Impossibility of phylogeny reconstruction from k-mer counts

Wai-Tong Louis Fan*† Brandon Legried‡
Sebastien Roch§

October 28, 2020

Abstract

We consider phylogeny estimation under a two-state model of sequence evolution by site substitution on a tree. In the asymptotic regime where the sequence lengths tend to infinity, we show that for any fixed k no statistically consistent phylogeny estimation is possible from k-mer counts of the leaf sequences alone. Formally, we establish that the joint leaf distributions of k-mer counts on two distinct trees have total variation distance bounded away from 1 as the sequence length tends to infinity. That is, the two distributions cannot be distinguished with probability going to one in that asymptotic regime. Our results are information-theoretic: they imply an impossibility result for any reconstruction method using only k-mer counts at the leaves.

1 Introduction

Molecular sequence comparisons are fundamental to many bioinformatics methods [Gus97, DEKM98, CP18]. In particular, the probabilistic analysis of sequences and their statistics has provided valuable insights, for instance, in

*Department of Mathematics, Indiana University, Bloomington.
†Center of Mathematical Sciences and Applications, Harvard University, Cambridge.
‡Department of Mathematics, University of Wisconsin–Madison.
§Department of Mathematics, University of Wisconsin–Madison.

comparative genomics [KA90, BC01, LHW02, RCSW09], population genetics [Tav84, PPP$^+$06, BAP05], and phylogenetics [Ste94, ESSW99, EKPS00, Mos04, RS17]. In this paper, we consider alignment-free phylogeny reconstruction [VA03, Hau14].

Alignment-free approaches are an important class of methods for estimating evolutionary trees that bypass the computationally hard multiple sequence alignment problem and avoid the need for a reference genome. Typically, these methods construct pairwise distances between sequences based on match lengths [UBTC06, HKP15] or k-mer counts [QWH04, Hau14, FISGC15]. Here a k-mer refers to a consecutive substring of length k in an input sequence. The pairwise distance matrix obtained is then used to reconstruct the phylogenetic relationships among the sequences. A variety of standard distance-based phylogenetic methods can be used for this purpose [War17, Ste16]. Numerous popular pipelines are available that implement these alignment-free approaches [HKP15, OTM$^+$16, LKP$^+$18, LHTH$^+$19], although they do not offer rigorous guarantees of accurate reconstruction.

In this paper, we consider the problem of phylogeny estimation under a two-state symmetric model of sequence evolution by site substitutions on a leaf-labeled tree. In the asymptotic regime where the sequence length tends to infinity, we show that:

for any fixed k, no statistically consistent phylogeny estimation is possible from the k-mer counts of the input sequences alone.

Formally, we establish that the joint leaf distributions of k-mer counts on two distinct trees have total variation distance bounded away from 1 as the sequence length tends to infinity. That is, the two distributions cannot be distinguished with probability going to one in that asymptotic regime. Our results are information-theoretic: they imply an impossibility result for any reconstruction method using only k-mer counts at the leaves, no matter how complex. To bound the total variation distance between the two distributions on well-chosen trees, our proof takes advantage of a multivariate local central limit theorem, an approach which is complicated by the probabilistic and algebraic dependencies of k-mers.

Related work A related impossibility result was established in our previous work [FLR20], where it was shown that no consistent distance estimation is possible from *sequence lengths alone* under the TKF91 model [TKF91], a more complex model of sequence evolution which also allows for insertions and deletions. On the other hand, sequence lengths are significantly simpler to analyze than k-mers and are not used in practice to infer phylogenies. Moreover the results in [FLR20] only apply to distance-based phylogeny estimation methods, while our current results are more general.

In a separate line of work, a computationally efficient algorithm for alignment-free phylogenetic reconstruction was developed and analyzed in [DR$^+$13]. Rigorous sequence length guarantees for high-probability reconstruction under an indel model (related to the TKF91 model) were established. While this method is based on 1-mers, it first divides up the input sequences into blocks of an appropriately chosen length and then compares the 1-mer counts on each block across sequences. The weak correlation between the blocks allows the use of concentration inequalities on a notion of pairwise distance proposed in [DR$^+$13]. In particular, this reconstruction method uses more information than 1-mer counts alone, so that our results do not apply to it. It also requires fine-tuning and has not been implemented in practice. Significantly improved bounds were obtained in [GZ19], who also exploited a connection to ancestral sequence reconstruction. In other related work, it was proved in [ARS15] that the tree topology as well as mutation parameters can be identified from pairwise joint k-mer count distributions under more general substitution-only models of sequence evolution using an appropriately defined notion of distance. See also [DS19] for extensions to coalescent-based models.

