Inverting random functions II: explicit bounds for parametric and non-parametric MLE, with applications. *

Michael A. Steel
Biomathematics Research Centre
University of Canterbury
Christchurch, New Zealand
E-mail: m.steel@math.canterbury.ac.nz

László A. Székely
Department of Mathematics
University of South Carolina
Columbia, South Carolina
E-mail: szekely@math.sc.edu

Abstract

In this paper we study inverting random functions under the maximum likelihood estimation (MLE) criterion. In particular, we consider how many independent evaluations of the random function at a particular element of the domain are needed for reliable reconstruction of that element. We provide explicit upper and lower bounds for MLE, both in the non-parametric and parametric setting, and give applications to coin-tossing and phylogenetic tree reconstruction.

*Michael A. Steel was supported by the New Zealand Marsden Fund. László A. Székely was supported by the NSF grants DMS 9701211 and 0072187, by a grant of SC CHE, and the Hungarian NSF grant T 032455. This research started when the second author visited the University of Canterbury under the support of the New Zealand Marsden Fund. The authors are indebted to Éva Czabarka for her invaluable comments on the manuscript. MAS also thanks Joe Chang for suggesting the usefulness of the Hellinger distance in the proof of Theorem 3.

1 Review of random functions

This paper is a sequel of our earlier paper [11]. We assume that the reader is familiar with that paper, however, we repeat the most important definitions.

For two finite sets, A and U, let us be given a U-valued random variable ξ_a for every $a \in A$. We call the vector of random variables $(\xi_a : a \in A)$ a *random function* $\Xi : A \to U$. Ordinary functions are specific instances of random functions. It is easy to see [11] that an equivalent definition of random functions is obtained by picking one of the $|U|^{|A|}$ ordinary functions from A to U according to some distribution.

Given another random function, Γ, from U to V, we can speak about the composition of Γ and Ξ, $\Gamma \circ \Xi : A \to V$, which is the vector variable $(\gamma_{\xi_a} : a \in A)$. In this paper we are concerned with inverting random functions. In other words, we look for random functions $\Gamma : U \to A$ in order to obtain the best approximations of the identity function $\iota : A \to A$ by $\Gamma \circ \Xi$. *We always assume that* Ξ *and* Γ *are independent.* This assumption holds for free if either Ξ or Γ is a deterministic function.

Our motivation for the study of random functions came from phylogeny reconstruction. Stochastic models define how biomolecular sequences are generated at the leaves of a binary tree. If all possible binary trees on n leaves come equipped with a model for generating biomolecular sequences of length k, then we have a random function from the set of binary trees with n leaves to the ordered n-tuples of biomolecular sequences of length k. *Phylogeny reconstruction* is a random function from the set of ordered n-tuples of biomolecular sequences of length k to the set of binary trees with n leaves. It is a natural assumption that random mutations in the past are independent from any random choices in the phylogeny reconstruction algorithm. Criteria for phylogeny reconstruction may differ according to what one wishes to optimize.

Consider the probability of returning a from a by the composition of two random functions, that is, $r_a = \mathbb{P}[\gamma_{\xi_a} = a]$. The assumption on the independence of Ξ and Γ immediately implies

$$r_a = \sum_{u \in U} \mathbb{P}[\xi_a = u] \cdot \mathbb{P}[\gamma_u = a]. \tag{1}$$

A natural criterion is to find Γ for a given Ξ in order to maximize $\sum_a r_a$. More generally, we may have a weight function $w : A \to \mathbb{R}^+$ and we may wish to maximize $\sum_a r_a w(a)$. This can happen if we give preference to returning certain a's, or, if we have a prior probability distribution on A and we want to maximize the expected return probability for a random element of A selected according to the prior distribution. A random function $\Gamma^* : U \to A$ can be defined in the following way: for any fixed $u \in U$,

$$\gamma_u^* = a^* \text{ for sure, if for all } a \in A, \quad \mathbb{P}[\xi_{a^*} = u]w(a^*) \geq \mathbb{P}[\xi_a = u]w(a). \tag{2}$$

In case there is more than one element a^* that satisfies (2), we may select uniformly at random from the set of such elements. This function Γ^* is called the *maximum a posteriori estimator* (MAP) in the literature [8]. The special case when the weight function w is constant, is known as the *maximum likelihood estimation* (MLE) [2, 8]. We showed in [11] that the MAP estimator Γ^* maximizes $\sum_a r_a w(a)$ for any given Ξ. However, it is at least as natural to look at a more conservative criterion: maximize the smallest value of r_a for $a \in A$. For this criterion MAP or MLE is, in general, not optimal. These results have been known in the context of statistical decision theory [2] but have not been discussed in the context of phylogeny before [11].

Our paper [11] introduced a new abstract model for phylogeny reconstruction: inverting parametric random functions. Most of the work done on the mathematics of phylogeny reconstruction can be discussed in this context. This model is more structured than random functions, and hence is better suited to describe details of models of phylogeny and the evolution of biomolecular sequences. The approach is likely to be applicable in other areas where "nuisance" parameters are involved.

Assume that for a finite set A, for every $a \in A$, a (arbitrary, finite or infinite) set $\Theta(a) \neq \emptyset$ is assigned, and moreover, $\Theta(a) \cap \Theta(b) = \emptyset$ for $a \neq b$. Set $B = \{(a, \theta) : \ a \in A, \theta \in \Theta(a)\}$ and let π_1 denote the natural projection from B to A. A *parametric random function* is the collection Ξ of random variables such that

(i) for $a \in A$ and $\theta \in \Theta(a)$, there is a (unique) U-valued random variable $\xi_{(a,\theta)}$ in Ξ.

