Evolutionary Trees can be Learned in Polynomial Time in the Two-State General Markov Model*

Mary Cryan   Leslie Ann Goldberg   Paul W. Goldberg.

July 20, 1998

Abstract

The j-State General Markov Model of evolution (due to Steel) is a stochastic model concerned with the evolution of strings over an alphabet of size j. In particular, the Two-State General Markov Model of evolution generalises the well-known Cavender-Farris-Neyman model of evolution by removing the *symmetry* restriction (which requires that the probability that a '0' turns into a '1' along an edge is the same as the probability that a '1' turns into a '0' along the edge). Farach and Kannan showed how to PAC-learn Markov Evolutionary Trees in the Cavender-Farris-Neyman model provided that the target tree satisfies the additional restriction that all pairs of leaves have a sufficiently high probability of being the same. We show how to remove both restrictions and thereby obtain the first polynomial-time PAC-learning algorithm (in the sense of Kearns et al.) for the general class of Two-State Markov Evolutionary Trees.

*Research Report RR347, Department of Computer Science, University of Warwick, Coventry CV4 7AL, United Kingdom. A preliminary version of this paper appears in the proceedings of FOCS '98. This work was partially supported by the ESPRIT Projects ALCOM-IT (Project 20244) and RAND-II (Project 21726) and by EPSRC grant GR/L60982. The affiliation of the authors is: Department of Computer Science, University of Warwick, Coventry CV4 7AL, United Kingdom, {maryc,leslie,pwg}@dcs.warwick.ac.uk

1 Introduction

The j-State General Markov Model of Evolution was proposed by Steel in 1994 [14]. The model is concerned with the evolution of strings (such as DNA strings) over an alphabet of size j. The model can be described as follows. A *j-State Markov Evolutionary Tree* consists of a *topology* (a rooted tree, with edges directed away from the root), together with the following parameters. The root of the tree is associated with j probabilities $\rho_0, \ldots, \rho_{j-1}$ which sum to 1, and each edge of the tree is associated with a stochastic transition matrix whose state space is the alphabet. A probabilistic experiment can be performed using the Markov Evolutionary Tree as follows: The root is assigned a letter from the alphabet according to the probabilities $\rho_0, \ldots, \rho_{j-1}$. (Letter i is chosen with probability ρ_i.) Then the letter propagates down the edges of the tree. As the letter passes through each edge, it undergoes a probabilistic transition according to the transition matrix associated with the edge. The result is a string of length n which is the concatenation of the letters obtained at the n leaves of the tree. A j-State Markov Evolutionary Tree thus defines a probability distribution on length-n strings over an alphabet of size j. (The probabilistic experiment described above produces a single sample from the distribution.[1])

To avoid getting bogged down in detail, we work with a binary alphabet. Thus, we will consider *Two-State* Markov Evolutionary Trees.

Following Farach and Kannan [9], Erdös, Steel, Székely and Warnow [7, 8] and Ambainis, Desper, Farach and Kannan [2], we are interested in the problem of learning a Markov Evolutionary Tree, given samples from its output distribution. Following Farach and Kannan and Ambainis et al., we consider the problem of using polynomially many samples from a Markov Evolutionary Tree M to "learn" a Markov Evolutionary Tree M' whose distribution is close to that of M. We use the *variation distance* metric to measure the distance between two distributions, $\mathcal{D}$ and $\mathcal{D}'$, on strings of length n. The variation distance between $\mathcal{D}$ and $\mathcal{D}'$ is $\sum_{s\in\{0,1\}^n} |\mathcal{D}(s) - \mathcal{D}'(s)|$. If M and M' are n-leaf Markov Evolutionary Trees, we use the notation $\mathrm{var}(M, M')$ to denote the variation distance between the distribution of M and the distribution of M'.

We use the "Probably Approximately Correct" (PAC) distribution learning model of Kearns, Mansour, Ron, Rubinfeld, Schapire and Sellie [11]. Our main result is the first polynomial-time PAC-learning algorithm for the class of Two-State Markov Evolutionary Trees (which we will refer to as METs):

Theorem 1 *Let δ and ϵ be any positive constants. If our algorithm is given* $\mathrm{poly}(n, 1/\epsilon, 1/\delta)$ *samples from any MET M with any n-leaf topology T, then with probability at least $1-\delta$, the MET M' constructed by the algorithm satisfies* $\mathrm{var}(M, M') \leq \epsilon$.

Interesting PAC-learning algorithms for biologically important restricted classes of METs have been given by Farach and Kannan in [9] and by Ambainis, Desper, Farach and Kannan in [2]. These algorithms (and their relation to our algorithm) will be discussed more fully in Section 1.1. At this point, we simply note that these algorithms only apply to METs which satisfy the following restrictions.

> **Restriction 1:** All transition matrices are symmetric (the probability of a '1' turning into a '0' along an edge is the same as the probability of a '0' turning into a '1'.)

[1]Biologists would view the n leaves as being existing species, and the internal nodes as being hypothetical ancestral species. Under the model, a single experiment as described above would produce a single bit position of (for example) DNA for all of the n species.

Restriction 2: For some positive constant α, every pair of leaves (x, y) satisfies $\Pr(x \neq y) \leq 1/2 - \alpha$.

We will explain in Section 1.1 why the restrictions significantly simplify the problem of learning Markov Evolutionary Trees (though they certainly do not make it easy!) The main contribution of our paper is to remove the restrictions.

While we have used variation distance (L_1 distance) to measure the distance between the target distribution $\mathcal{D}$ and our hypothesis distribution $\mathcal{D}'$, Kearns et al. formulated the problem of learning probability distributions in terms of the Kullback-Leibler divergence distance from the target distribution to the hypothesis distribution. This distance is defined to be the sum over all length-n strings s of $\mathcal{D}(s)\log(\mathcal{D}(s)/\mathcal{D}'(s))$. Kearns et al. point out that the KL distance gives an upper bound on variation distance, in the sense that the KL distance from $\mathcal{D}$ to $\mathcal{D}'$ is $\Omega(\mathrm{var}(\mathcal{D}, \mathcal{D}')^2)$. Hence if a class of distributions can be PAC-learned using KL distance, it can be PAC-learned using variation distance. We justify our use of the variation distance metric by showing that the reverse is true. In particular, we prove the following lemma in the Appendix.

Lemma 2 *A class of probability distributions over the domain $\{0,1\}^n$ that is PAC-learnable under the variation distance metric is PAC-learnable under the KL-distance measure.*

The lemma is proved using a method related to the ϵ-Bayesian shift of Abe and Warmuth [3]. Note that the result requires a discrete domain of support for the target distribution, such as the domain $\{0,1\}^n$ which we use here.

The rest of this section is organised as follows: Subsection 1.1 discusses previous work related to the General Markov Model of Evolution, and the relationship between this work and our work. Subsection 1.2 gives a brief synopsis of our algorithm for PAC-learning Markov Evolutionary Trees. Subsection 1.3 discusses an interesting connection between the problem of learning Markov Evolutionary Trees and the problem of learning mixtures of Hamming balls, which was studied by Kearns et al. [11].

1.1 Previous Work and Its Relation to Our Work

The Two-State General Markov Model [14] which we study in this paper is a generalisation of the Cavender-Farris-Neyman Model of Evolution [5, 10, 13]. Before describing the Cavender-Farris-Neyman Model, let us return to the Two-State General Markov Model. We will fix attention on the particular two-state alphabet $\{0,1\}$. Thus, the stochastic transition matrix associated with edge e is simply the matrix

$$\begin{pmatrix} 1-e_0 & e_0 \\ e_1 & 1-e_1 \end{pmatrix},$$

where e_0 denotes the probability that a '0' turns into a '1' along edge e and e_1 denotes the probability that a '1' turns into a '0' along edge e. The Cavender-Farris-Neyman Model is simply the special case of the Two-State General Markov Model in which the transition matrices are required to be symmetric. That is, it is the special case of the Two-State General Markov Model in which Restriction 1 (from page 1) holds (so $e_0 = e_1$ for every edge e).

We now describe past work on learning Markov Evolutionary Trees in the General Markov Model and in the Cavender-Farris-Neyman Model. Throughout the paper, we will define the *weight* $w(e)$ of an edge e to be $|1 - e_0 - e_1|$.

Steel [14] showed that if a j-State Markov Evolutionary Tree M satisfies (i) $\rho_i > 0$ for all i, and (ii) the determinant of every transition matrix is outside of $\{-1, 0, 1\}$, then the distribution of M uniquely determines its topology. In this case, he showed how to recover the topology, given the joint distribution of every pair of leaves. In the 2-state case, it suffices to know the exact value of the covariances of every pair of leaves. In this case, he defined the weight $\Lambda(e)$ of an edge e from node v to node w to be

$$\Lambda(e) = \begin{cases} w(e)\sqrt{\Pr(v=0)\Pr(v=1)}, & \text{if } w \text{ is a leaf, and} \\ w(e)\sqrt{\frac{\Pr(v=0)\Pr(v=1)}{\Pr(w=0)\Pr(w=1)}}, & \text{otherwise.} \end{cases} \tag{1}$$

Steel observed that these distances are multiplicative along a path and that the distance between two *leaves* is equal to their covariance. Since the distances are multiplicative along a path, their logarithms are additive. Therefore, methods for constructing trees from additive distances such as the method of Bandelt and Dress [4] can be used to reconstruct the topology. Steel's method does not show how to recover the *parameters* of a Markov Evolutionary Tree, even when the exact distribution is known and $j = 2$. In particular, the quantity that he obtains for each edge e is a one-dimensional distance rather than a *two*-dimensional vector giving the two transition probabilities e_0 and e_1. Our method shows how to recover the parameters exactly, given the exact distribution, and how to recover the parameters approximately (well enough to approximate the distribution), given polynomially-many samples from M.