Alignment-free sequence comparisons based on k-mer counts were also studied for independent sequences with i.i.d. sites or under certain hidden Markov models of sequences [LHW02, RCSW09, WRSW10, BC01]. Because they assume independent sequences, such results are not directly relevant to phylogeny reconstruction.

Organization The paper is organized as follows. In Section 2, we state our main results after providing the necessary background and definitions. We also sketch the main steps of the proof. The details of the proof can be found in Section 3. A few auxiliary results are in the appendix.

2 Definitions and main result

In this section, we state our result formally, after introducing the relevant concepts.

k-mers. Let k be a positive integer, fixed throughout. First, we define k-mers and introduce their frequencies in a binary sequence, which will serve as our main statistic.

Definition 1. *A k-**mer** is a string of length k, i.e., $y \in \{0,1\}^k$. For a binary sequence $\sigma = (\sigma_i)_{i=1}^m$ of length m, we let $f_\sigma(y) \in \mathbb{Z}_+$ be the number of times y appears in σ as a consecutive substring:*

$$f_\sigma(y) = \sum_{i=0}^{m-k} \mathbf{1}\{(\sigma_{i+1}, \ldots, \sigma_{i+k}) = y\}.$$

*The **frequency vector (or count vector) of k-mers in** σ is the vector*

$$f_\sigma = (f_\sigma(y))_{y \in \{0,1\}^k} \in \mathbb{Z}_+^{2^k}.$$

The coordinates of f_σ are ordered such that the j-th coordinate is the frequency of the k-mer that is the binary representation of $j-1$ (i.e., the base-2 numeral representation of $j-1$).

For example, when $k = 1$, the count vector of 1-mers of a binary sequence is (a, b) where a is the number of zeros and b is the number of ones. Hence, the count vector of 1-mers of 00111000 is $(5, 3)$. When $k = 2$, there are $2^k = 4$ binary strings, namely $\{(00), (01), (10), (11)\}$. So the count vector of 2-mers of the sequence 001111000 is $(3, 1, 1, 3)$ since (00) appears 3 times, (01) appears one time, etc. By convention, the count vector of k-mers for any binary sequence with length less than k is equal to $(0, \cdots, 0) \in \mathbb{Z}_+^{2^k}$.

Probabilistic model of sequence evolution. We consider a symmetric substitution model on phylogenies, also known as the **Cavender-Farris (CF) model** [Far73, Cav78], for binary sequences of fixed length m. The CF model on a single edge of a tree is a continuous-time Markov process with state space $\{0,1\}^m$ such that (i) the m digits are independent and (ii) each of the m digits follows a continuous-time Markov process with two states $\{0,1\}$ that switches state at rate $\alpha \in (0, \infty)$.

We are interested in this process on a rooted tree T, i.e., indexed by all points on T. The root vertex ρ is assigned a state $\mathcal{X}_\rho \in \{0,1\}^m$, drawn from the uniform distribution on $\{0,1\}^m$. This state then evolves down the tree (away from the root) according to the following recursive process. Moving away from the root, along each edge $e = (u,v)$ starting at u, we run the CF process for a time $\ell_{(u,v)}$ with initial state $\mathcal{X}_u$, described in the previous paragraph. Such processes along different edges starting at u are conditionally independent, given $\mathcal{X}_u$. Denote by $\mathcal{X}_t$ the resulting state at $t \in e$. Then the full process, denoted by $\{\mathcal{X}_t\}_{t\in T}$, is called the **CF model on tree** T. In particular, the set of leaf states is $\mathcal{X}_{\partial T} = \{\mathcal{X}_v : v \in \partial T\}$. It is clear that, under this process, the m digits remain independent. For more background on the CF model, see e.g. [Ste16].

An impossibility result. Our main result is the following. Recall that the total variation distance between two probability measures ν_1 and ν_2 on a countable space E is defined by

$$\|\nu_1 - \nu_2\|_{\mathrm{TV}} = \sup_{A\subseteq E} |\nu_1(A) - \nu_2(A)| . \tag{1}$$

Theorem 1. *Fix $k \in \mathbb{N}$. For any $n \geq 3$, there exists distinct trees $T_1 \neq T_2$ with n leaves such that*

$$\sup_{m\geq k} \|\mathcal{L}_m^{(1)} - \mathcal{L}_m^{(2)}\|_{\mathrm{TV}} < 1, \tag{2}$$

where $\mathcal{L}_m^{(i)}$ is the law of the k-mer frequencies of the leaf sequences of length m under the CF model on tree T_i. Furthermore, the trees $\{T_1, T_2\}$ can be chosen to be independent of k.

