

## Inverting Random Functions

Michael A. Steel<sup>1</sup> \* and László A. Székely<sup>2</sup> †

<sup>1</sup>Biomathematics Research Centre, University of Canterbury, Private Bag 4800, Christchurch, New Zealand  
 m.steel@math.canterbury.ac.nz

<sup>2</sup>Department of Mathematics, University of South Carolina, Columbia, SC 29208, USA  
 laszlo@math.sc.edu

Received November 23, 1998

*AMS Subject Classification:* 92D15, 62C20, 90C46, 90C47

**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 affect how we think about maximum likelihood estimation (MLE) in phylogeny. For inverting random functions, MLE is optimal under a first criterion, although it is not optimal under a second criterion which is at least equally natural but more conservative. Furthermore, MLE has to be used differently from the way it has been 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.

*Keywords:* random function, maximum likelihood estimation, minimax problem, phylogeny reconstruction

### 1. Introduction

For two finite sets,  $A$  and  $U$ , give a  $U$ -valued random variable  $\xi_a$  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,  $\Gamma$ , from  $U$  to  $V$ , we can speak about the composition of  $\Gamma$  and  $\Xi$ ,  $\Gamma \circ \Xi : A \rightarrow V$ , 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  $I : A \rightarrow A$  by  $\Gamma \circ \Xi$ . *We always assume*

\* Supported by the New Zealand Marsden Fund.

† Supported by the National Science Foundation grant DMS 9701211, and the Hungarian National Science Fund contract T 016 358.

that  $\Xi$  and  $\Gamma$  are independent, with the exception of Theorem 2.1. The reader may think that this is not the only or 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  $\Xi$  according to the first definition. Consider a fixed ordering of the elements of  $A$ ,  $A = \{a_1, a_2, \dots, a_{|A|}\}$ . For any sequence  $(u_1, u_2, \dots, u_{|A|})$  ( $u_i \in U$ ), set the probability

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

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

We might have also used a “weaker” definition for random functions, namely, a collection of distributions for the  $U$ -valued random variables for all  $a \in A$ , i.e., a collection of probability distributions but no joint distribution. In the 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  $\xi_a$  is independent of every  $\gamma_u$ , the results of this paper would still 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 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 we want to optimize.

Consider the probability of returning  $a$  from  $a$  by the composition of two random functions, i.e.,  $r_a = \mathbf{P}[\gamma_{\xi_a} = a]$ . A natural criterion is to find  $\Gamma$  for a given  $\Xi$  in order to maximize  $\sum_a r_a$ . We may have situations where we are given a weight function  $w : A \rightarrow \mathbf{R}$  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  $a \in A$ ,

$$\mathbf{P}[\xi_{a^*} = u]w(a^*) \geq \mathbf{P}[\xi_a = u]w(a).$$

This function  $\Gamma^*$  is called the *maximum likelihood estimation* (MLE) in the literature [7, 12] (in the case of ties, randomization is possible and usual.) We show that the MLE  $\Gamma^*$  maximizes  $\sum_a r_a w(a)$  for a given  $\Xi$ . However, it is at least natural to look at a more conservative criterion: maximize the smallest value of  $r_a$  for  $a \in A$ . Call this criterion the *mini-max criterion*. For the mini-max criterion MLE is not always optimal. These results have been known in the context of statistical decision theory but not 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  $(\Theta(a), \mu_a(\cdot))$  is assigned, so that  $\mu_a(\Theta(a)) < \infty$ ; 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  $\Xi$  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  $\Xi$ ;
- (ii) for all  $u \in U$ , the set  $\{\theta \in \Theta(a) : \xi_{(a,\theta)} = u\}$  is measurable in the measure space  $(\Theta(a), \mu_a(\cdot))$ .

We are interested in random functions  $\Gamma : U \rightarrow A$  independent from  $\Xi$  so that  $\gamma_{\xi_{(a,\theta)}}$  best approximates  $p$  under certain criteria. Call  $R_{a,\theta}$  the probability  $\mathbf{P}[\gamma_{\xi_{(a,\theta)}} = a]$ . MLE, would take the  $\Gamma^*$ , 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  $(a, \theta) \in B$ ,  $\mathbf{P}[\xi_{(a^*,\theta^*)} = u] \geq \mathbf{P}[\xi_{(a,\theta)} = u]$  (in the 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  $\Gamma^*$  maximizes

$$\sum_{a \in A} \int R_{a,\theta} d\mu_a(\theta). \tag{1.1}$$

This criterion is natural, since if  $\sum_{a \in A} \int d\mu_a(\theta) = 1$ , the formula (1.1) can be interpreted as are 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 [7, pp. 205–206] and [12]). In recent works on phylogeny, the mini-max criterion is gaining popularity; implicitly or explicitly, this criterion is followed in [2, 6, 8–10].

