# An Auto-validating Rejection Sampler for Differentiable Arithmetical Expressions: Posterior Sampling of Phylogenetic Quartets

Raazesh Sainudiin

Laboratory for Mathematical Statistical Experiments and
Department of Mathematics and Statistics
University of Canterbury,
Private Bag 4800, Christchurch, New Zealand 8041
r.sainudiin@math.canterbury.ac.nz
http://www.math.canterbury.ac.nz/~r.sainudiin

**Abstract.** We introduce an efficient extension of a recently introduced auto-validating rejection sampler that is capable of producing independent and identically distributed (IID) samples from a large class of target densities with locally Lipschitz arithmetical expressions. Our extension is restricted to target densities that are differentiable. We use the centered form, as opposed to the natural interval extension, to get tighter range enclosures of the differentiable multivariate target density using interval extended gradient differentiation arithmetic. By using the centered form we are able to sample one hundred times faster from the posterior density over the space of phylogenetic trees with four leaves (quartets).

**Keywords:** interval analysis, centered form, phylogentic inference

## Introduction

Obtaining independent and identical (IID) samples or realizations from a random vector $T$ with probability density function $f^\cdot$, denoted $T \sim f^\cdot$, is a basic problem in computational statistics. The density $f^\cdot(t) : \mathbb{T} \subseteq \mathbb{R}^d \to [0, \infty)$ allows one to obtain $\mathbb{P}(T \in B) = \int_B f^\cdot(t)dt$, the probability that $T$ belongs to any Borel set $B$. The density is absolutely continuous with respect to $\lambda^d$, the product of $d$ Lebesgue measures, i.e., $f^\cdot \ll \lambda^d$, and integrates to 1, i.e., $\int_\mathbb{T} f^\cdot(t)dt = 1$. A *sampler* is a randomized algorithm that transforms independent and identically distributed (IID) samples from $M$, the uniformly distributed random variable on the unit interval, to those from the desired random object, say the random vector $T$ with density $f^\cdot$.

In Bayesian estimation, we want to draw IID samples from a target posterior density $f^\cdot$ and more generally, in multivariate simulation, we want to draw IID samples from a random vector $T$ with probability density $f^\cdot$. These samples allow insights into the nature of the random vector itself. They are often used to estimate an integral of interest about the random vector, say, $\mathbb{E}_{f^\cdot}(h(T)) := \int_\mathbb{T} h(t)f^\cdot(t)dt$, where $h(t) : \mathbb{T} \to \mathbb{R}$ is bounded and $\mathbb{E}_{f^\cdot}(h^2(T)) < \infty$, using the

estimator $\widehat{h}_n = n^{-1} \sum_{i=1}^{n} h(t_i)$, where $t_1, t_2, \ldots, t_n$ are IID samples from $T$ with density $f^{\cdot}$. For example, such integrals of interest can be the posterior mean given by $\int_{\mathbb{T}} t f^{\cdot}(t) dt$ with $h(t) = t$, or the probability of an event $A$ given by $\mathbb{P}(A) := \int_{\mathbb{T}} \mathbb{1}_A(t) f^{\cdot}(t) dt$, with $h(t) = \mathbb{1}_A(t)$, where $\mathbb{1}_A(t)$ equals 1 if $t \in A$ and 0 otherwise. Due to the strong law of large numbers our estimator $\widehat{h}_n$ converges to the desired $\mathbb{E}_{f^{\cdot}}(h(T))$ with probability 1 as the number of samples $n$ approaches infinity. Furthermore, the condition $\mathbb{E}_{f^{\cdot}}(h^2(T)) < \infty$ ensures the asymptotic normality of our estimator due to the central limit theorem and provides a straightforward calculation of a confidence interval.

In Bayesian phylogenetic estimation, we want to draw independent and identically distributed samples from a target posterior density on the space of phylogenetic trees. The standard approaches to sampling from the posterior density, especially over phylogenetic trees, rely on Markov chain Monte Carlo (MCMC) methods. Despite their asymptotic validity, it is nontrivial to guarantee that an MCMC algorithm has converged to stationarity [1], and thus MCMC convergence diagnostics on phylogenetic tree spaces are heuristic and may lead to meaningless estimates [2].

