Enclosing the Maximum Likelihood of the Simplest DNA Model Evolving on Fixed Topologies: Towards a Rigorous Framework for Phylogenetic Inference

Raazesh Sainudiin

Cornell University, Ithaca, USA

Summary. An interval extension of the recursive formulation for the likelihood function of the simplest Markov model of DNA evolution on unrooted phylogenetic trees with a fixed topology is used to obtain rigorous enclosure(s) of the global maximum likelihood. Validated global maximizer(s) inside any compact set of the parameter space which is the set of all branch lengths of the tree are thus obtained. The algorithm is an adaptation of a widely applied global optimization method using interval analysis for the phylogenetic context. The method is applied to enclose the maximizer(s) and the global maximum for the simplest DNA model evolving on trees with 2, 3, and 4 taxa. The method is also applicable to larger trees.

1. INTRODUCTION

When one is given a homologous set of distinct deoxyriobnucleic acid (DNA) sequences of length v from s species and asked for an estimate of their inter-relationships back throught time under some model of DNA evolution, a phylogenetic tree estimation problem arises. This problem is two-fold. First, one has to estimate the shape or topology of the tree, which captures the set of "who is related to whom and in what order? and whose ancestors are related to whose and in what order?" questions. Second, one has to estimate the lengths of the branches when given a particular topology. The branch lengths of a tree usually represent a scaled product of mutation rate and number of generations between the nodes. The s extant species are represented by the external nodes or leaves and their ancestors are represented by the internal nodes of the tree. A rooted tree always has a bifurcation at the root, typically the most recent common ancestor of all s leaves, where as, an unrooted tree has m-furcations at all internal nodes with $m \geq 3$. This work focusses on the second problem, namely, estimating the branch lenghts for a given topology in a maximum likelihood framework.

When statistical inference is conducted in a maximum likelihood framework, one is interested in the global maximum of the likelihood function over the parameter space. Explicit analytical solutions for the maximum likelihood estimates of the branch lengths for a specified unrooted topology with more than 2 leaves are not available even for the simplest model of DNA evolution due to Jukes and Cantor (1969) without assuming a molecular clock. See 5.(b) of Yang (2000) for results on clocked 3-leaved rooted trees. Results are known for models with two character states superimposed on 3-leaved trees (Yang, 2000), as well as for specific observations on 4-leaved trees (Chor, 2000). In practice one settles for a local optimization algorithm to numerically approximate the global solution.

However, statistical inference procedures that rely on having found some global optimum through any numerical approach may suffer from at least five major sources of errors. They are *undirected rounding* and *catastrophic cancellation* (Cuyt et al., 2001; Loh and Walster, 2002), *discretization* of a problem originally posed in the continuum, *conversion* from the decimal format to a non-decimal floating-point format, *ill-posed statistical experiment or model* (unknown nonidentifiable parameter subspaces, for example), and *finite precision in measurement* of relevant empirical observations. Furthermore, traditional nonlinear programming techniques that use local information, such as, clustering methods, generalized ascent methods, and other stochastic search methods, including MCMC and simulated annealing, start from some approximate trial point(s) and iterate by sampling only finitely many points. Therefore, they can neither validate that the objective function has not soared between the sampled points, nor guarantee escape from a local maximum, albeit they can be made to increase the probability of such desired events. Methods that use local information at finitely many points and do not account for all five major sources of errors, cannot be expected to yield anything more than a possible, approximate, and local solution.

Controversy exists over the nonrigorous nature of such a numerically-based statistical inference procedure, especially in parameter-rich models that have not been shown to be identifiable and/or are adorned with multiple local optima. In some problems, a local approximate solution may be sufficient, but in others one may base statistical decisions that address a real biological problem on an incorrect solution. Unfortunately, it currently seems impossible to know this difference *apriori*. This paper shows an existing method toward such knowledge and applies it to enclose the maximum likelihood value as well as the estimate of the most likely unrooted multifurcating four taxa tree for any given data set of four homologous DNA sequences that are assumed to evolve according to the Jukes and Cantor model (Jukes and Cantor, 1969).

The global optimization method sketched below rigorously encloses the global maximum of the likelihood function through interval analysis. Such interval methods evaluate the likelihood function over a continuum of points including those that are not machine-representable and account for the five sources of errors described earlier. Thus, in contrast to local search methods, interval methods can enclose the global optimum with guaranteed accuracy by exhaustive search within any compact set of the parameter space. However, interval methods applied to phylogenetic inference are much slower on currently available hardware optimized for floating-point operations.

Section 2 contains a brief introduction to various enclosure arithmetics. For a recent introduction to such arithmetics see (Kulisch et al., 2001). Section 3 describes a problem that arises in phylogenetic inference through maximum likelihood. Section 4 gives the basic global optimization algorithm based on Hansen's method (Hansen, 1980, 1992) with Ratz's modifications (Ratz, 1992) as implemented in Hammer et al. (1995) with further extensions that account for non-stationary maxima at the boundaries and increase computational efficiency. Assuming the simplest model of DNA evolution, the method is applied in section 5 to rigorously estimate trees, with two, three, and four leaves, based on primate mitochondrial DNA sequences (Brown et al., 1982).

2. PRELIMINARIES

2.1. Interval arithmetic

Lower case letters denote real numbers, e.g. $x \in \mathbb{R}$, the set of real numbers. Upper case letters represent bounded and closed (compact) real intervals, e.g. $X = [\underline{x}, \overline{x}] = [\inf(X), \sup(X)]$. Any compact interval $X \in \mathbb{IR} := \{[a,b] : a \leq b,\ a,b \in \mathbb{R}\}$, the set of all compact real intervals. The diameter and the midpoint of X are $d(X) := \overline{x} - \underline{x}$ and $m(X) := (\underline{x} + \overline{x})/2$, respectively. The smallest and largest absolute value of an interval X are real numbers given by $\langle X \rangle := \min\{|x| : x \in X\} = \min\{|\underline{x}|, |\overline{x}|\}$ and $|X| := \max\{|x| : x \in X\} = \max\{|\underline{x}|, |\overline{x}|\}$, respectively, while the absolute value of an interval X is $|X|_{[\,]} := \{|x| : x \in X\} = [\langle X \rangle, |X|]$. The relative diameter of an interval X, denoted by d_{rel} is the diameter $d(X)$ itself if $0 \in X$, and $d(X)/\langle X \rangle$, otherwise. An interval X with zero diameter is called a thin interval with $\underline{x} = \overline{x} = x$. The hull of two intervals is $X \cup Y := [\min\{\underline{x}, \underline{y}\}, \min\{\overline{x}, \overline{y}\}]$. By the notation $X \Subset Y$, it is meant that X is completely contained in Y, i.e., $\underline{x} > \underline{y}$ and $\overline{x} < \overline{y}$. No notational distinction is made between a real number $x \in \mathbb{R}$ and a real vector $x = (x_i, \cdots, x_n)^T \in \mathbb{R}^n$ and between a real interval X and a real interval vector or box $X = (X_1, \cdots, X_n)^T \in \mathbb{IR}^n$, i.e. $X_i = [\underline{x_i}, \overline{x_i}] = [\inf(X_i), \sup(X_i)] \in \mathbb{IR}$, where, $i = 1, \cdots, n$. The dimension n should be clear from the context. For an interval vector X, the diameter, relative diameter, midpoint, and hull operations are defined component-wise to yield vectors, while the maximum over its components is taken to obtain the maximal diameter and the maximal relative diameter, $d_\infty(X) = \max_{1 \leq i \leq n} d(X_i)$ and $d_{rel,\infty}(X) = \max_{1 \leq i \leq n} d_{rel}(X_i)$, respectively.

It can be seen that $\mathbb{IR}$ under the metric $\mathfrak{h}$, given by $\mathfrak{h}(X,Y) := \max\{|\underline{x} - \underline{y}|, |\overline{x} - \overline{y}|\}$, is a complete metric space. Convergence of a sequence of intervals $\{X^{(i)}\}$ to an interval X under the metric $\mathfrak{h}$ is equivalent to the sequence $\mathfrak{h}(X^{(i)}, X)$ approaching 0 as i approaches ∞, which in turn is equivalent to both $\underline{x}^{(i)} \to \underline{x}$ and $\overline{x}^{(i)} \to \overline{x}$. Contuinity and differentiability of a function $f : \mathbb{IR}^n \to \mathbb{IR}^k$ are defined in the usual way.

