Inverting random functions [*]

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: laszlo@math.sc.edu

Abstract

In this paper we study how to invert random functions under different criteria. The motivation for this study is phylogeny reconstruction, since the evolution of biomolecular sequences may be considered as a random function from the set of possible phylogenetic trees to the set of collections of biomolecular sequences of observed species. Our results may effect how we think about the maximum likelihood estimation (MLE) in phylogeny. MLE is optimal to invert random functions under a first criterion, although it is not optimal under another, at least equally natural but more conservative second criterion. It turns out that MLE has to be used in a different way as it is used in the phylogeny literature, if we have a prior distribution on trees and mutation mechanisms and want to keep MLE optimal under the same first criterion. Some of the results of this paper have been known in the setting of statistical decision theory, but have never been discussed in the context of phylogeny.

[*] Michael A. Steel was supported by the New Zealand Marsden Fund. László A. Székely was supported by the National Science Foundation grant DMS 9701211, and the Hungarian National Science Fund contract T 016 358. 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.

1 Introduction

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* $\Xi : \ A \rightarrow U$. Ordinary functions are specific instances of random functions. Given another random function, Γ, from U to V, we can speak about the composition of Γ and Ξ, $\Gamma \circ \Xi : \ A \rightarrow V$, 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 $\Gamma : \ U \rightarrow A$ in order to obtain the best approximations of the identity function $\imath : \ A \rightarrow A$ by $\Gamma \circ \Xi$. *We always assume that* Ξ *and* Γ *are independent.* The reader may think that this is not the only or not the right definition of random functions. For example, a natural alternative is to consider all the $|U|^{|A|}$ ordinary functions from A to U, and pick one of them according to a certain probability distribution (second definition of random function). It is clear that this second definition of a random function yields the unique distribution of the vector variable $(\xi_a : \ a \in A)$. On the other hand, consider a Ξ according to the first definition. Consider a fixed ordering of the elements of A, $A = \{a_1, a_2, ..., a_{|A|}\}$. For any sequence $(u_1, u_2, ..., u_{|A|})$ $(u_i \in U)$ find the probability

$$p(u_1, u_2, ..., u_{|A|}) = \mathbb{P}[\xi_{a_i} = u_i \ \text{ for all } \ i = 1, 2, ..., |A|].$$

Now we see that Ξ can be seen as a random function according to the second definition, such that the ordinary function $a_i \mapsto u_i$ for all $i = 1, 2, ..., |A|$ is selected with probability $p(u_1, u_2, ..., u_{|A|})$.

We might have used as well a "weaker" definition for random functions, namely, just a collection of distributions the U-valued random variables for all $a \in A$, i.e. a collection of probability distributions but no joint distribution. In case of this weaker definition, a random function may have different representations as picking ordinary functions according to a probability distribution on ordinary functions. Requiring only that every ξ_a is independent of every γ_u, the results of this paper still would go through with this definition.

Our motivation for the study of random functions came from phylogeny reconstruction. Stochastic models define how biomolecular sequences develop along the edges of phylogenetic trees. 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 the coin tosses in the phylogeny reconstruction algorithm. Criteria for the phylogeny reconstruction may differ in what do we want to optimize.

Consider the probability of returning a from a by the composition of two random functions, i.e. $r_a = \mathbb{P}[\gamma_{\xi_a} = a]$. A natural criterion is to find Γ for a given Ξ in order to maximize $\sum_a r_a$. We may have situations where we are given an $w : A \rightarrow \mathbb{R}$ weight function and we want 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 $\Gamma^* : U \rightarrow A$ can be defined in the following way: for any fixed $u \in U$, $\gamma^*_u = a^*$ for sure if for all $a \in A$, $\mathbb{P}[\xi_{a^*} = u]w(a^*) \geq \mathbb{P}[\xi_a = u]w(a)$. This function Γ^* is called the *maximum likelihood estimation* (MLE) in the literature [12, 7]. (In case of ties, randomization is possible and usual.) We show that the maximum likelihood estimation Γ^* maximizes $\sum_a r_a w(a)$ for a given Ξ. However, it is at least as natural to look at a more conservative criterion: maximize the smallest value of r_a for $a \in A$. Call this criterion the *minimax criterion.* For the minimax criterion MLE is not always optimal. These results have been known in the context of statistical decision theory but have never been discussed in the context of phylogeny.