A more direct method for simulating IID samples from a random variable $T$ with density $f^{\cdot}(t)$ is the *Rejection Sampler* (RS) of von Neumann [3]. RS can produce IID samples from the target density $f^{\cdot}(t) := f(t)/(N_f)$ by only evaluating the *target shape* $f(t)$ — without knowing the normalising constant $N_f := \int_{\mathbb{T}} f(t) dt$. Briefly, the idea behind RS is as follows: produce a point uniformly distributed in the $(d+1)$-dimensional region under an envelope function that is strictly greater than or equal to the target shape and if this point is below the target shape then accept its first $d$ coordinates in $\mathbb{T}$ as a sample from $T$, otherwise reject it and try again.

RS can produce samples from $T \sim f^{\cdot}$ according to Algorithm 1 when provided with (i) a fundamental sampler that can produce independent samples from the Uniform$[0, 1]$ random variable $M$ with density $\mathbb{1}_{[0,1]}(m) : \mathbb{R} \mapsto \mathbb{R}$, (ii) a target shape $f(t) : \mathbb{T} \mapsto \mathbb{R}$, (iii) an envelope function $\widehat{g}(t) : \mathbb{T} \mapsto \mathbb{R}$, such that,

$$\widehat{g}(t) \geq f(t) \quad \text{for all } t \in \mathbb{T} \ , \tag{1}$$

(iv) a normalizing constant $N_{\widehat{g}} := \int_{\mathbb{T}} \widehat{g}(t) dt$, (v) a proposal density $g(t) := (N_{\widehat{g}})^{-1} \widehat{g}(t)$ over $\mathbb{T}$ from which independent samples can be drawn and finally (vi) $f(t)$ and $\widehat{g}(t)$ must be computable for any $t \in \mathbb{T}$.

The random variable $T$, if generated by Algorithm 1, is distributed according to $f^{\cdot}$ (e.g. [4]). Let $\boldsymbol{A}(\widehat{g})$ be the probability that a point proposed according to $g$ gets accepted as an independent sample from $f^{\cdot}$ through the envelope function $\widehat{g}$. Observe that the envelope-specific acceptance probability $\boldsymbol{A}(\widehat{g})$ is the ratio of the integrals

$$\boldsymbol{A}(\widehat{g}) = \frac{N_f}{N_{\widehat{g}}} := \frac{\int_{\mathbb{T}} f(t) \, dt}{\int_{\mathbb{T}} \widehat{g}(t) \, dt} \ ,$$

and the probability distribution over the number of samples from $g$ to obtain one sample from $f^{\cdot}$ is geometrically distributed with mean $1/\boldsymbol{A}(\widehat{g})$ (e.g. [4]).

---

**Algorithm 1:** von Neumann RS

   **input**    : (i) $f$; (ii) samplers for $V \sim g$ and $M \sim \mathbf{1}_{[0,1]}$; (iii) $\widehat{g}$; (iv) integer MaxTrials;

   **output**  : (i) possibly one sample $t$ from $T \sim f^{\cdot}$ and (ii) Trials

   **initialize**: Trials $\leftarrow 0$; Success $\leftarrow$ false; $t \leftarrow \emptyset$;

   **repeat**                         `// propose at most MaxTrials times until acceptance`
      |  $v \leftarrow \texttt{sample}(g)$ ;          `// draw a sample v from RV V with density g`
      |  $u \leftarrow \widehat{g}(v)\ \texttt{sample}(\mathbf{1}_{[0,1]})$;    `// draw a sample u from RV U with density` $\mathbf{1}_{[0,\widehat{g}(v)]}$
      |  **if** $u \le f(v)$ **then**          `// accept the proposed v and flag Success`
      |  |  $t \leftarrow v$; Success $\leftarrow$ true
      |  **end**
      |  Trials $\leftarrow$ Trials $+1$ ;        `// track the number of proposal trials so far`
   **until** Trials $\ge$ MaxTrials *or* Success = true;
   **return** $t$ *and* Trials

---

The crucial step in RS is the construction of an envelope function $\widehat{g}(t)$ that is not only greater than the target shape $f(t) := N_f f^{\cdot}(t)$ at every $t \in \mathbb{T} \subseteq \mathbb{R}^d$, but also easy to normalise and draw samples from. Moreover, a practical and efficient envelope function has to be as close to the target shape as possible from above in order to ensure a sufficiently high acceptance probability.

