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/

June 13, 2008

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 POSITIVE 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 VERSIONS OF A COMPLETE SET 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. There are certain data points which fit a fixed tree exactly for a given mutation model; I will call the collection of such points a
“tree image.” One of the main goals of the emerging field of phylogenetic geometry [1, 3, 8, 14] 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, 6, 14]. Although not completely understood for all models, a considerable amount of beautiful work has been done on these invariants.

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 [7] (equation (3)) where it is called the maximum likelihood problem. This is repeated again in [14] which 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 [4].

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 [3] 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 positive 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 positive 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 [9]. In the notation of Markov processes,

\[ P^{(e)} = \exp \left( t_e Q^{(e)} \right) \]  

(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 [13].

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."

Here 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 vectors satisfy the phylogenetic invariants, sit in the probability simplex, and the parameters are "biologically meaningful" in the terminology of [3].

For this example, we assume the two-state symmetric (CFN) model with uniform distribution at the root. We will label the two states 0 and 1. In the CFN model, there is only a single parameter per edge, which is typically called the branch length. It is a value such that the probability of mutation between distinct states along an edge of branch length $\gamma$ is $0.5(1 - \exp(-2\gamma))$. Consider the four-taxon tree with the 1234 split. For the example, we assign the branch length $\gamma$ to each pendant edge and $-\gamma$ to the internal edge. Formally, by the above, the off-diagonal entries of the mutation rate matrix of the internal edge will be negative. Let $\theta = \exp(-2\gamma)$. With the above notation, the nontrivial entry of the Fourier transform of the mutation matrix will be $\theta$ for the pendant edges and $\theta^{-1}$ 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 "$\mathbf{q}$-space." We will index $\mathbf{p}$ and $\mathbf{q}$ with state vectors $\mathbf{g}$.

Now we can use Hadamard conjugation (see [3] or [12]) to compute $\mathbf{q}$. For example, $q_{1010} = \theta \cdot \theta^{-1} \cdot \theta = \theta$. All of the other similar calculations are reported in Table 1. One more application of the Fourier transform gives the $\mathbf{p}$. 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 $\mathbf{g}$ equals zero.

It is clear that in Table 1 all $p_{\mathbf{g}}$ are positive for $0 \leq \theta \leq 1$ 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 $\theta$, which ensures that $\mathbf{p}$ 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 $\mathbf{q}$ 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, $\theta$ and $\theta^{-1}$) are positive for any real
\[ \gamma < 0, \text{ and thus the mutation parameters satisfy the two-state analog of the } \] 
\[ \text{‘biologically meaningful’ criterion of } [3]. \text{ This example begs the question of} \]
\[ \text{what conditions should be put on site-pattern frequency vectors or their Fourier} \]
\[ \text{transforms so that one can be assured that the corresponding trees are well-} \]
\[ \text{formed.} \]

\[ \text{This paper describes the set of ‘edge-parameter inequalities’ and shows that} \]
\[ \text{they are the exact conditions needed, namely that any solution of the phylogene-} \]
\[ \text{tic invariants for a given tree which satisfies these inequalities is guaranteed to} \]
\[ \text{come from a tree with positive edge parameters. For example, the ‘canonical’} \]
\[ \text{edge-parameter inequality for the internal edge of the quartet tree is} \]
\[ \frac{q_{1010} q_{0101} q_{1001} q_{0110}}{q_{0000} q_{1111} q_{1100} q_{0011}} \leq 1, \]

which is violated for our pathological example above.