Farach and Kannan [9] and Ambainis, Desper, Farach and Kannan [2] worked primarily in the special case of the Two-State General Markov Model satisfying the two restrictions on Page 1. Farach and Kannan's paper was a breakthrough, because prior to their paper nothing was known about the feasibility of reconstructing Markov Evolutionary Trees from samples. For any given positive constant α, they showed how to PAC-learn the class of METs which satisfy the two restrictions. However, the number of samples required is a function of $1/\alpha$, which is taken to be a constant. Ambainis et al. improved the bounds given by Farach and Kannan to achieve asymptotically tight upper and lower bounds on the number of samples needed to achieve a given variation distance. These results are elegant and important. Nevertheless, the restrictions that they place on the model do significantly simplify the problem of learning Markov Evolutionary Trees. In order to explain why this is true, we explain the approach of Farach et al.: Their algorithm uses samples from a MET M, which satisfies the restrictions above, to estimate the "distance" between any two leaves. (The distance is related to the covariance between the leaves.) The authors then relate the distance between two leaves to the amount of evolutionary time that elapses between them. The distances are thus turned into times. Then the algorithm of [1] is used to approximate the inter-leaf evolutionary times with times which are close, but form an additive metric, which can be fitted onto a tree. Finally, the times are turned back into transition probabilities. The symmetry assumption is essential to this approach because it is *symmetry* that relates a one-dimensional quantity (evolutionary time) to an otherwise two-dimensional quantity (the probability of going from a '0' to a '1' and the probability of going from a '1' to a '0'). The second restriction is also essential: If the probability that x differs from y were allowed to approach $1/2$, then the evolutionary time from x to y would tend to ∞. This would mean that in order to approximate the inter-leaf times accurately, the algorithm would have to get the distance estimates *very* accurately, which would require many samples. Ambainis et al. [2] generalised their results to a symmetric version of the j-state evolutionary model, subject to the two restrictions above.

Erdös, Steel, Székely and Warnow [7, 8] also considered the reconstruction of Markov Evolutionary Trees from samples. Like Steel [14] and unlike our paper or the papers of Farach et al. [9, 2], Erdös et al. were only interested in reconstructing the *topology* of a MET (rather than its parameters or distribution), and they were interested in using as few samples as possible to reconstruct the topology. They showed how to reconstruct topologies in the j-state General Markov Model when the Markov Evolutionary Trees satisfy (i) Every root probability is bounded above 0, (ii) every transition probability is bounded above 0 and below $1/2$, and (iii) for positive quantities λ and λ', the determinant of the transition matrix along each edge is between λ and $1-\lambda'$. The number of samples required is polynomial in the worst case, but is only polylogarithmic in certain cases including the case in which the MET is drawn uniformly at random from one of several (specified) natural distributions. Note that restriction (iii) of Erdös et al. is weaker than Farach and Kannan's Restriction 2 (from Page 1). However, Erdös et al. only show how to reconstruct the topology (thus they work in a restricted case in which the topology can be uniquely constructed using samples). They do not show how to reconstruct the parameters of the Markov Evolutionary Tree, or how to approximate its distribution.

1.2 A Synopsis of our Method

In this paper, we describe the first polynomial-time PAC-learning algorithm for the class of Two-State Markov Evolutionary Trees (METs). Our algorithm works as follows: First, using samples from the target MET, the algorithm estimates all of the pairwise covariances between leaves of the MET. Second, using the covariances, the leaves of the MET are partitioned into "related sets" of leaves. Essentially, leaves in different related sets have such small covariances between them that it is not always possible to use polynomially many samples to discover how the related sets are connected in the target topology. Nevertheless, we show that we can closely approximate the *distribution* of the target MET by approximating the distribution of each related set closely, and then joining the related sets by "cut edges". The first step, for each related set, is to discover an approximation to the correct topology. Since we do not restrict the class of METs which we consider, we cannot guarantee to construct the *exact* induced topology (in the target MET). Nevertheless we guarantee to construct a good enough approximation. The topology is constructed by looking at triples of leaves. We show how to ensure that each triple that we consider has large inter-leaf covariances. We derive quadratic equations which allow us to approximately recover the parameters of the triple, using estimates of inter-leaf covariances and estimates of probabilities of particular outputs. We compare the outcomes for different triples and use the comparisons to construct the topology. Once we have the topology, we again use our quadratic equations to discover the parameters of the tree. As we show in Section 2.4, we are able to prevent the error in our estimates from accumulating, so we are able to guarantee that each estimated parameter is within a small additive error of the "real" parameter in a (normalised) target MET. From this, we can show that the variation distance between our hypothesis and the target is small.

1.3 Markov Evolutionary Trees and Mixtures of Hamming Balls

A *Hamming ball distribution* [11] over binary strings of length n is defined by a *center* (a string c of length n) and a *corruption probability* p. To generate an output from the distribution, one starts with the center, and then flips each bit (or not) according to an independent Bernoulli experiment with probability p. A *linear mixture* of j Hamming balls is a distribution defined by j Hamming ball distributions, together with j probabilities $\rho_1, \ldots, \rho_j$ which sum to 1 and determine from which Hamming ball distribution a particular sample should be taken. For any fixed j, Kearns et al. give a polynomial-time PAC-learning algorithm for a mixture of j Hamming balls, *provided all j Hamming balls have the same corruption probability*[2].

A *pure* distribution over binary strings of length n is defined by n probabilities, $\lambda_1, \ldots, \lambda_n$. To generate an output from the distribution, the i'th bit is set to '0' independently with probability λ_i, and to '1' otherwise. A pure distribution is a natural generalisation of a Hamming ball distribution. Clearly, every linear mixture of j pure distributions can be realized by a j-state MET with a star-shaped topology. Thus, the algorithm given in this paper shows how to learn a linear mixture of any two pure distributions. Furthermore, a generalisation of our result to a j-ary alphabet would show how to learn any linear mixture of any j pure distributions.

2 The Algorithm

Our description of our PAC-learning algorithm and its analysis require the following definitions. For positive constants δ and ϵ, the input to the algorithm consists of poly$(n, 1/\epsilon, 1/\delta)$ samples from a MET M with an n-leaf topology T. We will let $\epsilon_1 = \epsilon/(20n^2)$, $\epsilon_2 = \epsilon_1/(4n^3)$, $\epsilon_3 = \epsilon_2^4/2^6$, $\epsilon_4 = \epsilon_1/(4n)$, $\epsilon_5 = \epsilon_2\epsilon_4/2^{10}$, and $\epsilon_6 = \epsilon_5\epsilon_2^3/2^7$. We have made no effort to optimise these constants. However, we state them explicitly so that the reader can verify below that the constants can be defined consistently. We define an ϵ_4*-contraction* of a MET with topology T' to be a tree formed from T' by contracting some internal edges e for which $\Lambda(e) > 1 - \epsilon_4$, where $\Lambda(e)$ is the edge-distance of e as defined by Steel [14] (see equation 1). If x and y are leaves of the topology T then we use the notation $\mathrm{cov}(x, y)$ to denote the covariance of the indicator variables for the events "the bit at x is 1" and "the bit at y is 1". Thus,

$$\mathrm{cov}(x, y) = \Pr(xy = 11) - \Pr(x = 1)\Pr(y = 1). \tag{2}$$

We will use the following observations.

Observation 3 *If MET M' has topology T' and e is an internal edge of T' from the root r to node v and T'' is a topology that is the same as T' except that v is the root (so e goes from v to r) then we can construct a MET with topology T'' which has the same distribution as M'. To do this, we simply set* $\Pr(v = 1)$ *appropriately (from the distribution of M'). If* $\Pr(v = 1) = 0$ *we set e_0 to be* $\Pr(r = 1)$ *(from the distribution of M'). If* $\Pr(v = 1) = 1$ *we set e_1 to be* $\Pr(r = 0)$ *(from the distribution of M'). Otherwise, we set* $e_0 = \Pr(r = 1)(\textit{old } e_1)/\Pr(v = 0)$ *and* $e_1 = \Pr(r = 0)(\textit{old } e_0)/\Pr(v = 1)$.

Observation 4 *If MET M' has topology T' and v is a degree-2 node in T' with edge e leading into v and edge f leading out of v and T'' is a topology which is the same as T' except that e*

[2] The kind of PAC-learning that we consider in this paper is *generation*. Kearns et al. also show how to do *evaluation* for the special case of the mixture of j Hamming balls described above. Using the observation that the output distributions of the subtrees below a node of a MET are independent, provided the bit at that node is fixed, we can also solve the evaluation problem for METs. In particular, we can calculate (in polynomial time) the probability that a given string is output by the hypothesis MET.

and f have been contracted to form edge g then there is a MET with topology T'' which has the same distribution as M'. To construct it, we simply set $g_0 = e_0(1-f_1)+(1-e_0)f_0$ and $g_1 = e_1(1-f_0)+(1-e_1)f_1$.

Observation 5 *If MET M' has topology T' then there is a MET M'' with topology T' which has the same distribution on its leaves as M' and has every internal edge e satisfy $e_0+e_1 \leq 1$.*

Proof of Observation 5: We will say that an edge e is "good" if $e_0+e_1 \leq 1$. Starting from the root we can make all edges along a path to a leaf good, except perhaps the last edge in the path. If edge e from u to v is the first non-good edge in the path we simply set e_0 to $1-(\text{old } e_0)$ and e_1 to $1-(\text{old } e_1)$. This makes the edge good but it has the side effect of interchanging the meaning of "0" and "1" at node v. As long as we interchange "0" and "1" an even number of times along every path we will preserve the distribution at the leaves. Thus, we can make all edges good except possibly the last one, which we use to get the parity of the number of interchanges correct. □

We will now describe the algorithm. In subsection 2.6, we will prove that with probability at least $1-\delta$, the MET M' that it constructs satisfies $\mathrm{var}(M,M') \leq \epsilon$. Thus, we will prove Theorem 1.

