilon\delta_{ij}\} - \min_{j\in[e]}\{v_j + \epsilon\delta_{ij}\},$$

where $\epsilon \geq 0$, and $\delta_{ij} = \mathbb{I}(i = j)$ with $\mathbb{I}$ being the indicator function. Three separate cases are considered.

i. If $v_i = M$, then

$$\max_{j\in[e]}\{v_j + \epsilon\delta_{ij}\} = v_i + \epsilon = M + \epsilon, \tag{3.A.1}$$
$$\min_{j\in[e]}\{v_j + \epsilon\delta_{ij}\} = m, \tag{3.A.2}$$

and so $d_{\text{tr}}(x + \epsilon E_i, y) = d_{\text{tr}}(x, y) + \epsilon$. Note that equations (3.A.1) and (3.A.2) hold for all $\epsilon > 0$.

ii. If $v_i = m$ **and** $v_i < v_k$ for all $k \neq i$, i.e. if v_i is the **unique** minimum component of vector v, then

$$\max_{j\in[e]}\{v_j + \epsilon\delta_{ij}\} = M, \text{ for all } \epsilon \leq M - m \tag{3.A.3}$$
$$\min_{j\in[e]}\{v_j + \epsilon\delta_{ij}\} = v_i + \epsilon = m + \epsilon, \text{ for all } \epsilon \leq m' - m, \tag{3.A.4}$$

where $m' := \min_{j:v_j>m}\{v_j\} > m$ is well-defined unless $v_j = m$ for all $j \in [e]$ i.e. for $v = m \cdot (1, \ldots, 1) = \mathbf{0}$, which falls under the first case. Clearly, $M \geq m'$, so for all $\epsilon \in [0, m' - m]$ equations (3.A.3) and (3.A.4) are satisfied and thence $d_{\text{tr}}(x + \epsilon E_i, y) = d_{\text{tr}}(x, y) - \epsilon$.

iii. Otherwise, if none of the first two cases hold then $\exists k \neq i$ such that $m = v_k \leq v_i <$

M and so

$$\min_{j\in[e]}\{v_j + \epsilon\delta_{ij}\} = v_k = m, \text{ for all } \epsilon > 0 \tag{3.A.5}$$

$$\max_{j\in[e]}\{v_j + \epsilon\delta_{ij}\} = M, \text{ if } \epsilon \leq M - v_i \tag{3.A.6}$$

Define $M' \coloneqq \max_{j:v_j<M}\{v_j\} < M$ which is well-defined for all $v \neq \mathbf{0}$ (first case). Since $v_i < M$, it follows by definition that $v_i \leq M'$ and so $M - v_i \geq M - M' > 0$. As a result, for all $\epsilon \in [0, M - M']$, equations (3.A.5) and (3.A.6) are satisfied and thence $d_{\text{tr}}(x + \epsilon E_i, y) = d_{\text{tr}}(x, y)$.

If $v = \mathbf{0}$, set $\eta = +\infty$. Otherwise, for $v \neq \mathbf{0}$, with m', M' being well-defined, set

$$\eta = \min(m' - m, M - M') > 0.$$

In all three cases and for all $\epsilon \in [0, \eta]$ the desired result is satisfied. □

Lemma 3.A.4. *Consider the function* $q : \mathbb{R}^e/\mathbb{R}\mathbf{1} \to \mathbb{R}$,

$$\begin{aligned} q(x) = \lambda_\alpha d_{\text{tr}}(x, \alpha) - \lambda_\beta d_{\text{tr}}(x, \beta) - \lambda_\gamma d_{\text{tr}}(x, \gamma) + \lambda_\delta d_{\text{tr}}(x, \delta) \\ + \log\left(\frac{\lambda_\beta}{\lambda_\alpha}\right) - \log\left(\frac{\lambda_\delta}{\lambda_\gamma}\right), \end{aligned}$$

where $\alpha, \beta, \gamma, \delta \in \mathbb{R}^e/\mathbb{R}\mathbf{1}$, $\lambda_\alpha, \lambda_\beta, \lambda_\gamma, \lambda_\delta > 0$ *and* $(\alpha, \lambda_\alpha) \neq (\beta, \lambda_\beta)$. *A set* $\mathcal{X}$ *contains neighbourhoods of* $\alpha, \beta, \gamma, \delta$. *If* $q(x) = 0$, $\forall x \in \mathcal{X}$ *then* $(\alpha, \lambda_\alpha) = (\gamma, \lambda_\gamma)$ *and* $(\beta, \lambda_\beta) = (\delta, \lambda_\delta)$.

Proof. According Lemma 3.A.3, there exists $\eta_1 > 0$ such that for all $\epsilon \in [0, \eta_1]$

$$d_{\text{tr}}(x + \epsilon E_i, y) = d_{\text{tr}}(x, y) + \epsilon\phi_i(x - y). \tag{3.A.7}$$

Moreover, $d_{\text{tr}}(x - \epsilon E_i, y) = d_{\text{tr}}(y, x - \epsilon E_i) = d_{\text{tr}}(y + \epsilon E_i, x)$ and so using Lemma 3.A.3

again (but with x and y swapped), there exists $\eta_2 > 0$ such that for all $\epsilon \in [0, \eta_2]$

$$d_{\mathrm{tr}}(x - \epsilon E_i, y) = d_{\mathrm{tr}}(x, y) + \epsilon\phi_i(y - x), \tag{3.A.8}$$

for all $\epsilon \in [0, \epsilon_0(y - x)]$. For all $\epsilon \in [0, \eta]$ where $\eta \coloneqq \min(\eta_1, \eta_2)$, equations (3.A.7), (3.A.8) are satisfied and so

$$\begin{aligned}
&q(x + \epsilon E_i) = q(x) + \\
&\quad \epsilon\left(\lambda_\alpha\phi_i(x - \alpha) - \lambda_\beta\phi_i(x - \beta) - \lambda_\gamma\phi_i(x - \gamma) + \lambda_\delta\phi_i(x - \delta)\right), \\
&q(x - \epsilon E_i) = q(x) + \\
&\quad \epsilon\left(\lambda_\alpha\phi_i(\alpha - x) - \lambda_\beta\phi_i(\beta - x) - \lambda_\gamma\phi_i(\gamma - x) + \lambda_\delta\phi_i(\delta - x)\right).
\end{aligned}$$

Consequently, for all $\epsilon \in [0, \eta]$,

$$\begin{aligned}
&q(x + \epsilon E_i) + q(x - \epsilon E_i) - q(x) = 0 \\
&\quad = \epsilon\left(\lambda_\alpha s_i(x - \alpha) - \lambda_\beta s_i(x - \beta) - \lambda_\gamma s_i(x - \gamma) + \lambda_\delta s_i(x - \delta)\right),
\end{aligned} \tag{3.A.9}$$

where

$$s_i(v) \coloneqq \phi_i(v) + \phi_i(-v) = \tag{3.A.10}$$
$$\begin{cases}
2, & \text{if } v = \mathbf{0} \\
1, & \text{if } v \neq \mathbf{0} \text{ and } v_i \text{ is the non-unique maximizer or minimizer of } v \\
0, & \text{otherwise}
\end{cases}$$

By summing equation (3.A.9) over $i \in [e]$ and defining $s(v) = \sum_{i=1}^{e} s_i(v)$,

$$\lambda_\alpha s(x - \alpha) - \lambda_\beta s(x - \beta) - \lambda_\gamma s(x - \gamma) + \lambda_\delta s(x - \delta) = 0, \tag{3.A.11}$$

$\forall x \in \mathcal{X}$.

Here we try to prove by contradiction that $\mathcal{S} := \{\alpha, \delta\} \cap \{\gamma, \beta\}$ is not empty. Suppose that $\mathcal{S} := \{\alpha, \delta\} \cap \{\gamma, \beta\} = \emptyset$. Then, setting $x = \alpha$ in equation (3.A.11) and noting that $s(0) = 2e$ and $0 \leq s(v) \leq e$ for $v \neq 0$, we get $2e\lambda_\alpha \leq e\lambda_\beta + e\lambda_\gamma$, since $\beta, \gamma \neq \alpha$. Applying the same argument to $x = \beta, \gamma, \delta$, the following system of inequalities holds

$$
\begin{aligned}
2\lambda_\alpha &\leq \lambda_\beta + \lambda_\gamma \\
2\lambda_\beta &\leq \lambda_\alpha + \lambda_\delta \\
2\lambda_\gamma &\leq \lambda_\alpha + \lambda_\delta \\
2\lambda_\delta &\leq \lambda_\beta + \lambda_\gamma.
\end{aligned}
$$

It follows that $\lambda_\alpha = \lambda_\beta = \lambda_\gamma = \lambda_\delta$. Then, rewrite equation (3.A.11) as

$$
s(x - \alpha) - s(x - \beta) - s(x - \gamma) + s(x - \delta) = 0, \tag{3.A.12}
$$

Note now equation (3.A.12) can only hold at $x = \alpha$ iff $s(\alpha - \gamma) = s(\alpha - \beta) = e$ and $s(\alpha - \delta) = 0$. But $s(v) = e$ if and only if all the components of v are non-unique minimizers and maximizers or $\{v_i : i \in [e]\} = \{\zeta, \kappa\}$, where $\zeta < \kappa$ and $|\{i : v_i = \zeta\}| = n_\zeta, |\{i : v_i = \kappa\}| = n_\kappa$, such that $n_\zeta + n_\kappa = e$ and $n_\zeta, n_\kappa \geq 2$.

Consider $z = v + \epsilon E_i$, where $v_i = \zeta$ and $0 < \epsilon < \kappa - \zeta$. The minimum and maximum components of z are ζ and κ, and $\{z_i : i \in [e]\} = \{\zeta, \zeta + \epsilon, \kappa\}$ with $|\{i : z_i = \zeta\}| = n_\zeta - 1, |\{i : z_i = \kappa\}| = n_\kappa$. It follows that,

$$
s(z) = |\{i : z_i = \zeta\}| + |\{i : z_i = \kappa\}| = e - 1.
$$

Now consider $z = v + \epsilon E_i$ where $v_i = \kappa$. The maximum is no longer unique, but the n_ζ minima are still unique. Therefore, $s(z) = n_\zeta \geq 2$. Combining the two cases, it is concluded that $s(v + \epsilon E_i) \geq 2$ for all $i \in [e]$.

Set $x = \alpha + \epsilon E_i$, where $\alpha_i - \beta_i = \min_k\{\alpha_k - \beta_k\}$. Then,

$$s(x - \alpha) = s(\epsilon E_i) = e - 1, \tag{3.A.13}$$

since there is a unique maximizer, but all the other $e - 1$ components are 0, which is the minimum. Furthermore,

$$s(x - \beta) = s(\alpha - \beta + \epsilon E_i) = e - 1,$$

since for $v = \alpha - \beta$ with $s(v) = e$, it corresponds to the first case examined. It is assumed that $\epsilon < \kappa - \zeta = d_{\text{tr}}(\alpha - \beta)$. Moreover,

$$s(x - \gamma) = s(\alpha - \gamma + \epsilon E_i) \geq 2,$$

for $v = \alpha - \gamma$ with $s(v) = e$. Finally, since $s(\alpha - \delta) = 0$ and so the components of $\alpha - \delta$ have a unique minimum and a unique maximum, there exists a neighborhood around $x = \alpha$ such that $x - \alpha$ still has that property, i.e.

$$s(x - \delta) = s(\alpha - \delta + \epsilon E_i) = 0 \tag{3.A.14}$$

for all $\epsilon < \eta$ for some $\eta > 0$.

From equations (3.A.13) – (3.A.14), it is concluded that

$$s(x - \alpha) - s(x - \beta) - s(x - \gamma) + s(x - \delta) \leq -2,$$

which contradicts equation (3.A.12). Therefore $\mathcal{S} = \{\alpha, \delta\} \cap \{\gamma, \beta\} \neq \emptyset$.

Define another set $\mathcal{T} = \{\alpha, \beta, \gamma, \delta\}$. Since $\mathcal{S} \neq \emptyset$, $|\mathcal{T}| \leq 3$. Suppose that $|\mathcal{T}| = 3$

with $\mathcal{T} = \{\tau, \upsilon, \phi\}$. Then, without loss of generality equation (3.A.11) becomes

$$\lambda_\tau s(x-\tau) + \lambda_\upsilon s(x-\upsilon) - \lambda_\phi s(x-\phi) = 0$$

Similarly to before, setting $x = \tau, \upsilon, \phi$ yields,

$$\begin{aligned} 2\lambda_\tau &\leq \lambda_\phi \\ 2\lambda_\upsilon &\leq \lambda_\phi \\ 2\lambda_\phi &\leq \lambda_\tau + \lambda_\upsilon, \end{aligned}$$

which is contradictory since $\lambda_\tau + \lambda_\upsilon > 0$. Therefore, $|\mathcal{T}| \leq 2$. There are 4 cases to consider

i. $\alpha = \delta \neq \beta = \gamma$, but then $\mathcal{S} = \emptyset$,

ii. $\alpha = \beta \neq \gamma = \delta$, but then equation (3.A.11) can only be satisfied $x = \alpha, \gamma$ if $\lambda_\alpha = \lambda_\beta$ and $\lambda_\gamma = \lambda_\delta$ which violates the statement that $(\alpha, \lambda_\alpha) \neq (\beta, \lambda_\beta)$,

iii. $\alpha = \gamma \neq \beta = \delta$ and from equation (3.A.11) at $x = \alpha, \gamma$ it follows that $\lambda_\alpha = \lambda_\gamma, \lambda_\beta = \lambda_\delta$ and hence the desired result,

iv. $\alpha = \beta = \gamma = \delta$, in which case

$$q(x) = (\lambda_\alpha - \lambda_\beta - \lambda_\gamma + \lambda_\delta) d_{\text{tr}}(x, \alpha) + \log\left(\frac{\lambda_\beta}{\lambda_\alpha}\right) - \log\left(\frac{\lambda_\delta}{\lambda_\gamma}\right),$$

which can only be uniformly 0 at $\mathcal{X}$ if and only if $\lambda_\alpha + \lambda_\delta = \lambda_\beta + \lambda_\gamma$. Observe that $(\lambda_\alpha, \lambda_\delta)$ and $(\lambda_\beta, \lambda_\gamma)$ are the two roots of the same quadratic $z^2 - (\lambda_\alpha + \lambda_\delta)z + \lambda_\alpha\lambda_\delta$ and noting that in this case $\lambda_\alpha \neq \lambda_\beta$, it follows that $\lambda_\alpha = \lambda_\gamma$ and $\lambda_\beta = \lambda_\delta$.

□

Lemma 3.A.5. *Consider a compact set* $\Sigma \subseteq \mathbb{R}_+ = (0, \infty)$. *Then the set* $\Lambda = \{\sigma^{-1} : \sigma \in \Sigma\} \in \mathbb{R}_+$ *is also compact.*

Proof. In metric spaces, a set is compact iff it is sequentially compact. Therefore, for every sequence $\sigma_n \in \Sigma$, $\sigma_n \to \sigma \in \Sigma$. Every sequence in Λ can be expressed as $1/\sigma_n$, which tends to $1/\sigma \in \Lambda$. Therefore, Λ is sequentially compact and hence compact. □

Proof of Theorem 3.3.5. This proof has been written for precision estimators $\lambda = 1/\sigma$ instead of deviation estimators. For the rest of the proof consider $\lambda_y = \sigma_y^{-1}$ for $y = 0, 1$ and define the set

$$\Lambda = \{\sigma^{-1} : \sigma \in \Sigma\} \in \mathbb{R}_+.$$

According to Lemma 3.A.5, Λ is also compact.

Define the functions f and h as

$$\begin{aligned} f &: \mathbb{R}^e/\mathbb{R}\mathbf{1} \times \{0,1\} \times \Omega^2 \times \Lambda^2 \to \mathbb{R}, \\ f((x,y),(\omega,\lambda)) &= y \log S(h(x,(\omega,\lambda))) + (1-y) \log S(-h(x,(\omega,\lambda))), \\ h &: \mathbb{R}^e/\mathbb{R}\mathbf{1} \times \Omega^2 \times \Lambda^2 \to \mathbb{R}, \\ h(x,(\omega,\lambda)) &= \lambda_0 d_{\mathrm{tr}}(x,\omega_0) - \lambda_1 d_{\mathrm{tr}}(x,\omega_1) + (e-1) \log \frac{\lambda_1}{\lambda_0}, \end{aligned}$$

where S is the logistic function. Also denote the empirical (Q_n) and expected (Q)

log-likelihood functions as

$$\begin{aligned}
Q_n(\omega, \lambda) &= \frac{1}{n} \sum_{i=1}^{n} f((X_i, Y_i), (\omega, \lambda)) \quad \text{with} \\
Q_n(\hat{\omega}_n, \hat{\lambda}_n) &= \sup_{\omega \in \Omega^2, \lambda \in \Lambda^2} Q_n(\omega), \quad \text{and} \\
Q(\omega, \lambda) &= \mathbb{E}_{(X,Y)} \left(f((X, Y), (\omega, \lambda)) \right) \\
&= \mathbb{E}_X \Big(S(h(X, (\omega^*, \lambda^*))) \log(S(h(X, (\omega, \lambda)))) \\
&\qquad + S(-h(X, (\omega^*, \lambda^*))) \log(S(-h(X, (\omega, \lambda))))\Big).
\end{aligned}$$

The last equation follows from conditioning on

$$Y \sim \text{Bernoulli}(S(h(X, (\omega^*, \lambda^*)))).$$

Before we move on, we need to prove that $f((X, Y), (\omega, \lambda))$ is integrable so that Q is well-defined. Without loss of generality assume that $\lambda_1 \geq \lambda_0$. It suffices to prove that $\mathbb{E}(f((X, Y), (\omega, \lambda)), Y = y)$ is integrable for both $y = 0, 1$. Observe that

$$\begin{aligned}
h(X, (\omega, \lambda)) &\leq (\lambda_0 - \lambda_1) d_{\text{tr}}(X, \omega_0) + \lambda_1 d_{\text{tr}}(\omega_0, \omega_1) + \text{const} \\
&\leq \lambda_1 d_{\text{tr}}(\omega_0, \omega_1) + \text{const}.
\end{aligned}$$

Since $h(X, (\omega, \lambda))$ is bounded above, $f((X, Y), (\omega, \lambda))$ is also bounded below on $Y = 0$ and is hence integral on $Y = 0$. Also, observe that

$$h(X, (\omega, \lambda)) \geq (\lambda_0 - \lambda_1) d_{\text{tr}}(X, \omega_1) - \lambda_0 d_{\text{tr}}(\omega_0, \omega_1) + \text{const}$$

and noting that $\log(S(x)) > x - 1$ for all $x < 0$

$$\log(S(h(X, (\omega, \lambda)))) \geq h(X, (\omega, \lambda)) - 1 \geq (\lambda_0 - \lambda_1) d_{\text{tr}}(X, \omega_1) + \text{const}.$$

Since $d_{\text{tr}}(X, \omega_1)$ is integrable on $Y = 1$, the LHS is integrable on $Y = 1$ too. It follows that $f(X, (\omega, \lambda))$ is integrable and hence Q is well-defined.

First, we prove that Q is maximised at $(\omega, \lambda) = (\omega^*, \lambda^*)$ and that this maximizer is unique. Consider the function

$$g : \mathbb{R} \to \mathbb{R},\ g(t) = S(\alpha) \log S(t) + S(-\alpha) \log S(-t),$$

where $\alpha \in \mathbb{R}$ is some constant. The function g is maximised at $t = \alpha$ and applying Taylor's theorem yields

$$g(x) = g(\alpha) - \frac{1}{2} S(\xi) S(-\xi)(x - \alpha)^2, \text{ for some } \xi \in (\alpha, x).$$

Setting $\alpha = h(X, (\omega^*, \lambda^*))$ and denoting ξ as a random variable

$$\xi(X) \in (h(X, (\omega^*, \lambda^*)), h(X, (\omega, \lambda)))$$

observe that

$$\begin{aligned} Q(\omega, \lambda) &= \mathbb{E}_X(g(h(X, (\omega, \lambda)))) \\ &= \mathbb{E}_X(g(h(X, (\omega^*, \lambda^*))) - \frac{1}{2} \mathbb{E}_X \left(S(\xi(X)) S(-\xi(X)) [h(X, (\omega, \lambda)) - h(X, (\omega^*, \lambda^*))]^2 \right) \\ &\leq Q(\omega^*, \lambda^*), \end{aligned} \tag{3.A.15}$$

Hence, from the expression above it is deduced that (ω^*, λ^*) is a maximizer. Now, consider the function $q : \mathcal{X} \to \mathbb{R}$

$$q(x) = h(x, (\omega^*, \lambda^*)) - h(x, (\omega, \lambda)),$$

where $\Omega \subset \mathcal{X} \subset \mathbb{R}^e/\mathbb{R}\mathbf{1}$ such that for some $\zeta > 0$

$$\mathcal{X} = \{x \in \mathbb{R}^e/\mathbb{R}\mathbf{1} : \inf_{\omega \in \Omega} d_{\text{tr}}(x, \omega) < \zeta\},$$

so that for any $\omega \in \Omega$ there is a neighborhood of ω within $\mathcal{X}$. Note that $\mathcal{X}$ is a bounded set since Ω is bounded too.

We will prove by contradiction that $q(x) = 0, \forall x \in \mathcal{X}$. Suppose there exists $x_0 \in \mathcal{X}$ such that $q(x_0) > 0$, then since q is continuous there exists a neighborhood U with $x_0 \in U$ such that $q(x) > 0$ for all $x \in U$ and so

$$\mathbb{E}(q^2(X)\mathbb{I}(X \in U)) > 0,$$

where $\mathbb{I}$ is the indicator function. Since $h(x, (\omega, \lambda))$ is continuous with respect to x and $\mathcal{X}$ is bounded, the function takes values on a bounded interval and hence $\xi(x)$ is bounded in $\mathcal{X}$ i.e. there exists $\epsilon > 0$ such that $\mathbb{P}(S(\xi(X)S(-\xi(X)) > \epsilon | X \in U) = 1$ and so equation (3.A.15) becomes

$$Q(\omega, \lambda) \leq Q(\omega^*, \lambda^*) - \frac{\epsilon}{2}\mathbb{E}(q^2(X)\mathbb{I}(X \in U)) < Q(\omega^*, \lambda^*),$$

since $\mathbb{P}(X \in U) > 0$ (X has positive density everywhere). Therefore, for (ω, λ) to be a maximizer, $q(x) = 0$ for all $x \in \mathcal{X}$. Apply Lemma 3.A.4 with $\omega^* = (\alpha, \beta)$, $\omega = (\gamma, \delta)$, $\lambda^* = (\lambda_\alpha, \lambda_\beta)$ and $\lambda = (\lambda_\gamma, \lambda_\delta)$ with the set $\mathcal{X}$ containing neighbourhoods of $\alpha, \beta, \gamma, \delta$ and $q(x) = 0$ for all x in those neighbourhoods. It is concluded that $\omega = \omega^*$ and $\lambda = \lambda^*$, thus proving the uniqueness of the maximizer.

Theorem 3.A.2 provides the uniform law of large numbers. The parameter space $\Omega^2 \times \Lambda^2$ is compact since Ω and Λ are compact. Moreover, $f((x, y), (\omega, \lambda))$ is clearly

continuous at each $(\omega, \lambda) \in \Omega^2 \in \Lambda^2$. Finally, consider the function

$$r(z) = \sup_{\omega \in \Omega^2, \lambda \in \Lambda^2} \{|f(z, (\omega, \lambda))|\} = -f(z, \omega(z), \lambda(z)),$$

since f is non-positive. The functions $\omega(z), \lambda(z)$ are chosen to minimize f. Using equation (3.A.15),

$$\mathbb{E}(r(X)) \leq -Q(\omega^*, \lambda^*) + \frac{1}{2}\mathbb{E}([h(X, (\omega(X), \lambda(X))) - h(X, (\omega^*, \lambda^*))]^2),$$

since the sigmoid function is bounded by 1. Note that

$$\mathbb{E}((Z + W)^2) \leq 2(\mathbb{E}(Z^2) + \mathbb{E}(W^2)),$$

and set $W = \log(\lambda_1(X)/\lambda_0(X)) - \log(\lambda_1^*/\lambda_0^*)$. Since $\lambda_y(X) \in \Lambda \subseteq [a, b]$ for some $b \geq a > 0$, it follows that W^2 is integrable and so now we just have to prove that Z is integrable, where $Z = Z_1 + Z_2 + Z_3 + Z_4$ with the four terms corresponding to tropical distance function $\lambda d_{\text{tr}}(X, \omega)$. It also holds

$$\mathbb{E}((Z_1 + Z_2 + Z_3 + Z_4)^2) \leq 2(\mathbb{E}(Z_1^2) + \mathbb{E}(Z_2^2) + \mathbb{E}(Z_3^2) + \mathbb{E}(Z_4^2))$$

and so $\mathbb{E}(Z^2)$ is bounded above by

$$\begin{aligned}
&\mathbb{E}\left(\sum_{i=0}^{1} \lambda_i^2 d_{\text{tr}}^2(X, \omega_i(X))) + (\lambda_i^*)^2 d_{\text{tr}}^2(X, \omega_i^*(X))\right) \\
&\leq \mathbb{E}_Y\left[2\left(\sum_{i=0}^{1} \lambda_i^2 + (\lambda_i^*)^2\right)\mathbb{E}\left(d_{\text{tr}}^2(X, \omega_Y^*)|Y\right) + 2\left(\sum_{i=0}^{1} \lambda_i^2 d_{\text{tr}}^2(\omega_i(X), \omega_Y^*) + (\lambda_i^*)^2 d_{\text{tr}}^2(\omega_i^*, \omega_Y^*)\right)\right],
\end{aligned}$$

where the second inequality came from applying the triangular inequality four times in the form $d_{\text{tr}}(X, \tau) \leq d_{\text{tr}}(X, \omega_Y^*) + d_{\text{tr}}(\omega_Y^*, \tau)$. The final expression is finite because Ω is compact and hence $d_{\text{tr}}(\omega_i(X), \omega_Y^*)$ is finite, $d_{\text{tr}}(X, \omega_Y)|Y$ is square-integrable. Therefore,

$\mathbb{E}(r(X))$ is finite.

All conditions of the theorem are satisfied and so

$$\sup_{\omega\in\Omega^2}\left|\frac{1}{n}\sum_{i=1}^{n} f((X_i,Y_i),\omega)-\mathbb{E}(f((X,Y),\omega))\right| = \sup_{\omega\in\Omega^2}|Q_n(\omega)-Q(\omega)| \xrightarrow{p} 0.$$

Finally, using Theorem 3.A.1 and combining the uniqueness of the maximizer with the uniform bound result, it is concluded that $\hat{\omega} \xrightarrow{p} \omega^*$. □

Proof of Proposition 3.3.6. First, define $\Delta_0 = \{C(X) \neq 1 | Y = 0\}$. By definition of $C(X)$,

$$\begin{aligned}\Delta_0 &= \left\{ (\sigma_0^{-1}-\sigma_1^{-1}) d_{\mathrm{tr}}(X,\hat{\omega}) - (e-1)\log\left(\frac{\sigma_1}{\sigma_0}\right) \geq 0 \,\middle|\, Y=0 \right\} \\ &= \left\{ d_{\mathrm{tr}}(X,\hat{\omega}) \geq \alpha\sigma_0\sigma_1 |\, Y=0 \right\}.\end{aligned}$$

Triangular inequality dictates that

$$d_{\mathrm{tr}}(X,\omega^*) - d_{\mathrm{tr}}(\omega^*,\hat{\omega}) \leq d_{\mathrm{tr}}(X,\hat{\omega}) \leq d_{\mathrm{tr}}(X,\omega^*) + d_{\mathrm{tr}}(\omega^*,\hat{\omega}),$$

and so it follows that

$$\begin{aligned}\Delta_0 &\supseteq \left\{ d_{\mathrm{tr}}(X,\omega^*) \geq \sigma_0\sigma_1 (\alpha+\epsilon) \,|\, Y=0 \right\} \\ \Delta_0 &\subseteq \left\{ d_{\mathrm{tr}}(X,\omega^*) \geq \sigma_0\sigma_1 (\alpha-\epsilon) \,|\, Y=0 \right\},\end{aligned}$$

and since $Z = \sigma_0^{-1} d_{\mathrm{tr}}(X,\omega^*) | Y=0 \sim F$,

$$\mathbb{P}(Z \geq \sigma_1(\alpha+\epsilon)) \leq \mathbb{P}(\Delta_0) \leq \mathbb{P}(Z \geq \sigma_1(\alpha-\epsilon)),$$

which yields the desired result.

Similarly, for $\Delta_1 = \{C(X) \neq 0 | Y = 1\} = \{d_{\text{tr}}(X, \hat{\omega}) \leq \sigma_0 \sigma_1 \alpha\}$,

$$\Delta_1 \supseteq \{d_{\text{tr}}(X, \omega^*) \leq \sigma_0 \sigma_1 (\alpha - \epsilon) | Y = 1\}$$
$$\Delta_1 \subseteq \{d_{\text{tr}}(X, \omega^*) \leq \sigma_0 \sigma_1 (\alpha + \epsilon) | Y = 1\},$$

and since $Z = \sigma_1^{-1} d_{\text{tr}}(X, \omega^*) | Y = 1 \sim F$,

$$\mathbb{P}(Z \leq \sigma_0(\alpha - \epsilon)) \leq \mathbb{P}(\Delta_1) \leq \mathbb{P}(Z \leq \sigma_0(\alpha + \epsilon)),$$

which is the desired interval.

For the second part of the proposition, $\hat{\omega} = \omega^*$ and so $\epsilon = 0$. Hence,

$$\mathbb{P}(\Delta_0) = 1 - F(\sigma_1 \alpha) = 1 - F(xu(x))$$
$$\mathbb{P}(\Delta_1) = F(\sigma_0 \alpha) = F(u(x)), \text{ where}$$
$$x = \frac{\sigma_1}{\sigma_0} \text{ and } u(x) = (e-1)\frac{\log x}{x-1}$$

Consider the function

$$g(x) = 1 - F(xu(x)) - F(u(x))$$

Proving that $g(x) < 0$ for all $x > 1$ is equivalent to proving the desired result that $\mathbb{P}(\Delta_0) < \mathbb{P}(\Delta_1)$ for $\sigma_1 > \sigma_0$. First,

$$\lim_{x \to 1} u(x) = \lim_{x \to 1} xu(x) = e - 1,$$

and so $\lim_{x \to 1} g(x) = 1 - 2F(e-1)$. It is a well-known fact that the median of the Gamma distribution is less than the mean. Hence, for $Z \sim \text{Gamma}(e-1, 1)$ with mean $e - 1$, $F(e-1) > \frac{1}{2}$ and so

$$\lim_{x \to 1} g(x) < 0. \tag{3.A.16}$$

Finally, the derivative of g is

$$g'(x) = -F'(u(x))u'(x) - F'(xu(x))(xu'(x) + u(x))$$

The following two inequalities

$$F'(u(x)) \geq F'(xu(x)), \tag{3.A.17}$$

$$u'(x) + xu'(x) + u(x) \geq 0, \tag{3.A.18}$$

imply that

$$g'(x) \leq -F'(xu(x))(u'(x) + xu'(x) + u(x)) \leq 0. \tag{3.A.19}$$

From (3.A.16) and (3.A.19) it follows that $g(x) < 0$ for all $x > 1$.

For inequality (3.A.17), remember that

$$F'(x) = \frac{x^{e-2}\exp(-x)}{\Gamma(e-1)}$$

and so

$$\begin{aligned} F'(u(x)) - F'(xu(x)) &= F'(u(x))\left(1 - x^{e-2}\exp(-(x-1)u(x))\right) \\ &= F'(u(x))\left(1 - x^{e-2}\exp(-(e-1)\log(x))\right) \\ &= F'(u(x))(1 - x^{-1}) > 0, \end{aligned}$$

for all $x > 1$.

For inequality (3.A.18),

$$u'(x) + xu'(x) + u(x) = \frac{e-1}{(x-1)^2}\left(x - x^{-1} - 2\log x\right),$$

is a non-negative function for $x > 1$ iff v is a non-negative function, where

$$v(x) = x - x^{-1} - 2\log x, \text{ with}$$
$$v'(x) = \frac{(x-1)^2}{x^2} \geq 0 \text{ and } v(1) = 0.$$

Clearly, v is a non-negative function for $x > 1$, so inequality (3.A.18) is satisfied. □

Proof of Proposition 3.3.7. For symbolic convenience, in this proof class 0 is referred to as class -1 and so $Y \in \{-1, 1\}$. Applying the triangular inequality twice,

$$\begin{aligned} D_X &= d_{\text{tr}}(X, \omega_Y^*) - d_{\text{tr}}(X, \omega_{-Y}^*) \\ &\geq (d_{\text{tr}}(X, \hat{\omega}_Y) - d_{\text{tr}}(\omega_Y^*, \hat{\omega}_Y)) \\ &\quad - \left(d_{\text{tr}}(X, \hat{\omega}_{-Y}) + d_{\text{tr}}(\omega_{-Y}^*, \hat{\omega}_{-Y})\right) \\ &= d_{\text{tr}}(X, \hat{\omega}_Y) - d_{\text{tr}}(X, \hat{\omega}_{-Y}) - \epsilon, \end{aligned}$$

it follows that

$$\{C(X) \neq Y\} = \{d_{\text{tr}}(X, \hat{\omega}_Y) - d_{\text{tr}}(X, \hat{\omega}_{-Y}) \geq 0\} \subseteq \{D_X \geq -\epsilon\}$$

and so the generalization error has the following upper bound

$$\mathbb{P}(C(X) \neq Y) \leq \mathbb{P}\left(D_X \geq -\epsilon\right).$$

Note that if $d_{\text{tr}}(X, \omega_Y^*) < \Delta_\epsilon$, then by the use of triangular inequality

$$\begin{aligned} D_X &= d_{\text{tr}}(X, \omega_Y^*) - d_{\text{tr}}(\omega_{-Y}^*, X) \\ &\leq d_{\text{tr}}(X, \omega_Y^*) - \left(d_{\text{tr}}(\omega_{-Y}^*, \omega_Y^*) - d_{\text{tr}}(\omega_Y^*, X)\right) \\ &< 2\Delta_\epsilon - d_{\text{tr}}(\omega_1^*, \omega_{-1}^*) = -\epsilon. \end{aligned}$$

Consequently,

$$\mathbb{P}(C(X) \neq Y) \leq \mathbb{P}\left(D_X \geq -\epsilon, Z_X \geq \Delta_\epsilon\right) \tag{3.A.20}$$

Since the distribution of X is symmetric around ω_Y^*, the random variable $2\omega_Y^* - X$ has the same distribution and so

$$\mathbb{P}\left(D_X \geq -\epsilon, Z_X \geq \Delta_\epsilon\right) = \mathbb{P}\left(D_{2\omega_Y^* - X} \geq -\epsilon, Z_{2\omega_Y^* - X} \geq \Delta_\epsilon\right). \tag{3.A.21}$$

It will be proved that

$$Z_{2\omega_Y^* - X} = Z_X, \tag{3.A.22}$$

$$D_X + D_{2\omega_Y^* - X} \leq 0, \tag{3.A.23}$$

and so $\{D_{2\omega_Y^* - X} \geq -\epsilon, Z_{2\omega_Y^* - X} \geq \Delta_\epsilon\} \subseteq \{D_X \leq \epsilon, Z_X \geq \Delta_\epsilon\}$. Then, using equation (3.A.21),

$$\mathbb{P}\left(D_X \geq -\epsilon, Z_X \geq \Delta_\epsilon\right) \leq \mathbb{P}\left(D_X \leq \epsilon, Z_X \geq \Delta_\epsilon\right),$$

and substituting it to inequality (3.A.20),

$$\begin{aligned}\mathbb{P}(C(X) \neq Y)) =& \frac{1}{2}(\mathbb{P}\left(D_X \geq -\epsilon, Z_X \geq \Delta_\epsilon\right) \\ &+ \ \mathbb{P}\left(D_X \leq \epsilon, Z_X \geq \Delta_\epsilon\right)) \\ =& \mathbb{P}\left(Z_X \geq \Delta_\epsilon\right) + h(\epsilon)\end{aligned}$$

where $h(\epsilon) = \mathbb{P}(Z_X \geq \Delta_\epsilon, |D_X| \leq \epsilon)$ is an increasing function with respect to ϵ, which completes the first part of the proof.

Equation (3.A.22) follows from the observation that

$$d_{\text{tr}}(2\omega_Y^* - x, \omega_Y^*) = d_{\text{tr}}(x, \omega_Y^*).$$

For equation (3.A.23),

$$\begin{aligned} D_{2\omega_Y^* - X} + D_X &= Z_{2\omega_Y^* - X} - d_{\mathrm{tr}}(2\omega_Y^* - X, \omega_{-Y}^*) \\ &\quad + Z_X - d_{\mathrm{tr}}(X, \omega_{-Y}^*) \\ &\overset{(3.A.22)}{=} 2Z_{2\omega_Y^* - X} - d_{\mathrm{tr}}(2\omega_Y^* - X, \omega_{-Y}^*) - d_{\mathrm{tr}}(\omega_{-Y}^*, X) \\ &\leq 2Z_{2\omega_Y^* - X} - d_{\mathrm{tr}}(2\omega_Y^* - X, X) = 0, \end{aligned}$$

where the last inequality comes from the triangular inequality. Finally, the consistency of the learning algorithm is proved. Under the conditions of Theoreom 3.3.5, the maximum likelihood estimator $\hat{\omega} = (\hat{\omega}_0, \hat{\omega}_1) \xrightarrow{\mathrm{p}} (\omega_0^*, \omega_1^*)$ as $n \to \infty$ where $(X_1, Y_1), \ldots, (X_n, Y_n)$ is the sample. For the rest of the proof, the test covariate-response pair (X, Y) is independent from the aforementioned training sample. Define the classifier,

$$C_\omega(x) = \mathrm{sgn}\left(d_{\mathrm{tr}}(x, \omega_0) - d_{\mathrm{tr}}(x, \omega_1)\right)$$

where $\omega = (\omega_0, \omega_1)$. The Bayes predictor is $C_{\omega_0^*, \omega_1^*}$. Noting that $C_{\omega_0^*, \omega_1^*}(X) = \mathrm{sgn}(D_X)Y$, the Bayes (or irreducible) error is

$$\mathrm{BE} = \mathbb{P}(\mathrm{sgn}(D_X)Y \neq Y) = \mathbb{P}(D_X > 0) = \mathbb{P}(D_X \geq 0),$$

since it is assumed that $\mathbb{P}(D_X = 0) = 0$. Using inequality 3.A derived earlier, it follows that the generalization error is bounded by

$$\mathbb{P}(D_X \geq 0) = \mathrm{BE} \leq \mathbb{P}(C_{\hat{\omega}}(X) \neq Y) \leq \mathbb{P}(D_X \geq -\epsilon(\hat{\omega})),$$

where $\epsilon(\hat{\omega}_0, \hat{\omega}_1) = d_{\mathrm{tr}}(\omega_0, \omega_0^*) + d_{\mathrm{tr}}(\omega_1, \omega_1^*) \xrightarrow{\mathrm{p}} 0.$ as the training sample size $n \to \infty$

according to Theorem 3.3.5. The complementary CDF of D_X, defined as

$$F^C_{D_X}(x) = \mathbb{P}(D_X \geq x),$$

is a continuous function and so it follows that $F^C_{D_X}(\epsilon(\hat{\omega})) \xrightarrow{\text{p}} F^C_{D_X}(0) = \text{BE}$ as $n \to \infty$. From the probability squeeze theorem,

$$\mathbb{P}(C_{\hat{\omega}}(X) \neq Y | (X_1, Y_1), \ldots, (X_n, Y_n)) \xrightarrow{\text{p}} \text{BE as } n \to \infty.$$

This concludes the proof of the consistency of the algorithm. □

Proof of Proposition 3.4.1. Consider the random variable $d_{\text{tr}}(X, \alpha)$. From the triangular inequality

$$d_{\text{tr}}(X, \alpha) \leq d_{\text{tr}}(X, \omega^*) + d_{\text{tr}}(\alpha, \omega^*),$$

it is deduced that $d_{\text{tr}}(X, \alpha)$ is integrable, bounded above by an integrable random variable.