From (1), we see that (2) implies the following: using only the k-mer frequencies of the leaf sequences for a fixed $k \geq 1$, there is no statistical test that can distinguish between T_1 and T_2 with probability going to 1 as the sequence length tends to $+\infty$. More precisely, by (2) and the reconstruction upper bound in part 1 of [FR18, Lemma 2], there exists $\epsilon > 0$ such that the probability that a tree estimator gives the correct estimate is at most $1 - \epsilon$, uniformly for all estimators and all integers $m \geq k$.

Proof sketch. Since our goal is to prove a negative result, we get to pick the trees. We consider two trees $\{T_1, T_2\}$ that have the same set of n leaves and are the same except for the edge that connects a single leaf A. These trees are depicted in Figure 1 and described in detail in Section 3.1 below. The topologies of $\{T_1, T_2\}$ differ only on the subtree containing three leaves $\{A, B, C\}$ that have the same distance from the root.

We seek to distinguish the law of the k-mer frequencies of the n leaf sequences between the two trees. This will be done in two steps, in Sections 3.2 and 3.3 respectively, and concluded in Section 3.4.

1. *Step 1 (Reductions):* By using the Markov property of the CF process on trees, we first reduce the problem from n leaf sequences to only 3 sequences (Lemma 1). We can assume the sequence length m is a multiple of k (Lemma 2). Then k-mer frequencies are functions of consecutive non-overlapping k-mer pairs, together with the first and the last k-mers (Lemma 3). Hence we can further reduce the problem to distinguishing the law of non-overlapping k-mer pairs (Lemma 4). The non-overlapping k-mer pairs satisfy certain algebraic relations (Lemma 5), which leads to algebraic redundancy that we need to get rid of (Lemma 6). Summing up, the problem is reduced to distinguishing the law of non-redundant, non-overlapping k-mer pairs on three points $\{A,\ B',\ C'\}$ as depicted in Figure 3 (Lemma 7).

2. *Step 2 (Applying a local CLT):* We apply a local central limit theorem for i.i.d. vectors to the law of non-redundant, non-overlapping k-mer pairs as $m \to \infty$ (Lemmas 8 and 9 and Theorem 2). Non-redundancy guarantees the non-degeneracy of the limit distribution (Lemmas 10, 11, 12 and 13). The two limit normal distributions, under the two trees respectively, have significant overlap (Lemmas 14 and 15) and therefore the laws of non-redundant, non-overlapping k-mer pairs cannot be distinguished with probability going to 1.

Section 3.4 concludes the proof of Theorem 1.

3 Proof

In this section, we give the details of the proof. Some standard results are stated in the appendix.

3.1 The two trees

We construct two trees $\{T_1, T_2\}$ such that

1. T_1 and T_2 have the same set of $n \geq 3$ leaves $\{A, B, C, X^4, ..., X^n\}$.
2. The (weighted) subtree of T_1 restricted to the $n-1$ leaves $\{B, C, X^4, ..., X^n\}$ is the same as that of T_2.
3. The subtrees of T_1 and T_2 below the most recent common ancestor (MRCA) of $\{A, B, C\}$ contain none of $\{X^4, ..., X^n\}$.
4. Leaves $\{A, B, C\}$ satisfy, for $i \in \{1, 2\}$,

$$\begin{aligned} \mathrm{dist}_{T_i}(\rho, C) &= \mathrm{dist}_{T_i}(\rho, B) \quad \text{and} \\ \mathrm{dist}_{T_1}(A, C) &= \mathrm{dist}_{T_2}(A, B) < \mathrm{dist}_{T_i}(B, C), \end{aligned} \tag{3}$$

   where dist_{T_i} denotes the distance on tree T_i, and ρ is the root vertex.

These trees are depicted in Figure 1, where $X = (X^4, ..., X^n)$ refers to the set of all leaves other than $\{A, B, C\}$. In *Newick tree format* (see, e.g., [War17]), the topology of T_1 restricted to $\{A, B, C\}$ is $((A, C), B)$, while the topology of T_2 restricted to $\{A, B, C\}$ is $((A, B), C)$. Clearly, $\{T_1, T_2\}$ does not depend on k, and their topologies differ only on the subtree containing three leaves $\{A, B, C\}$. The topology of the trees restricted to X is arbitrary and plays no role in the argument.