2.1 Step 1: Estimate the covariances of pairs of leaves

For each pair (x,y) of leaves, obtain an "observed" covariance $\widehat{\mathrm{cov}}(x,y)$ such that, with probability at least $1-\delta/3$, all observed covariances satisfy

$$\widehat{\mathrm{cov}}(x,y) \in [\mathrm{cov}(x,y)-\epsilon_3, \mathrm{cov}(x,y)+\epsilon_3].$$

Lemma 6 *Step 1 requires only* $\mathrm{poly}(n,1/\epsilon,1/\delta)$ *samples from M.*

Proof: Consider leaves x and y and let p denote $\Pr(xy=11)$. By a Chernoff bound (see [12]), after k samples the observed proportion of outputs with $xy=11$ is within $\pm\epsilon_3/4$ of p, with probability at least $1-2\exp(-k\epsilon_3^2/2^3)$. For each pair (x,y) of leaves, we estimate $\Pr(xy=11)$, $\Pr(x=1)$ and $\Pr(y=1)$ within $\pm\epsilon_3/4$. From these estimates, we can calculate $\widehat{\mathrm{cov}}(x,y)$ within $\pm\epsilon_3$ using Equation 2. □

2.2 Step 2: Partition the leaves of M into related sets

Consider the following *leaf connectivity graph* whose nodes are the leaves of M. Nodes x and y are connected by a "positive" edge if $\widehat{\mathrm{cov}}(x,y) \geq (3/4)\epsilon_2$ and are connected by a "negative" edge if $\widehat{\mathrm{cov}}(x,y) \leq -(3/4)\epsilon_2$. Each connected component in this graph (ignoring the signs of edges) forms a set of "related" leaves. For each set S of related leaves, let $s(S)$ denote the leaf in S with smallest index. METs have the property that for leaves x, y and z, $\mathrm{cov}(y,z)$ is positive iff $\mathrm{cov}(x,y)$ and $\mathrm{cov}(y,z)$ have the same sign. (To see this, use the following equation, which can be proved by algebraic manipulation from Equation 2.)

$$\mathrm{cov}(x,y) = \Pr(v=1)\Pr(v=0)(1-\alpha_0-\alpha_1)(1-\beta_0-\beta_1), \tag{3}$$

where v is taken to be the least common ancestor of x and y and α_0 and α_1 are the transition probabilities along the path from v to x and β_0 and β_1 are the transition probabilities along the path from v to y. Therefore, as long as the observed covariances are as accurate as stated in Step 1, the signs on the edges of the leaf connectivity graph partition the leaves of S into two sets S_1 and S_2 in such a way that $s(S) \in S_1$, all covariances between pairs of leaves in S_1 are positive, all covariances between pairs of leaves in S_2 are positive, and all covariances between a leaf in S_1 and a leaf in S_2 are negative.

For each set S of related leaves, let $T(S)$ denote the subtree formed from T by deleting all leaves which are not in S, contracting all degree-2 nodes, and then rooting at the neighbour of $s(S)$. Let $M(S)$ be a MET with topology $T(S)$ which has the same distribution as M on its leaves and satisfies the following.

- Every internal edge e of $M(S)$ has $e_0 + e_1 \leq 1$. (4)
- Every edge e to a node in S_1 has $e_0 + e_1 \leq 1$.
- Every edge e to a node in S_2 has $e_0 + e_1 \geq 1$.

Observations 3, 4 and 5 guarantee that $M(S)$ exists.

Observation 7 *As long as the observed covariances are as accurate as stated in Step 1 (which happens with probability at least* $1-\delta/3$*), then for any related set* S *and any leaf* $x \in S$ *there is a leaf* $y \in S$ *such that* $|\mathrm{cov}(x, y)| \geq \epsilon_2/2$.

Observation 8 *As long as the observed covariances are as accurate as stated in Step 1 (which happens with probability at least* $1-\delta/3$*), then for any related set* S *and any edge* e *of* $T(S)$ *there are leaves* a *and* b *which are connected through* e *and have* $|\mathrm{cov}(a, b)| \geq \epsilon_2/2$.

Observation 9 *As long as the observed covariances are as accurate as stated in Step 1 (which happens with probability at least* $1-\delta/3$*), then for any related set* S*, every internal node* v *of* $M(S)$ *has* $\Pr(v = 0) \in [\epsilon_2/2, 1 - \epsilon_2/2]$.

Proof of Observation 9: Suppose to the contrary that v is an internal node of $M(S)$ with $\Pr(v = 0) \in [0, \epsilon_2/2) \cup (1 - \epsilon_2/2, 1]$. Using Observation 3, we can re-root $M(S)$ at v without changing the distribution. Let w be a child of v. By equation 3, every pair of leaves a and b which are connected through (v, w) satisfy $|\mathrm{cov}(a, b)| \leq \Pr(v = 0)\Pr(v = 1) < \epsilon_2/2$. The observation now follows from Observation 8. □

Observation 10 *As long as the observed covariances are as accurate as stated in Step 1 (which happens with probability at least* $1-\delta/3$*), then for any related set* S*, every edge* e *of* $M(S)$ *has* $w(e) \geq \epsilon_2/2$.

Proof of Observation 10: This follows from Observation 8 using Equation 3. (Recall that $w(e) = |1 - e_0 - e_1|$.) □

2.3 Step 3: For each related set S, find an ϵ_4-contraction $T'(S)$ of $T(S)$.

In this section, we will assume that the observed covariances are as accurate as stated in Step 1 (this happens with probability at least $1 - \delta/3$). Let S be a related set. With probability at least $1 - \delta/(3n)$ we will find an ϵ_4-contraction $T'(S)$ of $T(S)$. Since there are at most n

related sets, all ϵ_4-contractions will be constructed with probability at least $1-\delta/3$. Recall that an ϵ_4-contraction of $M(S)$ is a tree formed from $T(S)$ by contracting some internal edges e for which $\Lambda(e) > 1-\epsilon_4$. We start with the following observation, which will allow us to redirect edges for convenience.

Observation 11 *If e is an internal edge of $T(S)$ then $\Lambda(e)$ remains unchanged if e is redirected as in Observation 3.*

Proof: The observation can be proved by algebraic manipulation from Equation 1 and Observation 3. Note (from Observation 9) that every endpoint v of e satisfies $\Pr(v=0) \in (0,1)$. Thus, the redirection in Observation 3 is not degenerate and $\Lambda(e)$ is defined. $\square$

We now describe the algorithm for constructing an ϵ_4-contraction $T'(S)$ of $T(S)$. We will build up $T'(S)$ inductively, adding leaves from S one by one. That is, when we have an ϵ_4-contraction $T'(S')$ of a subset S' of S, we will consider a leaf $x \in S - S'$ and build an ϵ_4-contraction $T'(S' \cup \{x\})$ of $T(S' \cup \{x\})$. Initially, $S' = \emptyset$. The precise order in which the leaves are added does not matter, but we will not add a new leaf x unless S' contains a leaf y such that $|\widehat{\mathrm{cov}}(x,y)| \geq (3/4)\epsilon_2$. When we add a new leaf x we will proceed as follows. First, we will consider $T'(S')$, and for every edge $e' = (u', v')$ of $T'(S')$, we will use the method in the following section (Section 2.3.1) to estimate $\Lambda(e')$. More specifically, we will let u and v be nodes which are adjacent in $T(S')$ and have $u \in u'$ and $v \in v'$ in the ϵ_4-contraction $T'(S')$. We will show how to estimate $\Lambda(e)$. Afterwards (in Section 2.3.2), we will show how to insert x.

2.3.1 Estimating $\Lambda(e)$

In this section, we suppose that we have a MET $M(S')$ on a set S' of leaves, all of which form a single related set. $T(S')$ is the topology of $M(S')$ and $T'(S')$ is an ϵ_4-contraction of $T(S')$. The edge $e' = (u', v')$ is an edge of $T'(S')$. $e = (u,v)$ is the edge of $T(S')$ for which $u \in u'$ and $v \in v'$. We wish to estimate $\Lambda(e)$ within $\pm\epsilon_4/16$. We will ensure that the overall probability that the estimates are not in this range is at most $\delta/(6n)$.

The proof of the following equations is straightforward. We will typically apply them in situations in which z is the error of an approximation.

$$\frac{x+z}{y-z} = \frac{x}{y} + \left(\frac{z}{y-z}\right)\left(1+\frac{x}{y}\right) \tag{5}$$

$$\frac{1+z}{1-z} \leq 1+4z \quad \text{if } z \leq 1/2 \tag{6}$$

$$\frac{1-z}{1+z} \geq 1-2z \quad \text{if } z \geq 0 \tag{7}$$

Case 1: e' is an internal edge

We first estimate e_0, e_1, $\Pr(u=0)$, and $\Pr(v=0)$ within $\pm\epsilon_5$ of the correct values. By Observation 9, $\Pr(u=0)$ and $\Pr(v=0)$ are in $[\epsilon_2/2, 1-\epsilon_2/2]$. Thus, our estimate of $\Pr(u=0)$ is within a factor of $(1 \pm 2\epsilon_5/\epsilon_2) = (1 \pm \epsilon_4 2^{-9})$ of the correct value. Similarly, our estimates of $\Pr(u=1)$, $\Pr(v=0)$ and $\Pr(v=1)$ are within a factor of $(1 \pm \epsilon_4 2^{-9})$ of the correct values. Now using Equation 1 we can estimate $\Lambda(e)$ within $\pm\epsilon_4/16$. In particular, our estimate of $\Lambda(e)$ is at most

$$\begin{aligned}
& (w(e)+2\epsilon_5)\sqrt{\frac{\Pr(v=0)\Pr(v=1)}{\Pr(w=0)\Pr(w=1)}\frac{(1+\epsilon_4 2^{-9})}{(1-\epsilon_4 2^{-9})}} \\
\leq \quad & (w(e)+2\epsilon_5)\sqrt{\frac{\Pr(v=0)\Pr(v=1)}{\Pr(w=0)\Pr(w=1)}}(1+\epsilon_4 2^{-7}) \\
\leq \quad & \Lambda(e)+\epsilon_4/16.
\end{aligned}$$

In the inequalities, we used Equation 6 and the fact that $\Lambda(e) \leq 1$. Similarly, by Equation 7, our estimate of $\Lambda(e)$ is at least

$$\begin{aligned}
& (w(e)-2\epsilon_5)\sqrt{\frac{\Pr(v=0)\Pr(v=1)}{\Pr(w=0)\Pr(w=1)}\frac{(1-\epsilon_4 2^{-9})}{(1+\epsilon_4 2^{-9})}} \\
\geq \quad & (w(e)-2\epsilon_5)\sqrt{\frac{\Pr(v=0)\Pr(v=1)}{\Pr(w=0)\Pr(w=1)}}(1-\epsilon_4 2^{-8}) \\
\geq \quad & \Lambda(e)-\epsilon_4/16.
\end{aligned}$$

We now show how to estimate e_0, e_1, $\Pr(u=0)$ and $\Pr(v=0)$ within $\pm\epsilon_5$. We say that a path from node α to node β in a MET is *strong* if $|\mathrm{cov}(\alpha,\beta)| \geq \epsilon_2/2$. It follows from Equation 3 that if node γ is on this path then

$$|\mathrm{cov}(\gamma,\beta)| \geq |\mathrm{cov}(\alpha,\beta)| \tag{8}$$

$$|\mathrm{cov}(\alpha,\beta)| \geq |\mathrm{cov}(\alpha,\gamma)|\,|\mathrm{cov}(\gamma,\beta)| \tag{9}$$

We say that a quartet $(c,b \mid a,d)$ of leaves a, b, c and d is a *good estimator* of the edge $e=(u,v)$ if e is an edge of $T(S')$ and the following hold in $T(S')$ (see Figure 1).