We are interested in random functions $\Gamma : U \to A$ independent from Ξ so that $\gamma_{\xi_{(a,\theta)}}$ best approximates π_1 under certain criteria. Call $R_{(a,\theta)}$ the probability $\mathbb{P}[\gamma_{\xi_{(a,\theta)}} = a]$. Maximum Likelihood Estimation, as it is used in the practice of phylogeny reconstruction, would take the Γ', for which for every fixed u, $\gamma'_u = a'$ for sure, if

$$\forall (a, \theta) \in B \ \ \exists \theta' \in \Theta(a') \ \ \mathbb{P}[\xi_{(a',\theta')} = u] \geq \mathbb{P}[\xi_{(a,\theta)} = u] \tag{3}$$

In case there is more than one element a' that satisfies (3), we may select uniformly at random from the set of such elements. We denote by $R'_{(a,\theta)}$ the probability that from the pair (a, θ) the Maximum Likelihood Estimation Γ' returns a, i.e.

$$R'_{(a,\theta)} = \mathbb{P}[\gamma'_{\xi_{(a,\theta)}} = a]. \tag{4}$$

In [11] we made further assumptions on parametric random functions that we do not make in this paper:

(ii) there is a measure space $(\Theta(a), \mu_a(.))$ defined on every $\Theta(a)$, such that $\mu_a(\Theta(a)) < \infty$;

(iii) for all $u \in U$, and for all $a \in A$, $\mathbb{P}[\xi_{(a,\theta)} = u] \in L^1(\Theta(a), \mu_a(.))$.

Under these additional conditions we showed in [11] that in the model of parametric random functions, the MLE criterion has to be modified to ensure the property that Γ' maximizes

$$\sum_{a \in A} \int R_{(a,\theta)} d\mu_a(\theta). \tag{5}$$

This criterion is natural, since if $\sum_{a \in A} \int d\mu_a(\theta) = 1$, the formula (5) can be interpreted as the expected probability of return of elements of A, given a prior distribution on A.

The purpose of this paper is to place explicit upper and lower bounds on the probability that MLE correctly reconstructs elements of A, in both the parametric and non-parametric settings. Our primary interest is in the situation where k independent experiments are carried out, and we wish to determine how large k needs to be in order to correctly recover the underlying element of A with high probability. To emphasise the role of k we will let $[r^{(k)}]^*_a$ (resp. $[R^{(k)}]'_{(a,\theta)}$) denote the probability that MLE correctly reconstructs a, in the non-parametric (resp. parametric) setting. We illustrate our bounds in the non-parametric setting by applications to coin-tossing and phylogeny reconstruction.

For the parametric setting, we first show, by way of an example, that the non-parametric upper bound on k does not extend in the way one might hope or expect. Nevertheless, we provide (in Theorem 4) an explicit upper bound on the number k of experiments required for MLE to reconstruct elements of A accurately. This result can be regarded as an extension of a discrete version of Wald's theorem ([14]). We describe some implications of this result for phylogeny reconstruction in the remarks following Theorem 4.

2 Distances between distributions

For $a, b \in A$, $\Xi : A \to U$, let

$$d(a,b) = \sum_{u \in U} |\mathbb{P}[\xi_a = u] - \mathbb{P}[\xi_b = u]|. \tag{6}$$

We will refer to $d(a,b)$ as the *variational distance* of the random variables ξ_a and ξ_b. We also use the *Hellinger distance* of the random variables ξ_a and ξ_b, defined by

$$d_H(a,b) = \sqrt{\sum_{u \in U} \left(\sqrt{\mathbb{P}[\xi_a = u]} - \sqrt{\mathbb{P}[\xi_b = u]|} \right)^2}. \tag{7}$$

These measures sometimes appear with slightly different definitions, terminology and normalization constants (for example, $\frac{1}{2}d(a,b)$ is sometimes referred to as the "variation distance"). It is well known [5] that $0 \le d(a,b) \le 2$ and

$$d_H^2(a,b) \le d(a,b) \le 2d_H(a,b). \tag{8}$$

We are going to use a well known, and elegant multiplicative property of the Hellinger distance. For any $\Xi : A \to U$ random function define the $\Xi^{(k)} : A \to U^k$ random function as a sequence of k independent trials of Ξ. Let $d_H^{(k)}(a,b)$ denote the Hellinger distance of the random variables $\xi_a^{(k)}$ and $\xi_b^{(k)}$. Then independence immediately implies the identity

$$1 - \frac{1}{2}\left(d_H^{(k)}(a,b)\right)^2 = \left(1 - \frac{1}{2}d_H^2(a,b)\right)^k, \tag{9}$$

by virtue of the formula

$$\sum_{u\in U}\left(\sqrt{\mathbb{P}[\xi_a = u]} - \sqrt{\mathbb{P}[\xi_b = u]}|\right)^2 = 2 - 2\sum_{u\in U}\sqrt{\mathbb{P}[\xi_a = u]}\sqrt{\mathbb{P}[\xi_b = u]}. \tag{10}$$