This paper introduces 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. Assume that for a finite set A, for every $a \in A$, a measure space $(\Theta(a), \mu_a(.))$ is assigned, so that $\mu_a(\Theta(a)) < \infty$, 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 p 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 $\theta \in \Theta(a)$, a U-valued random variable $\xi_{(a,\theta)}$ is in Ξ;

(ii) for all $u \in U$, the set $\{\theta \in \Theta(a) : \ \xi_{(a,\theta)} = u\}$ is measurable in the measure space $(\Theta(a), \mu_a(.))$.

We are interested in random functions $\Gamma : U \rightarrow A$ independent from Ξ so that $\gamma_{\xi_{(a,\theta)}}$ best approximates p under certain criteria. Call $R_{a,\theta}$ the probability $\mathbb{P}[\gamma_{\xi_{(a,\theta)}}] = a$. MLE, as it is used in the practice, would take the Γ^*, for which for every fixed u $\gamma^*_u = a^*$ for sure if there exists an $(a^*, \theta^*) \in B$, such that for all $a \in A$, and all $(a, \theta) \in B$, $\mathbb{P}[\xi_{(a^*,\theta^*)} = u] \geq \mathbb{P}[\xi_{(a,\theta)} = u]$. (In case of ties, randomization is possible and usual). We show that in the model of parametric random functions, the MLE criterion has to be modified to keep the property that Γ^* maximizes

$$\sum_{a \in A} \int R_{a,\theta} d\mu_a(\theta). \tag{1}$$

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

A general reference to phylogeny reconstruction is [20]. For the popular maximum likelihood estimation in phylogeny see [12], [7], esp. pp. 205–206. In recent works on phylogeny the minimax criterion is gaining popularity, implicitly or explicitly this criterion is followed in [10, 8, 9, 6, 2].

The only other work that studies necessary conditions for phylogeny reconstruction that we know about is [10]. Their setting is more specific than inverting random functions. We do not know if their results hold only for deterministic phylogeny reconstruction algorithms or extend to randomized phylogeny reconstruction algorithms.

2 General bounds

For completeness, we cite our first result on inverting random functions with proof from [8]. This theorem generalizes the fact that inverting any ordinary $A\rightarrow U$ function requires $|A|\le|U|$

Theorem 1 *Assume that we have finite sets A and U and random functions $\Xi : A\rightarrow U$ and $\Gamma : U\rightarrow A$.*
(i) If $r_a=\mathbb{P}[\gamma_{\xi_a}=a]>\epsilon$ for all $a\in A$ then $|U|>\epsilon|A|$, even without assuming the independence of Ξ and Γ.
(ii) If Ξ and Γ are independent, and $r_a=\mathbb{P}[\gamma_{\xi_a}=a]>1/2$ for all $a\in A$, then $|U|\ge|A|$.

Proof. Proof of (i): By hypothesis $\epsilon|A|<\sum_a\mathbb{P}[\gamma_{\xi_a}=a]=\sum_a\sum_u\mathbb{P}[\xi_a=u\ \&\ \gamma_u=a]\le\sum_u(\sum_a\mathbb{P}[\gamma_u=a])=\sum_u 1=|U|$.

Proof of (ii): First note that $\mathbb{P}[\gamma_{\xi_a}=b]=\sum_u\mathbb{P}[\gamma_u=b]\mathbb{P}[\xi_a=u]$ by independence. Arrange the numbers $\mathbb{P}[\gamma_u=b]$ into an $|A|\times|U|$ matrix M and the numbers $\mathbb{P}[\xi_a=u]$ into an $|U|\times|A|$ matrix N. Now we have ${}_b[MN]_a=\mathbb{P}[\gamma_{\xi_a}=b]$ and the column sums in MN equal to 1.