Now consider the function $F : \mathbb{R}^e/\mathbb{R}\mathbf{1} \to \mathbb{R}$,

$$F(x) = d_{\text{tr}}(x, \omega) + d_{\text{tr}}(2\omega^* - x, \omega) - 2d_{\text{tr}}(x, \omega^*).$$

Noting that $d_{\text{tr}}(2\omega^* - x, \omega) = d_{\text{tr}}(x, 2\omega^* - \omega)$, it follows that $F(X)$ is integrable as the sum of integrable random variables.

From triangular inequality and the fact that $d_{\text{tr}}(2\omega^* - x, x) = 2d_{\text{tr}}(x, \omega^*)$ it follows that $F(x) \geq 0$ for all $x \in \mathbb{R}^e/\mathbb{R}\mathbf{1}$. Furthermore, $F(\omega^*) > 0$ and since F is continuous, there exists a neighbourhood U that contains ω^* such that $F(x) > 0$ for all $x \in U$. Moreover, the function has positive density in a neighbourhood V that contains the centre ω^*. Therefore, there exists a neighbourhood $W = U \cap V$ such that $F(x) > 0$ for

all $x \in W$ and $\mathbb{P}(X \in W) > 0$. Hence, since $F(X) \geq 0$,

$$\mathbb{E}(F(X)) \geq \mathbb{E}(F(X)|X \in W)\mathbb{P}(X \in W) > 0.$$

In other words,

$$\mathbb{E}\left(d_{\mathrm{tr}}(X, \omega)\right) + \mathbb{E}\left(d_{\mathrm{tr}}(2\omega^* - X, \omega)\right) > 2\mathbb{E}(d_{\mathrm{tr}}(x, \omega^*)) \tag{3.A.24}$$

Moreover, consider the isometry $y = 2\omega^* - x$ and note that for symmetric probability density functions around ω^*, $f(\omega^* - \delta) = f(\omega^* + \delta)$ and so for $\delta = \omega^* - x$, we have $f(y) = f(x)$. Applying this transformation to the following integral yields

$$\begin{aligned} \mathbb{E}(d_{\mathrm{tr}}(2\omega^* - X, \omega)) &= \int_{\mathbb{R}^e/\mathbb{R}\mathbf{1}} d_{\mathrm{tr}}(2\omega^* - x, \omega) f(x)\, dx \\ &= \int_{\mathbb{R}^e/\mathbb{R}\mathbf{1}} d_{\mathrm{tr}}(y, \omega) f(y)\, dy = \mathbb{E}(d_{\mathrm{tr}}(X, \omega)). \end{aligned} \tag{3.A.25}$$

Combining equation (3.A.25) with inequality (3.A.24) shows that the function $Q(\omega) = \mathbb{E}(d_{\mathrm{tr}}(X, \omega))$ has a global minimum at ω^*.

From Theorem 3.A.2 (uniform law of large numbers), set $f(x, \omega) = d_{\mathrm{tr}}(x, \omega)$ and observe that $f(x, \omega)$ is always continuous w.r.t. ω. Setting $r(x) = \sup_{\omega \in \Omega} d_{\mathrm{tr}}(x, \omega)$, which is finite since Ω is compact, observe that

$$r(x) := \sup_{\omega \in \Omega} d_{\mathrm{tr}}(x, \omega) \leq d_{\mathrm{tr}}(x, \omega^*) + \sup_{\omega \in \Omega} d_{\mathrm{tr}}(\omega, \omega^*).$$

Since Ω is compact, the second term is finite and hence $r(X)$ is integrable, since $d_{\mathrm{tr}}(X, \omega^*)$ is integrable. All conditions of the theorem are satisfied so $Q(\omega) = \mathbb{E}(d_{\mathrm{tr}}(x, \omega))$ is continuous with respect to ω and

$$\sup_{\omega \in \Omega} |Q_n(\omega) - Q(\omega)| \xrightarrow{p} 0 \text{ as } n \to \infty,$$

where $Q_n(\omega) = n^{-1}\sum_{i=1}^n d_{\mathrm{tr}}(X_i, \omega)$. Since $Q(\omega)$ has a unique minimum at ω^*, all conditions of Theorem 3.A.1 are satisfied and so $\tilde{\omega}_n \to \omega^*$ as $n \to \infty$. □

Proof of Proposition 3.4.2. i. If $\omega - X_i$ has a unique maximum $M_i = \arg\max_j\{\omega_j - (X_i)_j\}$ and unique minimum $m_i = \arg\min_j\{\omega_j - (X_i)_j\}$, then the gradient is

$$(\nabla f(x))_j = |\{i : M_i = j\}| - |\{i : m_i = j\}|. \tag{3.A.26}$$

For the converse, assume that the gradient is well-defined. From equations (3.A.7)–(3.A.8) and following the first few sentences of Lemma 3.A.4

$$d_{\mathrm{tr}}(x + \epsilon E_j, y) + d_{\mathrm{tr}}(x - \epsilon E_j, y) - 2d_{\mathrm{tr}}(x, y) = \epsilon s_j(x - y),$$

where s_j is defined in equation (3.A.10) of Lemma 3.A.4. Consequently,

$$f(x + \epsilon E_j) + f(x - \epsilon E_j) - 2f(x) = \epsilon \sum_{i=1}^n s_j(X_i - \omega_i)$$

Since f has a well-defined gradient, $\sum_{i=1}^n s_j(X_i - \omega) = 0$ i.e. $s_j(X_i - \omega) = 0$ for all $(i, j) \in [n] \times [e]$. This can only happen iff $X_i - \omega$ has unique maximum and minimum component for all $i \in [n]$.

ii. Using equation (3.A.26), the gradient of f vanishes at $x = \omega$ if and only if

$$|\{i : M_i = j\}| = |\{i : m_i = j\}|. \tag{3.A.27}$$

Moreover,

$$\begin{aligned} f(\omega + v) &= \sum_{i=1}^{n} \max_k \{\omega_k - (X_i)_k + v_k\} - \min_k \{\omega_k - (X_i)_k + v_k\} \\ &\geq \sum_{i=1}^{n} \omega_{M_i} - (X_i)_{M_i} + v_{M_i} - \omega_{m_i} + (X_i)_{m_i} - v_{m_i} \\ &= f(\omega) + \sum_{i=1}^{n} v_{M_i} - v_{m_i} \end{aligned}$$

Finally, note that because of equation (3.A.27),

$$\begin{aligned} \sum_{i=1}^{n} v_{M_i} &= \sum_{j=1}^{e} v_j |\{i \in [n] : M_i = j\}| \\ &\overset{(3.A.27)}{=} \sum_{j=1}^{e} v_j |\{i \in [n] : m_i = j\}| = \sum_{i=1}^{n} v_{m_i}, \end{aligned}$$

and so $f(\omega + v) \geq f(\omega)$ for all $v \in \mathbb{R}^e/\mathbb{R}\mathbf{1}$.

□

3.B Space of ultrametrics

Theorem 3.B.1 (explained in Ardila and Klivans (2006); Page et al. (2020))**.** *Suppose we have a* classical *linear subspace $L_m \subset \mathbb{R}^e$ defined by the linear equations $x_{ij} - x_{ik} + x_{jk} = 0$ for $1 \leq i < j < k \leq m$. Let Trop$(L_m) \subseteq \mathbb{R}^e/\mathbb{R}\mathbf{1}$ be the* tropicalization *of the linear space $L_m \subset \mathbb{R}^e$, that is, classical operators are replaced by tropical ones (defined in Section 3.C in the supplement) in the equations defining the linear subspace L_m, so that all points $(v_{12}, v_{13}, \ldots, v_{m-1,m})$ in Trop(L_m) satisfy the condition that*

$$\max_{i,j,k \in [m]} \{v_{ij}, v_{ik}, v_{jk}\}.$$

is attained at least twice. Then, the image of $\mathcal{U}_m$ *inside of the tropical projective torus* $\mathbb{R}^e/\mathbb{R}\mathbf{1}$ *is equal to Trop*(L_m).

3.C Tropical Arithmetics and Tropical Inner Product

In tropical geometry, addition and multiplication are different than regular arithmetic. The arithmetic operations are performed in the max-plus tropical semiring $(\mathbb{R}\cup\{-\infty\}, \oplus, \odot)$ as defined in Pin (1998).

Definition 3.C.1 (Tropical Arithmetic Operations)**.** *In the tropical semiring, the basic tropical arithmetic operations of addition and multiplication are defined as:*

$$a \oplus b := \max\{a, b\}, \qquad a \odot b := a + b, \qquad \text{where } a, b \in \mathbb{R} \cup \{-\infty\}.$$

The element $-\infty$ *ought to be included as it is the identity element of tropical addition. Tropical subtraction is not well-defined and tropical division is classical subtraction.*

The following definitions are necessary for the definition of the tropical inner product

Definition 3.C.2 (Tropical Scalar Multiplication and Vector Addition)**.** *For any scalars* $a, b \in \mathbb{R} \cup \{-\infty\}$ *and for any vectors* $v, w \in (\mathbb{R} \cup \{-\infty\})^e$*, where* $e \in \mathbb{N}$*,*

$$a \odot v := (a + v_1, \ldots, a + v_e),$$
$$a \odot v \oplus b \odot w := (\max\{a + v_1, b + w_1\}, \ldots, \max\{a + v_e, b + w_e\}).$$

From the definitions above, it follows that the tropical inner product is $\omega^T \odot x = \max\{\omega + x\}$ for all vectors $\omega, x \in \mathbb{R}^e/\mathbb{R}\mathbf{1}$. In classical logistic regression a linear function in the form of a classical inner product $h_\omega(x) = \omega^T x$, $\omega \in \mathbb{R}^n$ is used. The tropical

symbolic equivalent is

$$h_\omega(x) = \omega^T \odot x = \max_{l\in[e]}\{\omega_l + x_l\}. \tag{3.C.1}$$

This expression is not well-defined, since the statistical parameter and covariate vectors $\omega, u \in \mathbb{R}^e/\mathbb{R}1$ are only defined up to addition of a scalar multiple of the vector $(1,\dots,1)$. To resolve this issue, we fix

$$-\min_{l\in[e]}\{\omega_l + x_l\} = c, \tag{3.C.2}$$

where $c \in \mathbb{R}$ is a constant for all observations. Combining equations (3.C.2), (3.C.1), and the definition of tropical distance (3.2.1),

$$h_\omega(x) = d_{\text{tr}}(x, -\omega) - c.$$

For simplicity, under the transformation $-\omega \to \omega$ the expression becomes

$$h_\omega(x) = d_{\text{tr}}(x, \omega) - c.$$

3.D Tropical Logistic Regression Algorithm

Algorithm 1 One-species tropical logistic regression

Input: distance matrix $D \in \mathbb{R}_+^{N\times e}$, labels $Y \in \{0,1\}^N$
$\tilde{\omega} = \text{FW\_point}(D)$
$\hat{\sigma}_0, \hat{\sigma}_1 = \arg\max_{\sigma_0,\sigma_1>0} l(\tilde{\omega}, \sigma_0, \sigma_1)|D, Y)$ with root solving.
Output: $(\tilde{\omega}, \hat{\sigma}_0, \hat{\sigma}_0)$

Algorithm 2 Two-species tropical logistic regression

Input: distance matrix $D \in \mathbb{R}_+^{N\times e}$, labels $Y \in \{0,1\}^N$
$\tilde{\omega}_0 = \text{FW\_point}(D[Y == 0])$
$\tilde{\omega}_1 = \text{FW\_point}(D[Y == 1])$
$\hat{\sigma} = \arg\max_{\sigma>0} l(\tilde{\omega}_0, \tilde{\omega}_1, \sigma|D, Y)$ with root solving.
Output: $(\tilde{\omega}_0, \tilde{\omega}_1, \hat{\sigma})$

3.E Fermat-Weber Point Visualization

As noted in Section 3.4, the gradient method is much faster than linear programming. Unfortunately, there is no guarantee that it will guide us to a Fermat-Weber point. However, in practice, the gradient method tends to work well. Figure 3.E.1 illustrates just that. Given, ten datapoint $X_1, \ldots, X_{10} \in \mathbb{R}^3/\mathbb{R}\mathbf{1} \cong \mathbb{R}^2$, the Fermat-Weber set is found to be a trapezoid. This is in agreement with (Lin and Yoshida, 2018a), which states that all Fermat-Weber sets are classical polytopes. The two-dimensional gradient vector, plotted as a vector field in Figure 3.E.1, always points towards the Fermat-Weber set. Therefore, the gradient algorithm should always guide us to a Fermat-Weber point.

3.F MLE Estimator for σ

If $Z_i \overset{\text{iid}}{\sim} \text{Gamma}(n, k)$, where n is constant and k is a statistical parameter, then it is well-known that the maximum likelihood estimator is

$$\hat{k} = \bar{Z}/n,$$

where $\bar{Z}$ is the sample average. In our case $Z_i = d(X_i, \omega^*)$ and $k = i\sigma^i$. From Proposition 3.3.3, $Z_i \sim \text{Gamma}(n/i, i\sigma^i)$ and by substituting these parameters in equation 3.F, it follows that the MLE for σ is

$$\hat{\sigma}^i = \bar{Z}/n,$$

where $\bar{Z}$ is the average distance of the covariates (gene trees) from their mean (species tree). This results holds for all $i \in \mathbb{N}$ and both Euclidean and tropical metrics. The only difference is that for Euclidean spaces $X \in \mathbb{R}^e$ and so $n = e$, while for the tropical projective torus $\mathbb{R}^e/\mathbb{R}\mathbf{1}$, $n = e - 1$.

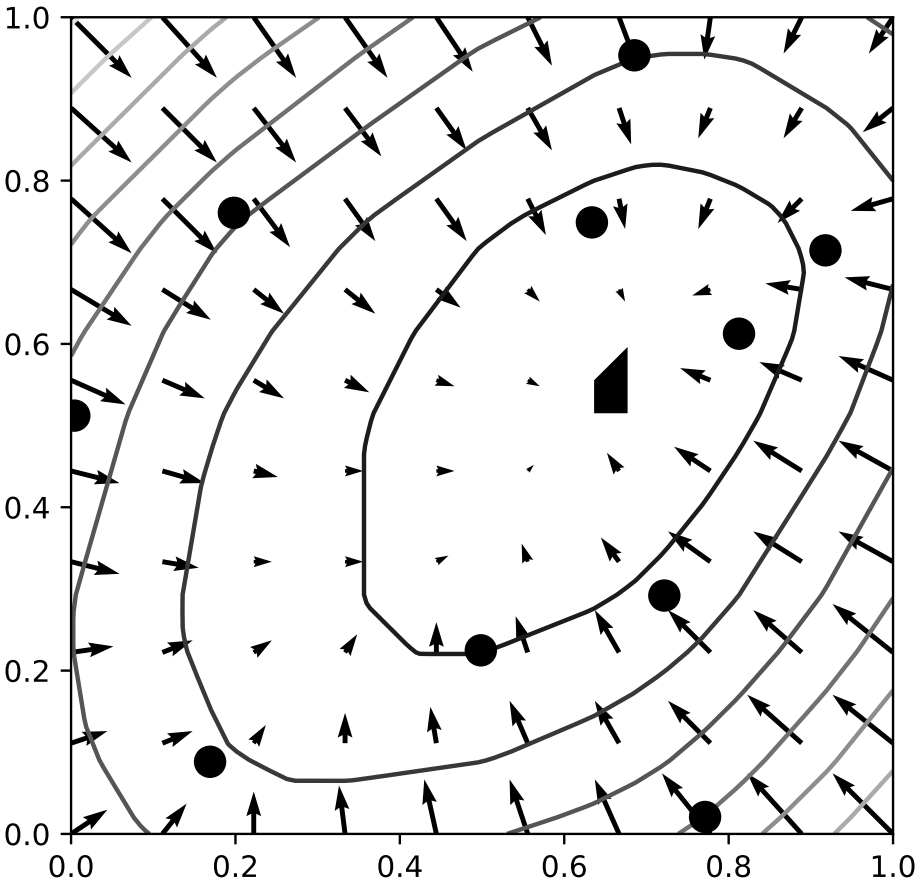


Figure 3.E.1: Visualization of the function $f(\omega) = \sum_{i=1}^{10} d_{\text{tr}}(X_i, \omega)$ for X_i. The black circles are the datapoints $X_1, \ldots, X_{10}$, the solid lines are contours of f, the vector field is the gradient and the small black trapezoid at $(0.65, 0.55)$ is the Fermat-Weber set.

3.G Approximate BHV Logistic Regression

Similar to the tropical Laplace distribution, in (Billera et al., 2001) the following distribution was considered

$$f_{\lambda,\omega}(x) = K_{\lambda,\omega} \exp\left(-\lambda d_{\text{BHV}}(x, \omega)\right),$$

where $\lambda = 1/\sigma$ is a concentration/precision parameter, d_{BHV} is the BHV metric and $K_{\lambda,\omega}$ is the normalization constant that depends on λ and ω. We consider an adaptation of the two-species model for this metric, where the data from the two classes have the same concentration rate but different centre. If $X|Y \sim f_{\lambda,\omega_Y^*}$, then

$$h_{\omega_0,\omega_1}(x) = \lambda \left(d_{\text{BHV}}(x,\omega_0^*) - d_{\text{BHV}}(x,\omega_1^*)\right) + \log \frac{K_{\lambda,\omega_0^*}}{K_{\lambda,\omega_1^*}}. \tag{3.G.1}$$

Unlike in the tropical projective torus or the euclidean space, in the BHV space $K_{\lambda,\omega_0^*} \neq K_{\lambda,\omega_1^*}$, because the space is not translation-invariant. However, if we assume that the two centres are far away from trees with bordering topologies, it may be assumed that the trees are mostly distributed in the Euclidean space and as a result $K_{\lambda,\omega_0^*} \approx K_{\lambda,\omega_1^*}$. Under this assumption, equation (3.G.1) becomes

$$h_{\omega_0,\omega_1}(x) \approx \lambda \left(d_{\text{BHV}}(x,\omega_0^*) - d_{\text{BHV}}(x,\omega_1^*)\right).$$

Therefore, the classification/decision boundary for the BHV is the BHV bisector $d_{\text{BHV}}(x,\omega_0^*) = d_{\text{BHV}}(x,\omega_1^*)$ and the most sensible classifier is

$$C(x) = \mathbb{I}\left(d_{\text{BHV}}(x,\omega_0^*) > d_{\text{BHV}}(x,\omega_1^*)\right),$$

where $\mathbb{I}$ is the indicator function.

3.H Graphs for Simulated Data under the Multi-Species Coalescent Model for different R

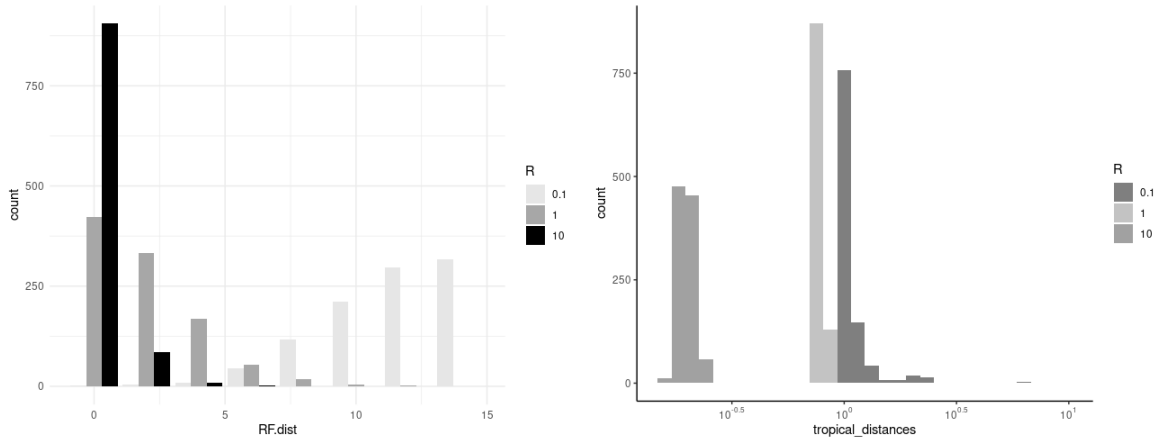


Figure 3.H.1: (left) Robinson-Foulds distances and (right) tropical distances of inferred species trees $\hat{\omega}$ from the actual species trees ω^* for $R = 0.1, 1, 10$.

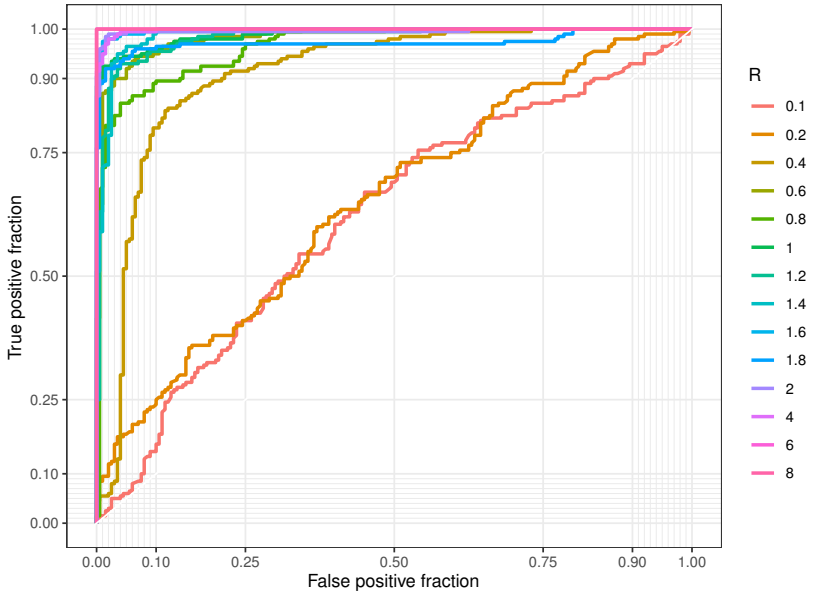


Figure 3.H.2: ROC curves for the tropical logistic regression with different values of R. Higher the value of R is the closer an estimated ROC curve for the tropical logistic regression model gets to the point $(0, 1)$.

Chapter 4

Tropical Neural Networks

Disclaimer: This chapter includes collaborative work with Dr Keiji Miura and Professor Ruriko Yoshida.

4.1 Introduction

A neural network is a learning method, called deep learning, to learn data in a way to mimic a brain system, i.e., which interconnects nodes, called neurons, in a layered structure like a human brain system (Goodfellow et al., 2016; Calin, 2020; Grohs and Kutyniok, 2023). In recent years, deep neural networks have achieved remarkable success in processing data that lie in Euclidean space (Ford, 2018). However, when input data are phylogenetic trees or time series with trends, represented as vectors in the *tropical projective torus* (Nye, 2011; Nye et al., 2017; Monod et al., 2019; Weyenberg et al., 2016; Yoshida et al., 2019, 2022a; Page et al., 2020; Yoshida et al., 2023c; Tang et al., 2020), classical neural networks show a poor performance. Therefore in this paper, we propose neural networks which process an input data as vectors over the tropical projective torus. The tropical projective torus denoted by $\mathbb{R}^d/\mathbb{R}\mathbf{1}$ is the d-dimensional real numbers, $\mathbb{R}^d$, mod by the vector with all ones, i.e., $\mathbf{1} := (1, 1, \ldots, 1) \in \mathbb{R}^d$. Over the tropical projective torus denoted by $\mathbb{R}^d/\mathbb{R}\mathbf{1}$, we define

$x := (x_1, x_2, \ldots, x_d) = (x_1 + c, x_2 + c, \ldots, x_d + c) \in \mathbb{R}^d/\mathbb{R}\mathbf{1}$ where $c \in \mathbb{R}$ (Maclagan and Sturmfels, 2021). Here we consider the *tropical metric*, also known as the *generalized Hilbert projective metric*, over the tropical projective torus as activation functions in a hidden layer of a neural network. It is important to keep the invariance of the input vector under the one-vector which is innate in the tropical projective torus (Maclagan and Sturmfels, 2021; Joswig, 2021; Speyer and Sturmfels, 2009; Lin and Yoshida, 2018b; Yoshida et al., 2022c; Richter-Gebert et al., 2003). Our strategy is to embed an input vector in the tropical projective torus into a vector in the classical Euclidean space in the first layer. This is analogous to the word embedding in the field of natural language processing (Vaswani et al., 2017; Onan, 2021). Then the following layers can be the same as the classical ones.

Although some previous works analyzed ReLU neural networks by using the tropical geometry, the neural networks themselves are defined on a classical Euclidean space (Zhang et al., 2018b; Alfarra et al., 2022; Montúfar et al., 2022). In this paper, on the other hand, we consider a tropical projective torus as an input space and keep the invariance under the one-vector. That is, our work is truly tropical.

In this paper, we first introduce a tropical embedding layer. We use the tropical embedding layer as the first layer of the classical neural networks to keep the invariance in the tropical projective space. To check if this tropical neural network has enough flexibility, we next prove that this tropical neural network is a universal approximator. Then we derive a backpropagation rule for the tropical neural networks. We provide TensorFlow 2 codes for implementing a tropical neural network in the same fashion as the classical one, where the weights initialization problem is considered according to extreme value statistics. We show its applications to phylogenomics, a new field in evolutionary biology which applies tools from phylogenetic trees to genome data. Applications includes simulated data under the multi-species coalescent model which is the most popular model to analyze gene tree analysis on a genome data (Maddison and

Maddison, 2009), and empirical data of influenza virus data set collected from the state of New York obtained from the GI-SAID EpiFlu database (www.gisaid.org). Finally we briefly show that a tropical neural network can be interpreted as a generalization of a tropical logistic regression.

4.2 Tropical Embedding for Tropical Neural Networks

The classical neural networks only accept an input vector in an Euclidean space in its original form. Thus they cannot accept a phylogenetic tree as an input since a space of phylogenetic trees is not Euclidean (Speyer and Sturmfels, 2009; Ardila and Klivans, 2006; Billera et al., 2001), for example. Therefore, we first consider a tropical embedding layer, which is analogous to the word embedding in natural language processing (Vaswani et al., 2017; Onan, 2021). Once a phylogenetic tree is embedded in an Euclidean space, a classical neural network is applied to analyzing it.

Definition 4.2.1 (tropical neural networks)**.** *A tropical neural network is a network where a tropical embedding layer as the first hidden layer is followed by a classical neural network (classical layers).*

Definition 4.2.2 (tropical embedding layer)**.** *Let x in $\mathbb{R}^d/\mathbb{R}\mathbf{1}$ be an input vector to the tropical embedding layer. Then, the activity of j-th neuron as an output of the tropical embedding layer is given by*

$$z_j = \max_i(x_i + w_{ji}^{(1)}) - \min_i(x_i + w_{ji}^{(1)}), \tag{4.2.1}$$

where $w_{ji}^{(1)}$ are the weights of the tropical neural network (see Figure 4.4.1).

Remark 10: Note that no activation function is executed for z as the "max - min"

operation is somehow regarded as the activation function of the neurons in the first hidden layer.

Remark 11: There is a geometric interpretation: "max - min" operation measures the distance between the points x and $w_j^{(1)}$. Therefore $z(x)$ is invariant along one vectors $\mathbf{1}$.

Remark 12: There are alternative ways to attain the invariance such as

$$z_j = \max_i(x_i + w_{ji}^{(1)}) - \text{2nd}\max_i(x_i + w_{ji}^{(1)}).$$

There is a geometric interpretation: "max - 2nd max" operation measures the distance between a point x and the tropical hyperplane whose normal vector is $w^{(1)}$ Joswig (2021). Therefore $z(x)$ is invariant along one vectors $\mathbf{1}$. You could even use j-th max in general. However, the repertoire of functions never increase by using alternative ones. That is, from the view point of universal approximation theorem, using Eq. (4.2.1) suffices. In addition, Eq. (4.2.1) seems to perform better than the alternative ones according to our numerical experiments (not shown). Therefore we solely use Eq. (4.2.1) as a tropical embedding layer in what follows.

Remark 13: Suppose $A \in \mathbb{Z}_+^{N\times d}$. With a ReLU activation function, the input vector x is transformed to

$$\max\{Ax + b, 0\},$$

in the first hidden layer. Assume that $A\mathbf{1} \neq 0$. Suppose $x \in \mathbb{R}^d/\mathbb{R}\mathbf{1}$. Then we have $x := x + c \cdot (1, \ldots, 1) = x + c \cdot \mathbf{1} \in \mathbb{R}^d/\mathbb{R}\mathbf{1}$. Then for $c \ll 0$ and fixed x, we have:

$$\max\{Ax + cA\mathbf{1} + b, 0\} \quad = \quad 0.$$

As $c \to -\infty$, we have

$$\frac{1}{1+\exp(-\max\{Ax+cA\mathbf{1}+b,0\})} \to \frac{1}{1+1} = 1/2$$

for any $x \in \mathbb{R}^d/\mathbb{R}\mathbf{1}$. Also for $c \gg 0$ and fixed x, we have:

$$\max\{Ax+cA\mathbf{1}+b,0\} \quad = \quad Ax+cA\mathbf{1}+b.$$

As $c \to \infty$, we have

$$\frac{1}{1+\exp(-(Ax+cA\mathbf{1}+b))} \to 1$$

for any $x \in \mathbb{R}^d/\mathbb{R}\mathbf{1}$. Therefore, neural networks with the ReLU cannot learn from observations in these cases. However, with the activation function defined in Eq. (4.2.1), we have

$$\max_i(x_i + c\cdot\mathbf{1} + w_{ji}^{(1)}) - \min_i(x_i + c\cdot\mathbf{1} + w_{ji}^{(1)}) = \max_i(x_i + w_{ji}^{(1)}) - \min_i(x_i + w_{ji}^{(1)}).$$

In other words, classical neural networks are not well-defined in the tropical projective torus, since the neuron values are not invariant under transformations of the form $x \to x + (c,\ldots,c)$. Meanwhile, the tropical embedding layer of Eq. (4.2.1) is invariant under such transformations.

4.3 Universal Approximation Theorems for Tropical Neural Networks

It is very important to check if the tropical embedding layer as in Eq. (4.2.1) followed by classical layers has enough varieties to represent considerable input-output relations (Calin, 2020). In this section, we show that the tropical neural network can approximate enough variety of functions so that we can safely use it.

Definition 4.3.1. *The norm* $\|\cdot\|_q$ *for* $q \geq 1$ *is defined by*

$$\|f\|_q = \int_{\mathbb{R}^n} |f(x)|^q dx$$

The space $L^q(\mathbb{R}^d), (1 \leq q < \infty)$, *is the set of Lebesgue integrable functions* f *from* $\mathbb{R}^d$ *to* $\mathbb{R}$ *for which* $\|f(x)\|_q < \infty$.

Definition 4.3.2. *The space* $C^0(\mathbb{R}^d)$ *is the set of continuous, compactly supported functions from* $\mathbb{R}^d$ *to* $\mathbb{R}$.

Remark 14: Note that $C_0(\mathbb{R}^d) \subset L^q(\mathbb{R}^d)$.

For the classical case, a universal approximation theorem for ReLU feedforward neural networks has been proved in Arora et al. (2016).

Theorem 4.3.3 (classical universal approximation theorem Arora et al. (2016))**.** *Any function of* x_j *for* $j = 1, \ldots, d$ *in* $L^q(\mathbb{R}^d), (1 < q < \infty)$, *can be arbitrarily well approximated in* $\|\cdot\|_q$ *by a ReLU feedforward neural network with at most* $L = 2(\lfloor \log_2 d \rfloor + 2)$ *layers.*

As the $d-1$ neurons in the tropical embedding layer can easily represent $(x_j - x_d)$ for $j = 1, \ldots, d-1$ and Theorem 4.3.3 can be applied to the second and later layers of a tropical neural network (that is equivalent to a classical neural network), we can prove the following theorem.

Theorem 4.3.4 (tropical universal approximation theorem)**.** *Any function of* $(x_j - x_d)$ *for* $j = 1, \ldots, d-1$ *in* $L^q(\mathbb{R}^d/\mathbb{R}\mathbf{1}) \simeq L^q(\mathbb{R}^{d-1}), (1 < q < \infty)$, *can be arbitrarily well approximated in the* $\|\cdot\|_q$ *by a tropical neural network with at most* $L = 2(\lfloor \log_2 d \rfloor + 2) + 1$ *layers (which include an tropical embedding layer as the first layer).*

Proof. For any $f \in L^q(\mathbb{R}^{d-1})$, $\exists g \in C_0(\mathbb{R}^{d-1})$ such that $\|f - g\|_q < \epsilon/2$ Calin (2020). Let K be the support of g and let M be $\max_{x \in K} \|x\|$. For $x \in K$, we can set $w_{jj}^{(1)} =$

$-w_{jd}^{(1)} = 2M$ and $w_{ji}^{(1)} = 0$ for $i \neq j, d$ to obtain $z_j = x_j - x_d + 4M$ for $j = 1, \ldots, d-1$. This means that a neuron in the first tropical embedding layer can represent $x_j - x_d$. Then $d-1$ neurons can represent z_1, z_2, …, z_{d-1}. Finally, simply apply Theorem 4.3.3 to the classical neural network $F(z_1, \ldots, z_{d-1})$ consisting of the second and later layers of a tropical neural network to obtain $\|g - F\|_q < \epsilon/2$. Taken together, $\|f - F\|_q < \|f - g\|_q + \|g - F\|_q < \epsilon$. □

There is another type of classical universal approximation theorems.

Definition 4.3.5. *The width d_m of a neural network is defined to be the maximal number of neurons in a layer.*

Theorem 4.3.6 (classical universal approximation theorem for width-bounded ReLU networks Lu et al. (2017))**.** *For any $f \in L^1(\mathbb{R}^d)$ and any $\epsilon > 0$, there exists a classical neural network $F(x)$ with ReLU activation functions with width $d_m \leq d+4$ that satisfies*

$$\int_{\mathbb{R}^d} |f(x) - F(x)| dx < \epsilon.$$

Again, as the $d-1$ neurons in the tropical embedding layer can easily represent $(x_j - x_d)$ for $j = 1, \ldots, d-1$ and Theorem 4.3.6 can be applied to the second and later layers of a tropical neural network (that is equivalent to a classical neural network), we can prove the following theorem.

Theorem 4.3.7 (tropical universal approximation theorem with bounded width)**.** *For any function f of $(x_j - x_d)$ for $j = 1, \ldots, d-1$ in $L^1(\mathbb{R}^d/\mathbb{R}\mathbf{1}) \simeq L^1(\mathbb{R}^{d-1})$ and any $\epsilon > 0$, there exists a tropical neural networks $F(x)$ with width $d_m \leq d+4$ that satisfies*

$$\int_{\mathbb{R}^{d-1}} |f(x) - F(x)| dx < \epsilon.$$

Proof. For any $f \in L^1(\mathbb{R}^{d-1})$, $\exists g \in C_0(\mathbb{R}^{d-1})$ such that $\|f - g\|_1 < \epsilon/2$ Calin (2020). Let K be the support of g and let M be $\max_{x \in K} \|x\|$. For $x \in K$, we can set $w_{jj}^{(1)} =$

$-w_{jd}^{(1)} = 2M$ and $w_{ji}^{(1)} = 0$ for $i \neq j, d$ to obtain $z_j = x_j - x_d + 4M$ for $j = 1, \ldots, d-1$. This means that a neuron in the first tropical embedding layer can represent $x_j - x_d$. Then $d-1$ neurons can represent z_1, z_2, $\ldots$, z_{d-1}. Finally, simply apply Theorem 4.3.6 to the classical neural network $F(z_1, \ldots, z_{d-1})$ consisting of the second and later layers of a tropical neural network to obtain $\|g - F\|_q < \epsilon/2$. Taken together, $\|f - F\|_q < \|f - g\|_q + \|g - F\|_q < \epsilon$. □

4.4 Backpropagation Rule for Simplest Tropical Neural Networks

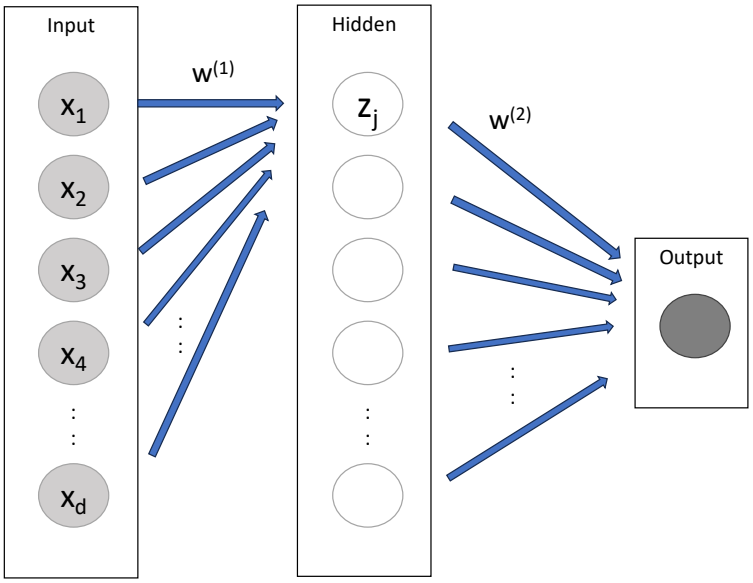


Figure 4.4.1: architecture of a simplest neural networks that accept a vector in $\mathbb{R}^d/\mathbb{R}\mathbf{1}$

Here we demonstrate that the gradients of the loss function with respect to weights exist for tropical neural networks. The gradient is computable through the chain rule for differentials called backpropagation rule in the similar fashion to the classical case. The gradients obtained in this way can guarantee the successful update of the weights at each iteration of learning.