1. a is a descendent of v.
2. The undirected path from c to a is strong and passes through u then v.
3. The path from u to its descendent b is strong and only intersects the (undirected) path from c to a at node u.
4. The path from v to its descendent d is strong and only intersects the path from v to a at node v.

We say that $(c,b \mid a,d)$ is an *apparently good estimator* of e' if the following hold in the ϵ_4-contraction $T'(S')$.

1. a is a descendent of v'.
2. The undirected path from c to a is strong and passes through u' then v'.
3. The path from u' to its descendent b is strong and only intersects the (undirected) path from c to a at node u'.
4. The path from v' to its descendent d is strong and only intersects the path from v' to a at node v'.

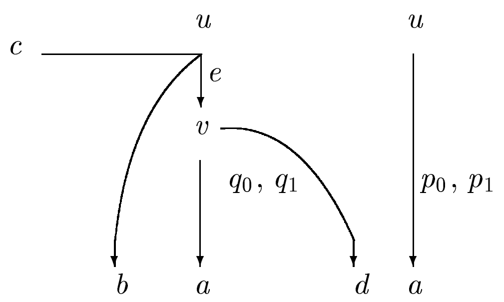

Figure 1: Finding $\Pr(u=1)$, e_0 and e_1

Observation 12 *If e is an edge of $T(S')$ and $(c,b \mid a,d)$ is a good estimator of e then any leaves $x,y \in \{a,b,c,d\}$ have $|\mathrm{cov}(x,y)| \geq (\epsilon_2/2)^3$.*

Proof: The observation follows from Equation 8 and 9 and from the definition of a good estimator. $\square$

Lemma 13 *If $(c,b \mid a,d)$ is a good estimator of e then it can be used (along with $\mathrm{poly}(n,1/\epsilon,1/\delta)$ samples from $M(S')$) to estimate e_0, e_1, $\Pr(u=0)$ and $\Pr(v=0)$ within $\pm\epsilon_5$. (If we use sufficiently many samples, then the probability that any of the estimates is not within $\pm\epsilon_5$ of the correct value is at most $\delta/(12n^7)$).*

Proof: Let q_0 and q_1 denote the transition probabilities from v to a (see Figure 1) and let p_0 and p_1 denote the transition probabilities from u to a. We will first show how to estimate p_0, p_1, and $\Pr(u=1)$ within $\pm\epsilon_6$. Without loss of generality (by Observation 3) we can assume that c is a descendant of u. (Otherwise we can re-root $T(S')$ at u without changing the distribution on the nodes or p_0 or p_1.) Let β be the path from u to b and let γ be the path from u to c. We now define

$$\begin{aligned}
\mathrm{cov}(b,c,0) &= \Pr(abc=011)\Pr(a=0) - \Pr(ab=01)\Pr(ac=01), \\
\mathrm{cov}(b,c,1) &= \Pr(abc=111)\Pr(a=1) - \Pr(ab=11)\Pr(ac=11).
\end{aligned} \tag{10}$$

(These do not quite correspond to the conditional covariances of b and c, but they are related to these.) We also define

$$\begin{aligned}
F &= \frac{1}{2}\left(\frac{\mathrm{cov}(b,c) + \mathrm{cov}(b,c,0) - \mathrm{cov}(b,c,1)}{\mathrm{cov}(b,c)}\right), \text{ and} \\
D &= F^2 - \mathrm{cov}(b,c,0)/\mathrm{cov}(b,c).
\end{aligned}$$

The following equations can be proved by algebraic manipulation from Equation 10, Equation 2 and the definitions of F and D.

$$\mathrm{cov}(b,c,0) = \Pr(u=1)\Pr(u=0)(1-\beta_0-\beta_1)(1-\gamma_0-\gamma_1)p_1(1-p_0) \tag{11}$$
$$\mathrm{cov}(b,c,1) = \Pr(u=1)\Pr(u=0)(1-\beta_0-\beta_1)(1-\gamma_0-\gamma_1)p_0(1-p_1)$$
$$F = \frac{1+p_1-p_0}{2} \tag{12}$$

$$D = \frac{(1-p_0-p_1)^2}{4} \tag{13}$$

Case $1a$**:** $a \in S_1$

In this case, by Equation 4 and by Observation 10, we have $1 - p_0 - p_1 > 0$. Thus, by Equation 13, we have

$$\sqrt{D} = \frac{1-p_0-p_1}{2}. \tag{14}$$

Equations 12 and 14 imply

$$p_1 = F - \sqrt{D} \tag{15}$$

$$p_0 = 1 - F - \sqrt{D} \tag{16}$$

Also, since $\Pr(a=0) = \Pr(u=1)p_1 + (1-\Pr(u=1))(1-p_0)$, we have

$$\Pr(u=1) = \frac{1}{2} + \frac{F - \Pr(a=0)}{2\sqrt{D}} \tag{17}$$

From these equations, it is clear that we could find p_0, p_1, and $\Pr(u=1)$ if we knew $\Pr(a=0)$, $\mathrm{cov}(b,c)$, $\mathrm{cov}(b,c,0)$ and $\mathrm{cov}(b,c,1)$ exactly. We now show that with polynomially-many samples, we can approximate the values of $\Pr(a=0)$, $\mathrm{cov}(b,c)$, $\mathrm{cov}(b,c,0)$ and $\mathrm{cov}(b,c,1)$ sufficiently accurately so that using our approximations and the above equations, we obtain approximations for p_0, p_1 and $\Pr(u=1)$ which are within $\pm\epsilon_6$. As in the proof of Lemma 6, we can use Equations 2 and 10 to estimate $\Pr(a=0)$, $\mathrm{cov}(b,c)$, $\mathrm{cov}(b,c,0)$ and $\mathrm{cov}(b,c,1)$ within $\pm\epsilon'$ for any ϵ' whose inverse is at most a polynomial in n and $1/\epsilon$. Note that our estimate of $\mathrm{cov}(b,c)$ will be non-zero by Observation 12 (as long as $\epsilon' \le (\epsilon_2/2)^3$), so we will be able to use it to estimate F from its definition. Now, using the definition of F and Equation 5, our estimate of $2F$ is at most

$$2F + \frac{3\epsilon'}{\mathrm{cov}(b,c) - 3\epsilon'}(1+2F).$$

By Observation 12, this is at most

$$2F + \frac{3\epsilon'}{(\epsilon_2/2)^3 - 3\epsilon'}(1+2). \tag{18}$$

The error is at most ϵ'' for any ϵ'' whose is inverse is at most polynomial in n and $1/\epsilon$. (This is accomplished by making ϵ' small enough with respect to ϵ_2 according to equation 18.) We can similarly bound the amount that we underestimate F. Now we use the definition of D to estimate D. Our estimate is at most

$$(F+\epsilon'')^2 - \frac{\mathrm{cov}(b,c,0) - \epsilon'}{\mathrm{cov}(b,c) + \epsilon'}.$$

Using Equation 5, this is at most

$$D + 2\epsilon'' F + \epsilon''^2 + \frac{\epsilon'}{\mathrm{cov}(b,c)+\epsilon'}\left(1 + \frac{\mathrm{cov}(b,c,0)}{\mathrm{cov}(b,c)}\right).$$

Once again, by Observation 12, the error can be made within $\pm\epsilon'''$ for any ϵ''' whose is inverse is polynomial in n and $1/\epsilon$ (by making ϵ' and ϵ'' sufficiently small). It follows that our estimate of $\sqrt{D}$ is at most $\sqrt{D}(1+\epsilon'''/(2D))$ and (since Observation 12 gives us an upper bound on the value of D as a function of ϵ_2), we can estimate $\sqrt{D}$ within $\pm\epsilon''''$ for any ϵ'''' whose inverse is polynomial in n and $1/\epsilon$. This implies that we can estimate p_0 and p_1 within $\pm\epsilon_6$. Observation 12 and Equation 3 imply that $w(p) \geq (\epsilon_2/2)^3$. Thus, the estimate for $\sqrt{D}$ is non-zero. This implies that we can similarly estimate $\Pr(u=1)$ within $\pm\epsilon_6$ using Equation 17.

Now that we have estimates for p_0, p_1, and $\Pr(u=1)$ which are within $\pm\epsilon_6$ of the correct values, we can repeat the trick to find estimates for q_0 and q_1 which are also within $\pm\epsilon_6$. We use leaf d for this. Observation 4 implies that

$$e_0 = \frac{p_0 - q_0}{1 - q_0 - q_1} \text{ and } e_1 = \frac{p_1 - q_1}{1 - q_0 - q_1}.$$

Using these equations, our estimate of e_0 is at most

$$\frac{p_0 - q_0 + 2\epsilon_6}{1 - q_0 - q_1 - 2\epsilon_6}.$$

Equation 5 and our observation above that $w(p) \geq (\epsilon_2/2)^3$ imply that the error is at most

$$\frac{2\epsilon_6}{(\epsilon_2/2)^3 - 2\epsilon_6}\left(1 + \frac{p_0 - q_0}{1 - q_0 - q_1}\right),$$

which is at most $2^7\epsilon_6/\epsilon_2^3 = \epsilon_5$. Similarly, the estimate for e_0 is at least $e_0 - \epsilon_5$ and the estimate for e_1 is within $\pm\epsilon_5$ of e_1. We have now estimated e_0, e_1, and $\Pr(u=0)$ within $\pm\epsilon_5$. As we explained in the beginning of this section, we can use these estimates to estimate $\Lambda(e)$ within $\pm\epsilon_4/16$.

