Ann. Comb. 13 (2009) 365–382
DOI 10.1007/s00026-009-0023-z
Published online October 23, 2009
© Birkhäuser Verlag Basel/Switzerland 2009

Annals of Combinatorics

Inverting Random Functions III: Discrete MLE Revisited*

Mike A. Steel[1] and László A. Székely[2]

[1]Biomathematics Research Centre, Mathematics and Statistics Department, University of Canterbury, Private Bag 4800, Christchurch 8041, New Zealand
m.steel@math.canterbury.ac.nz

[2]Department of Mathematics, University of South Carolina, LeConte College, 1523 Greene Street, Columbia, SC 29208, USA
szekely@math.sc.edu

Received August 3, 2006

AMS Subject Classification: 60C05, 62B10, 92B10, 94A17

Abstract. This paper continues our earlier investigations into the inversion of random functions in a general (abstract) setting. In Section 2, we investigate a concept of invertibility and the invertibility of the composition of random functions defined on finite sets. In Section 3, we resolve some questions concerning the number of samples required to ensure the accuracy of maximum likelihood estimation (MLE) in the presence of 'nuisance' parameters. A direct application to phylogeny reconstruction is given.

Keywords: random function, maximum likelihood estimation, phylogeny reconstruction

1. Review of Random Functions

This paper is a sequel of our earlier papers [14, 15]. We assume that the reader is familiar with those papers, 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 $(\xi_a\colon\ a \in A)$ a *random function* $\Xi\colon A \to U$. 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, $\Gamma \circ \Xi\colon A \to V$, which is the vector variable $\left(\gamma_{\xi_a} : a \in A\right)$. In this paper we are concerned with inverting random functions. In other words, we look for random functions $\Gamma\colon\ U \to A$ in order to obtain the best approximations of the identity function $\iota\colon A \to A$ 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.

* We thank the NZIMA Maclaurin Fellowship, the Phylogeny programme at the Isaac Newton Institute of Cambridge University, and Hungarian Bioinformatics MTKD-CT-2006-042794 for supporting this research. The second author was also supported in part by NSF DMS contracts 007 2187, 070 1111, and NIH NIGMS 1 R01 GM078991-01.

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

$$r_a = \sum_{u \in U} \mathbb{P}\left[\xi_a = u\right] \cdot \mathbb{P}\left[\gamma_u = a\right]. \qquad (1.1)$$

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 $w\colon A \to \mathbb{R}^+$ and we may wish to maximize $\sum_a r_a w(a)$. This can happen if we give preference to returning certain values of a, 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. The following random function $\Gamma^*\colon U \to A$, defined below, will perform this task. For any fixed $u \in U$,

$$\gamma_u^* = a^* \text{ for sure, if for all } a \in A, \quad \mathbb{P}\left[\xi_{a^*} = u\right] w\left(a^*\right) \geq \mathbb{P}\left[\xi_a = u\right] w(a). \qquad (1.2)$$

(If more than one element a^* satisfies (1.2), 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 [6]. The special case when the weight function w is constant is known as the *maximum likelihood estimation* (MLE) [2,6].

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

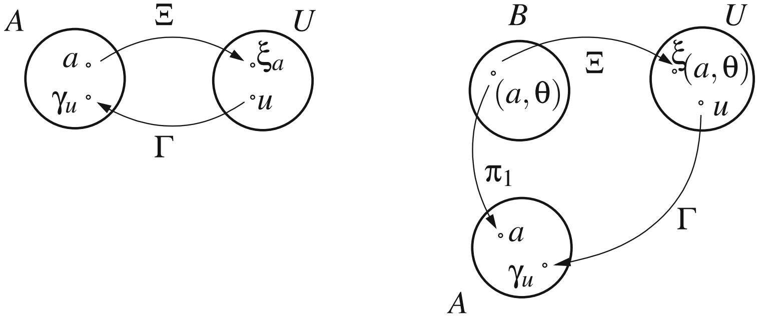
$$d(a,b) := d\left(\xi_a, \xi_b\right) = \sum_{u \in U} \left|\mathbb{P}\left[\xi_a = u\right] - \mathbb{P}\left[\xi_b = u\right]\right|, \qquad (1.3)$$

which is called the *variational distance* between the random variables ξ_a and ξ_b.

Any given $\Xi\colon A \to U$ will have an $|A| \times |U|$ *associated matrix* X, such that $x_{au} = \mathbb{P}\left[\xi_a = u\right]$. Given $\Gamma\colon U \to V$ with an associated matrix G, the composition of Γ and Ξ, $\Gamma \circ \Xi\colon A \to V$, will have the associated matrix XG.

Our motivation for the study of random functions came from phylogeny reconstruction [8,12]. 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* can be viewed as 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. However, there is an important extra complication in phylogeny reconstruction, namely, the random sequences that a tree generates (under some Markov model) typically depends on other (generally continuous) parameters associated with the tree, about which little, if anything, is known. With this motivation our paper [14] considered an abstract model for phylogeny reconstruction: Inverting random functions between finite sets, with associated additional ‘nuisance’ parameters. Most of the work done on the mathematics of phylogeny reconstruction can be discussed in this context. This model is more structured than when nuisance parameters are absent, and hence is better suited to describe details of phylogenetic models and the evolution of biomolecular sequences. We now describe this general setting (note that in [14, 15] we referred to the distinction between the presence and absence of nuisance parameters as the 'parametric setting' and the 'non-parametric setting', respectively; but here we adopt the more standard statistical terminology).

Assume that for a finite set A, for every $a \in A$, an (arbitrary, finite, or infinite) set $\Theta(a) \neq \emptyset$ is assigned, and moreover, $\Theta(a) \cap \Theta(b) = \emptyset$ for $a \neq b$. Set $B = \{(a, \theta)\colon a \in A, \theta \in \Theta(a)\}$ and let π_1 denote the natural projection from B to A. A *random function with nuisance parameters* is the collection Ξ of random variables such that for $a \in A$ and $\theta \in \Theta(a)$, there is a (unique) U-valued random variable $\xi_{(a,\theta)}$ in Ξ.



(a) (b)

Figure 1: Inversion of random functions without (a), and with (b) nuisance parameters.

We are interested in random functions $\Gamma\colon U \to A$ that are independent from Ξ so that $\gamma_{\xi_{(a,\theta)}}$ best approximates π_1 under certain criteria. Let $R_{(a,\theta)}$ denote the probability $\mathbb{P}\left[\gamma_{\xi_{(a,\theta)}} = a\right]$. Maximum Likelihood Estimation, as it is used in situations where there is a discrete parameter of interest to estimate but where additional nuisance parameters are also present (such as phylogeny reconstruction), corresponds to the random function Γ' for which, for every fixed u, $\gamma'_u = a'$ with probability 1 if,

$$\text{for all } (a, \theta) \in B, \text{ there exists } \theta' \in \Theta(a') \text{ with } \mathbb{P}\left[\xi_{(a',\theta')} = u\right] \geq \mathbb{P}\left[\xi_{(a,\theta)} = u\right]. \tag{1.4}$$

In case there is more than one element a' that satisfies (1.4), we may select uniformly at random from the set of such elements. (We avoided using the more natural looking quantification for (1.4) of requiring the existence of θ' in $\Theta(a')$ for all $(a, \theta) \in B$, since $\mathbb{P}\left[\xi_{(a',\theta')} = u\right]$ may not take a maximum value.) We denote by $R'_{(a,\theta)}$ the probability that from the pair (a, θ) the Maximum Likelihood Estimation Γ' returns a, i.e.,

$$R'_{(a,\theta)} = \mathbb{P}\left[\gamma'_{\xi_{(a,\theta)}} = a\right]. \tag{1.5}$$