Combining the inequality $1-(1-x)^k \le kx$ which holds for all $0 \le x \le 1$ and k positive integer, and (9), we obtain

$$\left(d_H^{(k)}(a,b)\right)^2 = 2\left[1 - \left(1 - \frac{1}{2}d_H^2(a,b)\right)^k\right] \le kd_H^2(a,b). \tag{11}$$

Using the notation $d^{(k)}(a,b)$ for the variational distance of the k independent trials, i.e. of the random variables $\xi_a^{(k)}$ and $\xi_b^{(k)}$, inequalities (8) and (11) imply

$$d^{(k)}(a,b) \le 2\sqrt{k}d_H(a,b). \tag{12}$$

The nonsymmetric Kullback-Leibler distance (or conditional entropy) of the random variables ξ_a and ξ_b is defined as

$$d_{KL}(a,b) = \sum_{u\in U}\mathbb{P}[\xi_a = u]\log\frac{\mathbb{P}[\xi_a = u]}{\mathbb{P}[\xi_b = u]}.$$

We will use the inequality [4]

$$d_{KL}(a,b) \ge \frac{1}{2}d^2(a,b). \tag{13}$$

3 Maximum likelihood estimation for inverting random functions

In this section we describe some lower and upper bounds on the probability that MLE correctly reconstructs elements of the set A. A classical upper bound on the average value of r_a over A – or more generally the value of $\sum_{a\in A} r_a w(a)$ for some probability distribution w on A – is given by "Fano's Inequality" (see for example [4]). Here we recall from [11] a different type of lower bound, that applies also to r_a for any particular value of a, and which is closely related to the variational distance.

Theorem 1 *Assume that we have finite sets A and U and random functions $\Xi : A \to U$ and $\Gamma : U \to A$. Suppose that there is an element $b \in A$ and a subset $N \subset A$ such that for all $a \in N$*

$$d(a, b) < \delta.$$

Then we have

$$\min_{a\in N} r_a \leq \frac{1}{|N|} + \delta(1 - \frac{1}{|N|}).$$

Now we can state the following *lower* bound for r_a in the setting of Theorem 1:

Theorem 2 *Assume that we have finite sets A and U and a random function $\Xi : A \to U$. Assume that $\Gamma^* : U \to A$ is the maximum likelihood estimation, and r^*_a is the return probability of $a \in A$ using Γ^*. Then we have*

$$r^*_a \geq 1 - \sum_{b\neq a} \left(1 - \frac{1}{2} d(a, b)\right). \tag{14}$$

If the maximum likelihood estimation $\Gamma^* : U^k \to A$ is applied to invert the random function $\Xi^{(k)} : A \to U^k$, which is a sequence of k independent trials of Ξ, then

$$[r^{(k)}]^*_a \geq 1 - \sum_{b\neq a} \left(1 - \frac{1}{2} d^2_H(a, b)\right)^k. \tag{15}$$

Proof. For $y \in A$ let

$$U_y = \Big\{ u \in U |\ \forall x \in A, x \neq y, \mathbb{P}[\xi_y = u] > \mathbb{P}[\xi_x = u] \Big\}$$

and similarly V_y with $\geq$ instead of $>$ in the definition. It is clear from independence (1) and the definition (2) that

$$r^*_a \geq \sum_{u\in U_a} \mathbb{P}[\xi_a = u]. \tag{16}$$

For $x, y \in A$ set $p^x_y = \sum_{u\in V_y} \mathbb{P}[\xi_x = u]$. Now we claim

$$r^*_a \geq 1 - \sum_{y\neq a} p^a_y. \tag{17}$$

Note that

$$\sum_{y \neq a} p_y^a = \sum_{y \neq a} \sum_{u \in V_y} \mathbb{P}[\xi_a = u] \geq \mathbb{P}[\xi_a \notin U_a],$$

since the complement of U_a is a subset of $\cup_{y \neq a} V_a$, and

$$\mathbb{P}[\xi_a \notin U_a] = 1 - \mathbb{P}[\xi_a \in U_a] \geq 1 - r_a^*$$

by (16). This establishes (17). Finally, we have

$$d(a, y) = \sum_{u \in U} |\mathbb{P}[\xi_a = u] - \mathbb{P}[\xi_y = u]| =$$

$$\sum_{u \in V_y} \Big(\mathbb{P}[\xi_y = u] - \mathbb{P}[\xi_a = u] \Big) + \sum_{u \notin V_y} |\mathbb{P}[\xi_a = u] - \mathbb{P}[\xi_y = u]| \leq$$

$$p_y^y - p_y^a + \sum_{u \notin V_y} \Big(\mathbb{P}[\xi_a = u] + \mathbb{P}[\xi_y = u] \Big) = p_y^y - p_y^a + (1 - p_y^a) + (1 - p_y^y) = 2 - 2p_y^a.$$

Hence $p_y^a \geq 1 - \frac{1}{2} d(a, y)$, and plugging this into (17) yields (14). To prove (15), apply (14) to $\Xi^{(k)}$ and invoke (9).