Case $1b$**:** $a \in S_2$

In this case, by Equation 4 and by Observation 10, we have $1 - p_0 - p_1 < 0$. Thus, by equation 13, we have

$$\sqrt{D} = -\left(\frac{1 - p_0 - p_1}{2}\right). \tag{19}$$

Equations 12 and 19 imply

$$p_1 = F + \sqrt{D} \tag{20}$$
$$p_0 = 1 - F + \sqrt{D} \tag{21}$$

Equation 17 remains unchanged. The process of estimating p_0, p_1 and $\Pr(u=1)$ (from the new equations) is the same as for Case $1a$. This concludes the proof of Lemma 13. □

Observation 14 *Suppose that e' is an edge from u' to v' in $T'(S')$ and that $e = (u,v)$ is the edge in $T(S')$ such that $u \in u'$ and $v \in v'$. There is a good estimator $(c, b \mid a, d)$ of e. Furthermore, every good estimator of e is an apparently good estimator of e'. (Refer to Figure 2.)*

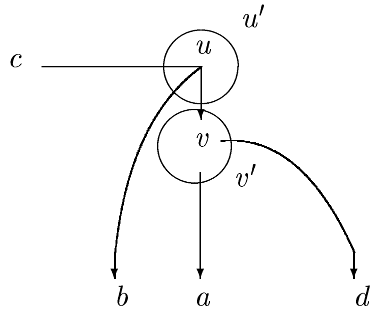

Figure 2: $(c, b \mid a, d)$ is a good estimator of $e = (u, v)$ and an apparently good estimator of $e' = (u', v')$.

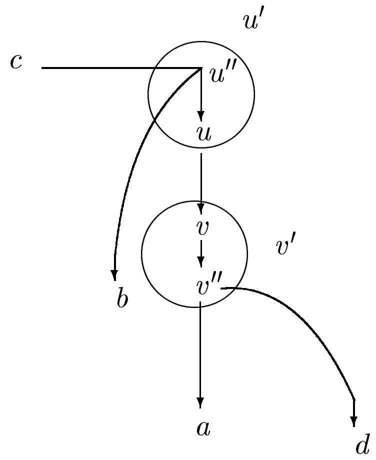

Figure 3: $(c, b \mid a, d)$ is an apparently good estimator of $e' = (u', v')$ and a good estimator of $p = (u'', v'')$. $\Lambda(p) \leq \Lambda(u, v)$.

Proof: Leaves c and a can be found to satisfy the first two criteria in the definition of a good estimator by Observation 8. Leaf b can be found to satisfy the third criterion by Observation 8 and Equation 8 and by the fact that the degree of u is at least 3 (see the text just before Equation 4). Similarly, leaf d can be found to satisfy the fourth criterion. $(c, b \mid a, d)$ is an apparently good estimator of e' because only internal edges of $T(S')$ can be contracted in the ϵ_4-contraction $T'(S')$. $\square$

Observation 15 *Suppose that e' is an edge from u' to v' in $T'(S')$ and that $e = (u, v)$ is an edge in $T(S')$ such that $u \in u'$ and $v \in v'$. Suppose that $(c, b \mid a, d)$ is an apparently good estimator of e'. Let u'' be the meeting point of c, b and a in $T(S')$. Let v'' be the meeting point of c, a and d in $T(S')$. (Refer to Figure 3.) Then $(c, b \mid a, d)$ is a good estimator of the path p from u'' to v'' in $T(S')$. Also, $\Lambda(p) \leq \Lambda(e)$.*

Proof: The fact that $(c, b \mid a, d)$ is a good estimator of p follows from the definition of good estimator. The fact that $\Lambda(p) \le \Lambda(e)$ follows from the fact that the distances Λ are multiplicative along a path, and bounded above by 1. $\square$

Observations 14 and 15 imply that in order to estimate $\Lambda(e)$ within $\pm\epsilon_4/16$, we need only estimate $\Lambda(e)$ using each apparently good estimator of e' and then take the maximum. By Lemma 13, the failure probability for any given estimator is at most $\delta/(12n^7)$, so with probability at least $1-\delta/(12n^3)$, all estimators give estimates within $\pm\epsilon_4/16$ of the correct values. Since there are at most $2n$ edges e' in $T'(S')$, and we add a new leaf x to S' at most n times, all estimates are within $\pm\epsilon_4/16$ with probability at least $1-\delta/(6n)$.

Case 2: e' is not an internal edge

In this case $v = v'$ since v' is a leaf of $T(S')$. We say that a pair of leaves (b,c) is a good estimator of e if the following holds in $T(S')$: The paths from leaves v, b and c meet at u and $|\mathrm{cov}(v,b)|$, $|\mathrm{cov}(v,c)|$ and $|\mathrm{cov}(b,c)|$ are all at least $(\epsilon_2/2)^2$. We say that (b,c) is an apparently good estimator of e' if the following holds in $T'(S')$: The paths from leaves v, b and c meet at u' and $|\mathrm{cov}(v,b)|$, $|\mathrm{cov}(v,c)|$ and $|\mathrm{cov}(b,c)|$ are all at least $(\epsilon_2/2)^2$. As in the previous case, the result follows from the following observations.

Observation 16 *If (b,c) is a good estimator of e then it can be used (along with* $\mathrm{poly}(n,1/\epsilon,1/\delta)$ *samples from $M(S')$) to estimate e_0, e_1, and* $\Pr(u=0)$ *within $\pm\epsilon_5$. (The probability that any of the estimates is not within $\pm\epsilon_5$ of the correct value is at most $\delta/(12n^3)$.)*

Proof: This follows from the proof of Lemma 13. $\square$

Observation 17 *Suppose that e' is an edge from u' to leaf v in $T'(S')$ and that $e=(u,v)$ is an edge in $T(S')$ such that $u \in u'$. There is a good estimator (b,c) of e. Furthermore, every good estimator of e is an apparently good estimator of e'.*

Proof: This follows from the proof of Observation 14 and from Equation 9. $\square$

Observation 18 *Suppose that e' is an edge from u' to leaf v in $T'(S')$ and that $e=(u,v)$ is an edge in $T(S')$ such that $u \in u'$. Suppose that (b,c) is an apparently good estimator of e'. Let u'' be the meeting point of b, v and c in $T(S')$. Then (b,c) is a good estimator of the path p from u'' to v in $T(S')$. Also, $\Lambda(p) \le \Lambda(e)$.*

Proof: This follows from the proof of Observation 15. $\square$

2.3.2 Using the Estimates of $\Lambda(e)$.

We now return to the problem of showing how to add a new leaf x to $T'(S')$. As we indicated above, for every internal edge $e'=(u',v')$ of $T'(S')$, we use the method in Section 2.3.1 to estimate $\Lambda(e)$ where $e=(u,v)$ is the edge of $T(S')$ such that $u \in u'$ and $v \in v'$. If the observed value of $\Lambda(e)$ exceeds $1-15\epsilon_4/16$ then we will contract e. The accuracy of our estimates will guarantee that we will not contract e if $\Lambda(e) \le 1-\epsilon_4$, and that we definitely contract e if $\Lambda(e) > 1-7\epsilon_4/8$. We will then add the new leaf x to $T'(S')$ as follows. We will insert a new edge (x,x') into $T'(S')$. We will do this by either (1) identifying x' with a node already in $T'(S')$, or (2) splicing x' into the middle of some edge of $T'(S')$.

We will now show how to decide where to attach x' in $T'(S')$. We start with the following definitions. Let S'' be the subset of S' such that for every $y \in S''$ we have $|\mathrm{cov}(x,y)| \geq (\epsilon_2/2)^4$. Let T'' be the subtree of $T'(S')$ induced by the leaves in S''. Let S''' be the subset of S' such that for every $y \in S'''$ we have $|\widehat{\mathrm{cov}}(x,y)| \geq (\epsilon_2/2)^4 - \epsilon_3$. Let T''' be the subtree of $T'(S')$ induced by the leaves in S'''.

Observation 19 *If $T(S' \cup \{x\})$ has x' attached to an edge $e = (u,v)$ of $T(S')$ and e' is the edge corresponding to e in $T'(S')$ (that is, $e' = (u',v')$, where $u \in u'$ and $v \in v'$), then e' is an edge of T''.*

Proof: By Observation 14 there is a good estimator $(c, b \mid a, d)$ for e. Since x is being added to S' (using Equation 8), $|\mathrm{cov}(x,x')| \geq \epsilon_2/2$. Thus, by Observation 12 and Equation 9, every leaf $y \in \{a,b,c,d\}$ has $|\mathrm{cov}(x,y)| \geq (\epsilon_2/2)^4$. Thus, a, b, c and d are all in S'' so e' is in T''. □

Observation 20 *If $T(S' \cup \{x\})$ has x' attached to an edge $e = (u,v)$ of $T(S')$ and u and v are both contained in node u' of $T'(S')$ then u' is a node of T''.*

Proof: Since u is an internal node of $T(S')$, it has degree at least 3. By Observation 8 and Equation 8, there are three leaves a_1, a_2 and a_3 meeting at u with $|\mathrm{cov}(u,a_i)| \geq \epsilon_2/2$. Similarly, $|\mathrm{cov}(u,v)| \geq \epsilon_2/2$. Thus, for each a_i, $|\mathrm{cov}(x,a_i)| \geq (\epsilon_2/2)^3$ so a_1, a_2, and a_3 are in S''. □

Observation 21 $S'' \subseteq S'''$.

Proof: This follows from the accuracy of the covariance estimates in Step 1. □

We will use the following algorithm to decide where to attach x' in T'''. In the algorithm, we will use the following tool. For any triple (a,b,c) of leaves in $S' \cup \{x\}$, let u denote the meeting point of the paths from leaves a, b and c in $T(S' \cup \{x\})$. Let M^u be the MET which has the same distribution as $M(S' \cup \{x\})$, but is rooted at u. (M^u exists, by Observation 3.) Let $\Lambda_c(a,b,c)$ denote the weight of the path from u to c in M^u. By observation 11, $\Lambda_c(a,b,c)$ is equal to the weight of the path from u to c in $M(S' \cup \{x\})$. (This follows from the fact that re-rooting at u only redirects internal edges.) It follows from the definition of Λ (Equation 1) and from Equation 3 that

$$\Lambda_c(a,b,c) = \sqrt{\frac{\mathrm{cov}(a,c)\mathrm{cov}(b,c)}{\mathrm{cov}(a,b)}}. \tag{22}$$

If a, b and c are in $S''' \cup \{x\}$, then by the accuracy of the covariance estimates and Equations 8 and 9, the absolute value of the pairwise covariance of any pair of them is at least $\epsilon_2^8/2^{10}$. As in Section 2.3.1, we can estimate $\mathrm{cov}(a,c)$, $\mathrm{cov}(b,c)$ and $\mathrm{cov}(a,b)$ within a factor of $(1 \pm \epsilon')$ of the correct values for any ϵ' whose inverse is at most a polynomial in n, and $1/\epsilon$. Thus, we can estimate $\Lambda_c(a,b,c)$ within a factor of $(1 \pm \epsilon_4/16)$ of the correct value. We will take sufficiently many samples to ensure that the probability that any of the estimates is outside of the required range is at most $\delta/(6n^2)$. Thus, the probability that any estimate is outside of the range for any x is at most $\delta/(6n)$.

