Fourier transform inequalities for phylogenetic trees

Frederick A. Matsen

Department of Statistics

University of California, Berkeley

367 Evans Hall #429

Berkeley, CA 94720-3860

USA

http://www.stat.berkeley.edu/~matsen/

October 29, 2018

Abstract

Phylogenetic invariants are not the only constraints on site-pattern frequency vectors for phylogenetic trees. A mutation matrix, by its definition, is the exponential of a matrix with non-negative off-diagonal entries; this positivity requirement implies non-trivial constraints on the site-pattern frequency vectors. We call these additional constraints "edge-parameter inequalities." In this paper, we first motivate the edge-parameter inequalities by considering a pathological site-pattern frequency vector corresponding to a quartet tree with a negative internal edge. This site-pattern frequency vector nevertheless satisfies all of the constraints described up to now in the literature. We next describe two complete sets of edge-parameter inequalities for the group-based models; these constraints are square-free monomial inequalities in the Fourier transformed coordinates. These inequalities, along with the phylogenetic invariants, form a complete description of the set of site-pattern frequency vectors corresponding to *bona fide* trees. Said in mathematical language, this paper explicitly presents two finite lists of inequalities in Fourier coordinates of the form "monomial ≤ 1," each list characterizing the phylogenetically relevant semialgebraic subsets of the phylogenetic varieties.

1 Introduction

The Bayesian and maximum-likelihood methods in phylogenetics can be classified as "model based." That is, at some stage in the analysis, one assumes a mutation model and calculates the likelihood of the observed data for a given tree and set of model parameters. We will call the set of site-pattern frequency vectors generated on a fixed tree by a mutation model under legal parameter

settings a "tree image." One of the main goals of the emerging field of phylogenetic geometry [1–5] is to locate these tree images in site pattern frequency space. Such work is foundational to understanding when model-based phylogenetics does and does not succeed.

The mutation models for sequences evolving on a tree are typically given in terms of nucleotide mutation models, which are stochastic matrices giving the probability of various mutations at an arbitrary site. One such matrix is associated with each edge; consequently one multiplies matrices along paths in the tree to get the mutation matrix along that path. Because a series of matrix multiplications is polynomial in the entries of the matrices, one can consider the tree image as a subset of an affine variety.

It is then natural to apply the well-developed tools of algebraic geometry to analyze these varieties. In particular, there has been a flourishing of interest in the corresponding ideals of these varieties; in the present setting these are called "phylogenetic invariants" [1,3,5–7]. Although not completely understood for all models, a considerable amount of beautiful work has been done on these invariants; a very nice overview has been published in [8].

One can then formulate a constrained optimization problem by optimizing the likelihood function across the set of site-pattern frequency vectors constrained to satisfy the phylogenetic invariants. This is the view taken by [9] (equation (3)) where it is called the maximum likelihood problem. Another article [5] says "exact computation of maximum likelihood estimates...can be formulated...as a constrained optimization problem where the probabilities are the decision variables and the phylogenetic invariants are the constraints." A similar statement has been made in a review article concerning the use of phylogenetic invariants for tree reconstruction [10].

These statements may be confusing for computational biologists thinking of phylogenetic trees as descriptions of mutational processes occurring in the evolutionary past. Indeed, there are solutions to the phylogenetic invariants sitting in the probability simplex which do not correspond to any reasonable assignment of branch lengths (or, more generally, edge parameters) to a tree. In the language of algebraic geometry, the tree image is not equal to its Zariski closure intersected with the probability simplex. This observation is not original to this paper: the authors of [2] define a useful notion of "biologically meaningful" solutions to the phylogenetic invariants. Their criterion is satisfied if the Fourier transform of the mutation matrices have non-negative diagonal entries. Positivity of Fourier transforms is indeed a necessary condition for a mutation matrix to come from a model (see Observation 2.3), but is not sufficient as we demonstrate below in our motivating example.

Our simple observation is this: *mutation matrices are the result of a continuous time Markov process operating for some non-negative period of time.* This fact is implicit in any description of mutation as a process in terms of rates, for example in the original description of the Kimura models [11]. In the notation of Markov processes,

$$P^{(e)} = \exp\left(t_e Q^{(e)}\right) \tag{1}$$

where $P^{(e)}$ is the mutation matrix for an edge e, $t_e \geq 0$ is elapsed time, and $Q^{(e)}$ is the mutation rate matrix. In this setting $Q^{(e)}$ must be a "Q-matrix", i.e. have non-negative off diagonal entries and zero row sums [12].

The observation (1) implies a collection of nontrivial square-free monomial inequalities in the Fourier transformed probability space which ensure that a solution to a complete set of phylogenetic invariants indeed corresponds to a *bona fide* tree. This paper develops a complete set of such inequalities; we call them "edge-parameter inequalities."

First we present a very simple motivating example on the quartet tree to illustrate the need for edge-parameter inequalities. This example has a negative internal branch length, or, said another way, the mutation rate matrix along that edge contains negative off-diagonal entries. Despite this nonsensical setup, the associated site-pattern frequency vector satisfies the phylogenetic invariants and sits in the probability simplex. Furthermore, the parameters satisfy the useful "biologically meaningful" criterion of [2], which as noted is necessary but not sufficient for a tree to have positive edge parameters. For our example we assume the two-state symmetric (CFN) model with uniform distribution at the root, labeling the two states 0 and 1. In the CFN model, there is only a single parameter per edge, called the branch length. It is the amount of time which we allow our binary Poisson mutation process to run, thus the probability that the endpoints of an edge are in different states is $0.5(1 - \exp(-2\gamma))$ for an edge of length γ. Let $\theta = \exp(-2\gamma)$; the Fourier transform [13] of the mutation matrix of length γ is thus $\mathrm{diag}(1, \theta)$.

Our motivating example is as follows: consider the tree on taxa $1, 2, 3$ and 4 with the $12|34$ split. Make each pendant edge of length γ and internal edge of length $-\gamma$. Thus formally, by the above, the off-diagonal entries of the mutation rate matrix for the internal edge will be negative. We now show that if $\gamma > 0.60938$ then the expected site-pattern frequency vector for this tree will satisfy all of the restrictions described up to now in the literature.