\[ \text{We will describe two distinct versions of the edge-parameter inequalities.} \]
\[ \text{The first version is derived by considering paths in the tree and thus we will call} \]
\[ \text{this version the “path” edge-parameter inequalities. This version is relatively} \]
\[ \text{simple to write down, involving two monomials of degree at most four for the} \]
\[ \text{two-state models and two monomials of degree at most six for the four-state} \]
\[ \text{models. We note that as this set of inequalities is derived on trees, they are} \]
\[ \text{only meaningful for } q \text{ which satisfy a complete set of phylogenetic invariants for} \]
\[ \text{a tree.} \]

\[ \text{Next we present the second version of the inequalities; these inequalities} \]
\[ \text{derive directly from the Székely-Steel-Erdős Fourier conjugation equation } [15]. \]
\[ \text{Because they are given directly by Fourier conjugation, we call these inequalities} \]
\[ \text{the “canonical” edge-parameter inequalities. These inequalities for group } G \]
\[ \text{-based models on trees of } m \text{ taxa carve out a subset of } q \text{-space which we denote} \]
\[ Y^+_{G,m}. \text{ The set of } q \text{'s corresponding to a given } m \text{-taxon tree is the set of solutions} \]
\[ \text{to that tree’s phylogenetic invariants intersected with } Y^+_{G,m}. \]

\[ \text{We then investigate some properties of } Y^+_{G,m}. \text{ The set } Y^+_{G,m} \text{ is the subset} \]
\[ \text{of } q \text{-space which corresponds precisely to the } q \text{ of splits networks with positive} \]
\[ \text{split parameters using an extension of the model of } [2]; \text{ thus it is contractible.} \]

| $k$  | $q_k$ | $8 \cdot p_k$ |
|------|------|--------------|
| 0000 | 1    | $1 + 4\theta + 2\theta^2 + \theta^4$ |
| 1001 | $\theta$ | $1 - \theta^4$ |
| 0101 | $\theta$ | $1 - \theta^4$ |
| 1100 | $\theta^2$ | $1 - 4\theta + 2\theta^2 + \theta^4$ |
| 0011 | $\theta^2$ | $1 - \theta^4$ |
| 1010 | $\theta$ | $(1 - \theta^2)^2$ |
| 0110 | $\theta$ | $(1 - \theta^2)^2$ |
| 1111 | $\theta^4$ | $1 - \theta^4$ |
It is not convex. Furthermore, the $q$ corresponding to phylogenetic trees sit on the boundary of $Y_{G,m}^+$, thus the complete space of phylogenetic “oranges” for group-based models lives on this boundary.

2 Technical introduction

In this section we fix notation and state two versions of the Fourier conjugation. Our notation combines that of [2] and [14].

As stated in the introduction, the simple observation of this paper is that the mutation matrices come from continuous-time Markov processes. Thus the mutation matrices $F^{(e)}$ must satisfy (1) for each edge $e$. 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 $h$. Fixing an edge $e$, we write

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

where $Q^{(e)}$ denotes the mutation rate matrix along an edge $e$ and $\psi^{(e)}$ is an arbitrary 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 $Q^{(e)}_{g,h} = Q^{(e)}_{h,g}$, which is equivalent to $\psi^{(e)}(g) = \psi^{(e)}(-g)$. Because exponentiation preserves symmetries of the matrices, we will also have

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

for some probability 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 $\hat{G}$.

We now come into a slight conflict of notation between [2] and [14]. Each paper uses “hat” to denote something to do with the dual group, but the specifics vary. We follow each of these papers by using the same letters for the group
and its dual \( \hat{G} \), but it is convenient to have a notation for the application of an element of \( \hat{G} \) to an element of \( G \). We will follow [2] in using “hat” for this purpose, such that \( \hat{g}(h) \) is the application of \( g \in \hat{G} \) to \( h \in G \). This conflicts with the notation of [14] for Fourier transform; we will use “check” as defined below.

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

\[
\hat{a}(g) := \sum_{h \in G} \hat{g}(h)a(h).
\]

By the definition \( \hat{f}^{(e)}(0) = 1 \) for any \( e \). Note

\[
\hat{f}^{(e)}(-g) = \sum_{h \in G} \hat{g}(h)f^{(e)}(h) = \sum_{h \in G} \hat{g}(-h)f^{(e)}(h) = \sum_{h \in G} \hat{g}(h)f^{(e)}(-h) = \hat{f}^{(e)}(g).
\]

By the definition of the Fourier transform, \( \hat{a}(-g) = \overline{\hat{a}(g)} \) for any real-valued function \( a \). Thus the fact that \( \hat{f}^{(e)}(g) = \hat{f}^{(e)}(-g) \) is equivalent to the fact that \( \hat{f}^{(e)}(g) \) is real.

