Tropical Logistic Regression Model on Space of Phylogenetic Trees

Georgios Aliatimis[1*], Ruriko Yoshida[2], Burak Boyacı[3], James A. Grant[4]

[1*]STOR-i Centre for Doctoral Training, Lancaster University, Lancaster, LA1 4YW, UK.
[2]Department of Operations Research, Naval Postgraduate School, 1411 Cunningham Road, Monterey, 93943, CA, USA.
[3]Management School, Lancaster University, Lancaster, LA1 4YX, UK.
[4]Department of Mathematics and Statistics, Lancaster University, Lancaster, LA1 4YX, UK.

*Corresponding author(s). E-mail(s): g.aliatimis@lancaster.ac.uk;
Contributing authors: ryoshida@nps.edu; b.boyaci@lancaster.ac.uk;
j.grant@lancaster.ac.uk;

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 paper, 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 (AUC) in numerical examples, including with data generated by the multi-species coalescent model. Theoretical properties such as statistical consistency have been proved and generalization error rates have been 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.

Keywords: coalescent model, classifications, gene trees and species trees, tropical geometry, ultrametrics

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 [1]. 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 [2]. 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 $e = \binom{m}{2}$ [3, 4]. This space is not Euclidean and even lacks convexity [4]. In fact, [3] showed that the space of phylogenetic trees is a *tropicalization* of linear subspaces defined by a system of tropical linear equations [5] and is therefore a tropical linear space.

Consequently, many researchers have applied tools from tropical geometry to statistical learning methods in phylogenomics, such as principal component analysis over the space of phylogenetic trees with a given set of leaves $[m]$ [5, 6], kernel density estimation [7], MCMC sampling [8], and support vector machines [9]. Recently, [10] 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 1 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 2. 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 2 and proved in Supplement A. Section 3 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 4 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 4.3. The paper concludes with a discussion in Section 5. The code developed and implemented for the proposed model can be found in [11].

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

1 Tropical Geometry and Phylogenetic Trees

1.1 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 [12].

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

Definition 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*

$$d_{\mathrm{tr}}(v, w) := \|v - w\|_{\mathrm{tr}} = \max_i \{v_i - w_i\} - \min_i \{v_i - w_i\}, \tag{1}$$

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

Remark 1. *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 $d_{\mathrm{tr}}(v, w) = 0$ 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 $(\mathbb{R} \cup \{-\infty\})^e/\mathbb{R}1$ achieves just that.*

Proposition 1. *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.*

1.2 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 2 (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 [13]. The proposed classification method is not restricted to equidistant trees, but all coalescent model gene trees produced in Section 4.2. are equidistant.

To conduct any mathematical analysis, a vector representation of trees is needed. A common way is to use BHV coordinates [14] 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 (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.

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

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 4 (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 1 (noted in [15])**.** *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 1, 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 5 (explained in [5, 16]) in Supplement B establishes the connection between phylogenetic trees and tropical geometry by stating that the ultrametric space is a tropical linear space.

2 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, $Y \sim$ Bernoulli$(p(x|\omega))$ where

$$p(x|\omega) = \mathbb{P}(Y = 1|x) = \frac{1}{1 + \exp\left(-h_\omega(x)\right)} = \sigma\left(h_\omega(x)\right), \tag{2}$$

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

$$l(\omega|x, y) = \frac{1}{N} \sum_{i=1}^{N} y^{(i)} \log p_\omega^{(i)} + (1 - y^{(i)}) \log(1 - p_\omega^{(i)}), \tag{3}$$

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.

2.1 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 (4) 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 1 and 2 illustrate this for non-tropical and tropical distributions respectively, and Remark 2 discusses the choice of tropical distribution in more detail.

Furthermore, the function h from (4) also minimizes the expected cross-entropy loss according to Proposition 2. Therefore, the *best model* to fit data that have been generated by tropical Laplace distribution (6) 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 $Y \sim$ 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, $Y|X \sim$ Bernoulli$(p(X))$ with $p(x) = \sigma(h(x))$, where*

$$h(x) = \log\left(\frac{r f_1(x)}{(1-r) f_0(x)}\right). \tag{4}$$

Proposition 2. *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) is $p(x) = \sigma(h(x))$, with h defined as in equation (4) of Lemma 1.*

Example 1 (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

$$h(x) = -\frac{\|x-\omega_1\|^2}{2\sigma^2} + \frac{\|x-\omega_0\|^2}{2\sigma^2} = \frac{(\omega_1-\omega_0)^T}{\sigma^2}(x-\bar{\omega}), \tag{5}$$

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

Example 2 (Tropical Laplace distribution)**.** *It may be assumed that the covariates are distributed according to the tropical version of the Laplace distribution, as presented in [8], with mean ω_Y and probability density functions*

$$f_Y(x) = \frac{1}{\Lambda}\exp\left(-\frac{d_{\mathrm{tr}}(x,\omega_Y)}{\sigma_Y}\right), \tag{6}$$

where Λ is the normalizing constant of the distribution.

Proposition 3. *In distribution (6), the normalizing factor is $\Lambda = e!\sigma_Y^{e-1}$.*

Proof. See Supplement A. □

Remark 2. *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*

$$f_{\mu,\Sigma}(x) \propto \exp\left(-\frac{1}{2}(x-\mu)^t\Sigma^{-1}(x-\mu)\right) \tag{7}$$

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 Laplacian distribution in (6) is tropicalization of the left hand side in (7) 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 [17] nicely surveys the many different definitions of tropical Gaussian distributions. Since the space of ultrametrics is a tropical linear space [3], it is natural to use tropical "linear algebra" for the definition of tropical "Gaussian" distribution defined

in (6) 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 [17], 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 Laplacian (7) 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 [7], and as the Bayes estimator [18] 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 3. 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 [19].