A real arithmetic operation $x \circ y$, where $\circ \in \{+, -, \cdot, /\}$ and $x, y \in \mathbb{R}$, is a continuous function $x \circ y := \circ(x,y) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, except when $y = 0$ under $/$ operation. An interval arithmetic operation $X \circ Y := \{x \circ y : x \in X, y \in Y\}$ thus yields the set that contains the result of the operation done for every real pair $(x,y) \in (X,Y)$. This definition of interval operation leads to the property of inclusion isotony which stipulates that $X \circ Y$ contain $V \circ W$ provided $V \subseteq X$ and $W \subseteq Y$. Since X and Y are simply connected compact intervals, so is their product $X \times Y$. On such a domain $X \times Y$, the continuity of $\circ(x,y)$ (except

when $\circ = /$ and $0 \in Y$) ensures the attainment of a minimum, a maximum and all intermediate values. In other words, with the exception of the case when $\circ = /$ and $0 \in Y$, the range $X \circ Y$ has an interval form $[\min(x \circ y), \max(x \circ y)]$, where the min and max are taken over all pairs $(x, y) \in X \times Y$. The particular forms of $X \circ Y$ for the elementary operations are,

$$X + Y = [\underline{x} + \underline{y}, \overline{x} + \overline{y}], \qquad X \cdot Y = [\min\{\underline{x}\underline{y}, \underline{x}\overline{y}, \overline{x}\underline{y}, \overline{x}\overline{y}\}, \max\{\underline{x}\underline{y}, \underline{x}\overline{y}, \overline{x}\underline{y}, \overline{x}\overline{y}\}],$$

$$X - Y = [\underline{x} - \overline{y}, \overline{x} - \underline{y}], \text{ and} \qquad X/Y = X \cdot [1/\overline{y}, 1/\underline{y}],\ 0 \notin Y.$$

The identity elements of $+$ and $\cdot$ are the thin intervals $[0,0]$ and $[1,1]$, respectively. Multiplicative and additive inverses do not exist except when X is also thin, since $[0,0] \subseteq X - X$, and $[1,1] \subseteq X/X$. Although the commutative and associative laws are satisfied by $+$ and $\cdot$, only a weaker notion of distributivity called subdistributivity, i.e., $X \cdot (Y + Z) \subseteq (X \cdot Y) + (X \cdot Z)$, is satisfied.

For any real function $f(x) : \mathbb{R}^n \to \mathbb{R}$ and some box $X \in \mathbb{IR}^n$, let the range of f over X be denoted by $f(X) := \{f(x) : x \in X\}$. Inclusion isotony also holds for interval evaluations that are compositions of arithmetic expressions and the elementary functions. When real variables and operations in f are replaced by their interval counterparts one obtains $F(X) : \mathbb{R} \to \mathbb{R}$, the natural interval extension of f. Guaranteed enclosures of the range $f(X)$ are obtained by $F(X)$, since inclusion isotony holds for F, i.e, if $X \subseteq Y$, then $F(X) \subseteq F(Y)$, and in particular, the inclusion property that $x \in X \implies f(x) \in F(X)$ holds. The natural interval extension $F(X)$ often overestimates $f(X)$, but can be shown under mild conditions to linearly approach the range as the maximal diameter of the box X goes to zero, i.e., $\mathfrak{h}(F(X), f(X)) \leq \alpha \cdot d_\infty(X)$ for some $\alpha \geq 0$. This implies that a partition of X into smaller boxes $\{X1, \cdots, Xm\}$, as done in section 4, gives better enclosures of $f(X)$ through the union $\bigcup_{i=1}^{m} F(Xi)$. Let $\nabla F(x)$ and $\nabla^2 F(x)$ represent the interval extensions of $\nabla f(x)$ and $\nabla^2 f(x)$, the gradient and Hessian of f. A better enclosure of $f(X)$ is possible for an f with the centered form,

$$f(x) = f(c) + \nabla f(b) \cdot (x - c) \in f(c) + \nabla f(X) \cdot (x - c) \subseteq F_c(X) := f(c) + \nabla F(X) \cdot (X - c),$$

where, $b, c, x \in X$ with b between c and x. $F_c(X)$ is the interval extension of the centered form of f with center c and decays quadratically to $f(X)$ as the maximal diameter of $X \to 0$. Finally, some interval extensions of f are better than others. Recall the implications of subdistributivity of interval arithmetic, for instance.

2.2. *Differentiation Arithmetic*

When it becomes too cumbersome or impossible to explicitly compute the derivative of a function $f : \mathbb{R}^n \to \mathbb{R}$, or when f itself is only available as an algorithm, one may employ a differentiation arithmetic, often known as automatic differentiation (see for e.g. Griewank and Corliss (1991)) to obtain any $\nabla^k f$, the k_{th}-order derivative of f. This approach circumvents the computation of a formal expression for f by defining a differentiation arithmetic on the ordered k-tuples $(f(x), \nabla f(x), \nabla^2 f(x), \cdots, \nabla^k f(x))$ (Berz, 1991). A brief sketch of such an arithmetic is given for the case when $k = 2$ as it will be used in section 4.

Consider a twice-continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ with the gradient vector and Hessian matrix given by $\nabla f(x) := (\partial f(x)/\partial x_1, \cdots, \partial f(x)/\partial x_n)^T \in \mathbb{R}^n$, and $\nabla^2 f(x) := ((\partial^2 f(x)/\partial x_i \partial x_j))_{i,j=\{1,\cdots,n\}} \in \mathbb{R}^{n \times n}$, respectively. For every, $f(x) : \mathbb{R}^n \to \mathbb{R}$, consider its corresponding ordered triple $(\, f(x),\ \nabla f(x),\ \nabla^2 f(x)\,)$. The ordered triples corresponding to a constant function, $c(x) = c : \mathbb{R}^n \to \mathbb{R}$, and a component identifying function (or variable), $I_j(x) = x_j : \mathbb{R}^n \to \mathbb{R}$, are $(\, c,\ 0,\ 0\,)$ and $(\, x_j,\ e^{(j)},\ 0\,)$, respectively, where, $e^{(j)}$ is the j-th unit vector and the 0's are additive identities in their appropriate spaces. To perform an elementary operation $\circ \in \{+, -, \cdot, /\}$ with a pair of such triples to obtain another, the rules of calculus apply as follows:

$$\begin{aligned}
&(\, h(x),\ \nabla h(x),\ \nabla^2 h(x)\,) := (\, f(x),\ \nabla f(x),\ \nabla^2 f(x)\,) \circ (\, g(x),\ \nabla g(x),\ \nabla^2 g(x)\,) \\
&= (\, f(x) \circ g(x),\ \nabla f(x) \circ \nabla g(x),\ \nabla^2 f(x) \circ \nabla^2 g(x)\,), && \text{if } \circ \in \{+, -\} \\
&= (\, f(x) \cdot g(x),\ f(x) \cdot \nabla g(x) + g(x) \cdot \nabla f(x), \\
&\quad g(x) \cdot \nabla^2 f(x) + \nabla f(x) \cdot \nabla g(x)^T + \nabla g(x) \cdot \nabla f(x)^T + f(x) \cdot \nabla^2 g(x)\,), && \text{if } \circ = \cdot \\
&= (\, f(x)/g(x),\ 1/g(x) \cdot \{\nabla f(x) - h(x) \cdot \nabla g(x)\}, \\
&\quad 1/g(x) \cdot \{\nabla^2 f(x) \cdot \nabla h(x) \cdot \nabla h(x)^T - \nabla g(x) \cdot \nabla h(x)^T - h(x) \cdot \nabla^2 h(x)\}, && \text{if } \circ = /,\ g(x) \neq 0
\end{aligned}$$

The arithmetic for composition of functions, such as, $h(x) = r(f(x)) : \mathbb{R} \to \mathbb{R}$, with the first and second derivative of r given by r' and r'', on their corresponding triples, given by,

$$(\, h(x),\ \nabla h(x),\ \nabla^2 h(x)\,) = (\, r(f(x)),\ r'(f(x)) \cdot \nabla f(x),\ r''(f(x)) \cdot \nabla f(x) \cdot \nabla f(x)^T + r'(f(x)) \cdot \nabla^2 f(x)\,),$$

yields the triples for the elementary functions, $\exp(x)$ and $ln(x)$, which are used to compute the likelihood in section 3.