## 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  $|A| \leq |U|$ .

**Theorem 2.1.** *Assume that we have finite sets  $A$  and  $U$  and random functions  $\Xi : A \rightarrow U$  and  $\Gamma : U \rightarrow A$ .*

- (i) *If  $r_a = \mathbf{P}[\gamma_{\xi_a} = a] > \varepsilon$  for all  $a \in A$ , then  $|U| > \varepsilon|A|$ , even without assuming the independence of  $\Xi$  and  $\Gamma$ .*
- (ii) *If  $\Xi$  and  $\Gamma$  are independent and  $r_a = \mathbf{P}[\gamma_{\xi_a} = a] > 1/2$  for all  $a \in A$ , then  $|U| \geq |A|$ .*

*Proof.* (i) By hypothesis,  $\varepsilon|A| < \sum_a \mathbf{P}[\gamma_{\xi_a} = a] = \sum_a \sum_u \mathbf{P}[\xi_a = u \ \& \ \gamma_u = a] \leq \sum_u (\sum_a \mathbf{P}[\gamma_u = a]) = \sum_u 1 = |U|$ .

(ii) First, note that  $\mathbf{P}[\gamma_{\xi_a} = b] = \sum_u \mathbf{P}[\gamma_u = b] \mathbf{P}[\xi_a = u]$  by independence. Arrange the numbers  $\mathbf{P}[\gamma_u = b]$  into an  $|A| \times |U|$  matrix  $M$  and the numbers  $\mathbf{P}[\xi_a = u]$  into an  $|U| \times |A|$  matrix  $N$ . Now, we have  $b[MN]_a = \mathbf{P}[\gamma_{\xi_a} = b]$  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  $i$ ,  $|c_{ii}| > \sum_{i \neq j} |c_{ij}|$ , 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  $\sum_{i \neq j} c_{ij} = 1 - c_{ii} < 1/2$ , and hence  $|A| = \text{rank}(MN) \leq \text{rank}(M) \leq |U|$ .

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

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

contradicting our assumption  $\mathbf{P}[\gamma_{\xi_a} = a] > 1/2$ . 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 are mapped to  $u$ , then  $1 = \sum_a \mathbf{P}[\gamma_u = a] > 1/2 + 1/2$ . ■

Note that the message of Theorem 2.1(ii) is that relaxing the requirement for reconstructing functions *for sure* to reconstructing functions *with probabilities exceeding 1/2* does not allow *any* relaxation on the size of  $U$ . Theorem 2.1(ii) is no longer valid if we give up the independence of  $\Xi$  and  $\Gamma$ , an observation due to Peter Winkler. Indeed, give a random variable  $\nu$  which has uniform distribution on  $\{1, 2, \dots, n\}$ . We give up independence by using the *same* experiment for  $\nu$  to define the random function and to design its inverse. Take  $A = \{1, 2, \dots, n+1\}$ ,  $U = \{1, 2, \dots, n\}$ , and define  $\xi_i = i$  for sure for  $i \leq n$ , and  $\xi_{n+1} = i$  if  $\nu = i$ . Define the random function  $\Gamma$  in the following way. If  $\nu = i$ , then make a coin toss  $\tau$  independently from  $\nu$ , so that  $\mathbf{P}[\tau = \text{HEAD}] = 1/3$  and  $\mathbf{P}[\tau = \text{TAIL}] = 2/3$ . Set

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

Using conditional probabilities, it is easy to see that

$$\mathbf{P}[\gamma_{\xi_{n+1}} = n+1] = \sum_{l=1}^n \frac{2}{3} \mathbf{P}[\nu = l] = 2/3,$$

and that for  $i \leq n$ ,

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

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

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

**Theorem 2.2.** *Assume that we have finite sets  $A$  and  $U$  and random functions  $\Xi : A \rightarrow U$  and  $\Gamma : U \rightarrow A$ . If, for all  $u \in U$ ,  $\max_a \mathbf{P}[\xi_a = u] \leq \lambda(u)$ , then for  $r_a = \mathbf{P}[\gamma_a = a]$ , 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

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

■

**Theorem 2.3.** *Assume that we have finite sets  $A$  and  $U$  and random functions  $\Xi : A \rightarrow U$  and  $\Gamma : U \rightarrow A$ . Let  $d(a, b)$  for  $a, b \in A$  denote the variational distance  $\sum_{u \in U} |\mathbf{P}[\xi_a = u] - \mathbf{P}[\xi_b = u]|$ . 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*