Example 3. *[Tropical Metric] Suppose we have equidistant trees T_1 and T_2 with leaf*

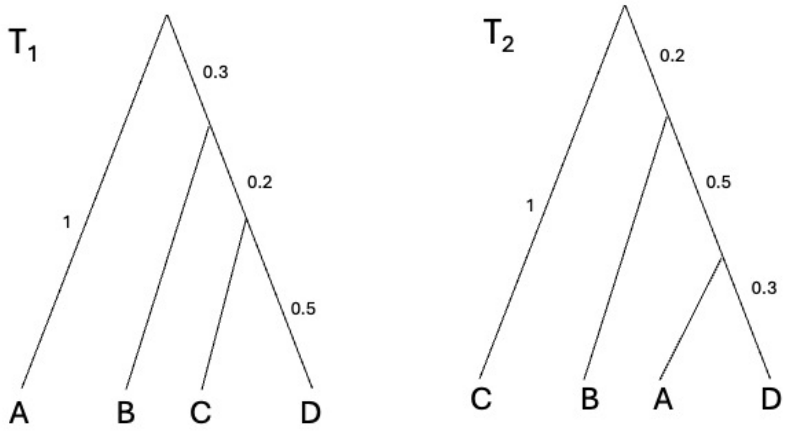

Fig. 1 Example for an interpretation of the tropical metric d_{tr} in Example 3.

labels $\{A, B, C, D\}$ shown in Fig. 1. Note that leaves A and C in T_1 and T_2 are switched. Thus, the pairwise distances from A and D in T_1 and T_2, as well as he

pairwise distances from C and D in T_1 and T_2 are the largest and second largest differences among all possible pairwise distances.

Let u be a dissimilarity may from T_1 and v be a dissimilarity map from T_2:

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

Then we have

$$u - v = (2 - 1.6, 2 - 2, 2 - 0.6, 1.4 - 2, 1.4 - 1, 1 - 2) = (0.4, 0, 1.4, -0.6, 0.4, -1).$$

Therefore

$$d_{\mathrm{tr}}(u, v) = (u - v)_{A,D} - (u - v)_{C,D}$$

which means the tropical metric measures the difference of divergence between A and D and difference of divergence between C and D.

Combining the result of Proposition 3 with Equations (4) and (6) yields

$$h_{\omega_0,\omega_1}(x) = \frac{d_{\mathrm{tr}}(x, \omega_0)}{\sigma_0} - \frac{d_{\mathrm{tr}}(x, \omega_1)}{\sigma_1} + (e - 1) \log\left(\frac{\sigma_0}{\sigma_1}\right). \tag{8}$$

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 (8) becomes

$$h_\omega(x) = \lambda \left(d_{\mathrm{tr}}(x, \omega) - c\right), \tag{9}$$

where $\lambda = (\sigma_0^{-1} - \sigma_1^{-1})$ and $\lambda c = \log(\sigma_1/\sigma_0)$. Symbolically, the expression in equation (9) 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 C in the supplement for more details. The classifier is $C(x) = \mathbb{I}(d_{\mathrm{tr}}(x, \hat{\omega}) > c)$, where $\hat{\omega}$ is the inferred estimator of ω^*. Note that the classification threshold and the probability contours $(p(x))$ are tropical circles, illustrated in Figure 2.

For the *two-species-tree model*, it is assumed that $\sigma_0 = \sigma_1$, and $\omega_0 \neq \omega_1$. Equation (8) reduces to

$$h_{\omega_0,\omega_1}(x) = \sigma^{-1}(d_{\mathrm{tr}}(x, \omega_0) - d_{\mathrm{tr}}(x, \omega_1)), \tag{10}$$

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 $y \in \{0, 1\}$. The classification boundary is the tropical bisector which is extensively studied in [20] 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 4(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.

2.2 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 2, 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 1 which is based on Proposition 4. Note that the distance metric corresponds to the geometry of the covariates. However, the arguments used in the proof of Corollary 1 do not work for distributions 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 4. *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, $d^i(X) \sim i\sigma^i \mathrm{Gamma}(n/i)$.*

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

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.

2.3 Consistency and Generalization Error

In this subsection, the consistency of the statistical estimators (in Theorem 2) and of the tropical logistic regression as a learning algorithm (in Propositions 5 and 6) 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 (11). Moreover, there is a higher misclassification rate from the more dispersed class (inequality (12)).

Theorem 2 (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* $X_i \in \mathbb{R}^e/\mathbb{R}\mathbf{1}$, $d_{\text{tr}}(X, \omega_Y)$ *being integrable and square-integrable and* $Y_i \sim$ *Bernoulli*$(S(h(X_i, (\omega^*, \sigma^*))))$. *Then,*

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

In other words, the model parameter estimator is consistent.

Proposition 5 (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 (9) and* $\hat{\omega}$ *is the estimate for* $\omega^\star$. *Define the covariate-response joint random variable* (X, Y) *with* $Z = \sigma_Y^{-1} d_{\text{tr}}(X, \omega_Y^*)$ *drawn from the same distribution with cumulative density function* F. *Then,*

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

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.*

$$\mathbb{P}(C(X) \neq Y) = \frac{1}{2}(1 - F(\sigma_1 \alpha) + F(\sigma_0 \alpha)) \tag{11}$$

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

$$\mathbb{P}(C(X) = 1 | Y = 0) < \mathbb{P}(C(X) = 0 | Y = 1). \tag{12}$$

Proposition 6 (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*

$$\mathbb{P}(C(X) \neq Y) \leq \frac{1}{2} F_Z^C(\Delta_\epsilon) + h(\epsilon), \tag{13}$$

where $\epsilon = d_{\mathrm{tr}}(\hat{\omega}_1, \omega_1^*) + d_{\mathrm{tr}}(\hat{\omega}_0, \omega_0^*)$, $2\Delta_\epsilon = (d_{\mathrm{tr}}(\omega_1^*, \omega_0^*) - \epsilon)$, 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* $\mathbb{P}(d_{\mathrm{tr}}(X, \omega_1^*)) = d_{\mathrm{tr}}(X, \omega_{-1}^*)) = 0$*. Moreover, under the conditions of Theorem 2, our proposed learning algorithm is consistent.*

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

Example 4. *The complementary cumulative distribution of* $\mathrm{Gamma}(n, \sigma)$ *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 (6) yields the following upper bound for the generalization error*

$$\frac{\Gamma\left(e-1, \frac{d_{\mathrm{tr}}(\omega_0^*, \omega_1^*)}{2\sigma}\right)}{2\Gamma(e-1)}, \tag{14}$$

under the assumptions of Proposition 6 and assuming that the estimators coincide with the theoretical parameters. This assumption is reasonable for large sample sizes and it follows from Theorem 2.

In subsequent sections, these theoretical results will guide us in implementing our model. Bounds on the generalization error from Propositions 5 and 6 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 4.

3 Optimization

As in the classical logistic regression, the parameter vectors $(\hat{\omega}, \hat{\sigma})$ maximising the log-likelihood (3), 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 [21] is revisited.

3.1 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.

$$\tilde{\omega}_n \in \arg\min_{\omega} \sum_{i=1}^{n} d_{\mathrm{tr}}(X_i, \omega). \tag{15}$$

This point is rarely unique for finite n, indeed there will often be an infinite set of Fermat-Weber points [21]. However, the proposition below gives conditions for asymptotic convergence.