If a random function $\Xi\colon A \to U$ (or $\Xi\colon B \to U$, as appropriate) is to have k independent evaluations, then we denote the resulting random function by $\Xi^{(k)}\colon A \to U^k$

$\big($or $\Xi^{(k)}\colon B \to U^k$, respectively$\big)$, and the random variable associated with a will be $\xi_a^{(k)}$. We will study the invertibility of $\Xi^{(k)}$ both with and without nuisance parameters. For a random function $\Gamma\colon U^k \to A$, we use the notation $r_a^{(k)} = \mathbb{P}\left[\gamma_{\xi_a^{(k)}} = a\right]$ in the absence of nuisance parameters, $R_{(a,\theta)}^{(k)} = \mathbb{P}\left[\gamma_{\xi_{(a,\theta)}^{(k)}} = a\right]$ in the presence of nuisance parameters, and $\left[R^{(k)}\right]'_{(a,\theta)}$ if Γ' is the Maximum Likelihood Estimation.

In Section 2, we will show that in the absence of nuisance parameters, several natural definitions of invertibility of a random function are, in fact, equivalent. The main result of Section 2 is an explicit bound on how invertibility "improves" as the variational distances between elements of A become increasingly separated from zero. Furthermore, we determine when the composition of invertible random functions is invertible.

In Section 3, we revisit our study of the worst-case behaviour of MLE in [15]. (This is a very natural question in situations where a prior distribution is not given on A, or the inverting of the random function is to be carried out only once. This situation arises in phylogeny reconstruction where, arguably, we do not have a prior distribution on alternative evolutionary scenarios, and the reconstruction is not going to be repeated — there is only one 'Tree of Life' that we want to explore.) A certain amount of controversy and debate has surrounded the statistical consistency of MLE in phylogeny, as described in [8, pp. 270–272]. Felsenstein's claim (from the early 1970s) of the consistency of MLE in phylogeny for simple ('identifiable') models is correct, but it was only formally established in 1996 by [3]. This result, like Wald's earlier result [17], relies on a compactness argument, continuity and limit theory, that does not give an explicit bound on k (the sequence length of i.i.d. observations). Other proofs in the biological literature have generally been less rigorous, and led to criticism and debate (see, e.g., [1,7,9,10,13,18,19]). One oversight has been to treat the MLE-estimated continuous parameters (branch lengths) of alternative trees as fixed rather than as random variables dependent on the data; such arguments are satisfying for practical purposes but call for more rigor. The significance of [15, Theorem 5.1] is that it gives the first explicit bounds for MLE, both in the phylogenetic setting and beyond. However, this result depended on an unnatural parameter, namely, the smallest positive probability that an image of the object to be reconstructed can have. Here in Theorem 3.3, we get rid of this dependence, and provide a simple and immediate application of this new result to phylogeny reconstruction.

We study two examples that reveal the subtleties of using MLE for inverting random functions in the presence of nuisance parameters. The first example shows that Theorem 3.3 is "near optimal" in one of its parameters. The second example shows that in contrast to the setting where there are no nuisance parameters, the vanishing of variational distance does not by itself preclude MLE (or any other estimation) for certain random functions.

Our approach is information-theoretic: We focus on the possibility or impossibility of inverting random functions, and not on the computational complexity issues. Our results can also be re-stated in the language of decision theory, by talking about the 'loss function' and 'risk function' associated with the decision rule.

2. Invertibility in the Absence of Nuisance Parameters

Let us say that a random function $\Xi\colon A \longrightarrow U$ is *invertible* if there exists a random function $\Gamma\colon U \longrightarrow A$ such that for all $a \in A$, $\mathbb{P}\left[\gamma_{\xi_a} = x\right]$ is strictly maximized when $x = a$, or equivalently:

$$\mathbb{P}\left[\gamma_{\xi_a} = a\right] - \max_{x \neq a}\left\{\mathbb{P}\left[\gamma_{\xi_a} = x\right]\right\} > 0, \qquad \text{for all } a \in A. \tag{2.1}$$

Informally, Ξ is invertible if there is some reconstruction method that is always more likely to pick the generating object in A than any other element of A.

A sufficient condition for Ξ to be invertible is that there exists a Γ so that for all $a \in A$, the following two conditions apply:

(I_1) $\mathbb{P}\left[\gamma_{\xi_a} = a\right] > \frac{1}{|A|}$,
(I_2) $\mathbb{P}\left[\gamma_{\xi_a} = b\right] < \frac{1}{|A|}$, for all $b \neq a$.

Note that invertibility implies (I_1), and is equivalent to this condition when $|A| = 2$, but not equivalent for $|A| \geq 3$.

We say Ξ *separates* A, if, for each distinct pair $a, b \in A$, the variational distance $d(a,b)$ of the probability distributions of ξ_a and ξ_b is strictly positive (that is, the parameter in A is identifiable from its probability distribution).

Proposition 2.1. *The following properties are equivalent for an $\Xi\colon A \longrightarrow U$ random function*:

(i) *Ξ separates A.*
(ii) *For all $\varepsilon > 0$ there is a value of k_ε so that for all $k \geq k_\varepsilon$ there is a random function $\Gamma^{\S}\colon U^k \longrightarrow A$ such that for all $a \in A$, $\mathbb{P}\left[\gamma^{\S}_{\xi_a^{(k)}} = a\right] > 1 - \varepsilon$.*
(iii) *Ξ is invertible.*
(iv) *For some $k \geq 1$, $\Xi^{(k)}$ is invertible.*
(v) *For some $k \geq 1$, $\Xi^{(k)}$ satisfies (I_1) and (I_2).*

Proof. We will show that (i) ⇔ (ii), (iv) ⇒ (ii), (i) ⇒ (iii), and (ii) ⇒ (v). Since (iii) ⇒ (iv) is trivial, and (v) ⇒ (iv) clearly holds, this will establish the five-way equivalence of (i) through (v).

(i) ⇔ (ii) The implication (i) ⇒ (ii) follows from the statistical consistency of maximum likelihood estimation under the separation condition (an explicit bounds on k_ε is provided by [14, Theorem 3.2]). For the implication (ii) ⇒ (i), let E_b be the event that $\gamma^{\S}_{\xi_b^{(k)}} = a$. Then $\mathbb{P}(E_a) > 1 - \varepsilon$, while for any $b \neq a$, $\mathbb{P}(E_b) < \varepsilon$. Consequently, if we select any $\varepsilon \in (0, 1/2)$, we have

$$d^{(k)}(a,b) \geq 2(1 - 2\varepsilon) > 0,$$

where $d^{(k)}$ denotes the variational distance between the random variables $\xi_a^{(k)}$ and $\xi_b^{(k)}$. However, $d^{(k)}(a,b) > 0$ implies that $d(a,b) > 0$ (for example, by [15, Equations (2.3) and (2.4)]) and since this holds for all distinct pairs a, b, then Ξ separates A.

(iv) $\Rightarrow$ (ii) Suppose that $\Xi^{(k)}$ is invertible. Select Γ to satisfy (2.1) for $\Xi^{(k)}$. For a positive integer m, generate km independent samples in U according to Ξ. Define $\Gamma^{\S}: U^{km} \rightarrow A$ as follows: Select the elements of A that are reconstructed most often according to Γ and choose one of them uniformly at random. By standard probability arguments, the probability that the correct element a will be selected by this process converges to 1 as m tends to infinity.