We consider a simplest three layer network whose weights in the first and the second layers are denoted as $w^{(1)} \in \mathbb{R}^{d\times N}$ and $w^{(2)} \in \mathbb{R}^{N\times 1}$. Suppose the activity in the first hidden layer is given by Eq. (4.2.1) and the output of the network is given as

$$y = \sum_{j}^{N} w_j^{(2)} z_j. \tag{4.4.1}$$

Note that although here we derive a backpropagation rule for this regression setting just for simplicity, the backpropagation rule can be derived in a similar manner for the classification setting with a sigmoid function, too.

Below is the summary of parameters of the neural network.

- $w^{(1)}$, $w^{(2)}$: weights in the first and second layers;
- z_j: activation of j-th neuron in the hidden layer; and
- y: activation of the output neuron.

Theorem 4.4.1. *The partial derivatives of the cost function $Q := \frac{1}{2}(y - y^{true})^2$ with respect to weights for the above tropical neural network $y = f(x)$ are given by*

$$\frac{\partial Q}{\partial w_{ji}^{(1)}} = (y - y^{true}) w_j^{(2)} (\delta(i = i_j^{max}) - \delta(i = i_j^{min})), \tag{4.4.2}$$

where i_j^{max} (or i_j^{min}) is the index i for which $(x_i + w_{ji}^{(1)})$ takes the maximum (or minimum) and

$$\frac{\partial Q}{\partial w_j^{(2)}} = (y - y^{true}) z_j.$$

Proof. Direct calculations. □

Example 13: As a simplest example of Eq. (4.4.2), let us consider the three dimensional input case ($d = 3$). Suppose the number of neurons in the middle layer is one and its activity is z, for simplicity. Assume $x_1 = 1$, $x_2 = 2$ and $x_3 = 3$ and $w_1^{(1)} = w_2^{(1)} = w_3^{(1)} =$

0. Then $(x_i + w_1^{(1)}) < (x_i + w_2^{(1)}) < (x_i + w_3^{(1)})$ and $i_{\max} = 3$ and $i_{\min} = 1$. Therefore,

$$\frac{\partial Q}{\partial w_i^{(1)}} = \begin{cases} -(y - y^{\text{true}})w^{(2)} & (i = 1) \\ 0 & (i = 2) \\ (y - y^{\text{true}})w^{(2)} & (i = 3). \end{cases}$$

In this case we have $z = 3 - 1 = 2$ If, furtheremore, $w^{(2)} = 1$, then, $y = w^{(2)}z = 2$ and

$$\frac{\partial Q}{\partial w_i^{(1)}} = \begin{cases} -(2 - y^{\text{true}}), & (i = 1) \\ 0, & (i = 2) \\ 2 - y^{\text{true}}. & (i = 3), \end{cases}$$

$w_i^{(1)}$ can be, for example, updated by the SGD rule: $\Delta w_i^{(1)} = -\eta \frac{\partial Q}{\partial w_i^{(1)}}$, where $\eta > 0$ is a learning rate. Then, $w_3^{(1)}$ increases (and $w_1^{(1)}$ decreases) if $2 > y^{\text{true}}$.

Remark 15: It is interesting that only two of $w_i^{(1)}$ are modified while the other remains. Note that $\Delta w^{(1)}$ is orthogonal to the one vector, $\mathbf{1} := (1, 1, \ldots, 1) \in \mathbb{R}^d$. It is interesting to elucidate how this learning rule works as a dynamical system.

4.5 TensorFlow2 Codes for Tropical Neural Networks

In order to boost computing with GPUs, we implement tropical neural networks in TensorFlow 2 Chollet (2021). As is the case for the classical neural networks, the auto-differential is the key for the GPU implementation of tropical neural networks. In order to guarantee fast auto-differentials, all the calculations must be implemented only with the math functions in TensorFlow2 such as top_k(v,d), which returns the maximum and the minimum of a vector v.

In practice, it is essential to create a user-friendly class for the tropical embedding

as the first layer of the tropical neural networks, that is scalable for big data. The following code defines a hand-made class called TropEmbed(), which enables us to easily implement the tropical neural networks in the Keras/Tensorflow style.

```
class TropEmbed(Layer):
    def __init__(self, units=2, input_dim=3):
        super(TropEmbed, self).__init__()
        self.w = self.add_weight(shape=(units, input_dim), \
                initializer="random_normal")
        self.units = units
        self.input_dim = input_dim

    def call(self, x):
        x_reshaped = tf.reshape(x,[-1, 1, self.input_dim])
        Bcast = repeat_elements(x_reshaped, self.units, 1)
        val, i = tf.math.top_k(Bcast + self.w, self.input_dim)
        return val[:,:,0] - val[:,:,-1]

# usage
model = Sequential([TropEmbed(10, d), Dense(1)])
```

The codes for TropEmbed() class and for reproducing all the figures in this paper are available at https://github.com/keiji-miura/TropicalNN.

4.6 Weight Initialization Based on Extreme Value Statistics

Weight initializations are important for avoiding the divergence and vanishment of neural activities after propagating many layers. For the classical neural networks, Xavier's and He's initializations are famous Glorot and Bengio (2010); He et al. (2016). The analysis relies on extreme value theory in the high dimensional regime $d \to \infty$, where the

tropical distance is governed by the maximum and minimum of Gaussian variables. The reader is reminded that the expected value of the Gumbel$(0,1)$ is the Euler-Mascheroni constant γ.

Definition 4.6.1 (Generalized Hilbert Projective Metric)**.** *For any points* $v := (v_1, \ldots, v_d)$, $w := (w_1, \ldots, w_d) \in \mathbb{R}^d/\mathbb{R}\mathbf{1}$*, the* tropical distance *(also known as* tropical metric*)* d_{tr} *between* v *and* w *is defined as:*

$$d_{\mathrm{tr}}(v, w) := \max_{i \in \{1,\ldots,d\}} \{v_i - w_i\} - \min_{i \in \{1,\ldots,d\}} \{v_i - w_i\}.$$

Lemma 4.6.2. *Suppose* $x_i, w_i \overset{\text{iid}}{\sim} N(0,1)$ *are independent for* $i = 1, \ldots, d$*. Then the expectation and variance of* $d_{\mathrm{tr}}(x, -w)$ *can be approximated by* $2\sqrt{2}(a_d\gamma + b_d)$ *and* $\frac{\pi^2}{3\log d}$*, respectively, where* $a_d = \frac{1}{\sqrt{2\log d}}$ *and* $b_d = \sqrt{2\log d} - \frac{\log\log d + \log(4\pi)}{2\sqrt{2\log d}}$*, as* $d \to \infty$*.*

Proof. As $x_i + w_i \sim N(0,2)$, $Z := \frac{\max\{x+w\}/\sqrt{2} - b_d}{a_d} \sim Gumbel(0,1)$ as $d \to \infty$. Therefore, $\mathrm{Ex}[d_{\mathrm{tr}}(x,-w)] = \mathrm{Ex}[2\max\{x+w\}] \xrightarrow[d\to\infty]{} 2\sqrt{2}(a_d\mathrm{Ex}[Z] + b_d)$. $\mathrm{Var}[d_{\mathrm{tr}}(x,-w)] = \mathrm{Var}[\max\{x+w\} - \min\{x+w\}] = 2\mathrm{Var}[\max\{x+w\}] + 2\mathrm{Cov}[\max\{x+w\}, -\min\{x+w\}] \xrightarrow[d\to\infty]{} 2 \times 2a_d^2\mathrm{Var}[Z] = 2 \times 2a_d^2\frac{\pi^2}{6}$ where $\mathrm{Cov}[\max\{x+w\}, -\min\{x+w\}] \xrightarrow[d\to\infty]{} 0$ follows from asymptotic independence of extremes. □

Here we confirm that the above scaling holds actually by numerical calculations.

One way for better weight initialization is to choose the scale of w so that the variance of the neural activity in the embedding layer becomes 1.

Lemma 4.6.3. *Suppose* $x_i \overset{\text{iid}}{\sim} N(0,1)$ *and* $w_i \overset{\text{iid}}{\sim} N(0, \frac{6\log d}{\pi^2} - 1)$ *for* $i = 1, \ldots, d$ *are independent of each other. Then the expectation and variance of* $d_{\mathrm{tr}}(x, -w)$ *can be approximated by* $2\sqrt{\frac{6\log d}{\pi^2}}(a_d\gamma + b_d)$ *and* 1*, respectively, as* $d \to \infty$*.*

Proof. As $x_i + w_i \sim N(0, \frac{6\log d}{\pi^2})$, $Z := \frac{\max\{x+w\}/\sqrt{\frac{6\log d}{\pi^2}} - b_d}{a_d} \sim Gumbel(0,1)$. Therefore, $\mathrm{Ex}[d_{\mathrm{tr}}(x,-w)] = \mathrm{Ex}[2\max\{x+w\}] \xrightarrow[d\to\infty]{} 2\sqrt{\frac{6\log d}{\pi^2}}(a_d\mathrm{Ex}[Z] + b_d)$. $\mathrm{Var}[d_{\mathrm{tr}}(x,-w)] \xrightarrow[d\to\infty]{} 2 \times \frac{6\log d}{\pi^2}a_d^2\mathrm{Var}[Z] = 2\frac{6\log d}{\pi^2}a_d^2\frac{\pi^2}{6} = 1$. □

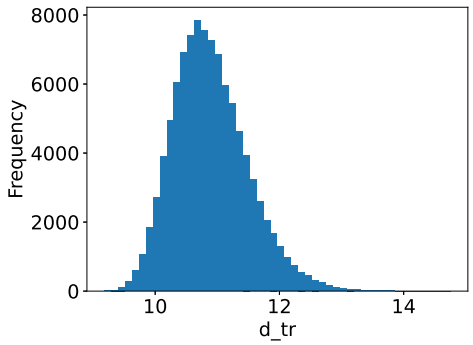


Figure 4.6.1: Histogram of simulated $d_{\text{tr}}(x, -w)$ for the same situation as in Lemma 4.6.2 with $d = 10000$. For the histogram, 100000 samples of $d_{\text{tr}}(x, -w)$ are used. The simulated mean is 10.893 while the theoretical prediction is 10.954. The simulated std is 0.604 while the theoretical prediction is 0.598. The mean and std are the same as predicted from the theory.

To control the standard deviation of the weights, you can customize an initializer (instead of simply specifying initializer = "random_normal") in the definition of TropEmbed class.

```
> ini = tf.keras.initializers.RandomNormal(mean=0., stddev=1.)
> self.w = self.add_weight(shape=(units, input_dim), \
      initializer=ini)
```

However, as the weight initialization should be done together with the data preprocessing, in this paper we entirely use the default value of stddev=0.05 for "random_normal" for simplicity.

```
> self.w = self.add_weight(shape=(units, input_dim), \
      initializer="random_normal")
```

4.7 Computational Experiments

In this section, we apply tropical neural networks with one hidden layer (the tropical embedded layer) to simulated data as well as empirical data. Then later we compare its performance against neural networks with one hidden layer with ReLU activator.

4.7.1 Small simulated data

First we illustrate our tropical neural networks with one hidden layer with 16 neurons and with one output Sigmoid function using a small example. First we generate two dimensional $16 + 16$ random points from the Gaussian distributions with means $(0.5, -0.5)$ and $(-0.5, 0.5)$ with the covariance matrix being the identity matrix. Then these points are randomly translated by (c, c) where c is a Gaussian random variable whose standard deviation is 4. The left and right figures in Figure 4.7.1 show the actual test labels and the predicted probabilities of the test data by the tropical neural networks with one hidden layer with 16 neurons.

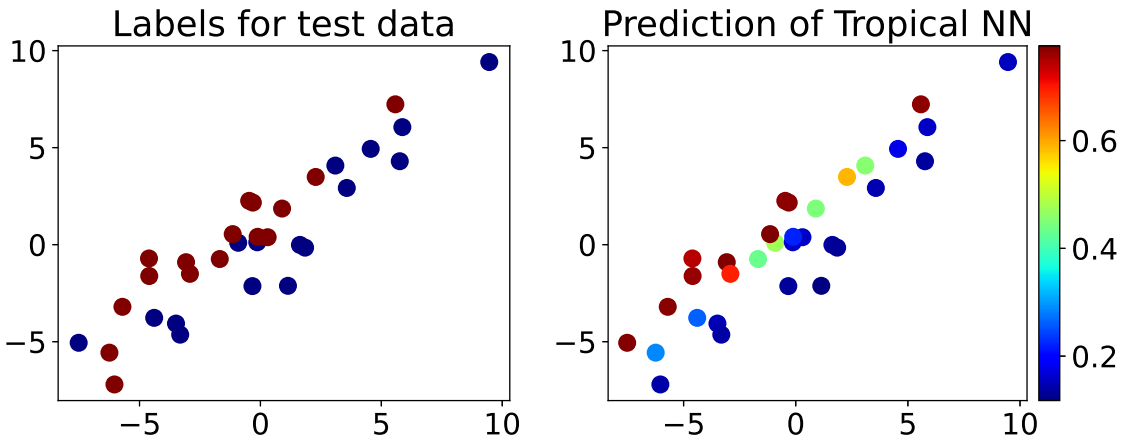


Figure 4.7.1: Predicted probabilities by the tropical neural networks on a small example.

4.7.2 High-dimensional simulated data

We now demonstrate that a tropical neural network with one hidden layer with 8 neurons and with one output sigmoid function works against the curse of dimensionality, where most of the variables in this high dimensional data are rather just noises Yoshida et al. (2023c). We generate d dimensional $16 + 16$ random points from the Gaussian distributions with means $(0.5, -0.5, 0, \ldots, 0)$ and $(-0.5, 0.5, 0, \ldots, 0)$ with the unit covariance matrix. Then these points are randomly translated by $(c, c, c, \ldots, c)$ where c

is a Gaussian random variable whose standard deviation is 6. The classification task for both classical and tropical neural networks is to assign points to the distributions from which they were generated. Figure 4.7.2 shows that, as the input dimension increases, the test accuracy of classical neural networks approaches 0.5, corresponding to random guessing, whereas tropical neural networks do not exhibit this behavior. The result demonstrates that the tropical neural networks work robustly against the curse of dimensionality.

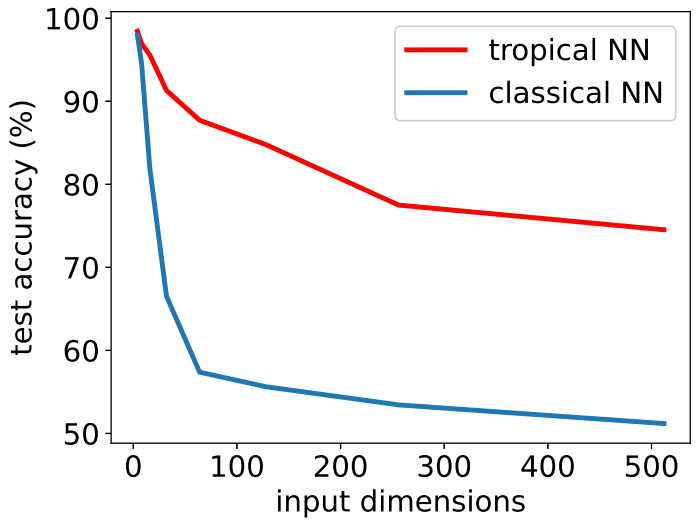


Figure 4.7.2: Application of the tropical neural networks to a high-dimensional example. The test accuracy averaged over 100 trials is plotted. The tropical neural networks work robustly against the curse of dimensionality.

4.7.3 Simulated data generated from the multi-species coalescent model

In this subsection we apply the tropical neural networks to a sample of phylogenetic trees generated under the *multi-species coalescent model*.

A phylogenetic tree is a weighted tree whose leaves are labeled with $[m] := \{1, 2, \ldots, m\}$, where m is the number of leaves, and whose internal nodes are unlabeled. A weight on each edge of a phylogenetic tree is considered as a distance from a node to another

node on the tree and in evolutionary biology, a weight on an edge can be considered as a product of an evolutionary time and mutation rate (Semple and Steel, 2003). In this paper we consider a rooted phylogenetic tree with a leaf label set $[m]$. A tree with m leaves is called an *equidistant tree* if the total weights on the unique path from its root to each leaves is the same for all leaf in $[m]$. Under the multispecies coalescent model which is commonly used to analyze gene trees (i.e. phylogenetic trees reconstructed from genes in a genome), gene trees are assumed to be equidistant. Therefore, throughout this chapter, we restrict attention to equidistant phylogenetic trees.

To conduct a statistical analysis on a set of phylogenetic trees, we consider a *space of phylogenetic trees* with fixed $[m]$. A space of phylogenetic trees on $[m]$ is a set of all possible phylogenetic trees with $[m]$ and it is well-known that it is not Euclidean Speyer and Sturmfels (2009). It is also well-known that the space of all possible equidistant trees on $[m]$ with the *tropical metric* under the max-plus algebra is a subspace of the tropical projective space Ardila and Klivans (2006); Yoshida et al. (2019). In order to define the space of equidistant trees, first we define *ultrametrics*. Consider a map $u : [m] \times [m] \to \mathbb{R}$ such that $u(i, j) = u(j, i)$ and $u(i, i) = 0$. This map is called a *dissimilarity map* on $[m]$. If a dissimilarity map u satisfies that

$$\max\{u(i, j), u(i, k), u(j, k)\}$$

is achieved at least twice, then we call u an *ultrametric*.

Example 14: Suppose $m = 3$, i.e., $[3] = \{1, 2, 3\}$ and suppose

$$u(1, 2) = u(2, 1) = 1, u(1, 3) = u(3, 1) = 1, u(2, 3) = u(3, 2) = 0.5$$

and $u(i, i) = 0$ for all $i = 1, 2, 3$. Since

$$\max\{u(1, 2), u(1, 3), u(2, 3)\} = 1$$

and it achieves twice, i.e., $u(1,2) = u(1,3) = 1$. Thus, u is an ultrametric.

Consider dissimilarity maps u_T on a phylogenetic tree T on $[m]$ such that $u(i,j)$ is the total weights on the unique path from a leaf i to a leaf j for all $i, j \in [m]$. Then we have the following theorem:

Theorem 4.7.1 (Buneman (1974)). *Consider an equidistant tree T on $[m]$. Then u_T realizes an equidistant tree T on $[m]$ if and only if a dissimilarity map u_T is ultrametric.*

Example 15: Suppose we have an ultrametric from Example 14. An equidistant tree whose dissimilarity maps are ultrametric in Example 14 is a rooted phylogenetic tree with leaves $[3] = \{1, 2, 3\}$ shown in Figure 4.7.3.

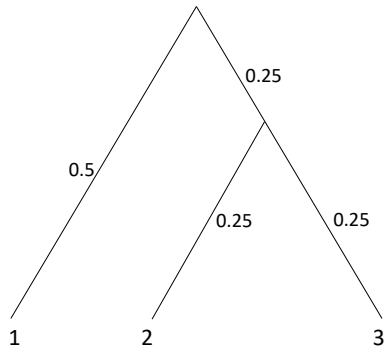


Figure 4.7.3: An equidistant tree with the dissimilarity maps which are ultrametric shown in Example 14.

Therefore, we consider the space of all possible ultrametrics on $[m]$ as the space of equidistant trees on $[m]$. Then we have the following theorem:

Theorem 4.7.2 (Ardila and Klivans (2006)). *The space of ultrametrics on $[m]$ is the tropicalization of the linear subspace defined by linear equations*

$$x_{ij} - x_{ik} + x_{jk} = 0$$

for $1 \leq i \leq j \leq k \leq m$ *by replacing sum with max operation and replacing multiplication by a classical summation.*

The space of ultrametrics on $[m]$ is a subspace of the tropical projective space $(\mathbb{R}^e \cup \{-\infty\})/\mathbb{R}\mathbf{1}$ where $e = \binom{m}{2}$. Therefore, we apply our method, tropical neural networks, to simulated data generated from the multi-species coalescent model using a software Mesquite Maddison and Maddison (2009).

The multi-species coalescent model has two parameters: species depth and effective population size. Here we fix the effective population size $N_e = 100000$ and we vary

$$R = \frac{SD}{N_e}$$

where SD is the species depth. We generate species trees using the Yule process. Then we use the multi-species coalescent model to generate gene trees with a given species tree. In this experiment, for each R, we generate two different set of 1000 gene trees: In each set of gene trees, we have a different species tree so that each set of gene trees is different from the other. We conduct experiments with $R = 0.25, 0.5, 1, 2, 5, 10$.

Note that when we have a small ratio R, then gene trees become more like random trees since the species tree constrains less on tree topologies of gene trees. Thus it is more difficult when we have a small R while if we have a large R, then it is easier to classify since the species tree constrains more on tree topologies of gene trees.

In this experiment, we set one hidden layer for each neural network: neural network with ReLU activators and neural network with tropical activators. We set the Sigmoid function in the output node in both neural networks. In each neural network, we set 1000 neurons in the hidden layer.

Figure 4.7.4 shows ROC curves for neural networks with ReLU and tropical neural networks. In general tropical neural networks perform better than neural networks with ReLU activation function.

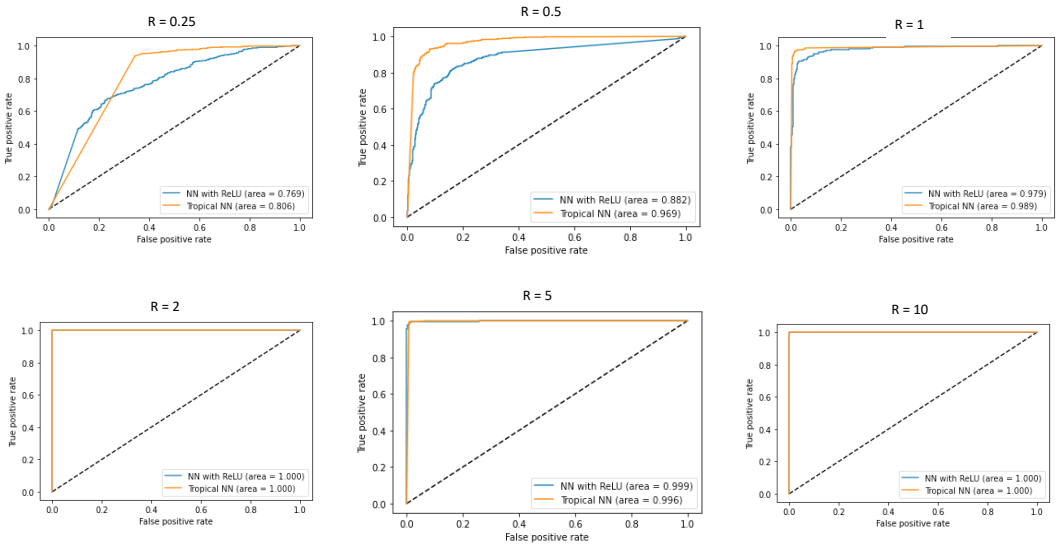


Figure 4.7.4: ROC curves for neural networks with ReLU and tropical neural networks with one hidden layer. We conduct experiments with $R = 0.25, 0.5, 1, 2, 5, 10$.

4.7.4 Influenza data

In this subsection we apply our method to genomic data for 1089 full length sequences of hemagglutinin (HA) for influenza A H3N2 from 1993 to 2017 collected in the state of New York obtained from the GI-SAID EpiFlu database (www.gisaid.org). These collected data were aligned using muscle developed by Edgar (2004) with the default settings. Then we apply the neighbor-joining method with the p-distance Saitou and Nei (1987b) to reconstruct a tree from each sequenced data. Each year corresponds to the first season. We also apply KDETrees Weyenberg et al. (2016) to remove outliers and a sample size of each year is about 20,000 trees.

We apply tropical neural networks and neural networks with ReLU with one hidden layer with 10 neurons to all pairs of different years to see if they are significantly different one year to the other. Heatmaps of accuracy rates with the probability threshold 0.5 and AUC values are shown in Figure 4.7.5. Again, tropical neural networks outperform classical neural networks.

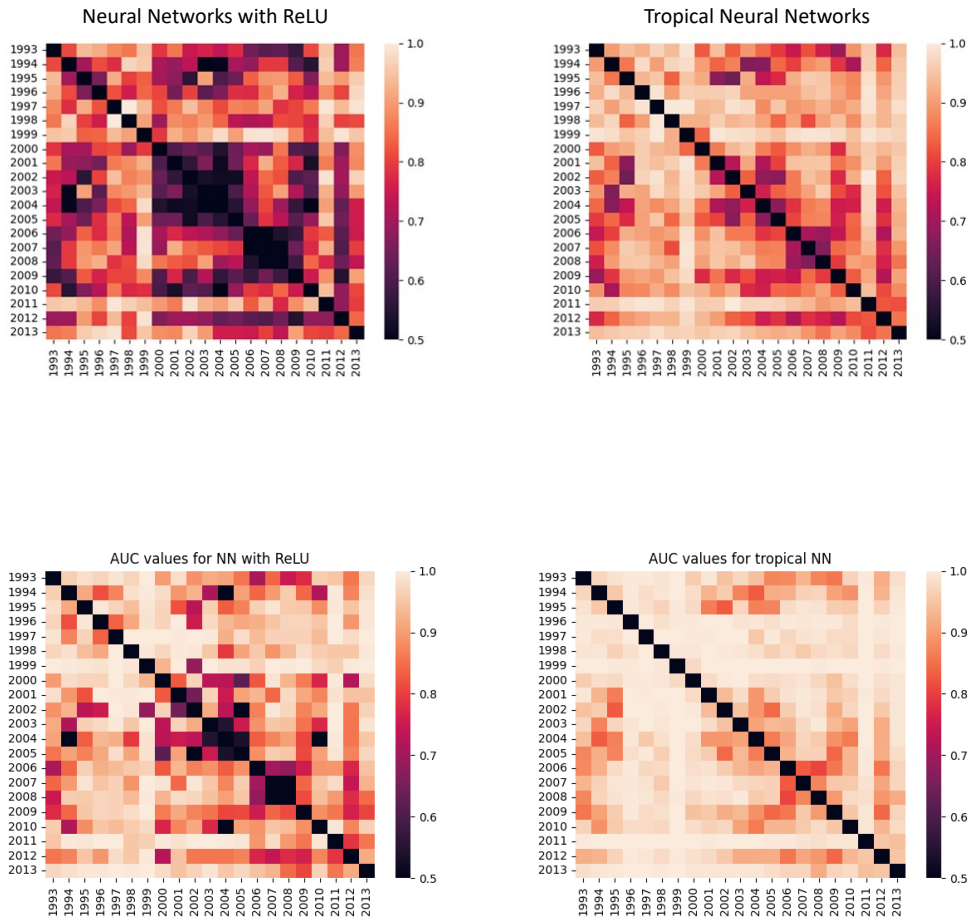


Figure 4.7.5: Heat maps for (top) classification rates with threshold 0.5 and (bottom) AUC values for classical neural networks with ReLU (left) and tropical neural networks (right).

4.7.5 Tropical Neural Network as a Generalization of Tropical Logistic Regression for Classification of Gene Trees

A tropical neural network can be interpreted as a generalization of a tropical logistic regression. The tropical logistic regression model Aliatimis et al. (2024) is developed for binary classification on gene trees and it has been proven to have higher predictive power than classical logistic regression to classify phylogenetic trees. It assumes that if X is

the covariate (gene tree), the binary response random variable is $Y \sim \text{Bernoulli}(p(X))$, with

$$p(x) = S\left(\lambda_0 d_{\text{tr}}(x, \omega_0) - \lambda_1 d_{\text{tr}}(x, \omega_1) + C\right),$$

S is the sigmoid function, ω_0, $\omega_1 \in \mathbb{R}^e/\mathbb{R}\mathbf{1}$, and $\lambda_0, \lambda_1, C \in \mathbb{R}$ with $\lambda_0\lambda_1 \geq 0$, and where m is the number of leaves in each phylogenetic tree in the given sample, and $e := \binom{m}{2}$. Note that this model is a special case for Eq. (4.4.1), with the sigmoid function as the link, and two neurons in the hidden layer whose weights are $w_0^{(2)} = \lambda_0, w_1^{(2)} = -\lambda_1$ and $w_0^{(1)} = -\omega_0, w_1^{(1)} = -\omega_1$. Therefore, tropical logistic regression is almost identical to a tropical neural network consisting of one tropical embedding layer with two neurons and a classical layer, with the additional assumption that $w_0^{(2)} w_1^{(2)} \leq 0$.

The one-species model described in Aliatimis et al. (2024) can be considered to be a neural network with e neurons in the input layer, no hidden layers and a unique output neuron. The activation function is the logistic function and the inner product used is tropical defined as

$$\langle x, -\omega \rangle := d_{\text{tr}}(x, \omega) - C, \tag{4.7.1}$$

where C can be considered to be a bias variable, similarly to the intercept variable in classical models. Tropical logistic regression returns the sigmoid of the tropical inner product. We define the tropical generalised linear model as an extension to tropical logistic regression, where instead of the sigmoid function, we may use a different link/activation function. If there are multiple outputs (multivariate generalized linear model (GLM)) and if we treat the output layer as the new input layer and iterate this L times, then we have an L-layer neural network. In the same way that classical neural networks are a stack/recursive application of classical multivariate GLMs, tropical neural networks can be a stack of tropical multivariate GLMs. Effectively, all is identical to classical networks, but instead of applying classical inner products, we apply tropical inner products as defined in Eq. (4.7.1). The i-th neuron of the l-th layer is defined as

$x_i^{(l)}$ and computed through the recursive formula,

$$x_i^{(l)} = d_{\mathrm{tr}}\left(x_i^{(l-1)}, \omega_i^{(l)}\right) - C_i^{(l)}, \tag{4.7.2}$$

where $\Omega^{(l)} = (\omega_1^{(l)}, \omega_2^{(l)}, \ldots, \omega_{N_l}^{(l)}) \in \mathbb{R}^{N_{l-1} \times N_l}$ is the weight matrix between layer $(l-1)$ and l for the number N_s of neurons in layer s, and $C^{(l)} \in \mathbb{R}^{N_l}$. By assuming that all neurons share the same bias variable $c = C_i^{(l)}$ for all $i \in [N_l]$, Eq. (4.7.2) reduces to Eq. (4.2.1), since vectors are defined up to an additive constant vector $(c, \ldots, c)$ in the tropical projective torus. When the last tropical embedding layer connects to the first classical layer, the constant bias vector is incorporated in the bias term of the classical layer. Hence, tropical bias terms are redundant and not considered in the development of tropical neural networks. Thus, the tropical neural network which we propose in this paper follows naturally as an extension of the tropical logistic regression model.

4.8 Summary and Discussion

In this paper, we first developed a tropical embedding layer. We used the tropical embedding layer as the first layer of the classical neural networks to keep the invariance in the tropical projective space. To check if this tropical neural network has enough flexibility, we next proved that this tropical neural network is a universal approximator. After we derived a backpropagation rule for the tropical neural networks, we provided TensorFlow 2 codes for implementing a tropical neural network in the same fashion as the classical one, where the weights initialization problem is considered according to extreme value statistics. Finally we showed some applications as examples.

The tropical neural networks with the tropical metric worked better than the classical neural networks when the input data are phylogenetic trees which is included in the tropical projective torus. This is partly because only the tropical neural network can keep the invariance of the input vector under the one-vector which is innate in the

tropical projective torus.

One of the nice properties of tropical neural networks is its tractability and interpretability in analysis. The tropical embedding can be interpreted as taking the tropical distance to a point in the space of the tropical projective torus. The backpropagation rule of the tropical neural networks can be derived and interpreted rather easily.

The TensorFlow 2 codes for the Python class for tropical embedding was provided in the paper. This makes it possible to implement a tropical neural network in the same familiar fashion as the classical one. This facilitates, for example, to compare tropical and classical neural networks for the same data by using a common code.

Recent work shows that neural networks are vulnerable against adversarial attacks (i.e. small, carefully chosen changes in high-dimensional inputs can exploit the model's learned decision boundaries, causing confident but incorrect predictions) (Biggio et al., 2013; Nguyen et al., 2014; Szegedy et al., 2014; Madry et al., 2018). However, our initial computational experiments on image data from computer vision show that tropical neural networks are robust against gradient based methods, such as the Fast Gradient Sign Method Goodfellow et al. (2014) and Ensemble Adversarial Training Tramèr et al. (2018). It is interesting to investigate why tropical neural networks are robust against such attacks. In addition, it is interesting to develop adversarial attacks toward tropical neural networks.

Chapter 5

Covariance Decomposition

5.1 Introduction

A core problem in phylogenomics is the reconstruction of a species tree from gene alignments. A species tree encodes the evolutionary relationships and divergence times among given species, while gene trees represent the evolutionary histories of individual genes. The structure of a gene tree may differ from the species tree due to variation induced by processes such as incomplete lineage sorting (ILS), horizontal gene transfer and hybridization (Maddison, 1997). Finally, gene alignments are comparisons of DNA, RNA, or protein sequences across different species, from which gene trees may be inferred. The task of species tree reconstruction is to infer the species tree from gene alignments.

Hence, species tree reconstruction is commonly formulated as a hierarchical statistical model with three levels, as illustrated in Fig. 5.1.1. At the top level, the species tree $\mathcal{S}$ is the parameter of ultimate interest, to be estimated either as a point estimate or via a (approximate) posterior distribution. Given $\mathcal{S}$, the second level consists of m gene trees $\{G_i : i \in [m]\}$, whose likelihoods $f(G_i \mid \mathcal{S})$ are defined by the multispecies coales-

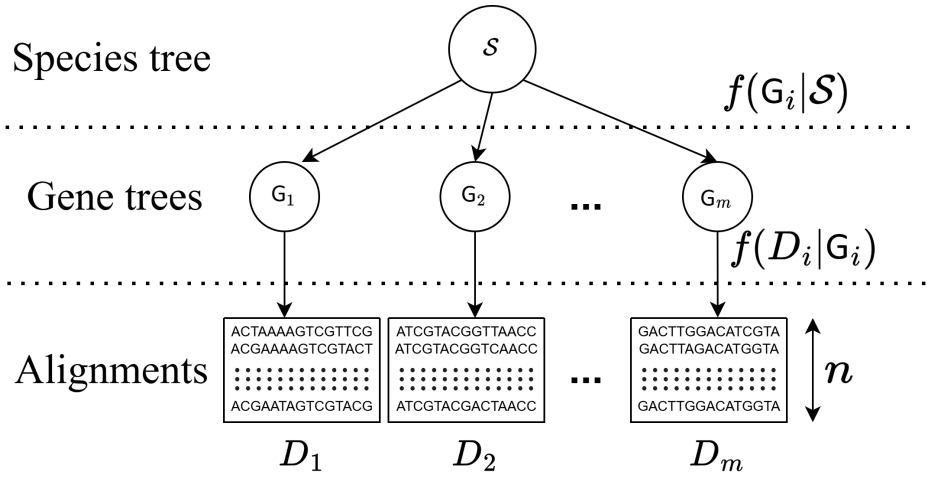


Figure 5.1.1: In phylogenetics, the goal is to reconstruct gene trees G_i for $i \in [m]$ (intermediate level) from m gene alignments (bottom level) through a nucleotide substitution model $f(D|G)$. In phylogenomics, the aim is to infer the species tree $\mathcal{S}$ (top level) from which these gene trees arose through the MSC model $f(G|\mathcal{S})$.

cent (MSC) model, which captures the stochasticity introduced by incomplete lineage sorting (ILS) under a given species tree (Degnan and Rosenberg, 2009). At the bottom level are the gene alignments $\{D_i : i \in [m]\}$, whose likelihoods $f(D_i \mid G_i)$ depend only on the corresponding gene trees, and are computed using nucleotide substitution models such as the Jukes-Cantor (JC69) model (Jukes, 1969), the Hasegawa-Kishino-Yano (HKY) model (Hasegawa et al., 1985), and the Generalized Time Reversible (GTR) model (Tavaré, 1986). In practice, we observe only the gene alignments. Depending on the method, one may first infer the gene trees $\{G_i : i \in [m]\}$ and then the species tree, or infer the species tree $\mathcal{S}$ directly from sequence data without explicit gene-tree reconstruction.

Viewing a species tree reconstruction via a hierarchical model has several conceptual advantages: it clarifies the distinct sources of randomness (under coalescent and substitution models), provides a foundation for Bayesian approaches (Yang and Rannala, 2012), and facilitates the development of statistical methods for quantifying uncertainty at each level. Moreover, it enables the integration of multiple data types and model

components within a coherent inferential framework, as seen in widely used tools such as MrBayes (Ronquist et al., 2012), BEAST (Bouckaert et al., 2019), and BPP (Flouri et al., 2018).

5.1.1 Species Tree Reconstruction Methods

In contrast to computationally intensive methods for species tree reconstruction, distance-based approaches based on summary statistics—such as METAL (Dasarathy et al., 2014), GLASS (Mossel and Roch, 2008), and STEAC (Liu et al., 2009)—offer simpler and more computationally efficient alternatives. These methods rely on estimated coalescent times or Hamming distances between species pairs, and then define summary statistics over these estimates across all genes. GLASS, which uses the *minimum* coalescent times across genes as its summary statistic, can provide more accurate species tree estimates than METAL or STEAC when gene tree uncertainty is minimal. In contrast, METAL and STEAC, which rely on *averages* as their summary measure, tend to be more robust in the presence of greater variance in gene tree estimates or forms of gene tree heterogeneity other than ILS.

Both STEAC and GLASS were originally developed as methods that take gene trees as input and produce an inferred species tree as output. Their application in phylogenomics thus depends on the accurate reconstruction of gene trees from gene sequence data. In contrast, METAL was specifically designed to infer species trees directly from gene sequences without requiring intermediate gene-tree estimation. As a result, METAL typically achieves accurate species tree reconstruction with fewer loci than methods that rely on gene-tree estimation, such as GLASS and STEAC. Notably, when working directly with sequence data, METAL requires only $\mathcal{O}(\tau^{-2})$ total sites to recover the correct species tree topology with a a specified high probability, where τ is the shortest internal branch length in the species tree. By comparison, GLASS and

STEAC require $\mathcal{O}(\tau^{-3})$ and $\mathcal{O}(\tau^{-4})$ total sites respectively (Dasarathy et al., 2014). However, GLASS can outperform METAL in regimes when the number of bases per gene is large; as substitutional noise diminishes, GLASS can yield more accurate estimates than METAL due to its minimum distance criterion.