We will now determine where in T''' to attach x'. Choose an arbitrary internal root u' of T'''. We will first see where x' should be placed with respect to u'. For each neighbour v' of u' in T''', each pair of leaves (a_1, a_2) on the "u'" side of (u',v') and each leaf b on the "v'" side of (u',v') (see Figure 4) perform the following two tests.

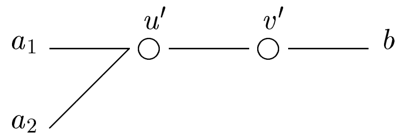

Figure 4: The setting for Test1(u', v', a_1, a_2, b) and Test2(u', v', a_1, a_2, b) when v' is an internal node of T'''. (If v' is a leaf, we perform the same tests with $v' = b$.).

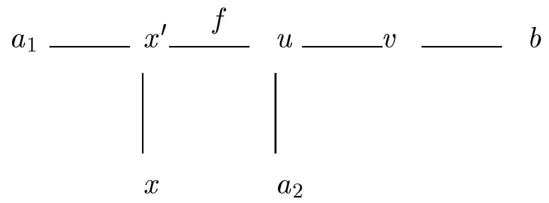

Figure 5: Either Test1(u', v', a_1, a_2, b) fails or Test2(u', v', a_1, a_2, b) fails.

- **Test1**(u', v', a_1, a_2, b)**:** The test succeeds if the observed value of $\Lambda_x(a_1, x, b)/\Lambda_x(a_2, x, b)$ is at least $1 - \epsilon_4/4$.
- **Test2**(u', v', a_1, a_2, b)**:** The test succeeds if the observed value of $\Lambda_b(a_1, a_2, b)/\Lambda_b(a_1, x, b)$ is at most $1 - 3\epsilon_4/4$.

We now make the following observations.

Observation 22 *If x is on the "u side" of (u, v) in $T(S''' \cup \{x\})$ and u is in u' in T''' and v is in $v' \neq u'$ in T''' then some test fails.*

Proof: Since u' is an internal node of T''', it has degree at least 3. Thus, we can construct a test such as the one depicted in Figure 5. (If $x' = u$ then the figure is still correct, that would just mean that $\Lambda(f) = 1$. Similarly, if v' is a leaf, we simply have $\Lambda(f') = 1$ where f' is the edge from v to b.) Now we have

$$\frac{1}{\Lambda(f)} = \frac{\Lambda_x(a_1, x, b)}{\Lambda_x(a_2, x, b)} = \frac{\Lambda_b(a_1, a_2, b)}{\Lambda_b(a_1, x, b)}.$$

However, Test1(u', v', a_1, a_2, b) will only succeed if the left hand fraction is at least $1 - \epsilon_4/4$. Furthermore, Test2(u', v', a_1, a_2, b) will only succeed if the right hand fraction is at most $1 - 3\epsilon_4/4$. Since our estimates are accurate to within a factor of $(1 \pm \epsilon_4/16)$, at least one of the two tests will fail. □

Observation 23 *If x is between u and v in $T(S''' \cup \{x\})$ and the edge f from u to x' has $\Lambda(f) \leq 1 - 7\epsilon_4/8$ then Test1(u', v', a_1, a_2, b) and Test2(u', v', a_1, a_2, b) succeed for all choices of a_1, a_2 and b.*

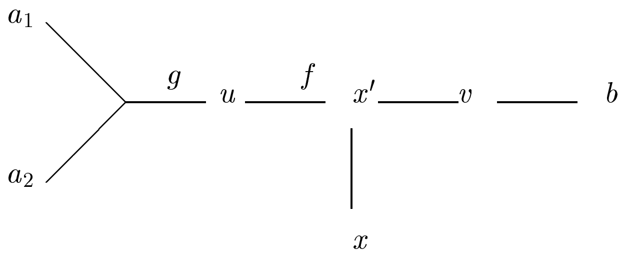

Figure 6: Test1(u', v', a_1, a_2, b) and Test2(u', v', a_1, a_2, b) succeed for all choices of a_1, a_2 and b.

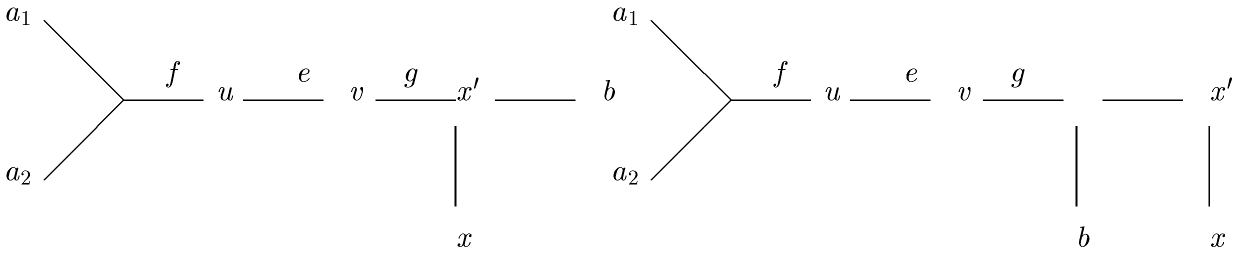

Figure 7: Test1(u', v', a_1, a_2, b) and Test2(u', v', a_1, a_2, b) succeed for all choices of a_1, a_2 and b.

Proof: Every such test has the form depicted in Figure 6, where again g might be degenerate, in which case $\Lambda(g) = 1$. Observe that $\Lambda_x(a_1, x, b)/\Lambda_x(a_2, x, b) = 1$, so its estimate is at least $1 - \epsilon_4/4$ and Test1 succeeds. Furthermore,

$$\frac{\Lambda_b(a_1, a_2, b)}{\Lambda_b(a_1, x, b)} = \Lambda(f)\Lambda(g) \leq \Lambda(f) \leq 1 - 7\epsilon_4/8,$$

so the estimate is at most $1 - 3\epsilon_4/4$ and Test2 succeeds. □

Observation 24 *If x is on the "v side" of (u, v) in $T(S''' \cup \{x\})$ and $\Lambda(e) \leq 1 - 7\epsilon_4/8$ (recall from the beginning of Section 2.3.2 that $\Lambda(e)$ is at most $1 - 7\epsilon_4/8$ if u and v are in different nodes of T'''), then Test1(u', v', a_1, a_2, b) and Test2(u', v', a_1, a_2, b) succeed for all choices of a_1, a_2 and b.*

Proof: Note that this case only applies if v is an internal node of $T(S''')$. Thus, every such test has one of the forms depicted in Figure 7, where some edges may be degenerate. Observe that in both cases $\Lambda_x(a_1, x, b)/\Lambda_x(a_2, x, b) = 1$, so its estimate is at least $1 - \epsilon_4/4$ and Test1 succeeds. Also in both cases

$$\frac{\Lambda_b(a_1, a_2, b)}{\Lambda_b(a_1, x, b)} = \Lambda(e)\Lambda(f)\Lambda(g) \leq \Lambda(e) \leq 1 - 7\epsilon_4/8,$$

so the estimate is at most $1 - 3\epsilon_4/4$ and Test2 succeeds. □

Now note (using Observation 22) that node u' has at most one neighbour v' for which all tests succeed. Furthermore, if there is no such v', Observations 23 and 24 imply that x' can be merged with u'. The only case that we have not dealt with is the case in which there is exactly one v' for which all tests succeed. In this case, if v' is a leaf, we insert x' in the middle of edge (u', v'). Otherwise, we will either insert x' in the middle of edge (u', v'), or we will insert it in the subtree rooted at v'. In order to decide which, we perform similar tests from node v', and we check whether $\text{Test1}(v', u', a_1, a_2, b)$ and $\text{Test2}(v', u', a_1, a_2, b)$ both succeed for all choices of a_1, a_2, and b. If so, we put x' in the middle of edge (u', v'). Otherwise, we recursively place x' in the subtree rooted at v'.

2.4 Step 4: For each related set S, construct a MET $M'(S)$ which is close to $M(S)$

For each set S of related leaves we will construct a MET $M'(S)$ with leaf-set S such that each edge parameter of $M'(S)$ is within $\pm\epsilon_1$ of the corresponding parameter of $M(S)$. The topology of $M'(S)$ will be $T'(S)$. We will assume without loss of generality that $T(S)$ has the same root as $T'(S)$. The failure probability for S will be at most $\delta/(3n)$, so the overall failure will be at most $\delta/3$.

We start by observing that the problem is easy if S has only one or two leaves.

Observation 25 *If $|S| < 3$ then we can construct a MET $M'(S)$ such that each edge parameter of $M'(S)$ is within $\pm\epsilon_1$ of the corresponding parameter of $M(S)$.*

We now consider the case in which S has at least three leaves. Any edge of $T(S)$ which is contracted in $T'(S)$ can be regarded as having e_0 and e_1 set to 0. The fact that these are within $\pm\epsilon_1$ of their true values follows from the following lemma.

Lemma 26 *If e is an internal edge of $M(S)$ from v to w with $\Lambda(e) > 1 - \epsilon_4$ then $e_0 + e_1 < 2\epsilon_4 = \epsilon_1/(2n)$.*

Proof: First observe from Observation 9 that $\Pr(w = 0) \notin \{0, 1\}$ and from Observation 10 that $e_0 + e_1 \neq 1$. Using algebraic manipulation, one can see that

$$\begin{aligned} \Pr(v = 1) &= \frac{\Pr(w = 1) - e_0}{1 - e_0 - e_1} \\ \Pr(v = 0) &= \frac{\Pr(w = 0) - e_1}{1 - e_0 - e_1}. \end{aligned}$$

Thus, by Equation 1,

$$\Lambda(e)^2 = \left(1 - \frac{e_0}{\Pr(w = 1)}\right)\left(1 - \frac{e_1}{\Pr(w = 0)}\right).$$

Since $\Lambda(e)^2 \geq 1 - 2\epsilon_4$, we have $e_0 \leq 2\epsilon_4 \Pr(w = 1)$ and $e_1 \leq 2\epsilon_4 \Pr(w = 0)$, which proves the observation. $\square$

Thus, we need only show how to label the remaining parameters within $\pm\epsilon_1$. Note that we have already shown how to do this in Section 2.3.1. Here the total failure probability is at most $\delta/(3n)$ because there is a failure probability of at most $\delta/(6n^2)$ associated with each of the $2n$ edges.