For dyadic reasons, the differentiation arithmetic has been explained above only in terms of reals. By replacing the real x's above by interval X's and performing all operations in the real interval arithmetic with the interval extension F of f, as discussed in section 2.1, one can rigorously enclose the components of the interval triple $(F(X),\ \nabla F(X),\ \nabla^2 F(X))$ through interval differentiation arithmetic, such that, for every $x \in X \in \mathbb{IR}^n$, $f(x) \in F(X) \in \mathbb{IR}$, $\nabla f(x) \in \nabla F(X) \in \mathbb{IR}^n$, and $\nabla^2 f(x) \in \nabla^2 F(X) \in \mathbb{IR}^{n \times n}$.

2.3. *Extended interval Newton method*

By including two ideal points $+\infty$ and $-\infty$ to $\mathbb{R}$, it becomes possible to extend interval arithmetic to $\mathbb{IR}^* := \mathbb{IR} \cup \{(-\infty, \overline{x}] : \overline{x} \in \mathbb{R}\} \cup \{[\underline{x}, +\infty) : \underline{x} \in \mathbb{R}\} \cup (-\infty, +\infty)$, the set of intervals with end points in the complete lattice $\mathbb{R}^* := \mathbb{R} \cup \{+\infty\} \cup \{-\infty\}$, with respect to the ordering relation $\leq$. Since, division is the inverse operation of multiplication, obtaining any $x/y \in X/Y := \{x/y : x \in X, y \in Y\}$ is equivalent to solving the equation $y \cdot s = x$ for s, i.e., $X/Y := \{s : y \cdot s = x, x \in X, y \in Y\}$. Let $[\,]$ denote the empty interval. With the following rules, division by intervals containing 0 becomes possible.

$$X/Y := \begin{cases} (-\infty, +\infty) & \text{if } 0 \in X, \text{or } Y = [0,0] \\ [\,] & \text{if } 0 \notin X, \text{and } Y = [0,0] \\ [\overline{x}/\underline{y}, +\infty) & \text{if } \overline{x} \leq 0, \text{ and } \overline{y} = 0 \\ [\underline{x}/\overline{y}, +\infty) & \text{if } 0 \leq \underline{x}, \text{ and } 0 = \underline{y} < \overline{y} \\ (-\infty, \overline{x}/\overline{y}] & \text{if } \overline{x} \leq 0, \text{ and } 0 = \underline{y} < \overline{y} \\ (-\infty, \underline{x}/\underline{y}] & \text{if } 0 \leq \underline{x}, \text{ and } \underline{y} < \overline{y} = 0 \\ (-\infty, \overline{x}/\overline{y}] \cup [\overline{x}/\underline{y}, +\infty) & \text{if } \overline{x} \leq 0, \text{ and } [0,0] \Subset Y \\ (-\infty, \underline{x}/\underline{y}] \cup [\underline{x}/\overline{y}, +\infty) & \text{if } 0 \leq \underline{x}, \text{ and } [0,0] \Subset Y \end{cases}$$

When X is a thin interval with $x = \underline{x} = \overline{x}$ and Y has $+\infty$ or $-\infty$ as one of its bounds, then extended interval subtraction is also necessary for the following interval Newton algorithm, and is defined as follows,

$$[\underline{x}, \overline{x}] - Y := \begin{cases} (-\infty, +\infty) & \text{if } Y = (-\infty, +\infty) \\ (-\infty, x - \underline{y}] & \text{if } Y = (\underline{y}, +\infty) \\ [x - \overline{y}, +\infty) & \text{if } Y = (-\infty, \overline{y}] \end{cases}$$

The analog of the Newton method that approximates a zero of a continuously differentiable scalar-valued real function f is the extended interval Newton method that encloses all the zeros of F, an interval extension of f, in any given interval X. The method sketched below uses the extended interval arithmetic described above and is a variant of the method based on Hansen and Sengupta (1981) with Ratz's modifications (Ratz, 1992) as implemented in Hammer et al. (1995). Due to the mean value theorem, $f(m(X)) - f(x^*) = \nabla f(c) \cdot (m(X) - x^*)$, for some $c, x^* \in X$. Interest in the solution x^*, such that, $f(x^*) = 0$, yields the following relations.

$$\begin{aligned} f(m(X)) &= \nabla f(c) \cdot (m(X) - x^*) \\ x^* &= m(X) - (\nabla f(c))^{-1} \cdot f(m(X)), \quad \text{provided } \forall\ \nabla f(c) \text{ is invertible} \\ &\in m(X) - (\nabla f(X))^{-1} \cdot f(m(X)) \\ &\subseteq m(X) - (\nabla F(X))^{-1} \cdot F(m(X)) =: \mathcal{N}(X) \\ &\subseteq \mathcal{N}(X) \cap X \end{aligned}$$

An iteration scheme $X^{(j+1)} := \mathcal{N}(X^{(j)}) \cap X^{(j)}$, where $j = 0, 1, \cdots$, and $X^{(0)} := X$, will enclose the zeros of f contained in the interval X. To relax the assumption that every matrix in $\nabla F(X)$ be invertible, the inverse of the midpoint of $\nabla F(X)$, i.e., $(m(\nabla F(X)))^{-1} =: p \in \mathbb{R}^{n \times n}$, is used as a matrix preconditioner. The extended interval Gauss-Seidel iteration is used to solve the preconditioned interval linear equation,

$$\begin{aligned} p \cdot f(m(X)) &= p \cdot \nabla f(X) \cdot (m(X) - x^*) \\ a &= G \cdot (c - x^*), \text{ where, } a \in A := p \cdot F(m(X)),\ G := p \cdot \nabla f(X), \text{ and, } c := m(X). \end{aligned}$$

Thus, the solution set $\mathbf{S} := \{x \in X : g \cdot (c - x) = a, \forall g \in G\}$ of the interval linear equation $a = G \cdot (c - x)$ has the component-wise solution set $\mathbf{S}_i = \{x_i \in X_i : \sum_{j=1}^{n} (g_{i,j} \cdot (c_j - x_j)) = a_i, \forall g \in G\}$, $\forall i \in \{1, \cdots, n\}$. Now, set $Y = X$, and solve the ith equation for the ith variable, iteratively for each i, as follows:

$$y_i = c_i - \frac{1}{g_{ii}} \left(a_i + \sum_{j=1, j \neq i}^{n} (g_{i,j} \cdot (y_j - c_j)) \right) \in \left(c_i - \frac{1}{G_{ii}} \left(A_i + \sum_{j=1, j \neq i}^{n} (G_{i,j} \cdot (Y_j - c_j)) \right) \right) \cap Y_i$$

The interval vector(s) Y obtained at the end of such an iteration is the set, $\mathcal{N}(X)$, resulting from one extended interval Newton Gauss-Seidel step, such that, $\mathbf{S} \subseteq \mathcal{N}(X) \subseteq X$. Every 0 of f that lies in X also lies in $\mathcal{N}(X)$. If $\mathcal{N}(X) = [\,]$, the empty interval, then f has no solution in X. If $\mathcal{N}(X) \Subset X$, then f has a unique solution in X. For proofs of the above three statements see Hansen (1992).

Note that when X contains two or more roots of f, the enclosure of its derivative $\nabla F(X)$, computed by differentiation arithmetic of section 2.2, will contain at least one zero, since it is the superset of all slopes of tangents to $f(x)$ over the interval X. Thus, one may obtain disjoint finite intervals subsequent to the Newton operator $\mathcal{N}(X)$ over X, that is obtained through extended interval arithmetic, being intersected with the finite interval X. In such cases, the iteration is applied to each resulting sub-interval. See Kulisch (2001) for a geometric interpretation of the interval Newton method.

2.4. Machine interval arithmetic

All interval arithmetic was done above with real intervals. However, there are only finitely many floating-point numbers available on a computing machine. Let $\mathcal{R}$ be this set of floating-point numbers. A machine interval is a real interval with floating-point bounds. Thus, on a computer, one works with $\mathbb{IR} := \{X \in \mathbb{IR} : \underline{x}, \overline{x} \in \mathcal{R}\}$, the set of all machine intervals. Inspite of the finiteness of $\mathbb{IR}$, the strength of interval arithmetic lies in a machine interval X being able to eclose the entire continuum of reals between its machine-representable boundaries. Through rounding controlled floating-point arithmetic provided by the IEEE arithmetic standard, operations with real intervals can be tightly enclosed by the rounding directed operations with the smallest machine intervals containing them. See Kulisch (2001) for a recent description of machine interval arithmetic. The errors resulting from converting a decimal number, usually a constant or input data, which in general does not have a finite binary representation, to a binary floating-point number is controlled by passing the decimal numbers as strings and then converting it to the floating-point number by proper rounding. The program is written in *C++* using the C-XSC class libraries. The differentiation arithmetic of section 2.2 is implemented using the *hess.ari* module provided in Hammer et al. (1995).