Let $C=(c_{ij})$ denote a complex square matrix. A theorem of Lévy and Desplanques asserts that if for all i, $|c_{ii}|>\sum_{i:\ i\neq j}|c_{ij}|$, then $\det(C)\neq 0$ ([15], p. 146). By our assumptions $C=MN$ satisfies the conditions of the Lévy-Desplanques theorem: $c_{ii}>1/2$ and $\sum_{i:\ i\neq j}c_{ij}=1-c_{ii}<1/2$, and hence $|A|=rank(MN)\le rank(M)\le|U|$.

There is another proof which avoids linear algebra. Observe that for each a, there exists an $u = u_a$ for which $\mathbb{P}[\gamma_{u_a} = a] > 1/2$. Assume that for contrary, exists an $a \in A$ such that for all $u \in U$ $\mathbb{P}[\gamma_u = a] \leq 1/2$. Then we have

$$\begin{aligned} r_a &= \sum_u \mathbb{P}[\xi_a = u \ \& \ \gamma_u = a] \quad \text{(here we use independence)} \\ &= \sum_u \mathbb{P}[\xi_a = u]\mathbb{P}[\gamma_u = a] \leq \frac{1}{2}\sum_u \mathbb{P}[\xi_a = u] = \frac{1}{2} \end{aligned}$$

contradicting our assumption $\mathbb{P}[\gamma_{\xi_a} = a] > 1/2$. Now, the map sending a to u_a is one-to-one from A into U (and so $|A| \leq |U|$ as required) since otherwise, if two elements get mapped to u, then $1 = \sum_a \mathbb{P}[\gamma_u = a] > 1/2 + 1/2$. □

Note that the message of Theorem 1 (ii) is that relaxing the requirement for reconstructing functions *for sure* to reconstructing functions *with probabilities exceeding half* does not allow *any* relaxation on the size of U. Theorem 1 (ii) is no longer valid if we give up the independence of Ξ and Γ, this observation was made by Peter Winkler. Indeed, let us be given a random variable υ which has uniform distribution in $\{1, 2, ..., n\}$. We give up independence by using the *same* experiment for υ to define the random function and to design its inverse. Take $A = \{1, 2, ..., n+1\}$, $U = \{1, 2, ..., n\}$, and define $\xi_i = i$ for sure for $i \leq n$, and $\xi_{n+1} = i$ if $\upsilon = i$. Define the random function Γ in the following way. If $\upsilon = i$, then make a coin toss τ independently from υ, so that $\mathbb{P}[\tau = \text{HEAD}] = 1/3$ and $\mathbb{P}[\tau = \text{TAIL}] = 2/3$. Set

$$\gamma_j = \begin{cases} n+1, & \text{if } \upsilon = j, \ \tau = \text{TAIL}, \\ j, & \text{if } \upsilon = j, \ \tau = \text{HEAD}, \\ j, & \text{if } \upsilon \neq j. \end{cases}$$

Using conditional probabilities, it is easy to see that

$$\mathbb{P}[\gamma_{\xi_{n+1}} = n+1] = \sum_{l=1}^{n} \frac{2}{3}\mathbb{P}[\upsilon = l] = 2/3;$$

and that for $i \leq n$,

$$\begin{aligned} \mathbb{P}[\gamma_{\xi_i} = i] &= \mathbb{P}[\gamma_{\xi_i} = i|\upsilon = i]\mathbb{P}[\upsilon = i] + \mathbb{P}[\gamma_{\xi_i} = i|\upsilon \neq i]\mathbb{P}[\upsilon \neq i] \\ &= \frac{1}{3} \cdot \frac{1}{n} + 1 \cdot (1 - \frac{1}{n}). \end{aligned}$$

Any $n \geq 2$ yields the example required.

The condition $|A| \leq |U|$ is not sufficient to invert *every* ordinary $A \to U$ function. In the following two theorems we show a finer analysis for random functions, measuring how close they are to injections.