With the above notation, the nontrivial entry of the Fourier transform of the mutation matrix will be θ for the pendant edges and θ^{-1} for the internal edges. In this and the following sections, we use $\mathbf{p}$ to denote points of the probability simplex and $\mathbf{q}$ to denote points of the Fourier transform of the probability simplex. We will call the $\mathbf{p}$ "site-pattern frequency vectors" and the image of the probability simplex under the Fourier transform "q-space." We will index $\mathbf{p}$ and $\mathbf{q}$ with taxon state vectors $\mathbf{g}$.

We use Hadamard conjugation to compute $\mathbf{q}$ for the pathological tree. The formulation for general group-based models is given in (4), but for the CFN model the calculation of $\mathbf{q}$ is quite simple. To find a given $q_{\mathbf{g}}$, first let $S_{\mathbf{g}}$ be the set of all taxa in state 1 according to $\mathbf{g}$. Second, let $E_{\mathbf{g}}$ be the set of edges in the (unique) collection of disjoint paths connecting the taxa in $S_{\mathbf{g}}$ to each other. Then $q_{\mathbf{g}}$ is simply the product of all nontrivial entries of the Fourier transform of the mutation matrices for edges in $E_{\mathbf{g}}$ [13]. For example, the path collection corresponding to $\mathbf{g} = 1010$ is the single path connecting taxa 1 and 3, going through the internal edge. Thus $q_{1010} = \theta \cdot \theta^{-1} \cdot \theta = \theta$. All of the other similar calculations are reported in Table 1. An application of the inverse

3

$\mathbf{g}$	$q_{\mathbf{g}}$	$8 \cdot p_{\mathbf{g}}$
0000	1	$1 + 4\theta + 2\theta^2 + \theta^4$
1001	θ	$1 - \theta^4$
0101	θ	$1 - \theta^4$
1100	θ^2	$1 - 4\theta + 2\theta^2 + \theta^4$
0011	θ^2	$1 - \theta^4$
1010	θ	$(1 - \theta^2)^2$
0110	θ	$(1 - \theta^2)^2$
1111	θ^4	$1 - \theta^4$

Table 1: Site pattern frequencies and their Fourier transforms for the example mentioned in the text.

Fourier transform gives the **p**. Note that because our root distribution is taken to be uniform, the Fourier transform of the root distribution is nonzero only at the identity. Thus the only nonzero $q_{\mathbf{g}}$ are those for which the $\mathbb{Z}_2$ sum of the components of **g** equals zero.

It is clear that in Table 1 all $p_{\mathbf{g}}$ are positive for $0 \le \theta \le 1$ with the possible exception of p_{1100}. One can ensure positivity of p_{1100} by choosing $0 < \theta < 0.2955$, corresponding to a branch length $\gamma > 0.60938$. We fix such a choice of θ, which ensures that **p** sits in the probability simplex. (Note that a less stringent constraint on the branch lengths could be achieved by taking the absolute value of the internal branch length to be smaller than the pendant branch lengths.) Because our **q** comes from Hadamard conjugation, it satisfies the two phylogenetic invariants in this setting: $q_{1001} \cdot q_{0110} = q_{1010} \cdot q_{0101}$ and $q_{0000} \cdot q_{1111} = q_{1100} \cdot q_{0011}$. Furthermore, the diagonal entries of the Fourier transform of the mutation matrices (i.e. 1, θ and θ^{-1}) are positive for any real $\gamma < 0$, and thus the mutation parameters satisfy the two-state analog of the 'biologically meaningful" criterion of [2]. However, this **q** came from a phylogenetic tree with a negative internal edge. Thus the example begs the question of what conditions should be put on site-pattern frequency vectors or their Fourier transforms so that one can be assured that the corresponding trees are well-formed.

This paper describes the set of "edge-parameter inequalities" and shows that they are the exact conditions needed, namely that any solution of the phylogenetic invariants for a given tree which satisfies these inequalities is guaranteed to come from a tree with non-negative edge parameters. For example, an edge-parameter inequality for the internal edge of the quartet tree is

$$q_{0000}\, q_{1111}\, q_{1100}\, q_{0011} \;\ge\; q_{1010}\, q_{0101}\, q_{1001}\, q_{0110} \tag{2}$$

which is equivalent to the inequality $1 \ge \theta^4$ or $\gamma \ge 0$. Thus (2) specifically rules out the pathological example above.

We will describe two distinct versions of the edge-parameter inequalities. The first version is derived by considering paths in the tree and thus will be called the "path" edge-parameter inequalities. This version is relatively simple

to write down, involving two monomials of degree at most four for the two-state models and two monomials of degree at most six for the four-state models. We note that as this set of inequalities is derived on trees, they are only meaningful for $\mathbf{q}$ which satisfy a complete set of phylogenetic invariants for a tree.

Next we present the second version of the inequalities; these inequalities derive directly from the Székely-Steel-Erdős Fourier conjugation equation [14]. Because they are given directly by Fourier conjugation, we call these inequalities the "canonical" edge-parameter inequalities. These inequalities for group G-based models on trees of m taxa carve out a subset of q-space which we denote $Y_{G,m}$. The set of $\mathbf{q}$'s corresponding to a given m-taxon tree is the set of solutions to that tree's phylogenetic invariants intersected with $Y_{G,m}$.

We then investigate some properties of $Y_{G,m}$. The set $Y_{G,m}$ is the subset of q-space which corresponds precisely to the $\mathbf{q}$ of splits networks with non-negative split parameters using an extension of the model of [15]; thus it is contractible. It is not convex. Furthermore, the $\mathbf{q}$ corresponding to phylogenetic trees sit on the boundary of $Y_{G,m}$, thus the complete space of phylogenetic "oranges" [4,16] for group-based models lives on this boundary.

Before getting into details, we would like to note that the idea of constraint inequalities goes back to the remarkable paper of Cavender and Felsenstein [3]. Indeed, they anticipate such inequalities, the (phylogenetic) Fourier transform, and problems with phylogenetic mixtures. Our paper can be seen as a completion of their investigation of phylogenetic inequalities for the group-based models.