3. A PHYLOGENETIC PROBLEM

Let D denote a homologous set of distinct DNA sequences of length v from s species. The objective of this paper is to find the maximum likelihood estimates of branch lengths for the best tree under a particular topology. Recall that the branch lengths usually represent a scaled product of mutation rate and number of generations. Let b denote the number of branches and n denote the number of nodes of a tree with topology τ. For an s-leaved unrooted tree of a given topology, there are at most $2s - 3$ branches, i.e., $b \leq 2s - 3$. Since the number of topologies for multifurcating unrooted trees grows as a factorial of the number of leaves, it is difficult to exhaustively search through all possible topologies to find the most likely tree even when the number of leaves is reasonably small. For example, there are 12,818,912 topologies when $s = 10$. See chapter 3 of Felsenstein (2003) to appreciate this problem. However, one can find the most likely tree among a small set of specified topologies, by first computing the most likely tree under each topology of interest and then choosing the tree with the highest likelihood.

Thus, for a given unrooted topology τ with s leaves and b branches, the unknown parameter $\theta = (\theta_1, \cdots, \theta_b)$ is the real vector of branch lengths in the positive orthant, where each positive branch length $\theta_q \in [\theta_\delta, R] \to \mathbb{R}_+$, as $\theta_\delta \to 0$ and $R \to \infty$. An explicit model of DNA evolution is needed to construct the likelihood function which gives the probability of observing data D as a function of the parameter θ. The simplest such continuous time Markov chain model (JC69) on the state space $\mathcal{S} := \{A, G, C, T\}$ is due to Jukes and Cantor (1969) with the rate of mutation between nucleotides i and j given by $q_{i,j} = 1/3$, if $i \neq j$, and -1, otherwise. Its stationary distribution $\pi = (1/4, 1/4, 1/4, 1/4)$, and $P_{i,j}(t)$, the probability

of transition from i to j in time t is, $1/4 + 3/4\exp(-4t/3)$ if $i = j$, and $1/4 - 1/4\exp(-4t/3)$, otherwise. Felsenstein's algorithm (Felsenstein, 1981) to compute $\ell^{(k)}(\theta)$, the likelihood at site $k \in \{1, \cdots, v\}$, is the following postorder traversal:

(a) Associate with each node $q \in \{1, \cdots, n\}$ a real vector $\ell_q := (\ell_q^A, \ell_q^C, \ell_q^G, \ell_q^T) \in \mathbb{R}^4$, and let the length of the branch leading to its ancestor be θ_q.
(b) For a leaf node q with nucleotide i, set $\ell_q^i = 1$ and $\ell_q^j = 0$ for all $j \neq i$. For any internal node q, set $\ell_q := (1, 1, 1, 1)$.
(c) For an internal node q with descendants $s_1, s_2, \cdots, s_m$,

$$\ell_q^i = \sum_{j_1, \cdots, j_m \in \mathcal{S}} \{ \ell_{s_1}^{j_1} \cdot P_{i,j_1}(\theta_{s_1}) \cdot \ell_{s_2}^{j_2} \cdot P_{i,j_2}(\theta_{s_2}) \cdot \cdots \cdot \ell_{s_m}^{j_m} \cdot P_{i,j_m}(\theta_{s_m}) \}$$

(d) Compute ℓ_q for each sub-terminal node q, then those of their ancestors recursively to finally compute ℓ_r for the root node r and obtain $\ell^{(k)}(\theta) = \sum_{i \in \mathcal{S}} (\pi_i \cdot \ell_r^i)$ for each site k.

Assuming independence across sites one obtains the likelihood function for the entire sequence by multiplying the site-specific likelihoods together. The problem of finding the global maximum of this likelihood function is equivalent to finding the global minimum of $l(\theta)$, the negative of the natural logarithm of the likelihood function given by,

$$l(\theta) = -\sum_{k=1}^{v} \ln \ell^{(k)}(\theta).$$

$l(\theta)$ is of interest because algorithms in the optimization literature are usually addressed in terms of minimization. Replacing θ, a positive real vector of branch lengths, in the above algorithm by a positive real interval vector or box Θ and all real operations by their interval counterparts, yields $L(\Theta)$, the natural interval extension of the negative log likelihood function $l(\theta)$ over Θ. Since $\nabla L(\Theta)$ and $\nabla^2 L(\Theta)$, the enclosures of the gradient and the Hessian of $l(\theta)$ over Θ, respectively, are needed in Section 4, one may use the constant triples, $(C, 0, 0)$, variable triples, $(\Theta_j, e^{(j)}, 0)$, appropriate triples for the elementary functions, exp and ln, and perform all operations in the interval differentiation arithmetic of Section 2.2, in order to obtain the negative log likelihood triple $(L(\Theta), \nabla L(\Theta), \nabla^2 L(\Theta))$.

4. GLOBAL OPTIMIZATION

4.1. Branch-and-bound

The most basic strategy in global optimization through enclosure methods is to employ rigorous branch-and-bound techniques. Such techniques recursively partition (branch) the original compact space of interest into compact subspaces and discard (bound) those subspaces that are guaranteed to not contain the global optimizer(s). For the real scalar-valued multi-dimensional objective function $l(\theta)$, the interval branch-and-bound technique can be applied to its natural interval extension $L(\Theta)$ to obtain an interval enclosure L^* of the global minimum value l^* as well as the set of minimizer(s) to a specified accuracy ϵ. Note that this set of minimizer(s) of $L(\theta)$ is the set of maximizer(s) of the likelihood function for the observed data D. The strength of such methods arises from the algorithmic ability to discard large sub-boxes from the original search region,

$$\Theta^{(0)} = (\Theta_1^{(0)}, \cdots, \Theta_b^{(0)}) := ([\underline{\theta}_1^{(0)}, \overline{\theta}_1^{(0)}], \cdots, [\underline{\theta}_b^{(0)}, \overline{\theta}_b^{(0)}]) \subset \mathbb{IR}^b,$$

that are not candidates for global minimizer(s). Four tests that help discard sub-regions are described below. Let $\mathfrak{L}$ denote a list of ordered pairs of the form $(\Theta^{(i)}, \underline{L}_{\Theta^{(i)}})$, where, $\Theta^{(i)} \subseteq \Theta^{(0)}$, and $\underline{L}_{\Theta^{(i)}} := \min(L(\Theta^{(i)}))$ is a lower bound for the range of the negative log likelihood function l over $\Theta^{(i)}$. Let $\tilde{l}$ be an upper bound for l^* and $\nabla L(\Theta^{(i)})_k$ denote the k-th interval of the gradient box $\nabla L(\Theta^{(i)})$. If no information is available for $\tilde{l}$, then $\tilde{l} = \infty$.

(a) *Midpoint Cut-off test:* The basic idea of the *midpoint cut-off test* is to discard sub-boxes of the search space $\Theta^{(0)}$ with the lower bound for their range enclosures above $\tilde{l}$, the current best estimate of an upper bound for l^*. Figure 1 shows a multi-modal l as a function of a scalar valued θ over $\Theta^{(0)} = \cup_{i=1}^{9} \Theta^{(i)}$. For this illustrative example, $\tilde{l}$ is set as the upper bound of the range enclosure of l over the smallest