Moore rejection sampler (MRS) of [5, 6] uses the natural interval extension of $f$ over an adaptive partition of $\mathbb{T}$ to rigorously produce IID samples from the posterior distribution over phylogenetic tree spaces. Informally, MRS [5, 6] partitions the domain into boxes and uses interval analysis to rigorously enclose the range of the target shape in each box; then it uses as envelope the piece-wise constant function given by the upper bound of the range in each box. More formally, the method employs the natural interval extension of the target posterior shape $f(t) : \mathbb{T} \mapsto \mathbb{R}$ to produce rigorous enclosures of the range of $f$ over each interval vector or box in an adaptive partition $\mathfrak{T} := \{\boldsymbol{t}^{(1)}, \boldsymbol{t}^{(2)}, \ldots, \boldsymbol{t}^{(|\mathfrak{T}|)}\}$ of the tree space $\mathbb{T} = \cup_i \boldsymbol{t}^{(i)}$. This partition is adaptively constructed by a priority queue. The interval extended target shape maps boxes in $\mathbb{T}$ to intervals in $\mathbb{R}$. This image interval provides an upper bound for the global maximum and a lower bound for the global minimum of $f$ over each element of the partition of $\mathbb{T}$. This information is used to construct an envelope as a simple function over the partition $\mathfrak{T}$. Using the Alias method [7] samples are proposed efficiently from this normalized piece-wise constant function envelope for von Neumann rejection sampling. Unlike many conventional samplers, each sample produced by MRS is equivalent to a computer-assisted proof that it is drawn from the desired target density.

**An improved approach**

Suppose our target shape $f : \mathbb{T} \to \mathbb{R}$, with $\mathbb{T} \subseteq \mathbb{R}^d$, is differentiable. Then the centered form [8, 9] (of first order) of $f$ over a box $\boldsymbol{x} \in \mathbb{T}$ is

$$\boldsymbol{f}_c(\boldsymbol{x}) := (f(c) + [\nabla f(\boldsymbol{x})](\boldsymbol{x} - c)) \cap \boldsymbol{f}(\boldsymbol{x})$$

where, $c := \mathrm{mid}\,(\boldsymbol{x})$ is the midpoint of the box $\boldsymbol{x}$, $\boldsymbol{f}(\boldsymbol{x})$ is the natural interval extension of $f$ over $\boldsymbol{x}$ and $[\nabla f(\boldsymbol{x})]$ is the enclosure of the gradient of $f$ over $\boldsymbol{x}$. For any box $\boldsymbol{x} \subseteq \mathbb{T}$ it is well known that the centered form encloses the range, i.e.,

$$f(\boldsymbol{x}) := \{f(x) : x \in \boldsymbol{x}\} \subseteq \boldsymbol{f}_c(\boldsymbol{x}) \subseteq \boldsymbol{f}(\boldsymbol{x})\ \ ,$$

and the range enclosure given by $\boldsymbol{f}_c(\boldsymbol{x})$ is usually sharper than $\boldsymbol{f}(\boldsymbol{x})$ especially as $\boldsymbol{x}$ shrinks. Instead of using the natural interval extension $\boldsymbol{f}(\boldsymbol{x})$ to enclose $f(\boldsymbol{x})$, the range of $f$ over $\boldsymbol{x}$, as done in [5, 6], our new approach employs the centered form $\boldsymbol{f}_c(\boldsymbol{x})$ to enclose the range of the target shape over each box $\boldsymbol{x}$ in the partition of $\mathbb{T}$.

We use gradient differentiation arithmetic using interval-extended automatic differentiation (e.g. [10]) to obtain $[\nabla f(\boldsymbol{x})]$, the enclosure of the gradient of $f$ over $\boldsymbol{x}$. We use the implementation of this arithmetic in the `grad_ari` module of `C-XSC 2.0`, a `C++` class library for extended scientific computing [11]. Using the centered form to bound the range in MRS allows us to sample from the posterior density over the space of phylogenetic quartets 100 times faster than using the natural interval extension alone as in [5]. An open source `C++` class library for MRS is available from `www.math.canterbury.ac.nz/~r.sainudiin/codes/mrs` under the terms of the GNU general public license (GPL).