Remarks. First, note that (15) immediately implies that if $d_a = \min_{b \neq a} d_H(a, b)$, then $[r^{(k)}]_a^* > 1 - |A| \exp(-k d_a^2 / 2)$. Consequently, if

$$k > \frac{2}{d_a^2} \log \frac{|A|}{\epsilon},$$

then $[r^{(k)}]_a^* > 1 - \epsilon$. Second, note that an analogue of (15) also holds, if instead of k independent trials of Ξ, we take independent $A \to U$ random functions $\Xi_1, \Xi_2, ..., \Xi_k$. Now the lower bound on $[r^{(k)}]_a^*$ is

$$1 - \sum_{b \neq a} \prod_{i=1}^{k} \left(1 - \frac{1}{2} d_H^2((\xi_i)_a, (\xi_i)_b) \right).$$

4 Applications

4.1 Solving biased coin tossing with MLE

We want to show an example where our upper and lower bounds for reconstructing random functions are nearly tight. Assume that $U = \{T, H\}$, i.e. we are tossing coins. Let a set A consist of $n + 1$ biased coins, denoted by $0, 1, 2, ..., n + 1$. The random function Ξ is as follows: and coin i takes H with probability i/n and takes T with probability $1 - i/n$. We show the following: there is a constant c_1 such that for $k = c_1 n^2$, for k independent trials of Ξ,

$\Xi^{(k)}$, $[r^{(k)}]_i$ cannot be uniformly close to 1, no matter which method is used for inverting $\Xi^{(k)}$. However, there is a constant c_2 such that for $k = c_2 n^2$, using MLE, we find $[r^{(k)}]_i^*$ uniformly close to 1.

For simplicity we assume that n is odd. We are going to use Theorem 1 in the following setting: $b = \frac{n-1}{2}$, $N = \{\frac{n-3}{2}, \frac{n+1}{2}\}$. Then,

$$\min_{a \in N} [r^{(k)}]_a \leq \frac{1}{2}(1+\delta),$$

where δ is the smaller variational distance for $\Xi^{(k)}$ among b and the elements of N. Observe that for Ξ, by formula (10), we have

$$d_H^2(i,j) = 2\left(1 - \frac{\sqrt{ij}}{n} - \frac{\sqrt{(n-i)(n-j)}}{n}\right). \tag{18}$$

It is easy to see that for $i = b$, $j \in N$, (18) is maximized by $j_0 = \frac{n+1}{2}$ at the value $2\big(1 - \sqrt{1 - \frac{1}{n^2}}\big) \sim 1/n^2$. By (12), $d^{(k)}(b,x) \leq 2\sqrt{k} d_H(b,x)$, for every x, and therefore $\delta \leq 2\sqrt{k} d_H(b, j_0) \sim 2\sqrt{k}/n$. Any choice of $c_1 < 1/4$ suffices to keep either $r_{\frac{n-3}{2}}$ or $r_{\frac{n+1}{2}}$ separated from 1.

In the other direction we use Theorem 2. By (15) and (18) we have

$$[r^{(k)}]_i^* \geq 1 - \sum_{\substack{j=0 \\ j \neq i}}^{n} \left(1 - \frac{1}{2} d_H^2(i,j)\right)^k = 1 - \sum_{\substack{j=0 \\ j \neq i}}^{n} \left(\frac{\sqrt{ij}}{n} + \frac{\sqrt{(n-i)(n-j)}}{n}\right)^{k/2}. \tag{19}$$

By the classical inequality

$$\sqrt{a_1 a_2} + \sqrt{b_1 b_2} \leq \sqrt{a_1 + b_1} \cdot \sqrt{a_2 + b_2},$$

the generic subtracted term in the summation (19) is estimated from above by

$$\left(1 - \frac{(j-i)^2}{n^2}\right)^{k/2}.$$

Hence,

$$[r^{(k)}]_i^* \geq 1 - 2 \sum_{m=1}^{n} \left(1 - \frac{m^2}{n^2}\right)^{k/2}. \tag{20}$$

Now observe that

$$\sum_{m=1}^{n} \left(1 - \frac{m^2}{n^2}\right)^{k/2} \leq \left(1 - \frac{1}{n^2}\right)^{k/2} + n \int_{1/n}^{1} (1 - x^2)^{k/2} dx \tag{21}$$

and

$$n \int_{1/n}^{1} (1-x^2)^{k/2} dx \leq n \int_{1/n}^{1} e^{-k \cdot x^2/2} dx \leq \frac{n}{\sqrt{k}} \int_{\sqrt{k}/n}^{\sqrt{k}} e^{-t^2/2} dt \leq \frac{n}{\sqrt{k}} \cdot \frac{1}{2\sqrt{2\pi}}.$$

Therefore, for a sufficiently large c_2, selecting $k = c_2 n^2$, both terms in the RHS of (21) will be as small as wanted, and hence in (20) $[r^{(k)}]_i^*$ will be as close to 1 as wanted.

4.2 Phylogeny reconstruction

As a second application, we consider a problem arising in phylogenetic analysis. In this setting we have a model for generating sequences at the leaves of a tree, and the question is how long such sequences need to be in order to correctly reconstruct the tree from just the generated sequences.

The simplest stochastic model, for two-state sequences, is the symmetric model, due to Neyman [9] and which we call the Neyman-2 model (related models also arise in statistical physics and in the theory of noisy communication—see for example [7]). Let $\{0, 1\}$ denote the two states. Let us be given a binary tree T (a tree in which each vertex has degree 1 or 3) with n labelled leaves. We describe how a single site in the sequence develops on T, and then we assume that the sites are independently and identically distributed (i.i.d.).