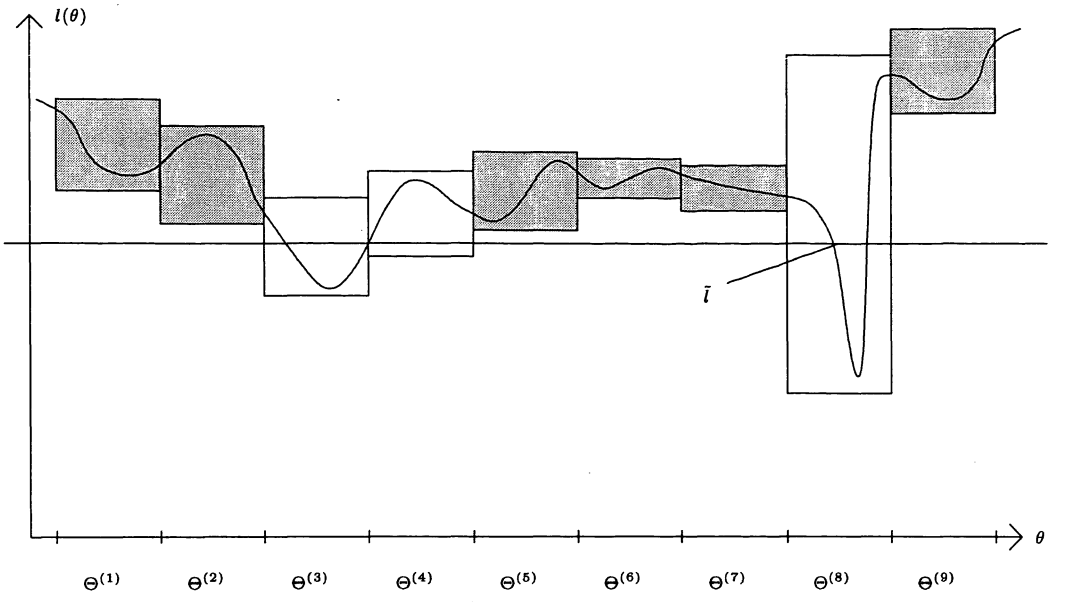

Fig. 1. *Midpoint Cut-off test*

machine interval containing the midpoint of $\Theta^{(8)}$, the interval with the smallest lower bound of its range enclosure. The *midpoint cut-off test* would discard the intervals $\Theta^{(1)}$, $\Theta^{(2)}$, $\Theta^{(5)}$, $\Theta^{(6)}$, $\Theta^{(7)}$, and $\Theta^{(9)}$ corresponding to the shaded boxes as the lower bound of their range enclosures is strictly above the current best estimate for an upper bound of l^*.

- Given a list $\mathfrak{L}$ and $\tilde{l}$
- Choose an element j of $\mathfrak{L}$, such that, $j = \underset{i}{argmin}\, \underline{L}_{\Theta^{(i)}}$, since $\Theta^{(j)}$ is likely to contain a minimizer.
- Find its midpoint $c = m(\Theta^{(j)})$ and let C be the smallest machine interval containing c.
- Compute a possibly improved $\tilde{l} = \min\{\tilde{l}, \overline{L}_C\}$, where, $\overline{L}_C := \max(L(C))$
- Discard any i-th element of $\mathfrak{L}$ for which $\underline{L}_{\Theta^{(i)}} > \tilde{l} \geq l^*$

(b) *Monotonicity test:* For a continuously differentiable function $l(\theta)$, the *monotonicity test* determines whether $l(\theta)$ is strictly monotone over an entire sub-box $\Theta^{(i)} \subset \Theta^{(0)}$. If l is strictly monotone over $\Theta^{(i)}$, then a global minimizer cannot lie in the interior of $\Theta^{(i)}$. Therefore, $\Theta^{(i)}$ can only contain a global minimizer as a boundary point if this point also lies in the boundary of $\Theta^{(0)}$. Figure 2 illustrates the *monotonicity test* for the one-dimensional case. In this example the search space of interest, $\Theta^{(0)} = [\underline{\theta}^{(0)}, \overline{\theta}^{(0)}] = \cup_{i=1}^{9}\Theta^{(i)}$, can be reduced considerably. One may delete $\Theta^{(2)}$, $\Theta^{(3)}$, $\Theta^{(5)}$, $\Theta^{(7)}$, and $\Theta^{(8)}$ since $l(\theta)$ is monotone over them and they belong to the interior of $\Theta^{(0)}$. Since $l(\theta)$ is monotonically increasing over $\Theta^{(9)}$ one can also deleted it since we are only interested in minimization. $\Theta^{(1)}$ may be pruned to its left boundary point $\theta^{(1)} = \underline{\theta}^{(0)}$ due to the strictly decreasing nature of $l(\theta)$ over it. Thus, the *monotonicity test* has pruned $\Theta^{(0)}$ to the smaller candidate set $\{\underline{\theta}^{(0)}, \Theta^{(4)}, \Theta^{(6)}\}$ for a global minimizer.

- Given $\Theta^{(0)}$, $\Theta^{(i)}$, and $\nabla L(\Theta^{(i)})$
- Iterate for $k = 1, \cdots, b$
  - If $0 \in \nabla L(\Theta^{(i)})_k$, then leave $\Theta_k^{(i)}$ unchanged, as it may contain a stationary point of l.
  - Otherwise, $0 \notin \nabla L(\Theta^{(i)})_k$. This implies that $\Theta^{(i)}$ can be pruned, since $l^* \notin \Theta^{(i)}$ except possibly at the boundary points, as follows:

    (i) if $\min(\nabla L(\Theta^{(i)})_k) > 0$ and $\underline{\theta}_k^{(0)} = \underline{\theta}_k^{(i)}$, then $\Theta_k^{(i)} = [\underline{\theta}_k^{(i)}, \underline{\theta}_k^{(i)}]$,

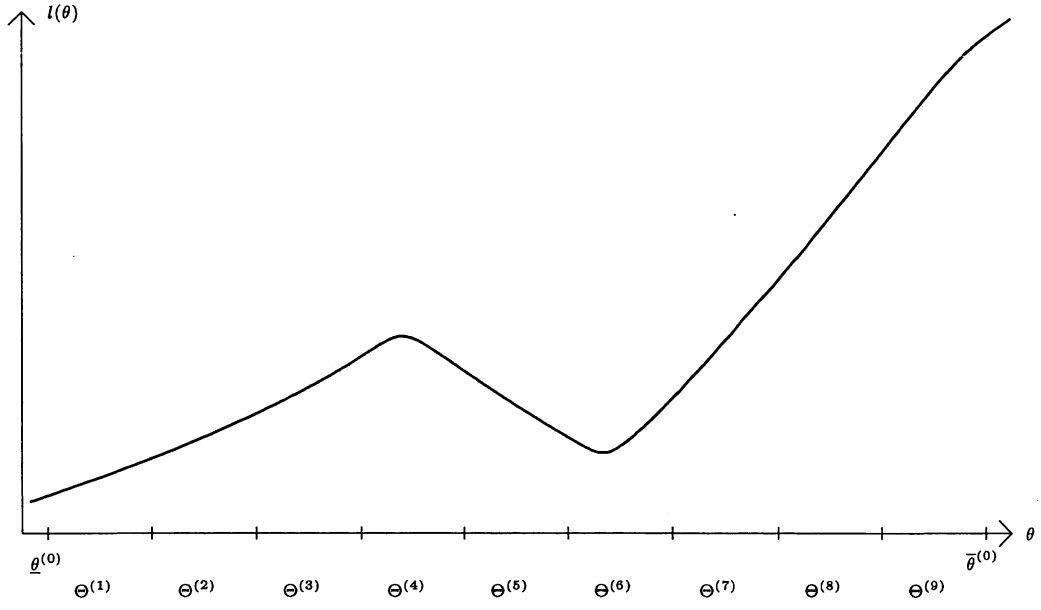

Fig. 2. *Monotonicity test*

(ii) Else if $\max(\nabla L(\Theta^{(i)})_k) < 0$ and $\overline{\theta}_k^{(0)} = \overline{\theta}_k^{(i)}$, then $\Theta_k^{(i)} = [\overline{\theta}_k^{(i)}, \overline{\theta}_k^{(i)}]$.
(iii) Else, delete the i-th element of $\mathfrak{L}$ and stop the iteration.

(c) *Concavity test:* Given $\Theta^{(i)} \Subset \Theta^{(0)}$, and the diagonal elements $(\nabla^2 L(\Theta^{(i)}))_{kk}$ of $\nabla^2 L(\Theta^{(i)})$, note that if $\min((\nabla^2 L(\Theta^{(i)}))_{kk}) < 0$ for some k, then, $\nabla^2 L(\Theta^{(i)})$ cannot be positive semidefinite, and therefore $l(\theta)$ cannot be convex over $\Theta^{(i)}$ and thus cannot contain a minimum in its interior. In the one-dimensional example shown in Figure 2, an application of the *concavity test* to the candidate set $\{\underline{\theta}^{(0)}, \Theta^{(4)}, \Theta^{(6)}\}$ for a global minimizer returned by the *monotonocity test*, would result in the deletion of $\Theta^{(4)}$ due to the concavity of $l(\theta)$ over it.

- Given $\Theta^{(i)} \Subset \Theta^{(0)}$ and $\nabla^2 L(\Theta^{(i)})$
- If $\min((\nabla^2 L(\Theta^{(i)}))_{kk}) < 0$ for any $k \in \{1, \cdots, b\}$, then delete the i-th element of $\mathfrak{L}$.