Proposition 7. *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 (15). Then, $\tilde{\omega}_n \xrightarrow{p} \omega^*$ as $n \to \infty$.*

The significance of Proposition 7 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 2 and Proposition 7, $\hat{\omega}_n - \tilde{\omega}_n \xrightarrow{p} 0$ as $n \to \infty$. Furthermore, empirical evidence in Figure 5, 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 D.

There is also another another benefit of using Fermat-Weber points. Proposition 8 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 8. *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*

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

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 [21], 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 4.

Remark 3. *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 [22]. 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. [4] 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 2.

4 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.

4.1 Toy Example

In this example, a set of data points is generated from the tropical normal distribution as defined in Equation (6) 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 2 a scatterplot of of the two classes is shown, where misclassified points are highlighted. As anticipated from Proposition 5 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 (11) of Proposition 5 are 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

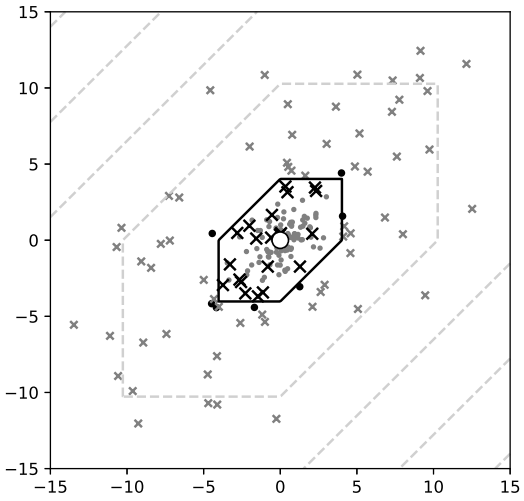

Fig. 2 Scatterplot of 200 points - 100 dots for class 0 and 100 Xs for class 1, black for misclassified and grey otherwise - imposed upon a contour plot of the probability of inclusion in class 0, where the black contour is the classification threshold. The deviation parameters used in data generation 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 $\omega_0 = \omega_1$.

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(left) verifies the bounds of Proposition 5.

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, using h as defined in equations (5) and (10) for the classical and tropical logistic regression respectively. Fig. 4 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(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 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

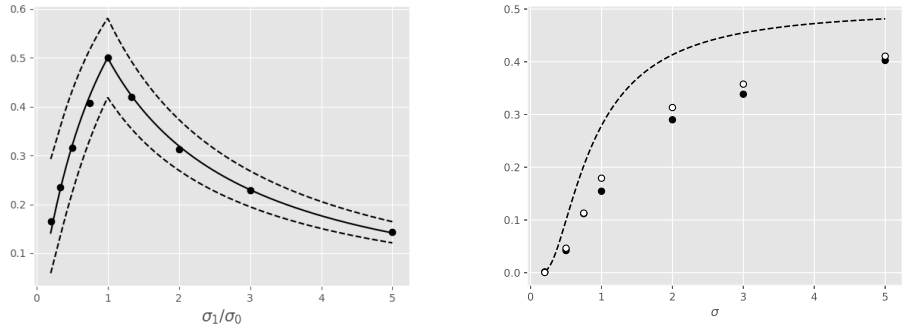

Fig. 3 (left) Generalization error for 9 different deviation ratios. The estimator $\hat{\omega} = (0.3, 0, 3)$ differs from the true parameter $\omega = (0, 0)$. The upper and lower bounds of Proposition 5 are plotted in dashed lines and the generalization error for the correct estimator $\hat{\omega} = \omega^*$ 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 (14) of Proposition 6 is plotted in dashed line.

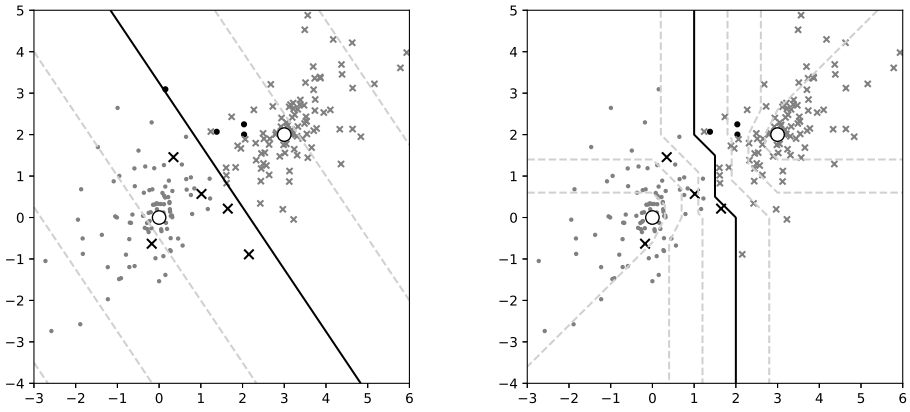

Fig. 4 Scatterplot of points - dots for class 0 and X for class 1, black for misclassified according to (left) **classical logistic regression** or (right) **tropical logistic regression**, and grey otherwise - alongside a contour plot of the probabilities, where the black contour is the classification threshold. The centres, drawn as big white dots, are $\omega_0 = (0, 0, 0), \omega_1 = (3, 2, 0)$ and $\sigma = 0.5$.

repeat this procedure 100 times to get an estimate of the mean error rate $r_N = \mathbb{E}(d_N)$. Figure 5 shows that for both methods, $r_N\sqrt{N} \to C$ as $N \to \infty$, with $C_{\mathrm{FW}} < C_{\mathrm{MLE}}$. 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 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.

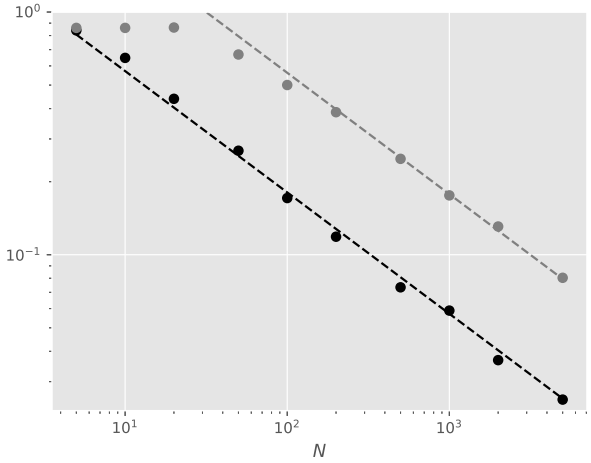

Fig. 5 Expected asymptotic error for FW points $(\tilde{\omega}_0)_N$ (in black) and MLE points $(\hat{\omega}_0)_N$ (in grey) for different values of N. Error is defined as the tropical distance from the true centre ω_0^* i.e. $d_{\text{tr}}(\omega_N, \omega_0^*)$. The dashed lines are $y \propto N^{-0.5}$, so this figure illustrates that $d_{\text{tr}}((\omega_0)_N, \omega_0^*) = \mathcal{O}_p(1/\sqrt{N})$ as $N \to \infty$.