2 Technical introduction

In this section we fix notation and state two versions of Fourier conjugation. The application of discrete Fourier transform ideas to phylogenetics was pioneered in [17, 18] for the CFN model, then generalized to group-based models in [14] and [19]. Our notation combines that of [5] and [15]. We note that because Fourier conjugation is our primary tool, we will only be considering group-based mutation models (defined below), in particular $G = \mathbb{Z}_2$ or $\mathbb{Z}_2 \times \mathbb{Z}_2$.

As stated in the introduction, the simple observation of this paper is that mutation transition matrices come from continuous-time Markov processes. Thus the mutation matrices $P^{(e)}$ must satisfy (1) for each edge e, with t_e and the off-diagonal elements of $Q^{(e)}$ being non-negative. We allow the rate matrices $Q^{(e)}$ to vary from edge to edge; thus we can (and do) incorporate t_e into $Q^{(e)}$ and so assume $t_e = 1$ for any e. We call the resulting entries of the mutation rate matrices $Q^{(e)}$ for an edge "edge parameters." We note that in phylogenetic practice one often assumes a fixed rate matrix Q for the whole tree and the only parameters of a given edge are the branch lengths t_e; here we make no such restriction.

Fourier conjugation applies to the "group-based models." Each state in such a model is uniquely labeled with an element of an Abelian group. We will write our group G additively, with 0 denoting the identity element. The essential

point in the definition of a group-based model is that such that the rate of transition from state g to h is only a function of the difference of g and h in G. Fixing an edge e, we write

$$Q_{g,h}^{(e)} = \psi^{(e)}(h - g)$$

where $Q^{(e)}$ denotes the mutation rate matrix along an edge e and $\psi^{(e)}$ is an arbitrary $|G|$-vector with components summing to zero such that $\psi^{(e)}(g) \geq 0$ for $g \neq 0$. The group-based models considered in the literature are also time reversible, i.e. one requires that $Q_{g,h}^{(e)} = Q_{h,g}^{(e)}$, which is equivalent to $\psi^{(e)}(g) = \psi^{(e)}(-g)$. Because exponentiation preserves symmetries of the matrices, we will also have

$$P_{g,h}^{(e)} = f^{(e)}(h - g)$$

for some probability $|G|$-vector $f^{(e)}$. Time reversibility similarly implies $f^{(e)}(g) = f^{(e)}(-g)$.

The discrete Fourier transform is constructed via the "dual group" of an Abelian group. The elements of $\hat{G}$, the dual group to G, are the homomorphisms of G to the multiplicative group of complex numbers of magnitude one. The groups G and $\hat{G}$ are isomorphic; such an isomorphism is canonical after choosing an identification of G with a direct product of finite cyclic groups. We make such a choice, and because of the resulting isomorphism we will use the same letters $g, h, \ldots$ to denote elements of G and $\hat{G}$. However, we will follow [15] in using "hat" for the application of an element of the dual group, such that $\hat{g}(h)$ is the application of $g \in \hat{G}$ to $h \in G$. (This conflicts with traditional notation for Fourier transform; we will use "check" for this purpose as defined below.) We also note that because G is isomorphic to a direct product of cyclic groups we have $\hat{g}(h) = \hat{h}(g)$.

The Fourier transform of a function $a : G \to \mathbb{C}$ is

$$\check{a}(g) := \sum_{h \in G} \hat{g}(h) a(h).$$

By the definitions $\check{f}^{(e)}(0) = 1$ for any e. Note

$$
\begin{aligned}
\check{f}^{(e)}(-g) &= \sum_{h \in G} \widehat{-g}(h) f^{(e)}(h) = \sum_{h \in G} \hat{g}(-h) f^{(e)}(h) \\
&= \sum_{h \in G} \hat{g}(h) f^{(e)}(-h) = \sum_{h \in G} \hat{g}(h) f^{(e)}(h) = \check{f}^{(e)}(g)
\end{aligned}
\tag{3}
$$

where the fourth equality is by time reversibility. By the definition of the Fourier transform, $\check{a}(-g) = \overline{\check{a}(g)}$ for any real-valued function a. Thus the fact that $\check{f}^{(e)}(g) = \check{f}^{(e)}(-g)$ is equivalent to the fact that $\check{f}^{(e)}(g)$ is real.

The formulas for the phylogenetic Fourier transform are simplified by re-rooting the tree at a leaf, which eliminates the need for a special root distribution [5, 14]. Specifically, we extend an edge from the root terminating in a new leaf; the previous root distribution is then replaced by a transition matrix along the

new edge. Thus, without loss of generality, we assume our given tree $\mathscr{T}$ on m leaves is rooted at a leaf and that the root distribution puts all mass at the identity.

Phylogenetic Fourier conjugation is an invertible transformation between the collection of edge parameters $\psi^{(e)}(g)$ and the corresponding site-pattern frequency vector for a given tree. This site-pattern frequency vector is the joint distribution of states at the leaves defined as follows. Start at the root, and move towards the leaves, changing state along an edge e according to $P^{(e)}$. The induced joint distribution on the leaves will be denoted $\mathbf{p}$ where the component $p_{\mathbf{g}}$ of $\mathbf{p}$ is the probability of seeing $\mathbf{g} \in G^m$ by the above process.

The Fourier transform of the $\mathbf{p}$ vector using the group G^m will be denoted $\mathbf{q}$. The matrix representation of the Fourier transform for the group G will be denoted K, i.e. $K_{g,h} := \hat{g}(h)$ for any $g, h \in G$. The analogous matrix for G^m will be denoted H. Note that H is the m-fold Kronecker product of K. In this notation, $\mathbf{q} = H\mathbf{p}$. We note that when K (and thus H) is a matrix with entries ± 1, the Fourier transform is often called the Hadamard transform.