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

*Proof.* Set  $r_{ac} = \mathbf{P}[\gamma_{\xi_a} = c]$ . 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

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

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

Using independence, we obtain

$$x_u = \mathbf{P}[\xi_a = u] \mathbf{P}[\gamma_u = a] \text{ and } y_u = \mathbf{P}[\xi_b = u] \mathbf{P}[\gamma_u = a].$$

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

### 3. Optimization Criteria for Inverting Random Functions

**Theorem 3.1.** *Assume that we have finite sets  $A$  and  $U$ , a function  $w : A \rightarrow \mathbf{R}$ , and a random function  $\Xi : A \rightarrow U$ . A random function  $\Gamma^* : U \rightarrow A$  maximizing  $\sum_a r_a w(a)$  can be defined in the following way: For any fixed  $u \in U$ , set  $\gamma_u^* = a^*$  for sure if, for all  $a \in A$ ,  $\mathbf{P}[\xi_{a^*} = u] w(a^*) \geq \mathbf{P}[\xi_a = u] w(a)$ .*

*Proof.* This theorem occurs in the setting of statistical decision theory in [3, p.159]. For completeness, we give a proof. A more general result will be given in Section 4. We have to solve the following linear program in order to find  $\Gamma^*$ :

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

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 in the following setting: Collect the values  $\mathbf{P}[\gamma_u = a]$  in a column vector  $y$  of length  $|U||A|$ , the values  $-\mathbf{P}[\xi_a = u] w(a)$  into a row vector  $b$  of length  $|U||A|$ ,  $c^T = (1, 1, \dots, 1)$  of length  $|U|$ , and define a  $|U||A| \times |U|$  matrix  $M$  by setting

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

The dual problem, in variables  $x_u$  ( $u \in U$ ), becomes the following linear program:

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

Clearly,

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

and hence, for every feasible solution  $x_u$ , we have

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

This bound for the dual objective function is attained by setting the values of  $\Gamma^*$  as it is in the statement of this theorem. ■

**Theorem 3.2.** *Assume that we have finite sets  $A$  and  $U$  and a random function  $\Xi : A \rightarrow U$ . The random function  $\Gamma^\dagger : U \rightarrow A$  maximizing  $\min_{a \in A} r_a$  has the following good characterization:*

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

where  $\mu$  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:

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

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  $A = \{a, b\}$ ,  $p_i = \mathbf{P}[\xi_a = u_i]$ , and  $q_i = \mathbf{P}[\xi_b = u_i]$ . Denote  $x_i = \mathbf{P}[\gamma_{u_i} = a]$  and  $1 - x_i = \mathbf{P}[\gamma_{u_i} = b]$ . 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 solution has 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  $j_1, j_2, \dots, j_{|U|}$ . 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

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

Setting

$$x_{j_i} = \begin{cases} 1, & \text{if } i < i^*, \\ r^*, & \text{if } i = i^*, \\ 0, & \text{if } i > i^* \end{cases} \tag{3.1}$$

provides the required solution.

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

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

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

$$0 \leq x_i \leq c_i \tag{3.2}$$

$$\sum_i p_i x_i = \sum_i (c_i - x_i) q_i \tag{3.3}$$

$$\max \sum_i p_i x_i. \tag{3.4}$$

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 (3.1) to  $c_i$ .

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

**Problem 3.3.** *Is there a clear combinatorial way to give an optimal  $\Gamma^\dagger$  solution also for  $|A| > 2$ ? If a  $\Gamma$  has the property that, for all  $a, b \in A$ ,  $x_i = \mathbf{P}[\gamma_{u_i} = a]$  is an optimal solution for (3.2)–(3.4), is  $\Gamma$  then necessarily 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 with high probability

under the mini-max criterion. Take  $A = \{1, 2, \dots, n + 1\}$ ,  $U = \{1, 2, \dots, n\}$ , and define  $\xi_i = i$  for sure for  $i \leq n$ , and  $\xi_{n+1} = i$  with probability  $1/n$  for  $i = 1, 2, \dots, n$ . Clearly, using MLE, we *never* retrieve  $(n + 1)$ . Define  $\Gamma$  by

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

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

**Problem 3.4.** *How much can the MLE criterion differ from the mini-max criterion? Which situations are the worst?*

**Problem 3.5.** *There is a natural candidate for a  $\Gamma : U \rightarrow A$ , namely,*

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

*Is there any optimization criterion under which this  $\Gamma$  is the best random function to invert  $\Xi$ ?*

#### 4. How to Use Maximum Likelihood in Phylogeny

Let us turn now to inverting parametric random functions.