We follow [14, 15] by re-rooting the tree by extending an edge from the root terminating in a node which we still call the root; the previous root distribution is then replaced by a mutation matrix along the new edge. We assume that the new root is always in the identity state. Note that this new edge may not have the same symmetries as the edges in the rest of the tree: for example, if the rest of the edges in the tree have the symmetries corresponding to the Jukes-Cantor (JC) DNA model, the new edge need not have these symmetries. This poses no problems for what follows. Fix a tree \( T \) on \( m \) leaves. We will write elements \( g \) of \( G^m \) in bold.

Fourier conjugation is an invertible transformation between the 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 \( e \) with probabilities \( P^{(e)} \). The induced joint distribution on the leaves will be denoted \( p \) where \( p_g \) is the probability of seeing \( g \) by the above process. We note that this notation differs from that of [15] though agrees with more recent publications such as [12] and [14].

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

Following [14], use \( \Lambda(e) \) to denote the set of leaves \( i \) such that the path between \( i \) and the root goes through \( e \). We also define

\[
^g e = \sum_{i \in \Lambda(e)} g_i.
\]
The vector $\vec{g}$ is a natural lift of a $g \in G^m$ to an assignment of $G$ to all nodes of the tree. In this notation, the first version of Fourier conjugation can be written

**Theorem 2.1** (Evans and Speed, 1993 [5]).

$$q_g = \prod_{e \in E} \vec{f}(e)^*(\vec{g}_e).$$

(3)

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

**Lemma 2.2.**

$$\hat{f}(h) = \exp(\tilde{\psi}(h)).$$

Proof. By the same argument as in [2] (though for right rather than left eigenvalues),

$$(QK)_{g,h} = \sum_{x \in G} \psi(x - g)\hat{\psi}(h) = \sum_{y \in G} \psi(y)\hat{g}(h) = \hat{g}(h) \sum_{y \in G} \psi(y)\hat{\psi}(h) = K_{g,h} \tilde{\psi}(h).$$

(4)

Thus the $h$th column of $K$ is a right eigenvector of $Q$ with eigenvalue $\tilde{\psi}(h)$. The same argument with $f$ in place of $\psi$ shows that the $h$th column of $K$ is a right eigenvector of $P$ with eigenvalue $\hat{f}(h)$. However, $P = \exp(Q)$ so the eigenvalues of $P$ are the exponentials of the corresponding eigenvalues of $Q$.

As noted in the discussion after (2), $\tilde{\psi}(g)$ is real for any $g$. Thus Lemma 2.2 implies

**Observation 2.3.** Any edge with positive parameters will have real and positive Fourier transform $\tilde{\psi}(e)$.

Thus any tree with positive edge parameters has “biologically meaningful” parameters in the language of [3], though converse does not hold. We also note that by [3] the $q_g$ are real; thus the log of (6) 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 eigenvalues,

$$\psi(h) = [K^{-1} \log K\vec{f}]_h.$$

(5)

The following theorem is Theorem 6 of [15] in the presence of (5).

**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

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

(6)

Note that the log in equation (6) is entry-wise.
3 Fourier transform inequalities: path version

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

\[
F(i,j;g) = \prod_{e \in p(i,j)} \hat{f}(e)(g).
\]

We record the following facts for future use:

**Lemma 3.1.**

(i) Let \( \nu \) 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) \( F(i,j;g) = F(i,j;-g) \).