(i) $\Rightarrow$ (iii) Suppose that $\Xi\colon A \rightarrow U$ separates A. Let X denote the associated matrix of Ξ, and let $\mathbf{a}_i$, $i \in A$ denote the rows of X. Recall that $\mathbf{a}_i$ gives the distribution of ξ_i. We will describe the inverse random function $\Gamma\colon U \rightarrow A$ with its associated matrix, i.e., in the form of a $|U| \times |A|$ matrix G, whose rows represent the distribution of the element of U corresponding to the row.

For $\mu > 0$ we write $G_\mu = \mu V + \frac{1}{|A|}J$, where J is the $|U| \times |A|$ matrix that every entry equals to 1, and where V will be explicitly described shortly $\big($if we were to take V as being equal to the zero matrix, then (2.1) yields zero uniformly instead of the desired strictly positive value by $XG_\mu = X \cdot \frac{1}{|A|}J = \frac{1}{|A|}J\big)$. We denote the columns of V by $\mathbf{v}_i$, $i \in A$. We define each vector $\mathbf{v}_i$ as follows:

$$\mathbf{v}_i = \frac{\mathbf{a}_i}{|\mathbf{a}_i|} - \frac{1}{|A|} \sum_{j=1}^{|A|} \frac{\mathbf{a}_j}{|\mathbf{a}_j|},$$

where $|\cdot|$ is the usual euclidean vector norm (note that $|\mathbf{a}_i| > 0$). It is easy to check now that $\mathbf{a}_i \cdot \mathbf{v}_i - \mathbf{a}_i \cdot \mathbf{v}_j = |\mathbf{a}_i| - \mathbf{a}_i \cdot \frac{\mathbf{a_j}}{|\mathbf{a}_j|} > 0$. $\big($The Cauchy-Schwartz inequality, (i), and the fact that $\mathbf{a}_i$ is a probability distribution imply that $|\mathbf{a}_i| - \mathbf{a}_i \cdot \frac{\mathbf{a_j}}{|\mathbf{a}_j|} > 0$ for $i \neq j$.$\big)$ It is easy to see that $\sum_{l \in A} \mathbf{v}_l = 0$, and therefore the row sums of G_μ are equal to 1. Hence G_μ is the matrix of a random function, if all of its entries are non-negative. This can be achieved by selecting a sufficiently small positive μ. Let Γ denote the random function whose matrix is G_μ. Now for i, $j \in A$, we have $\mathbb{P}\left[\gamma_{\xi_i} = j\right] = {}_i\left[XG_\mu\right]_j = \frac{1}{|A|} + \mu \mathbf{a}_i \cdot \mathbf{v}_j$, and this together with the inequality $\mathbf{a}_i \cdot \mathbf{v}_i - \mathbf{a}_i \cdot \mathbf{v}_j > 0$ imply (2.1), which is the definition of invertibility.

(ii) $\Rightarrow$ (v) Suppose that Ξ satisfies (ii). Let $\varepsilon = 1/|A|$ and select any value $k \geq k_\varepsilon$, for which the inequality in part (ii) holds. Then $\Xi^{(k)}$ satisfies (I_1) and (I_2). ∎

2.1. Explicit Bounds

From Proposition 2.1, if Ξ separates A then there is a random function $\Gamma\colon U \rightarrow A$ for which

$$\mathbb{P}\left[\gamma_{\xi_a} = a\right] - \frac{1}{|A|} > 0.$$

We now consider putting an explicit lower bound on the right hand side of this inequality. That is, we show that for a specific continuous positive function $h\colon \mathbb{R} \rightarrow \mathbb{R}$ (dependent only on $|A|$), the following holds: Suppose that $d(a, b) > \delta$ for all $a, b \in A$, $a \neq b$, then there is a random function $\Gamma\colon U \rightarrow A$ for which

$$\mathbb{P}\left[\gamma_{\xi_a} = a\right] - \frac{1}{|A|} > h(\delta),$$

for all $a \in A$.

Note that we cannot insist that Γ is MLE (maximum likelihood estimation), even when $|A| = 2$. To see this, let $A = \{1, 2\}, U = \{u_1, u_2\}$, ξ_1 take the value u_1 with probability 1, and let ξ_2 take the values u_1, u_2 with probabilities $\frac{2}{3}$ and $\frac{1}{3}$, respectively; then if $\Gamma = \Gamma^*$ is MLE, we have $\mathbb{P}\left[\gamma_{\xi_2} = 2\right] = \frac{1}{3}$.

Theorem 2.2. *For every random function $\Xi\colon A \to U$, with $|A| > 1$, there exists a $\Gamma\colon U \to A$, such that*

$$\min_{a\in A} r_a \geq \frac{1}{|A|} + \frac{1}{2|A|(|A|-1)} \min_{a\in A} \sum_{b\in A} d(a, b). \tag{2.2}$$

In particular, if for all $a \neq b \in A$, $d(a, b) \geq \delta$, then $\min_{a\in A} r_a \geq \frac{1}{|A|} + \frac{\delta}{2|A|}$.

Proof. Recall the characterization of the random inverse function maximizing $\min_{a\in A} r_a$ from [14, Theorem 5]

$$\min_{a\in A} r_a = \min_{\mu} \sum_{u\in U} \max_{a\in A} \mu(a)\mathbb{P}\left[\xi_a = u\right], \tag{2.3}$$

where μ is a probability distribution on A. In the rest of the proof, μ refers to this minimizing distribution. (Note that [14, Theorem 5] contains an annoying typo, it shows maximization for μ instead of minimization.) We will use the following lemma:

Lemma 2.3. *Suppose that we have real numbers $b_1, b_2, \ldots, b_n$ for which*

$$\sum_{1\leq i<j\leq n} \left|b_i - b_j\right| \geq (n-1)\varepsilon.$$

Then $\max_j \left[b_j - \frac{1}{n}\sum_{i=1}^{n} b_i\right] \geq \frac{\varepsilon}{n}$.

Proof. Without loss of generality, we may assume $b_1 \geq b_2 \geq \cdots \geq b_n$. The conditions of the Lemma can be rewritten as the conditions of the following primal linear program:

$$\text{maximize } \frac{1}{n}\sum_{i}(b_i - b_1),$$

subject to

$$b_2 - b_1 \leq 0,$$

$$b_3 - b_2 \leq 0,$$

$$\vdots$$

$$b_n - b_{n-1} \leq 0,$$

$$\sum_{i<j} b_i - b_j \leq (n-1)\varepsilon.$$

Recall the Duality Theorem of linear programming [11]: max $\left\{c^T b\colon Mb \le d\right\}$ = min $\left\{x^T d\colon x \ge 0, x^T M = c\right\}$, if both optimization problems have feasible solutions. The dual linear program is as follows:

$$\text{minimize } (n-1)\varepsilon x_n$$

subject to

$$(n-1)x_n - x_1 = -\frac{n-1}{n},$$

$$x_i - x_{i+1} + (n-2i-1)x_n = \frac{1}{n}, \quad \text{for } i = 1,2,\ldots,n-2;$$

$$x_{n-1} + (1-n)x_n = \frac{1}{n},$$

$$x_1, x_2, \ldots, x_n \ge 0.$$

It is easy to see that for the dual problem, a feasible solution is the following setting: $x_i = 1 - \frac{i(i-1)}{n(n-1)}$ for $i = 1,2,\ldots,n-1$, and $x_n = \frac{1}{n(n-1)}$; with the value $\frac{\varepsilon}{n}$. Observe that $\max \frac{1}{n}\sum(b_i - b_1) = -\min \frac{1}{n}\sum(b_i - b_1) = \min(n-1)\varepsilon x_n \le \frac{\varepsilon}{n}$. This implies that $\frac{\varepsilon}{n} \le \max_j b_j - \frac{1}{n}\sum_{i=1}^n b_i$ for any feasible solution of the primal problem. ∎