2.5 Step 5: Form M' from the METs $M'(S)$

Make a new root r for M' and set $\Pr(r = 1) = 1$. For each related set S of leaves, let u denote the root of $M'(S)$, and let $\overline{p}$ denote the probability that u is 0 in the distribution of $M'(S)$. Make an edge e from r to u with $e_1 = \overline{p}$.

2.6 Proof of Theorem 1

Let M'' be a MET which is formed from M as follows.

- Related sets are formed as in Step 2.
- For each related set S, a copy $M''(S)$ of $M(S)$ is made.
- The METs $M''(S)$ are combined as in Step 5.

Theorem 1 follows from the following lemmas.

Lemma 27 *Suppose that for every set S of related leaves, every parameter of $M'(S)$ is within $\pm\epsilon_1$ of the corresponding parameter in $M(S)$. Then* $\mathrm{var}(M'', M') \leq \epsilon/2$.

Proof: First, we observe (using a crude estimate) that there are at most $5n^2$ parameters in M'. (Each of the (at most n) METs $M'(S)$ has one root parameter and at most $4n$ edge parameters.) We will now show that changing a single parameter of a MET by at most $\pm\epsilon_1$ yields at MET whose variation distance from the original is at most $2\epsilon_1$. This implies that $\mathrm{var}(M'', M') \leq 10n^2\epsilon_1 = \epsilon/2$. Suppose that e is an edge from u to v and e_0 is changed. The probability that the output has string s on the leaves below v and string s' on the remaining leaves is

$$\begin{aligned} & \Pr(u=0)\Pr(s' \mid u=0)(e_0 \Pr(s \mid v=1) + (1-e_0)\Pr(s \mid v=0)) \\ + \quad & \Pr(u=1)\Pr(s' \mid u=1)(e_1 \Pr(s \mid v=0) + (1-e_1)\Pr(s \mid v=1)). \end{aligned}$$

Thus, the variation distance between M'' and a MET obtained by changing the value of e_0 (within $\pm\epsilon_1$) is at most

$$\begin{aligned} & \epsilon_1 \sum_s \sum_{s'} \Pr(u=0)\Pr(s' \mid u=0)(\Pr(s \mid v=1) + \Pr(s \mid v=0)) \\ \leq \quad & \epsilon_1 \Pr(u=0)\left(\sum_{s'} \Pr(s' \mid u=0)\right)\left((\sum_s \Pr(s \mid v=1)) + (\sum_s \Pr(s \mid v=0))\right) \\ \leq \quad & 2\epsilon_1. \end{aligned}$$

Similarly, if ρ_1 is the root parameter of a MET then the probability of having output s is

$$\rho_1 \Pr(s \mid r=1) + (1-\rho_1)\Pr(s \mid r=0).$$

So the variation distance between the original MET and one in which ρ_1 is changed within $\pm\epsilon_1$ is at most

$$\sum_s \epsilon_1(\Pr(s \mid r=1) + \Pr(s \mid r=0)) \leq 2\epsilon_1.$$

□

Lemma 28 $\mathrm{var}(M'', M) \leq \epsilon/2$.

Before we prove Lemma 28, we provide some background material. Recall that the *weight* $w(e)$ of an edge e of a MET is $|1 - e_0 - e_1|$ and define the weight $w(\ell)$ of a leaf ℓ to be the product of the weights of the edges on the path from the root to ℓ. We will use the following lemma.

Lemma 29 *In any MET with root r, the variation distance between the distribution on the leaves conditioned on $r=1$ and the distribution on the leaves conditioned on $r=0$ is at most $2\sum_\ell w(\ell)$, where the sum is over all leaves ℓ.*

Proof: We proceed by induction on the number of edges in the MET. In the base case there are no edges so r is a leaf, and the result holds. For the inductive step, let e be an edge from r to node x. For any string s_1 on the leaves below x and any string s_2 on the other leaves,

$$\Pr(s_1 s_2 \mid r=0) = \Pr(s_2 \mid r=0)(e_0 \Pr(s_1 \mid x=1) + (1-e_0)\Pr(s_1 \mid x=0).$$

Algebraic manipulation of this formula shows that $\Pr(s_1 s_2 \mid r=1) - \Pr(s_1 s_2 \mid r=0)$ is

$$\begin{aligned} & (1-e_0-e_1)\ \Pr(s_2 \mid r=1)\,(\Pr(s_1 \mid x=1) - \Pr(s_1 \mid x=0)) \\ + \quad & \Pr(s_1 \mid r=0)\,(\Pr(s_2 \mid r=1) - \Pr(s_2 \mid r=0)). \end{aligned} \tag{23}$$

It follows that the variation distance is at most the sum over all $s_1 s_2$ of the absolute value of the quantity in Equation 23, which is at most

$$\begin{aligned} & |1-e_0-e_1| \left(\sum_{s_2} \Pr(s_2 \mid r=1)\right) \left(\sum_{s_1} |\Pr(s_1 \mid x=1) - \Pr(s_1 \mid x=0)|\right) \\ + \quad & \left(\sum_{s_1} \Pr(s_1 \mid r=0)\right) \left(\sum_{s_2} |\Pr(s_2 \mid r=1) - \Pr(s_2 \mid r=0)|\right). \end{aligned}$$

The result follows by induction. □

Lemma 30 *Suppose that m is a MET with n leaves and that e is an edge from node u to node v. Let m' be the MET derived from m by replacing e_0 with $\Pr(v=1)$ and e_1 with $\Pr(v=0)$. Then $\mathrm{var}(m,m') \le n^2 z$, where z is the maximum over all pairs (x,y) of leaves which are connected via e in m of $|\mathrm{cov}(x,y)|$.*

Proof: By Observation 3, we can assume without loss of generality that u is the root of m. For any string s_1 on the leaves below v and any string s_2 on the remaining leaves, we find (via a little algebraic manipulation) that the difference between the probability that m outputs $s_1 s_2$ and the probability that m' does is

$$\Pr(u=1)\Pr(u=0)(1-e_0-e_1)(\Pr(s_2 \mid u=1) - \Pr(s_2 \mid u=0))(\Pr(s_1 \mid v=1) - \Pr(s_1 \mid v=0)).$$

Thus, the variation distance between m and m' is $\Pr(u=1)\Pr(u=0)(1-e_0-e_1)$ times the product of the variation distance between the distribution on the leaves below v conditioned on $v=1$ and the distribution on the leaves below v conditioned on $v=0$ and the variation distance between the distribution on the remaining leaves conditioned on $u=1$ and the distribution on the remaining leaves conditioned on $u=0$. By Lemma 29, this is at most

$$\Pr(u=0)\Pr(u=1)\left(2 \sum_{\ell \text{ below } v} w(\ell)\right)\left(2 \sum_{\text{other } \ell} w(\ell)\right),$$

which by Equation 3 is

$$4 \sum_{(x,y) \text{ connected via } e} |\mathrm{cov}(x,y)|,$$

which is at most $4(n/2)^2 z = n^2 z$. □

Lemma 31 *If, for two different related sets, S and S', an edge e from u to v is in $M(S)$ and in $M'(S)$, then $e_0 + e_1 \leq n^2\epsilon_2/(n+1)$.*

Proof: By the definition of the leaf connectivity graph in Step 2, there are leaves $a, a' \in S$ and $b, b' \in S'$ such that the path from a' to a and the path from b' to b both go through $e = u \to v$ and

$$|\widehat{\operatorname{cov}}(a,a')| \geq (3/4)\epsilon_2 \quad \text{and} \quad |\widehat{\operatorname{cov}}(b,b')| \geq (3/4)\epsilon_2,$$

and the remaining covariance estimates amongst leaves a, a', b and b' are less than $(3/4)\epsilon_2$. Without loss of generality (using Observation 3), assume that u is the root of the MET. Let $p_{u,a'}$ denote the path from u to a' and use similar notation for the other leaves. By Equation 3 and the accuracy of the estimates in Step 1,

$$\begin{aligned}
\Pr(u=0)^2 \Pr(u=1)^2 w(e)^2 w(p_{u,a'}) w(p_{v,a}) w(p_{u,b'}) w(p_{v,b}) &\geq ((3/4)\epsilon_2 - \epsilon_3)^2 \\
\Pr(u=0)\Pr(u=1) w(p_{u,a'}) w(p_{u,b'}) &< (3/4)\epsilon_2 + \epsilon_3 \\
\Pr(v=0)\Pr(v=1) w(p_{v,a}) w(p_{v,b}) &< (3/4)\epsilon_2 + \epsilon_3.
\end{aligned}$$

Thus,

$$w(e) \geq \left(1 - \frac{2\epsilon_3}{(3/4)\epsilon_2 + \epsilon_3}\right) \sqrt{\frac{\Pr(v=1)\Pr(v=0)}{\Pr(u=1)\Pr(u=0)}}.$$

By Equation 1,

$$\Lambda(e) \geq 1 - \frac{2\epsilon_3}{(3/4)\epsilon_2 + \epsilon_3}.$$

The result now follows from the proof of Lemma 26. (Clearly, the bound in the statement of Lemma 31 is weaker than we can prove, but it is all that we will need.) □

Proof of Lemma 28: Let M^* be the MET which is the same as M except that every edge e which is contained in $M(S)$ and $M(S')$ for two different related sets S and S' is contracted. Similarly, let M''^* be the MET which is the same as M'' except that every such edge has *all* of its copies contracted in M''^*. Clearly, $\operatorname{var}(M, M'') \leq \operatorname{var}(M, M^*) + \operatorname{var}(M^*, M''^*) + \operatorname{var}(M''^*, M'')$. By Lemma 31, $\operatorname{var}(M, M^*) + \operatorname{var}(M''^*, M'') \leq \ell n^2 \epsilon_2$, where ℓ is the number of edges in M that are contracted. We now wish to bound $\operatorname{var}(M^*, M''^*)$. By construction, $M^*(S)$ and $M^*(S')$ do not intersect in an edge (for any related sets S and S'). Now suppose that $M^*(S)$ and $M^*(S')$ both contain node u. We can modify M^* *without changing the distribution* in a way that avoids this overlap. To do this, we just replace node u with two copies of u, and we connect the two copies by an edge e with $e_0 = e_1 = 0$. Note that this change will not affect the operation of the algorithm. Thus, without loss of generality, we can assume that for any related sets S and S', $M^*(S)$ and $M^*(S')$ do not intersect. Thus, M^* and M''^* are identical, except on edges which go between the sub-METs $M^*(S)$. Now, any edge e going between two sub-METs has the property that for any pair of leaves, x and y connected via e, $|\operatorname{cov}(x,y)| \leq \epsilon_2$. (This follows from the accuracy of our covariance estimates in Step 1.) Thus, by Lemma 30, changing such an edge according to Step 5 adds at most $n^2\epsilon_2$ to the variation distance. Thus, $\operatorname{var}(M^*, M''^*) \leq \ell' n^2 \epsilon_2$, where ℓ' is the number of edges that are modified according to Step 5. We conclude that $\operatorname{var}(M, M'') \leq (2n)n^2\epsilon_2 = \epsilon_1/2 \leq \epsilon/2$. □

Acknowledgements: We thank Mike Paterson for useful ideas and discussions.