**Proof.** Part (i) is clear from the definition. Equation (2) implies (ii).

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 \( g \) have \( g_i = h \), \( g_j = -h \) and all other components zero. Then \( q_g = F(i,j;h) \).

The first identity is for pendant edges.

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

\[
\left[ \hat{f}(e)(h) \right]^2 = \frac{q_w(h,-h,0) \cdot q_w(-h,0,h)}{q_w(0,-h,h)}.
\]

**Proof.** Lemmas 3.1 and 3.2 show

\[
q_w(h,-h,0) = \hat{f}(e)(h) \cdot F(\nu,j;h)
\]
\[
q_w(-h,0,h) = \hat{f}(e)(h) \cdot F(\nu,k;h)
\]
\[
q_w(0,h,-h) = F(\nu,j;h) F(\nu,k;h).
\]

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 p(i,j) (resp. p(i′,j′))
contains ν but not ν′ (resp. ν′ but not ν). Let z(g_i,g_j,g_i′,g_j′) assign state g_x
to leaf x for x ∈ {i,j,i′,j′} and the identity to all other leaves. Then
\[
[\hat{f}(e)(h)]^2 = \frac{q_z(h,0,-h,0) \cdot q_z(0,-h,0,h)}{q_z(-h,0,0) \cdot q_z(0,0,-h,h)}.
\] (8)

Now, constraints on the  \( \hat{f}(e)(h) \) will imply inequalities in the  \( q_k \). 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 com-
ponent  \( \hat{f}(e)(1) \) 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) \ implies
\[
\hat{f}(e)(1) \leq 1.
\] (9)

Inserting the values for  \( \hat{f}(e)(1) \) from Propositions 3.3 and 3.4 into this equation
give the edge-parameter inequalities for each edge.

Proposition 3.5. Assume that  \( q \) is the \( \mathbb{Z}_2 \)-Fourier transform of a site-pattern
frequency vector under the CFN model. If  \( q \) satisfies a complete set of phylo-
genetic invariants for a tree  \( T \) and a set of inequalities gained by substituting
an instance of (7) or (8) into the square of (9) for each edge  \( e \) of  \( T \), then  \( q \) is
the expected site-pattern frequency vector of  \( T \) for some assignment of positive
branch lengths to  \( T \).

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
\[
\frac{q_{1001}q_{0110}}{q_{1100}q_{0011}} = \hat{f}(e)(1) \leq 1.
\]

However, by substituting in values from Table 1 the above ratio is \( \theta^{-2} \), 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
\[
A = (0,0) \quad C = (1,0) \quad G = (0,1) \quad T = (1,1).
\] (10)

Assume that the rates of transition from A to C, G, and T are a, b and c,
respectively, i.e. that  \( \psi \) is the column vector \([-a+b+c, a, b, c]^T \). Then by
Lemma 2.2 we have
\[
\hat{f}^{(e)}(C) = \exp(-2(a + c))
\]
\[
\hat{f}^{(e)}(G) = \exp(-2(b + c))
\]
\[
\hat{f}^{(e)}(T) = \exp(-2(a + b)).
\]

We ignore the fact that \(\hat{f}^{(e)}(A) = 1\); this equality leads to trivial inequalities. However, the fact that we require \(a\), \(b\), and \(c\) to be non-negative leads to nontrivial ones. The following equations are equivalent to requiring \(a\), \(b\), and \(c\) to be non-negative via (11):
\[
\hat{f}^{(e)}(C) \hat{f}^{(e)}(T) \leq \hat{f}^{(e)}(G)
\]
\[
\hat{f}^{(e)}(G) \hat{f}^{(e)}(T) \leq \hat{f}^{(e)}(C)
\]
\[
\hat{f}^{(e)}(C) \hat{f}^{(e)}(G) \leq \hat{f}^{(e)}(T).
\]

In summary,

\textbf{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 \(T\) and a set of inequalities gained by substituting an instance of (7) or (8) into the square of (12), (13), and (14) for each edge \(e\) of \(T\), then \(\mathbf{q}\) is the expected site-pattern frequency vector of \(T\) for some assignment of positive branch lengths to \(T\).