**Phylogenetic Estimation**

In this section we briefly review phylogenetic estimation as we will apply our improved method to sample from challenging phylogenetic posterior densities. Introduction to phylogenetics can be found in [12–14]. Inferring the ancestral relationship among a set of extant (presently surviving) species based on their DNA sequences is a basic problem in phylogenetic estimation. A phylogenetic tree relates the extant species represented by its leaf nodes with ancestral species represented by its internal nodes. The topology or shape of the tree specifies the order of speciation or branching events. The length of an edge (branch length) connecting two nodes (species) in the phylogenetic tree represents the amount of evolutionary time (divergence) between the two species as measured by the differences in their DNA sequence due to mutation. One can obtain the likelihood of a particular phylogenetic tree that relates the extant species of interest at its leaves by superimposing a continuous time Markov chain model of DNA mutation along the lengths of the branches on that tree. During the likelihood computation, one needs to sum over all possible states of the DNA sequence at the unobserved ancestral nodes. In [5] MRS was used to draw IID posterior samples from small phylogenetic tree spaces of the same dimension (number of

branches) based on primate DNA sequence data. In [6] this was generalized to the trans-dimensional setting where the number of branch length parameters are allowed to vary between models of phylogenetic trees. However, only the natural interval extension of the posterior density was used in [5] and [6] to obtain upper bounds for the range before Moore rejection sampling. Here, we will employ interval extended gradient differentiation arithmetic to obtain much tighter enclosures of the posterior density, which is the product of a uniform prior density and the likelihood function over phylogenetic trees.

**Likelihood of a phylogenetic tree** Let $d$ denote a homologous set of sequences of length $v$ with character set $\mathfrak{A} = \{a_1, a_2, \ldots, a_{|\mathfrak{A}|}\}$ from $n$ taxa. We think of $d$ as an $n \times v$ matrix with entries from $\mathfrak{A}$. We are interested in estimating the branch lengths and topologies of the tree underlying our observed $d$. Let $b_k$ denote the number of branches and $s_k$ denote the number of nodes of a tree with a specific topology or branching order labeled by $k$. Thus, for a given topology label $k$, $n$ labeled leaves and $b_k$ many branches, the labeled tree $^k t$ is the topology-labeled vector of branch lengths $(^k t_1, \ldots, ^k t_{b_k})$ contained in the topology-labeled tree space $^k \mathbb{T}$, i.e.,

$$^k \mathbb{T} := \{^k t := (^k t_1, \ldots, ^k t_{b_k}) \in \mathbb{R}_+^{b_k} : {^k t_i} > 0 \text{ for terminal branches}\} \ .$$

The tree space with $|\mathfrak{K}|$ many topologies in the topology label set $\mathfrak{K}$ can be defined as follows:

$$^{\mathfrak{K}} \mathbb{T} := \bigcup_{k \in \mathfrak{K}} {^k \mathbb{T}} \ .$$

An explicit model of sequence evolution is prescribed in order to obtain the likelihood of observing data $d$ at the leaf nodes as a function of the parameter $^k t \in {^{\mathfrak{K}} \mathbb{T}}$ for each topology label $k \in \mathfrak{K}$. Such a model prescribes $P_{a_i, a_j}(t)$, the probability of mutation from a character $a_i \in \mathfrak{A}$ to another character $a_j \in \mathfrak{A}$ in time $t$. Using such a transition probability we may compute $\ell_q(^k t)$, the log-likelihood of the data $d$ at site $q \in \{1, \ldots, v\}$ or the $q$-th column of $d$, via the post-order traversal over the labeled tree with branch lengths $^k t := (^k t_1, {^k t_2}, \ldots, {^k t_{b_k}})$. This amounts to the sum-product Algorithm 2 [15] that associates with each node $h \in \{1, \ldots, s_k\}$ of $^k t$ subtending $\hbar$ many descendants, a partial likelihood vector, $l_h := (l_h^{(a_1)}, l_h^{(a_2)}, \ldots, l_h^{(a_{|\mathfrak{A}|})}) \in \mathbb{R}^{|\mathfrak{A}|}$, and specifies the length of the branch leading to its ancestor as $^k t_h$.