We are going to apply Lemma 2.3 in the following setting. Fix an arbitrary $u \in U$ and for $i \in A$, let $b_i = \mu(i)\mathbb{P}[\xi_i = u]$. Then Lemma 2.3 $\big($with $n = |A|$ and $\varepsilon = \sum_{1\le i<j\le n} |b_i - b_j|\big)$ yields

$$\max_{a\in A}\left(\mu(a)\mathbb{P}[\xi_a = u] - \frac{1}{|A|}\sum_{i\in A}\mu(i)\mathbb{P}[\xi_i = u]\right) \tag{2.4}$$

$$\ge \frac{1}{|A|(|A|-1)}\sum_{1\le i<j\le |A|}\big|\mu(i)\mathbb{P}[\xi_i = u] - \mu(j)\mathbb{P}[\xi_j = u]\big|. \tag{2.5}$$

Observe the identity

$$\sum_{u\in U}\frac{1}{|A|}\sum_{i\in A}\mu(i)\mathbb{P}[\xi_i = u] = \frac{1}{|A|}\sum_{i\in A}\mu(i)\sum_{u\in U}\mathbb{P}[\xi_i = u] = \frac{1}{|A|}. \tag{2.6}$$

Now (2.3) and identity (2.6) imply that

$$\min_{a\in A} r_a = \frac{1}{|A|} + \sum_{u\in U}\max_{a\in A}\left\{\mu(a)\mathbb{P}[\xi_a = u] - \frac{1}{|A|}\sum_{i\in A}\mu(i)\mathbb{P}[\xi_i = u]\right\},$$

and so inequality (2.5) implies

$$\min_{a\in A} r_a \ge \frac{1}{|A|} + \frac{1}{|A|(|A|-1)}\sum_{u\in U}\sum_{1\le i<j\le |A|}\big|\mu(i)\mathbb{P}[\xi_i = u] - \mu(j)\mathbb{P}[\xi_j = u]\big|. \tag{2.7}$$

Fix arbitrary $a, b \in A$, and set $Q = \sum_{u\in U}|\mu(a)\mathbb{P}[\xi_a = u] - \mu(b)\mathbb{P}[\xi_b = u]|$.

Define

$$\begin{aligned} U^{>} &= \{u \in U \colon \mathbb{P}[\xi_a = u] > \mathbb{P}[\xi_b = u]\}, \\ U^{=} &= \{u \in U \colon \mathbb{P}[\xi_a = u] = \mathbb{P}[\xi_b = u]\}, \\ U^{<} &= \{u \in U \colon \mathbb{P}[\xi_a = u] < \mathbb{P}[\xi_b = u]\}. \end{aligned}$$

Define further $A^+ = \sum_{u\in U^>} \mathbb{P}[\xi_a = u]$, $A^- = \sum_{u\in U^<} \mathbb{P}[\xi_a = u]$, $B^+ = \sum_{u\in U^>} \mathbb{P}[\xi_b = u]$, $B^- = \sum_{u\in U^<} \mathbb{P}[\xi_b = u]$. Observe that

$$d(a,b) = \sum_{u\in U} \big|\mathbb{P}[\xi_a = u] - \mathbb{P}[\xi_b = u]\big| = A^+ - B^+ + B^- - A^-.$$

On the other hand,

$$A^+ + A^- = 1 - \sum_{u\in U^=} \mathbb{P}[\xi_a = u] = 1 - \sum_{u\in U^=} \mathbb{P}[\xi_b = u] = B^+ + B^-.$$

From the last two equations, we conclude that $d(a,b) = 2(A^+ - B^+) = 2(B^- - A^-)$. We finish the proof by setting a lower bound on Q with a case analysis.

- If $\mu(b) = \mu(a)$, $Q = \mu(a)d(a,b)$.
- If $\mu(b) > \mu(a)$,

$$Q \geq \mu(a) \sum_{u\in U^<} \mathbb{P}[\xi_b = u] - \mathbb{P}[\xi_a = u] = \frac{1}{2}\mu(a)d(a,b).$$

- If $\mu(b) < \mu(a)$,

$$Q \geq \mu(a) \sum_{u\in U^>} \mathbb{P}[\xi_a = u] - \mathbb{P}[\xi_b = u] = \frac{1}{2}\mu(a)d(a,b).$$

In all cases, we have $Q \geq \frac{1}{2}\mu(a)d(a,b)$. Returning to (2.7), we find

$$\sum_{1\leq i<j\leq |A|} \sum_{u\in U} \big|\mu(i)\mathbb{P}[\xi_i = u] - \mu(j)\mathbb{P}[\xi_j = u]\big| \geq \frac{1}{2} \sum_{a\in A} \mu(a) \sum_{b\in A} d(a,b), \tag{2.8}$$

and from (2.7) and (2.8), we have

$$\begin{aligned} \min_{a\in A} r_a &\geq \frac{1}{|A|} + \frac{1}{2|A|(|A|-1)} \sum_{a\in A} \mu(a) \sum_{b\in A} d(a,b) \\ &\geq \frac{1}{|A|} + \frac{1}{2|A|(|A|-1)} \min_{a\in A} \sum_{b\in A} d(a,b). \end{aligned}$$ ∎

2.2. Composition of Invertible Functions

A natural question is whether the composition of invertible functions is also invertible. The next result shows that in general the answer is ‘no’, though we can provide a precise characterization based on the rank of an associated matrix.

Theorem 2.4. *Let $\Upsilon\colon U \to Z$ be a random function, let Y denote the associated matrix $\big($with $Y_{uz} = \mathbb{P}[\upsilon_u = z]\big)$, and let Y^+ denote the extension of Y by an all-1 row. If rank $(Y^+) = |U|$, then for any invertible random function $\Xi\colon A \to U$, the composition $\Upsilon \circ \Xi\colon A \to Z$ is invertible, and if the rank is less than $|U|$, then there exist invertible random functions $\Xi\colon A \to U$ such that $\Upsilon \circ \Xi\colon A \to Z$ is not invertible.*

Proof. First assume that $\Upsilon \circ \Xi$ is not invertible, i.e., there exist $a, b \in A$, $a \neq b$ such that the distributions υ_{ξ_a} and υ_{ξ_b} are identical. Then we consider the following homogeneous system of linear equations, where the coefficients are the numbers $\mathbb{P}[\upsilon_u = z]$ and 1's, and the variables are the x_u's

$$\sum_{u\in U} \mathbb{P}[\upsilon_u = z]\, x_u = 0, \quad \text{for all } z \in Z, \tag{2.9}$$

$$\sum_{u\in U} x_u = 0. \tag{2.10}$$

The matrix Y^+ is the matrix of the system of homogeneous linear Equations (2.9)–(2.10). Observe that $x_u = \mathbb{P}[\xi_a = u] - \mathbb{P}[\xi_b = u]$ solves the system (2.9)–(2.10). If the rank of Y^+ is $|U|$, then it has only the trivial solution, i.e., for all $u \in U$, $x_u = 0$. This amounts to ξ_a and ξ_b having the same distribution, contrary to the assumption of Ξ being invertible.