4.2 Coalescent Model

The data that have been used in our simulations were generated under the multispecies coalescent model, using the python library `dendropy` [23]. 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.

Two distinct species trees are used, which were randomly generated under a Yule model. 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 ([14]) 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. 6, the distribution of the radius $d(X, \omega)$ as given by Proposition 4, 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 4, for both the classical and tropical Laplace distributed covariates, $d(X, \omega^*) \sim \sigma\text{Gamma}(n)$, shown in solid lines in Fig. 6, 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 Laplacian distributions produce better fits in both geometries and that the tropical

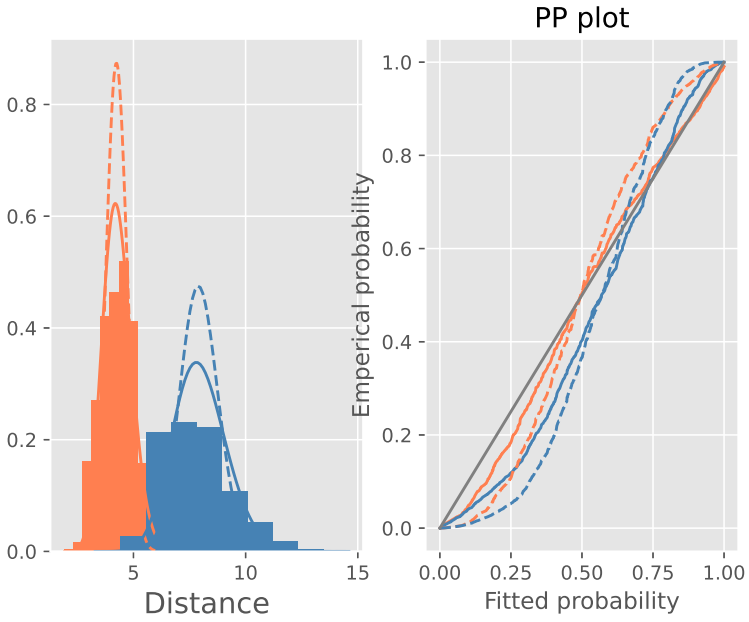

Fig. 6 (Left) Histograms of the distances of 1000 gene trees from the species trees that generated them under the coalescent model with $R = 0.7$. Coral and blue corresponds to tropical and euclidean geometries respectively. The solid and dashed lines are fitted distributions $\sigma\text{Gamma}(n)$ 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).

Laplacian fits the data best. As discussed in Section 2.2, the same analysis can not be applied to the BHV metric, because the condition of Proposition 4 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. H3 and Fig. H2 in Supplement 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. 7 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 both the BHV and classical logistic regression. Finally, note that the generalization error upper bound as given in equation (14) is satisfied but it not very tight (dashed line in Fig. 7).

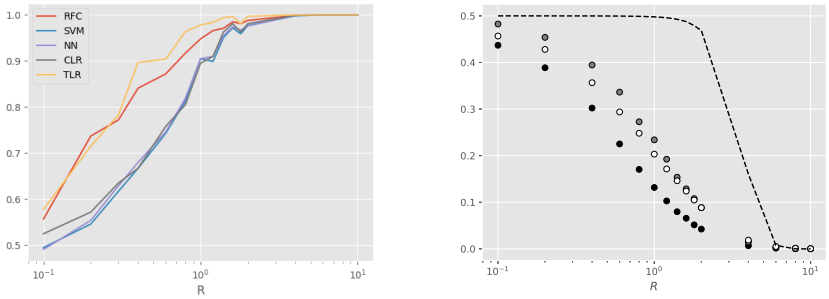

Fig. 7 (left) Average AUCs against R. Five classification models which we considered are the tropical two species-tree model (TLR), random forest classifier (RFC), support vector machines (SVM), neural networks (NN) and classical logistic regression (CLR). We used default set up for TLR, SVM, NN and 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 6.

4.3 Convergence of Mr Bayes

`Mr Bayes` ([24]) 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 [25]) 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 [26]. Figure 8 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 [27]) could be applied instead at the cost of simplicity and interpretability.

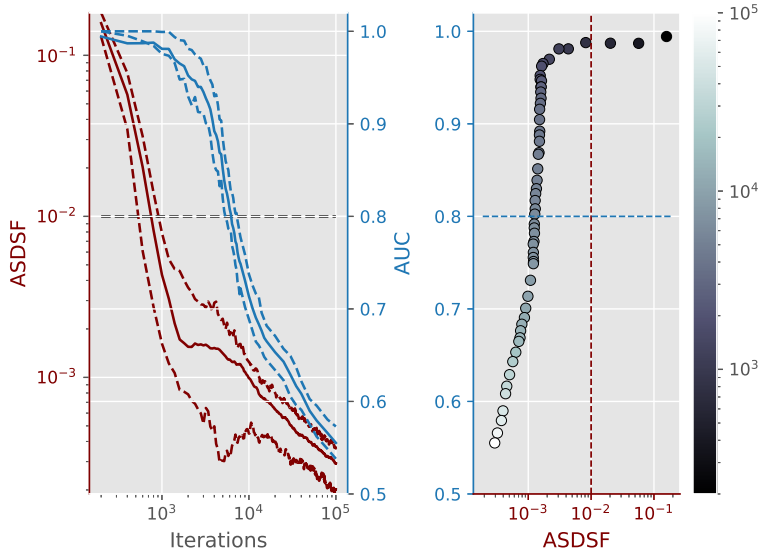

Fig. 8 (Left) Average ASDSF (in red) and AUC (in blue) values plotted against the number of iterations of the MCMC chains. The coloured dashed lines correspond to the first and third quartile. The grey dashed line indicates the `Mr Bayes` threshold for ASDSF and our provisional AUC threshold of 80%. (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.