Following [5], use $\Lambda(e)$ to denote the set of leaves i such that the path from i to the root goes through e; $\Lambda(e)$ can be thought of the set of leaves "below" e. We also define

$$^*g_e = \sum_{i \in \Lambda(e)} g_i.$$

The vector $^*\mathbf{g}$ is a natural lift of a $\mathbf{g} \in G^m$ to an assignment of G to the edges of the tree. We will be using two versions of Fourier conjugation. In this notation, version one can be written

Theorem 2.1 (Hendy, 1989 [18]; Evans and Speed, 1993 [19])**.**

$$q_{\mathbf{g}} = \prod_{e \in E} \check{f}^{(e)}(^*g_e). \tag{4}$$

The second version of the edge-parameter inequalities will use a different version of the Fourier conjugation. In order to express this second version, we state the following lemma:

Lemma 2.2.
$$\check{f}(h) = \exp\left(\check{\psi}(h)\right).$$

Proof. We begin as for Lemma 17.2 of [15] (though for right rather than left eigenvalues),

$$
\begin{aligned}
(QK)_{g,h} &= \sum_{x \in G} \psi(x - g)\hat{x}(h) = \sum_{y \in G} \psi(y)\widehat{y + g}(h) \\
&= \hat{g}(h) \sum_{y \in G} \psi(y)\hat{y}(h) = K_{g,h}\,\check{\psi}(h).
\end{aligned} \tag{5}
$$

Thus the hth column of K is a right eigenvector of Q with eigenvalue $\check{\psi}(h)$. The same argument with f in place of ψ shows that the hth column of K is a right

eigenvector of P with eigenvalue $\check{f}(h)$. However, $P = \exp(Q)$ so the eigenvalues of P are the exponentials of the corresponding eigenvalues of Q. $\qquad\square$

As noted in the discussion after (3), $\check{\psi}(g)$ is real for any g. Thus Lemma 2.2 implies

Observation 2.3. *Any edge with real edge parameters will have real and non-negative Fourier transform $\check{f}^{(e)}$.*

Thus any tree with non-negative edge parameters has "biologically meaningful" parameters in the language of [2], though the converse does not hold. We also note that by (4) the $q_{\mathbf{g}}$ are real; thus the logarithm in (7) retains its usual meaning as a mapping between real numbers.

We will now present a second version of Fourier conjugation. By Lemma 2.2 and the definition of Fourier transform,

$$\psi(h) = \left[K^{-1} \log K f \right]_h \tag{6}$$

where the subscript h denotes the h component of the vector. The following theorem is Theorem 6 of [14] in the presence of (6).

Theorem 2.4 (Székely, Steel, and Erdős, 1993)**.** *Let $\rho(e, h)$ be the element of G^m which assigns h to all leaves in $\Lambda(e)$ and 0 to all others. Then*

$$\psi^{(e)}(h) = \left[H^{-1} \log \mathbf{q} \right]_{\rho(e,h)}. \tag{7}$$

Note that the log in equation (7) is entry-wise.

3 Fourier transform inequalities: path version

In this section we show first that one can very easily extract specific $\check{f}^{(e)}(g)$ terms by taking ratios of certain $q_{\mathbf{g}}$ terms. Then basic inequalities for the $\check{f}^{(e)}(g)$ terms will lead to inequalities in the $q_{\mathbf{g}}$. Let $\mathsf{p}(i, j)$ be the set of edges on the path between nodes i and j in the tree (i and j may or may not be leaves). Now define

$$F(i, j; g) = \prod_{e \in \mathsf{p}(i,j)} \check{f}^{(e)}(g).$$

We record the following facts for future use:

Lemma 3.1.

(i) *Let ν be a node on the path from i to j in a tree. Then*

$$F(i, \nu; g) \cdot F(\nu, j; g) = F(i, j; g).$$

(ii) $F(i, j; g) = F(j, i; g).$

8

(iii) $F(i, j; g) = F(i, j; -g)$.

Proof. Parts (i) and (ii) are clear from the definition. Equation (3) implies (iii). □

The following fact is a simple application of the above lemma and Theorem 2.1.

Lemma 3.2. *Let i and j be leaves and let $\mathbf{g}$ have $g_i = h$, $g_j = -h$ and all other components zero. Then $q_{\mathbf{g}} = F(i, j; h)$.*

The first identity is for pendant edges. Denote the set of leaves by $\mathscr{L}$.

Proposition 3.3. *Given some pendant edge e, let i denote the leaf on e and let ν be the internal node on e. Pick j and k any leaves distinct from i such that the path $\mathsf{p}(j, k)$ contains ν. Let $w(g_i, g_j, g_k) \in G^{\mathscr{L}}$ assign state g_x to leaf x for $x \in \{i, j, k\}$ and the identity to all other leaves. Then*

$$\left[\check{f}^{(e)}(h) \right]^2 = \frac{q_{w(h,-h,0)} \cdot q_{w(-h,0,h)}}{q_{w(0,-h,h)}}. \tag{8}$$

Proof. Lemmas 3.1 and 3.2 show

$$q_{w(h,-h,0)} = \check{f}^{(e)}(h) \cdot F(\nu, j; h)$$
$$q_{w(-h,0,h)} = \check{f}^{(e)}(h) \cdot F(\nu, k; h)$$
$$q_{w(0,-h,h)} = F(\nu, j; h) \cdot F(\nu, k; h).$$

□

A similar proof implies the next identity, which is for internal edges.

Proposition 3.4. *Pick some internal edge e; say the two nodes on either side of e are ν and ν'. Choose i, j (resp. i', j') such that $\mathsf{p}(i, j)$ (resp. $\mathsf{p}(i', j')$) contains ν but not ν' (resp. ν' but not ν). Let $z(g_i, g_j, g_{i'}, g_{j'}) \in G^{\mathscr{L}}$ assign state g_x to leaf x for $x \in \{i, j, i', j'\}$ and the identity to all other leaves. Then*

$$\left[\check{f}^{(e)}(h) \right]^2 = \frac{q_{z(h,0,-h,0)} \cdot q_{z(0,-h,0,h)}}{q_{z(h,-h,0,0)} \cdot q_{z(0,0,-h,h)}}. \tag{9}$$