(d) *Interval Newton test:* Given $\Theta^{(i)} \Subset \Theta^{(0)}$, and $\nabla L(\Theta^{(i)})$, attempt to solve the system, $\nabla L(\theta) = 0$, in terms of $\theta \in \Theta^{(i)}$.

- Apply one extended interval Newton Gauss-Seidel step of Section 2.3 to the linear interval equation $a = G \cdot (c - \theta)$, where, $a := p \cdot L(m(\Theta^{(i)}))$, $G := p \cdot \nabla^2 L(\Theta^{(i)})$, $c := m(\Theta^{(i)})$, and $p := (m(\nabla^2 F(X)))^{-1}$, in order to obtain $\mathcal{N}'(\Theta^{(i)})$.
- One of the following can happen,

(i) If $\mathcal{N}'(\Theta^{(i)})$ is empty, then discard $\Theta^{(i)}$.
(ii) If $\mathcal{N}'(\Theta^{(i)}) \Subset \Theta^{(i)}$, then replace $\Theta^{(i)}$ by the contraction $\mathcal{N}'(\Theta^{(i)}) \cap \Theta^{(i)}$.
(iii) If $0 \in G_{jj}$, and the extended interval division splits $\Theta_j^{(i)}$ into a non-empty union of $\Theta_j^{(i),1}$ and $\Theta_j^{(i),2}$, then the iteration is continued on $\Theta_j^{(i),1}$, while $\Theta_j^{(i),2}$, if non-empty, is stored in $\mathfrak{L}$ for future processing. Thus, one extended interval Newton Gauss-Seidel step can add at most $b+1$ sub-boxes to $\mathfrak{L}$.

4.2. Verification

Given a collection of sub-boxes, $\{\Theta^{(1)}, \cdots, \Theta^{(n)}\}$, each of width $\leq \epsilon$, that could not be discarded by the tests in Section 4.1, one can attempt to verify the existence and uniqueness of a local minimizer within each

sub-box $\theta^{(i)}$ by checking whether the conditions of the following two theorems are satisfied. For proof of these two theorems see Hansen (1992) and Ratz (1992).

(a) If $\mathcal{N}'(\Theta^{(i)}) \Subset \Theta^{(i)}$, then there exists a unique stationary point of L, i.e., a unique zero of ∇L exists in $\Theta^{(i)}$.
(b) If $(I + \frac{1}{\kappa} \cdot (\nabla^2 L(\Theta^{(i)}))) \cdot Z \Subset Z$, where $(\nabla^2 L(\Theta^{(i)}))_{d,\infty} \leq \kappa \in \mathbb{R}$, for some $Z \in \mathbb{IR}^n$, then, the spectral radius $\rho(s) < 1$ for all $s \in (I - \frac{1}{\kappa} \cdot (\nabla^2 L(\Theta^{(i)})))$, and all symmetric matrices in $\nabla^2 L(\Theta^{(i)})$ are positive definite.

If the conditions of the above two theorems are satisfied by some $\Theta^{(i)}$, then a unique stationary point exists in $\Theta^{(i)}$ and this stationary point is a local minimizer. Therefore, if exactly one candiate sub-box for minimizer(s) remained after pruning the search box $\Theta^{(0)}$ with the tests in Section 4.1, and if this sub-box satisfies the above two conditions for the existence of a unique local minimizer within it, then one has rigorously enclosed the global minimizer in the search interval. On the other hand, if there are two or more sub-boxes in our candidate list for minimizer(s) that satisfy the above two conditions, then one may conclude that each sub-box contains a candidate for a global minimizer which may not necessarily be unique (disconnected sub-boxes, for example). Observe that failure to verify the uniqueness of a local minimizer in a sub-box can occur if it contains two or more points or even a continuum of points that are stationary (nonidentifiable manifolds in the sub-box, for example).

4.3. Algorithm

- *Initialization:*

(a) Let the search region be a single box $\Theta^{(0)}$ or a collection of not necessarily connected, but pair-wise disjoint boxes, $\Theta^{(i)}$, $i \in \{1, \cdots, r\}$.
(b) Initialize the list $\mathfrak{L}$ which may just contain one element $(\Theta^{(0)}, \underline{L}_{\Theta^{(0)}})$ or several elements

$$\{(\Theta^{(1)}, \underline{L}_{\Theta^{(1)}}), (\Theta^{(2)}, \underline{L}_{\Theta^{(2)}}), \cdots, (\Theta^{(r)}, \underline{L}_{\Theta^{(r)}})\}.$$

(c) Let ϵ be a specified tolerance.
(d) Let $\max_{\mathfrak{L}}$ be the maximal length allowed for list $\mathfrak{L}$.
(e) Set the noninformative lower bound for l^*, i.e., $\tilde{l} = \infty$

- *Iteration:*

(a) (i) Improve $\tilde{l} = \min\{\tilde{l}, \max(L(m(\Theta^{(j)})))\}$, where $j = \underset{i}{argmin}\{\underline{L}_{\Theta^{(i)}}\}$.
(ii) Perform the *midpoint cut-off test* to $\mathfrak{L}$.
(iii) Set $L^* = [\underline{L}_{\Theta^{(j)}}, \tilde{l}]$.
(b) Bisect $\Theta^{(j)}$ along its longest side k, i.e., $d(\Theta_k^{(j)}) = d_\infty(\Theta^{(j)})$, to obtain sub-boxes $\Theta^{(j_q)}$, $q \in \{1, 2\}$.
(c) For each sub-box $\Theta^{(j_q)}$, evaluate its triple $(L(\Theta^{(j_q)}), \nabla L(\Theta^{(j_q)}), \nabla^2 L(\Theta^{(j_q)}))$, and do the following:

(i) Perform *monotonicity test* to possibly discard $\Theta^{(j_q)}$.
(ii) *Centered form cut-off test:*
Improve the range enclosure of $L(\Theta^{(j_q)})$ by replacing it with its centered form $L_c(\Theta^{(j_q)})$,

$$L_c(\Theta^{(j_q)}) := \{L(m(\Theta^{(j_q)})) + \nabla L(\Theta^{(j_q)}) \cdot (\Theta^{(j_q)} - m(\Theta^{(j_q)}))\} \cap L(\Theta^{(j_q)}),$$

and then discarding $\Theta^{(j_q)}$, if $\tilde{l} < \underline{L}_{\Theta^{(j_q)}}$.
(iii) Perform *concavity test* to possibly discard $\Theta^{(j_q)}$.
(iv) Apply an *extended interval Newton Gauss-Seidel step* to $\Theta^{(j_q)}$, in order to either entirely discard it or shrink it into v sub-sub-boxes, where v is at most $2s - 2$.
(v) For each one of these sub-sub-boxes $\Theta^{(j_{q,u})}$, $u \in \{1, \cdots, v\}$

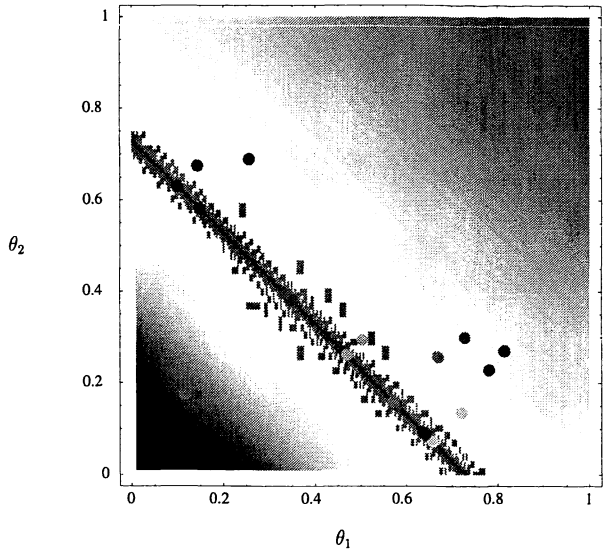

Fig. 3. For a pair of homologous sequences of 600 nucleotides out of which 280 sites are polymorphic, the nonidentifiable subspace of minimizers $\theta_1 + \theta_2 = \frac{3}{4}\log(45/17) = 0.730087$ of the negative log likelihood function under the JC69 model evolving on a rooted two-leaved tree is enclosed by a union of upto 30,000 boxes. The larger grey, and smaller black boxes have tolerances of $\epsilon = 1.0 \times 10^{-4}$ and $\epsilon = 1.0 \times 10^{-6}$, respectively. The 10 pairs of colored circles are the initial and final points of 10 BFGS searches with random initializations.