Theorem 2 *Assume that we have finite sets A and U and random functions $\Xi : A \to U$ and $\Gamma : U \to A$. If for all $u \in U$, $\max_a \mathbb{P}[\xi_a = u] \leq \lambda(u)$, then for $r_a = \mathbb{P}[\gamma_{\xi_a} = a]$ we have*

$$\min_a r_a \leq \frac{1}{|A|} \sum_{a \in A} r_a \leq \frac{1}{|A|} \sum_{u \in U} \lambda(u).$$

Proof. We have

$$\begin{aligned} r_a &= \sum_u \mathbb{P}[\xi_a = u]\mathbb{P}[\gamma_u = a] \leq \sum_u \lambda(u)\mathbb{P}[\gamma_u = a], \\ r_a &\leq \frac{1}{|A|} \sum_{a \in A} r_a \leq \frac{1}{|A|} \sum_{a \in A} \sum_{u \in U} \lambda(u)\mathbb{P}[\gamma_u = a] \\ &= \frac{1}{|A|} \sum_{u \in U} \lambda(u) \sum_{a \in A} \mathbb{P}[\gamma_u = a] = \frac{1}{|A|} \sum_{u \in U} \lambda(u). \Box \end{aligned}$$

Theorem 3 *Assume that we have finite sets A and U and random functions $\Xi : A \to U$ and $\Gamma : U \to A$. Let $d(a, b)$ for $a, b \in A$ denote the variational distance $\sum_{u \in U} |\mathbb{P}[\xi_a = u] - \mathbb{P}[\xi_b = u]|$. 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

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

Proof. Set $r_{ac} = \mathbb{P}[\gamma_{\xi_a} = c]$. We have

$$\sum_{a \in N} r_a = \sum_{a \in N} r_{aa} = r_{bb} + \sum_{\substack{a \in N \\ a \neq b}} r_{aa}.$$

Now we have

$$r_{bb} = 1 - \sum_{\substack{a \in A \\ a \neq b}} r_{ba} \leq 1 - \sum_{\substack{a \in N \\ a \neq b}} r_{ba}.$$

We have

$$\begin{aligned} \sum_{a \in N} r_a &\leq r_b + \sum_{\substack{a \in N \\ a \neq b}} r_a \leq 1 - \sum_{\substack{a \in N \\ a \neq b}} r_{ba} + \sum_{\substack{a \in N \\ a \neq b}} r_{aa} \\ &= 1 + \sum_{\substack{a \in N \\ a \neq b}} (r_{aa} - r_{ba}) \leq 1 + \sum_{\substack{a \in N \\ a \neq b}} |r_{aa} - r_{ba}|. \end{aligned}$$

Now take

$$r_{aa} = \sum_u \mathbb{P}[\xi_a = u \ \& \ \gamma_u = a] =: \sum_u x_u,$$

$$r_{ba} = \sum_u \mathbb{P}[\xi_b = u \ \& \ \gamma_u = a] =: \sum_u y_u.$$

Using the independence, we obtain $x_u = \mathbb{P}[\xi_a = u]\mathbb{P}[\gamma_u = a]$ and $y_u = \mathbb{P}[\xi_b = u]\mathbb{P}[\gamma_u = a]$. We have $|x_u - y_u| \leq \mathbb{P}[\gamma_u = a]|\mathbb{P}[\xi_a = u] - \mathbb{P}[\xi_b = u]| \leq |\mathbb{P}[\xi_a = u] - \mathbb{P}[\xi_b = u]|$, and hence $|r_{aa} - r_{ba}| \leq \sum_u |x_u - y_u| \leq d(a,b)$. Thus $\sum_{a\in N} r_a \leq 1 + \sum_{\substack{a\in N\\ a\neq b}} d(a,b)$ and we obtain the claimed result.□

3 Optimization criteria for inverting random functions

Theorem 4 *Assume that we have finite sets A and U, a function $w : A \to \mathbb{R}$, and a random function $\Xi : A \to U$. A random function $\Gamma^* : U \to A$ maximizing $\sum_a r_a w(a)$ can be defined in the following way: for any fixed $u \in U$, $\gamma^*_u = a^*$ for sure if for all $a \in A$, $\mathbb{P}[\xi_{a^*} = u]w(a^*) \geq \mathbb{P}[\xi_a = u]w(a)$.*