Now, constraints on the $\check{f}^{(e)}(h)$ will imply inequalities in the $q_{\mathbf{g}}$. Such non-trivial constraints exist; we review these constraints now for the usual group based models. First we investigate the two-state symmetric (CFN) model, which was described in the introduction. There is only one non-trivial component $\check{f}^{(e)}(1)$ of the Fourier transform along an edge, which is $\exp(-2\gamma(e))$, where $\gamma(e)$ is the "branch length" of that edge. Now $0 \leq \gamma(e)$ is equivalent to

$$\check{f}^{(e)}(1) \leq 1. \tag{10}$$

Inserting the values for $\check{f}^{(e)}(1)$ from Propositions 3.3 and 3.4 into this equation give the edge-parameter inequalities for each edge. In summary,

Proposition 3.5. *Assume that* $\mathbf{q}$ *is the* $\mathbb{Z}_2$*-Fourier transform of a site-pattern frequency vector under the CFN model. If* $\mathbf{q}$ *satisfies a complete set of phylogenetic invariants for a tree* $\mathscr{T}$ *and a set of inequalities gained by substituting an instance of (8) or (9) into the square of (10) for each edge e of* $\mathscr{T}$*, then* $\mathbf{q}$ *is the expected site-pattern frequency vector of* $\mathscr{T}$ *for some assignment of non-negative branch lengths to* $\mathscr{T}$*. Conversely, any tree with non-negative branch lengths will satisfy such a set of inequalities.*

As a quick application, we demonstrate how these inequalities exclude the pathological example described in the introduction. For the internal edge of this quartet tree under the CFN model, we should have

$$\frac{q_{1010}q_{0101}}{q_{1100}q_{0011}} = \left[\check{f}^{(e)}(1)\right]^2 \leq 1.$$

However, by substituting in values from Table 1 the above ratio is θ^{-2}, which is greater than one.

For the four-state models we will only discuss the Kimura three parameter (K3P) model. It is the most general group-based four-state model; results for this model extend to less general models by choosing transition matrices with extra symmetries. The K3P model is associated with the group $\mathbb{Z}_2 \times \mathbb{Z}_2$. Thus K for this model is the Hadamard matrix of order four, which is the Kronecker product of two Hadamard matrices of order two. We make the identifications

$$A = (0,0) \quad C = (1,0) \quad G = (0,1) \quad T = (1,1). \tag{11}$$

We write the column vector ψ as

$$[-(\psi(C) + \psi(G) + \psi(T)), \psi(C), \psi(G), \psi(T)]^T$$

Then by Lemma 2.2 we have that $\check{f}^{(e)}(A) = 1$ and

$$\begin{aligned} \check{f}^{(e)}(C) &= \exp(-2(\psi(C) + \psi(T))) \\ \check{f}^{(e)}(G) &= \exp(-2(\psi(G) + \psi(T))) \\ \check{f}^{(e)}(T) &= \exp(-2(\psi(C) + \psi(G))). \end{aligned} \tag{12}$$

The following equations are equivalent to requiring $\psi(C)$, $\psi(G)$, and $\psi(T)$ to be non-negative via (12):

$$\check{f}^{(e)}(C)\check{f}^{(e)}(T) \leq \check{f}^{(e)}(G) \tag{13}$$

$$\check{f}^{(e)}(G)\check{f}^{(e)}(T) \leq \check{f}^{(e)}(C) \tag{14}$$

$$\check{f}^{(e)}(C)\check{f}^{(e)}(G) \leq \check{f}^{(e)}(T). \tag{15}$$

In summary,

Proposition 3.6. *Assume that* $\mathbf{q}$ *is the* $\mathbb{Z}_2 \times \mathbb{Z}_2$ *Fourier transform of a site-pattern frequency vector under the K3P model. If* $\mathbf{q}$ *satisfies a complete set of*

phylogenetic invariants for a tree $\mathscr{T}$ and a set of inequalities gained by substituting an instance of (8) or (9) into the square of (13), (14), and (15) for each edge e of $\mathscr{T}$, then $\mathbf{q}$ is the expected site-pattern frequency vector of $\mathscr{T}$ for some assignment of non-negative edge parameters to $\mathscr{T}$. Conversely, any tree with non-negative edge parameters will satisfy such a set of inequalities. $\quad\square$

For example, say we substitute (9) into the square of (13). This gives

$$\frac{q_{z(C,0,C,0)} \cdot q_{z(0,C,0,C)}}{q_{z(C,C,0,0)} \cdot q_{z(0,0,C,C)}} \cdot \frac{q_{z(T,0,T,0)} \cdot q_{z(0,T,0,T)}}{q_{z(T,T,0,0)} \cdot q_{z(0,0,T,T)}} \leq \frac{q_{z(G,0,G,0)} \cdot q_{z(0,G,0,G)}}{q_{z(G,G,0,0)} \cdot q_{z(0,0,G,G)}}$$

which is equivalent to a monomial inequality of degree six.

Before moving on, we highlight that (8) is essentially concerned with induced subtrees on only 3 taxa, and (9) is concerned with induced subtrees on only 4 taxa. Inequalities on the collection of these subtrees imply positivity of edge parameters for the entire tree.

4   Fourier transform inequalities: canonical version

The previous section described a relatively simple set of inequalities which can be computed for any edge of a tree. However, some readers may feel uncomfortable with the fact that these inequalities involve some arbitrary choice. In this section we give a "canonical" version of the edge parameter inequalities which is a simple consequence of Theorem 2.4. This version of the inequalities also gives a clearer understanding of the underlying geometry.

We now specialize to the case of either the CFN model or the K3P model (this again includes K3P with extra symmetries, such as JC DNA and K2P). In these cases, the entries of the Fourier transform matrix K are ± 1.

Proposition 4.1. *Let $G = \mathbb{Z}_2$ or $\mathbb{Z}_2 \times \mathbb{Z}_2$ and $\rho(e, h)$ be the element of G^m which assigns h to all leaves in $\Lambda(e)$ and 0 to all others. Then for any $\mathbf{q}$ generated on a tree with non-negative edge parameters,*

$$\prod_{\mathbf{g}:\widehat{\rho(e,h)}(\mathbf{g})=1} q_{\mathbf{g}} \;\geq\; \prod_{\mathbf{g}:\widehat{\rho(e,h)}(\mathbf{g})=-1} q_{\mathbf{g}} \tag{16}$$

Conversely, any tree (with edge parameters) whose $\mathbf{q}$ satisfies (16) for any e and h has non-negative edge parameters.