A. Perform *monotonicity test* to possibly discard $\Theta^{(j_{q,u})}$.
B. Try to discard $\Theta^{(j_{q,u})}$ by applying the *centered form cut-off test* in cii to it.
C. Append $(\Theta^{(j_{q,u})}, \underline{L}_{\Theta^{(j_{q,u})}})$ to $\mathfrak{L}$ if $\Theta^{(j_{q,u})}$ could not be discarded by steps c(v)A and c(v)B.

- *Termination:*

(a) Terminate iteration if $d_{rel,\infty}(\Theta^{(j)}) < \epsilon$, or $d_{rel,\infty}(L^*) < \epsilon$, or $\mathfrak{L}$ is empty, or $\text{Length}(\mathfrak{L}) > \max_{\mathfrak{L}}$
(b) Verify uniqueness of minimizer(s) in the final list $\mathfrak{L}$ by applying algorithm of section 4.2 to each of its elements.

5. APPLICATIONS

5.1. Enclosing nonidentifiable subspaces

For time reversible Markov chains, such as JC69, evolving on a rooted tree, only the sum of the branch lengths emanating from the root is identifiable. Identifiability is a prerequisite for statistical consistency of estimators. To demonstrate the ability of interval methods, unlike the local search methods, to enclose the nonidentifiable ridge along $\theta_1 + \theta_2$, in the simplest case of a 2-leaved tree, a nonidentifiable negative log likelihood function $l(\theta)$ is formulated and its global minimizers along $\theta_1 + \theta_2 = \frac{3}{4}\log(45/17) = 0.730087$ are enclosed as shown in Figure 3 for a fictitious dataset for which 280 out of 600 sites were polymorphic. Observe that the basin of attraction for each point on $\theta_1 + \theta_2 = 0.730087$ under the BFGS local search algorithm is the line running orthogonal to it. This trivial example is only chosen for pedantic reasons. Enclosing possibly nonidentifiable submanifolds, that may not even be simply connected, within any compact subset of higher dimensional parameter spaces, may be accomplished, at least partly, by studying the rates of decay of the hyper-volume of the union of all pending boxes as the algorithm progresses, for instance.

Table 1. Enclosures of the Maximum log likelihood and their corresponding parameter estimates for 3 taxa tree relating Chimpanzee, Gorilla, and Orangutan.

Tree	$\Theta^{(0)}$	$\Theta^* \supset \theta^*$	$-L(\Theta^*) \supset -l(\theta^*)$
star	$[1.0 \times 10^{-11}, 1.0 \times 10^{9}]^{\otimes 3}$	$5.98162213840^{2}_{0} \times 10^{-2}$	
		$5.41674167940^{2}_{0} \times 10^{-2}$	
		$1.32990896859^{9}_{8} \times 10^{-1}$	$-2.15031806585656^{5}_{6} \times 10^{3}$

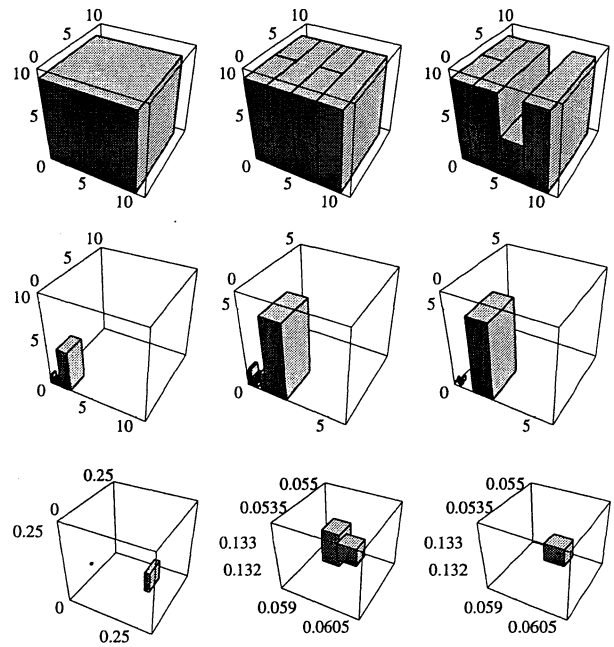

Fig. 4. Progress of the algorithm as it prunes $[0.001, 10.0]^{\otimes 3}$.

5.2. Unrooted 3-leaved Tree

The global maximum of the log likelihood function for the JC69 model of DNA evolution on the three taxa unrooted tree with data from the mitochondria of Chimpanzee, Gorilla, and Orangutan (Brown et al., 1982) is enclosed. There is only one unrooted multifurcating topology for three species with all three branches emanating from the root like a star. The data set for this problem can be summarized by the following 29 data patterns:

```
PATTERN COUNTS :
232 71 229 168 13 31 16 18 9 20 1 8 22 3 10 8 7 1 9 2 4 2 1 2 1 1 2 1 3
PATTERNS:
agctatcacccatctgccgtactaagcgt
agctgttatcaacacgcaaaatccggtat
agctaccgttcccataataataaagcgca
```

The parameter space is three dimensional corresponding to the three branch lengths of the 3-leaved tree. The algorithm is given a large search box $\Theta^{(0)}$. The results are summarized in Table 1. The notation x^b_a means the interval $[xa, xb]$, for e.g. $5.98162213840^{2}_{0} \times 10^{-2} = [5.98162213840 \times 10^{-2}, 5.98162213842 \times 10^{-2}]$. Figure 4 shows the the parameter space being rigorously pruned as the algorithm progresses according to section 4.

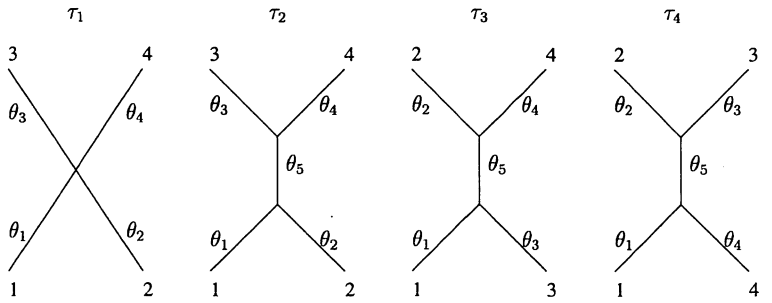

Fig. 5. The four different topologies, τ_1, τ_2, τ_3, and τ_4 with 4 leaves. The four leaves 1, 2, 3, and 4 denote the four primates Chimpanzee, Gorilla, Orangutan, and Gibbon, respectively.

5.3. Four Unrooted 4-leaved Trees

By adding the homologous mitochondrial sequence from Gibbon to the previous problem, one has the simplest phylogeny estimation problem with the following 61 data patterns:

```
PATTERN COUNTS :
209 71 192 157 28 5 11 20 2 10 10 15 5 15 1 5 1 2 15 3 14 3 4 2 5 4 5
4 9 2 4 5 1 1 7 6 3 1 1 4 2 3 3 1 1 2 3 1 1 1 4 1 1 1 1 1 1 1 1 2 1
PATTERNS:
agctccatatcacctacaaatccatatctgtccgggattacccctatttacgcgtcgcttc
agctccgtgttatctaaaaacatacacctgccaaagaatactttcccatgcgtattacctt
agctccacaccgttcaccaacatctccctatataaagattaccaaacatgtcgcagattaa
agcttaacgtcaccactcgtcatgcatgcgtcctaagtgaatacaaaaactgaaagcaata
```

Four topologies are considered for a tree with four leaves (Figure 5). The star tree τ_1 has all four lineages coalescing at the same time, while the other three trees have an additional parameter θ_5 representing the only internal branch length. They differ due to the order in which the leaves relate to one another as shown in Figure 5. The parameter space is four dimensional for the star topology τ_1, and five dimensional for each of the unrooted topologies τ_2, τ_3, and τ_4.

Observe that τ_1 is really a special case of the other three trees, τ_2, τ_3, and τ_4, as their internal branch θ_5 vanishes. Since we assume that the branches are $\geq \theta_\delta > 0$, and let $\theta_\delta \to 0$ on a sequence of floating-point numbers, it is convenient to treat the star tree τ_1 separately. The algorithm is given a large search box $\Theta^{(0)}$ for each topology and the results are summarized in Table refT:4runs.

Within each one of the four topologies there exists a unique global maximum. However, the global maximizer over all five topologies falls under topology τ_2 with the global maximum $-l^*$ contained in the interval $-L^* = -2.656936470946^5_6 \times 10^3$.