For each edge e of T we have an associated *transition probability*, which lies strictly between 0 and 0.5. Let $p : E(T) \to (0, 0.5)$ denote the associated map. Select one of the leaves, and assign it state 0 or state 1 with probability 0.5. Direct all edges away from this leaf and recursively assign random states to the vertices of T as follows: if $e = \{u, v\}$ is directed from u to v, and u (but not v) has a state assignment, then v is assigned the same state as u with probability $1 - p_e$ or the other state with probability p_e (in this latter case, we say there is a *transition on* e). It is assumed that all assignments are made independently, and so the pair (T, p) determines the joint probability of any assignment of states to the vertices of T, and thereby the marginal probability of any assignment of states to the leaves of T. If we independently generate k such assignments of states to the leaves of T, we obtain n sequences of length k. For this model, upper bounds on the sequence length k required to reconstruct the underlying tree were given in [6, 11]. These papers showed that, for accurate tree reconstruction, k needs to grow only quadratically in $1/f$ where f is the smallest transition probability in the tree, when other parameters are fixed. We now show that this rate of growth is not only sufficient, but is also necessary.

Consider binary trees on having four labelled leaves and two unlabelled interior vertices. There are three such trees (up to equivalence) and we will denote them as a, b, c. Each tree has four leaf edges (an edge incident with a leaf) and one interior edge.

Theorem 3 *For the three binary trees a, b, c on four leaves, suppose we have a Neyman-2 model, in which all the leaf edges have fixed transition probabilities, while the interior edge has transition probability f. Under any method for inverting this random function Ξ from k independent trials (i.e. sequences of length k) with success probability near 1 for all three trees, $k = \Omega\left(\frac{1}{f^2}\right)$.*

Proof.

We are going to prove that for f sufficiently close to 0, for some constant $C > 0$

$$d_H(a,b) \leq Cf. \tag{22}$$

Now (12) and (22) imply $d^{(k)}(a,b) \leq 2Cf\sqrt{k}$, and one similarly obtains $d^{(k)}(c,b) \leq 2Cf\sqrt{k}$. So if we apply Theorem 1 with $N = \{a,c\}$:

$$\min\{r_a, r_c\} \leq \frac{1}{3} + \frac{4}{3}Cf\sqrt{k}, \tag{23}$$

and the right-hand side of (23) is well separated from 1 as k is a small constant over f^2.

To complete the proof, we have to verify (22). Assume that a is the tree in which the interior edge separates leaves $1,2$ from leaves $3,4$; and b is the tree in which the interior edge separates leaves $1,3$ from leaves $2,4$. By (7)

$$d_H(a,b)^2 = \sum_{u \in U} \left(\sqrt{\mathbb{P}[\xi_a = u]} - \sqrt{\mathbb{P}[\xi_b = u]|} \right)^2 \tag{24}$$

where the summation goes for 16 terms which correspond to the 16 elements of U: functions with domain $\{1,2,3,4\}$ and co-domain $\{0,1\}$. We are going to condition on the event Φ denoting that there is transition on the interior edge of the tree, and also for the complement of this event. For $x = a,b$ define

$$\begin{aligned} A(x,u) &= \mathbb{P}[\xi_x = u \mid \neg\Phi] \\ B(x,u) &= \mathbb{P}[\xi_x = u \mid \Phi] - \mathbb{P}[\xi_x = u \mid \neg\Phi], \end{aligned}$$

where $A(x,u)$ and $B(x,u)$ are just positive constants. Observe that

$$\begin{aligned} \mathbb{P}[\xi_x = u] &= \mathbb{P}[\xi_x = u| \neg\Phi] \cdot (1-f) + \mathbb{P}[\xi_x = u| \Phi] \cdot f \\ &= \mathbb{P}[\xi_x = u| \neg\Phi] + f \cdot \left(\mathbb{P}[\xi_x = u \mid \Phi] - \mathbb{P}[\xi_x = u \mid \neg\Phi] \right) \\ &= A(x,u) + fB(x,u). \end{aligned}$$

It easily follows from the geometry of the trees a and b that $A(a,u) = A(b,u)$. Furthermore, it is easily seen that $A(a,u) \neq 0$ for all values of u, which ensures (below) that we may divide expressions by $A(a,u)$. Hence, by the Taylor expansion of the squareroot function, we have:

$$\begin{aligned} \sqrt{\mathbb{P}[\xi_a = u]} - \sqrt{\mathbb{P}[\xi_b = u]} &= \sqrt{A(a,u)} \left(\sqrt{1 + \frac{fB(a,u)}{A(a,u)}} - \sqrt{1 + \frac{fB(b,u)}{A(a,u)}} \right) \\ &= f\frac{B(a,u) - B(b,u)}{2\sqrt{A(a,u)}} + O(f^2), \end{aligned} \tag{25}$$

and summing up 16 terms like (25) we obtain

$$d_H^2(a,b) = f^2 \sum_{u \in U} \frac{(B(a,u) - B(b,u))^2}{4A(a,u)} + O(f^3),$$

and this proves (22) for all

$$C > \sqrt{\sum_{u \in U} \frac{(B(a,u) - B(b,u))^2}{4A(a,u)}}.$$

5 Maximum likelihood estimation for inverting parametric random functions

We start with an example showing that for parametric MLE there is no counterpart of (15), that is, there is no function $f = f(\delta, k)$ such that for all $\delta > 0$, $\lim_{k\to\infty} f(\delta, k) = 0$ and

$$[R^{(k)}]'_{(a,\theta)} \geq 1 - \sum_{b \neq a} f\big(\delta((a,\theta), b), k\big), \tag{26}$$

where

$$\delta((a,\theta), b) = \inf_{\theta' \in \Theta(b)} d_H\big((a,\theta),(b,\theta')\big).$$