Assume now that Y^+ has rank less than $|U|$. Then the system (2.9)–(2.10) has a non-trivial solution x_u. Set $P = \sum_{u\colon x_u>0} x_u$ and $N = \sum_{u\colon x_u<0} x_u$. Clearly, $P = -N > 0$. Take $A = \{a, b\}$, $\mathbb{P}[\xi_a = u] = \frac{x_u}{P}$ if $x_u \geq 0$, and 0 otherwise; and $\mathbb{P}[\xi_b = u] = \frac{x_u}{N}$ if $x_u \leq 0$, and 0 otherwise. It is clear that this Ξ is invertible, as it separates a and b. However, the distributions υ_{ξ_a} and υ_{ξ_b} are identical, as

$$\begin{aligned}
\mathbb{P}\left[\upsilon_{\xi_a} = z\right] &= \sum_{u\in U} \mathbb{P}[\upsilon_u = z] \cdot \mathbb{P}[\xi_a = u] \\
&= \sum_{u\in U\colon x_u>0} \mathbb{P}[\upsilon_u = z] \frac{x_u}{P} \\
&= \sum_{u\in U\colon x_u<0} \mathbb{P}[\upsilon_u = z] \frac{x_u}{N} \\
&= \sum_{u\in U} \mathbb{P}[\upsilon_u = z] \cdot \mathbb{P}[\xi_b = u] \\
&= \mathbb{P}\left[\upsilon_{\xi_b} = z\right].
\end{aligned}$$

∎

3. Maximum Likelihood Estimation (MLE) in the Finite Parameter Setting with Nuisance Parameters

In this section, we reconsider the question of how many i.i.d. samples are required in order for maximum likelihood to recover elements of a finite set accurately, when additional nuisance parameters are present. Assume $B = \{(a, \theta)\colon a \in A, \theta \in \Theta(a)\}$

and that $\Xi\colon B \to U$ is a random function, where A and U are finite sets. Define

$$U^+ := \left\{u\colon \mathbb{P}\left[\xi_{(a,\theta)} = u\right] > 0\right\}, \tag{3.1}$$

$$\alpha := \alpha_{(a,\theta)} = \min_{u \in U^+} \left\{\mathbb{P}\left[\xi_{(a,\theta)} = u\right]\right\}, \tag{3.2}$$

and assume

$$d := d_{(a,\theta)} = \inf_{b \neq a,\, \theta' \in \Theta(b)} \sum_{u \in U} \left|\mathbb{P}\left[\xi_{(a,\theta)} = u\right] - \mathbb{P}\left[\xi_{(b,\theta')} = u\right]\right| > 0. \tag{3.3}$$

In our earlier work, in [15, Theorem 5], we showed that for

$$k \geq f(\alpha, d) \log\left(\frac{2\left|U^+\right|}{\varepsilon}\right), \tag{3.4}$$

k samples suffice to reconstruct $a \in A$, from (a, θ) with probability at least $1-\varepsilon$ using MLE. More formally, for $\Xi^{(k)}\colon B \to U^k$, $\left[R^{(k)}\right]'_{(a,\theta)} \geq 1-\varepsilon$. Our function f in (3.4) tends to infinity when either (or both) $\alpha \to 0$ or $d \to 0$. This dependence on d is reasonable (though not always necessary; see Subsection 3.2); however, the dependence on α is not clear and raises two questions:

Q1. Is there an bound on k (as in (3.4)) which depends only on $\left|U^+\right|$, ε, and d but not on α?

Q2. Moreover, can the function f in (3.4) be replaced by a function of just d and ε $\left(\text{and not } \alpha \text{ and } U^+\right)$ so that the resulting function is still a valid bound for k?

In this section we show that the answer to the first question is ‘yes’ (Theorem 3.3) while the answer to the second is ‘no’ (in Subsection 3.1).

We begin by introducing some further notation. For any two probability distributions p, p' on a set U, let $d_{KL}\left(p, p'\right) = \sum_{u \in U\colon p_u > 0} p_u \log\left(\frac{p_u}{p'_u}\right) \in [0, \infty) \cup \{\infty\}$ denote the Kullback-Leibler distance of p and p', and recall the standard inequality (see, for example, [4])

$$d_{KL}\left(p, p'\right) \geq \frac{1}{2} d\left(p, p'\right)^2, \tag{3.5}$$

where $d\left(p, p'\right)$ denotes, as usual, the variational distance, $\sum_{u \in U} \left|p_u - p'_u\right|$. We will also use $d_2\left(p, p'\right) = \left(\sum_{u \in U} \left|p_u - p'_u\right|^2\right)^{1/2}$.

Lemma 3.1. *Let $X_1, X_2, \ldots, X_k$ be a sequence of i.i.d. random variables taking values in a finite set U. For each $u \in U$, let $\hat{p}_u := \frac{1}{k}\sum_{i=1}^{k} \mathbb{I}\left(X_i = u\right)$ (the normalized multinomial counts) and let $p_u = \mathbb{P}\left[X_1 = u\right]$. Let $U^+ := \{u\colon p_u > 0\}$. Then,*

(i) $\mathbb{P}\left[d_{KL}(\hat{p}, p) \geq \delta\right] \leq \frac{\left|U^+\right|}{k\delta}$,

(ii) $\mathbb{P}\left[d(\hat{p}, p) \geq \delta\right] \leq \frac{\left|U^+\right|}{k\delta^2}$.

Proof. Part (i). Let $\hat{\Delta}_u = \hat{p}_u - p_u$. For $u \in U^+$, set $\hat{Q}_u = 0$ if $\hat{p}_u = 0$, while if $\hat{p}_u > 0$ set

$$\hat{Q}_u := \hat{p}_u \log\left(\frac{\hat{p}_u}{p_u}\right) = \left(p_u + \hat{\Delta}_u\right) \log\left(1 + \frac{\hat{\Delta}_u}{p_u}\right)$$

$$\leq \left(p_u + \hat{\Delta}_u\right) \cdot \frac{\hat{\Delta}_u}{p_u} = \hat{\Delta}_u + \frac{\hat{\Delta}_u^2}{p_u}. \tag{3.6}$$

Recall Markov's inequality, which states that if X is a non-negative random variable, and $a > 0$, then

$$\mathbb{P}[X \geq a] \leq \frac{\mathbb{E}[X]}{a}. \tag{3.7}$$

Note that $\mathbb{E}\left[(\hat{p}_u - p_u)^2\right] = Var[\hat{p}_u] = \frac{p_u(1-p_u)}{k}$, and applying (3.7) to

$$X = \sum_{u \in U^+} \frac{\hat{\Delta}_u^2}{p_u} \geq 0$$

and noting that $\mathbb{E}[X] = \frac{|U^+|-1}{k}$ gives $\mathbb{P}[X \geq \delta] \leq \frac{|U^+|}{k\delta}$. By definition, $d_{KL}(\hat{p}, p) = \sum_{u:\, \hat{p}_u \neq 0} \hat{Q}_u = \sum_{u \in U^+} \hat{Q}_u$ and this is less or equal to X $\big($by (3.6), and the identity $\sum_{u \in U^+} \hat{\Delta}_u = 0\big)$, which leads to the required inequality.