Proof. This theorem occurs in the setting of statistical decision theory in the book of Berger [3] p. 159. For completeness we give a proof. We will prove an even more general result in Section 4. We have to solve the following linear program in order to find Γ^*:

$$\begin{array}{rcl} \text{for all} \quad a \in A,\ u \in U,\ \ \mathbb{P}[\gamma_u = a] & \geq & 0; \\ \text{for all} \quad u \in U,\ \ \displaystyle\sum_a \mathbb{P}[\gamma_u = a] & = & 1; \\ \max \quad \displaystyle\sum_{u\in U}\sum_{a\in A} \mathbb{P}[\gamma_u = a]\mathbb{P}[\xi_a = u]w(a). & & \end{array}$$

The Duality Theorem of linear programming [17] applies:

$$\max\{c^T x : \ Mx \leq b\} = \min\{y^T b : \ y \geq 0, y^T M = c\},$$

if both optimizations are taken over nonempty sets. The Duality Theorem applies with the following setting: collect the values $\mathbb{P}[\gamma_u = a]$ in a column vector y of length $|U||A|$, the values $-\mathbb{P}[\xi_a = u]w(a)$ into a row vector b of length $|U||A|$, $c^T = (1, 1, ..., 1)$ of length $|U|$, and set the $|U||A| \times |U|$ matrix M by

$${}_{(a,u)}M_{u'} = \begin{cases} 1, & \text{if } u = u', \\ 0, & \text{otherwise.} \end{cases}$$

We see that the dual problem comes in variables x_u $(u \in U)$,

$$\begin{aligned} \text{for all} \quad a \in A, \quad x_u &\leq -\mathbb{P}[\xi_a = u]w(a), \\ \max \sum_{u \in U} x_u. & \end{aligned}$$

Clearly

$$x_u \leq -\max_{a \in A} \mathbb{P}[\xi_a = u]w(a),$$

and hence for every feasible solution x_u we have

$$\sum_{u \in U} x_u \leq -\sum_{u \in U} \max_{a \in A} \mathbb{P}[\xi_a = u]w(a).$$

This bound for the dual objective function is attained with setting the values of Γ^* as it is in the statement of the Theorem. □

Theorem 5 *Assume that we have finite sets A and U and a random function $\Xi : A \to U$. The random function $\Gamma^\dagger : U \to A$ maximizing $\min_{a \in A} r_a$ has the following good characterization:*

$$\min_{a \in A} r_a = \max_{\mu} \sum_{u \in U} \max_{a \in A} \mu(a) \mathbb{P}[\xi_a = u],$$

where μ is a probability distribution on A, i.e. $\sum_{a \in A} \mu(a) = 1$ and $\mu(a) \geq 0$ for every $a \in A$. An optimal $\Gamma^\dagger$ can be computed by linear programming.

Proof. Finding $\Gamma^\dagger$ can be written as the following linear program:

$$\begin{aligned} \text{for all} \quad a \in A,\ u \in U, \quad \mathbb{P}[\gamma_u = a] &\geq 0; \\ \text{for all} \quad u \in U, \quad \sum_a \mathbb{P}[\gamma_u = a] &= 1; \\ \text{for all} \quad a \in A, \quad h_a &\geq 0; \\ s &\geq 0; \\ -h_a - s + \sum_{u \in U} \mathbb{P}[\xi_a = u]\mathbb{P}[\gamma_u = a] &= 0; \\ \min \quad -s. & \end{aligned}$$

The Duality Theorem of linear programming [17] states that

$$\max\{c^T x : \ Mx \leq b\} = \min\{y^T b : \ y \geq 0, y^T M = c\},$$

if both optimizations are taken over nonempty sets. From here the required characterization immediately follows. Note that this characterization is present in a different terminology in [3] Chapter 5. □