Proof. Recall that $H^{-1} = |G|^m H$. Thus (7) is

$$|G|^{-m} \psi^{(e)}(h) = [H \log \mathbf{q}]_{\rho(e,h)}, \tag{17}$$

the left hand side of which is non-negative by our main assumption. Exponentiate (17); the left hand side will be not less than one, and the right hand side becomes a ratio with those $q_{\mathbf{g}}$ such that $\widehat{\rho(e, h)}(\mathbf{g}) = 1$ on top and those $q_{\mathbf{g}}$ such that $\widehat{\rho(e, h)}(\mathbf{g}) = -1$ on the bottom. Then multiply to clear denominators. $\quad\square$

Although we have specialized to groups where K has real entries, we note here that equivalent (though more complex) such inequalities exist in all cases. First, we claim that $q_{\mathbf{h}} = q_{-\mathbf{h}}$ for any $\mathbf{h}$. Indeed, assuming time reversibility we have $f^{(e)}(g) = f^{(e)}(-g)$, thus $q_{\mathbf{h}} = q_{-\mathbf{h}}$ by (4). It follows that the coefficients of the $q_{\mathbf{h}}$ in $H^{-1}\mathbf{q}$ are real. Therefore the same exponentiation process in Proposition 4.1 works, although the $q_{\mathbf{h}}$ may now have exponents different than ± 1.

The "path" inequalities of Propositions 3.3 and 3.4, and the "canonical" inequalities of Proposition 4.1, are equivalent. Indeed, they each express the equation $\psi^{(e)}(h) \geq 0$ for various e and h. However, the expressions are different, but by the definition of invariants one can go from one to the other formulation via a complete set of phylogenetic invariants [5].

The previous paragraph establishes equivalence between the two formulations in principle; we present an example here to show how the transformation works. Assume a quartet tree of topology $12|34$; use notation as in the introduction. First we investigate the pendant edge leading to taxon 1. By (16), that edge having non-negative edge length is equivalent to

$$q_{0000}\, q_{0110}\, q_{0011}\, q_{0101} \geq q_{1100}\, q_{1010}\, q_{1001}\, q_{1111}. \qquad (18)$$

A couple of algebraic steps using the phylogenetic invariant $q_{1100}q_{0011} = q_{1111}$ and the fact that $q_{0000} = 1$ shows that (18) is equivalent to

$$1 \geq \left(\frac{q_{1100}\, q_{1001}}{q_{0101}} \right) \left(\frac{q_{1100}\, q_{1010}}{q_{0110}} \right),$$

which is the product of the two "path" pendant edge length inequalities. Similarly, the internal edge being non-negative is equivalent to

$$1 \geq \frac{q_{1010}\, q_{0101}\, q_{1001}\, q_{0110}}{q_{0000}\, q_{1111}\, q_{1100}\, q_{0011}} = \left(\frac{q_{1010}\, q_{0101}}{q_{1100}\, q_{0011}} \right) \left(\frac{q_{1001}\, q_{0110}}{q_{1100}\, q_{0011}} \right)$$

where the right hand side of the equality is the product of the two "path" internal edge length inequalities.

The canonical construction generalizes the inequalities to the more general setting of group-based mutation models on split networks as formulated by David Bryant [15]. Assume the set of splits is labeled Σ. In his elegant formulation, one assigns mutation probabilities to each possible split, i.e. a probability distribution on the group G for each split. Assuming independence of these distributions, one gets a probability distribution on G^{Σ} by multiplication. From there the probability of a single site-pattern $\mathbf{h}$ (i.e. the assignment of a group element to each taxon) is the sum of the probabilities of all elements of G^{Σ} which give $\mathbf{h}$ on the leaves.

Fourier conjugation also works in this setting. Although Bryant's paper [15] only develops the conjugation in the case of models with a fixed rate matrix and "branch length" varying among splits, there is also an invertible transformation for the setting where one allows the whole rate matrix to vary. We will apply

this extended version and call the set of $\psi^{(e)}$ for splits e "split parameters" analogous to the edge parameters we have been describing so far. Although we do not go into details here, the proof of the Fourier conjugation formula in the extended case is similar to that in [15]. One can then obtain an equation for the Fourier conjugation written exactly as in (7) but with a generalized definition of the terms: "root" the splits network at the taxon n, and so redefine $\Lambda(e)$ to be all of the taxa on the opposite "side" of the split from n. For example, $\Lambda(12|34)$ is the set $\{1, 2\}$ as in this case $n = 4$.

Definition 4.2. *Let $Y_{G,m}$ be the points of q-space which satisfy inequalities (16) for each split e and each $h \in G$.*

Observation 4.3.

(i) *$Y_{G,m}$ is the image of the non-negative split parameter splits networks under Hadamard conjugation.*

(ii) *$Y_{G,m}$ is contractible.*

(iii) *The points of q-space corresponding to trees of topology $\mathscr{T}$ with non-negative edge parameters are the zero set of the phylogenetic invariants for $\mathscr{T}$ intersected with $Y_{G,m}$. These points sit on the boundary of $Y_{G,m}$ for $m > 3$.*

Proof. We note that $Y_{G,m}$ is the (injective) image of the set of non-negative split parameter vectors in $(\mathbb{R}_{\geq 0})^{2^{m-1} \cdot (|G|-1)}$. For (i), the inequalities (16) precisely specify positivity of split parameters. For (ii) the required homotopy simply uniformly shrinks every split parameter to zero. The first sentence of (iii) is equivalent to Proposition 4.1. For the second sentence, the boundary of $Y_{G,m}$ consists of the image of splits networks with at least one zero split parameter. Phylogenetic trees are simply split networks such that only a compatible set of split parameters are nonzero. $\qquad\square$

This series of observations suggests that rather than phylogenetic "orange" [4] with one orange slice for each tree topology, one might think of a phylogenetic "soccer ball" with one panel of the soccer ball for each tree topology. Indeed, the set of Fourier transformed points corresponding to any tree live on the boundary of a higher dimensional contractible object. However, it should be noted that not every point of the boundary of $Y_{G,m}$ corresponds to a phylogenetic tree, and in fact the panels are of strictly lower dimension than the boundary of the soccer ball.