Assuming independence across all $v$ sites we obtain the likelihood function for the given data $d$, by multiplying the site-specific likelihoods

$$l_d(^k t) = \prod_{q=1}^{v} l_{d_{.,q}}(^k t) \ . \tag{2}$$

The maximum likelihood estimate is a point estimate (single best guess) of the unknown phylogenetic tree on the basis of the observed data $d$ and it is

$$\operatorname*{argmax}_{^k t \in {^{\mathfrak{K}} \mathbb{T}}} l_d(^k t) \ .$$

---

**Algorithm 2:** Likelihood by post-order traversal

    **input**     : (i) a labeled tree with branch lengths $^k t := (^k t_1, {}^k t_2, \ldots, {}^k t_{b_k})$, (ii) transition probability $P_{a_i,a_j}(t)$ for any $a_i, a_j \in \mathfrak{A}$, (iii) stationary distribution $\pi(a_i)$ over each character $a_i \in \mathfrak{A}$, (iv) site pattern or data $d_{\cdot,q}$ at site $q$

    **output**  : $l_{d_{\cdot,q}}(^k t)$, the likelihood at site $q$ with pattern $d_{\cdot,q}$

    **initialize**: For a leaf node $h$ with observed character $a_i = d_{h,q}$ at site $q$, set $l_h^{(a_i)} = 1$ and $l_h^{(a_j)} = 0$ for all $j \neq i$. For any internal node $h$, set $l_h := (1, 1, \ldots, 1)$.

    **recurse**  : compute $l_h$ for each sub-terminal node $h$, then those of their ancestors recursively to finally compute $l_r$ for the root node $r$ to obtain the likelihood for site $q$,

$$l_{d_{\cdot,q}}(^k t) = l_r = \sum_{a_i \in \mathfrak{A}} \left( \pi(a_i) \cdot l_r^{(a_i)} \right) .$$

For an internal node $h$ with descendants $s_1, s_2, \ldots, s_{\hbar}$,

$$l_h^{(a_i)} = \sum_{j_1, \ldots, j_{\hbar} \in \mathfrak{A}} \{ l_{s_1}^{(j_1)} \cdot P_{a_i,j_1}(^k t_{s_1}) \cdot l_{s_2}^{(j_2)} \cdot P_{a_i,j_2}(^k t_{s_2}) \ldots l_{s_{\hbar}}^{(j_{\hbar})} \cdot P_{a_i,j_{\hbar}}(^k t_{s_{\hbar}}) \}.$$

---

The simplest probability models for character mutation are continuous time Markov chains with finite state space $\mathfrak{A}$. We introduce the simplest such model with just two characters as it is thought to well-represent the core problems in phylogenetic estimation (see for e.g. [16]).

**Posterior density of a tree** The posterior density $f^{\cdot}(^k t)$ conditional on data $d$ at tree $^k t$ is the normalized product of the likelihood $l_d(^k t)$ and the prior density $p(^k t)$ over a given tree space $^{\mathfrak{K}}\mathbb{T}$:

$$f^{\cdot}(^k t) = \frac{l_d(^k t) p(^k t)}{\int_{^{\mathfrak{K}}\mathbb{T}} l_d(^k t) p(^k t) \, \partial(^k t)} \ . \tag{3}$$

We assume a uniform prior density over a large box or a union of large boxes in a given tree space $^{\mathfrak{K}}\mathbb{T}$. Typically, the sides of the box giving the range of branch lengths, are extremely long, say, $[0, 10]$ or $[10^{-10}, 10]$. The branch lengths are measured in units of expected number of DNA substitutions per site and therefore the support of our uniform prior density over $^{\mathfrak{K}}\mathbb{T}$ contains the biologically relevant branch lengths. If $^{\mathfrak{K}}\mathbb{T}$ is a union of distinct topologies then we let our prior be an equally weighted finite mixture of uniform densities over large boxes in each topology. Naturally, other prior densities are possible especially in the presence of additional information. We choose flat priors for the convenient interpretation of the target posterior shape $f(^k t) = f^{\cdot}(^k t) \int_{^{\mathfrak{K}}\mathbb{T}} l_d(^k t) p(^k t) \, \partial(^k t)$ to

be the likelihood function in the absence of prior information beyond a compact support specification.