It seems to be an interesting new observation, that for $|A| = 2$, $\Gamma^\dagger$ can be computed by the greedy algorithm. Let $A = \{a, b\}$, $p_i = \mathbb{P}[\xi_a = u_i]$, and $q_i = \mathbb{P}[\xi_b = u_i]$. Denote $x_i = \mathbb{P}[\gamma_{u_i} = a]$ and $1 - x_i = \mathbb{P}[\gamma_{u_i} = b]$. We have to solve:

$$\begin{aligned} 0 \leq x_i &\leq 1 \\ \sum_i p_i x_i &= \sum_i (1 - x_i) q_i \\ \max \sum_i p_i x_i, & \end{aligned}$$

since some optimal solutions have to satisfy $r_a = r_b$. The middle line of the linear program can be rewritten as

$$\sum_i (p_i + q_i) x_i = \sum_i q_i.$$

From here, the following greedy algorithm gives the solution: sort the numbers $\frac{p_i}{p_i + q_i}$ into decreasing order. Assume now without loss of generality that this order is $j_1, j_2, ..., j_{|U|}$. Let i^* denote the smallest index for which $\sum_{l=1}^{i^*} p_{j_l} + q_{j_l} > \sum_{i=1}^{|U|} q_i$, and let

$$r^* = \frac{1}{p_{i^*} + q_{i^*}} \left(\sum_{i=1}^{|U|} q_i - \sum_{l=1}^{i^*-1} (p_{j_l} + q_{j_l}) \right).$$

Set

$$x_{j_i} = \begin{cases} 1, & \text{if } i < i^*, \\ r^*, & \text{if } i = i^*, \\ 0, & \text{if } i > i^*. \end{cases} \tag{2}$$

Furthermore, if $|A| > 2$, and for every every $a, b \in A$ *we knew* all the numbers $c_i(a, b)$,

$$c_i(a, b) = \mathbb{P}[\gamma_{u_i} = a] + \mathbb{P}[\gamma_{u_i} = b]$$

for a $\Gamma^\dagger$ solution, then the argument above makes it possible to find $\mathbb{P}[\gamma_{u_i} = a]$ and $\mathbb{P}[\gamma_{u_i} = b]$. Just modify the linear progam to

$$\begin{aligned} 0 \leq x_i &\leq c_i & (3) \\ \sum_i p_i x_i &= \sum_i (c_i - x_i) q_i & (4) \\ \max \sum_i p_i x_i, & & (5) \end{aligned}$$

Let i^* denote the smallest index for which $\sum_{l=1}^{i^*} (p_{j_l} + q_{j_l}) c_{j_l} > \sum_{i=1}^{|U|} c_i q_i$, and let

$$r^* = \frac{1}{p_{i^*} + q_{i^*}} \left(\sum_{i=1}^{|U|} c_i q_i - \sum_{l=1}^{i^*-1} (p_{j_l} + q_{j_l}) c_{j_l} \right),$$

and change 1 in the first line of (2) to c_i.

The modification of the greedy algorithm is straightforward. We are left with the following

Problem 1 *Is there a clear combinatorial way to give an optimal $\Gamma^\dagger$ solution also for $|A| > 2$? If a Γ has the property, that for all $a, b \in A$, $x_i = \mathbb{P}[\gamma_{u_i} = a]$ is an optimal solution for (3,4,5), is then necessarily Γ an optimal $\Gamma^\dagger$?*

The following example shows that some elements of A may never be recovered by the MLE principle, although all elements of A may be returned by high probability under the minimax criterion. Take $A = \{1, 2, ..., n+1\}$, $U = \{1, 2, ..., n\}$, and define $\xi_i = i$ for sure for $i \leq n$, and $\xi_{n+1} = i$ with probability $\frac{1}{n}$ for $i = 1, 2, ..., n$. Clearly, using MLE we *never* get back $(n+1)$. Define Γ by

$$\gamma_j = \begin{cases} n+1, & \text{with probability } 1/2, \\ j, & \text{with probability } 1/2 \end{cases}$$

for $j = 1, 2, ..., n$. It is not difficult to see that $\mathbb{P}[\gamma_{\xi_i} = i] = 1/2$ for all $i = 1, 2, ..., n+1$.

Problem 2 *How much the MLE criterion can differ from the minimax criterion? Which situations are the worst?*

Problem 3 *There is a natural definition for a $\Gamma : U \to A$. Namely,*

$$\mathbb{P}[\gamma_u = a] = \frac{\mathbb{P}[\xi_a = u]}{\sum_{b \in A} \mathbb{P}[\xi_b = u]}.$$

Is there any optimization criterion under which this Γ is the best random function to invert Ξ?