5 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

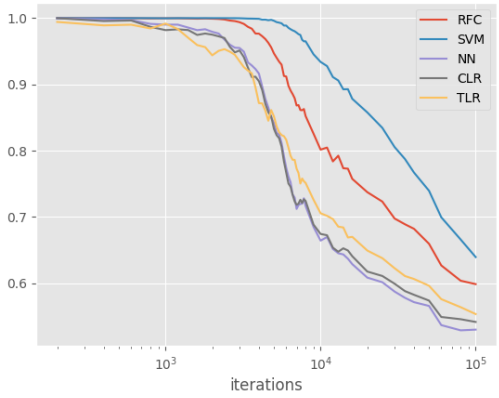

Fig. 9 Average AUC values plotted against the number of MCMC iterations for the 5 supervised learning methods considered.

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 [28].

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 Laplacian 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.

Acknowledgments. RY is partially funded by NSF Division of Mathematical Sciences: Statistics Program DMS 1916037. GA is funded by EPSRC through the STOR-i Centre for Doctoral Training under grant EP/L015692/1.

Declarations

Some journals require declarations to be submitted in a standardised format. Please check the Instructions for Authors of the journal to which you are submitting to see if

you need to complete this section. If yes, your manuscript must contain the following sections under the heading 'Declarations':

- Funding: RY is partially funded by NSF Division of Mathematical Sciences: Statistics Program DMS 1916037. GA is funded by EPSRC through the STOR-i Centre for Doctoral Training under grant EP/L015692/1.
- Conflict of interest: No conflict of interest.
- Ethics approval: NA
- Consent to participate: NA
- Consent for publication: NA
- Availability of data and materials: DRYAD with DOI: 10.5061/dryad.tht76hf65
- Code availability: DRYAD with DOI: 10.5061/dryad.tht76hf65
- Authors' contributions: GA contributed theoretical work and computations. RY directed this project. BB and JG supervised GA.

Appendix 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 2. 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 4. 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. From the proof of Proposition 4, 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 1. 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 Proposition 4, it follows that $d^i(X - \omega^*, \mathbf{0}) = d^i(X, \omega^*) \sim i\sigma^i \mathrm{Gamma}(n/i)$. 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 ($d = d_{\mathrm{tr}}$), 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 2

Theorem 3. *(Theorem 4.2.1 in [29]) 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 4. *(Lemma 2.4 in [30]) 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 1. Consider two points $x, y \in \mathbb{R}^e/\mathbb{R}\mathbf{1}$. There exists $\eta > 0$ such that

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

$$\phi_i(v) = \begin{cases} 1, & \text{if } v_i \geq v_j\ \forall j \in [e] \\ -1, & v_i < v_j\ \forall j \in [e] \backslash \{i\} \\ 0, & \text{otherwise} \end{cases}, \tag{A1}$$

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 $m := \min_{j \in [e]}\{v_j\}$,

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

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

$$\max_{j \in [e]}\{v_j + \epsilon \delta_{ij}\} = v_i + \epsilon = M + \epsilon, \tag{A2}$$

$$\min_{j \in [e]}\{v_j + \epsilon \delta_{ij}\} = m, \tag{A3}$$

and so $d_{\text{tr}}(x + \epsilon E_i, y) = d_{\text{tr}}(x, y) + \epsilon$. Note that equations (A2) and (A3) 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

$$\max_{j \in [e]}\{v_j + \epsilon \delta_{ij}\} = M, \text{ for all } \epsilon \leq M - m \tag{A4}$$

$$\min_{j \in [e]}\{v_j + \epsilon \delta_{ij}\} = v_i + \epsilon = m + \epsilon, \text{ for all } \epsilon \leq m' - m, \tag{A5}$$

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 (A4) and (A5) 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

$$\min_{j \in [e]} \{v_j + \epsilon \delta_{ij}\} = v_k = m, \text{ for all } \epsilon > 0 \tag{A6}$$

$$\max_{j \in [e]} \{v_j + \epsilon \delta_{ij}\} = M, \text{ if } \epsilon \leq M - v_i \tag{A7}$$

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 (A6) and (A7) 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 2. Consider the function $q : \mathbb{R}^e / \mathbb{R}\mathbf{1} \to \mathbb{R}$,

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

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 1, there exists $\eta_1 > 0$ such that for all $\epsilon \in [0, \eta_1]$

$$d_{\text{tr}}(x + \epsilon E_i, y) = d_{\text{tr}}(x, y) + \epsilon \phi_i(x - y). \tag{A8}$$

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

$$d_{\text{tr}}(x - \epsilon E_i, y) = d_{\text{tr}}(x, y) + \epsilon \phi_i(y - x), \tag{A9}$$

for all $\epsilon \in [0, \epsilon_0(y - x)]$. For all $\epsilon \in [0, \eta]$ where $\eta \coloneqq \min(\eta_1, \eta_2)$, equations (A8), (A9) are satisfied and so

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

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

$$\begin{aligned} & q(x + \epsilon E_i) + q(x - \epsilon E_i) - q(x) = 0 \\ & = \epsilon \left(\lambda_\alpha s_i(x - \alpha) - \lambda_\beta s_i(x - \beta) - \lambda_\gamma s_i(x - \gamma) + \lambda_\delta s_i(x - \delta) \right), \end{aligned} \tag{A10}$$

where

$$s_i(v) \coloneqq \phi_i(v) + \phi_i(-v) = \tag{A11}$$
$$\begin{cases} 2, & \text{if } v = \mathbf{0} \\ 1, & \text{if } v \neq \mathbf{0} \text{ and } v_i \text{ is the non-unique maximizer or minimizer of } v \\ 0, & \text{otherwise} \end{cases}$$

By summing equation (A10) over $i \in [e]$ and defining $s(v) = \sum_{i=1}^{e} s_i(v)$,

$$\lambda_\alpha s(x - \alpha) - \lambda_\beta s(x - \beta) - \lambda_\gamma s(x - \gamma) + \lambda_\delta s(x - \delta) = 0, \tag{A12}$$

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

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 (A12) 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 (A12) as

$$s(x - \alpha) - s(x - \beta) - s(x - \gamma) + s(x - \delta) = 0, \tag{A13}$$

Note now equation (A13) 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,

$$s(x - \alpha) = s(\epsilon E_i) = e - 1, \tag{A14}$$

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

$$s(x - \beta) = s(\alpha - \beta + \epsilon E_i) = e - 1, \tag{A15}$$

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,

$$s(x - \gamma) = s(\alpha - \gamma + \epsilon E_i) \geq 2, \tag{A16}$$

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.

$$s(x - \delta) = s(\alpha - \delta + \epsilon E_i) = 0 \tag{A17}$$

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

From equations (A14) – (A17), it is concluded that

$$s(x - \alpha) - s(x - \beta) - s(x - \gamma) + s(x - \delta) \leq -2, \tag{A18}$$

which contradicts equation (A13). 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 (A12) becomes

$$\lambda_\tau s(x - \tau) + \lambda_\upsilon s(x - \upsilon) - \lambda_\phi s(x - \phi) = 0 \tag{A19}$$