References

[1] R. Agarwala, V. Bafna, M. Farach, B. Narayanan, M. Paterson and M. Thorup, On the Approximability of Numerical Taxonomy, *Proceedings of the 7th Annual ACM-SIAM Symposium on Discrete Algorithms*, (1996) 365–372.

[2] A. Ambainis, R. Desper, M. Farach and S. Kannan, Nearly Tight Bounds on the Learnability of Evolution, *Proceedings of the 38th Annual IEEE Symposium on the Foundations of Computer Science*, (1997) 524–533.

[3] N. Abe and M.K. Warmuth, Polynomial Learnability of Probabilistic Concepts with Respect to the Kullback-Leibler Divergence, *Proceedings of the 1992 Conference on Computational Learning Theory,* (1992) 277–289.

[4] H.J. Bandelt and A. Dress, Reconstructing the shape of a tree from observed dissimilarity data, *Adv. Appl. Math.* **7** (1987) 309–343.

[5] J.A. Cavender, Taxonomy with confidence. *Math. Biosci.*, **40** (1978), 271-280.

[6] T.H. Cover and J.A.Thomas, *Elements of Information Theory*, John Wiley & Sons, 1991.

[7] P.L. Erdös, M.A. Steel, L.A. Székely and T.J. Warnow, A few logs suffice to build (almost) all trees (I), DIMACS Technical Report 97-71, October 1997.

[8] P.L. Erdös, M.A. Steel, L.A. Székely and T.J. Warnow, A few logs suffice to build (almost) all trees (II), DIMACS Technical Report 97-72, October 1997. (A preliminary version appeared as "Inferring big trees from short quartets" in *Proceedings of the 24th International Colloquium on Automata, Languages and Programming*, (1997) 827–837)

[9] M. Farach and S. Kannan, Efficient algorithms for inverting evolution, *Proc. of the 28th Ann. ACM Symp. on Theory of Computing,* (1996) 230–236.

[10] J.S. Farris, A probability model for inferring evolutionary trees, *Syst. Zool.*, **22** (1973), 250-256.

[11] M. Kearns, Y. Mansour, D. Ron, R. Rubinfeld, R.E. Schapire and L. Sellie, On the Learnability of Discrete Distributions, *Proceedings of the 26th Annual ACM Symposium on the Theory of Computing*, (1994) 273–282.

[12] C. McDiarmid, On the method of bounded differences, *London Mathematical Society Lecture Note Series* **141**, Cambridge University Press, 1989, 148–188.

[13] 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, 1971) 1–27.

[14] M. Steel, Recovering a tree from the leaf colourations it generates under a Markov model, *Appl. Math. Lett.* **7(2)** 19–24 (1994).

3 Appendix

3.1 Proof of Lemma 2

Lemma 2 *A class of probability distributions over the domain $\{0,1\}^n$ that is PAC-learnable under the variation distance metric is PAC-learnable under the KL-distance measure.*

Proof: Let K be a polynomial in three inputs and let A be an algorithm which takes as input $K(n, 1/\epsilon, 1/\delta)$ samples from a distribution $\mathcal{D}$ from the class of distributions and, with probability at least $1-\delta$, returns a distribution $\mathcal{D}'$ such that $\mathrm{var}(\mathcal{D},\mathcal{D}') \leq \epsilon$. Without loss of generality, we can assume that ϵ is sufficiently small. For example, it will suffice to have $\epsilon \leq 2/15$.

Define algorithm A' as follows. Let $\xi = \epsilon^2/(12n)$. Run A with sample size $K(n, 1/\xi, 1/\delta)$ (note that the sample size is polynomial in n, $1/\epsilon$, and $1/\delta$). Let $\mathcal{D}'$ be the distribution returned by A. Let $\mathcal{U}$ denote the uniform distribution on $\{0,1\}^n$ and let $\mathcal{D}''$ be the distribution defined by

$$\mathcal{D}''(s) = (1-(\xi))\mathcal{D}'(s) + \xi\,\mathcal{U}(s).$$

With probability at least $1-\delta$, $\mathrm{var}(\mathcal{D},\mathcal{D}') \leq \xi$. By the definition of $\mathcal{D}''$, $\mathrm{var}(\mathcal{D}',\mathcal{D}'') \leq 2\xi$. Thus, with probability at least $1-\delta$, $\mathrm{var}(\mathcal{D},\mathcal{D}'') < 3\xi$. Note that for all s, $\mathcal{D}''(s) \geq \xi\, 2^{-n}$. Let S be the set of all output strings s satisfying $\mathcal{D}''(s) < \mathcal{D}(s)$. S contains all the strings which contribute positively to the KL-distance from $\mathcal{D}$ to $\mathcal{D}''$. Thus,

$$\begin{aligned}
\mathrm{KL}(\mathcal{D},\mathcal{D}'') &\leq \sum_{s\in S} \mathcal{D}(s)(\log \mathcal{D}(s) - \log \mathcal{D}''(s)) \\
&= \sum_{s\in S} (\mathcal{D}(s) - \mathcal{D}''(s))(\log \mathcal{D}(s) - \log \mathcal{D}''(s)) + \sum_{s\in S} \mathcal{D}''(s)(\log \mathcal{D}(s) - \log \mathcal{D}''(s)).
\end{aligned}$$

We have seen that $\mathrm{var}(\mathcal{D},\mathcal{D}'') \leq 3\xi$. Thus, $\sum_{s\in S}(\mathcal{D}(s) - \mathcal{D}''(s)) \leq 3\xi$. So, the first term is at most

$$\begin{aligned}
& \max_{s\in S}(\log \mathcal{D}(s) - \log \mathcal{D}''(s)) \sum_{s\in S}(\mathcal{D}(s) - \mathcal{D}''(s)) \\
\leq\;& 3\xi \max_{s\in S}(\log \mathcal{D}(s) - \log \mathcal{D}''(s)) \\
\leq\;& 3\xi \max_{s\in S}(-\log \mathcal{D}''(s)) \\
\leq\;& 3\xi(-\log(\xi\, 2^{-n})) \\
=\;& 3\xi(n - \log(\xi)).
\end{aligned}$$

Furthermore, the second term is at most

$$\begin{aligned}
& \sum_{s\in S} \mathcal{D}''(s)(\log \mathcal{D}(s) - \log \mathcal{D}''(s)) \\
=\;& \sum_{s\in S} \mathcal{D}''(s)(\log(\mathcal{D}''(s) + h_s) - \log \mathcal{D}''(s)),
\end{aligned}$$

where $h_s = \mathcal{D}(s) - \mathcal{D}''(s)$, which is a positive quantity for $s \in S$. By concavity of the logarithm function, the above quantity is at most

$$\sum_{s\in S} \mathcal{D}''(s) h_s \Big[\frac{d}{dx}(\log(x))\Big]_{x=\mathcal{D}''(s)} = \sum_{s\in S} h_s \leq 3\xi.$$

Thus, $\mathrm{KL}(\mathcal{D},\mathcal{D}'') \leq 3\xi(1+n-\log\xi)$. This quantity is at most ϵ for all $n \geq 1$ by the definition of ξ. □

The method in the proof of Lemma 2 converts a hypothesis distribution which is close (in variation distance) to the target distribution to a hypothesis distribution which is close (in KL-distance) to the target distribution. However, if the original hypothesis is given as a 2-state MET, then the modified hypothesis would require a 3-state MET to realize it. We conclude the paper by explaining how to perform a similar trick using only 2-state METs. The distribution obtained is not quite the same as the one used in the proof of Lemma 2, but it has the properties needed to show that small KL-distance is achieved.

Let M be the target Markov Evolutionary Tree. We run the PAC learning algorithm with accuracy parameter $\xi = \epsilon^2/(12n^3)$ to obtain MET M'. Now we construct a new hypothesis M'' by adjusting some of the parameters of M' as follows:

For each edge $e = (u, l)$ of M' where l is a leaf, let e_0 and e_1 be its parameters. If $e_0 < \xi$ then we set $e_0 = \xi$ and if $e_0 > 1 - \xi$ then set $e_0 = 1 - \xi$. We make the same change to e_1. By the proof of Lemma 27, $\mathrm{var}(M', M'') \leq 4n\xi$, since $2n$ parameters have each been changed by at most ξ. Hence, with probability at least $1 - \delta$, $\mathrm{var}(M, M'') \leq (1 + 4n)\xi$.

For each string $s \in \{0,1\}^n$, $M''(s) \geq \xi^n$ (where $M''(s)$ denotes the probability that M'' outputs s). Using a similar argument to the proof of lemma 2,

$$\mathrm{KL}(M, M'') \leq (1 + 4n)\xi(1 - \log(\xi^n)) = (1 + 4n)\xi(1 - n\log\xi)$$

$$= (1 + 4n)\frac{\epsilon^2}{12n^3}(1 - n(2\log\epsilon - 3\log n - \log 12))$$

which as before is at most ϵ for all $n \geq 1$.