Part (ii). Applying the Cauchy-Schwartz inequality, $d^2(\hat{p}, p) \leq d_2^2(\hat{p}, p) \cdot |U^+|$, therefore,

$$\mathbb{P}[d(\hat{p}, p) \geq \delta] \leq \mathbb{P}\left[d_2^2(\hat{p}, p) \geq \frac{\delta^2}{|U^+|}\right] \leq \frac{|U^+|}{\delta^2} \mathbb{E}\left[d_2^2(\hat{p}, p)\right],$$

by Markov's inequality (3.7). Now,

$$\mathbb{E}\left[d_2^2(\hat{p}, p)\right] = \mathbb{E}\left[\sum_{u \in U} (\hat{p}_u - p_u)^2\right] = \sum_{u \in U} Var[\hat{p}_u] = \sum_{u \in U} \frac{1}{k} p_u (1 - p_u) \leq \frac{1}{k}. \quad \blacksquare$$

Corollary 3.2. *Under the assumptions of Lemma 3.1, suppose that* $\delta < 1$, $\varepsilon > 0$, *and* $k \geq \frac{2|U^+|}{\varepsilon\delta^2}$. *Then, with probability at least* $1 - \varepsilon$, *the inequalities* $d_{KL}(\hat{p}, p) < \delta$ *and* $d(\hat{p}, p) < \delta$ *simultaneously hold.*

Theorem 3.3. *Assume* $B = \{(a, \theta)\colon a \in A, \theta \in \Theta(a)\}$ *and that* $\Xi\colon B \to U$ *is a random function, where* A *and* U *are finite sets. Recall definition* (3.1) *and condition* (3.3). *Provided that* $k \geq \frac{c_1 |U^+|}{\varepsilon d_{(a,\theta)}^4}$ *with* $c_1 = \frac{2}{\left(2-\sqrt{3}\right)^2}$, *the probability that MLE correctly returns* a *from* $\Xi^{(k)}$ *is at least* $1 - \varepsilon$, *i.e.,* $\left[R^{(k)}\right]'_{(a,\theta)} \geq 1 - \varepsilon$.

Proof. Let p be the probability distribution on U induced by $\xi_{(a,\theta)}$, let $c = 2 - \sqrt{3}$, and let E be the event that $d(\hat{p}, p) \leq c \cdot d_{(a,\theta)}$. For the probability distribution q induced by $\xi_{(b,\theta')}$ where $b \neq a$, by the triangle inequality, we have

$$d(\hat{p}, q) \geq |d(p, q) - d(\hat{p}, p)|.$$

Now, by assumption $d(p,q) \geq d_{(a,\theta)}$, and so, conditional on E, $d(\hat{p},q) \geq (1-c)d_{(a,\theta)}$. Invoking the inequality (3.5) gives

$$d_{KL}(\hat{p},q) \geq \frac{1}{2}d(\hat{p},q)^2 \geq \frac{1}{2}(1-c)^2 d^2_{(a,\theta)}.$$

Thus, conditional on E, we have

$$\sum_{u\in U^+} \hat{p}_u \log q_u \leq \sum_{u\in U^+} \hat{p}_u \log \hat{p}_u - \frac{1}{2}(1-c)^2 d^2_{(a,\theta)}. \tag{3.8}$$

For $x\in A$, $\omega \in \Theta(x)$, consider

$$L(x,\omega) = \sum_{u\in U^+} \hat{p}(u) \log \mathbb{P}\left[\xi_{x,\omega} = u\right]. \tag{3.9}$$

$L(x,\omega)$ is $\frac{1}{k}$ times the natural logarithm of the probability of generating the observed sequence of U-elements under (x,ω). Therefore, $L(x,\omega) \leq 0$ is proportional to the log-likelihood of (x,ω). Now consider the log likelihood ratio

$$\Delta L := L(a,\theta) - L(b,\theta') = \sum_{u\in U^+} \hat{p}_u \log\left(p_u/q_u\right).$$

Conditional on E, we have, by (3.8),

$$\Delta L \geq -\sum_{u\in U^+} \hat{p}_u \log\left(\frac{\hat{p}_u}{p_u}\right) + \frac{1}{2}(1-c)^2 d^2_{(a,\theta)} = \frac{1}{2}(1-c)^2 d^2_{(a,\theta)} - d_{KL}(\hat{p},p). \tag{3.10}$$

So if we select $\delta = c\cdot d^2_{(a,\theta)}$ in Corollary 3.2, we can ensure that with probability at least $1-\varepsilon$ that event E occurs and also $\left(\text{since } \frac{1}{2}(1-c)^2 = c\right)$ that

$$d_{KL}(\hat{p},p) < \delta = c\cdot d^2_{(a,\theta)} = \frac{1}{2}(1-c)^2 d^2_{(a,\theta)}.$$

Therefore, by (3.10), we have $\Delta L > 0$. The value of k that Corollary 3.2 requires is precisely that given in the statement of this theorem. This completes the proof. ∎

Remark 3.4.

- Theorem 3.3 also implies that for MLE in the setting where nuisance parameters are absent, the number k of i.i.d. samples required to reconstruct an element $a\in A$ correctly with probability at least $1-\varepsilon$ is bounded above by a function that depends only on $|U^+|$, ε and $d_a := \min_{b\neq a} d(a,b)$. In [14], an upper bound on k was also derived; however, it depended solely on $|A|, \varepsilon$ and d_a. Comparing these results suggests an interesting question: Is there an upper bound for k (in absence of nuisance parameters) which depends just on d_a and ε?
- We show below that the linear dependence of k on $|U^+|$ in Theorem 3.3 is best possible in the sense that no sublinear dependence is possible. However, the exponent of 4 for d in Theorem 3.3 could possibly be reduced.

3.1. Construction Showing That k Must Grow Linearly with $|U^+|$

We now show that Theorem 3.3 cannot be improved by replacing the dependence of k on $|U^+|$ with a sublinear function (such as the logarithmic dependence on $|U|^+$ in [15, Theorem 5.1]), even when $d_{(a,\theta)}$ and ε are held constant.

Let $A = \{a, b\}$, with $\Theta(a) = \{*\}$, and

$$\Theta(b) = \left\{ \theta = (\lambda_1, \ldots, \lambda_n) : \sum_{i=1}^{n} \lambda_i = 1, \forall\, i,\ \lambda_i \geq 0 \right\}.$$

Let $U = \{0, 1, \ldots, n\}$. Fix $\delta > 0$ and consider the random function Ξ defined as follows.

$$\mathbb{P}\left[\xi_{(a,*)} = u\right] = \begin{cases} \delta, & \text{if } u = 0; \\ \frac{1-\delta}{n}, & \text{if } u \in \{1, \ldots, n\}; \end{cases}$$

$$\mathbb{P}\left[\xi_{(b,(\lambda_1,\ldots,\lambda_n))} = u\right] = \begin{cases} 2\delta, & \text{if } u = 0; \\ \lambda_u(1-2\delta), & \text{if } u \in \{1, \ldots, n\}. \end{cases}$$