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 (A12) 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 (A12) 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

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

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. 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 2. 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, 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

$$\begin{aligned} Q(\omega, \lambda) &= \mathbb{E}_X(g(h(X, (\omega, \lambda)))) \\ &= \mathbb{E}_X(g(h(X, (\omega^*, \lambda^*))) - \frac{1}{2} \mathbb{E}_X \left(S(\xi(X)) S(-\xi(X)) [h(X, (\omega, \lambda)) - h(X, (\omega^*, \lambda^*))]^2 \right) \\ &\leq Q(\omega^*, \lambda^*), \end{aligned} \tag{A20}$$

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 (A20) 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 2 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 4 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 (A20),

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

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

Proof of Proposition 5. First, define $\Delta_0 = \{C(X) \neq 1|Y = 0\}$. By definition of $C(X)$,

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

Triangular inequality dictates that

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

and so it follows that

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

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

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

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

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

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,

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

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

$$\lim_{x \to 1} g(x) < 0. \tag{A21}$$

Finally, the derivative of g is

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

The following two inequalities

$$F'(u(x)) \geq F'(xu(x)), \tag{A22}$$

$$u'(x) + xu'(x) + u(x) \geq 0, \tag{A23}$$

imply that

$$g'(x) \leq -F'(xu(x))(u'(x) + xu'(x) + u(x)) \leq 0. \tag{A24}$$

From (A21) and (A24) it follows that $g(x) < 0$ for all $x > 1$.
For inequality (A22), 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 (A23),

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

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

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

Clearly, v is a non-negative function for $x > 1$, so inequality (A23) is satisfied. □

Proof of Proposition 6. 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

$$\mathbb{P}(C(X) \neq Y) \leq \mathbb{P}\left(D_X \geq -\epsilon\right). \tag{A25}$$

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,

$$\mathbb{P}(C(X) \neq Y) \leq \mathbb{P}\left(D_X \geq -\epsilon, Z_X \geq \Delta_\epsilon\right) \tag{A26}$$

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

$$\mathbb{P}\left(D_X \geq -\epsilon, Z_X \geq \Delta_\epsilon\right) = \mathbb{P}\left(D_{2\omega_Y^* - X} \geq -\epsilon, Z_{2\omega_Y^* - X} \geq \Delta_\epsilon\right). \tag{A27}$$

It will be proved that

$$Z_{2\omega_Y^* - X} = Z_X, \tag{A28}$$
$$D_X + D_{2\omega_Y^* - X} \leq 0, \tag{A29}$$

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 (A27),

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

and substituting it to inequality (A26),

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

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 (A28) follows from the observation that

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

For equation (A29),

$$\begin{aligned}D_{2\omega_Y^* - X} + D_X &= Z_{2\omega_Y^* - X} - d_{\mathrm{tr}}(2\omega_Y^* - X, \omega_{-Y}^*) \\ &\quad + Z_X - d_{\mathrm{tr}}(X, \omega_{-Y}^*) \\ &\overset{(A28)}{=} 2Z_{2\omega_Y^* - X} - d_{\mathrm{tr}}(2\omega_Y^* - X, \omega_{-Y}^*) - d_{\mathrm{tr}}(\omega_{-Y}^*, X) \\ &\leq 2Z_{2\omega_Y^* - X} - d_{\mathrm{tr}}(2\omega_Y^* - X, X) = 0,\end{aligned}$$

where the last inequality comes from the triangular inequality. Finally, the consistency of the learning algorithm is proved. Under the conditions of Theoreom 2, 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 afore 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 $C_{\omega_0^*,\omega_1^*}(X) = \text{sgn}(D_X)Y$, the Bayes (or irreducible) error is

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

since it is assumed that $\mathbb{P}(D_X = 0) = 0$. Using inequality A25 derived earlier, it follows that the generalization error is bounded by

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

where $\epsilon(\hat{\omega}_0, \hat{\omega}_1) = d_{\text{tr}}(\omega_0, \omega_0^*) + d_{\text{tr}}(\omega_1, \omega_1^*) \xrightarrow{\text{p}} 0.$ as the training sample size $n \to \infty$ according to Theorem 2. The complementary CDF of D_X, defined as

$$F_{D_X}^C(x) = \mathbb{P}(D_X \geq x),$$

is a continuous function and so it follows that $F_{D_X}^C(\epsilon(\hat{\omega})) \xrightarrow{\text{p}} F_{D_X}^C(0) = \text{BE}$ 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 7. 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,

$$\mathbb{E}\left(d_{\text{tr}}(X, \omega)\right) + \mathbb{E}\left(d_{\text{tr}}(2\omega^* - X, \omega)\right) > 2\mathbb{E}(d_{\text{tr}}(x, \omega^*)) \tag{A30}$$

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

$$\begin{aligned}\mathbb{E}(d_{\mathrm{tr}}(2\omega^* - X, \omega)) &= \int_{\mathbb{R}^e/\mathbb{R}\mathbf{1}} d_{\mathrm{tr}}(2\omega^* - x, \omega) f(x)\, dx \\ &= \int_{\mathbb{R}^e/\mathbb{R}\mathbf{1}} d_{\mathrm{tr}}(y, \omega) f(y)\, dy = \mathbb{E}(d_{\mathrm{tr}}(X, \omega)).\end{aligned} \tag{A31}$$

Combining equation (A31) with inequality (A30) shows that the function $Q(\omega) = \mathbb{E}(d_{\mathrm{tr}}(X, \omega))$ has a global minimum at ω^*.

From Theorem 4 (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 are satisfied and so $\tilde{\omega}_n \to \omega^*$ as $n \to \infty$. □

Proof of Proposition 8. 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

$$(\nabla f(x))_j = |\{i : M_i = j\}| - |\{i : m_i = j\}|. \tag{A32}$$

For the converse, assume that the gradient is well-defined. From equations (A8)–(A9) and following the first few sentences of Lemma 2

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

where s_j is defined in equation (A11) of Lemma 2. 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 (A32), the gradient of f vanishes at $x = \omega$ if and only if

$$|\{i : M_i = j\}| = |\{i : m_i = j\}|. \tag{A33}$$

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 (A33),

$$\begin{aligned} \sum_{i=1}^{n} v_{M_i} &= \sum_{j=1}^{e} v_j |\{i \in [n] : M_i = j\}| \\ &\overset{(A33)}{=} \sum_{j=1}^{e} v_j |\{i \in [n] : m_i = j\}| = \sum_{i=1}^{n} v_{m_i}, \end{aligned}$$

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

□

Appendix B Space of ultrametrics

Theorem 5 (explained in [5, 16])**.** *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 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).