Despite the computational efficiency and practical appeal of distance-based methods, their behavior under different sources of uncertainty has remained only partially understood. Prior work has noted limitations of early approaches such as GLASS and proposed refinements (e.g., iGLASS (Jewett and Rosenberg, 2012)) to mitigate sensitivity to mutational noise. Other studies have explicitly examined the impact of gene-tree variance and substitutional error on inference accuracy, including simulation-based analyses that teased apart their relative contributions (Huang et al., 2010). What remains less clear, however, is how these two sources of uncertainty—coalescent variation versus substitutional noise—interact across parameter regimes, and under which conditions methods such as STEAC, METAL, or more recent variants of GLASS will be most reliable. The relative influence of these error sources can shift dramatically with mutation rate, sequence length, and tree height, making a systematic understanding essential.

Although bootstrap resampling of concatenated sequences is common for assessing split support, it conflates coalescent and substitutional uncertainties and can yield misleading confidence values under extreme variance regimes. By deriving the full covariance matrix of METAL's pairwise distances, we introduce a Gaussian-sampling procedure that produces more accurate split support estimates than traditional bootstrapping.

5.1.2 Contributions

Considering the hierarchical nature of species tree reconstruction, uncertainty arises from two distinct sources; (1) variation in tree topologies governed by MSC, and (2) sequence-level stochasticity introduced by the substitution process used to infer gene trees from alignments (e.g., as JC69). By explicitly deriving the covariance matrices corresponding to each source, we obtain a quantitative framework for understanding how uncertainty propagates through distance-based estimates. Related work by Guerra and Nielsen (Guerra and Nielsen, 2022) has also derived the covariance of pairwise distances under the multispecies coalescent with mutation, in a more general setting that accommodates arbitrary species trees and varying ancestral population sizes. Our analysis differs in focusing specifically on distance-based species tree reconstruction methods (METAL, GLASS, STEAC), and in decomposing the total covariance into coalescent and substitutional components to study how their relative magnitudes govern method performance. In particular, we demonstrate that METAL outperforms GLASS when substitutional variance dominates the total uncertainty, and vice versa.

This decomposition enables informed decisions about study design, such as whether to prioritize collecting longer sequences or more independent loci. When coalescent variance dominates, increasing the number of loci is most effective; when substitutional noise is dominant, longer alignments (perhaps via imputation strategies) are preferable.

In Section 5.2, we analyze species tree reconstruction from known gene trees under METAL, GLASS, and STEAC. We compute the covariance structure of METAL and STEAC, and show that METAL effectively interpolates between STEAC and GLASS as a function of mutation rate. In Section 5.3, we study species tree reconstruction from gene base sequences, analyzing the statistical properties (signal and noise) of the distance estimates used by METAL and STEAC under finite sequence lengths. Section 5.4

presents our theoretical results: the decomposition of the total covariance of METAL and STEAC into coalescent and substitutional components, analysis of the spectrum of the METAL covariance matrix, and asymptotic expressions quantifying their relative magnitudes across a range of model parameters. Section 5.5 contains two empirical applications. Section 5.5.1 compares the performance of METAL, GLASS, and STEAC across regimes where substitutional or coalescent variance dominates, showing that the relative error of GLASS versus METAL depends critically on the proportion of variance attributable to substitutional noise. In Section 5.5.2, we introduce a Gaussian-sampling procedure for computing support values for METAL trees, enabled by our covariance derivations. We show empirically that this method provides more reliable split confidence than traditional bootstrapping. Finally, Section 5.6 concludes the paper with a discussion of the implications, limitations, and future directions of this work.

5.2 Species tree reconstruction from gene trees

First, we introduce the notation that will be used in the rest of the paper. The task is to successfully reconstruct the species tree $S = (V, E)$, where V and E is the set of nodes and edges in the tree respectively. The leaf set $L \subseteq V$ corresponds to the $n = |L|$ extant species examined. We assume throughout that the species tree is ultrametric, consistent with a molecular clock model in which all lineages evolve at the same rate, so that all root-to-leaf path lengths are equal. We define τ_e to be the length of $e \in E$ in coalescent units and the diameter of the tree as $\Delta = \max_{a,b \in L} \tau_{ab}$ where ab here refers to the edge connecting nodes a and b.

For simplicity, we sample a single individual per species, following the approach implemented in agglomerative, distance-based frameworks such as METAL (Braun et al., 2024). We consider m loci with associated gene trees $G^{(i)}, i \in [m]$ with the same leaf

set L, and define $g_{ab}^{(i)}$ as twice the time in coalescent units to the most recent ancestor of a and b in $G^{(i)}$. This is the evolutionary distance between the two leaves. In this section, we assume that the gene trees are known. Each of the three methods examined in this paper uses a dissimilarity map $d : L^2 \to \mathbb{R}_+$ and subsequently performs hierarchical clustering on that distance matrix. The resulting dendrogram is an estimate of the species tree. Pseudocode for these algorithms can be found in Appendix 5.C, Algorithm 4.

The dissimilarity maps are defined as follows:

$$d^{(\text{GLASS})}(a,b) = \min_{i\in[m]} g_{ab}^{(i)}, \tag{5.2.1}$$

$$d^{(\text{STEAC})}(a,b) = \frac{1}{m}\sum_{i=1}^{m} g_{ab}^{(i)},$$

$$d^{(\text{METAL})}(a,b) = \frac{1}{m}\sum_{i=1}^{m} p_{ab}^{(i)} = \frac{3}{4}\left(1 - \frac{1}{m}\sum_{i=1}^{m}\exp\left(-\mu g_{ab}^{(i)}\right)\right), \tag{5.2.2}$$

for all pairs $(a,b) \in \binom{L}{2}$ where $p_{ab}^{(i)}$ is the probability that the bases of species a and b differ at any given site in locus i under the Jukes-Cantor evolutionary model, and μ is rate of substitution.

Note that METAL does not require prior knowledge of the mutation rate μ when applied to empirical sequence alignments. In practice, one simply computes the normalized Hamming distance $\hat{p}_{ab}^{(i)}$ between alignments of species a and b for each locus i, which is the proportion of sites that differ between those two alignments, and then averages over loci to obtain

$$\hat{d}^{(\text{METAL})}(a,b) = \frac{1}{m}\sum_{i=1}^{m}\hat{p}_{ab}^{(i)}.$$

As the number of sites per locus tends to infinity, $\hat{p}_{ab}^{(i)} \to p_{ab}^{(i)}$, and thus $\hat{d}^{(\mathrm{METAL})}(a,b) \to d^{(\mathrm{METAL})}(a,b)$. Under the Jukes–Cantor model (Jukes, 1969), the p-distance p_{ab} is related to the Jukes-Cantor distance $d_{\mathrm{JC}}(a,b)$, which is measured in expected number of substitution time units, through

$$p_{ab} = \frac{3}{4}\left(1 - \exp\left(-\frac{4}{3}d_{\mathrm{JC}}(a,b)\right)\right),$$

which, yields Equation (5.2.2) by noting that $d_{\mathrm{JC}}(a,b) = \frac{3}{4}\mu g_{ab}$.

Proposition 5.2.1 shows that, under perfect gene-tree knowledge, the METAL distance "interpolates" between STEAC and GLASS: as the substitution rate $\mu \to 0$, it reduces to STEAC's average coalescent time, whereas as $\mu \to \infty$, it collapses to GLASS's minimum-coalescent-time criterion.

Proposition 5.2.1. *Let $\hat{T}_{\mathrm{METAL}}(\mu), \hat{T}_{\mathrm{STEAC}}, \hat{T}_{\mathrm{GLASS}}$ be the reconstructed species tree topologies of each method provided that the correct gene trees are given, assuming that all methods use the same hierarchical clustering method, and that all the minima attained by GLASS are unique. Then,*

$$\lim_{\mu \to 0} \hat{T}_{\mathrm{METAL}}(\mu) = \hat{T}_{\mathrm{STEAC}},$$
$$\lim_{\mu \to \infty} \hat{T}_{\mathrm{METAL}}(\mu) = \hat{T}_{\mathrm{GLASS}}.$$

For the second limit, it is assumed that the hierarchical clustering method is single or complete linkage clustering.

Proof. All proofs can be found in Appendix 5.A. □

Proposition 5.2.2 then establishes that, for any four leaves $a, b, c, d \in L$, the pairwise distances produced by STEAC and by METAL are both nonnegatively correlated, with

STEAC correlations always at least as large as those of METAL. The covariance matrix of METAL pairwise distances—assembled from the individual covariance components derived in the proof of Proposition 5.2.2—will be used for uncertainty quantification in Section 5.4 and to formulate the total covariance in our simulation results (Section 5.5).

Proposition 5.2.2. *Let $a, b, c, d \in L$ be leaves in the species tree. Then, under MSC,*

$$\text{Cor}\left(g_{ab}^{(1)}, g_{cd}^{(1)}\right) \geq \text{Cor}\left(e^{tg_{ab}^{(1)}}, e^{tg_{cd}^{(1)}}\right) \geq 0, \forall t \leq 0,$$

where Cor *denotes the correlation between two random variables. Moreover,* $\text{Cor}\left(e^{tg_{ab}^{(1)}}, e^{tg_{cd}^{(1)}}\right)$ *as a function of t is continuous and strictly increasing on $(-\infty, 0)$ with limit*

$$\lim_{t \to 0} \text{Cor}\left(e^{tg_{ab}^{(1)}}, e^{tg_{cd}^{(1)}}\right) = \text{Cor}\left(g_{ab}^{(1)}, g_{cd}^{(1)}\right).$$

Finally,

$$\text{Cor}\left(d^{(\text{STEAC})}(a,b), d^{(\text{STEAC})}(c,d)\right) = \text{Cor}\left(g_{ab}^{(1)}, g_{cd}^{(1)}\right)$$
$$\text{Cor}\left(d^{(\text{METAL})}(a,b), d^{(\text{METAL})}(c,d)\right) = \text{Cor}\left(e^{tg_{ab}^{(1)}}, e^{tg_{cd}^{(1)}}\right)$$

5.3 Gene tree reconstruction from base sequences

In the previous section, we considered properties of the distribution of gene trees $G_i|S$ under the MSC model. In this section, we investigate statistical properties of gene base sequence conditional on the gene tree or $\chi^i|G_i$. Let $\chi_a^{ij} \in \{A, C, T, G\}$ be the j^{th} base of the i^{th} locus of species $a \in L$, where $i \in [m]$, $j \in [K_i]$, m is the number of loci studied and K_i is the sequence length of loci i. For notational simplicity in the formulae that follow, we assume that all loci have the same number of bases $K = K_i$, $\forall i \in [m]$. The analysis can be readily extended to the case of loci with varying sequence lengths.

The normalized Hamming distance, $\hat{p}_{ab}^{(i)}$ and Jukes-Cantor distance $\hat{\nu}_{ab}^{(i)}$ between species $a, b \in L$ are then defined as

$$\hat{p}_{ab}^{(i)} = \frac{1}{K}\sum_{j=1}^{K} \mathbb{1}\left(\chi_a^{ij} \neq \chi_b^{ij}\right), \text{ and}$$
$$\hat{\nu}_{ab}^{(i)} = \frac{3}{4}\mu\hat{g}_{ab}^{(i)} = -\frac{3}{4}\log\left(1 - \frac{4}{3}\hat{p}_{ab}^{(i)}\right).$$

METAL computes the average of the normalized Hamming distances, while STEAC and GLASS consider the average and minimum Jukes-Cantor distances respectively. However, a problem with using the Jukes-Cantor distances directly is that they are infinite if $\hat{p}_{ab}^{(i)} \geq \frac{3}{4}$ for any $a, b \in L$, $i \in [m]$, which has a non-zero chance of occurring.

5.3.1 Hamming Distances

Given the gene tree $G^{(i)}$ of locus i, the distribution of $\hat{p}_{ab}^{(i)}$ is

$$\hat{p}_{ab}^{(i)} = \frac{1}{K}\sum_{j=1}^{K} \mathbb{1}\left(\chi_a^{ij} \neq \chi_b^{ij}\right) \sim \frac{1}{K}\text{Bin}\left(K, p_{ab}^{(i)}\right), \text{ where}$$
$$p_{ab}^{(i)} = \frac{3}{4}\left(1 - \exp\left(-\mu g_{ab}^{(i)}\right)\right).$$

Effectively, $p_{ab}^{(i)}$ is the true probability that species a and b would share the same base in locus i and $\hat{p}_{ab}^{(i)}$. The number of bases, K, is typically large and so we can apply the common Normal approximation of the Binomial distribution,

$$\hat{p}_{ab}^{(i)} \sim \mathcal{N}\left(p_{ab}^{(i)}, \frac{p_{ab}^{(i)}\left(1 - p_{ab}^{(i)}\right)}{K}\right).$$

While the normal approximation above applies to each pairwise distance $\hat{p}_{ab}^{(i)}$ individually, we require a joint multivariate description of all $\binom{L}{2}$ distances at locus i. In

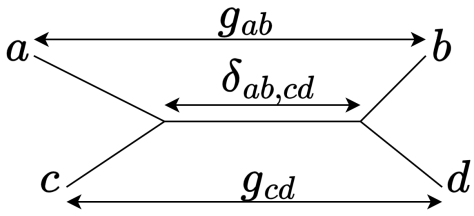


Figure 5.3.1: The evolutionary tree used in proposition 5.3.1. Here $\delta_{ab,cd} = \delta_{ac,bd}$ is the length of the intersection of the shortest path from a to b and the shortest path from c to d. All trees with 4 distinct leaves $a, b, c, d \in L$ can be drawn in this way. For cherry trees, the pairs $(a, c), (b, d)$ are sisters/cherries and the root of the 4-leaf tree is contained in the δ-segment. For comb trees, without loss of generality, the topology is $(((a, c), b), d)$ with a, c being sisters and the root of the tree contained in the shortest path from b to d. Note that since the paths from a to c and b, d are disconnected $\delta_{ac,bd} = 0$.

fact, the multivariate extension of that result is

$$\hat{p}^{(i)}_{\binom{L}{2}} \sim \mathcal{N}\Big(p^{(i)}_{\binom{L}{2}}, \tfrac{1}{K}\Sigma^{(i)}\Big),$$

where $\hat{p}^{(i)}_{\binom{L}{2}}, p^{(i)}_{\binom{L}{2}} \in \mathbb{R}^{\binom{n}{2}}$ are vectors containing the estimated and true pairwise distances respectively and $\Sigma^{(i)} \in \mathbb{R}^{\binom{n}{2}\times\binom{n}{2}}$ is the covariance matrix whose entries are of the form given by Proposition 5.3.1.

Proposition 5.3.1. *Let $a, b, c, d \in L$ be leaves of a gene tree, where the pairs a, c and b, d are sisters as shown in Fig. 5.3.1. Letting $\delta_{ab,cd}$ be the length of the intersection of the shortest path from a to b the shortest path from c to d, the covariance is*

$$\mathrm{Cov}\left(\mathbb{1}\left(\chi^{(i)}_a \neq \chi^{(i)}_b\right), \mathbb{1}\left(\chi^{(i)}_c \neq \chi^{(i)}_d\right)|G\right) = \frac{3}{16} e^{-\mu\left(g^{(i)}_{ab}+g^{(i)}_{cd}\right)}\left(e^{2\mu\delta^{(i)}_{ab,cd}} + 2e^{\mu\delta^{(i)}_{ab,cd}} - 3\right) \tag{5.3.1}$$

Using Proposition 5.3.1, we see in particular, that the diagonal entries of $\Sigma^{(i)}$ coincide with the univariate variances: by expanding Equation (5.3.1) and noting that $\delta_{ab,ab} = g_{ab}$, we have,

$$\mathrm{Var}\left(\mathbb{1}(\chi^{(i)}_a \neq \chi^{(i)}_b)|G\right) = p^{(i)}_{ab}\left(1 - p^{(i)}_{ab}\right).$$

5.3.2 Jukes-Cantor Distances

For the STEAC method, instead of using normalized Hamming distances, $\hat{p}_{ab}$, it uses estimates for the gene tree branch lengths

$$\hat{g}_{ab} = -\frac{1}{\mu}\log\left(1 - \frac{4}{3}\hat{p}_{ab}\right), \forall a, b \in L. \tag{5.3.2}$$

The major issue with this approach is that if at any loci $i \in [m]$ we get $\hat{p}_{ab}^{(i)} \geq 3/4$, then $\hat{g}_{ab}^{(i)} = \infty$ and so $\hat{d}_{ab}^{(\text{STEAC})} = \infty$. Clearly, this is undesirable, especially when the number of genes is large and the number of sites per gene K is low, when such events are most common. Hierarchical clustering methods such as UPGMA would not allow clusters containing leaves a and b to ever merge.

In this paper we consider the case when $K \gg 1$, when the Delta method is typically used to approximate the distribution of $\hat{g}_{ab}$ in Equation (5.3.2), which dictates that

$$\hat{g}^{(i)}_{\binom{L}{2}} \sim \mathcal{N}\left(g^{(i)}_{\binom{L}{2}}, \Sigma^{(i)}_{\text{STEAC}}\right),$$

where $\hat{g}^{(i)}_{\binom{L}{2}}$, $g^{(i)}_{\binom{L}{2}} \in \mathbb{R}^{\binom{n}{2}}$ are vectors containing the estimated and true Jukes-Cantor distances respectively, and the row ab, column cd element of the covariance matrix $\Sigma^{(i)}_{\text{STEAC}}$ is

$$\begin{aligned}\text{Cov}\left(\hat{g}^{(i)}_{ab}, \hat{g}^{(i)}_{cd} | G\right) &\approx \frac{\text{Cov}\left(\mathbb{1}\left(\chi_a^{(i)} \neq \chi_b^{(i)}\right), \mathbb{1}\left(\chi_c^{(i)} \neq \chi_d^{(i)}\right)\right)}{K\mu^2\left(\frac{3}{4} - p^{(i)}_{ab}\right)\left(\frac{3}{4} - p^{(i)}_{cd}\right)} \\ &= \frac{e^{2\mu\delta^{(i)}_{ab,cd}} + 2e^{\mu\delta^{(i)}_{ab,cd}} - 3}{3\mu^2 K},\end{aligned}$$

where the last equality follows from Proposition 5.3.1 and the definition of $p^{(i)}_{ab}$. In the following section, we make use of these derived covariance matrices to compute the total covariance matrix of the METAL and STEAC pair distances.

5.4 Theoretical results

In this section, we analyze both the *signal* and *noise* of distance-based inference under the MSC+Jukes–Cantor model. The signal is quantified by the *expected gap* in the *four-point condition* for quartets, which reflects the intrinsic separation of sister pairs in the species tree. Intuitively, the greater the expected gap, the easier it is to correctly reconstruct the quartet. The noise of METAL and STEAC estimators is captured by their total covariance which can be decomposed into coalescent and substitution components, thereby quantifying the relative contributions of coalescent/genealogical and substitution/mutational uncertainty. The spectral properties of these matrices are also analyzed.

5.4.1 Expected Gap

The following Proposition formalizes the expected four-point gap for a quartet of leaves. The four-point gap is a quantity used to differentiate quarter topologies based on pairwise distances. For a four leaf set $\{a, b, c, d\}$, it compares the sums of distances corresponding to alternative splits. Intuitively, if a, b form a sister pair (i.e. they share a common ancestor not shared with c or d, or equivalently $S_{ab} = \min_{p \in \binom{\{a,b,c,d\}}{2}} S_p$), then the sum $\hat{d}_{ad} + \hat{d}_{bc}$ is expected to be larger than $\hat{d}_{ab} + \hat{d}_{cd}$, whose difference is the four-point positive gap. The greater the gap is, the greater confidence we typically have about the quarter topology during inference (the probability of correctly estimating the topology is an increasing function of the four-point gap).

Proposition 5.4.1. *Let $\{a, b, c, d\}$ be four leaves with a and b forming a sister pair. Then the expected four-point gap is*

$$\mathbb{E}\left(\hat{d}_{ad} + \hat{d}_{bc} - \hat{d}_{ab} - \hat{d}_{cd}\right) = \frac{1}{1 + 2\mu}\left(e^{-\mu S_{ab}} + e^{-\mu S_{cd}} - e^{-\mu S_{ad}} - e^{-\mu S_{bc}}\right), \tag{5.4.1}$$

where $\hat{d}$ is the pairwise distances of the METAL estimator, and S_{xy} denotes the to-

tal branch lengths between leaves x and y in the underlying species tree. If the tree has cherry topology $((a,b),(c,d))$ with quartet diameter $\Delta_{\{a,b,c,d\}} = S_{ad} = S_{bc}$, equation (5.4.1) simplifies to

$$\mathbb{E}\Big(\hat{d}_{ad} + \hat{d}_{bc} - \hat{d}_{ab} - \hat{d}_{cd}\Big) = \left(e^{-\mu S_{ab}} + e^{-\mu S_{cd}} - 2e^{-\mu \Delta_{\{a,b,c,d\}}}\right)/(1+2\mu).$$

For a comb topology, w.l.o.g. taking the topology $(((a,b),c),d)$, the expression reduces to

$$\mathbb{E}\Big(\hat{d}_{ad} + \hat{d}_{bc} - \hat{d}_{ab} - \hat{d}_{cd}\Big) = \left(e^{-\mu S_{ab}} - e^{-\mu S_{ac}}\right)/(1+2\mu),$$

and the same formula applies if c and d are swapped.

5.4.2 Decomposition of Covariance

We begin by defining the total covariance of METAL distance estimates as

$$(\Sigma_{\text{total}})_{ab,cd} \coloneqq \text{Cov}\left(\hat{d}^{(\text{METAL})}(a,b), \hat{d}^{(\text{METAL})}(c,d)\right) = \frac{1}{m}\text{Cov}\left(\hat{p}^{(1)}_{ab}, \hat{p}^{(1)}_{cd}\right), \tag{5.4.2}$$

where $\Sigma_{\text{total}} \in \mathbb{R}^{\binom{n}{2}\times\binom{n}{2}}$. The following proposition shows how the total covariance can be decomposed as a coalescent and substitution covariance.

Proposition 5.4.2. *The total covariance can be expressed as*

$$\Sigma_{\text{total}} = \frac{1}{m}\Sigma_{\text{coal}} + \frac{1}{mK}\Sigma_{\text{sub}}, \tag{5.4.3}$$

where

$$(\Sigma_{\text{coal}})_{ab,cd} = \frac{9}{16}\text{Cov}\left(e^{-\mu g^{(1)}_{ab}}, e^{-\mu g^{(1)}_{cd}}\right), \tag{5.4.4}$$

$$(\Sigma_{\text{sub}})_{ab,cd} = \frac{3}{16}\mathbb{E}_{G^{(1)}|\mathcal{S}}\left(e^{-\mu\left(g^{(i)}_{ab}+g^{(i)}_{cd}\right)}\left(e^{2\mu\delta^{(i)}_{ab,cd}} + 2e^{\mu\delta^{(i)}_{ab,cd}} - 3\right)\right), \tag{5.4.5}$$

and $\Sigma_{\text{coal}}, \Sigma_{\text{sub}}, \Sigma_{\text{total}} \in \mathbb{R}^{\binom{n}{2}\times\binom{n}{2}}$ *are positive semi-definite matrices.*

The decomposition of Proposition 5.4.2 arises from conditioning on the latent gene trees and utilizing the law of total covariance: variation across loci due to random coalescent histories contributes the Σ_{coal} term (scaled by $1/m$), while conditional variation in the site patterns within each gene tree contributes the Σ_{sub} term (scaled by $1/(mK)$). Although both matrices are influenced by the coalescent and substitution processes, we use the terminology "coalescent covariance" and "substitutional covariance" to highlight which source of randomness dominates each component: Σ_{coal} reflects the variability across gene trees even with infinitely long sequences, while Σ_{sub} reflects the additional noise from having only finitely many sites per gene.

Expressions for Σ_{coal} defined in Equation (5.4.4) have been derived in the proof of Proposition 5.2.2. Specific expressions for Σ_{sub} defined in Equation (5.4.5), for all tree topology cases can be found in Appendix 5.B.

The decomposition in Proposition 5.4.2 sets the stage for Proposition 5.4.3, in which we specialize to the case of a star-shaped species tree and derive explicit formulae and asymptotics for each component of covariance. These results allow us both to compare the magnitude of coalescent versus substitution uncertainty and to study how their ratio varies with the mutation rate μ and tree diameter Δ. While the star-tree setting of Proposition 5.4.3 does not capture the full complexity of species-tree inference, it provides a tractable case that highlights the relative contributions of coalescent and substitutional variance.

Proposition 5.4.3. *Suppose that the species tree is a star tree with diameter* Δ *i.e.*

$\tau_{ab} = \Delta, \forall a, b \in L = [n], a \neq b$. *Then*

$$(\Sigma_{\text{mode}})_{ab,cd} = \begin{cases} \sigma^{(2)}_{\text{mode}}(\mu, \Delta), & \text{if } |\{a,b\} \cap \{c,d\}| = 2 \\ \sigma^{(1)}_{\text{mode}}(\mu, \Delta), & \text{if } |\{a,b\} \cap \{c,d\}| = 1 \\ \sigma^{(0)}_{\text{mode}}(\mu, \Delta), & \text{if } |\{a,b\} \cap \{c,d\}| = 0, \end{cases}$$

for $\text{mode} \in \{\text{coal}, \text{sub}, \text{total}\}$ *with*

$$\sigma^{(i)}_{\text{total}}(\mu, \Delta) = \frac{1}{m}\left(\sigma^{(i)}_{\text{coal}}(\mu, \Delta) + \frac{1}{K}\sigma^{(i)}_{\text{sub}}(\mu, \Delta)\right), \forall i \in \{0,1,2\},$$

where m *is the number of loci and* K *is the number of sites per loci. Moreover,* $\sigma^{(2)}_{\text{mode}}(\mu, \Delta) \geq \sigma^{(1)}_{\text{mode}}(\mu, \Delta) \geq \sigma^{(0)}_{\text{mode}}(\mu, \Delta) > 0$, *and*

$$\sigma^{(i)}_{\text{coal}}(\mu, \Delta) = \begin{cases} \mathcal{O}(\mu^2), & \text{as } \mu \to 0, \\ \mathcal{O}(\mu^{-3+i} e^{-2\mu\Delta}), & \text{as } \mu \to \infty \end{cases}$$

$$\sigma^{(i)}_{\text{sub}}(\mu, \Delta) = \begin{cases} \mathcal{O}(\mu), & \text{as } \mu \to 0, \\ \mathcal{O}(\mu^{-2+i} e^{(-2+i)\mu\Delta}), & \text{as } \mu \to \infty \end{cases}$$

for all $i \in \{0,1,2\}$ *and consequently*

$$\frac{\sigma^{(i)}_{\text{sub}}}{\sigma^{(i)}_{\text{coal}}} = \mathcal{O}(\mu^{-1}) \quad \text{as } \mu \to 0, \forall i \in \{0,1,2\}, \text{and}$$
$$\frac{\sigma^{(i)}_{\text{sub}}}{\sigma^{(i)}_{\text{coal}}} = \mathcal{O}(\mu e^{i\mu\Delta}) \quad \text{as } \mu \to \infty, \forall i \in \{0,1,2\}.$$

Proposition 5.2.1 shows that in the low-mutation regime ($\mu \ll 1$), STEAC and

METAL agree to first order, since $p_{ab} = \frac{3}{4}\big(1 - e^{-\mu g_{ab}}\big) \approx \frac{3}{4}\,\mu\, g_{ab}$. However, a fundamental drawback of STEAC is that, for any positive value of μ, the distribution of the estimate $\hat{g}_{ab}^{(i)}$ has strictly positive probability of being infinite, because $\hat{p}_{ab}^{(i)}$ can exceed $3/4$ with nonzero probability. Consequently, the variance of the STEAC estimator $\hat{g}_{ab}$ is infinite for all μ. This implies that with many loci the STEAC average is highly unstable: a single locus producing an infinite estimate suffices to drive the average to infinity. Although one can truncate or discard these infinite estimates, doing so introduces downward bias by systematically removing the highest inferred branch lengths. We conjecture that, in the presence of substitution uncertainty—i.e. when gene trees must be estimated rather than known a priori—STEAC is uniformly inferior to METAL for all μ; indeed, STEAC was originally developed under the assumption of almost perfect gene-tree knowledge. We assess this conjecture empirically in Section 5.5.1.

5.4.3 Principal components of variance

As we have seen in Proposition 5.4.3, the coalescent, substitution and total covariance matrices of the METAL estimator for the star species tree only depend on $|\{a,b\} \cap \{c,d\}|$. In this subsection we study the spectrum of these matrices. Specifically, for a covariance matrix $C \in \mathbb{R}^{\binom{n}{2}\times\binom{n}{2}}$, with entries

$$C_{ab,cd} = \operatorname{Cov}\left(\hat{d}_{ab}^{(\mathrm{METAL})}, \hat{d}_{cd}^{(\mathrm{METAL})}\right) = \begin{cases} \alpha, & \text{if } |\{a,b\} \cap \{c,d\}| = 2 \\ \beta, & \text{if } |\{a,b\} \cap \{c,d\}| = 1 \\ \gamma, & \text{if } |\{a,b\} \cap \{c,d\}| = 0, \end{cases} \tag{5.4.6}$$

for pairs $(a,b),(c,d) \in \binom{L}{2}$ where $\alpha > \beta > \gamma > 0$, Proposition 5.4.4 derives its spectrum.

Proposition 5.4.4. *The matrix $C \in \mathbb{R}^{\binom{n}{2}\times\binom{n}{2}}, n \geq 3$, defined in Equation (5.4.6), has*

eigenvalues

- *the dominant eigenvalue* $\gamma\binom{n}{2} + 2(\beta-\gamma)(n-1) + \gamma + \alpha - 2\beta$ *with multiplicity* 1 *and corresponding first principal component* $\mathrm{PC1} = \mathbf{1}$*,*
- $(\beta-\gamma)(n-2)+\gamma+\alpha-2\beta$ *with corresponding principal components* $\mathrm{PC2}, \ldots, \mathrm{PC}n$ *dependent on* n *and independent of* α, β, γ*, and*
- $\gamma + \alpha - 2\beta$ *with multiplicity* $n(n-3)/2$*.*

Note that these eigenvalues are presented in decreasing magnitude and that C *is positive definite if and only if* $\alpha + \gamma > 2\beta$*. It follows that the variance ratio of the principal component tends to* $\frac{\gamma}{\alpha}$ *as* $n \to \infty$*. The variance ratio of the first* n *most dominant eigenvectors tends to* $\frac{2\beta-\gamma}{\alpha}$ *as* $n \to \infty$*.*

Proposition 5.4.4 shows that each covariance matrix in our model has exactly three distinct eigenvalues. In particular, the all-ones vector $\mathbf{1}$ is an eigenvector corresponding to the largest eigenvalue. This direction may reflect the star tree structure, where all pairwise distances are the same. From the perspective of tropical geometry, $\mathbf{1}$ plays a canonical role: the tropical projective torus $\mathbb{R}^{\binom{n}{2}}/\mathbb{R}\mathbf{1}$ is precisely the quotient that identifies any vector $\mathbf{v}$ with its translate $\mathbf{v} + c\,\mathbf{1}$ for all real c. Equivalently, adding a constant to every coordinate does not change the point in the toric quotient. Statistical analysis on the tropical projective torus has been explored in various works (Aliatimis et al., 2024; Barnhill and Yoshida, 2023; Lee et al., 2022).

Corollary 5.4.5 (to Proposition 5.4.4) further quantifies that up to one-third of the total variance of our distance-vector estimates lies along the $\mathbf{1}$-direction. Since the tropical projective torus "modes out" precisely the $\mathbf{1}$-direction, restricting our inference to $\mathbb{R}^{\binom{n}{2}}/\mathbb{R}\mathbf{1}$ removes this single principal-component subspace. As a result, the remaining variance—now confined to the $(\binom{n}{2} - 1)$-dimensional tropical torus—drops by up to

one-third. Working in this quotient removes a large portion of extraneous variation while preserving all information relevant to tree topology, since adding any constant multiple of $\mathbf{1}$ to the distance vector does not alter the inferred tree topology.

Moreover, Corollary 5.4.5 shows that when $\mu \ll 1$, the relative magnitude of substitution versus coalescent uncertainty scales with μK, the expected number of substitutions per locus: as μK increases, substitution noise becomes negligible. In particular, if

$$K^{-1} \ll \mu \ll \frac{1}{2\Delta} \log\left(\frac{1+\sqrt{8K\Delta}}{2}\right),$$

then the coalescent covariance term dominates the total uncertainty. Note that as the number of bases per gene K increases, the interval widens on both ends, whereas increasing Δ narrows the interval from the upper bound. Figure 5.4.1 confirms this behavior: substitution-driven variance is largest for very small and very large μ, whereas for intermediate μ the coalescent variance is predominant. Figure 5.4.2 plots the ratio of Frobenius norms $\|\Sigma_{\text{sub}}\|_F / \|\Sigma_{\text{coal}}\|_F$, which attains its minimum at $\mu = \mathcal{O}(1/\Delta)$ when Δ is large. Finally, Figure 5.4.3 displays the cumulative variance explained by the leading n principal components of Σ_{sub}. Remarkably, for $\mu \ll 1$ and large Δ, these n components capture nearly all of the METAL estimator's variance, while for larger μ the explained fraction vanishes, in agreement with Proposition 5.4.6.

It is worth mentioning that Proposition 5.4.6 further establishes that as $\mu \to \infty$, the limiting coalescent correlation matrix, as defined in Equation (5.4.7), has rank $n-1$, and its nonzero principal components are the split-indicator vectors which are exactly the generators of the unique maximal cone of the ultrametric fan that contains the species-tree ultrametric. In other words, all variation of the METAL distance vector lies within the cone corresponding to the true species-tree ultrametric.

Corollary 5.4.5. *Consider the matrices* $\Sigma_{\text{coal}}(\mu), \Sigma_{\text{sub}}(\mu), \Sigma_{\text{total}}(\mu) \in \mathbb{R}^{\binom{n}{2}\times\binom{n}{2}}$ *defined in Equations* (5.4.2), (5.4.4), (5.4.5) *for a star species tree with diameter* Δ. *They have the following asymptotic properties*

$$\frac{\operatorname{Tr}(\Sigma_{\text{sub}})}{\operatorname{Tr}(\Sigma_{\text{coal}})}, \sqrt[\binom{n}{2}]{\frac{\det(\Sigma_{\text{sub}})}{\det(\Sigma_{\text{coal}})}}, \frac{\|\Sigma_{\text{sub}}\|_2}{\|\Sigma_{\text{coal}}\|_2}, \frac{\|\Sigma_{\text{sub}}\|_{\text{F}}}{\|\Sigma_{\text{coal}}\|_{\text{F}}} = \begin{cases} \mathcal{O}(\mu^{-1}K^{-1}), & \text{as } \mu \to 0, \\ \mathcal{O}(\mu e^{2\mu\Delta}K^{-1}), & \text{as } \mu \to \infty \end{cases}$$

As $\mu \to 0$ *and* $n \to \infty$, *the first principal component PC1=***1** *explains* $2/(6+3\Delta)$ *of the variance of* Σ_{sub} *and* Σ_{total}, *and* $2/9$ *of the variance of* Σ_{coal}, *while the first* n *principal components explain* $(4+3\Delta)/(6+3\Delta)$ *and* $4/9$ *of the variance respectively.*

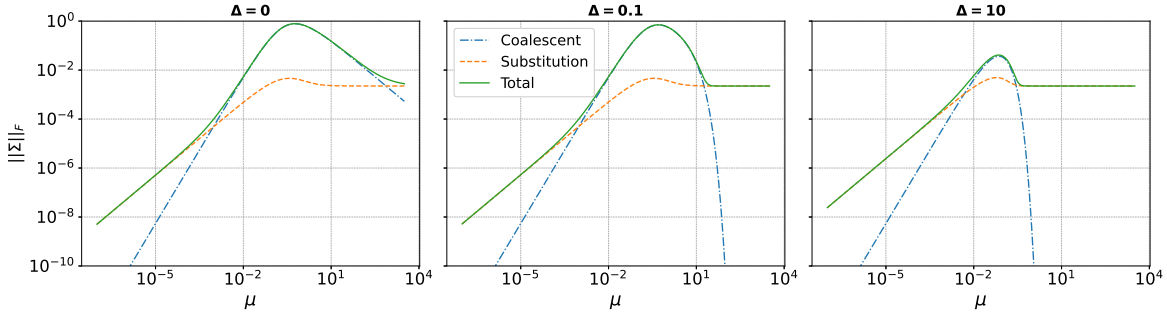


Figure 5.4.1: Frobenius norms of $\Sigma_{\text{coal}}(\mu)$, $\Sigma_{\text{sub}}(\mu)$, and $\Sigma_{\text{total}}(\mu)$ as functions of the mutation rate μ, shown for three values of the species-tree diameter Δ with $K = 1000$. For $\mu \ll 1$ and for $\mu \gg 1$, the substitution variance Σ_{sub} dominates; in the intermediate μ range, the coalescent variance Σ_{coal} prevails. As Δ increases, the interval in which Σ_{coal} dominates becomes narrower.