Furthermore, we now show that the soccer ball $Y_{G,m}$ is not convex. Recall that $\check{f}^{(e)}(g)$ is real by the discussion after (3). Then:

Lemma 4.4. *The components of the Fourier transformed mutation probability vector $\check{f}^{(e)}(g)$ are less than or equal to one for any edge e with non-negative edge parameters.*

Proof. By Lemma 2.2, it suffices to show that $\check{\psi}^{(e)}(g)$ is non-positive. By the definition of ψ,

$$\psi(0) = -\sum_{g \neq 0} \psi(g)$$

which implies that $\check{\psi}^{(e)}(g)$ is non-positive by the definition of the discrete Fourier transform. $\qquad\square$

Proposition 4.5. $Y_{G,m}$ *is not convex for* $m \geq 3$ *and* $G = \mathbb{Z}_2$ *or* $\mathbb{Z}_2 \times \mathbb{Z}_2$.

Proof. We report the argument for the case of $G = \mathbb{Z}_2 \times \mathbb{Z}_2$ (i.e. K3P); the case of $G = \mathbb{Z}_2$ is analogous but easier. We label the sates A, C, G, T as in (11). Pick an arbitrary tree $\mathscr{T}$ on m taxa; Find a cherry (two-taxon rooted subtree) of $\mathscr{T}$ and label the leaves of $\mathscr{T}$ with $1, 2$. Number the edge leading to taxon 1 with 1, the edge leading to taxon 2 with 2, and the edge meeting 1 and 2 with 3. Pick arbitrary $0 \leq \theta_1, \theta_2, \theta_3 \leq 1$ such that

$$\theta_1 \theta_2 < \theta_3^2 \left((\theta_1 + \theta_2)/2 \right)^2 ; \tag{19}$$

this is easily achieved by fixing θ_2 and θ_3 and taking θ_1 to be small.

We will construct two vectors $\mathbf{q}', \mathbf{q}'' \in Y_{G,m}$ such that $\mathbf{q} := (\mathbf{q}' + \mathbf{q}'')/2$ is not in $Y_{G,m}$. The vectors $\mathbf{q}'$ and $\mathbf{q}''$ will be defined via the Fourier transform by specifying their $\check{f}^{(e)}(g)$. It can be checked that $\mathbf{q}'$ and $\mathbf{q}''$ sit in $Y_{G,m}$ using Lemma 2.2, then taking the logarithm and the inverse Fourier transform.

Let $V = \{C, T\}$. For $\mathbf{q}'$ set

$$\check{f}^{(1)}(g) = \theta_1 \quad \check{f}^{(2)}(g) = \theta_2 \quad \check{f}^{(3)}(g) = \theta_3$$

for $g \in V$, and $\check{f}^{(e)}(g) = 1$ otherwise. For $\mathbf{q}''$ set

$$\check{f}^{(1)}(g) = \theta_2 \quad \check{f}^{(2)}(g) = \theta_1 \quad \check{f}^{(3)}(g) = \theta_3$$

for $g \in V$, and $\check{f}^{(e)}(g) = 1$ otherwise.

We claim that $\mathbf{q}$ violates (16) with $e = 3$ and $h = C$, and thus does not sit in $Y_{G,m}$. To establish this claim, we calculate each side of (16). First note that $\hat{C}(g) = -1$ for $g \in V$ and is 1 otherwise. Thus $\widehat{\rho(3, C)}(\mathbf{g}) = -1$ exactly when $|\{g_1, g_2\} \cap V|$ is odd, and is 1 otherwise (here and below the notation g_i denotes the ith-taxon component of $\mathbf{g}$).

Define $q_{u(x_1, x_2)}$ to be $q_{\mathbf{g}}$ for any $\mathbf{g}$ such that $g_1 = x_1$ and $g_2 = x_2$. This $q_{u(x_1, x_2)}$ is well defined via (4) because all $\check{f}^{(e)}(g) = 1$ except when $e = 1, 2, 3$. Noting that $C + C = 0$, we see that $q_{u(C,C)} = \theta_1 \theta_2$ by (4). Similarly,

$$q_{u(C,A)} = q_{u(A,C)} = \theta_3(\theta_1 + \theta_2)/2.$$

Because we have arranged that $\check{f}^{(e)}(A) = \check{f}^{(e)}(G) = 1$ and $\check{f}^{(e)}(C) = \check{f}^{(e)}(T)$ for both $\mathbf{q}'$ and $\mathbf{q}''$, there are three cases for $q_{u(x_1, x_2)}$. If x_1 and x_2 are in V then $q_{u(x_1, x_2)} = q_{u(C,C)}$. If $|\{x_1, x_2\} \cap V|$ is one then $q_{u(x_1, x_2)} = q_{u(C,A)}$. If neither x_1 nor x_2 are in V then $q_{u(x_1, x_2)} = 1$.

Thus (16) is in this case

$$\left(q_{u(C,C)}^4\right)^{4^{m-2}} \geq \left(q_{u(C,A)}^8\right)^{4^{m-2}}.$$

Taking both sides to the power of 4^{1-m} and substituting gives

$$\theta_1\theta_2 \geq \theta_3^2 \left((\theta_1 + \theta_2)/2\right)^2,$$

violating (19). □

Proposition 4.5 has an interesting phylogenetic interpretation along the lines of [20]: there are mixtures of two site pattern frequency vectors corresponding to trees such that the splits network corresponding to the mixture has negative edge parameters. However, the trees used in the proof had many edge-parameters zero; this is not strictly necessary though it greatly simplifies the proof.