Appendix 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 [31].

Definition 5 (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 6 (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 $e \in \mathbb{N}$,*

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

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

$$h_\omega(x) = \omega^T \odot x = \max_{l \in [e]} \{\omega_l + x_l\}. \tag{C34}$$

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 $(1, \ldots, 1)$. To resolve this issue, we fix

$$-\min_{l \in [e]} \{\omega_l + x_l\} = c, \tag{C35}$$

where $c \in \mathbb{R}$ is a constant for all observations. Combining equations (C35), (C34), and the definition of tropical distance (1),

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

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

$$h_\omega(x) = d_{\mathrm{tr}}(x, \omega) - c. \tag{C36}$$

Appendix 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: $(\tilde{\omega}_0, \tilde{\omega}_1, \hat{\sigma})$

Appendix E Fermat-Weber Point Visualization

As noted in Section 3, 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 E1 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 [21], which states that all Fermat-Weber sets are classical polytopes. The two-dimensional gradient vector, plotted as a vector field in Figure E1, always points towards the Fermat-Weber set. Therefore, the gradient algorithm should always guide us to a Fermat-Weber point.

Appendix F MLE Estimator for $\boldsymbol{\sigma}$

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

$$\hat{k} = \bar{Z}/n, \tag{F37}$$

where $\bar{Z}$ is the sample average. In our case $Z_i = d(X_i, \omega^*)$ and $k = i\sigma^i$. From Proposition 4, $Z_i \sim \text{Gamma}(n/i, i\sigma^i)$ and by substituting these parameters in equation F37, it follows that the MLE for σ is

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

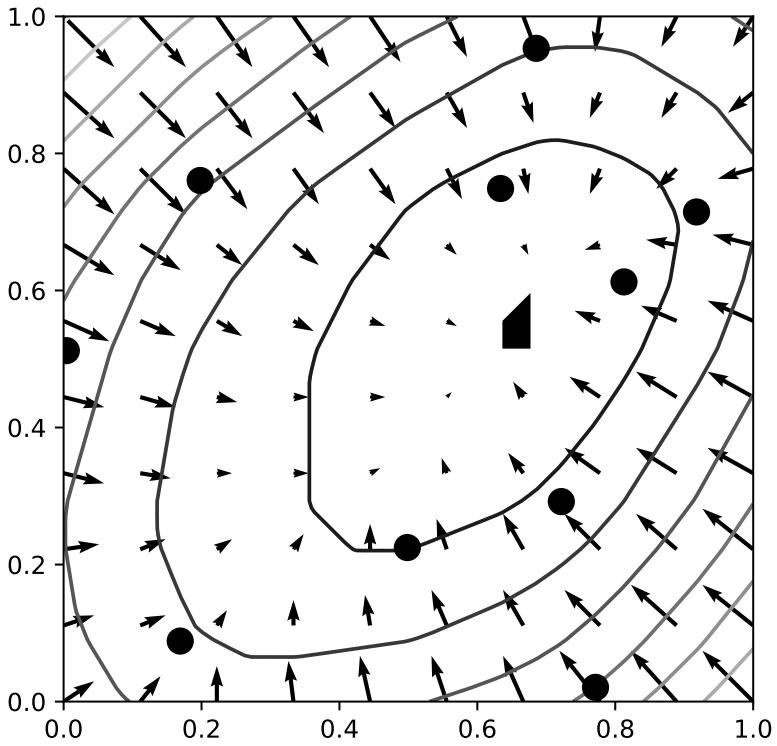

Fig. E1 Visualization of the function $f(\omega) = \sum_{i=1}^{10} d_{\text{tr}}(X_i, \omega)$ for X_i. The black circles are the datapoints $X_1, \ldots, X_{10}$, the solid lines are contours of f, the vector field is the gradient and the small black trapezoid at $(0.65, 0.55)$ is the Fermat-Weber set.

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 $\mathbb{R}^e/\mathbb{R}\mathbf{1}$, $n = e - 1$.

Appendix G Approximate BHV Logistic Regression

Similar to the tropical Laplace distribution, in [14] 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

$$h_{\omega_0,\omega_1}(x) = \lambda \left(d_{\text{BHV}}(x,\omega_0^*) - d_{\text{BHV}}(x,\omega_1^*)\right) + \log \frac{K_{\lambda,\omega_0^*}}{K_{\lambda,\omega_1^*}}. \tag{G38}$$

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 (G38) 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.

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

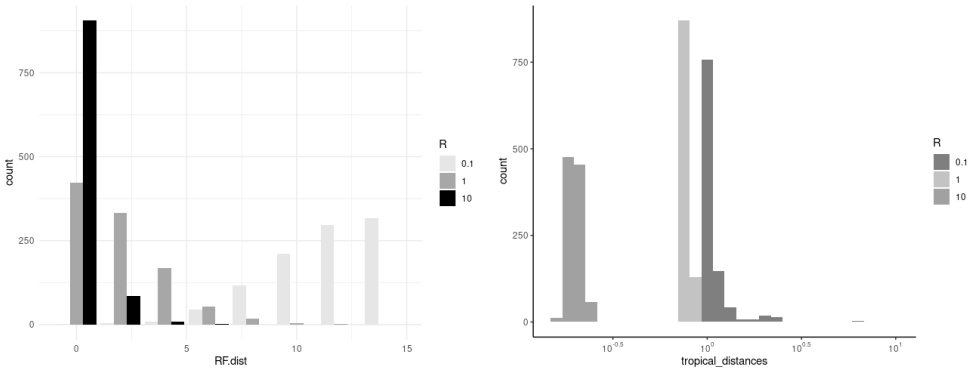

Fig. H2 (left) Robinson-Foulds distances and (right) tropical distances of inferred species trees $\hat{\omega}$ from the actual species trees ω^* for $R = 0.1, 1, 10$.