Take $A = \{a_1, a_2\}$, $U = \{u_1, u_2, ..., u_{2k^2}\}$, $\Theta(a_1) = \Theta(a_2) = U^k$. (This is the trick in the construction: we select codomains and parameter spaces whose size *depends* on the sequence length k). We denote a generic element of U^k by $\mathbf{u}$, and $\mathrm{supp}(\mathbf{u})$ denotes the set of elements of U which occur as coordinates in $\mathbf{u}$. Let $B = (\{a_1\} \times \Theta(a_1)) \cup (\{a_2\} \times \Theta(a_2))$. Define the parametric random function $\Xi : B \to U$ as follows. Set $\mathbb{P}[\xi_{(a_1,\mathbf{u})} = v] = 1/|U|$ for each $v \in U$. For $\mathbf{u} \in U^k$ and $v \in U$, set $\mathbb{P}[\xi_{(a_2,\mathbf{u})} = v] = i/k$, if v occurs at each $i = i(v)$ coordinates in $\mathbf{u}$. Now for any $\mathbf{w}, \mathbf{u} \in U^k$ we have

$$d\Big((a_1, \mathbf{w}), (a_2, \mathbf{u})\Big) \geq 2 - \frac{1}{k} \tag{27}$$

by the calculation

$$\sum_{v \in supp(\mathbf{u})} \left(\frac{i(v)}{k} - \frac{1}{|U|}\right) + \sum_{v \notin supp(\mathbf{u})} \frac{1}{|U|} = 2 - 2 \sum_{v \in supp(\mathbf{u})} \frac{1}{|U|} \geq 2 - \frac{2k}{|U|} = 2 - \frac{1}{k}.$$

Consider now k independent trials of Ξ, $\Xi^{(k)}$. We study inverting $\Xi^{(k)}$ with parametric MLE. Note that for any $\mathbf{u} \in U^k$,

$$\mathbb{P}[\xi^{(k)}_{(a_2,\mathbf{u})} = \mathbf{u}] = \prod_{i=1}^{k} \mathbb{P}[\xi_{(a_2,\mathbf{u})} = u_i] \geq \big(\frac{1}{k}\big)^k;$$

and for any $\mathbf{w} \in U^k$,

$$\mathbb{P}[\xi^{(k)}_{(a_1,\mathbf{w})} = \mathbf{u}] = \prod_{i=1}^{k} \mathbb{P}[\xi_{(a_1,\mathbf{w})} = u_i] = \big(\frac{1}{2k^2}\big)^k < \big(\frac{1}{k}\big)^k.$$

Therefore, one *always* has $[R^{(k)}]'_{(a_1,\mathbf{w})} = 0$ (see (4)), while by (27) and (8) the d_H distances between the the random variables corresponding to a and b are well separated from zero. This establishes our claim at the start of this section, regarding the non-existence of an analogue of (15) from Theorem 2.

However, with mild extra conditions we can state a positive result. This positive result provides explicit bounds on the convergence of the MLE in the parametric setting.

Theorem 4 *Assume $B = \{(a,\theta) : a \in A, \theta \in \Theta(a)\}$, and $\Xi : B \to U$ is a parametric random function, where A and U are finite sets. Assume that for a particular $(a,\theta) \in B$, there exists a $d_0 > 0$ such that for all $b \in A$, $b \neq a$, and $\theta' \in \Theta(b)$*

$$d\Big((a,\theta),(b,\theta')\Big) \geq d_0, \tag{28}$$

where d, as usual, denotes the variational distance. If the maximum likelihood estimation is applied to invert the parametric random function $\Xi^{(k)} : A \to U^k$, which is a sequence of k independent trials of Ξ, then

$$\lim_{k\to\infty} [R^{(k)}]'_{(a,\theta)} = 1. \tag{29}$$

For a more precise result, set $U^+ = \{u \in U : \mathbb{P}[\xi_{(a,\theta)} = u] > 0\}$, and $m = \min_{u\in U^+} \mathbb{P}[\xi_{(a,\theta)} = u]$. If

$$k > f(m,d_0)\log\left(\frac{2|U^+|}{\epsilon}\right) \tag{30}$$

then MLE estimation returns a with probability at least $1-\epsilon$, where

$$f(m,d_0) = \max\left\{\frac{16}{m}, \frac{17\log^2 m(1+\frac{2}{m})^2}{d_0^4}\right\}.$$

Proof. For $u \in U$, define $p(u) = \mathbb{P}[\xi_{(a,\theta)} = u]$, and then $m = \min_{u\in U^+}\{p(u)\} > 0$. Define $\hat{p}(u)$ as the corresponding relative frequency, i.e.

$$\hat{p}(u) = \frac{1}{k}\#\{j : (\xi_j)_{(a,\theta)} = u\}, \tag{31}$$

where ξ_j is the j^{th} trial of the random function. Let $\delta = \frac{4}{\sqrt{17}}$, and let

$$\eta = \min\{\frac{1}{2}, \frac{\delta d_0^2}{2|\log m|(1+\frac{2}{m})}\}.$$

Then,

$$\eta|\log m| + \frac{\eta|\log m|}{m(1-\eta)} \leq \frac{\delta}{2}d_0^2. \tag{32}$$