Proposition 5.4.6. *Consider a species tree* S *with positive pair distances* $\tau_{i,j} = \rho_{i,j}\Delta$, *where* Δ *is the diameter of the tree and* $\rho_{ij} \in (0, 1]$. *From the covariance matrix defined in Equation* (5.4.4), *if we construct the correlation matrix* $C_{\text{coal}}(\Delta) = D_{\text{coal}}^{-1}\Sigma_{\text{coal}}(\Delta)D_{\text{coal}}^{-1}$, *where* $D_{\text{coal}} = \sqrt{\operatorname{diag}(\Sigma_{\text{coal}}(\Delta))}$ *and consider the limiting case for large trees, then we let*

$$C_{\text{coal}}^{\infty} := \lim_{\Delta\to\infty} C_{\text{coal}}(\Delta). \tag{5.4.7}$$

Let $i \in V\backslash L$ *be an internal node of the species tree* S, *and the sets* $L_i, R_i \subset L = [n]$ *are*

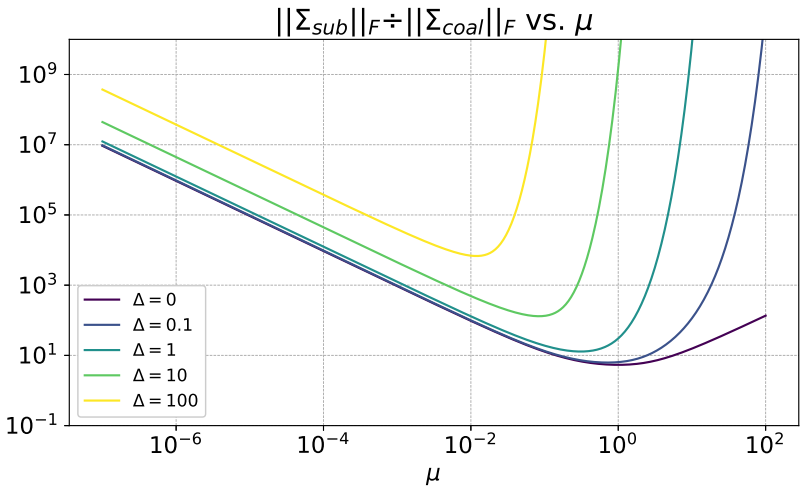


Figure 5.4.2: Ratio $\|\Sigma_{\text{sub}}(\mu)\|_F / \|\Sigma_{\text{coal}}(\mu)\|_F$ plotted against μ for several values of Δ. Substitution variance (Σ_{sub}) dominates when $\mu \ll 1$ or $\mu\Delta \gg 1$, whereas coalescent variance dominates in the intermediate μ regime.

the sets of leaf nodes left and right of i respectively. Define the vector,

$$\begin{aligned} v_{kl}^{(i)} &:= \mathbb{I}\left(i \text{ is the most recent common ancestor of } k \text{ and } l\right) \\ &= \mathbb{I}\left((k,l) \in (L_i \times R_i) \cup (R_i \times L_i)\right), \end{aligned}$$

which are also known as split-indicator vectors. Then, $C_{\text{coal}}^{\infty} v^{(i)} = |L_i||R_i| v^{(i)}$ i.e. $v^{(i)}$ is an eigenvector with corresponding eigenvalues $|L_i||R_i|$. These are exactly $n-1$ principal components corresponding to distinct internal nodes in $V \backslash L$. The remaining eigenvalues are zero, and so $\operatorname{rank}(C_{\text{coal}}^{\infty}) = n-1$. Finally, let the species tree pairwise distance vector $d^{\text{S}} \in \mathbb{R}^{\binom{n}{2} \times \binom{n}{2}}$, $d_{ab}^{\text{S}} = \tau_{ab}$, $\forall a, b \in L$ be an ultrametric i.e. $d \in \mathcal{U}_n$ and suppose it does not lie on topological boundaries. Then, there exists $\delta_i > 0$ such that for all $\epsilon_i \in (0, \delta_i)$,

$$d^{\text{S}} + \sum_{i=1}^{n-1} \epsilon_i v^{(i)} \in \mathcal{U}_n$$

Finally, $C_{\text{sub}}^{\infty} = C_{\text{total}}^{\infty} = I_{\binom{n}{2}}$, where the correlation matrices are defined as above, and $I_N \in \mathbb{R}^{N \times N}$ is the identity matrix.

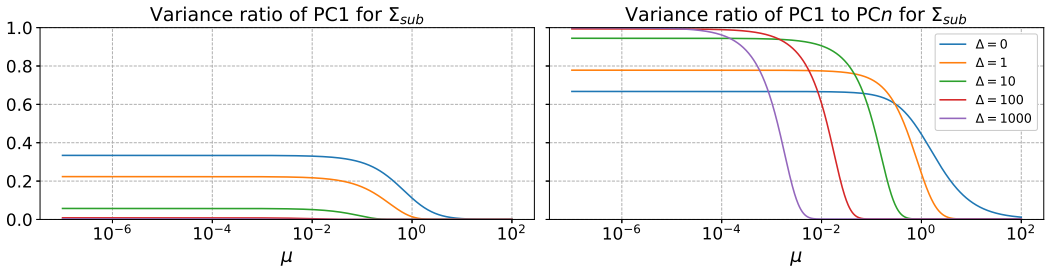


Figure 5.4.3: Variance ratios for Σ_{sub} plotted against mutation rate μ and species-tree diameter Δ. (Left) Fraction of total variance explained by the first principal component. (Right) Fraction of total variance explained by the first n principal components. For corresponding plots of Σ_{coal} and Σ_{total}, see Figure 5.D.1 in the Appendix.

5.5 Simulation Results

Two applications of the derived covariance matrix in phylogenetic analysis are considered in this section. In the first application (Section 5.5.1), our derivations of the covariance matrix enable us to quantify uncertainty from the substitution model, which predicts the relative performance of GLASS against METAL. In the second application (Section 5.5.2), the covariance matrix is used to compute node-support values as confidence for each bipartition of the estimated METAL tree. We compare our method to the standard bootstrap approach. Our repository can be found in Aliatimis (2024a).

5.5.1 Species-Tree Estimation

We begin by simulating data under the multi-species coalescent (MSC) combined with a site-substitution model. Using default parameters

$$\mu = 1, \quad \Delta = 1, \quad K = 100, \quad n = 10, \quad m = 100,$$

we first generate a species tree of diameter Δ (in coalescent units) via the Yule model. Then m gene trees are sampled under MSC on that species tree. Finally, sequences of length K are evolved along each gene tree under the Jukes-Cantor substitution model

with rate μ, producing m alignments.

Under these default conditions, the simulated alignments contain approximately 0.3% constant sites, corresponding to a relatively high mutation rate ($\mu = 1$). Although this rate is higher than what is typically observed in empirical datasets, we use it deliberately to keep the simulations tractable. Using a more realistic mutation rate (e.g., $\mu = 0.01$) would result in over 99% constant sites. In simulated data, compensating for this would require increasing the number of sites per genes K by two orders of magnitude, making simulations computationally impractical. Moreover, the additional constant sites would contribute no information to any inference method. Hence, we adopt the higher mutation rate solely to ensure that the simulated datasets contain a reasonable proportion of informative sites, and not because the method requires unrealistically high mutation rates in practice.

From these m alignments, we infer the species tree using three approaches (see Algorithm 4): STEAC, METAL, and GLASS. In STEAC and GLASS, each gene's pairwise Hamming distances $\hat{p}_{ab}^{(i)}$ are transformed via the inverse Jukes–Cantor formula to branch-length estimates $\hat{g}_{ab}^{(i)}$. STEAC then averages across all m genes before applying UPGMA. GLASS takes the minimum branch gene tree branch lengths before applying UPGMA. METAL concatenates all m alignments, computes Hamming distances on the full superalignment, and reconstructs a UPGMA tree from that distance matrix.

Figure 5.5.1 illustrates how the relative accuracy of GLASS versus METAL depends on the fraction of variance contributed by the substitution model. As this ratio increases, METAL (which explicitly pools substitution noise across genes) consistently outperforms GLASS. To quantify reconstruction accuracy, each inferred tree $\widehat{T}_{\text{method}}$ is

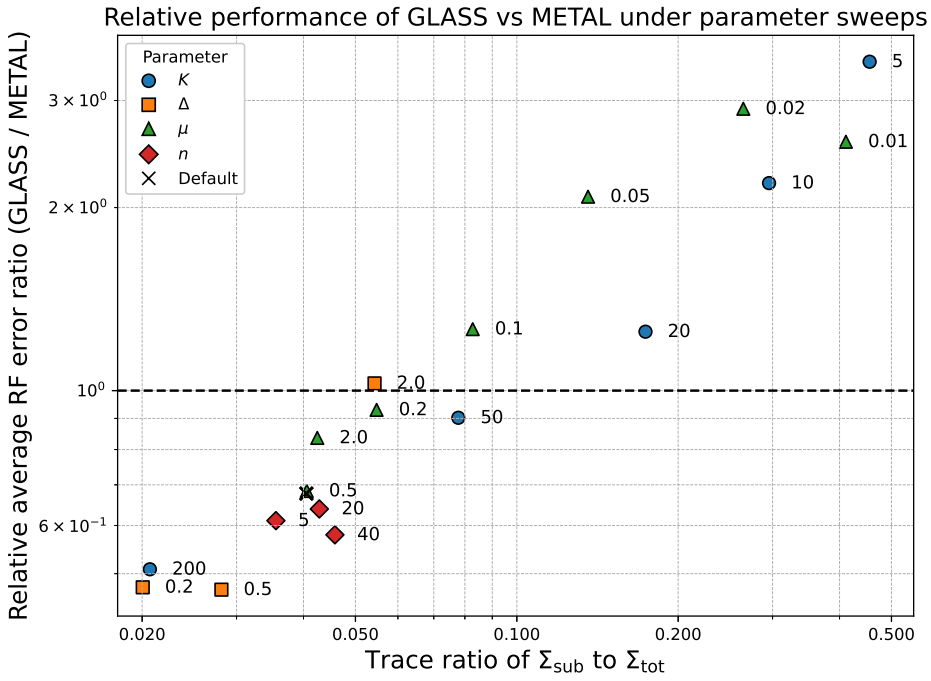


Figure 5.5.1: Comparison of GLASS and METAL performance as a function of the ratio $\mathrm{tr}(\Sigma_{\mathrm{sub}})/\mathrm{tr}(\Sigma_{\mathrm{total}})$. When substitution-model variance constitutes a large fraction of total variance, METAL outperforms GLASS.

compared against the true species tree T_{true} via the Robinson–Foulds (RF) distance. Over R replicates (here $R = 100$), we obtain RF distances $\mathrm{RF}^{(r)}_{\mathrm{STEAC}}$, $\mathrm{RF}^{(r)}_{\mathrm{METAL}}$, $\mathrm{RF}^{(r)}_{\mathrm{GLASS}}$. The mean RF error for each method is

$$\overline{\mathrm{RF}}_{\mathrm{method}} \;=\; \frac{1}{R}\sum_{r=1}^{R} \mathrm{RF}^{(r)}_{\mathrm{method}}.$$

We consistently observe

$$\overline{\mathrm{RF}}_{\mathrm{STEAC}} \;>\; \overline{\mathrm{RF}}_{\mathrm{METAL}},$$

so STEAC underperforms METAL in all tested regimes (see Figure 5.D.2 in the Appendix). In contrast,

$$\frac{\overline{\mathrm{RF}}_{\mathrm{GLASS}}}{\overline{\mathrm{RF}}_{\mathrm{METAL}}}$$

increases with $\mathrm{tr}(\Sigma_{\mathrm{sub}})/\mathrm{tr}(\Sigma_{\mathrm{total}})$. When substitution variance dominates, GLASS yields higher RF error than METAL. However, if coalescent variance from MSC is the primary source (for instance, when K is large or μ lies in an intermediate regime), GLASS approaches the Kullback–Leibler-optimal species-tree estimator and outperforms METAL, i.e.

$$\overline{\mathrm{RF}}_{\mathrm{GLASS}} \;<\; \overline{\mathrm{RF}}_{\mathrm{METAL}}.$$

In summary, the ratio

$$\frac{\mathrm{tr}(\Sigma_{\mathrm{sub}})}{\mathrm{tr}(\Sigma_{\mathrm{total}})}$$

determines the preferable method: if this ratio exceeds a threshold (when μ is very low or very high, Δ is large, or K is small), METAL—by aggregating substitution noise—yields lower RF error. Otherwise, for moderate μ and large K, the coalescent term dominates and GLASS (leveraging accurate gene-tree estimates) gives superior accuracy.

5.5.2 Bipartition Confidence

The procedure begins with data generation: we first simulate a species tree, then simulate gene trees under the multi-species coalescent (MSC) model using `DendroPy` (Moreno et al., 2024), and finally perform sequence evolution along those gene trees using `Pyvolve` (Spielman and Wilke, 2015). This pipeline is illustrated in Fig. 5.1.1.

Once we have simulated data, we obtain the METAL estimate by computing Hamming distances between the concatenated sequences. These distances serve as pairwise dissimilarities, and we reconstruct a tree via the UPGMA algorithm, following the METAL methodology.

Our main objective is to assign confidence scores to tree splits (bipartitions). The

standard approach uses bootstrapping: sites from the concatenated alignment are resampled (with replacement) to generate pseudo-replicate alignments. For each replicate, we compute a new METAL tree, yielding a collection of bootstrap trees. The proportion of replicates in which a given split appears provides its bootstrap support.

As an alternative, we propose a Gaussian-sampling approach. The METAL estimator of the average pairwise distances can, for a sufficiently large number of genes, be approximated by a multivariate normal distribution. The mean of this distribution is the METAL estimate itself, and the covariance is derived from the METAL tree (in coalescent units to scale Hamming distances appropriately). The quality of this normal approximation improves as the number of genes increases.

We draw samples from this multivariate distribution to obtain synthetic vectors of pairwise distances. Each sampled distance matrix is then converted into a tree via UPGMA, exactly as in the bootstrap procedure. We estimate split support by computing the frequency of each bipartition across these "Gaussian-sampled" trees.

Figure 5.5.2 provides an illustrative comparison. The leftmost tree is the METAL estimate with bootstrap-derived support values, the middle tree shows the same METAL topology but with support values from Gaussian sampling, and the rightmost tree is the true species topology. Note that Gaussian sampling typically yields more conservative (lower) support values than bootstrapping.

To compare the two confidence-estimation methods quantitatively, we exploit the fact that the true species tree is known in simulation. With $n = 40$ taxa, there are $n - 2 = 38$ non-trivial splits. We label each split in the METAL estimate as correct (1) if it appears in the true tree, or incorrect (0) otherwise. Using the support scores

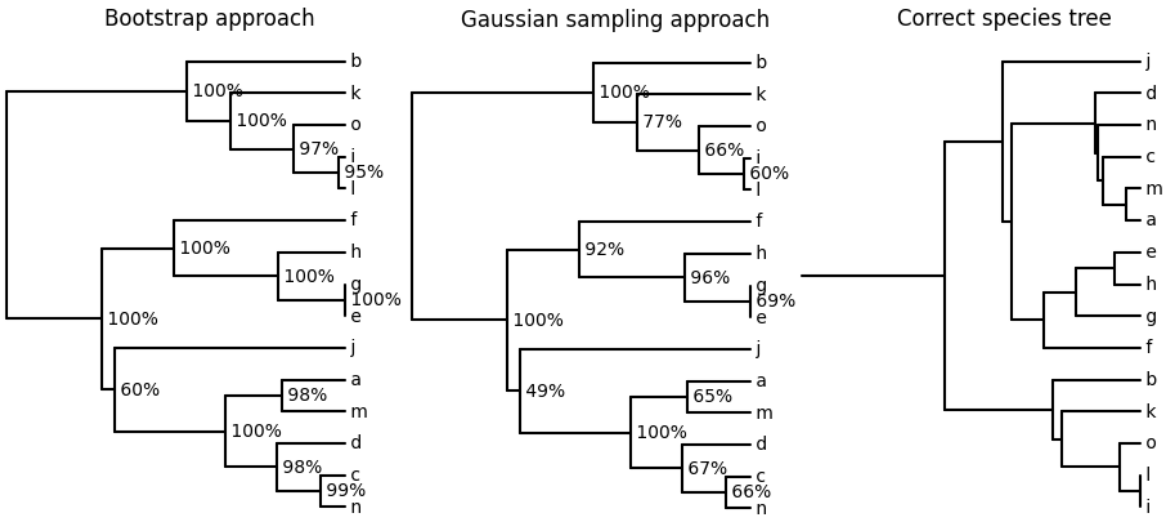


Figure 5.5.2: Example of METAL estimate (left) with bootstrap support (percentages at each node), METAL estimate (center) with Gaussian-sampling support, and the true species tree (right). Bootstrap supports tend to be higher than Gaussian supports, but may also overestimate confidence for incorrect splits.

assigned by each method, we compute the area under the ROC curve (AUC) to measure how well each set of support values discriminates correct splits from incorrect ones.

This experiment is repeated 100 times. In each series of 100 replicates, we vary exactly one simulation parameter (e.g., mutation rate μ) while holding all other parameters (species-tree diameter Δ, number of sites per gene K, and number of genes m, number of taxa n) fixed. We then repeat separately for each of the other parameters (Δ, K, m, and n), always keeping the remaining factors constant.

Figure 5.5.3 shows boxplots of AUC scores for three methods: standard bootstrapping (blue), Gaussian sampling (green), and multilocus bootstrapping (magenta). In the left panel, we vary the number of genes m; in the right panel, we vary the number of sites per gene K. In both cases, Gaussian-sampling and multilocus bootstrapping yield higher median AUC and lower variance than standard bootstrapping. Similar figures for the other parameters can be found in Appendix 5.D. What is interesting about param-

eters m and K is that the concatenated sequence length has mK bases. As the number of genes m or the number of sites per gene K increases, the runtime of standard and multilocus bootstrap procedures grows roughly in proportion to $m \cdot K$, the size of the concatenated sequence. In contrast, the computational cost of the Gaussian-sampling approach is essentially unaffected by changes in m or K, since it depends only on the number of taxa n. Moreover, although standard bootstrapping requires more time as m and K increase, its ability to correctly classify true splits does not improve—in fact, its classification accuracy becomes more erratic.

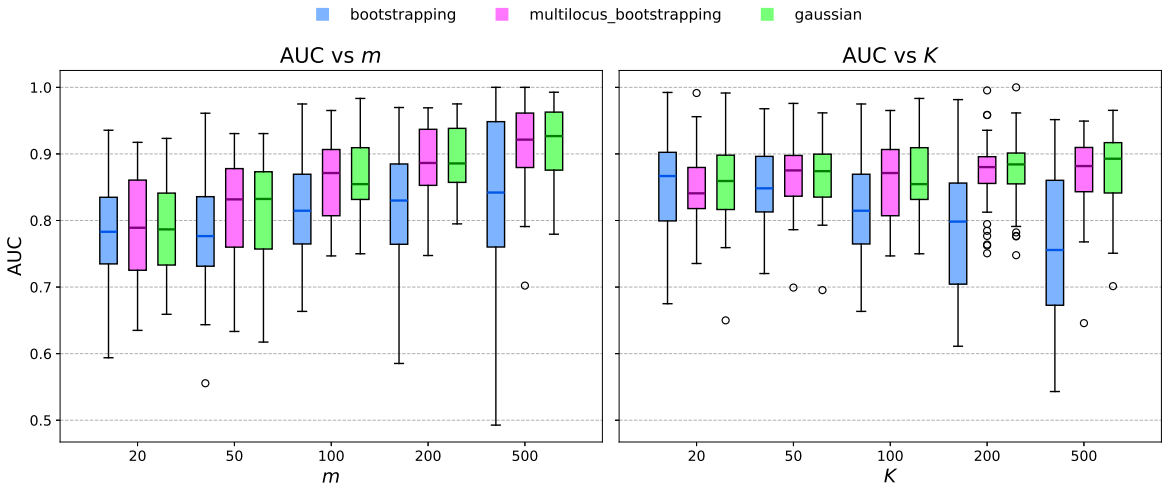


Figure 5.5.3: Boxplots of AUC values comparing standard bootstrap (blue), Gaussian sampling (green), and multilocus bootstrap (magenta) support scores. (Left:) AUC distributions as the number of genes m varies (with μ, Δ, n, and K held constant). (Right:) AUC distributions as the number of sites per gene K varies (with μ, Δ, n, and m held constant). Gaussian sampling and multilocus bootstrapping yield similar median AUC and lower variability than standard bootstrapping, while Gaussian sampling retains a clear computational advantage.

5.5.2.1 Computational Complexity Comparison

For bootstrapping, each replicate requires recomputing the METAL tree from a resampled alignment of total length $m \cdot K$. Computing all pairwise Hamming distances is $\mathcal{O}(n^2 \cdot m \cdot K)$, since there are $\binom{n}{2} = \mathcal{O}(n^2)$ taxon pairs. To ensure METAL is accurate

with probability $1 - \varepsilon$, one needs

$$m \;=\; O\big(\tfrac{\log(n/\varepsilon)}{\tau^2}\big),$$

where τ is the shortest internal branch length. Since $\tau \leq \Delta/(n-1)$, one obtains $m = \Omega(n^2 \log n)$. Hence each bootstrap replicate costs $\mathcal{O}(n^4 \log n \cdot K)$, and obtaining B replicates is B times that.

In contrast, the Gaussian sampling approach requires computing a covariance matrix over

$$N = \binom{n}{2} \;=\; \mathcal{O}(n^2)$$

entries, resulting in a full covariance matrix of size $N \times N$, containing $\mathcal{O}(n^4)$ elements. Computing the Cholesky decomposition of this matrix takes $\mathcal{O}(N^3) = \mathcal{O}(n^6)$ time, but only once. Each subsequent multivariate-normal sample (to produce one synthetic distance vector) costs $\mathcal{O}(N^2) = \mathcal{O}(n^4)$. Crucially, this Gaussian method is independent of the sequence length K, making it far more efficient when K is large.

Nonetheless, when the number of taxa is very large ($n \gg 1$), the one-time cost of the Cholesky decomposition becomes prohibitive, and bootstrapping may be more efficient. However, if the mutation rate is low $\mu \ll 1$ or if $\mu\Delta \gg 1$, we can instead leverage the theoretical results of Section 5.4, specifically Corollary 5.4.5 and Proposition 5.4.6. These results describe the asymptotic spectrum of the total covariance matrix of the METAL estimator. In both of these limiting cases, we have shown in Proposition 5.4.3 that $\Sigma_{\text{coal}} \ll \Sigma_{\text{sub}}$, and so that $\Sigma_{\text{total}} \approx \Sigma_{\text{sub}}$. For $\mu \ll 1$, there are three distinct eigenvalues with known variance ratios. As illustrated in Figure 5.5.4, the asymptotic predictions of the eigenvalues of Corollary 5.4.5 closely match the empirical ones of a randomly generated species tree with $\Delta = 1, n = 30, \mu = 0.2$. Knowing the principal components in advance allows us to bypass the Cholesky decomposition entirely,

eliminating the most computationally expensive step of the method.

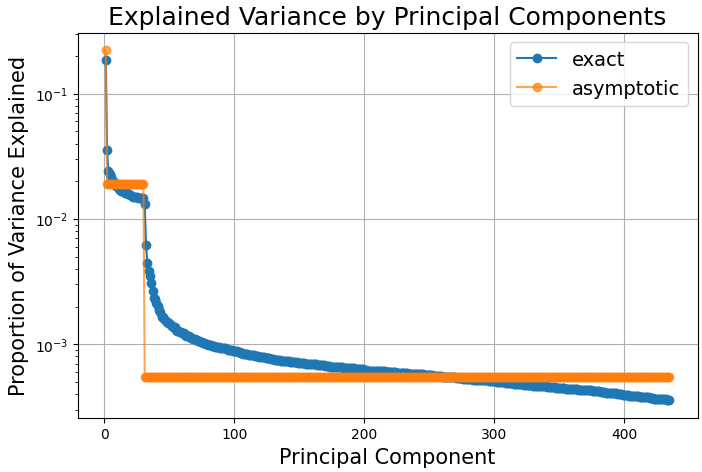


Figure 5.5.4: Explained variance by principal components of the total covariance matrix for a species tree with $\Delta = 1, n = 30, \mu = 0.2, K = 100$. The exact curve (blue) shows the empirical proportion of variance explained by each principal component, while the "asymptotic" curve (orange) represents theoretical predictions from Corollary 5.4.5. The close agreement validates the asymptotic spectrum in the low mutation rate regime ($\mu \ll 1$).

5.6 Discussion

In this work, we explicitly formulate the covariance matrix of all pairwise distances under the MSC and substitution models for the species tree reconstruction problem. METAL constructs a distance matrix from concatenated multi-locus data Dasarathy et al. (2014), and then applies a clustering algorithm to infer the species tree. Our first main result is the derivation of the exact covariance matrix of METAL's pairwise distance estimates. This covariance naturally decomposes into a term arising from coalescent genealogy variance and a term from sequence-substitution variance. We give explicit formulas for both components. We then analyze the asymptotic regime as the number of loci grows large; we show that METAL distances concetrate around expectations, and we study the spectrum of the limiting covariance matrix. The spectral analysis suggests that much of the phylogenetic signal lies in the first few principal

components of the distance matrix when the number of leaves is large.

Our simulation study illustrates how the relative performance of distance-based methods depends on the underlying source of variation. In particular, METAL tends to outperform GLASS when the substitutional noise dominates, while GLASS becomes more accurate as coalescence variance takes over. This sensitivity reflects the design of each estimator; METAL effectively aggregates substitutional uncertainty across loci, whereas GLASS leverages gene-tree precision when it is available. Likewise, our results show METAL retaining strong performance even under rate variation and short branches. Finally, we introduce a new Gaussian-sampling procedure to estimate METAL split-support values, which uses the derived covariance to quantify support for inferred clades.

There are several natural extensions of our work. Our current analysis assumed a strict molecular clock, so that all lineages evolve at the same rate and that the gene and species trees are equidistant. Relaxing this assumption is important for real data. Generalizing our covariance derivation to non-clock settings would be a major next step of our current analysis. Moreover, although we derived the expected four-point gap of a quartet (i.e., the signal) in Section 5.4.1, we have not yet considered the corresponding signal-to-noise ratio. Quantifying this ratio could provide insight into the likelihood of correctly reconstructing the species tree, as a higher signal relative to noise should generally increase reconstruction accuracy.

In conclusion, we have offered a systematic decomposition of the covariance structure in distance-based methods like METAL and STEAC, separating the contributions of gene-tree (coalescent) variability from sequence-level (substitutional) noise. Our framework not only illuminates how these two sources of uncertainty compare to each other, but also lays the groundwork for extending this analysis to a broader class of

estimators. An important direction for future research is to determine the extent to which our findings on the balance between coalescent and substitutional effects hold under alternative models and estimators. Addressing this question will deepen our understanding of estimator performance across diverse biological scenarios.

5.A Proofs

Proof of Proposition 5.2.1. The definitions of the dissimilarity metrics can be found in Equations (5.2.1)–(5.2.2). We start by considering low mutation rates. Since,

$$\lim_{\mu\to 0} \frac{3}{4} \cdot \frac{1 - \exp\left(-\frac{4}{3}\mu g_{ab}^{(i)}\right)}{\mu} = g_{ab}^{(i)}, \ \forall a, b \in L$$

and by expressing Equation (5.2.2) as the partial sum of terms $1 - \exp\left(-\frac{4}{3}\mu g_{ab}^{(i)}\right)$, it follows that

$$\lim_{\mu\to 0} \frac{d^{(\text{METAL})}(a,b)}{\mu} = \frac{1}{m}\sum_{i=1}^{m} g_{ab}^{(i)} = d^{(\text{STEAC})}(a,b), \ \forall a, b \in L.$$

Therefore, in the limit $\mu \to 0$, the pairwise METAL distances become proportional to the STEAC distances. Since hierarchical clustering dendrograms are invariant under scalar multiplication of the distance matrix,

$$\lim_{\mu\to 0} \hat{T}_{\text{METAL}}(\mu) = \hat{T}_{\text{STEAC}}.$$

For high mutation rates, note that if the GLASS minimum coalescence times are unique, i.e. for any given pair of leaves $a, b \in L, a \neq b$ there exists $i^* \in [m]$ such that $\min_{j\in[m]} g_{ab}^{(j)} = g_{ab}^{(i^*)} < g_{ab}^{(i)}$ for all $i \in [m]\backslash\{i^*\}$, then

$$\lim_{\mu\to\infty} m\frac{1 - \frac{4}{3}d^{(\text{METAL})}(a,b)}{\exp\left(-\frac{4}{3}\mu d^{(\text{GLASS})}(a,b)\right)} = \lim_{\mu\to\infty}\sum_{i=1}^{m} \exp\left(-\frac{4}{3}\mu\left(g_{ab}^{(i)} - g_{ab}^{(i^*)}\right)\right) = 1.$$

In other words,

$$d^{(\text{METAL})}(a,b) = \frac{3}{4}\left(1 - \left(\frac{1}{m} + o(1)\right)\exp\left(-\frac{4}{3}\mu d^{(\text{GLASS})}(a,b)\right)\right).$$

Note that $f(x) = \frac{3}{4}\left(1 - a\exp\left(-\frac{4}{3}\mu x\right)\right)$ is a strictly increasing function. Hence, for a sufficiently large μ,

$$d^{(\text{GLASS})}(a,b) < d^{(\text{GLASS})}(c,d) \Rightarrow d^{(\text{METAL})}(a,b) < d^{(\text{METAL})}(c,d).$$

Since the number of combinations $a, b, c, d \in L$ is finite, there exists a sufficiently large μ such that the above equation holds for all $a, b, c, d \in L$. Hence, the asymptotic ordering of METAL distances agrees with the order of GLASS distances. Another way of reaching the same conclusion is using the monotone admissibility of hierarchical clustering methods such as single linkage or complete linkage Dugad and Ahuja (1998), which allows monotonic transformation of the elements of the distance matrix without altering the clustering. □

Proof of Proposition 5.2.2. Instead of writing $g_{ab}^{(1)}$ to refer to locus 1 specifically, we write g_{ab} for the remainder of the proof. First, we address the trivial case where $a = b$ or $c = d$. Assume that wlog $a = b$, then $g_{ab} = 0$ and $e^{tg_{ab}} = 1$, which implies that

$$\text{Cov}(e^{tg_{aa}}) = \text{Cov}(g_{aa}) = 0,$$

which implies that the corresponding correlations will be zero too, validating the results immediately in this case.

For the rest of the proof, assume that $a \neq b$ and $c \neq d$.

There are four cases to consider regarding the choice of the two pairs.

- **Two leaves only**, where wlog $a = c$ and $b = d$.

- **Three leaves only**, where wlog $d = a$ and $S_{ab} \leq S_{ac} = S_{bc}$.
- **Four leaves forming a cherry tree**, where $S_{ab} \leq S_{cd} \leq S_{ad} = S_{bd} = S_{ac} = S_{bc}$.
- **Four leaves forming a comb tree**, where $S_{ab} \leq S_{ac} = S_{bc} \leq S_{ad} = S_{bd} = S_{cd}$.

These are the only cases to be considered; if a, b, c, d do not satisfy them, they can always be permuted so that at least one of the conditions above are satisfied.

Fig. 5.A.1 illustrates those four cases.

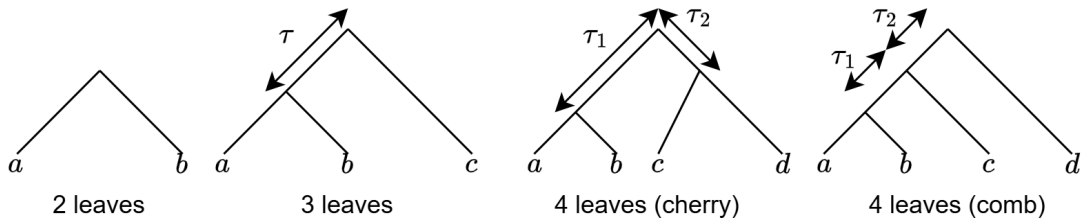


Figure 5.A.1: The four possible shapes of tree containing $a, b, c, d \in L$ with $a \neq b, c \neq d$.

Two leaf tree For the case where $(a, b) = (c, d)$, the covariance expressions becomes variances, which are positive. Under the MSC $g_{ab} = S_{ab} + 2E_1$, where $E_1 \sim \text{Exp}(1)$, and so

$$\text{Cov}(g_{ab}, g_{ab}) = \text{Var}\,(g_{ab}) = 4, \tag{5.A.1}$$

$$\begin{aligned}\text{Cov}(e^{tg_{ab}}, e^{tg_{ab}}) = \text{Var}\,(e^{tg_{ab}}) &= \mathbb{E}\left(e^{2tg_{ab}}\right) - \left(\mathbb{E}(e^{tg_{ab}})\right)^2 \\ &= e^{2tS_{ab}}\mathbb{E}\left(e^{2tE_1}\right) - e^{2tS_{ab}}\mathbb{E}\left(e^{tE_1}\right)^2 \\ &= \frac{4t^2 e^{2tS_{ab}}}{(1-4t)(1-2t)^2},\end{aligned} \tag{5.A.2}$$

where for the last result we use that the moment generating function (MGF) of Exp(1) is $M(t) = \mathbb{E}\left(e^{tE_1}\right) = 1/(1-t)$ and $\mathbb{E}\left(e^{2tE_1}\right) = 1/(1-2t)$.

For the remaining three cases, we consider the variable

$$E_{vw,xy} = \frac{(g_{vw} + g_{xy} - S_{vw} - S_{xy})}{2},$$

where $v, w, x, y \in L$, and its moment-generating function (MGF)

$$M_{vw,xy}(t) = \mathbb{E}\left(\exp\left(tE_{vw,xy}\right)\right).$$

The moment-generating function allows access to the covariances of interest, as the two auxiliary results below (Equations (5.A.3),(5.A.4)) show. We now show these results.

For each pair (i, j), define

$$Y_{ij} := \frac{g_{ij} - S_{ij}}{2} \sim \text{Exp}(1),$$

with $\mathbb{E}[Y_{ij}] = 1$, $\text{Var}(Y_{ij}) = 1$ and hence $\mathbb{E}[Y_{ij}^2] = 2$. Writing

$$E_{vw,xy} = Y_{vw} + Y_{xy},$$

we have

$$\frac{\partial^2 M_{vw,xy}}{\partial t^2}(0) = \mathbb{E}\big[(Y_{vw} + Y_{xy})^2\big] = \mathbb{E}[Y_{vw}^2] + \mathbb{E}[Y_{xy}^2] + 2\,\mathbb{E}[Y_{vw}Y_{xy}].$$

But,

$$\mathbb{E}[Y_{vw}Y_{xy}] = \text{Cov}(Y_{vw}, Y_{xy}) + \mathbb{E}[Y_{vw}]\,\mathbb{E}[Y_{xy}] = \text{Cov}(Y_{vw}, Y_{xy}) + 1.$$

Therefore

$$\frac{\partial^2 M_{vw,xy}}{\partial t^2}(0) = 2 + 2 + 2\big(\text{Cov}(Y_{vw}, Y_{xy}) + 1\big) = 6 + 2\,\text{Cov}(Y_{vw}, Y_{xy}).$$

Noting that $\mathrm{Cov}(Y_{vw}, Y_{xy}) = \frac{1}{4}\mathrm{Cov}(g_{vw}, g_{xy})$ recovers

$$\frac{\partial^2 M_{vw,xy}}{\partial t^2}(0) = 6 + \tfrac{1}{2}\mathrm{Cov}\big(g_{vw}, g_{xy}\big). \tag{5.A.3}$$

Similarly, one shows that

$$\mathrm{Cov}\big(e^{tg_{vw}}, e^{tg_{xy}}\big) = \Big(M_{vw,xy}(2t) - \tfrac{1}{(1-2t)^2}\Big)\, e^{t(S_{vw}+S_{xy})}. \tag{5.A.4}$$

Indeed, since $g_{ij} = S_{ij} + 2Y_{ij}$ with $Y_{ij} \sim \mathrm{Exp}(1)$,

$$\mathbb{E}[e^{tg_{vw}} e^{tg_{xy}}] = e^{t(S_{vw}+S_{xy})}\,\mathbb{E}[e^{2t(Y_{vw}+Y_{xy})}] = e^{t(S_{vw}+S_{xy})}\,M_{vw,xy}(2t),$$

and

$$\mathbb{E}[e^{tg_{ij}}] = e^{tS_{ij}}\,(1-2t)^{-1},$$

where $1/(1-2t)^2$ is the MGF of $2\,\mathrm{Exp}(1)$, so $\mathbb{E}[e^{tg_{vw}}]\,\mathbb{E}[e^{tg_{xy}}] = e^{t(S_{vw}+S_{xy})}(1-2t)^{-2}$. Subtracting gives (5.A.4).

Three leaf tree Consider the case of a three-leaf tree where $S_{ab} \le S_{ac} = S_{bc}$, where $\tau = \frac{S_{ac}-S_{ab}}{2}$ as shown in the second leftmost tree in Fig. 5.A.1. Define the joint random variable $(T, \mathbf{p})$, where T is the first time that a pair coalesces and $\mathbf{p}$ is that pair. In other words,

$$T = \min_{x,y\in\binom{\{a,b,c\}}{2}} \{g_{xy}\},$$

$$\mathbf{p} = \operatorname*{arg\,min}_{(x,y)\in\binom{\{a,b,c\}}{2}} \{g_{xy}\}.$$

Since the first speciation event happens at time $S_{ab}/2$, we require that $T \ge S_{ab}/2$. If

$T \in (S_{ab}/2, S_{ac}/2)$, then $\mathbf{p} = (a, b)$. The joint distribution function of $(T, \mathbf{p})$ is

$$f_{T,\mathbf{p}}(T, \mathbf{p}) = \begin{cases} \exp\left(-\left(T - \frac{S_{ab}}{2}\right)\right), & \text{if } T \in \left(\frac{S_{ab}}{2}, \frac{S_{ac}}{2}\right), \mathbf{p} = (a, b) \\ \exp(-\tau)\exp\left(-3\left(T - \frac{S_{ac}}{2}\right)\right), & \text{if } T \geq \frac{S_{ac}}{2}, \forall \mathbf{p} \in \binom{\{a,b,c\}}{2}, \\ 0, & \text{otherwise,} \end{cases}$$

where for the remainder of this proof $\binom{\mathcal{A}}{2}$ is the set containing pairs of distinct elements of the set $\mathcal{A}$.

Note that $T - \frac{S_{ac}}{2} | T > \frac{S_{ac}}{2} \sim \text{Exp}(3)$, but since there are three choices of pairs, the pdf of Exp(3) is divided by 3 for each pair in the joint pdf above. Moreover, the conditioning probability $\mathbb{P}(T > \frac{S_{ac}}{2}) = \exp(-\tau)$ is included.