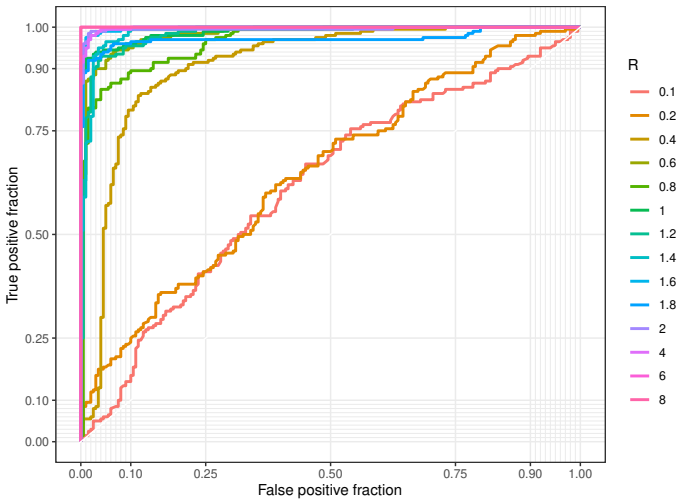

Fig. H3 ROC curves for the tropical logistic regression with different values of R. Higher the value of R is the closer an estimated ROC curve for the tropical logistic regression model gets to the point $(0, 1)$.

References

[1] Maddison, W.P.: Mesquite: a modular system for evolutionary analysis. Evolution **62**, 1103–1118 (2008)

[2] Ané, C., Larget, B., Baum, D., Smith, S., Rokas, A.: Bayesian estimation of concordance among gene trees. Mol Biol Evol. **24**(2), 412–26 (2007)

[3] Speyer, D., Sturmfels, B.: Tropical mathematics. Mathematics Magazine **82**, 163–173 (2009)

[4] Lin, B., Sturmfels, B., Tang, X., Yoshida, R.: Convexity in tree spaces. SIAM Discrete Math **3**, 2015–2038 (2017)

[5] Page, R., Yoshida, R., Zhang, L.: Tropical principal component analysis on the space of phylogenetic trees. Bioinformatics **36**(17), 4590–4598 (2020) https://doi.org/10.1093/bioinformatics/btaa564 https://academic.oup.com/bioinformatics/article-pdf/36/17/4590/34220689/btaa564.pdf

[6] Yoshida, R., Zhang, L., Zhang, X.: Tropical principal component analysis and its application to phylogenetics. Bulletin of Mathematical Biology **81**, 568–597 (2019)

[7] Yoshida, R., Miura, K., Barnhill, D., Howe, D.: Tropical Density Estimation of Phylogenetic Trees. https://arxiv.org/abs/2206.04206 (2022)

[8] Yoshida, R., Miura, K., Barnhill, D.: Hit and run sampling from tropically convex sets. arXiv preprint arXiv:2209.15045 (2022)

[9] Yoshida, R., Takamori, M., Matsumoto, H., Miura, K.: Tropical Support Vector Machines: Evaluations and Extension to Function Spaces. https://arxiv.org/abs/2101.11531 (2021)

[10] Akian, M., Gaubert, S., Qi, Y., Saadi, O.: Tropical linear regression and mean payoff games: or, how to measure the distance to equilibria. https://arxiv.org/abs/2106.01930 (2021)

[11] Aliatimis, G.: Tropical logistic regression. GitHub (2024)

[12] Maclagan, D., Sturmfels, B.: Introduction to Tropical Geometry. Graduate Studies in Mathematics, vol. 161. Graduate Studies in Mathematics, 161, American Mathematical Society, Providence, RI (2015)

[13] Maddison, W.P., Maddison, D.R.: Mesquite: a modular system for evolutionary analysis. Version 2.72. Available at http://mesquiteproject.org (2009). http://mesquiteproject.org

[14] Billera, L.J., Holmes, S.P., Vogtmann, K.: Geometry of the space of phylogenetic trees. Adv Appl Math **27**(4), 733–767 (2001)

[15] Buneman, P.: A note on the metric properties of trees. J. Combinatorial Theory Ser. B. **17**, 48–50 (1974)

[16] Ardila, F., Klivans, C.J.: The Bergman complex of a matroid and phylogenetic trees. journal of combinatorial theory. Series B **96**(1), 38–49 (2006)

[17] Tran, N.: Tropical gaussians: a brief survey. Algebraic Statistics **11**(2), 155–168 (2020)

[18] Huggins, P.M., Li, W., Haws, D., Friedrich, T., Liu, J., Yoshida, R.: Bayes Estimators for Phylogenetic Reconstruction. Systematic Biology **60**(4), 528–540 (2011) https://doi.org/10.1093/sysbio/syr021 https://academic.oup.com/sysbio/article-pdf/60/4/528/24555331/syr021.pdf

[19] Garba, M.K., Nye, T.M.W., Lueg, J., Huckemann, S.F.: Information geometry for phylogenetic trees. J. Math. Biol. **81**(19) (2021)

[20] Criado, F., Joswig, M., Santos, F.: Tropical bisectors and Voronoi diagrams. Foundations of Computational Mathematics, 1–38 (2021)

[21] Lin, B., Yoshida, R.: Tropical Fermat–Weber points. SIAM Journal on Discrete Mathematics **32**(2), 1229–1245 (2018)

[22] Com˘ aneci, A., Joswig, M.: Tropical medians by transportation. Math. Program

205, 813–839 (2023)

[23] Sukumaran, J., Holder, M.T.: Dendropy: a python library for phylogenetic computing. Bioinformatics **26**(12), 1569–1571 (2010)

[24] Huelsenbeck, J.P., Ronquist, F., Nielsen, R., Bollback, J.P.: Bayesian inference of phylogeny and its impact on evolutionary biology. science **294**(5550), 2310–2314 (2001)

[25] Lakner, C., Van Der Mark, P., Huelsenbeck, J.P., Larget, B., Ronquist, F.: Efficiency of Markov chain Monte Carlo tree proposals in bayesian phylogenetics. Systematic biology **57**(1), 86–103 (2008)

[26] Ronquist, F., Huelsenbeck, J.P., Mark, P.: MrBayes 3.1 Manual (2005)

[27] Yoshida, R., Takamori, M., Matsumoto, H., Miura, K.: Tropical support vector machines: Evaluations and extension to function spaces. Neural Networks **157**, 77–89 (2023) https://doi.org/10.1016/j.neunet.2022.10.002

[28] Yoshida, R., Aliatimis, G., Miura, K.: Tropical neural networks and its applications to classifying phylogenetic trees. arXiv preprint arXiv:2309.13410 (2023)

[29] Bierens, H.J.: Topics in Advanced Econometrics: Estimation, Testing, and Specification of Cross-section and Time Series Models. Cambridge University Press, ??? (1996)

[30] Newey, W.K., McFadden, D.: Large sample estimation and hypothesis testing. Handbook of econometrics **4**, 2111–2245 (1994)

[31] Pin, J.-E.: Tropical semirings. Cambridge Univ. Press, Cambridge (1998)