By the large deviation inequality [1] Appendix A formula (14), we have

$$\mathbb{P}\left[|p(u)-\hat{p}(u)|>\eta p(u)\right]< 2e^{-c_\eta k p(u)} \tag{33}$$

where $c_\eta = \min\{-\log\left[e^\eta(1+\eta)^{-(1+\eta)}\right], \frac{\eta^2}{2}\}$. Note that for $0<\eta<1/2$ we have $-\log[e^\eta(1+\eta)^{-(1+\eta)}] \geq \frac{\eta^2(1-\eta)}{2}$ by Taylor expansion, and hence $c_\eta \geq \eta^2/4$. Therefore, formula (33) holds if we change c_η to $\eta^2/4$ in the exponent. Now suppose k satisfies inequality (30). Then,

$$k > \frac{4}{m\eta^2}\log\left(\frac{2|U^+|}{\epsilon}\right)$$

by the definition of f and η. Consequently, $2|U^+|e^{-\eta^2 km/4} < \epsilon$ and so, with probability at least $1-\epsilon$ we have,

$$\forall u \in U \quad |p(u)-\hat{p}(u)| \leq \eta p(u). \tag{34}$$

(We also used the Bonferroni inequality, and the fact that with probability 1, $p(u)=\hat{p}(u)=0$ for all $u \in U\setminus U^+$). For $x\in A, \omega\in\Theta(x)$, consider

$$L(x,\omega) = \sum_{u\in U}\hat{p}(u)\log\mathbb{P}[\xi_{x,\omega}=u]. \tag{35}$$

(Here, as always in this kind of calculations, we use the convention $0\times(-\infty)=0$, which is supported by $\lim_{x\to 0^+} x\log x = 0$.) $L(x,\omega)$ is $\frac{1}{k}$ times the natural logarithm of the probability that the observed sequence of U-elements came from (x,ω). Therefore $L(x,\omega)\leq 0$ is proportional to the log-likelihood of (x,ω).

Consider now a fixed $b\in A$, $b\neq a$ and a fixed $\theta'\in\Theta(b)$. For $u\in U$, we use the notation $q(u)=\mathbb{P}[\xi_{(b,\theta')}=u]$.

We finish the proof conditional to the following event:

$$(34) \text{ holds and } u\notin U^+ \text{ implies } \hat{p}(u)=0. \tag{36}$$

Note that the second part of the condition holds with probability 1, and so event (36) occurs with probability at least $1-\epsilon$.

We distinguish two cases. In both cases we show $L(a,\theta)-L(b,\theta')>0$. Since $L(a,\theta)-L(b,\theta')$ is proportional to the log-likelihood ratio of getting the observed sequence from (a,θ) resp. (b,θ'), this means the correct reconstruction of a from the observed data by MLE by (4). Since this holds (with probability 1) for all θ', conditional on event (36), and event (36) occurs with probability at least $1-\epsilon$, the probability that MLE correctly reconstructs a will be least $1-\epsilon$, as required.

Case 1: there exists a $v\in U^+$ with $q(v)<\exp(\frac{\log m}{m(1-\eta)})$. In this case $L(b,\theta')\leq\hat{p}(v)\log q(v)<\log m$, so $L(b,\theta')<\log m$. On the other hand,

$L(a,\theta) = \sum_{u\in U} \hat{p}(u)\log p(u) \geq \sum_{u\in U} \hat{p}(u)\log m = \log m$. Therefore, $L(a,\theta) > L(b,\theta')$.

Case 2: for all $u \in U^+$, $q(u) \geq \exp(\frac{\log m}{m(1-\eta)})$. We have for all $u \in U^+$, $|\log q(u)| \leq \frac{|\log m|}{m(1-\eta)}$. Consider

$$L(a,\theta) - L(b,\theta') = \sum_{u\in U} \hat{p}(u)\log\frac{p(u)}{q(u)} = \sum_{u\in U^+} \hat{p}(u)\log\frac{p(u)}{q(u)} =$$

$$\sum_{u\in U} p(u)\log\frac{p(u)}{q(u)} + \sum_{u\in U^+} \big(\hat{p}(u) - p(u)\big)\log\frac{p(u)}{q(u)}. \tag{37}$$

Notice that the first sum in (37) is exactly the Kullback-Leibler distance $d_{KL}((a,\theta),(b,\theta'))$. By formulae (13, 28) this first sum is at least $\frac{1}{2}d_0^2$. Since we conditionize on (34), $|\hat{p}(u) - p(u)| \leq \eta p(u)$. Hence we can estimate the absolute value of the second sum in (37) by

$$\sum_{u\in U} \eta p(u)\big(|\log p(u)| + |\log q(u)|\big) \leq \sum_{u\in U} \eta p(u)(|\log m| + \frac{|\log m|}{m(1-\eta)}) =$$

$$\eta|\log m| + \frac{\eta|\log m|}{m(1-\eta)} \leq \frac{\delta}{2}d_0^2 \tag{38}$$

by (32) and so, $L(a,\theta) - L(b,\theta') > 0$.

Remarks.

1. Notice that, because $|U^+|m \leq 1$, inequality (30) will hold whenever $k \geq f(m,d_0)|\log(\frac{2}{m\epsilon})|$. Notice that this bound on k (that suffices for parametric MLE to reconstruct a with probability at least $1-\epsilon$) depends only on ϵ, d_0 and m, and it is independent of the cardinality of A and U (*cf.* the bound we described for non-parametric MLE in the Remark following Theorem 2).