**Theorem 4.1.** *For a parametric random function  $\Xi : B \rightarrow U$ , a random function  $\Gamma^* : U \rightarrow A$  maximizing  $\sum_{a \in A} \int R_{a,\theta} d\mu_a(\theta)$  can be obtained by the following rule: For any fixed  $u \in U$ , set  $\gamma_u^* = a^*$  for sure if, for all  $a \in A$ ,*

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

*Proof.* Apply Theorem 3.1 in the following way: For a parametric random function  $\Xi : B \rightarrow U$ , assign a random function  $\Psi : A \rightarrow U$  in the following way:

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

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

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

and Theorem 3.1 applies to  $\Gamma \circ \Psi$  with  $w(a) = \int d\mu_a(\theta)$ . ■

Theorem 4.1 is interesting from the following point of view. For phylogeny reconstruction, several stochastic models have been employed, for example, the Neyman 2-state (or Cavender–Farris) model [5, 16], the Kimura 3-parameter model [14], and even more general stochastic models [9, 19]. All these models are easily described as

parametric random functions. A  $\theta$  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 applied in phylogeny to suggest that not all trees are equally likely as phylogenetic trees [1, 4, 13]. A prior distribution may be convenient to describe this situation.

MLE, as it is used in the practice of phylogeny reconstruction, would take the  $\Gamma^*$ , 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  $(a, \theta) \in B$ ,  $\mathbf{P}[\xi_{(a^*, \theta^*)} = u] \geq \mathbf{P}[\xi_{(a, \theta)} = u]$  (in the case of ties, randomization is possible and usual). Theorem 4.1 shows that this  $\Gamma^*$  does *not necessarily maximize*  $\sum_{a \in A} \int R_{a, \theta} d\mu_a(\theta)$  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 4.1. 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 modification 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 whether MLE-disadvantaged examples, mentioned in the previous section, can also occur in phylogenetics. Recently Siddall [18] exhibited an example where the parsimony principle [20] beats MLE. This example also shows the difference between mini-max and MLE reconstruction.

Take, for example, the set  $A$  to be the set of three unrooted binary trees on four leaves, each having equal prior probability  $1/3$ . For each tree  $a \in A$  with its associated set of five 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 positive everywhere, but which concentrates all but  $\delta$  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 independently evolve  $k$  sites on these three trees under the Neyman 2-state model [5], in which  $\theta_e$  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)  $1/3$  for  $\epsilon$  and  $\delta$  sufficiently small (and  $k$  fixed), yet for the maximum parsimony method ([20]) the expected reconstruction probability (over  $B$ ) is (approximately,  $\epsilon, \delta$  sufficiently small)  $1 - (3/4)^k$  and so can be arbitrarily close to 1. Note that this example also give a phylogenetic example where Maximum Likelihood can also fail to maximize the minimum expected reconstruction probability  $\min_{a \in A} \int R_{(a, \theta)} d\mu_a(\theta)$ .

**Acknowledgments.** The authors are indebted to Éva Czabarka for her invaluable comments on the manuscript. This research started when the second author visited the University of Canterbury

under the support of the New Zealand Marsden Fund.

## References

1. D.J. Aldous, Probability distributions on cladograms, In: *Discrete Random Structures*, D.J. Aldous and R. Pemantle, Eds., IMA Vol. in Mathematics and its Applications, Vol. 76, Springer-Verlag, 1995, pp. 1–18.
2. A. Ambainis, R. Desper, M. Farach, and S. Kannan, Nearly tight bounds on the learnability of evolution, *Proc. of the 38th IEEE Conference on the Foundations of Computer Science (FOCS'97)*, 1997, pp. 524–533.
3. J.O. Berger, *Statistical Decision Theory and Bayesian Analysis*, 2nd Ed., 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, and P.W. Goldberg, Evolutionary trees can be learned in polynomial time in the two-state general Markov model, *Proc. 39th Foundations of Computer Science (FOCS'98)*, 1998, pp. 436–445.
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.
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) 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, pp. 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*, S.S. Gupta and J. Yackel, Eds., Academic Press, New York, 1971, pp. 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** (1998) 209–220.
19. M. Steel, M.D. Hendy, and D. Penny, Reconstructing trees from nucleotide pattern probabilities: A survey and some new results, *Discrete Appl. Math.* **88** (1998) 367–396.
20. D.L. Swofford, G.J. Olsen, P.J. Waddell, and D.M. Hillis, Chapter 11: Phylogenetic inference, In: *Molecular Systematics*, D.M. Hillis, C. Moritz, and B.K. Mable, Eds., 2nd Ed., Sinauer Associates, Inc., Sunderland, 1996, pp. 407–514.
21. Z. Yang and B. Rannala, A Markov Chain Monte Carlo Method, *Mol. Biol. Evol.* **14** (1997) 714–724.