**Likelihood under Cavender-Farris-Neyman (CFN) model** We now describe the simplest model for the evolution of binary sequences under a symmetric transition matrix over all branches of a tree. This model has been used by authors in various fields including molecular biology, information theory, operations research and statistical physics; for references see [17, 12]. This model is referred to as the Cavender-Farris-Neyman (CFN) model in molecular biology, although in other fields it has been referred to as 'the on-off machine', 'symmetric binary channel' and the 'symmetric two-state Poisson model'.

Under the CFN mutation model, only pyrimidines and purines, denoted respectively by $\mathtt{Y} := \{\mathtt{C}, \mathtt{T}\}$ and $\mathtt{R} := \{\mathtt{A}, \mathtt{G}\}$, are distinguished as evolutionary states among the four nucleotides $\{\mathtt{A}, \mathtt{G}, \mathtt{C}, \mathtt{T}\}$, i.e. $\mathfrak{A} = \{\mathtt{Y}, \mathtt{R}\}$. Time $t$ is measured by the expected number of substitutions in this homogeneous continuous time Markov chain with rate matrix:

$$Q = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} \ ,$$

and transition probability matrix $P(t) = e^{Qt}$ :

$$P(t) = \begin{pmatrix} 1 - (1 - e^{-2t})/2 & (1 - e^{-2t})/2 \\ (1 - e^{-2t})/2 & 1 - (1 - e^{-2t})/2 \end{pmatrix} \ .$$

Thus, the probability that $\mathtt{Y}$ mutates to $\mathtt{R}$, or vice versa, in time $t$ is $a(t) := (1 - e^{-2t})/2$. The stationary distribution is uniform on $\mathfrak{A}$, i.e. $\pi(\mathtt{R}) = \pi(\mathtt{Y}) = 1/2$.

To get a concrete idea of the likelihood function let us consider the case when there are only three taxa. Consider the unrooted tree space with a single topology labeled 0 and three non-negative terminal branch lengths ${}^0t = ({}^0t_1, {}^0t_2, {}^0t_3) \in \mathbb{R}^3_+$ as shown in Figure 1 $(i)$. There are $2^3 = 8$ possible site patterns, i.e. for each site $q \in \{1, 2, \ldots, v\}$, the $q$-th column of the data $d$, denoted by $d_{\bullet,q}$, is one of eight possibilities, numbered $0, 1, \ldots, 7$ for convenience:

$$d_{\bullet,q} \in \begin{Bmatrix} 0 \ , \ 1 \ , \ 2 \ , \ 3 \ , \ 4 \ , \ 5 \ , \ 6 \ , \ 7 \\ \\ \mathtt{R} \quad \mathtt{Y} \quad \mathtt{R} \quad \mathtt{Y} \quad \mathtt{R} \quad \mathtt{Y} \quad \mathtt{R} \quad \mathtt{Y} \\ \mathtt{R} \ , \ \mathtt{Y} \ , \ \mathtt{R} \ , \ \mathtt{Y} \ , \ \mathtt{Y} \ , \ \mathtt{R} \ , \ \mathtt{Y} \ , \ \mathtt{R} \\ \mathtt{R} \quad \mathtt{Y} \quad \mathtt{Y} \quad \mathtt{R} \quad \mathtt{Y} \quad \mathtt{R} \quad \mathtt{R} \quad \mathtt{Y} \end{Bmatrix} \ . \tag{4}$$