2. Note also that the example described at the beginning of Section 5 shows that one cannot strengthen Theorem 4 by simply dropping the role of m. That is, Theorem 4 fails if we replace replace (30) with the weaker condition that that
$$k \geq f_1(d_0)\log\left(\frac{2|U|^+}{\epsilon}\right)$$
for some suitable function f_1 (that does not depend on m), since in the example described, any such inequality will be satisfied for sufficiently large k ($|U|$ grows only quadratically with k) yet MLE fails to recover a_1. A closer examination of this example shows that m converges to zero sufficiently fast with k for the bound in (30) to be violated.

3. Suppose that, for each $b \in A$ we have (i) the set $\Theta(b)$ is a compact topological space, and (ii) the mapping from $\Theta(b)$ to the interval $[0,1]$ defined by $(b, \theta) \mapsto \mathbb{P}(\xi_{(b,\theta)=a} = u)$ is continuous for each element $u \in U$. Then the separation property (28) required in Theorem 4 becomes equivalent to the (in general weaker) condition that for all $b \in A, b \neq a$, and $\theta' \in \Theta(b)$

$$d((a, \theta), (b, \theta')) > 0. \tag{39}$$

For example, for most models in the phylogenetic setting, assumptions (i) and (ii) will apply, and so MLE will be statistically consistent (that is, satisfy (29)) provided the model satisfies (39). In particular, the detailed analysis, and additional assumption required by Chang [3] in order to establish (for a general Markov model on trees) a strengthening of (39) to the case $b = a, \theta \neq \theta'$ is unncessary if one wishes simply to establish the statistical consistency of MLE in the estimation of a binary tree (and not the associated transition matrices of the model). There are also other models in use that satisfy (39), and thereby justify the statistical consistency of MLE. For example, consider a model in which sites evolve i.i.d. on a binary tree according to a stationary, reversible Markov process (with an unknown rate matrix), and with a rate factor (constant across the tree) drawn from a distribution $\mathcal{D}$. Such models satisfy (39) if $\mathcal{D}$ is known and therefore the same for each possible tree ([13], Section 3.3), however (39) may fail if $\mathcal{D}$ is unknown [12]. We note that Theorem 4 also provides the first explicit upper bounds on the sequence length required for MLE to accurately reconstruct a binary tree in the phylogenetic setting.

5.1 Correction

In our earlier paper ([11]) we made some comments concerning the analysis of MLE under the Neyman 2-tree model, for a 4-species tree with transition probability $p(e) < \epsilon$ on two adjacent pendant edges, and $p(e) > 0.5 - \epsilon$ on the remaining three edges. We stated that the expected reconstruction probability for MLE is (approximately) 1/3 for ϵ and δ sufficiently small (and k fixed). The words "(approximately) 1/3" should be replaced by "at most 2/3" (a proof of this assertion is given in [10]). This does not, however, affect the conclusions described in our earlier paper.

References

[1] N. Alon and J. H. Spencer, *The Probabilistic Method*, John Wiley and Sons, New York, 1992.

[2] G. Casella and R. L. Berger, *Statistical Inference*, The Wadsworth & Brooks/Cole Statistics/Probability Series, Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA, 1990. -chapter 10

[3] J. T. Chang, Full reconstruction of Markov models on evolutionary trees: identifiability and consistency, *Math. Biosci.* **137** (1996), 51–73.

[4] T. M. Cover and J. A. Thomas, *Elements of information theory*, John Wiley and Sons, New York, 1991.

[5] P. Diaconis, *Group representations in probability and statistics,* Institute of Mathematical Statistics Lecture Series vol. 11, 1988.

[6] P. L. Erdős, M. A. Steel, L. A. Székely, and T. Warnow, A few logs suffice to build (almost) all trees I, *Random Structures and Algorithms*, **14** (1999), 153–184.

[7] W. Evans, C. Kenyon, Y. Peres, L. J. Schulman, Broadcasting on trees and the Ising model, *Ann. Appl. Probab.* **10** (2) (2000), 410–433.

[8] B. S. Everitt, *The Cambridge Dictionary of Statistics*, Cambridge Univ. Press, Cambridge, UK, 1998.

[9] J. Neyman, Molecular studies of evolution: a source of novel statistical problems. In: Gupta, S.S., and J. Yackel J., eds., *Statistical Decision Theory and Related Topics*, 1–27, New York, Academic Press, 1971.

[10] M. Steel, and Penny, D., Parsimony, likelihood, and the role of models in molecular phylogenetics, *Molecular Biology and Evolution* **17** (6) (2000) 839–850.

[11] M. A. Steel and L. A. Székely, Inverting random functions, *Annals of Combinatorics* **3** (1999), 103–113.

[12] M. A. Steel, L. A. Székely and M. D. Hendy, Reconstructing trees when sequence sites evolve at variable rates, *J. Comput. Biol.* **1(2)** (1994), 153–163.

[13] C. Tuffley and M. Steel, Modelling the covarion hypothesis of nucleotide substitution, *Math. Biosci.* **147**, (1998), 63-91.

[14] A. Wald, Note on the consistency of the maximum likelihood estimate, *Ann. Math. Statistics* **20** (1949) 595–600.