5 Consequences and Conclusions

In summary, we have presented a collection of inequalities in the Fourier transformed site-pattern frequency space which are equivalent to the assumption that group-based mutation rate matrices have non-negative off-diagonal entries. We are motivated in part by the idea of formulating maximum likelihood as a constrained optimization problem [5, 9]. We noted in the introduction that the previously known constraints are not sufficient to ensure that the result of the constrained optimization is in fact a proper tree. As described in Propositions 3.5, 3.6, and 4.1, our inequalities complete the set of constraints: if a **q** satisfies a complete set of phylogenetic invariants and the inequalities described here, then it does indeed correspond to a *bona fide* tree. Thus phylogenetic invariants along with the edge-parameter inequalities could indeed be safely used to formulate maximum-likelihood phylogenetic estimation as a constrained optimization problem.

We also defined $Y_{G,m}$, which is the set of **q** which come from splits networks with non-negative edge parameters. We noted that the tree images for each tree topology sit on the boundary of $Y_{G,m}$. Here we showed that $Y_{G,m}$ is not convex at a number of points. Note that because $Y_{G,m}$ is cut out by monomial inequalities (16) one would expect that $Y_{G,m}$ would be non-convex at "most" points.

As the edge-parameter inequalities are the second component of the constraints for phylogenetic trees, one might wonder if they could be used for phylogenetic inference in an manner analogous to phylogenetic invariants [3, 10]. In a sense these inequalities appear more natural than phylogenetic invariants for the purpose of determining the tree corresponding to a data set: given a real-world data set, one might actually hope that the inequalities presented here could be satisfied, whereas phylogenetic invariants (which are equalities) will essentially never be. Using the terminology above, one might hope that data

would sit in the interior of $Y_{G,m}$ even though one would never expect data to sit on its boundary.

This hope is not justified for simulated data on a tree. Indeed, one can think of the simulated data points as some distribution centered on the expected distribution. Recall that the set of trees are simply the set of splits networks with some edge parameters set to zero. If the simulation distribution does not have support on some lower-dimensional surface, then the pre-image of the distribution will almost certainly have points with negative coordinates in parameter space. Said another way, it is improbable that a sample from a distribution centered on a "corner" of the boundary of $Y_{G,m}$ would sit in the interior of $Y_{G,m}$. As an example one might look at Figure 17.1 of [7] where negative split parameters (besides that for the trivial split) are encountered in a simulation. Despite these challenges, edge-parameter inequalities may well prove useful for inference.

We acknowledge that all of the work presented here is for group-based models. This is a rather strong restriction as all group-based models must have uniform stationary distribution; real data sets rarely have this feature. Presumably, there are inequalities corresponding to those presented here for non-group based models. However, as no Fourier transform is available for those models the formulation may be very complex.

Acknowledgments

References

[1] E. S. Allman and J. A. Rhodes, "Phylogenetic invariants for the general Markov model of sequence mutation," *Math. Biosci.*, vol. 186, no. 2, pp. 113–144, 2003.

[2] M. Casanellas and J. Fernandez-Sanchez, "Geometry of the Kimura 3-parameter model," 2007, arXiv:math/0702834.

[3] J. A. Cavender and J. Felsenstein, "Invariants of phylogenies in a simple case with discrete states," *J. Classif.*, vol. 4, pp. 57–71, 1987.

[4] J. Kim, "Slicing hyperdimensional oranges: the geometry of phylogenetic estimation," *Mol. Phylogenet. Evol.*, vol. 17, no. 1, pp. 58–75, Oct 2000.

[5] B. Sturmfels and S. Sullivant, "Toric ideals of phylogenetic invariants," *J. Comput. Biol.*, vol. 12, no. 2, pp. 204–228, Mar 2005.

[6] E. S. Allman and J. A. Rhodes, "Phylogenetic ideals and varieties for the general Markov model," *Adv. Appl. Math.*, 2007, to appear, `arXiv:math.AG/0410604`.

[7] J. Felsenstein, *Inferring Phylogenies.* Sunderland, MA: Sinauer Press, 2004.

[8] E. Allman and J. Rhodes, "Phylogenetic invariants," in *Reconstructing Evolution: New Mathematical and Computational Advances*, O. Gascuel and M. Steel, Eds. Oxford, UK: Oxford University Press.

[9] S. Hoşten, A. Khetan, and B. Sturmfels, "Solving the likelihood equations," *Found. Comput. Math.*, vol. 5, no. 4, pp. 389–407, 2005.

[10] N. Eriksson, "Using invariants for phylogenetic tree construction," 2007, arXiv:0709.2890.

[11] M. Kimura, "A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences," *J. Mol. Evol.*, vol. 16, no. 2, pp. 111–120, Dec 1980.

[12] D. W. Stroock, *An introduction to Markov processes*, ser. Graduate Texts in Mathematics. Berlin: Springer-Verlag, 2005, vol. 230.

[13] C. Semple and M. Steel, *Phylogenetics.*, ser. Oxford Lecture Series in Mathematics and its Applications. Oxford: Oxford University Press, 2003, vol. 24.

[14] L. A. Székely, M. A. Steel, and P. L. Erdős, "Fourier calculus on evolutionary trees," *Adv. in Appl. Math.*, vol. 14, no. 2, pp. 200–210, 1993.

[15] D. Bryant, "Extending tree models to split networks," in *Algebraic statistics for computational biology*, L. Pachter and B. Sturmfels, Eds. Cambridge: Cambridge University Press, 2005.

[16] V. Moulton and M. Steel, "Peeling phylogenetic 'oranges'," *Adv. Appl. Math.*, vol. 33, no. 4, pp. 710–727, 2004.

[17] M. D. Hendy and D. Penny, "A framework for the quantitative study of evolutionary trees," *Syst. Zool.*, vol. 38, no. 4, pp. 297–309, 1989.

[18] M. D. Hendy, "The relationship between simple evolutionary tree models and observable sequence data." *Syst. Zool.*, vol. 38, no. 4, pp. 301–321, 1989.

[19] S. N. Evans and T. P. Speed, "Invariants of some probability models used in phylogenetic inference," *Ann. Statist.*, vol. 21, no. 1, pp. 355–377, 1993.

[20] F. A. Matsen and M. Steel, "Phylogenetic mixtures on a single tree can mimic a tree of another topology," *Syst. Biol.*, vol. 56, no. 5, pp. 767–775, Oct 2007.