Given a multiple sequence alignment data $d$ from 3 taxa at $v$ homologous sites, i.e. $d \in \{\mathtt{Y}, \mathtt{R}\}^{3 \times v}$, Algorithm 2 can be used to compute the likelihood of any tree ${}^0t \in {}^0\mathbb{T}$ as follows:

$$l_d({}^kt) = \prod_{q=1}^v l_{d_{\bullet,q}}({}^kt) = \prod_{i=0}^7 \left(l_i({}^kt)\right)^{c_i} \ , \tag{5}$$

where $l_i({}^kt)$ is the likelihood of the the $i$-th site pattern as in (4) and $c_i$ is the count of sites with pattern $i$. In fact, $l_i({}^kt) = P(i|{}^kt)$ is the probability of

observing site pattern $i$ given topology label $k$ and branch lengths $t$ and similarly $l_d(^k t) = P(d|^k t)$. By using gradient differentiation arithmetic in Algorithm 2 we can obtain the enclosure of $l_d(^k \boldsymbol{t})$, the likelihood of a box $^k \boldsymbol{t}$ in the tree space $^k \mathbb{T}$, using the centered form. When there are four taxa we have three unrooted topologies, each with five branch length parameters, as shown in Figure 1 (*ii*), (*iii*) and (*iv*). In this case there are sixteen site patterns and we can similarly obtain the likelihood for each topology using Algorithm 2.
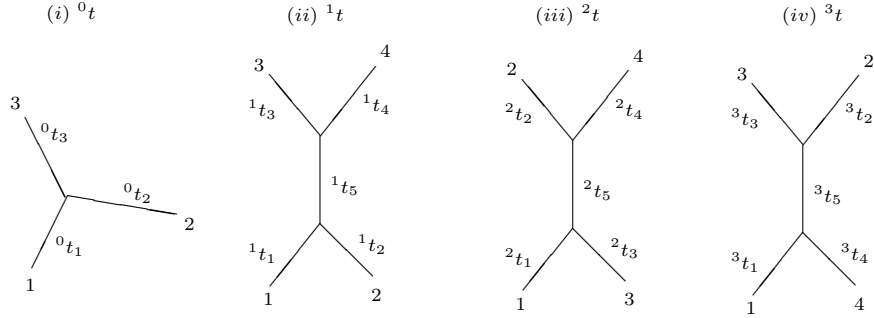


**Fig. 1.** The only topology of the unrooted tree with three taxa is shown in (*i*) and the three unrooted trees on four taxa with five branch length parameters each are shown by (*ii*), (*iii*) and (*iv*), respectively.

### Efficiency of MRS with Centered Form

For typical data sets on three taxa, including those in [6, Table 3], posterior samples of size $10^5$ using a partition of size $10^3$ can be obtained about 100 times faster by using the centered form as opposed to the natural interval extension of the likelihood function. This speed-up is particularly significant for typical data sets on four taxa as posterior samples of size $10^6$ are available in about $30 - 150$ minutes as opposed to a few days.

We were able to produce $10^7$ posterior samples in under two hours for all of the synthetic data sets in [18]. These data sets are well-known in phylogenetics for producing multiple maxima. The matrix plot of the first $10^4$ posterior samples in tree space $^1 \mathbb{T}$ is shown in Figure 2 for the following site pattern counts from [18, Proof of Thm. 2]:

$$c_{\text{YYYY}} = 1400, \qquad c_{\text{RYYY}} = 1, \qquad c_{\text{RYRR}} = 1, \qquad c_{\text{RRYY}} = 300,$$
$$c_{\text{RRYR}} = 1, \qquad c_{\text{RYRY}} = 200, \qquad c_{\text{YRRY}} = 100, \qquad c_{\text{RRRY}} = 1 \ . \qquad (6)$$

The likelihood function for this data is known to attain the maximum at two distinct points in $^1 \mathbb{T}$. This is evident from the two clusters of posterior samples in the matrix plot of Figure 2.
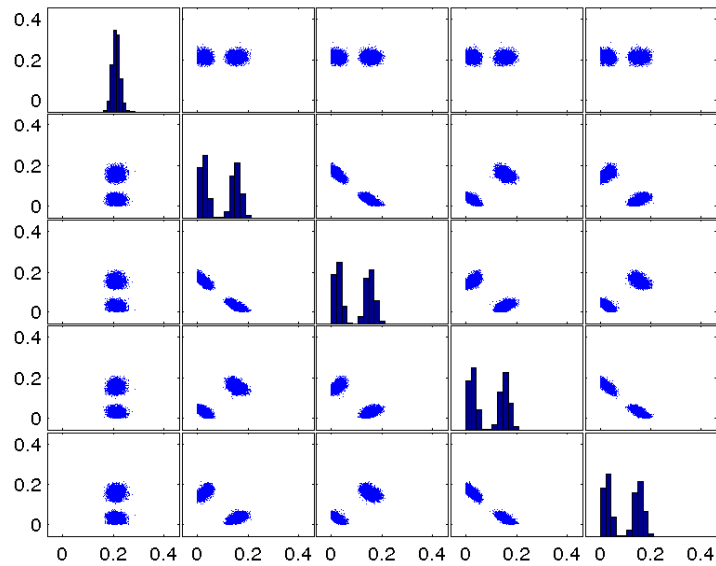
**Fig. 2.** The matrix plot of $10^4$ posterior samples for the site pattern counts in (6).