We assume that $k \leq n$; otherwise, we have nothing to prove. For $\mathbf{u} = (u_1, \ldots, u_k) \in U^k$, let $x(\mathbf{u}) = |\{i \in \{1, \ldots, k\} \colon u_i = 0\}|$. We have

$$L_1 := \sup_{\theta \in \Theta(a)} \mathbb{P}\left[\xi^{(k)}_{(a,\theta)} = \mathbf{u}\right] = \delta^{x(\mathbf{u})} \left(\frac{1-\delta}{n}\right)^{k-x(\mathbf{u})}$$

and

$$L_2 := \sup_{\theta' \in \Theta(b)} \mathbb{P}\left[\xi^{(k)}_{(b,\theta')} = \mathbf{u}\right] \geq (2\delta)^{x(\mathbf{u})} \left(\frac{1-2\delta}{k-x(\mathbf{u})}\right)^{k-x(\mathbf{u})}, \tag{3.11}$$

since we are free to select $\theta \in \Theta(b)$ to be the uniform distribution on $\{i \colon u_i \neq 0\}$. We will select a sufficiently small value of δ so that

$$2(1-2\delta)^{\delta/2} > 1. \tag{3.12}$$

Now, suppose we generate u randomly from $(a, *)$. Note that the value of $d_{(a,*)}$ is at least δ, since

$$d((a,*),(b,\theta')) \geq \left|\mathbb{P}\left[\xi_{(a,*)} = 0\right] - \mathbb{P}\left[\xi_{(b,\theta')} = 0\right]\right| = \delta.$$

Then MLE will (incorrectly) reconstruct b whenever $R := L_2/L_1 > 1$. We will show that this occurs with probability at least $1-\varepsilon$, if k is less than $\frac{1}{2}|U^+|$, for any δ satisfying (3.12) and any sufficiently large $|U^+|$.

Note that by replacing L_2 by its lower bound (3.11), we can write $R \geq Y^k$ where

$$Y = 2^{\rho} \left[\frac{n}{k} \cdot \frac{(1-2\delta)}{(1-\delta)(1-\rho)}\right]^{1-\rho},$$

where $\rho := x(\mathbf{u})/k$. Now, if $k \leq \frac{1}{2}n$, then since $((1-\delta)(1-\rho))^{-(1-\rho)} \geq 1$,

$$Y \geq 2(1-2\delta)^{1-\rho}.$$

Now, for δ, ε fixed, a value of k exists for which we have $\rho > \frac{1}{2}\delta$ with probability at least $1-\varepsilon$. Thus for this value of k, and any $n > 2k$, inequality (3.12) gives

$$Y \geq 2(1-2\delta)^{\delta/2} > 1.$$

Consequently $R > 1$, and so MLE will make an incorrect decision. Thus, we must have $k \geq \frac{1}{2}n = \frac{1}{2}\left(|U^+|-1\right)$ in order to avoid this.

3.2. Example to Show That MLE with Nuisance Parameters Can Still Succeed When Variational Distance Vanishes on Each Element of A

In the absence of nuisance parameters and given a random function $\Xi\colon A \to U$, suppose that $d\left(a_1, a_2\right) = 0$ for two elements $a_1, a_2 \in A$. Then for *any* random function $\Gamma\colon U \to A$, it is easily shown (for example, by [15, Theorem 3.1]) that

$$\min\left\{\mathbb{P}\left[\gamma_{\xi_{a_1}} = a_1\right], \mathbb{P}\left[\gamma_{\xi_{a_2}} = a_2\right]\right\} \leq \frac{1}{2}. \tag{3.13}$$

That is, if the probability distribution induced by a_1 and a_2 is the same, no method can recover both a_1 and a_2 more accurately than by a toss of a fair coin. We can ask if a similar result holds for MLE when nuisance parameters are present. That is, suppose that $A = \{a_1, a_2\}$ and that for a value $\theta_1 \in \Theta(a_1)$, and $\theta_2 \in \Theta(a_2)$, we have

$$d_{(a_1,\theta_1)} = d_{(a_2,\theta_2)} = 0, \tag{3.14}$$

where $d_{(a,\theta)}$ is defined as in (3.3). Note that Theorem 3.3 does not give a finite bound on k for MLE to accurately reconstruct a_1 or a_2. However it turns out that for certain random functions satisfying (3.14), if nuisance parameters are present and MLE is used to estimate a_1 and a_2 from k independent trials, then for any parameter (a_i, θ_i) chosen, and for even values of k, the probability that the selection is correct is always strictly greater than $\frac{1}{2}$.

Let $A = \{a_1, a_2\}$, $U = \{(1,0), (1,1), (2,0), (2,1)\}$, $\Theta(a_1) = [\pi/4, 3\pi/4)$, and $\Theta(a_2) = (\pi/4, 3\pi/4]$. For $t \in \Theta(a_1)$, let

$$\mathbb{P}\left[\xi_{(a_1,t)} = (1, \lfloor 2t/\pi \rfloor)\right] = \sin^2 t; \quad \mathbb{P}\left[\xi_{(a_1,t)} = (2, \lfloor 2t/\pi \rfloor)\right] = \cos^2 t;$$

and for $t \in \Theta(a_2)$, let

$$\mathbb{P}\left[\xi_{(a_2,t)} = (1, \lfloor 2t/\pi \rfloor)\right] = \cos^2 t; \quad \mathbb{P}\left[\xi_{(a_2,t)} = (2, \lfloor 2t/\pi \rfloor)\right] = \sin^2 t.$$

The key observation for the argument that follows is that $\sin^2 t > \cos^2 t$ in $(\pi/4, 3\pi/4)$, while in the endpoints $\sin^2 t = 1/2 = \cos^2 t$. It is easy to see that $\lim_{t\to\frac{\pi}{4}^+} d\big((a_1, \pi/4), (a_2, t)\big) = 0$, and hence $d_{(a_1,\pi/4)} = 0$. A similar argument shows that $d_{(a_2,3\pi/4)} = 0$. It is also easy to see that the distributions of all $\xi_{(a_i,t)}$ random variables are different. The only possible problem would be the distributions of $\xi_{(a_1,\pi/4)}$ and $\xi_{(a_2,3\pi/4)}$. However, in this case, the second coordinates in the elements of U separate these distributions. There is a pedestrian way to guess where an element of U came from. Count the one's and two's in the first coordinates after k independent trials. If there are more one's, then select a_1; if there are more two's then select a_2 and in the case of a tie, if the second coordinate (which has to be constant over the trials!) $\lfloor 2t/\pi \rfloor = 0$ select a_1, otherwise select a_2. MLE pretty much does the same; the only thing that requires more careful analysis is whether MLE correctly returns $(a_1, \pi/4)$ and $(a_2, 3\pi/4)$. Let us focus on $(a_1, \pi/4)$, as the other problem is analogous. Let #1 and #2 denote the number of one's and two's in the first coordinates in $\xi^{(k)}_{(a_1,\pi/4)}$, respectively. Let p be the probability of the event $X_1 =$ "# 1 > # 2"; by symmetry, it is also the probability of the event $X_2 =$ "# 1 < # 2", and let q be the probability of the event $X_3 =$ "# 1 = # 2". Note that MLE correctly returns a_1 for events X_1 and X_3 (but not for X_2), and hence $\left[R^{(k)}\right]'_{(a_1,\pi/4)} \geq p+q = \frac{1+q}{2} > \frac{1}{2}$. The claim holds for X_3 for the following reason. For any k-sequence that satisfies event X_3, the probability that $\xi^{(k)}_{(a_1,\pi/4)}$ generates this sequence is 2^{-k}, while the probability that (a_2, θ_2) generates this sequence is $\lambda^{k/2}(1-\lambda)^{k/2}$ for some $\lambda \neq 1/2$, and the second probability is strictly smaller than 2^{-k}.