Conditional on $(T, \mathbf{p})$,

$$E_{ab,ac} | (T, \mathbf{p}) \stackrel{d}{=} \begin{cases} \left(T - \frac{S_{ab}}{2}\right) + \text{Exp}(1), & \text{if } T \in \left(\frac{S_{ab}}{2}, \frac{S_{ac}}{2}\right), \mathbf{p} = (a, b) \\ 2\left(T - \frac{S_{ac}}{2}\right) + \tau + 2\,\text{Exp}(1), & \text{if } T \geq \frac{S_{ac}}{2}, \mathbf{p} = (b, c) \\ 2\left(T - \frac{S_{ac}}{2}\right) + \tau + \text{Exp}(1), & \text{if } T \geq \frac{S_{ac}}{2}, \mathbf{p} = (a, b) \text{ or } (a, c) \end{cases}$$

and so

$$\mathbb{E}\left(e^{tE_{ab,ac}} | (T, \mathbf{p})\right) = \begin{cases} \frac{\exp\left(t\left(T - \frac{S_{ab}}{2}\right)\right)}{1-t}, & \text{if } T \in \left(\frac{S_{ab}}{2}, \frac{S_{ac}}{2}\right), \mathbf{p} = (a, b) \\ \frac{\exp\left(2t\left(T - \frac{S_{ac}}{2}\right) + t\tau\right)}{1-2t}, & \text{if } T \geq \frac{S_{ac}}{2}, \mathbf{p} = (b, c) \\ \frac{\exp\left(2t\left(T - \frac{S_{ac}}{2}\right) + t\tau\right)}{1-t}, & \text{if } T \geq \frac{S_{ac}}{2}, \mathbf{p} = (a, b) \text{ or } (a, c) \end{cases}$$

Finally, combining those two results gives us the MGF,

$$\begin{aligned}
M_{ab,ac}(t) &= \mathbb{E}_T\left(\mathbb{E}(\exp(tE_{ab,ac})|T\right) \\
&= \int_{\frac{S_{ab}}{2}}^{\infty} \sum_{\mathbf{p}\in\{(a,b),(b,c),(a,c)\}} f_{T,\mathbf{p}}(T,\mathbf{p})\mathbb{E}\left(e^{tE_{ab,ac}|(T,\mathbf{p})}\right)\, dT \\
&= \int_{S_{ab}/2}^{S_{ac}/2} e^{-\left(T-\frac{S_{ab}}{2}\right)} \frac{e^{t\left(T-\frac{S_{ab}}{2}\right)}}{1-t}\, dT \\
&\quad + \int_{S_{ac}/2}^{+\infty} e^{-\tau} e^{\left(-3\left(T-\frac{S_{ac}}{2}\right)\right)} \frac{e^{\left(2t\left(T-\frac{S_{ac}}{2}\right)+t\tau\right)}}{1-2t}\, dT \qquad (\text{for } \mathbf{p}=(b,c)) \\
&\quad + \int_{S_{ac}/2}^{+\infty} 2e^{-\tau} e^{\left(-3\left(T-\frac{S_{ac}}{2}\right)\right)} \frac{e^{\left(2t\left(T-\frac{S_{ac}}{2}\right)+t\tau\right)}}{1-t}\, dT \qquad (\text{for } \mathbf{p}\neq(b,c)) \\
&= \frac{1}{(1-t)^2} + e^{(t-1)\tau}\frac{t^2}{(1-t)^2(1-2t)(3-2t)}.
\end{aligned}$$

Taking the second derivative of this expression with respect to t at $t=0$ yields

$$M''_{ab,cd}(0) = 6 + \frac{2}{3}e^{-\tau},$$

and so, using equations (5.A.3), and (5.A.4), it follows that

$$\begin{aligned}
\mathrm{Cov}(g_{ab}, g_{ac}) &= \mathrm{Cov}(g_{ab}, g_{bc}) = \frac{4}{3}e^{-\tau} \\
\mathrm{Cov}(e^{tg_{ab}}, e^{tg_{ac}}) &= \mathrm{Cov}(e^{tg_{ab}}, e^{tg_{bc}}) = e^{-\tau+2tS_{ac}}\frac{4t^2}{(1-2t)^2(1-4t)(3-4t)}
\end{aligned}$$

which are all positive. Note that the reason these results hold for the pairs (a,b) and (b,c) is that $M_{ab,bc}(t) = M_{ab,ac}(t)$, since a and b are interchangeable. To get the correlations, we need the expression of the variances from Equations (5.A.1), (5.A.2), which yield

$$\begin{aligned}
\mathrm{Cor}(g_{ab}, g_{ac}) &= \mathrm{Cor}(g_{ab}, g_{bc}) = \frac{1}{3}e^{-\tau} \\
\mathrm{Cor}(e^{tg_{ab}}, e^{tg_{ac}}) &= \mathrm{Cor}(e^{tg_{ab}}, e^{tg_{bc}}) = \frac{e^{(2t-1)\tau}}{3-4t},
\end{aligned}$$

We also need to examine $M_{ac,bc}(t)$ which is distinct from the other two cases. Here,

$$E_{ac,bc}|(T,\mathbf{p}) \overset{d}{=} \begin{cases} 2\,\mathrm{Exp}(1), & \text{if } T \in \left(\frac{S_{ab}}{2}, \frac{S_{ac}}{2}\right), \mathbf{p} = (a,b) \\ 2\left(T - \frac{S_{ac}}{2}\right) + 2\,\mathrm{Exp}(1), & \text{if } T \geq \frac{S_{ac}}{2}, \mathbf{p} = (a,b) \\ 2\left(T - \frac{S_{ac}}{2}\right) + \mathrm{Exp}(1), & \text{if } T \geq \frac{S_{ac}}{2}, \mathbf{p} \neq (a,b). \end{cases}$$

Following the same approach as before,

$$\begin{aligned} M_{ac,bc}(t) &= \int_{S_{ab}/2}^{S_{ac}/2} e^{-\left(T - \frac{S_{ab}}{2}\right)} \frac{1}{1-2t}\, dT \\ &+ \int_{S_{ac}/2}^{+\infty} e^{-\tau} e^{\left(-3\left(T - \frac{S_{ac}}{2}\right)\right)} \frac{e^{2t\left(T - \frac{S_{ac}}{2}\right)}}{1-2t}\, dT \quad (\text{for } \mathbf{p} = (a,b)) \\ &+ 2\int_{S_{ac}/2}^{+\infty} e^{-\tau} e^{\left(-3\left(T - \frac{S_{ac}}{2}\right)\right)} \frac{e^{2t\left(T - \frac{S_{ac}}{2}\right)}}{1-t}\, dT \quad (\text{for } \mathbf{p} \neq (a,b)) \\ &= \frac{1}{1-2t} - e^{-\tau} \frac{2t^2}{(1-t)(1-2t)(3-2t)}. \end{aligned}$$

Hence, it follows that

$$\begin{aligned} \mathrm{Cov}(g_{ac}, g_{bc}) &= 4 - \frac{8}{3}e^{-\tau} \geq 0 \\ \mathrm{Cov}(e^{tg_{ac}}, e^{tg_{bc}}) &= \frac{4t^2}{(1-2t)(1-4t)} e^{2tS_{ac}} \left(\frac{1}{1-2t} - e^{-\tau}\frac{2}{3-4t}\right) \\ &\geq e^{2tS_{ac}} \frac{4t^2}{(1-2t)^2(1-4t)(3-4t)} \geq 0 \\ \mathrm{Cor}(g_{ac}, g_{bc}) &= 1 - \frac{2}{3}e^{-\tau} \\ \mathrm{Cor}(e^{tg_{ab}}, e^{tg_{ac}}) &= 1 - \frac{2(1-2t)}{3-4t}e^{-\tau}. \end{aligned}$$

Cherry tree We continue by considering the cherry tree as shown in Figure 5.A.1, with $S_{ab} \leq S_{cd} \leq S_{ad} = S_{bd} = S_{ac} = S_{bc} = \Delta$, where Δ is the diameter of the subtree containing a, b, c, d. We use this for notational convenience for the remainder of the proof and should not be confused with the diameter of the species tree contatining all

leaves. We also defined, $\tau_1 = \frac{\Delta - S_{ab}}{2}$ and $\tau_2 = \frac{\Delta - S_{cd}}{2} \leq \tau_1$. Once again, define the joint random variable $(T, \mathbf{p})$ as before. For the cherry tree case, it has density function

$$f_{T,\mathbf{p}}(T,\mathbf{p}) = \begin{cases} \exp\left(-\left(T - \frac{S_{ab}}{2}\right)\right), & \text{if } T \in \left(\frac{S_{ab}}{2}, \frac{S_{ac}}{2}\right), \mathbf{p} = (a,b) \\ \exp(-(\tau_1 - \tau_2)) \exp\left(-2\left(T - \frac{S_{ac}}{2}\right)\right), & \text{if } T \in \left(\frac{S_{ac}}{2}, \frac{\Delta}{2}\right), \mathbf{p} = (a,b) \text{ or } (c,d) \\ \exp(-\tau_1 - \tau_2)) \exp\left(-6\left(T - \frac{S_{ac}}{2}\right)\right), & \text{if } T \geq \frac{\Delta}{2} \, \forall \mathbf{p} \in \binom{\{a,b,c,d\}}{2} \\ 0, & \text{otherwise.} \end{cases}$$

Conditional on $(T, \mathbf{p})$ and the species tree S,

$$E_{ab,cd}|(T,\mathbf{p}), S \stackrel{d}{=} \begin{cases} \left(T - \frac{S_{ab}}{2}\right) + \text{Exp}(1), & \text{if } T \in \left(\frac{S_{ab}}{2}, \frac{S_{cd}}{2}\right), \mathbf{p} = (a,b) \\ 2\left(T - \frac{S_{cd}}{2}\right) + \tau_1 - \tau_2 + \text{Exp}(1), & \text{if } T \in (\frac{S_{cd}}{2}, \frac{\Delta}{2}), \mathbf{p} = (a,b),(c,d) \\ 2\left(T - \frac{\Delta}{2}\right) + \tau_1 + \tau_2 + \text{Exp}(1), & \text{if } T \geq \frac{\Delta}{2}, \mathbf{p} = (a,b),(c,d) \\ 2\left(T - \frac{\Delta}{2}\right) + \tau_1 + \tau_2 + E_{xy,xz}|S = *, & \text{if } T \geq \frac{\Delta}{2}, \mathbf{p} \neq (a,b),(c,d). \end{cases}$$

Note the last case where the first pair to coalesce is neither (a,b) nor (a,c). Let's assume without loss of generality that the first pair was (a,c). Since species a and c have coalesced, we can treat them as one leaf $a = c$ in a gene tree of 3 leaves $a = c, b, d$ that has not been resolved yet. This gene tree is generated by the Multispecies coalescent model assuming that the species tree is the star tree, denoted $S = *$, since all three species can coalesce after $T \geq \Delta/2$. Therefore,

$$E_{ab,cd}|(T,\mathbf{p}) \stackrel{d}{=} 2(T - \Delta/2) + \tau_1 + \tau_2 + E_{ab,ad}|S = *.$$

We have already computed the MGF of E for three-leaf trees. When $S = *$, the internal

branch length vanishes i.e. $\tau = 0$. Hence,

$$\mathbb{E}\left(e^{t(E_{xy,xz}|S=*)}\right) = \frac{1}{(1-t)^2} + \frac{t^2}{(1-t)^2(1-2t)(3-2t)} = \frac{3-5t}{(1-t)(1-2t)(3-2t)}, \text{and}$$

$$\mathbb{E}\left(e^{tE_{ab,cd}}|(T,\mathbf{p})\right) = \begin{cases} \frac{\exp\left(t\left(T-\frac{S_{ab}}{2}\right)\right)}{1-t}, & \text{if } T \in \left(\frac{S_{ab}}{2}, \frac{S_{cd}}{2}\right), \mathbf{p} = (a,b) \\ \frac{\exp\left(2t\left(T-\frac{S_{cd}}{2}\right)+t(\tau_1-\tau_2)\right)}{1-t}, & \text{if } T \in \left(\frac{S_{ac}}{2}, \frac{\Delta}{2}\right), \mathbf{p} = (a,b) \text{ or } (c,d) \\ \frac{\exp\left(2t\left(T-\frac{\Delta}{2}\right)+t(\tau_1+\tau_2)\right)}{1-t}, & \text{if } T \geq \frac{\Delta}{2}, \mathbf{p} = (a,b) \text{ or } (c,d) \\ \frac{\exp\left(2t\left(T-\frac{\Delta}{2}\right)+t(\tau_1+\tau_2)\right)(3-5t)}{(1-t)(1-2t)(3-2t)}, & \text{if } T \geq \frac{\Delta}{2}, \mathbf{p} \neq (a,b) \text{ or } (c,d). \end{cases}$$

Hence,

$$\begin{aligned} M_{ab,cd}(t) &= \int_{S_{ab}/2}^{S_{ac}/2} e^{-\left(T-\frac{S_{ab}}{2}\right)} \frac{e^{t\left(T-\frac{S_{ab}}{2}\right)}}{1-t}\, dT \\ &+ 2\int_{S_{ac}/2}^{\Delta/2} e^{-(\tau_1-\tau_2)} e^{\left(-2\left(T-\frac{S_{ac}}{2}\right)\right)} \frac{e^{2t\left(T-\frac{S_{ac}}{2}\right)} e^{t(\tau_1-\tau_2)}}{1-2t}\, dT \\ &+ 2\int_{\Delta/2}^{+\infty} e^{-\tau_1-\tau_2} e^{\left(-6\left(T-\frac{\Delta}{2}\right)\right)} \frac{e^{2t\left(T-\frac{\Delta}{2}\right)} e^{t(\tau_1+\tau_2)}}{1-t}\, dT \qquad (\text{for } \mathbf{p} = (a,b),(c,d)) \\ &+ 4\int_{\Delta/2}^{+\infty} e^{-\tau_1-\tau_2} e^{\left(-6\left(T-\frac{\Delta}{2}\right)\right)} \frac{e^{2t\left(T-\frac{\Delta}{2}\right)} e^{t(\tau_1+\tau_2)}(3-5t)}{(1-t)(1-2t)(3-2t)}\, dT \quad (\text{for } \mathbf{p} \neq (a,b),(c,d)) \\ &= \frac{1}{(1-t)^2} + e^{-\tau_1-\tau_2} \frac{2t^2}{(1-t)^2(1-2t)(3-2t)(3-t)}. \end{aligned}$$

From this we conclude that

$$\begin{aligned} \text{Cov}(g_{ab}, g_{cd}) &= \frac{8}{9} e^{-(\tau_1+\tau_2)} \geq 0 \\ \text{Cov}(e^{tg_{ab}}, e^{tg_{cd}}) &= e^{t(S_{ab}+S_{cd})-(\tau_1+\tau_2)} \frac{8t^2}{(1-2t)^2(1-4t)(3-4t)(3-2t)} \\ \text{Cor}(g_{ac}, g_{bc}) &= \frac{2}{9} e^{-(\tau_1+\tau_2)} \\ \text{Cor}(e^{tg_{ab}}, e^{tg_{ac}}) &= e^{-(\tau_1+\tau_2)} \frac{2}{(3-4t)(3-2t)}. \end{aligned}$$

A different approach is used for finding the MGF of $E_{ac,bd}$. Define the random variable $N \in \{0, 1, 2\}$ to be the number of pairs that coalesce before time $\Delta/2$. The only pairs that could coalesce in that time interval are $\mathbf{p} = (a, b)$ or (c, d) and the probabilities of each one of these pairs coalescing is independent of the other. Therefore,

$$\begin{aligned}
\mathbb{P}(N = 0) &= e^{-\tau_1 - \tau_2}, \\
\mathbb{P}(N = 1) &= (1 - e^{-\tau_1})e^{-\tau_2} + (1 - e^{-\tau_2})e^{-\tau_1}, \\
\mathbb{P}(N = 2) &= (1 - e^{-\tau_1})(1 - e^{-\tau_2}).
\end{aligned}$$

If $N = 0$, then $E_{ac,bd} = E_{xy,zw}|S = *$ i.e. the MGF of $E_{ac,bd}$ will be the same as $M_{ab,cd}(t)$ with $\tau_1 = \tau_2 = 0$ or

$$\mathbb{E}\left(e^{tE_{ac,bd}}|N = 0\right) = \frac{1}{(1-t)^2} + \frac{2t^2}{(1-t)^2(1-2t)(3-2t)(3-t)}.$$

If $N = 1$, either (a, b) or (c, d) have coalesced before time $\Delta/2$ but not both, and so at time $\Delta/2$ there will be three nodes that have not coalesced yet. Consequently, $E_{ac,bd} = E_{xy,xw}|S = *$ i.e. the MGF of $E_{ac,bd}$ will be the same as $M_{ab,ac}(t)$ with $\tau = 0$ or

$$\mathbb{E}\left(e^{tE_{ac,bd}}|N = 1\right) = \frac{1}{(1-t)^2} + \frac{t^2}{(1-t)^2(1-2t)(3-2t)}.$$

For $N = 2$, both (a, b) and (c, d) have coalesced, $E_{ac,bd} = 2\,\mathrm{Exp}(1)$, and thus

$$\mathbb{E}\left(e^{tE_{ac,bd}}|N = 2\right) = \frac{1}{1-2t} = \frac{1}{(1-t)^2} + \frac{t^2}{(1-t)^2(1-2t)}.$$

Combining those three results,

$$\begin{aligned} M_{ac,bd}(t) &= \mathbb{E}_N(\mathbb{E}(e^{tE_{ac,bd}}|N)) = \sum_{n=0}^{2} \mathbb{E}(e^{tE_{ac,bd}}|N=n)\mathbb{P}(N=n) \\ &= \frac{1}{(1-t)^2} + \frac{2t^2 e^{-\tau_1-\tau_2}}{(1-t)^2(1-2t)(3-2t)(3-t)} \\ &+ \frac{t^2\left((1-e^{-\tau_1})e^{-\tau_2} + (1-e^{-\tau_2})e^{-\tau_1}\right)}{(1-t)^2(1-2t)(3-2t)} + \frac{t^2(1-e^{-\tau_1})(1-e^{-\tau_2})}{(1-t)^2(1-2t)}. \end{aligned}$$

The same argument can be made for the pair (a,d) and (b,c), since a, b are interchangeable. We conclude that

$$\mathrm{Cor}(g_{ad}, g_{bc}) = \mathrm{Cor}(g_{ac}, g_{bd}) = 1 - \frac{2}{3}\left(e^{-\tau_1} + e^{-\tau_2}\right) + \frac{5}{9}e^{-(\tau_1+\tau_2)} \tag{5.A.5}$$

$$\begin{aligned} \mathrm{Cor}(e^{tg_{ad}}, e^{tg_{bc}}) = \mathrm{Cor}(e^{tg_{ac}}, e^{tg_{bd}}) &= 1 - \frac{2(1-2t)}{3-4t}(e^{-\tau_1} + e^{-\tau_2}) \\ &+ e^{-(\tau_1+\tau_2)}\frac{(5-4t)(1-2t)}{(3-4t)(3-2t)}. \end{aligned} \tag{5.A.6}$$

We now need to prove that the correlations in Equation (5.A.5) and (5.A.6) are non-negative and that the correlation of $\mathrm{Cor}(e^{tg_{ac}}, e^{tg_{bd}})$ is an increasing function of $t \in (-\infty, 0]$. Note that since,

$$\begin{aligned} \lim_{t\to-\infty} \mathrm{Cor}(e^{tg_{ad}}, e^{tg_{bc}}) &= (1-e^{-\tau_1})(1-e^{-\tau_2}) \geq 0, \text{ and} \\ \lim_{t\to 0} \mathrm{Cor}(e^{tg_{ad}}, e^{tg_{bc}}) &= \mathrm{Cor}(g_{ad}, g_{bc}), \end{aligned}$$

proving the monotonicity of $\mathrm{Cor}(e^{tg_{ac}}, e^{tg_{bd}})$ on $(-\infty, 0]$ is sufficient for the positivity of (5.A.5) and (5.A.6).

To do that, we compute the derivative,

$$\begin{aligned} \frac{d(\mathrm{Cor}(e^{tg_{ac}}, e^{tg_{bd}}))}{dt} &= \frac{4}{(3-4t)^2}\left((e^{-\tau_1} + e^{-\tau_2}) - \frac{8t^2-16t+9}{(3-2t)^2}e^{-\tau_1-\tau_2}\right) \\ &\geq \frac{4}{(3-4t)^2}\left((e^{-\tau_1} + e^{-\tau_2}) - 2e^{-\tau_1-\tau_2}\right) \geq 0. \end{aligned}$$

The first inequality comes from the observation that

$$\frac{8t^2 - 16t + 9}{(3-2t)^2} \geq 2 \quad \Leftrightarrow \quad t < \frac{9}{8}.$$

The second inequality stems from the fact that the function

$$g : [0,1]^2 \to \mathbb{R},\ g(x,y) = x + y - 2xy$$

has non-negative range, which can be proved by rewriting g as

$$g(x,y) = 1/2 - \left(x - \frac{1}{2}\right)(2y-1),$$

which is clearly minimized at $x, y = (1,1)$, and so $\min_{(x,y)\in[0,1]^2} g(x,y) = g(1,1) = 0$. This concludes the proof for the case of the cherry tree.

Comb tree Finally, we conclude this proof with the last tree shape of Fig. 5.A.1, the comb tree, where $S_{ab} \leq S_{ac} = S_{bc} \leq S_{ad} = S_{bd} = S_{cd} = \Delta$ and $\tau_1 = \frac{S_{ac} - S_{ab}}{2}$, $\tau_2 = \frac{\Delta - S_{ac}}{2}$. The joint density function of $(T, \mathbf{p})$ is

$$f_{T,\mathbf{p}}(T,\mathbf{p}) = \begin{cases} \exp\left(-\left(T - \frac{S_{ab}}{2}\right)\right), & \text{if } T \in \left(\frac{S_{ab}}{2}, \frac{S_{ac}}{2}\right), \mathbf{p} = (a,b) \\ \exp(-\tau_1)\exp\left(-3\left(T - \frac{S_{ac}}{2}\right)\right), & \text{if } T \in \left(\frac{S_{ac}}{2}, \frac{\Delta}{2}\right), \forall \mathbf{p} \in \binom{\{a,b,c\}}{2} \\ \exp(-\tau_1 - 3\tau_2))\exp\left(-6\left(T - \frac{\Delta}{2}\right)\right), & \text{if } T \geq \frac{\Delta}{2}\, \forall \mathbf{p} \in \binom{\{a,b,c,d\}}{2} \\ 0, & \text{otherwise.} \end{cases}$$

Furthermore,

$$E_{ab,cd}|(T,\mathbf{p}),S \overset{d}{=} \begin{cases} \left(T-\frac{S_{ab}}{2}\right)+\mathrm{Exp}(1), & \text{if } T\in\left(\frac{S_{ab}}{2},\frac{S_{ac}}{2}\right),\mathbf{p}=(a,b) \\ \left(T-\frac{S_{ab}}{2}\right)+\mathrm{Exp}(1), & \text{if } T\in(\frac{S_{ac}}{2},\frac{\Delta}{2}),\mathbf{p}=(a,b) \\ \left(T-\frac{S_{ab}}{2}\right)+E_{xy,xz}|S=\mathcal{A}_{\Delta/2-T}, & \text{if } T\in(\frac{S_{ac}}{2},\frac{\Delta}{2}),\mathbf{p}=(a,c),(b,c) \\ 2\left(T-\frac{\Delta}{2}\right)+\tau_1+\tau_2+\mathrm{Exp}(1), & \text{if } T\geq\frac{\Delta}{2},\mathbf{p}=(a,b),(c,d) \\ 2\left(T-\frac{\Delta}{2}\right)+\tau_1+\tau_2+E_{xy,xz}|S=*, & \text{if } T\geq\frac{\Delta}{2},\mathbf{p}\neq(a,b),(c,d). \end{cases}$$

where the three-leaf tree $\mathcal{A}_\tau$, for $\tau>0$, is drawn in Fig. 5.A.2.

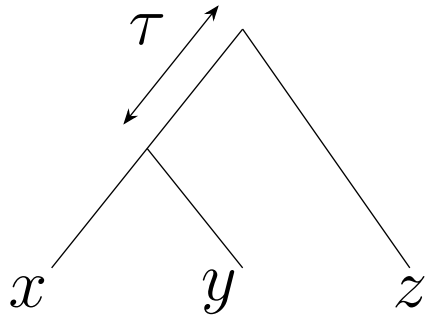


Figure 5.A.2: Three-leaf tree $\mathcal{A}_\tau$ with internal branch length τ. Note that $A_0=*$.

The MGF of $E_{xy,xz}|S=\mathcal{A}_\tau$ and $E_{xy,xz}|S=*$ has already been computed before (three-leaf tree) as

$$M_{xy,xz|\mathcal{A}_\tau}(t)=\frac{1}{(1-t)^2}+e^{(t-1)\tau}\frac{t^2}{(1-t)^2(1-2t)(3-2t)},$$
$$M_{xy,xz|*}(t)=M_{xy,xz|\mathcal{A}_0}(t)=\frac{3-5t}{(1-t)(1-2t)(3-2t)}.$$

Hence,

$$\begin{aligned}
M_{ab,cd}(t) &= \int_{S_{ab}/2}^{S_{ac}/2} e^{-\left(T-\frac{S_{ab}}{2}\right)} \frac{e^{t\left(T-\frac{S_{ab}}{2}\right)}}{1-t}\, dT \\
&+ \int_{S_{ac}/2}^{\Delta/2} e^{-\tau_1} e^{\left(-3\left(T-\frac{S_{ac}}{2}\right)\right)} \frac{e^{t\left(T-\frac{S_{ac}}{2}\right)} e^{t\tau_1}}{1-t}\, dT \quad (\text{for } \mathbf{p} = (a,b)) \\
&+ \int_{S_{ac}/2}^{\Delta/2} e^{-\tau_1} e^{\left(-3\left(T-\frac{S_{ac}}{2}\right)\right)} e^{2t\left(T-\frac{S_{ac}}{2}\right)} e^{t\tau_1} M_{xy,xz|\mathcal{A}_{\Delta/2-T}}(t)\, dT \quad (\text{for } \mathbf{p} = (a,c),(b,c)) \\
&+ 2\int_{\Delta/2}^{+\infty} e^{-\tau_1-3\tau_2} e^{\left(-6\left(T-\frac{\Delta}{2}\right)\right)} \frac{e^{2t\left(T-\frac{\Delta}{2}\right)} e^{t(\tau_1+\tau_2)}}{1-t}\, dT \quad (\text{for } \mathbf{p} = (a,b),(c,d)) \\
&+ 4\int_{\Delta/2}^{+\infty} e^{-\tau_1-3\tau_2} e^{\left(-6\left(T-\frac{\Delta}{2}\right)\right)} e^{2t\left(T-\frac{\Delta}{2}\right)} e^{t(\tau_1+\tau_2)} M_{xy,xz|*}(t)\, dT \quad (\text{for } \mathbf{p} \neq (a,b),(c,d)) \\
&= \frac{1}{(1-t)^2} + e^{(t-1)(\tau_1+\tau_2)} \frac{t^2\left((3-t)-(1-t)e^{-2\tau_2}\right)}{(1-t)^2(1-2t)(3-2t)(3-t)},
\end{aligned}$$

and so

$$\begin{aligned}
\text{Cor}(g_{ab}, g_{cd}) &= \frac{1}{3}e^{-\tau_1-\tau_2} - \frac{1}{9}e^{-\tau_1-3\tau_2} \in \left(0, \frac{1}{3}\right), \\
\text{Cor}(e^{tg_{ab}}, e^{tg_{cd}}) &= \frac{e^{(2t-1)(\tau_1+\tau_2)}}{3-4t}\left(1 - \frac{1-2t}{3-2t}e^{-2\tau_2}\right).
\end{aligned}$$

Note that $\text{Cor}(e^{tg_{ab}}, e^{tg_{cd}})$ is an increasing function of t on $(-\infty, 0)$, since the first term is increasing and $-\frac{1-2t}{3-2t}$ is also increasing.

$$E_{ac,bd}|(T,\mathbf{p}), S \overset{d}{=} \begin{cases} E_{xy,xz}|S = \mathcal{A}_{\tau_2}, & \text{if } T \in \left(\frac{S_{ab}}{2}, \frac{S_{ac}}{2}\right), \mathbf{p} = (a,b) \\ E_{ab,cd}|(T,\mathbf{p}), S, & \text{if } T \geq \frac{S_{ac}}{2}. \end{cases}$$

The last case follows from the fact that a, b, c are interchangeable conditional on (a, b) not coalescing before time $S_{ac}/2$. Moreover, note that $E_{ab,cd}|T > S_{ac}/2, S$ has MGF equal to the above $M_{ab,cd}(t)$, but with $\tau_1 = 0$ i.e.

$$M_{ab,cd|\tau_1=0}(t) = \frac{1}{(1-t)^2} + e^{(t-1)\tau_2} \frac{t^2\left((3-t)-(1-t)e^{-2\tau_2}\right)}{(1-t)^2(1-2t)(3-2t)(3-t)}.$$

Hence, it follows that

$$\begin{aligned} M_{ac,bd}(t) &= \left(1 - e^{-\tau_1}\right) M_{xy,xz|\mathcal{A}_{\tau_2}}(t) + e^{-\tau_2} M_{ab,cd|\tau_1=0}(t) \\ &= \frac{1}{(1-t)^2} + \frac{t^2}{(1-t)^2(1-2t)(3-2t)} e^{(t-1)\tau_2} \left(1 - e^{-2\tau_2-\tau_1} \frac{1-t}{3-t}\right). \end{aligned}$$

Also note that $M_{ac,bd}(t) = M_{ad,bc}(t)$. It follows that the correlations are as follows

$$\begin{aligned} \mathrm{Cor}(g_{ac}, g_{bd}) = \mathrm{Cor}(g_{ad}, g_{bc}) &= \frac{1}{3} e^{-\tau_2} - \frac{1}{9} e^{-\tau_1 - 3\tau_2} \in \left(0, \frac{1}{3}\right), \\ \mathrm{Cor}(e^{tg_{ac}}, e^{tg_{bd}}) = \mathrm{Cor}(e^{tg_{ad}}, e^{tg_{bc}}) &= \frac{e^{(2t-1)\tau_2}}{3-4t} \left(1 - \frac{1-2t}{3-2t} e^{-2\tau_2-\tau_1}\right). \end{aligned}$$

Similarly, the correlation is an increasing function of t. This concludes the proof for the case of the comb tree. □

Proof of Proposition 5.3.1. The covariance of interest can be rewritten as follows, using the properties of indicator random variables,

$$\begin{aligned} \mathrm{Cov}(\mathbb{1}(\chi_a \neq \chi_b), \mathbb{1}(\chi_c \neq \chi_d)) &= \mathrm{Cov}(\mathbb{1}(\chi_a = \chi_b), \mathbb{1}(\chi_c = \chi_d)) \\ &= \mathbb{P}(\chi_a = \chi_b, \chi_c = \chi_d) - \mathbb{P}(\chi_a = \chi_b)\mathbb{P}(\chi_c = \chi_d) \end{aligned} \tag{5.A.7}$$

Under the Jukes-Cantor evolutionary model, when there is a substitution, it is equally likely that the current base is substituted by any one of the four bases A, C, T, G. Since the probability of no substitution across a branch of time length t is $\exp(-\mu t)$, the probability that a randomly selected base stays the same across the same branch is

$$\exp(-\mu t) + \frac{1}{4}(1 - \exp(-\mu t)) = \frac{1}{4} + \frac{3}{4}\exp(-\mu t).$$

It follows that,

$$\mathbb{P}(\chi_i = \chi_j) = \frac{3}{4} e^{-\mu g_{ij}} + \frac{1}{4}, \forall i, j \in L. \tag{5.A.8}$$

The second term of Equation (5.A.7) can then be computed through a product of terms computed from Equation (5.A.8). We now focus on the first term of Equation (5.A.7). Let r and q be the internal nodes connecting a and c and b and d respectively, such that the length of the path from q to r is equal to $\delta_{ab,cd}$. Similarly to Equation (5.A.8),

$$\mathbb{P}(\chi_r = \chi_q) = \frac{3}{4}e^{-\mu\delta_{ab,cd}} + \frac{1}{4}. \tag{5.A.9}$$

We replace the states A, C, G, T with numbers $0, 1, 2, 3$ respectively. Define the random variable $N_{v,w} \equiv |\chi_v - \chi_w| \mod 4$ and note that

$$\begin{aligned}
\{\chi_r = \chi_q\} &= \{N_{rq} \equiv 0 \mod 4\},\\
\{\chi_a = \chi_b\} &= \{N_{ab} \equiv 0 \mod 4\},\\
\{\chi_c = \chi_d\} &= \{N_{cd} \equiv 0 \mod 4\}, \quad \text{where}\\
N_{ab} &\equiv N_{ar} + N_{rq} + N_{qb} \mod 4,\\
N_{cd} &\equiv N_{cr} + N_{rq} + N_{qd} \mod 4
\end{aligned}$$

Moreover, $N_{ar}, N_{rq}, N_{qb}, N_{cr}, N_{qd}$ are all independent of each other since their corresponding paths are disjoint. Hence, $N_{ar} + N_{qb}$ is independent of $N_{cr} + N_{qd}$. For the remainder of the proof $\mod 4$ will be dropped when $\equiv$ is used.

Conditioning on N_{rq},

$$\begin{aligned}
\mathbb{P}(\chi_a = \chi_b, \chi_c = \chi_d) &= \sum_{i=0}^{3} \mathbb{P}(\chi_a = \chi_b, \chi_c = \chi_d | N_{rq} \equiv i)\mathbb{P}(N_{rq} \equiv i) \\
&= \sum_{i=0}^{3} \mathbb{P}(N_{ar} + N_{qb} \equiv -i, N_{cr} + N_{qd} \equiv -i)\mathbb{P}(N_{rq} \equiv i) \\
&= \sum_{i=0}^{3} \mathbb{P}(N_{ar} + N_{qb} \equiv -i)\mathbb{P}(N_{cr} + N_{qd} \equiv -i)\mathbb{P}(N_{rq} \equiv i) \\
&= \mathbb{P}(N_{ar} + N_{qb} \equiv 0)\mathbb{P}(N_{cr} + N_{qd} \equiv 0)\mathbb{P}(\chi_r = \chi_q) \\
&\quad + \frac{1}{9}\mathbb{P}(N_{ar} + N_{qb} \not\equiv 0)\mathbb{P}(N_{cr} + N_{qd} \not\equiv 0)\mathbb{P}(\chi_r \neq \chi_q), \quad (5.A.10)
\end{aligned}$$

where the third equality holds because of the independence of $N_{ar} + N_{qb}$ and $N_{cr} + N_{qd}$, and the fourth equality holds because all the terms in the summand are equal and each probabilities involving sums of N being $\equiv i$ are a third of the corresponding probabilities with $\not\equiv 0$.

Now the only thing left is to find the distribution of $N_{ar} + N_{qb}$ and $N_{cr} + N_{qd}$. Without loss of generality, we will focus on the former. The probability that there are no substitutions across the paths from a to r and q to b is $\exp(-\mu(g_{ab} - \delta_{ab,cd}))$, since the length of the sum of the paths is $g_{ab} - \delta_{ab,cd}$. If there has been no substitution along those paths, then $N_{ar} + N_{qb} \equiv 0$. Otherwise, if there has been at least one substitution, $N_{ar} + N_{qb} \equiv i$, where $i = 0, 1, 2, 3$ are all equiprobable. Therefore,

$$\mathbb{P}(N_{ar} + N_{qb} \equiv 0) = \frac{1}{4} + \frac{3}{4}e^{-\mu(g_{ab} - \delta_{ab,cd})} \quad (5.A.11)$$

$$\mathbb{P}(N_{cr} + N_{qd} \equiv 0) = \frac{1}{4} + \frac{3}{4}e^{-\mu(g_{cd} - \delta_{ab,cd})} \quad (5.A.12)$$

Substituting Equations (5.A.11)–(5.A.12) and Equation (5.A.9) into Equation (5.A.10)

gives

$$\mathbb{P}(\chi_a = \chi_b, \chi_c = \chi_d) = \frac{1}{16}\left(1 + 3e^{-\mu(g_{ab}+g_{cd}-2\delta_{ab,cd})} + 3e^{-\mu g_{ab}} \right. \\ \left. +3e^{-\mu g_{cd}} + 6e^{-\mu(g_{ab}+g_{cd}-\delta_{ab,cd})}\right)$$

Substituting the above to Equation (5.A.7) yields the desired result. □

Proof of Proposition 5.4.1. Under the Jukes–Cantor model, the expected estimated distance satisfies $\mathbb{E}[\hat{d}_{xy}] = 1 - \mathbb{E}(e^{-\mu g_{xy}})$, where g_{xy} is the gene-tree distance between x and y and μ is the substitution rate. Under the multispecies coalescent (MSC), we may write $g_{xy} = S_{xy} + 2E_{xy}$, where S_{xy} is the species-tree path length and $E_{xy} \sim \text{Exp}(1)$ is the coalescent waiting time after speciation. Hence,

$$\mathbb{E}(e^{-\mu g_{xy}}) = e^{-\mu S_{xy}}\,\mathbb{E}\left(e^{-2\mu E_{xy}}\right) = \frac{1}{1+2\mu}\,e^{-\mu S_{xy}}.$$

The constant terms (the leading 1's in each $\mathbb{E}[\hat{d}_{xy}] = 1 - \mathbb{E}(e^{-\mu g_{xy}})$) cancel in the four-point expression, giving

$$\mathbb{E}\Big(\hat{d}_{ad} + \hat{d}_{bc} - \hat{d}_{ab} - \hat{d}_{cd}\Big) = \mathbb{E}(e^{-\mu g_{ab}}) + \mathbb{E}(e^{-\mu g_{cd}}) - \mathbb{E}(e^{-\mu g_{ad}}) - \mathbb{E}(e^{-\mu g_{bc}}),$$

and substituting the expectation above yields (5.4.1). For the cherry topology, $S_{ad} = S_{bc} = \Delta_{\{a,b,c,d\}}$, which gives the stated simplification. For the comb tree case with topology $(((a,b),c),d)$, $S_{cd} = S_{ad}$ and $S_{bc} = S_{ac}$, and so the expression collapses to $\frac{1}{1+2\mu}(e^{-\mu S_{ab}} - e^{-\mu S_{ac}})$. □

Proof of Proposition 5.4.2. Using the law of total covariance, the total covariance

between pairs of METAL distance estimates for a single loci is

$$\begin{aligned}
&\text{Cov}\left(\hat{p}_{ab}^{(1)}, \hat{p}_{cd}^{(1)}\right) \\
&= \mathbb{E}_{G^{(1)}|\mathcal{S}}\left(\text{Cov}\left(\hat{p}_{ab}^{(1)}, \hat{p}_{cd}^{(1)}|G^{(1)}\right)\right) + \text{Cov}_{G^{(1)}|\mathcal{S}}\left(\mathbb{E}\left(\hat{p}_{ab}^{(1)}|G^{(1)}\right), \mathbb{E}\left(\hat{p}_{cd}^{(1)}|G^{(1)}\right)\right) \\
&= \frac{3}{16K}\mathbb{E}_{G^{(1)}|\mathcal{S}}\left(e^{-\mu\left(g_{ab}^{(i)}+g_{cd}^{(i)}\right)}\left(e^{2\mu\delta_{ab,cd}^{(i)}} + 2e^{\mu\delta_{ab,cd}^{(i)}} - 3\right)\right) + \frac{9}{16}\text{Cov}\left(e^{-\mu g_{ab}}, e^{-\mu g_{cd}}\right) \qquad (5.A.13)
\end{aligned}$$

To derive the last line we use Proposition 5.3.1 for the first term and the Jukes-Cantor formula

$$\mathbb{E}\left(\hat{p}_{xy}^{(1)}|G^{(1)}\right) = p_{xy|G^{(1)}}^{(1)} = \frac{3}{4}\left(1 - e^{-\mu g_{xy}^{(1)}}\right).$$

for the second term. Using the definitions in Equations (5.4.2), (5.4.4), (5.4.5), Equation (5.4.3) is equivalent to (5.A.13).