For example, say we substitute (8) into the square of (12). This gives
\[
\frac{q_z(C,0,C,0) \cdot q_z(0,C,0,C) \cdot q_z(T,0,T,0) \cdot q_z(0,T,0,T)}{q_z(C,C,0,0) \cdot q_z(0,0,C,C) \cdot q_z(T,T,0,0) \cdot q_z(0,0,T,T)} \leq \frac{q_z(G,0,G,0) \cdot q_z(0,G,0,G)}{q_z(G,G,0,0) \cdot q_z(0,0,G,G)}
\]
which is equivalent to a monomial inequality of degree six.

\section{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 choice. In this section we give a “canonical” version of the positive 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 (we also include K3P with extra symmetries, such as JC DNA and K2P). In these cases, the entries of the Fourier transform matrix \(K\) are \(\pm 1\). When we exponentiate the image of a vector under a matrix whose entries are \(\pm 1\), each component of the result is a ratio where the vector components assigned 1 are on the top, and those assigned \(-1\) are on the bottom. Thus
Proposition 4.1. Let $G = \mathbb{Z}_2$ or $\mathbb{Z}_2 \times \mathbb{Z}_2$ and $\rho(e, h)$ be the element of $G^n$ which assigns $h$ to all leaves in $\Lambda(e)$ and $0$ to all others. Then for any $q$ generated on a tree with positive edge parameters,

$$\prod_{\gamma : \rho(e, h)(\gamma) = 1} q_\gamma \geq \prod_{\gamma : \rho(e, h)(\gamma) = -1} q_\gamma$$

Conversely, any tree (with edge parameters) whose $q$ satisfies (15) has positive edge parameters.

**Proof.** Recall that $K^{-1} = |G|^m K$. Exponentiate $|G|^m$ times the equation in Theorem 2.4; by definition of positive edge-parameter the left hand side will be non-negative. Then multiply to clear denominators.

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 $q_h = q_{-h}$ for any $h$. Indeed, assuming time reversibility we have $f(e)(g) = f(e)(-g)$, thus $p_s = p_s$ which implies $q_h = q_{-h}$ by the definition of $q$. It follows that the coefficients of the $q_h$ in $H^{-1}q$ are real. Therefore the same exponentiation process in Proposition 4.1 works, although the $q_h$ may have exponents different than $\pm 1$.

We now note equivalence between the “path” inequalities in Propositions 3.3 and 3.4 on one hand, and the “canonical” inequalities of Proposition 4.1 on the other. This equivalence can be achieved by cancellation using a complete set of phylogenetic invariants, as in [14]. Let us assume the K3P model, as the CFN case is simpler. The ratios in (12), (13), and (14) are expressions for the exponential of $-4$ times $a = \psi(C)$, $b = \psi(G)$, and $c = \psi(T)$, respectively. By substituting in values for the $f(e)(g)$ in these ratios one gets a ratio of $q_h$’s which by definition is equal to the ratio obtained by exponentiating $-4$ times $q_h$’s. The resulting equality is a phylogenetic invariant by definition and so can be written as a product of the assumed generating set of invariants. These invariants specify the cancellations.

Although the previous paragraph establishes equivalence in principle, we present an example here to show how it 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 (15), that edge having positive edge length is equivalent to

$$q_{0000} q_{0110} q_{0011} q_{0101} \geq q_{1100} q_{1010} q_{1001} q_{1111}.$$  \(16\)

A couple of algebraic steps using the phylogenetic invariant $q_{1100}q_{0011} = q_{1111}$ and the fact that $q_{0000} = 1$ shows that (16) is equivalent to

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

which is the product of the two “path” pendant edge length inequalities. Similarly, the internal edge being positive is equivalent to

$$\frac{q_{1010} q_{0101} q_{1001} q_{0110}}{q_{0000} q_{1111} q_{1100} q_{0011}} = \left( \frac{q_{1010} q_{0101}}{q_{1100} q_{0011}} \right) \left( \frac{q_{1001} q_{0110}}{q_{1100} q_{0011}} \right) \leq 1.$$
where the right hand side of the equality is the product of the two “path” pendant edge length inequalities.