Informally, the reason for this phenomenon is that the parameter space associated to a_i is tuned for 'fitting' data that are produced by the pair (a_i, θ_i).

Notice that in all but one choice of the nuisance parameter settings (associated to a_1), the probability that the selection is correct tends to 1 as $k \to \infty$ (in the other setting it tends to $\frac{1}{2}$ from above). For the pedestrian approach for estimating a_1 or a_2 from the k independent trials (described above) the probability of making the correct reconstruction tends to 1 as k tends to infinity for all parameter settings (in contrast to MLE which has problems at one particular parameter setting — this illustrates again the care required in consistency arguments for MLE).

Notice also that, in this example, with any parameters θ_1, θ_2, we have the strict inequality $d\left((a_1, \theta_1), (a_2, \theta_2)\right) > 0$.

Despite this somewhat surprising result, one can easily derive an analogue of (3.13) for any random function $\Xi\colon B \to U$ (where $B = \{(a, \theta)\colon \theta \in \Theta(a)\}$ as usual) under the stronger condition that there exists $(a_1, \theta_1), (a_2, \theta_2)$ such that $d\left((a_1, \theta_1),\right.$ $\left.(a_2, \theta_2)\right) = 0$. In this case, for any random function (not just MLE) $\Gamma \to U$ that is independent of Ξ it is easily shown that

$$\min\left\{\mathbb{P}\left[\gamma_{\xi_{(a_1,\theta_1)}} = a_1\right], \mathbb{P}\left[\gamma_{\xi_{(a_2,\theta_2)}} = a_2\right]\right\} \leq \frac{1}{2}.$$

Of course this bound applies also for k i.i.d. trial experiments.

3.3. Application of Theorem 3.3

As a simple illustration of the use of Theorem 3.3, we describe an application to the reconstruction of phylogenetic trees from binary sequences according to a simple Markov process (the CFN model). Such processes are central to much of molecular biology (see, e.g., [8]). Let A denote the three binary phylogenetic trees that have a leaf set $X = \{1, 2, 3, 4\}$. For a tree $T = (V_T, E_T) \in A$, $\Theta(a)$ is a function $p = p_T\colon E_T \to [0, 0.5]$ which assigns to each edge e of T an associated *substitution probability*. Under the CFN model a state is assigned uniformly at random to a leaf (e.g., leaf 1) and states are assigned recursively to the remaining vertices of the tree by (independently) changing the state (0 to 1 or 1 to 0) across each edge e of T with probability $p(e)$. This gives a (marginal) probability distribution on each of the 16 site patterns $c\colon X \to \{0, 1\}$ (further details concerning this model can be found in [15] or [12]). Thus if we generate k site patterns i.i.d. from the pair (T, p), we can ask how large k must be in order for MLE to reconstruct T accurately. To ensure that $d_{(T,p)} > 0$, one must impose the following condition on p:

(P) For each of the four edges e of T incident with a leaf we have $p(e) \leq g < \frac{1}{2}$; and for the central edge e of T, $p(e) \geq f > 0$.

From [16, Lemma 6.3], we have $d_{(T,p)} \geq H(f, g) > 0$ for a continuous function H. Note that condition (P) can allow arbitrarily small values for $\alpha_{(T,p)} : = \min_{u \in U^+} \{\mathbb{P}\left[\xi_{(T,p)} = u\right]\}$ even when f and g take fixed values (since condition (P) allows two adjacent edges incident with leaves of T to both have arbitrarily small $p(e)$ values, and the probability of any site pattern that assigns these two leaves different states can therefore be made as close to zero as we wish). Consequently, the main result from [15] does not provide any (finite) estimate for the site patterns required for MLE to reconstruct a tree correctly. However, we may apply Theorem 3.3 in this setting. Since $|U^+| \leq 16$, we obtain an explicit upper bound on the number of site patterns required to reconstruct each phylogenetic tree on four leaves correctly with probability at least $1 - \varepsilon$.

Of course, the number k of site patterns required for MLE to accurately reconstruct each binary tree T on four leaves under condition (P) also depends on f and g. These quantities enter into the upper bound on k in Theorem 3.3 via the term $d_{(a,\theta)}$. As f tends to 0 or g tends to $\frac{1}{2}$, $d_{(a,\theta)}$ converges to 0, and so the resulting upper bound on k goes to infinity. Indeed Theorem 3.3 requires k to grow at the rate $1/f^4$ as $f \to 0$. However, we suspect that for MLE in the phylogenetic setting this can be improved to $1/f^2$, as this rate can be achieved by other tree reconstruction methods [5], and this rate is best possible ([15, Theorem 4.1]).

Acknowledgments. We would like to thank Linyuan (Lincoln) Lu, for suggesting a shorter proof of the implication (i) ⇒ (iii) of Proposition 2.1. We would also like to thank the two anonymous referees for numerous helpful suggestions.

References

1. Allman, E.S., Ané, C., Rhodes, J.A.: Identifiability of a Markovian model of molecular evolution with gamma-distributed rates. Adv. in Appl. Probab. 40, 229–249 (2008)
2. Casella, G., Berger, R.L.: Statistical Inference. Duxbury Press, Belmont (1990)
3. Chang, J.T.: Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. Math. Biosci. 137, 51–73 (1996)
4. Cover, T.M., Thomas, J.A.: Elements of Information Theory. John Wiley & Sons, Inc., New York (1991)
5. Erdös, P.L., Steel, M.A., Székely, L.A., Warnow, T.: A few logs suffice to build (almost) all trees (Part 1). Random Structures Algorithms 14, 153–184 (1999)
6. Everitt, B.S.: The Cambridge Dictionary of Statistics. Cambridge University Press, Cambridge, UK (1998)
7. Farris, J.S.: Likelihood and inconsistency. Cladistics 15, 199–204 (1999)

8. Felsenstein, J.: Inferring Phylogenies. Sinauer Associates Inc., Sunderland, MA. (2004)
9. Rogers, J.S.: On the consistency of maximum likelihood estimation of phylogenetic trees from nucleotide sequences. Syst. Biol. 46, 354–357 (1997)
10. Rogers, J.S.: Maximum likelihood estimation of phylogenetic trees is consistent when substitution rates vary according to the invariable sites plus gamma distribution. Syst. Biol. 50, 713–722 (2001)
11. Schrijver, A.: Theory of Linear and Integer Programming. John Wiley & Sons Ltd., Chichester (1986)
12. Semple, C., Steel, M.: Phylogenetics. Oxford University Press, Oxford (2003)
13. Siddall, M.E.: Success of parsimony in the four-taxon case: long-branch repulsion by likelihood in the Farris zone. Cladistics 14, 209–220 (1998)
14. Steel, M.A., Székely, L.A.: Inverting random functions. Ann. Combin. 3, 103–113 (1999)
15. Steel, M.A., Székely, L.A.: Inverting random functions II: explicit bounds for the discrete maximum likelihood estimation, with applications. SIAM J. Discrete Math. 15, 562–575 (2002)
16. Steel, M.A., Székely, L.A.: Teasing apart two trees. Combin. Probab. Comput. 16, 903–922 (2007)
17. Wald, A.: Note on the consistency of the maximum likelihood estimate. Ann. Math. Statist. 20, 595–601 (1949)
18. Yang, Z.: Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. Syst. Biol. 43, 329–342 (1994)
19. Yang, Z.: Phylogenetic analysis using parsimony and likelihood methods. J. Mol. Evol. 42, 294–307 (1996)