It is easy to see that Σ_{coal} is positive semi-definite, as a covariance matrix of the random vector $\exp\left(-\mu g_{\binom{L}{2}}\right)$. For Σ_{sub}, observe that

$$\Sigma_{\hat{p}_{\binom{L}{2}}^{(1)}|G^{(1)}} \coloneqq \text{Cov}\left(\hat{p}_{ab}^{(1)}, \hat{p}_{cd}^{(1)}|G^{(1)}\right),$$

is positive semidefinite i.e. for any vector $\mathbf{v} \in \mathbb{R}^{\binom{n}{2}}$, $\mathbf{v}\Sigma_{\hat{p}_{\binom{L}{2}}^{(1)}|G^{(1)}}\mathbf{v}^T \geq 0$ and so

$$\mathbf{v}\Sigma_{\text{sub}}\mathbf{v}^T = \mathbb{E}\left(\mathbf{v}\Sigma_{\hat{p}_{\binom{L}{2}}^{(1)}|G^{(1)}}v^T\right) \geq 0$$

This proves that Σ_{sub} is a positive semi-definite matrix. □

Proof of Proposition 5.4.3. Using the formulas from the Proof of Proposition 5.2.2 for the coalescent covariance matrix and the formulas from Appendix 5.B for the substitution covariance matrix. Observe that for the star species tree, the three leaf covariances agree with each other for both the coalescent and the substitution matrix. So

do the results for the four leaf covariances. In other words, $(\Sigma_{\text{coal}})_{ab,cd}$ and $(\Sigma_{\text{sub}})_{ab,cd}$ only depend on the total number of leaves $|\{a,b,c,d\}| \in \{2,3,4\}$. Also, note that $|\{a,b\} \cap \{c,d\}| = 4 - |\{a,b,c,d\}| \in \{0,1,2\}$ and so $\sigma^{(0)}, \sigma^{(1)}, \sigma^{(2)}$ now refer to 4-leaf, 3-leaf, and 2-leaf covariances respectively.

First, we derive the formulas of $\sigma^{(i)}_{\text{mode}}$ for mode $\in \{\text{coal}, \text{sub}\}, i \in \{0,1,2\}$.

$$\sigma^{(0)}_{\text{coal}}(\mu) = \frac{8\mu^2}{(1+4\mu)(1+2\mu)^2(3+4\mu)(3+2\mu)} e^{-2\mu\Delta} \tag{5.A.14}$$

$$\sigma^{(1)}_{\text{coal}}(\mu) = \frac{4\mu^2}{(1+4\mu)(1+2\mu)^2(3+4\mu)} e^{-2\mu\Delta} = \frac{3+2\mu}{2}\sigma^{(0)}_{\text{coal}}(\mu)$$

$$\sigma^{(2)}_{\text{coal}}(\mu) = \frac{4\mu^2}{(1+4\mu)(1+2\mu)^2} e^{-2\mu\Delta} = (3+4\mu)\sigma^{(1)}_{\text{coal}}(\mu) \tag{5.A.15}$$

$$\begin{aligned}
\sigma^{(0)}_{\text{sub}}(\mu) &= \frac{e^{-2\mu\Delta}}{(1+2\mu)(3+2\mu)} - \sigma^{(0)}_{\text{coal}}(\mu) - \frac{e^{-2\mu\Delta}}{(1+2\mu)^2} \\
&= \frac{4\mu(8\mu^2+17\mu+6)}{3(\mu+1)(2\mu+1)(2\mu+3)(4\mu+1)(4\mu+3)} e^{-2\mu\Delta} \\
\sigma^{(1)}_{\text{sub}}(\mu) &= \frac{2}{3}\frac{e^{-3\mu\Delta/2}}{(1+2\mu)(1+\mu)} + \frac{1}{3}\frac{e^{-\mu\Delta}}{1+2\mu} - \sigma^{(1)}_{\text{coal}}(\mu) - \frac{e^{-2\mu\Delta}}{(1+2\mu)^2} \\
&\geq e^{-2\mu\Delta}\frac{2\mu(8\mu^2+17\mu+6)}{3(\mu+1)(4\mu+1)(4\mu+3)(2\mu+1)} = \frac{3+2\mu}{2}\sigma^{(0)}_{\text{sub}}(\mu) \\
\sigma^{(2)}_{\text{sub}}(\mu) &= \frac{2}{3}\frac{e^{-\mu\Delta}}{1+2\mu} + \frac{1}{3} - \sigma^{(2)}_{\text{coal}}(\mu) - \frac{e^{-2\mu\Delta}}{(1+2\mu)^2} \geq \frac{8}{3}\frac{\mu(\mu+1)}{(4\mu+1)(2\mu+1)} e^{-2\mu\Delta}
\end{aligned}$$

The inequalities are derived by using $e^{-i\mu\Delta} \leq e^{-2\mu\Delta}$ for all $i \leq 2$. From here, the asymptotic results as $\mu \to 0$ and $\mu \to \infty$ are easy to derive. It is also clear that $\sigma^{(i)}_{\text{coal}}$ is decreasing with $i \in \{0,1,2\}$, and that $\sigma^{(1)}_{\text{sub}}(\mu) \geq \sigma^{(0)}_{\text{sub}}(\mu)$. We now prove that $\sigma^{(2)}_{\text{sub}}(\mu) \geq \sigma^{(1)}_{\text{sub}}(\mu)$,

$$\begin{aligned}
\sigma^{(2)}_{\text{sub}}(\mu) - \sigma^{(1)}_{\text{sub}}(\mu) &\geq \frac{1}{3}e^{-3\mu\Delta/2}\frac{2\mu(\mu+2)}{(1+2\mu)(1+\mu)} - 2(1+2\mu)\sigma^{(1)}_{\text{coal}}(\mu) \\
&\geq e^{-2\mu\Delta}\frac{2\left(16\mu^3+36\mu^2+23\mu+6\right)}{3(\mu+1)(2\mu+1)(4\mu+1)(4\mu+3)} > 0.
\end{aligned}$$

This concludes the proof. □

Proof of Proposition 5.4.4. [1] First we find the spectrum of the following matrix;

$$A \in \mathbb{R}^{\binom{n}{2}\times\binom{n}{2}}, A_{ab,cd} = |\{a,b\} \cap \{c,d\}| \in \{0,1,2\}$$

We need to prove the following results for A

i. $r(A) \leq n$.

ii. Take two distinct leaves $x, y \in [n]$. Consider the vector $v^{(xy)} \in \mathbb{R}^{\binom{n}{2}}$ with $v^{(xy)}_{xy} = 2n-4$, $v^{(xy)}_{xz} = v^{(xy)}_{yz} = n-4$ and $v^{(xy)}_{zw} = -4$ for all distinct $z, w \neq x, y$. Then $v^{(xy)}$ is an eigenvector of A with eigenvalue $n-2$. In particular, the eigenvectors $v^{(12)}, v^{(13)}, \ldots, v^{(1n)}$ are linearly independent and so $\dim(E_{n-2}(A)) \geq n-1$, where $E_\rho(Y)$ is the eigenspace of matrix Y associated with eigenvalue ρ.

The first result shows that the dimension of the eigenspace $E_0(A)$ is at least $\binom{n}{2} - n = n(n-3)/2$ by the rank-nullity theorem. The second result shows that $\dim(E_{n-2}(A)) \geq n-1$. The dominant eigenvalue is $2n-2$ (with corresponding eigenvector $\mathbf{1}$) and the corresponding eigenspace has dimension at least 1. Since the sum of the dimensions of eigenspaces is $\binom{n}{2}$, it follows that

$$\begin{aligned}\dim(E_{2n-2}(A)) &= 1,\\ \dim(E_{n-2}(A)) &= n-1,\\ \dim(E_0(A)) &= \frac{n(n-3)}{2}.\end{aligned}$$

Proof of i. First, observe that the sum of rows is a multiple of the all-ones vector

[1] It has been noted that this proof can be shortened by using the fact that C can be constructed from the adjacency matrix of a Johnson graph, whose spectral properties are well-known Brouwer and Haemers (2011).

$\mathbf{1} = (1, \ldots, 1)$, since

$$\sum_{a<b} A_{ab,cd} = \sum_{a<b} |\{a,b\} \cap \{c,d\}| = 2(n-1), \ \forall (c,d) \in \binom{[n]}{2}.$$

Hence, it suffices to prove that the space spanned by the first $n-1$ rows $A_{1l,\cdot}$ for $l \in \{2, \ldots, n\}$ along with the row $(1, \ldots, 1)$ (which is the sum of all the rows of A) contains all the other rows $A_{xy,\cdot}$ for $x, y \neq 1$. To do that, our goal is to write $A_{xy,\cdot}$ as a linear combination of the first $n-1$ rows the vector $(1, \ldots, 1)$ for any $x, y > 1$, and proving element-wise that it holds. Consider the coefficients $\lambda_j = \lambda + |\{x,y\} \cap \{j\}|, j \in \{2, \ldots, n\}$, corresponding to row $A_{1j,\cdot}$, where λ is a free parameter to be determined. Also let -2λ be the coefficient corresponding to the newly created row $\mathbf{1} = (1, 1, \ldots, 1)$. Then, for all components zw with $z \neq w$ and $z, w \in [n]$,

$$\begin{aligned}
\sum_{j=2}^{n} \lambda_j A_{1j,zw} + (-2\lambda)\mathbf{1}_{zw} &= -2\lambda + \sum_{j=2}^{n} (\lambda + |\{j\} \cap \{x,y\}|) \, |\{1,j\} \cap \{z,w\}| \\
&= -2\lambda + \sum_{j=2}^{n} (\lambda + |\{j\} \cap \{x,y\}|) \, (|\{1\} \cap \{z,w\}| + |\{j\} \cap \{z,w\}|) \\
&= -2\lambda - 2\lambda \, |\{1\} \cap \{z,w\}| \\
&\quad + \sum_{j=1}^{n} (\lambda + |\{j\} \cap \{x,y\}|) \, (|\{1\} \cap \{z,w\}| + |\{j\} \cap \{z,w\}|) \\
&= -2\lambda - 2\lambda \, |\{1\} \cap \{z,w\}| \, (\lambda n + 2) \, |\{1\} \cap \{z,w\}| + 2\lambda \\
&\quad + \sum_{j=1}^{n} |\{j\} \cap \{z,w\}| \cdot |\{j\} \cap \{x,y\}| \\
&= |\{1\} \cap \{z,w\}| \, (\lambda(n-2) + 2) + |\{z,w\} \cap \{x,y\}| = A_{xy,zw}
\end{aligned}$$

For the last equality to hold, note that first term vanishes if we choose $\lambda = \frac{-2}{n-2}$ and the summation term is equal to $|\{z,w\} \cap \{x,y\}|$. Hence, we have proved the following

result

$$\sum_{j=2}^{n} \lambda_j^{(x,y)} A_{1j,.} + \lambda_1^{(x,y)} \mathbf{1} = A_{xy,.}, \forall x, y \in \{2, 3, \ldots, n\}, x \neq y$$

which means that for all distinct $x, y \in [n]$ we have $A_{xy,.} \in \text{span}\{A_{12,.}, \ldots, A_{1n,.}, \mathbf{1}\}$. Hence, the rank of A cannot exceed n.

Proof of ii. We now need to verify that $Av^{(xy)} = (n-2)v^{(xy)}$. First, observe that

$$(A^2)_{ab,cd} = \sum_{x<y} A_{ab,xy} A_{xy,cd} = \begin{cases} 2n, & \text{if} |\{a,b\} \cap \{c,d\}| = 2, \\ n+2, & \text{if} |\{a,b\} \cap \{c,d\}| = 1, \\ 4, & \text{if} |\{a,b\} \cap \{c,d\}| = 0 \end{cases}$$

and so it follows that

$$A^2 = (n-2)A + 4 \cdot \mathbf{1}\mathbf{1}^T,$$

where $\mathbf{1} = (1, \ldots, 1)$ is a column vector i.e. $\mathbf{1}\mathbf{1}^T \in \mathbb{R}^{\binom{n}{2} \times \binom{n}{2}}$ is a matrix of all-ones. Given that $A\mathbf{1} = (2n-2)\mathbf{1}$, we conclude that

$$A(nA - 4 \cdot \mathbf{1}\mathbf{1}^T) = (n-2)(nA - 4 \cdot \mathbf{1}\mathbf{1}^T),$$

which implies that the columns of $V \coloneqq nA - 4 \cdot \mathbf{1}\mathbf{1}^T$ are all eigenvectors of A with eigenvalue $n-2$. Note that $v^{(xy)} = V_{.,xy}$ i.e. the vectors constructed in statement (ii) are precisely the columns of V and hence elements in the eigenspace $E_{n-2}(A)$.

To prove that $v^{(12)}, \ldots, v^{(1n)}$ are linearly independent, consider the square matrix $B \in \mathbb{R}^{(n-1)\times(n-1)}$ with $B_{i,j} = v_i^{(1j)}$. It suffices to prove that B is non-singular. Note that $B_{i,i} = 2n-4 > n-4 = B_{i,j} > 0$ for all $j \neq i$ and $i \in \{2, \ldots, n\}$ and denote them as $B_{ii} = x, B_{ij} = y$ for all $i \in [N], j \neq i$ where $N = n-1$ and $x > y > 0$. Observing

that B is a circulant matrix, its eigenvalues are

$$\rho_k = \sum_{j=0}^{N-1} c_j \omega^{kj}, k \in \{0, \ldots, N-1\}$$

where $c_0 = x$, $c_1 = \cdots = c_{N-1} = y$, and ω is the N^{th}-root of unity. Hence, $\rho_0 = x + (n-1)y > 0$ and $\rho_k = x + y\sum_{j=1}^{N-1} \omega^{kj} = x - y > 0$ for $k \in \{1, \ldots, N-1\}$. All eigenvalues are positive and so B is non-singular.

Having found the spectrum of A, the task is to now find the spectrum of C. Observe that

$$C = (\beta - \gamma)A + (\gamma + \alpha - 2\beta)I + \gamma\mathbf{1},$$

where $\mathbf{1}$ is a matrix whose entries are all 1.

Every eigenvalue and eigenvector that is not $\mathbf{1}$ is transformed

$$\tilde{\lambda} = (\beta - \gamma)\lambda + (\gamma + \alpha - 2\beta)$$
$$\tilde{v} = v - \mathbf{1}\frac{\gamma v^T \mathbf{1}}{(\beta - \gamma)(2n - 2 - \lambda) + \gamma n}$$

In particular, if $Av = \lambda v$, then $C\tilde{v} = \tilde{\lambda}\tilde{v}$. This argument doesn't work for $v = \mathbf{1}$ since $\tilde{v} = 0$. But it's straightforward to compute the eigenvalue directly

$$C\mathbf{1} = \left((\beta - \gamma)(2n-2) + \gamma + \alpha - 2\beta + \binom{n}{2}\gamma\right)\mathbf{1}$$

The eigenvalue $n-2$ of A becomes $(\beta - \gamma)(n-2) + \gamma + \alpha - 2\beta$ and the corresponding eigenvectors do not change because $v^T\mathbf{1} = 0$ for all the vectors in $E_{n-2}(A)$. The remaining eigenvalue is $\gamma + \alpha - 2\beta$. □

Proof of Corollary 5.4.5. For the trace and Frobenius norm, the results of Propositions 5.4.3 suffice to prove them. For the determinant, the spectral norm, and the results about variance ratios, the spectrum of those matrices is given by Proposition 5.4.4.

Regarding the variance ratio, note that Proposition 5.4.4 provides limits of variance ratios as $n \to \infty$. The variance ratio of PC1 is γ/α and the variance variance of the first n prinipal components is $\frac{2\beta-\gamma}{\alpha}$. For the covariance matrix Σ_{mode} where mode is either coal or sub, $\alpha = \sigma^{(2)}_{\text{mode}}, \beta = \sigma^{(1)}_{\text{mode}}, \gamma = \sigma^{(0)}_{\text{mode}}$.

We start with Σ_{coal}, whose components α, β, γ are given by Equations (5.A.14)–(5.A.15). As $\mu \to 0$,

$$\frac{\sigma^{(i)}_{\text{coal}}}{\mu^2} \to \begin{cases} 4, & \text{if} \quad i = 2 \\ \frac{4}{3}, & \text{if} \quad i = 1 \\ \frac{8}{9}, & \text{if} \quad i = 0 \end{cases}.$$

From that it follows that

$$\lim_{\mu \to 0} \frac{\gamma_{\text{coal}}}{\alpha_{\text{coal}}} = \frac{2}{9}$$
$$\lim_{\mu \to 0} \frac{2\beta_{\text{coal}} - \gamma_{\text{coal}}}{\alpha_{\text{coal}}} = \frac{4}{9}$$

Similarly, for Σ_{sub}, as $\mu \to 0$,

$$\frac{\sigma^{(i)}_{\text{sub}}}{\mu} \to \begin{cases} \frac{8}{3} + \frac{4}{3}\Delta, & \text{if} \quad i = 2 \\ \frac{4}{3} + \frac{2}{3}\Delta, & \text{if} \quad i = 1 \\ \frac{8}{9}, & \text{if} \quad i = 0 \end{cases}.$$

From that it follows that

$$\lim_{\mu \to 0} \frac{\gamma_{\text{sub}}}{\alpha_{\text{sub}}} = \frac{2}{6 + 3\Delta}$$
$$\lim_{\mu \to 0} \frac{2\beta_{\text{sub}} - \gamma_{\text{sub}}}{\alpha_{\text{sub}}} = \frac{4 + 3\Delta}{6 + 3\Delta}$$

This concludes the proof of the Corollary. □

Proof of Proposition 5.4.6. Two statements need to be proved; first, that $v^{(i)}$ is an

eigenvector of C^{∞}_{coal} for all interior nodes i, and second, that the remaining eigenvalues are zero. Observe that C^{∞}_{coal} only takes values $\{0, 1\}$. The value of $(C^{\infty}_{\text{coal}})_{ab,cd}$ is 1 if and only if the topology of the leaves $\{a, b, c, d\}$ is either $((a, c), (b, d))$ or $((a, d), (b, c))$. Hence,

$$\begin{aligned}
\left(C^{\infty}_{\text{coal}} v^{(i)}\right)_{ab} &= \sum_{cd} \mathbb{I}\left(((a,d),(b,c)) \text{ or } ((a,c),(b,d))\right) \cdot \mathbb{I}\left((c,d) \in (L_i \times R_i)\right) \\
&= \sum_{c \in L_i, d \in R_i} \mathbb{I}\left(((a,d),(b,c)) \text{ or } ((a,c),(b,d))\right) \\
&= \sum_{c \in L_i, d \in R_i} \mathbb{I}\left((a,d),(b,c)\right) + \sum_{c \in L_i, d \in R_i} \mathbb{I}\left((a,c),(b,d)\right) \\
&= \mathbb{I}\left((a,b) \in R_i \times L_i\right) |L_i||R_i| + \mathbb{I}\left((a,b) \in L_i \times R_i\right) |L_i||R_i| \\
&= |L_i||R_i| \mathbb{I}\left((a,b) \in (L_i \times R_i) \cup (R_i \times L_i)\right) \\
&= |L_i||R_i| v^{(i)}_{ab}
\end{aligned}$$

To prove the second statement, it suffices to prove that $\sum_i |L_i||R_i| = \text{tr}(C^{\infty}_{\text{coal}})$ where the sum is over all internal nodes; that's because the trace is the sum of eigenvalues and C^{∞}_{coal} is a symmetric semi-positive definite matrix. The diagonal of C^{∞}_{coal} only includes ones, and so $\text{tr}(C^{\infty}_{\text{coal}}) = \binom{m}{2}$. Let r be the root of the tree with $|L_r| + |R_r| = m$. Denote $\xi_{\mathcal{A}}$ be the sum of eigenvalues (corresponding to interior nodes) for the tree formed from the leaf set $\mathcal{A} \subset [m]$. It follows that

$$\xi_{[m]} = |L_r|(m - |L_r|) + \xi_{L_i} + \xi_{R_i}$$

Using strong induction, we write $\xi_{L_r} = \binom{|L_r|}{2}$ and $\xi_{R_r} = \binom{m-|L_r|}{2}$. Substituting those to the formula above yields $\xi_{[m]} = \binom{m}{2}$. That concludes the inductive step. □

5.B Covariance calculations

Recall that using Equation (5.A.13) we decompose the covariance into an MSC and substitution component. Formulae for the former can be found in the Proof of Proposition 5.2.2. In this section of the Appendix, the task is to compute the latter component, namely the substitution covariance matrix for all four subtree topology cases shown in Figure 5.A.1. This substitution component can be further decomposed as follows

$$\begin{aligned}\frac{16}{9}K\left(\Sigma_{\text{sub}}\right)_{ab,cd} &= \mathbb{E}_{G^{(1)}|\mathcal{S}}\left(e^{-\mu(g_{ab}+g_{cd})}\left(\frac{1}{3}e^{2\mu\delta_{ab,cd}}+\frac{2}{3}e^{\mu\delta_{ab,cd}}-1\right)\right)\\ &= \frac{1}{3}e^{(2)}_{ab,cd}+\frac{2}{3}e^{(1)}_{ab,cd}\\ &\quad -\left(\text{Cov}\left(e^{-\mu g_{ab}},e^{-\mu g_{cd}}\right)+\frac{e^{-\mu(S_{ab}+S_{cd}))}}{(1+2\mu)^2}\right), \quad \text{where}\\ e^{(1)}_{ab,cd} &\coloneqq \mathbb{E}_{G^{(1)}|\mathcal{S}}\left(e^{-\mu(g_{ab}+g_{cd}-\delta_{ab,cd})}\right)\\ e^{(2)}_{ab,cd} &\coloneqq \mathbb{E}_{G^{(1)}|\mathcal{S}}\left(e^{-\mu(g_{ab}+g_{cd}-2\delta_{ab,cd})}\right)\end{aligned}$$

where $\text{Cov}\left(e^{-\mu g_{ab}},e^{-\mu g_{cd}}\right)$ is the covariance under MSC derived in Proposition 5.2.2. Therefore, the task now is to compute $e^{(1)}_{ab,cd}$ and $e^{(2)}_{ab,cd}$ for all different tree topologies.

5.B.1 Computing $e^{(1)}$ terms

First, we will state all the results for $e^{(1)}_{ab,cd}$, and then provide their proofs.

- **Two-leaves.** If $\{a,b\}=\{c,d\}$, then $\delta_{ab,ab}=g_{ab}$ and so

$$e^{(1)}_{ab,ab}=\frac{e^{-\mu S_{ab}}}{1+2\mu}.$$

- **Three-leaf tree** with tree topology $((a,b),c)$ has

$$e^{(1)}_{ab,ac}=e^{(1)}_{ac,bc}=e^{(1)}_{ab,bc}=\frac{e^{-\mu\left(\Delta+\frac{S_{ab}}{2}\right)}}{(1+\mu)(1+2\mu)}$$

- **Four leaves (Cherry tree)** with tree topology $((a,b),(c,d))$ has

$$e^{(1)}_{ab,cd} = \frac{e^{-\mu(S_{ab}+S_{cd})}}{(1+2\mu)^2} + e^{-(2\mu+1)\Delta+\frac{S_{ab}+S_{cd}}{2}}\frac{2\mu}{(\mu+1)(2\mu+1)^2(2\mu+3)}$$
$$e^{(1)}_{ac,bd} = e^{(1)}_{ad,bc} = \frac{e^{-\mu\left(\Delta+\frac{S_{ab}+S_{cd}}{2}\right)}}{(1+2\mu)(1+\mu)^2} + e^{-(2\mu+1)\Delta+\frac{S_{ab}+S_{cd}}{2}}\frac{\mu(\mu+2)}{(\mu+1)^2(2\mu+1)(2\mu+3)}$$

- **Four leaves (Comb tree)** with tree topology $(((a,b),c),d)$ has

$$e^{(1)}_{ab,cd} = \frac{e^{-\mu(S_{ab}+\Delta)}}{(1+2\mu)^2} + e^{-\mu(\Delta+S_{ac})-\frac{S_{ac}-S_{ab}}{2}}\frac{2\mu}{(\mu+1)(2\mu+1)^2(2\mu+3)}$$
$$e^{(1)}_{ac,bd} = e^{(1)}_{ad,bc} = \frac{e^{-\mu\left(\Delta+\frac{S_{ab}+S_{ac}}{2}\right)}}{(1+2\mu)(1+\mu)^2} + e^{-\mu(\Delta+S_{ac})-\frac{S_{ac}-S_{ab}}{2}}\frac{\mu(\mu+2)}{(\mu+1)^2(2\mu+1)(2\mu+3)}$$

Here and for the remainder of the Appendix, Δ is the diameter of the subtree with leaves a,b,c,d and should not be confused with the diameter of the total tree, which may include more leaves and thus be larger. A more accurate notation would be Δ_{abcd}, but we will write Δ for convenience.

For the **2-leaf** case, note that $\delta_{ab,ab} = g_{ab}$ and so

$$e^{(1)}_{ab,ab} = \mathbb{E}(e^{-\mu g_{ab}}) = e^{\mu S_{ab}}\mathbb{E}\left(e^{-2\mu E_1}\right) = \frac{e^{\mu S_{ab}}}{1+2\mu},$$

where $E_1 \coloneqq (g_{ab} - S_{ab})/2 \sim \text{Exp}(1)$ under the MSC, whose moment generating function is $1/(1-t)$ which gives the $1/(1+2\mu)$ term in the expression above.

For the **3-leaf** case, note that

$$\Lambda \coloneqq g_{ab} + g_{ac} - \delta_{ab,ac} = g_{ab} + g_{bc} - \delta_{ab,bc} = g_{ac} + g_{bc} - \delta_{ac,bc},$$

which is equal to the sum of the branches of the gene tree containing the three leaves $\{a,b,c\}$, regardless of the topology of the gene tree. Let T be the first coalescent time

in the gene tree i.e. $T = \frac{1}{2}\min\{g_{ab}, g_{ac}, g_{bc}\}$. Then, the total branch length of the gene tree generated by $\{a, b, c\}$ is

$$\Lambda = \begin{cases} T + \Delta + 2E_1, & \text{if } T \in \left(\frac{S_{ab}}{2}, \frac{\Delta}{2}\right) \\ 3T + 2E_1, & \text{if } T \geq \frac{\Delta}{2} \end{cases}$$

where $E_1 \sim \text{Exp}(1)$ is independent of T. Figure 5.B.1 illustrates this result.

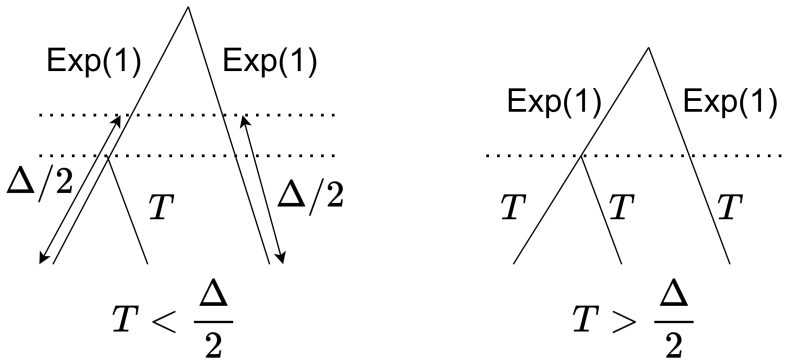


Figure 5.B.1: The total length of this gene tree is $T + 2\frac{\Delta}{2} + 2\,\text{Exp}(1)$ if $T < \Delta/2$ (left) and $3T + 2\,\text{Exp}(1)$ if $T \geq \frac{\Delta}{2}$ (right).

The pdf of T is

$$f_T(T) = \begin{cases} \exp\left(-\left(T - \frac{S_{ab}}{2}\right)\right), & \text{if } T \in \left(\frac{S_{ab}}{2}, \frac{\Delta}{2}\right) \\ 3\exp\left(-\frac{\Delta - S_{ab}}{2} - 3\left(T - \frac{\Delta}{2}\right)\right), & \text{if } T \geq \frac{\Delta}{2} \\ 0, & \text{otherwise} \end{cases}$$

Therefore, we have that

$$\begin{aligned} e^{(1)}_{ab,ac} = e^{(1)}_{ac,bc} = e^{(1)}_{ab,bc} &= \mathbb{E}_T\left(\mathbb{E}(e^{t\Lambda}|T)\right) \\ &= \frac{1}{1+2\mu}\left[\int_{\frac{S_{ab}}{2}}^{\frac{\Delta}{2}} \exp\left(-\left(T-\frac{S_{ab}}{2}\right)-\mu T-\Delta\right)\,dT\right. \\ &\left.+\int_{\frac{\Delta}{2}}^{\infty} 3\exp\left(-\frac{\Delta-S_{ab}}{2}-3\left(T-\frac{\Delta}{2}\right)-3\mu T\right)\,dT\right] \\ &= \frac{e^{-\mu\left(\Delta+\frac{S_{ab}}{2}\right)}}{(1+\mu)(1+2\mu)}, \end{aligned} \tag{5.B.1}$$

as stated for the **3-leaf tree**.

For the **4-leaf cherry tree** with topology $((a,b),(c,d))$, again denote T as the first coalescence time and without loss of generality let $S_{ab} \leq S_{cd}$. We also define the random variable Λ_0 as the sum of branch lengths of a 3-leaf gene tree, whose species tree is the star tree with zero diameter, and whose MGF is a special case of Equation (5.B.1) with $\Delta = S_{ab} = 0$, i.e.

$$\mathbb{E}\left(e^{-\mu\Lambda_0}\right) = \frac{1}{(1+2\mu)(1+\mu)}. \tag{5.B.2}$$

The central idea of this derivation is that after 2 leaves have coalesced into one, the problem is reduced to a 3-leaf case that is already computed. Note that if $T < \frac{\Delta}{2}$, then either the first pair to coalesce is either (a,b) or (c,d). Either way, $\delta_{ab,cd} = 0$. Conditioning on T and the corresponding first coalescent pair $\mathbf{p}$,

$$g_{ab}+g_{cd}-\delta_{ab,cd}|(T,\mathbf{p}) \overset{d}{=} \begin{cases} 2T+\Delta+2E_1, & \text{if } T \in \left(\frac{S_{ab}}{2},\frac{\Delta}{2}\right) \\ 4T+2E_1, & \text{if } T > \frac{\Delta}{2}, \mathbf{p} = (a,b) \text{ or } (c,d) \\ 4T+\Lambda_0, & \text{if } T > \frac{\Delta}{2}, \mathbf{p} \neq (a,b) \text{ or } (c,d) \end{cases}$$

where $E_1 \sim \text{Exp}(1)$ and Λ_0 are independent of T. It follows that,

$$\begin{aligned}
e^{(1)}_{ab,cd} &= \mathbb{E}\left(e^{-\mu(g_{ab}+g_{cd}-\delta_{ab,cd})}\right) \\
&= \int_{\frac{S_{ab}}{2}}^{\frac{\Delta}{2}} \frac{\exp\left(-\left(T-\frac{S_{ab}}{2}\right)-\mu(2T+\Delta)\right)}{1+2\mu}\,dT \\
&\quad + e^{-\frac{\Delta-S_{ab}}{2}} \int_{\frac{S_{cd}}{2}}^{\frac{\Delta}{2}} \frac{\exp\left(-\left(T-\frac{S_{cd}}{2}\right)-\mu(2T+\Delta)\right)}{1+2\mu}\,dT \\
&\quad + e^{-\left(\Delta-\frac{S_{ab}+S_{cd}}{2}\right)} \left(\int_{\frac{\Delta}{2}}^{\infty} \frac{6\exp\left(-6(T-\Delta)-4\mu T\right)\right)}{1+2\mu}\,dT \cdot \mathbb{P}\left(\mathbf{p}=(a,b),(c,d)\right) + \right. \\
&\quad \left. + \int_{\frac{\Delta}{2}}^{\infty} 6\exp\left(-6(T-\Delta)-4\mu T\right)\right)\mathbb{E}\left(e^{-\mu\Lambda_0}\right)\,dT \cdot \mathbb{P}\left(\mathbf{p}\neq(a,b),(c,d)\right) \Bigg)
\end{aligned}$$

which yields the desired result, once we substitute Equation (5.B.2) and note that $\mathbb{P}(\mathbf{p}=(a,b),(c,d)) = 2/6 = 1/3$. Similarly, note that

$$g_{ac}+g_{bd}-\delta_{ac,bd}|(T,\mathbf{p}) \overset{d}{=} \begin{cases} 4T+\Lambda(\Delta-2T, S_{cd}-2T), & \text{if } T\in\left(\frac{S_{ab}}{2},\frac{S_{cd}}{2}\right), \mathbf{p}=(a,b) \\ 4T+\Lambda(\Delta-2T,0), & \text{if } T\in\left(\frac{S_{cd}}{2},\frac{\Delta}{2}\right), \mathbf{p}=(a,b),(c,d) \\ 4T+2E_1, & \text{if } T>\frac{\Delta}{2}, \mathbf{p}=(a,c) \text{ or } (b,d) \\ 4T+\Lambda_0, & \text{if } T>\frac{\Delta}{2}, \mathbf{p}\neq(a,c) \text{ or } (b,d) \end{cases}$$

where $E_1 \sim \text{Exp}(1)$ is independent of T. It follows that

$$\begin{aligned}
e^{(1)}_{ac,bd} &= \mathbb{E}\left(e^{-\mu(g_{ac}+g_{bd}-\delta_{ac,bd})}\right) \\
&= \int_{\frac{S_{ab}}{2}}^{\frac{S_{cd}}{2}} \exp\left(-\left(T-\frac{S_{ab}}{2}\right)-4\mu T)\right)\frac{\exp\left(-\mu\left(\Delta+\frac{S_{ab}}{2}\right)+3\mu T\right)}{(1+\mu)(1+2\mu)}\,dT \\
&\quad + e^{-\frac{S_{cd}-S_{ab}}{2}}\int_{\frac{S_{cd}}{2}}^{\frac{\Delta}{2}} 2\exp\left(-2\left(T-\frac{S_{cd}}{2}\right)-4\mu T)\right)\frac{\exp\left(-\mu\Delta+2\mu T\right)}{(1+\mu)(1+2\mu)}\,dT \\
&\quad + e^{-\left(\Delta-\frac{S_{ab}+S_{cd}}{2}\right)}\left(\int_{\frac{\Delta}{2}}^{\infty}\frac{6\exp\left(-6(T-\Delta)-4\mu T)\right)}{1+2\mu}\,dT\cdot\mathbb{P}\left(\mathbf{p}=(a,c),(b,d)\right)\right. \\
&\quad \left. + \int_{\frac{\Delta}{2}}^{\infty} 6\exp\left(-6(T-\Delta)-4\mu T)\right)\frac{dT}{(1+\mu)(1+2\mu)}\cdot\mathbb{P}\left(\mathbf{p}\neq(a,c),(b,d)\right)\right) \\
&= \frac{e^{-\mu\left(\Delta+\frac{S_{ab}+S_{ac}}{2}\right)}}{(1+2\mu)(1+\mu)^2} + e^{-\mu(\Delta+S_{ac})+\frac{\Delta-S_{ab}}{2}}\frac{\mu(\mu+2)}{(\mu+1)^2(2\mu+1)(2\mu+3)}
\end{aligned}$$

Note that since a and b are interchangeable in this tree topology, $e^{(1)}_{ad,bc} = e^{(1)}_{ac,bd}$. This concludes the proof for the cherry tree case.

Finally, for the **4-leaf comb tree** case with topology $(((a,b),c),d)$, and noting that

$$g_{ab}+g_{cd}-\delta_{ab,cd}|(T,\mathbf{p}) \overset{d}{=} \begin{cases}
2T+\Delta+2E_1, & \text{if } T\in\left(\frac{S_{ab}}{2},\frac{S_{ac}}{2}\right) \\
2T+\Delta+2E_1, & \text{if } T\in\left(\frac{S_{ac}}{2},\frac{\Delta}{2}\right),\mathbf{p}=(a,b) \\
4T+\Lambda(\Delta-2T,0), & \text{if } T\in\left(\frac{S_{ac}}{2},\frac{\Delta}{2}\right),\mathbf{p}=(a,c),(b,c) \\
4T+2E_1, & \text{if } T>\frac{\Delta}{2},\mathbf{p}=(a,c) \text{ or } (b,d) \\
4T+\Lambda_0, & \text{if } T>\frac{\Delta}{2},\mathbf{p}\neq(a,c) \text{ or } (b,d)
\end{cases}$$

$$g_{ac}+g_{bd}-\delta_{ac,bd}|(T,\mathbf{p}) \stackrel{d}{=} \begin{cases} 4T+\Lambda(\Delta-2T, S_{ac}-2T), & \text{if } T \in \left(\frac{S_{ab}}{2}, \frac{S_{ac}}{2}\right) \\ 2T+\Delta+2\,\text{Exp}(1), & \text{if } T \in \left(\frac{S_{ac}}{2}, \frac{\Delta}{2}\right), \mathbf{p}=(a,c) \\ 4T+\Lambda(\Delta-2T, 0), & \text{if } T \in \left(\frac{S_{ac}}{2}, \frac{\Delta}{2}\right), \mathbf{p}=(a,b),(b,c) \\ 4T+2\,\text{Exp}(1), & \text{if } T > \frac{\Delta}{2}, \mathbf{p}=(a,c) \text{ or } (b,d) \\ 4T+\Lambda_0, & \text{if } T > \frac{\Delta}{2}, \mathbf{p} \neq (a,c) \text{ or } (b,d) \end{cases}$$

where $E_1 \sim \text{Exp}(1)$ is independent of T, the desired results are derived by following the same process in the cherry tree case.

5.B.2 Computing $e^{(2)}$ terms

First, we will state all the results for $e^{(2)}_{ab,cd}$, and then provide key ideas for their proof.

- **Two-leaves.** If $\{a,b\}=\{c,d\}$, then $\delta_{ab,ab}=g_{ab}$ and so the random variable in the exponent is zero, i.e.

$$e^{(2)}_{ab,ab}=1.$$

- **Three-leaf tree** with tree topology $((a,b),c)$ has

$$e^{(2)}_{ab,ac}=e^{(2)}_{ab,bc}=\frac{e^{-\mu S_{ac}}}{1+2\mu}$$
$$e^{(2)}_{ac,bc}=\frac{e^{-\mu S_{ab}}}{1+2\mu}$$

- **Four leaves (Cherry tree)** with tree topology $((a,b),(c,d))$ has

$$e^{(2)}_{ab,cd}=e^{(2)}_{ac,bd}=e^{(2)}_{ad,bc}=\frac{e^{-\mu(S_{ab}+S_{cd})}}{(1+2\mu)^2}+e^{-(2\mu+1)\Delta+\frac{S_{ab}+S_{cd}}{2}}\frac{4\mu}{(2\mu+1)^2(2\mu+3)}$$

- **Four leaves (Comb tree)** with tree topology $(((a,b),c),d)$ has

$$e^{(2)}_{ab,cd}=e^{(2)}_{ac,bd}=e^{(2)}_{ad,bc}=\frac{e^{-\mu(S_{ab}+\Delta)}}{(1+2\mu)^2}+e^{-\mu(\Delta+S_{ac})-\frac{S_{ac}-S_{ab}}{2}}\frac{4\mu}{(2\mu+1)^2(2\mu+3)}$$

To prove the three leaf case, observe that

$$g_{xy} + g_{xz} - 2\delta_{xy,xz} = g_{yz}, \quad \forall (x, y, z) \in \binom{L}{3}$$

In other words,

$$e^{(2)}_{xy,xz} = \mathbb{E}(e^{-\mu g_{yz}}) = \frac{e^{-\mu S_{yz}}}{1 + 2\mu}$$

For the four-leaf comb and cherry case, we observe that

$$g_{ab} + g_{cd} - 2\delta_{ab,cd} = g_{\mathbf{p}} + g_{\mathbf{p}^c},$$

where $\mathbf{p} \in \binom{\{a,b,c,d\}}{2}$ is the first pair in the gene tree to coalesce and $\mathbf{p}^c$ is the complimentary pair; the remaining two leaves. Conditioning on $T, \mathbf{p}$ just like before yields the desired results.