The canonical construction also generalizes the inequalities to the more general setting of group-based mutation models on split networks as formulated by David Bryant [2]. Assume the set of splits is labeled $\Sigma$. 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 $G^\Sigma$. From there the probability of a single site-pattern $h$ (i.e. the assignment of a group element to each taxon) is the sum of the probabilities of all elements of $G^\Sigma$ which give $h$ on the leaves.

Fourier conjugation also exists in this setting. Although Bryant’s paper [2] 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 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 this case is similar to that in [2]. One can then obtain an equation for the Fourier conjugation written exactly as in (6) 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 $\{1,2\}$.

**Definition 4.2.** Let $Y^+_{G,m}$ be the points of q-space which satisfy inequalities (15) for each split $e$ and each $h \in G$.

**Observation 4.3.**

(i) $Y^+_{G,m}$ is the image of the positive split-parameter splits networks under Hadamard conjugation.

(ii) $Y^+_{G,m}$ is contractible.

(iii) The points of q-space corresponding to a tree $T$ with positive edge parameters are the zero set of the phylogenetic invariants for $T$ 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 positive split parameter vectors in $(\mathbb{R}_+)^{2^m-1.\left|\Sigma\right|-1}$. For (i), the inequalities (15) 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.

We will now show that $Y^+_{G,m}$ is not convex. Recall that $f^{(e)}(g)$ is real by the discussion after (2). Then:
Lemma 4.4. The Fourier transformed mutation probability vector $\tilde{f}^{(e)}(g)$ is less than or equal to one for any edge $e$ with positive edge parameters.

Proof. By Lemma 2.2 it suffices to show that $\tilde{\psi}^{(e)}(g)$ is negative. By the definition of $\psi$, 

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

which implies that $\tilde{\psi}(e)$ is negative. ∎

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 states $A, C, G, T$ as in [10]. Pick an arbitrary tree $T$ on $m$ taxa; Find a cherry (two-taxon rooted subtree) of $T$ and label the leaves of $T$ 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

$$\theta_1 \theta_2 < \theta_3^6 ((\theta_1 + \theta_2)/2)^6; \quad (17)$$

this is easily achieved by fixing $\theta_2$ and $\theta_3$ and taking $\theta_1$ 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 $\tilde{f}^{(e)}(g)$. For $\mathbf{q}'$ set

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

and all other $\tilde{f}^{(i)}(g) = 1$. For $\mathbf{q}''$ set

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

and all other $\tilde{f}^{(i)}(g) = 1$.

We claim that $\mathbf{q}$ violates Proposition 4.1 with $e = 3$ and $h = C$, and thus does not sit in $Y_{G,m}^+$. To establish this claim, we calculate each side of (15). Because $C + C = 0$,

$$\rho(3, C)(\mathbf{g}) = \hat{C}(g_1) \cdot \hat{C}(g_2).$$

Thus $\rho(3, C)(\mathbf{g}) = 1$ when $g_1 = g_2 = C$ or neither $g_1$ or $g_2$ is equal to $C$, while $\rho(3, C)(\mathbf{g}) = -1$ exactly when precisely one of $g_1$ or $g_2$ is equal to $C$.