4 How to use maximum likelihood in phylogeny?

Let us turn now to inverting parametric random functions.

Theorem 6 *For a parametric random function $\Xi : B \to U$, a random function $\Gamma^* : U \to A$ maximizing $\sum_{a \in A} \int R_{a,\theta} d\mu_a(\theta)$ can be obtained by the following rule: for any fixed $u \in U$, $\gamma^*_u = a^*$ for sure if for all $a \in A$,*

$$\int \mathbb{P}[\xi_{(a^*,\theta)} = a^*] d\mu_{a^*}(\theta) \geq \int \mathbb{P}[\xi_{(a,\theta)} = a] d\mu_a(\theta).$$

Proof. Apply Theorem 4 in the following way: for a parametric random function $\Xi : B \to U$, assign a random function $\Upsilon : A \to U$ in the following way:

$$\mathbb{P}[\upsilon_a = u] = \frac{\int \xi_{(a,\theta)} d\mu_a(\theta)}{\int d\mu_a(\theta)}.$$

The value of r_a regarding $\Gamma \circ \Upsilon$ is

$$\frac{\int R_{(a,\theta)} d\mu_a(\theta)}{\int d\mu_a(\theta)},$$

and Theorem 4 applies to $\Gamma \circ \Upsilon$ with $w(a) = \int d\mu_a(\theta)$. □

Theorem 6 is interesting from the following point of view. For phylogeny reconstruction, several stochastic models are in use, like the Cavender-Farris (or Neyman) model [5, 16], the Kimura 3-parameter model [14], and even more general stochastic models [19, 9]. All these models are easily described as parametric random functions. A θ associated to a fixed tree would represent the stochastic mutation mechanism associated with the tree, the models mentioned above associate numbers or matrices to the edges of tree. The model probabilities with which certain biomolecular sequences occur in the leaves of the tree, also depend on these numbers or matrices, and not just on the tree itself. The numbers $\int d\mu_a(\theta)$ can be interpreted as a prior distribution on A if $\sum_{a\in A} \int d\mu_a(\theta) = 1$. Arguments and models have been made in phylogeny to explain that not all trees are equally likely as phylogenetic trees [1, 4, 13]. A prior distribution may be convenient to describe this situation.

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 there exists an $(a^*, \theta^*) \in B$, such that for all $a \in A$, and all $(a, \theta) \in B$, $\mathbb{P}[\xi_{(a^*,\theta^*)} = u] \geq \mathbb{P}[\xi_{(a,\theta)} = u]$. (In case of ties, randomization is possible and usual). Theorem 6 shows that this Γ^* is *not necessarily maximizing* $\sum_{a\in A} \int R_{a,\theta} d\mu_a(\theta)$, since the Duality Theorem yields a different criterion. If one would like to keep this maximizing property, then one has to use MLE in a slightly different way, as it is described in Theorem 6. There are certainly difficulties in doing so,

(1) one needs a measure on the parameters—how can we convince ourselves that we have an appropriate measure?

(2) evaluating the integrals associated with the modified selection criterion can be difficult. Actually people are now starting to use biologically motivated priors on trees [1, 4, 13] and on the edge parameters (but for slightly different things than the modication of MLE that we mention). They [21] are also estimating the integrals (with some success), so perhaps both difficulties can be overcome.

It is a natural question if bad examples like in the previous section can also happen in phylogenetics. Recently Siddall [18] exhibited an example where the parsimony principle [20] beats MLE. This example also shows the difference between the minimax and MLE reconstruction.