6. CONCLUSIONS

A general procedure has been provided to rigorously enclose the maximum likelihood value, as well as the most likely branch lengths of a tree (with a specified topology) upon which the simplest Markov model of DNA evolution is superimposed. The global optimization algorithm is general and can be used by frequentists to conduct rigorous numerical inference in a likelihood framework when the statistical experiment is indexed by a compact subset of a finite dimensional continuum. This procedure is not suceptible to errors caused by *undirected rounding, catastrophic cancellation, discretization, conversion,* and *ill-posed model.* Modifications of this algorithm are also applicable to Markov models with unknown parameters. For models that have parameter constraints, several constrained global optimization algorithms already exist (Hansen, 1992). When analytical spectral decompositions are not available for more complicated Markov models, one may use one of several rigorous eigen system solvers (e.g. Mayer (1994)) to compute the transition probabilities.

Table 2. Enclosures of the Maximum log likelihood and their corresponding parameter estimates for 4 taxa trees relating Chimpanzee, Gorilla, Orangutan, and Gibbon.

Tree	$\Theta^{(0)}$	$\Theta^* \supset \theta^*$	$-L(\Theta^*) \supset -l(\theta^*)$
τ_1	$[1.0 \times 10^{-11}, 1.0 \times 10^{9}]^{\otimes 4}$	$6.57882493333^{5}_{4} \times 10^{-2}$	
		$6.2361625124032^{8}_{2} \times 10^{-2}$	
		$1.324874902248^{5}_{4} \times 10^{-1}$	
		$1.6359125624763^{4}_{3} \times 10^{-1}$	$-2.7027434501964^{1}_{4} \times 10^{3}$
τ_2	$[1.0 \times 10^{-11}, 1.0 \times 10^{9}]^{\otimes 5}$	$4.962819343269^{9}_{8} \times 10^{-2}$	
		$5.899264246907^{8}_{7} \times 10^{-2}$	
		$5.518490773873^{4}_{3} \times 10^{-2}$	
		$9.097140075962^{3}_{2} \times 10^{-2}$	
		$1.2315160183101^{2}_{1} \times 10^{-1}$	$-2.6569364709465^{5}_{8} \times 10^{3}$
τ_3	$[1.0 \times 10^{-11}, 1.0 \times 10^{9}]^{\otimes 5}$	$9.071770476^{7}_{6} \times 10^{-3}$	
		$6.1423911113^{4}_{3} \times 10^{-2}$	
		$1.2963838225^{5}_{4} \times 10^{-1}$	
		$5.6506921810^{3}_{0} \times 10^{-2}$	
		$1.600054316561^{5}_{1} \times 10^{-1}$	$-2.699878136170^{0}_{5} \times 10^{3}$
τ_4	$[1.0 \times 10^{-11}, 1.0 \times 10^{9}]^{\otimes 5}$	$1.1495164302969^{9}_{5} \times 10^{-2}$	
		$5.825806134317^{8}_{7} \times 10^{-2}$	
		$1.5888166092521^{3}_{1} \times 10^{-1}$	
		$5.7069581801992^{9}_{2} \times 10^{-2}$	
		$1.2932141694890^{1}_{0} \times 10^{-1}$	$-2.6985586285405^{5}_{9} \times 10^{3}$

The *extended interval Newton* method in combination with the *midpoint cut-off*, *monotonicity*, and *concavity* tests may be used to study the shape of the likelihood surface itself. For instance, one could rigorously enclose all the local maxima, or fish for nonidentifiable subspaces, above a given level-set of the likelihood function within any compact subset of the parameter space. Several efficiency increasing steps could be taken. Pre-enclosing the transition probabilities and accessing them through hash functions can save computational effort. Asynchronous parallelization of the algorithm across 6 processors is also observed to increase the rate of convergence to the global maximum. It also provides a natural framework to manage the memory requirements for larger trees through partial likelihood evaluations for non-overlapping subtrees in parallel prior to obtaining the full likelihood.

Finally, it is worth noting that inclusion isotony does indeed hold by the continuity of the likelihood function in the CAT(0) space of trees (Billera et al., 2004), and thus in conjunction with interval analysis may be made to provide a rigorous numerical framework for global maximizaton of the likelihood over compact sets containing distinct topologies. Preliminary results indicate that efficiency increases when one starts with a disjoint union of compact subsets of branch lengths from finitely many topologies, i.e, $\Theta^{(0)} = \cup_i \Theta^{(0,\tau_i)}$ and simultaneously prunes away sub-boxes from distinct $\Theta^{(0,\tau_i)}$ with a variant of the above algorithm that allows for compact sets contained in each $\Theta^{(0,\tau_i)}$ with its corresponding τ_i-specific post-order traversal to specify its topology-specific likelihood function. Thus, interval methods may be able to enclose the global maximum more efficiently when several topologies are considered simultaneously than when the global maximum is enclosed for each member of a finite set of topologies, one at a time, and finally compared.

7. Acknowledgements

The seeds for this work were sown in the inter-disciplinary environment of integrative graduate education and research traineeship program in complex non-linear systems funded by the NSF grant DGE-9870631 and completed with support from a joint NSF/NIGMS grant to Durrett, Aquadro, and Nielsen DMS 0201037. Many thanks to Dave Capella for insights into accelerations based on local search(es), Mike Steel for warning about the nonstationary optima at boundaries, Warwick Tucker for an introduction to interval analysis,

Tandy Warnow for posing the problem, Wendy Shuk Wan Wong for sharing local search and data parsing codes, and Ziheng Yang for discussions about analytical results.

References

Berz, M. (1991). Forward algorithms for high orders and many variables with application to beam physics. See Griewank and Corliss (1991), pp. 147–156.

Billera, L. J., S. Holmes, and K. Vogtmann (2004). Geometry of the space of phylogenetic trees. *Advances in Applied Math.*.

Brown, W. M., E. M. Prager, A. Wang, and A. C. Wilson (1982). Mitochondrial DNA sequences of primates, tempo and mode of evolution. *Jnl. Mol. Evol. 18*, 225–239.

Chor, B. (2000). Multiple maxima of likelihood in phylogenetic trees: An analytic approach. *Mol. Biol. Evol. 17*, 1529–1541.

Cuyt, A., B. Verdonk, S. Becuwe, and P. Kuterna (2001). A remarkable example of catastrophic cancellation unraveled. *Computing 66*, 309–320.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Jnl. Mol. Evol. 17*, 368–376.

Felsenstein, J. (2003). *Inferring phylogenies.* Sunderland, MA: Sinauer Associates.

Griewank, A. and G. Corliss (Eds.) (1991). *Automatic differentiation of algorithms: theory, implementation and applications.* Philadelphia: SIAM.

Hammer, R., M. Hocks, U. Kulisch, and D. Ratz (1995). *C++ toolbox for verified computing: basic numerical problems.* Berlin: Springer-Verlag.

Hansen, E. (1980). Global optimization using interval analysis - the multi-dimensional case. *Numerische Mathematik 34*, 247–270.

Hansen, E. (1992). *Global optimization using interval analysis.* New York: Marcel Dekker.

Hansen, E. and S. Sengupta (1981). Bounding solutions of systems of equations using interval analysis. *BIT 21*, 203–211.

Jukes, T. H. and C. R. Cantor (1969). Evolution of protein molecules. In H. N. Munro (Ed.), *Mammalian protein metabolism*, Volume 3, pp. 21–123. New York: Academic Press.

Kulisch, U. (2001). Advanced arithmetic for the digital computer, interval arithmetic revisited. See Kulisch et al. (2001), pp. 50–70.

Kulisch, U., R. Lohner, and A. Facius (Eds.) (2001). *Perspectives on encolsure methods.* New York: Springer-Verlag.

Loh, E. and G. W. Walster (2002). Rump's example revisited. *Reliable Computing 8*, 245–248.

Mayer, G. (1994). Result verification for eigenvectors and eigenvalues. In J. Herzberger (Ed.), *Topics in validated computations*, Volume 5 of *Studies in computational mathematics*, pp. 209–276. New York: North-Holland.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery (1992). *Numerical Recipes: The Art of Scientific Computing.* Cambridge, UK: Cambridge University Press.

Ratz, D. (1992). *Automatische ergebnisverifikation bei globalen optimierungsproblemen.* pHd dissertation, Universitat Karlsruhe, Karlsruhe.

Yang, Z. (2000). Complexity of the simplest phylogenetic estimation problem. *Proceedings Royal Soc. London B Biol. Sci. 267*, 109–119.