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 because all $\tilde{f}^{(e)}(g) = 1$ except when $e = 1, 2, 3$. Because $\tilde{f}^{(e)}(g) = 1$ unless $g = C$, $q_u(C, g) = q_u(C, A)$ for all $g \neq C$. Similarly, $q_\mathbf{g} = 1$ unless $g_1$ or $g_2$ is equal to $C$. Now

$$q_u(C, A) = q_u(A, C) = \frac{\theta_1 + \theta_2}{2} \theta_3 \quad \text{and} \quad q_u(C, C) = \theta_1 \theta_2. \quad (18)$$
Thus (15) is in this case
\[
(q_u(C,C))^{4m-3} \geq (q_u(C,A)q_u(A,C))^{4m-3}.
\]
Taking both sides to the power of \(4^{3-m}\) and substituting (18) gives
\[
\theta_1 \theta_2 \geq \theta_3^6 ((\theta_1 + \theta_2)/2)^6,
\]
vioating (17).

Proposition 4.5 has an interesting phylogenetic interpretation along the lines of [10]: there are mixtures of two site pattern frequency vectors corresponding to trees such that the splits network corresponding to the mixture does not have positive edge parameters. Note that the trees used in the above 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 follow from the assumption that the mutation rate matrix has non-negative off-diagonal entries. We were motivated in part by the idea of formulating maximum likelihood as a constrained optimization problem [7, 14]. We noted in the introduction that the present 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 proper tree.

We also defined \(Y^+_{G,m}\), which corresponds to the set of \(q\) vectors which come from splits networks with positive 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, although given that \(Y^+_{G,m}\) is cut out by monomial inequalities (15) one would expect that some general strong version of non-convexity would hold.

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 [4]. 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 notation 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 [6] 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; it is known that 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

The starting point for this paper came from discussions with a number of very helpful people, including Elizabeth Allman, David Bryant, Mike Steel, and Bernd Sturmfels. The author would like to thank Seth Sullivan for clearing up a technical point from their paper, Steve Evans and Lior Pachter for comments on an early version of the manuscript, and Dustin Cartwright for several helpful discussions. Frederick Matsen is funded by the Miller Institute for Basic Research in Science at the University of California, Berkeley.

References

[1] E. S. Allman and J. A. Rhodes. Phylogenetic invariants for the general Markov model of sequence mutation. *Math. Biosci.*, 186(2):113–144, 2003.

[2] D. Bryant. Extending tree models to split networks. In L. Pachter and B. Sturmfels, editors, *Algebraic statistics for computational biology*. Cambridge University Press, Cambridge, 2005.

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

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

[5] S. N. Evans and T. P. Speed. Invariants of some probability models used in phylogenetic inference. *Ann. Statist.*, 21(1):355–377, 1993.

[6] J. Felsenstein. *Inferring Phylogenies*. Sinauer Press, Sunderland, MA, 2004.
[7] S. Hoşten, A. Khetan, and B. Sturmfels. Solving the likelihood equations. *Found. Comput. Math.*, 5(4):389–407, 2005.

[8] J. Kim. Slicing hyperdimensional oranges: the geometry of phylogenetic estimation. *Mol. Phylogenet. Evol.*, 17(1):58–75, Oct 2000.

[9] M. Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.*, 16(2):111–120, Dec 1980.

[10] F. A. Matsen and M. Steel. Phylogenetic mixtures on a single tree can mimic a tree of another topology. *Syst. Biol.*, 56(5):767–775, Oct 2007.

[11] V. Moulton and M. Steel. Peeling phylogenetic ‘oranges’. *Adv. Appl. Math.*, 33(4):710–727, 2004.

[12] C. Semple and M. Steel. *Phylogenetics.*, volume 24 of *Oxford Lecture Series in Mathematics and its Applications*. Oxford University Press, Oxford, 2003.

[13] D. W. Stroock. *An introduction to Markov processes*, volume 230 of *Graduate Texts in Mathematics*. Springer-Verlag, Berlin, 2005.

[14] B. Sturmfels and S. Sullivant. Toric ideals of phylogenetic invariants. *J. Comput. Biol.*, 12(2):204–228, Mar 2005.

[15] L. A. Székely, M. A. Steel, and P. L. Erdős. Fourier calculus on evolutionary trees. *Adv. in Appl. Math.*, 14(2):200–210, 1993.