5.C Algorithms

Algorithm 3 describes the procedure to compute the Hamming distance between two alignment sequences. This is a sub-routine within the distance-based approaches for species tree clustering - METAL, GLASS and STEAC - which are presented in Algorithm 4.

Algorithm 3 Hamming Distance Between Two Sequences

Require: Sequences X and Y of equal length K
Require: $X, Y \in \{\mathtt{A}, \mathtt{C}, \mathtt{T}, \mathtt{G}\}^K$

```
function HAMMING(X, Y)
    d ← 0
    for i ← 1 to K do
        if X[i] ≠ Y[i] then
            d ← d + 1
        end if
    end for
    return d/K                          ▷ Normalize Hamming Distances
end function
```

5.D Additional plots

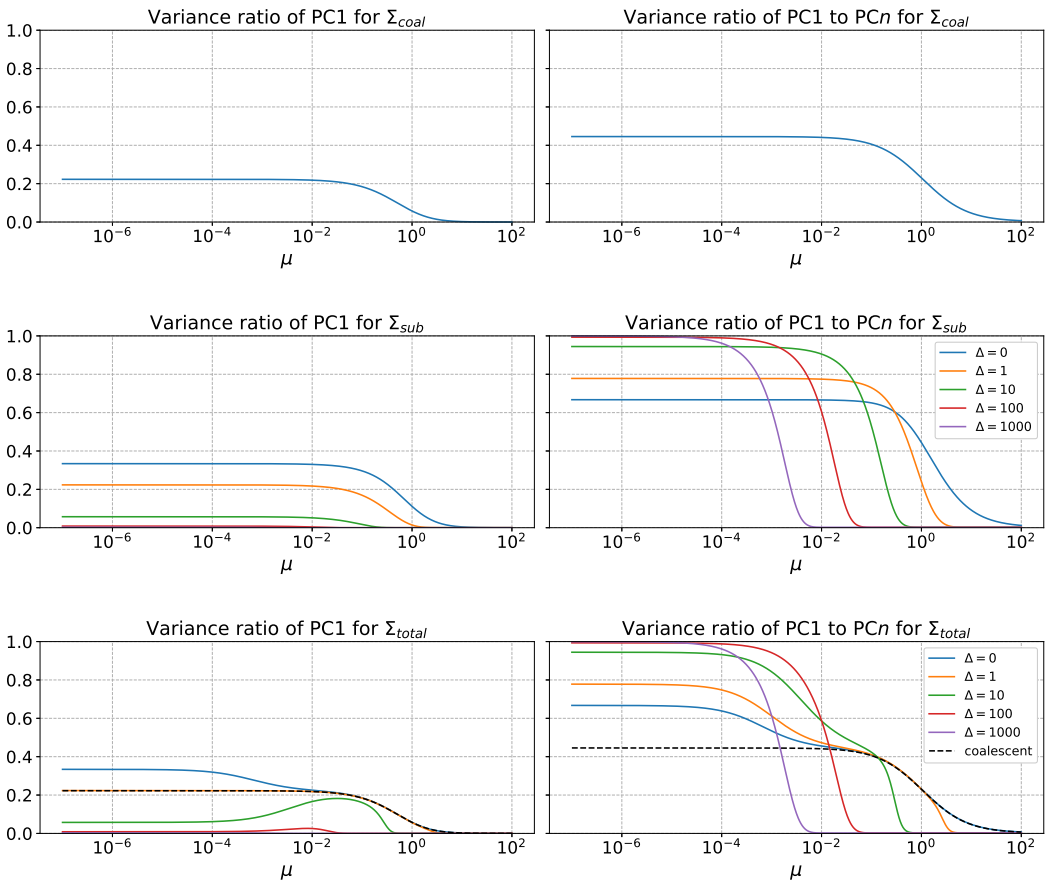


Figure 5.D.1: Variance ratios for (top) Σ_{coal}, (middle) Σ_{sub}, and (bottom) Σ_{total}, plotted against mutation rate μ and species-tree diameter Δ. In each row, the left panel shows the fraction of total variance explained by the first principal component, and the right panel shows the fraction of total variance explained by the first n principal components. Here, the number of sites per gene is $K = 1000$. Since Σ_{coal} does not depend on Δ, its row contains only a single curve. For Σ_{total}, note that when $\mu \ll 1$ and $\mu\Delta \gg 1$ (i.e., substitution variance dominates), its curves resemble those of Σ_{sub}, whereas for intermediate μ (when MSC variance dominates), all three models' curves converge to that of the coalescent covariance.

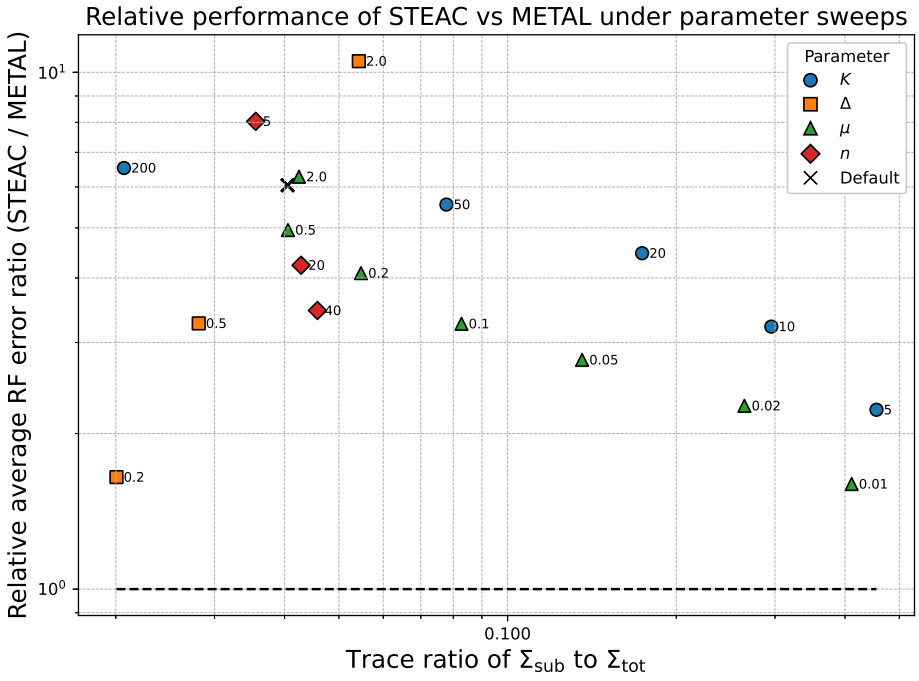


Figure 5.D.2: This Figures compares STEAC to METAL, similarly to Figure 5.5.1. In all cases METAL outperforms STEAC

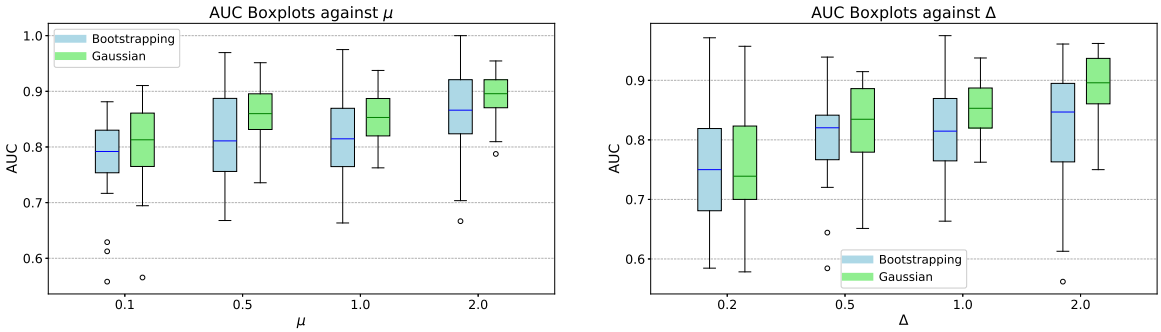


Figure 5.D.3: Boxplots of AUC values for different mutation rates μ (left), and different species tree diameter Δ (right).

Algorithm 4 Species-Tree Estimation via Agglomerative Distance-Based Methods

Require: gene sequences $\chi_i^{(g)}$ for genes $g \in G = [m]$ and leaves/species $i \in L$
Require: method $\in \{\textbf{METAL}, \textbf{STEAC}, \textbf{GLASS}\}$
Ensure: species tree T

if method == **METAL then** ▷ Concatenate gene sequences
  for all leaves $i \in L$ **do**
    $\chi_i \leftarrow \text{Concatenate}\left(\chi_i^{(1)}, \chi_i^{(2)}, \ldots, \chi_i^{(|G|)}\right)$
  end for ▷ Compute distances on concatenated alignment
  for all leaf pairs $(i,j) \in \binom{L}{2}$ **do**
    $D[i,j] \leftarrow \text{Hamming}\left(\chi_i, \chi_j\right)$ ▷ Normalized Hamming Distances
  end for
else ▷ Compute gene-by-gene distances
  for $g \in G$ **do**
    for all leaf pairs $(i,j) \in \binom{L}{2}$ **do**
      $\theta \leftarrow \text{Hamming}\left(\chi_i^{(g)}, \chi_j^{(g)}\right)$
      if $\theta \geq \frac{3}{4}$ **&** method == **STEAC then**
        G.delete(g) ▷ Delete gene
        break
      else if method == **STEAC then**
        $D_g[i,j] \leftarrow -\frac{3}{4}\log\left(1 - \frac{4}{3}\theta\right)$ ▷ Convert to coalescent units
      else if method == **GLASS then**
        $D_g[i,j] \leftarrow \theta$
      end if
    end for
  end for ▷ Aggregate across genes
  for all leaf pairs $(i,j) \in \binom{L}{2}$ **do**
    if method == **STEAC then**
      $D[i,j] \leftarrow \frac{1}{|G|}\sum_{g \in G} D_g[i,j]$
    else if method == **GLASS then**
      $D[i,j] \leftarrow \min_{g \in G} D_g[i,j]$
    end if
  end for
end if
$T \leftarrow \text{HierarchicalClustering}(D)$ ▷ Hierarchical clustering (e.g., UPGMA or NJ)
return T

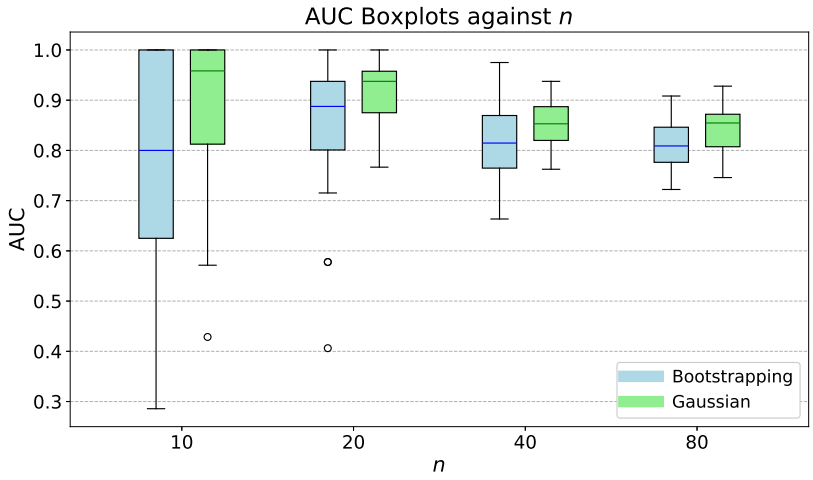


Figure 5.D.4: Boxplots of AUC values for different number of taxa n.

Chapter 6

Conclusions and Further Work

In this chapter, we summarise the main results of Chapters 3 to 5 and suggest future research directions.

In Chapter 3, we introduced the *tropical logistic regression* (TLR) model, a classifier designed for data that lie in the tropical projective torus, $\mathbb{R}^e/\mathbb{R}\mathbf{1}$. If we view classical logistic regression as equivalent to linear discriminant analysis (LDA) with equal class covariances, the decision boundary corresponds to the Euclidean bisector of the class means. Likewise, the decision boundary of TLR is the *tropical bisector* of those class centres. The model parameters of TLR are the centres of the two classes and their dispersion parameters. Tropical Fermat–Weber points are used as estimates of the class centres, so the optimization problem simplifies to optimizing over the two dispersion parameters. Theoretical results establish statistical consistency and provide bounds on the generalization error of TLR. Empirically, TLR outperformed both Euclidean and BHV logistic regression models on simulated datasets generated under the multispecies coalescent model.

We also demonstrated that TLR can serve as a diagnostic tool for Bayesian phy-

logenetic inference. By training the TLR classifier to distinguish trees sampled from two independent MCMC chains, we evaluated MCMC convergence based on the classifier's ability to separate those samples. A high area under the curve (AUC) indicates lack of convergence, whereas an AUC around 0.5 suggests that the two chains are indistinguishable and may have both converged to the equilibrium distribution. Unlike conventional convergence diagnostics such as `MrBayes`'s average standard deviation of split frequencies (ASDSF), which rely solely on topological frequencies, our TLR-based method is jointly sensitive to both topology and branch lengths. Importantly, TLR produced high AUC values for trees from chains that the ASDSF criterion considered converged. In other words, the classifier could confidently differentiate between the two chains, indicating that they were not drawn from the same distribution and had not yet converged to the equilibrium distribution, despite ASDSF suggesting otherwise. The reason is that the sampled trees had converged in topological terms only, while their branch lengths had not. In this sense, the TLR diagnostic provides a more holistic measure of convergence on tree space, albeit at a higher computational cost.

A natural next step for the TLR model is to restrict the model parameters to be ultrametric trees, which correspond to equidistant trees under a molecular clock assumption. While the tropical projective torus is more restrictive than Euclidean space, it remains a superset of the space of ultrametric trees. Sampling techniques for distributions over tropical convex sets—and in particular over the space of ultrametric trees—have already been developed by Yoshida et al. (2023b). Moreover, tropical Fermat–Weber points over the space of ultrametrics have recently been studied in Cox et al. (2025), which may provide more robust estimates of the class centres. Other promising directions for future work include obtaining tighter generalization bounds (the current ones are relatively loose) and establishing a stronger theoretical justification for the tropical Laplacian distribution—specifically, why it provides a better description of

empirical data and trees generated from the multispecies coalescent model.

Chapter 4 extends TLR by developing *tropical neural networks* (TNNs), using tropical generalized linear models (GLMs) as an intermediate step. The key idea is to construct a *tropical embedding layer*, a mapping from tree space—represented as points in the tropical projective torus—into a Euclidean space. Subsequent layers are classical: they apply standard linear transformations and nonlinear activation functions. Therefore, the network is hybrid, combining an initial tropical layer with subsequent classical layers. We also prove that any continuous function on the tropical projective torus can be approximated by a sufficiently deep TNN.

From a practical perspective, we implemented TNNs in Tensorflow 2 using the backpropagation rule adapted to the initial tropical layer and a weight initialization ensuring numerical stability across forward propagation. TNNs outpeformed conventional neural networks in classification tasks by scoring higher AUCs in both simulated and empirical datasets, such as in the classification of trees generated by MSC and the analysis of influenza hemagglutinin sequences. It is worth noting that when TNNs are restricted to a single hidden tropical layer they reduce to the TLR model. In summary, TNNs provide a deep learning framework that conform to the geometry of the tree space and theoretically guarantees expressive power in the tropical setting as ordinary neural networks in Euclidean space.

Nonetheless, one of the key limitations of the current architecture is that the AUCs were not considerably higher than those achieved by TLR. Further experimentation with different architectures is required to develop a TNN that consistently outperforms TLR. Moreover, a notable extension of tropical neural networks was presented in Pasque et al. (2024); building on our definition of TNN, they developed a convolutional TNN

with a tropical decision boundary and proved that it was robust against adversarial attacks.

Chapter 5 shifted focus from developing statistical models based on tropical geometry to analyzing the statistical properties of fast, distance-based phylogenomic inference methods such as STEAC, GLASS, and METAL, and highlighting some theoretical connections to tropical geometry. In particular, we derived the *covariance matrix* of estimated pairwise taxon distances under the combined multispecies coalescent and Jukes–Cantor substitution models. The central result was the analytical decomposition of the total covariance into two additive components: the *coalescent covariance*, which captures stochasticity in gene trees, and the *substitution covariance*, which corresponds to random mutational noise in gene sequence evolution. This decomposition enables us to explore which source of variation dominates across the parameter space. Specifically, at very low or very high mutation rates the substitutional variance dominates, whereas at intermediate rates there is a considerable contribution from the coalescent variance. Moreover, as the species-tree diameter increases, the region where coalescent variance dominates becomes narrower. In both covariance components, the first principal component is the vector $\mathbf{1} = (1, \ldots, 1)$, which explains between two ninths and one third of the total variance regardless of the number of taxa. Therefore, by working in the tropical projective torus, the amount of variance is substantially lower without losing any useful information for the reconstruction of the species tree topology.

The covariance decomposition has a concrete implication, namely that it provides guidance for experimental design. If substitution variance dominates, longer loci should be sequenced or missing sites should be imputed, whereas if coalescent variance dominates, more loci should be added. We also proposed a Gaussian sampling approach for estimating clade confidence in METAL instead of non-parametric bootstrap replicates.

The resulting confidence values more accurately reflect true support, avoiding overconfidence in incorrect clades. Its accuracy exceeds that of standard bootstrapping and is comparable to multilocus bootstrapping. Nonetheless, it is substantially faster and supported by stronger computational complexity guarantees.

The current analysis of covariance decomposition assumes a strict molecular clock, and hence equidistant gene and species trees. Although this assumption helps draw connections to the tropical projective torus, it represents a rather limited framework. Most real datasets do not follow a strict molecular clock, so generalizing the derivation to the non-clock case would be a major next step.

Bibliography

Akian, M., Gaubert, S., Qi, Y., and Saadi, O. (2021). Tropical linear regression and mean payoff games: or, how to measure the distance to equilibria. https://arxiv.org/abs/2106.01930.

Alfarra, M., Bibi, A., Hammoud, H., Gaafar, M., and Ghanem, B. (2022). On the decision boundaries of neural networks: A tropical geometry perspective. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Aliatimis, G. (2024a). Metal covariance matrix. https://github.com/GeorgiosAliatimis/metal_covariance_matrix. Accessed: 2025-06-19.

Aliatimis, G. (2024b). Tropical logistic regression. https://github.com/GeorgiosAliatimis/tropical_logistic_regression.

Aliatimis, G., Yoshida, R., Boyacı, B., and Grant, J. A. (2024). Tropical logistic regression model on space of phylogenetic trees. *Bulletin of Mathematical Biology*, 86(8):99.

Allamigeon, X., Bœuf, V., and Gaubert, S. (2015). Performance evaluation of an emergency call center: tropical polynomial systems applied to timed petri nets. In *International Conference on Formal Modeling and Analysis of Timed Systems*, pages 10–26. Springer.

Ané, C., Larget, B., Baum, D., Smith, S., and Rokas, A. (2007). Bayesian estimation of concordance among gene trees. *Mol Biol Evol.*, 24(2):412–26.

Ardila, F. and Klivans, C. J. (2006). The Bergman complex of a matroid and phylogenetic trees. *Journal of Combinatorial Theory, Series B*, 96(1):38–49.

Arora, R., Basu, A., Mianjy, P., and Mukherjee, A. (2016). Understanding deep neural networks with rectified linear units. *arXiv preprint arXiv:1611.01491*.

Baldwin, E. and Klemperer, P. (2019). Understanding preferences:"demand types", and the existence of equilibrium with indivisibilities. *Econometrica*, 87(3):867–932.

Barnhill, D. and Yoshida, R. (2023). Clustering methods over the tropical projective torus. *Mathematics*, 11(15):3433.

Barnhill, D., Yoshida, R., Aliatimis, G., and Miura, K. (2024). Tropical geometric tools for machine learning: The TML package. *Journal of Software for Algebra and Geometry*, 14(1):133–174.

Bierens, H. J. (1996). *Topics in advanced econometrics: estimation, testing, and specification of cross-section and time series models*. Cambridge University Press.

Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Giacinto, G., and Roli, F. (2013). Evasion attacks against machine learning at test time. In Blockeel, H., Kersting, K., Nijssen, S., and Železný, F., editors, *Machine Learning and Knowledge Discovery in Databases*, pages 387–402, Berlin, Heidelberg. Springer Berlin Heidelberg.

Billera, L., Holmes, S., and Vogtmann, K. (2001). Geometry of the space of phylogenetic trees. *Adv Appl Math*, 27(4):733–767.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard,

M. A., Rambaut, A., and Drummond, A. J. (2014). BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS computational biology*, 10(4):e1003537.

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kühnert, D., De Maio, N., et al. (2019). BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS computational biology*, 15(4):e1006650.

Braun, E. L., Oliveros, C. H., White Carreiro, N. D., Zhao, M., Glenn, T. C., Brumfield, R. T., Braun, M. J., Kimball, R. T., and Faircloth, B. C. (2024). Testing the mettle of metal: A comparison of phylogenomic methods using a challenging but well-resolved phylogeny. *BioRxiv*, pages 2024–02.

Brouwer, A. E. and Haemers, W. H. (2011). *Spectra of graphs*. Springer Science & Business Media.

Buneman, P. (1974). A note on the metric properties of trees. *J. Combinatorial Theory Ser. B.*, 17:48–50.

Calin, O. (2020). *Deep Learning Architectures: A Mathematical Approach*. Springer.

Camin, J. H. and Sokal, R. R. (1965). A method for deducing branching sequences in phylogeny. *Evolution*, 19(3):311–326.

Chollet, F. (2021). *Deep Learning with Python 2nd Edition*. Manning.

Comaneci, A. and Joswig, M. (2023). Tropical medians by transportation. *Math. Program*, 205:813–839.

Cox, S., Sabol, J., Talbut, R., and Yoshida, R. (2025). Tropical fermat-weber points over bergman fans. *arXiv preprint arXiv:2505.09584*.

Criado, F., Joswig, M., and Santos, F. (2021). Tropical bisectors and Voronoi diagrams. *Foundations of Computational Mathematics*, pages 1–38.

Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection*. John Murray, London.

Darwin Tree of Life Project Consortium (2022). Sequence locally, think globally: the darwin tree of life project. *Proceedings of the National Academy of Sciences*, 119(4):e2115642118.

Dasarathy, G., Nowak, R., and Roch, S. (2014). Data requirement for phylogenetic inference from multiple loci: a new distance method. *IEEE/ACM transactions on computational biology and bioinformatics*, 12(2):422–432.

Degnan, J. H. and Rosenberg, N. A. (2006). Discordance of species trees with their most likely gene trees. *PLoS genetics*, 2(5):e68.

Degnan, J. H. and Rosenberg, N. A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in ecology & evolution*, 24(6):332–340.

Develin, M. and Sturmfels, B. (2004). Tropical convexity. *Documenta Mathematica*, 9:1–27.

Drummond, A. J. and Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC evolutionary biology*, 7(1):214.

Dugad, R. and Ahuja, N. (1998). Unsupervised multidimensional hierarchical clustering. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, volume 5, pages 2761–2764. IEEE.

Edgar, R. (2004). Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32:1792–1797.

Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Biology*, 27(4):401–410.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376.

Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *evolution*, 39(4):783–791.

Felsenstein, J. (2004). *Inferring phylogenies*. Sinauer Associates.

Fitch, W. M. (1971). Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Biology*, 20(4):406–416.

Flouri, T., Jiao, X., Rannala, B., and Yang, Z. (2018). Species tree inference with bpp using genomic sequences and the multispecies coalescent. *Molecular Biology and Evolution*, 35(10):2585–2593.

Ford, M. (2018). *Architects of Intelligence: The truth about AI from the people building it*. Packt Publishing.

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using Bayesian networks to analyze expression data. In *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology*, RECOMB '00, page 127–135, New York, NY, USA. Association for Computing Machinery.

Garba, M., Nye, T., Huckmann, S., and Lueg, J. (2021). Information geometry for phylogenetic trees. *J. Math. Biol.*, 82(19):https://doi.org/10.1007/s00285–021–01553–x.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472.

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, PMLR*, 9:249–256.

Goloboff, P. A., Farris, J. S., and Nixon, K. C. (1999). Improvements to resampling measures of group support. *Cladistics*, 15(4):407–414.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. The MIT Press.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Grenfell, B. T., Pybus, O. G., Gog, J. R., Wood, J. L., Daly, J. M., Mumford, J. A., and Holmes, E. C. (2004). Unifying the epidemiological and evolutionary dynamics of pathogens. *science*, 303(5656):327–332.

Grohs, P. and Kutyniok, G. (2023). *Mathematical Aspects of Deep Learning*. Cambridge Univ Press.

Guerra, G. and Nielsen, R. (2022). Covariance of pairwise differences on a multispecies coalescent tree and implications for F_{ST}. *Philosophical Transactions of the Royal Society B*, 377(1852):20200415.

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.

Hampe, S. (2015). Tropical linear spaces and tropical convexity. *The Electronic Journal of Combinatorics*, 22(4):P4.43–P4.43.

Hasegawa, M., Kishino, H., and Yano, T.-a. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22:160–174.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Heled, J. and Drummond, A. J. (2009). Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution*, 27(3):570–580.

Hennig, W. (1999). *Phylogenetic systematics*. University of Illinois Press.

Höhna, S., Landis, M. J., Heath, T. A., Boussau, B., Lartillot, N., Moore, B. R., Huelsenbeck, J. P., and Ronquist, F. (2016). RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Systematic Biology*, 65(4):726–736.

Hood, L. and Rowen, L. (2013). The human genome project: big science transforms biology and medicine. *Genome medicine*, 5(9):79.

Huang, H., He, Q., Kubatko, L. S., and Knowles, L. L. (2010). Sources of error inherent in species-tree estimation: impact of mutational and coalescent effects on accuracy and implications for choosing among different methods. *Systematic Biology*, 59(5):573–583.

Huelsenbeck, J. P., Ronquist, F., Nielsen, R., and Bollback, J. P. (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, 294(5550):2310–2314.

Huggins, P. M., Li, W., Haws, D., Friedrich, T., Liu, J., and Yoshida, R. (2011). Bayes Estimators for Phylogenetic Reconstruction. *Systematic Biology*, 60(4):528–540.

Jewett, E. M. and Rosenberg, N. A. (2012). iglass: an improvement to the glass method for estimating species trees from gene trees. *Journal of Computational Biology*, 19(3):293–315.

Joswig, M. (2021). *Essentials of Tropical Combinatorics*. Springer, New York, NY.

Jukes, T. (1969). Evolution of protein molecules. *Mammalian Protein Metabolism*, 3.

Kingman, J. F. C. (1982). The coalescent. *Stochastic processes and their applications*, 13(3):235–248.

Lakner, C., Van Der Mark, P., Huelsenbeck, J. P., Larget, B., and Ronquist, F. (2008). Efficiency of Markov chain Monte Carlo tree proposals in Bayesian phylogenetics. *Systematic Biology*, 57(1):86–103.

Lartillot, N., Lepage, T., and Blanquart, S. (2009). Phylobayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics*, 25(17):2286–2288.

Lee, W., Li, W., Lin, B., and Monod, A. (2022). Tropical optimal transport and wasserstein distances. *Information Geometry*, 5(1):247–287.

Lin, B., Sturmfels, B., Tang, X., and Yoshida, R. (2017). Convexity in tree spaces. *SIAM Discrete Math*, 3:2015–2038.

Lin, B. and Yoshida, R. (2018a). Tropical Fermat–Weber points. *SIAM Journal on Discrete Mathematics*, 32(2):1229–1245.

Lin, B. and Yoshida, R. (2018b). Tropical Fermat-Weber points. *SIAM Discrete Math.*, page https://arxiv.org/abs/1604.04674.

Liu, L. and Pearl, D. K. (2007). Species Trees from Gene Trees: Reconstructing Bayesian Posterior Distributions of a Species Phylogeny Using Estimated Gene Tree Distributions. *Systematic Biology*, 56(3):504–514.

Liu, L., Yu, L., and Pearl, D. K. (2010). Maximum tree: a consistent estimator of the species tree. *Journal of mathematical biology*, 60:95–106.

Liu, L., Yu, L., Pearl, D. K., and Edwards, S. V. (2009). Estimating species phylogenies using coalescence times among sequences. *Systematic Biology*, 58(5):468–477.

Lu, Z., Pu, H., Wang, F., Hu, Z., and Wang, L. (2017). The expressive power of neural networks: A view from the width. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Maclagan, D. and Sturmfels, B. (2021). *Introduction to tropical geometry*, volume 161. American Mathematical Society.

Maddison, W. P. (1997). Gene trees in species trees. *Systematic Biology*, 46(3):523–536.

Maddison, W. P. (2008). Mesquite: a modular system for evolutionary analysis. *Evolution*, 62:1103–1118.

Maddison, W. P. and Maddison, D. (2009). Mesquite: a modular system for evolutionary analysis. version 2.72. Available at http://mesquiteproject.org.

Madigan, D., York, J., and Allard, D. (1995). Bayesian graphical models for discrete data. *International statistical review/revue internationale de statistique*, pages 215–232.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Mau, B., Newton, M. A., and Larget, B. (1999). Bayesian phylogenetic inference via markov chain monte carlo methods. *Biometrics*, 55(1):1–12.

Mendes, F. K. and Hahn, M. W. (2018). Why concatenation fails near the anomaly zone. *Systematic Biology*, 67(1):158–169.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092.

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., and Lanfear, R. (2020). Iq-tree 2: new models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution*, 37(5):1530–1534.

Mirarab, S., Reaz, R., Bayzid, M. S., Zimmermann, T., Swenson, M. S., and Warnow, T. (2014). Astral: genome-scale coalescent-based species tree estimation. *Bioinformatics*, 30(17):i541–i548.

Monod, A., Lin, B., Kang, Q., and Yoshida, R. (2019). Tropical foundations for probability & statistics on phylogenetic tree space.

Montúfar, G., Ren, Y., and Zhang, L. (2022). Sharp bounds for the number of regions of maxout networks and vertices of minkowski sums. *SIAM J. APPL. ALGEBRA GEOMETRY*, 6(4):618–649.

Moreno, M. A., Holder, M. T., and Sukumaran, J. (2024). Dendropy 5: a mature python library for phylogenetic computing. *Journal of Open Source Software*, 9(101):6943.

Mossel, E. and Roch, S. (2008). Incomplete lineage sorting: consistent phylogeny estimation from multiple loci. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 7(1):166–171.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245.

Nguyen, A. M., Yosinski, J., and Clune, J. (2014). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *CoRR*, abs/1412.1897.

Nye, T. M. W. (2011). Principal components analysis in the space of phylogenetic trees. *Ann. Stat.*, 39(5):2716–2739.

Nye, T. M. W., Tang, X., Weyenberg, G., and Yoshid, R. (2017). Principal component analysis and the locus of the fréchet mean in the space of phylogenetic trees. *Biometrika*, 104:901–922.

Ogilvie, H. A., Heled, J., Xie, D., and Drummond, A. J. (2016). Computational performance and statistical accuracy of* BEAST and comparisons with other methods. *Systematic Biology*, 65(3):381–396.

Onan, A. (2021). Sentiment analysis on product reviews based on weighted word embeddings and deep neural networks. *Concurrency and Computation: Practice and Experience*, 33(23):e5909.

Page, R., Yoshida, R., and Zhang, L. (2020). Tropical principal component analysis on the space of phylogenetic trees. *Bioinformatics*, 36(17):4590–4598.

Pasque, K., Teska, C., Yoshida, R., Miura, K., and Huang, J. (2024). Tropical decision boundaries for neural networks are robust against adversarial attacks. *arXiv preprint arXiv:2402.00576*.

Patterson, C. (1982). Morphological characters and homology. *Problems of phylogenetic reconstruction*, pages 21–74.

Pin, J.-E. (1998). Tropical semirings.

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., and Suchard, M. A. (2018). Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Systematic Biology*, 67(5):901–904.

Rannala, B. and Yang, Z. (1996). Probability distribution of molecular evolutionary

trees: a new method of phylogenetic inference. *Journal of Molecular Evolution*, 43(3):304–311.

Rannala, B. and Yang, Z. (2003). Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics*, 164(4):1645–1656.

Richter-Gebert, J., Sturmfels, B., and Theobald, T. (2003). First steps in tropical geometry.

Robinson, D. F. and Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical biosciences*, 53(1-2):131–147.

Roch, S., Nute, M., and Warnow, T. (2019). Long-branch attraction in species tree estimation: inconsistency of partitioned likelihood and topology-based summary methods. *Systematic Biology*, 68(2):281–297.

Roch, S. and Steel, M. (2015). Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. *Theoretical population biology*, 100:56–62.

Ronquist, F., Huelsenbeck, J. P., and van der Mark, P. (2005). Mrbayes 3.1 manual.

Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., and Huelsenbeck, J. P. (2012). Mrbayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, 61(3):539–542.

Saitou, N. and Nei, M. (1987a). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425.

Saitou, N. and Nei, M. (1987b). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol.*, 4:406–425.

Semple, C. and Steel, M. (2003). *Phylogenetics*, volume 24 of *Oxford Lecture Series in Mathematics and its Applications*. Oxford University Press, 2003.

Simon, I. (1988). Recognizable sets with multiplicities in the tropical semiring. In *International Symposium on Mathematical Foundations of Computer Science*, pages 107–120. Springer.

Simon, I. (1994). On semigroups of matrices over the tropical semiring. *RAIRO-Theoretical Informatics and Applications*, 28(3-4):277–294.

Snijders, T. (2002). Markov chain monte carlo estimation of exponential random graph models. *Journal of Social Structure*, 3.

Sokal, R. R. and Michener, C. D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38:1409–1438.

Sokal, R. R. and Sneath, P. H. A. (1963). *Principles of Numerical Taxonomy*. W. H. Freeman, San Francisco.

Speyer, D. and Sturmfels, B. (2009). Tropical mathematics. *Mathematics Magazine*, 82:163–173.

Spielman, S. J. and Wilke, C. O. (2015). Pyvolve: A flexible python module for simulating sequences along phylogenies. *PLOS ONE*, 10(9):e0139047.

Stamatakis, A. (2014). Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313.

Sukumaran, J. and Holder, M. T. (2010). Dendropy: a python library for phylogenetic computing. *Bioinformatics*, 26(12):1569–1571.

Swofford, D. L. (1998). Phylogenetic analysis using parsimony.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. J., and Fergus, R. (2014). Intriguing properties of neural networks. In Bengio, Y. and LeCun, Y., editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Tamura, K. (1992). Estimation of the number of nucleotide substitutions when there are strong transition-transversion and g+c-content biases. *Molecular Biology and Evolution*, 9(4):678–687.

Tamura, K. and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution*, 10(3):512–526.

Tang, X., Wang, H., and Yoshida, R. (2020). Tropical support vector machine and its applications to phylogenomics.

Tavaré, S. (1986). Some probabilistic and statistical problems on the analysis of DNA sequence. *Lecture of mathematics for life science*, 17:57.

Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I. J., Boneh, D., and McDaniel, P. D. (2018). Ensemble adversarial training: Attacks and defenses. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Tran, N. (2020). Tropical gaussians: a brief survey. *Algebraic Statistics*, 11(2):155–168.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Weyenberg, G., Yoshida, R., and Howe, D. (2016). Normalizing kernels in the Billera-Holmes-Vogtmann treespace. *IEEE ACM T. Comput. Bi.*, page doi:10.1109/TCBB.2016.2565475.

Woodman, W. M. and Nye, T. M. (2025). Brownian motion, bridges and Bayesian inference in phylogenetic tree space. *arXiv preprint arXiv:2506.22135*.

Yang, Z. (2015). The bpp program for species tree estimation and species delimitation. *Current Zoology*, 61(5):854–865.

Yang, Z. and Rannala, B. (2012). Molecular phylogenetics: principles and practice. *Nature reviews genetics*, 13(5):303–314.

Yoshida, R., Aliatimis, G., and Miura, K. (2023a). Tropical neural networks and its applications to classifying phylogenetic trees. *arXiv preprint arXiv:2309.13410*.

Yoshida, R., Barnhill, D., Miura, K., and Howe, D. (2022a). Tropical density estimation of phylogenetic trees.

Yoshida, R., Miura, K., and Barnhill, D. (2022b). Hit and run sampling from tropically convex sets. *arXiv preprint arXiv:2209.15045*.

Yoshida, R., Miura, K., and Barnhill, D. (2022c). Hit and run sampling from tropically convex sets.

Yoshida, R., Miura, K., Barnhill, D., and Howe, D. (2022d). Tropical density estimation of phylogenetic trees. https://arxiv.org/abs/2206.04206.

Yoshida, R., Owada, T., and Miura, K. (2023b). Hit-and-run sampling for tropical convex sets and applications to ultrametric trees. In *Proceedings of the 2023 International Conference on Tropical Geometry and Applications*, pages 115–130.

Yoshida, R., Takamori, M., Matsumoto, H., and Miura, K. (2023c). Tropical support vector machines: Evaluations and extension to function spaces. *Neural Networks*, 157:77–89.

Yoshida, R., Takamori, M., Matsumoto, H., and Miura, K. (2023d). Tropical support vector machines: Evaluations and extension to function spaces. *Neural Networks*, 157:77–89.

Yoshida, R., Zhang, L., and Zhang, X. (2019). Tropical principal component analysis and its application to phylogenetics. *Bulletin of Mathematical Biology*, 81:568–597.

Zhang, C., Rabiee, M., Sayyari, E., and Mirarab, S. (2018a). Astral-iii: polynomial time species tree reconstruction from partially resolved gene trees. *BMC bioinformatics*, 19(Suppl 6):153.

Zhang, L., Naitzat, G., and Lim, L. (2018b). Tropical geometry of deep neural networks. *International Conference on Machine Learning*, pages 5824–5832.

Zuckerkandl, E. and Pauling, L. (1965). Evolutionary divergence and convergence in proteins. In *Evolving genes and proteins*, pages 97–166. Elsevier.