Take for example, the set A to be the set of 3 resolved (binary) trees on 4 leaves, each having equal prior probability 1/3. For each tree $a \in A$ with its associated set of 5 edges $E(a)$, randomly select the parameter $\theta = \{(e, \theta_e) : e \in E(a)\}$ by selecting $\theta_e \in (0, 0.5)$ according to a joint probability density function that is everywhere positive, but which concentrates all but δ of its measure into a region for which $p(e) > 0.5 - \epsilon$ for three edges all incident with a single vertex, and $p(e) < \epsilon$ for the other two edges. Suppose we now indendently evolve k sites on these three trees under the Neyman 2-state model [5], in which θ_e is interpreted as the probability that a change of state occurs on edge e. Then the expected reconstruction probability for the (ordinary) maximum likelihood method is (approximately) $\frac{1}{3}$ for ϵ and δ sufficiently small (and k fixed), yet for the maximum parsimony method ([20]) the expected reconstruction probability (over B) is (approximately, ϵ, δ sufficiently small) $1 - (\frac{3}{4})^k$ and so can be arbitrarily close to 1. Note this example also give a phylogenetic example where Maximum Likelihood can also fail to maximize the minimum expected reconstruction probability $\min_{a \in A} \int R_{(a,\theta)} d\mu_a(\theta)$.

References

[1] D. J. Aldous, Probability distributions on cladograms, in: *Discrete Random Structures*, eds. D. J. Aldous and R. Permantle, Springer-Verlag, IMA Vol. in Mathematics and its Applications. Vol. 76, 1995, 1–18.

[2] A. Ambainis, R. Desper, M. Farach, and S. Kannan. Nearly tight bounds on the learnability of evolution. To appear, *Proc. of the 1998 Foundations of Computer Science.*

[3] James O. Berger, *Statistical Decision Theory and Bayesian Analysis* 2nd edition, Springer Series in Statistics, Springer-Verlag, 1985.

[4] J. K. M. Brown, Probabilities of evolutionary trees, *Syst. Biol.* **43** (1994), 78–91.

[5] J. A. Cavender, Taxonomy with confidence, *Math. Biosci.* **40** (1978), 270–280.

[6] M. Cryan, L. A. Goldberg, P. W. Goldberg, Evolutionary trees can be learned in polynomial time in the two-state general Markov model, to appear in *FOCS'98.*

[7] R. Durbin, S. Eddy, A. Krogh and G. Mitchison, *Biological sequence analysis: probabilistic models of proteins and nucleic acids,* Cambridge University Press, Cambridge, 1998. p.205-206.

[8] 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)(2) 153–184.

[9] P. L. Erdős, M. A. Steel, L. A. Székely and T. Warnow, A few logs suffice to build (almost) all trees II, to appear in *Theoretical Computer Science.*

[10] M. Farach, and S. Kannan, Efficient algorithms for inverting evolution, *Proceedings of the ACM Symposium on the Foundations of Computer Science*, 1996, 230–236.

[11] J. Felsenstein, Cases in which parsimony or compatibility methods will be positively misleading, *Syst. Zool.*, **27** (1978), 401–410.

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

[13] E. F. Harding, The probabilities of rooted tree shapes generated by random bifurcation, *Adv. Appl. Probab.* **3** (1971), 44–77.

[14] M. Kimura, Estimation of evolutionary distances between homologous nucleotide sequences. *Proc. Nat. Acad. Sci. USA*, **78** (1981), 454–458.

[15] M. Marcus and H. Minc, *A Survey of Matrix Theory and Matrix Inequalities*, Dover Publications Inc., New York, 1992.

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

[17] A. Schrijver, *Theory of Linear and Integer Programming,* Wiley-Interscience Series in Discrete Mathematics, John Wiley & Sons Ltd., Chichester, 1986.

[18] M. E. Siddall, Success of parsimony in the four-taxon case: long-branch repulsion by likelihood in the Farris Zone, *Cladistics* **14**(3) (1998), 209–220.

[19] M. Steel, M. D. Hendy, D. Penny, Reconstructing trees from nucleotide pattern probabilities: a survey and some new results, to appear in *Discrete Appl. Math.*

[20] D. L Swofford, G. J. Olsen, P. J. Waddell, D. M. Hillis, Chapter 11: Phylogenetic inference, in: *Molecular Systematics*, D. M. Hillis, C. Moritz, B. K. Mable, eds., 2nd edition, Sinauer Associates, Inc., Sunderland, 1996, 407–514.

[21] Z. Yang, B. Rannala, A Markov Chain Monte Carlo Method, *Mol. Biol. Evol.* **14**(7) (1997), 714–724.