## Conclusion

In this paper, we use the centered form of the posterior density using gradient differentiation arithmetic to get tighter range enclosures and thereby increase the acceptance probability of the naive Moore rejection sampler [5, 6] that is only based on the natural interval extension of the posterior density. When we apply this centered form to produce IID samples from phylogenetic posterior densities over three and four taxa tree spaces we observe a hundred-fold speedup. Higher-order centered forms [9] in conjunction with constraint propagation [19] may further improve the sampler efficiency enough to produce IID posterior samples from five taxa phylogenetic trees with fifteen topologies in seven dimensions.

## References

1. Jones, G., Hobert, J.: Honest exploration of intractable probability distributions via Markov chain Monte Carlo. Statistical Science **16**(4) (2001) 312–334
2. Mossel, E., Vigoda, E.: Phylogenetic MCMC algorithms are misleading on mixtures of trees. Science **309** (2005) 2207–2209
3. von Neumann, J.: Various techniques used in connection with random digits. In: John Von Neumann, Collected Works. Volume V. Oxford University Press (1963)
4. Williams, D.: Weighing the Odds: A Course in Probability and Statistics. Cambridge University Press (2001)

5. Sainudiin, R., York, T.: Auto-validating von Neumann rejection sampling from small phylogenetic tree spaces. Algorithms for Molecular Biology **4** (2009) 1
6. Sainudiin, R., York, T.: An auto-validating, trans-dimensional, universal rejection sampler for locally Lipschitz arithmetical expressions. Reliable Computing **18** (2013) 15–54
7. Walker, A.J.: An efficient method for generating discrete random variables with general distributions. ACM Trans. Math. Softw. **3** (1977) 253–256
8. Moore, R.E.: Interval analysis. Prentice-Hall (1967)
9. Ratschek, H.: Centered forms. SIAM Journal on Numerical Analysis **17**(5) (1980) pp. 656–662
10. Rall, L.: Automatic differentiation, techniques and applications. Volume 120 of Springer lecture notes in computer science. Springer-Verlag (1981)
11. Hofschuster, W., Krämer, W.: C-XSC 2.0: A C++ library for extended scientific computing. In Alt, R., Frommer, A., Kearfott, R., Luther, W., eds.: Numerical Software with Result Verification. Volume 2991 of Lecture Notes in Computer Science. Springer-Verlag (2004) 15–35
12. Semple, C., Steel, M.: Phylogenetics. Oxford University Press (2003)
13. Felsenstein, J.: Inferring phylogenies. Sinauer Associates, Sunderland, MA (2003)
14. Yang, Z.: Computational Molecular Evolution. Oxford University Press, UK (2006)
15. Felsenstein, J.: Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. **17** (1981) 368–376
16. Yang, Z.: Complexity of the simplest phylogenetic estimation problem. Proceedings Royal Soc. London B Biol. Sci. **267** (2000) 109–119
17. Evans, W., Kenyon, C., Peres, Y., Schulman, L.J.: Broadcasting on trees and the Ising model. Advances in Applied Probability **10** (2000) 410–433
18. Chor, B., Hendy, M., Holland, B., Penny, D.: Multiple maxima of likelihood in phylogenetic trees: An analytic approach. Mol. Biol. Evol. **17** (2000) 1529–1541
19. Schichl, H., Neumaier, A.: Interval analysis on directed acyclic graphs for global optimization. Journal of Global Optimization **33**(4) (2005) 541–562