Notation. For $i \in \{1, 2\}$, we let $\mathbb{P}^{(i)} = \mathbb{P}^{(i),m}$ be the probability measure of the CF model on T_i with sequence length m and $\mathbb{P}^{(i)}_\Theta$ be the law of a random variable Θ under $\mathbb{P}^{(i)}$. For a binary sequence $\sigma_Z = \sigma(Z)$ at a point $Z \in T_i$ where $i \in \{1, 2\}$, we let $f_Z := f_{\sigma(Z)} \in \mathbb{Z}_+^{2^k}$ be the k-mer count vector in $\sigma(Z)$ (see Definition 1). For a finite ordered set of points $U = (u_j)$ on the tree T_i, we let $f_U = (f_{u_j}) \in \mathbb{Z}_+^{2^k \times |U|}$. With this notation, in Theorem 1, $\mathcal{L}_m^{(i)} = \mathbb{P}^{(i)}_{f_X, f_A, f_B, f_C}$. We also write $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$.

3.2 Reductions

Our argument proceeds through a series of reductions.

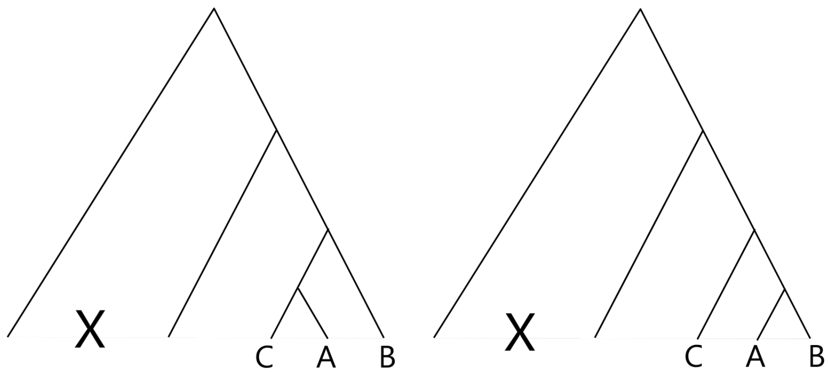

Figure 1: The trees T_1 (left) and T_2 (right) on n leaves. Here X refers to a set of $n-3$ leaves.

3.2.1 Reduction to three vertices

First we shall reduce the complexity of the problem from n to three vertices. For this we define two internal vertices B' and C' on both T_1 and T_2 as follows. Let C' be the MRCA of A and C in T_1 and label C' as well the point on T_2 with the same location on the path between C and B. Similarly, we let B' be the MRCA of A and B in T_2 and label B' the corresponding point on T_1. This setup is depicted in Figure 2.

We can now state our first reduction lemma. The proof of this lemma uses simple information-theoretic inequalities which are stated in the appendix.

Lemma 1 (Reduction to 3 vertices)**.** *Let T_1 and T_2 be the modified trees with points C' and B' as described above. Then for all $m \in \mathbb{N}$,*

$$\|\mathcal{L}_m^{(1)} - \mathcal{L}_m^{(2)}\|_{\mathrm{TV}} \leq \|\mathbb{P}^{(1)}_{f_A, f_{B'}, f_{C'}} - \mathbb{P}^{(2)}_{f_A, f_{B'}, f_{C'}}\|_{\mathrm{TV}}.$$

Proof. First, (f_X, f_A, f_B, f_C) is of course a function of $(f_X, f_A, f_B, f_C, f_{B'}, f_{C'})$, so Lemma 19 in the appendix implies

$$\begin{aligned}\|\mathcal{L}_m^{(1)} - \mathcal{L}_m^{(2)}\|_{\mathrm{TV}} &= \|\mathbb{P}^{(1)}_{f_X, f_A, f_B, f_C} - \mathbb{P}^{(2)}_{f_X, f_A, f_B, f_C}\|_{\mathrm{TV}} \\ &\leq \|\mathbb{P}^{(1)}_{f_X, f_A, f_B, f_C, f_{B'}, f_{C'}} - \mathbb{P}^{(2)}_{f_X, f_A, f_B, f_C, f_{B'}, f_{C'}}\|_{\mathrm{TV}}.\end{aligned}$$

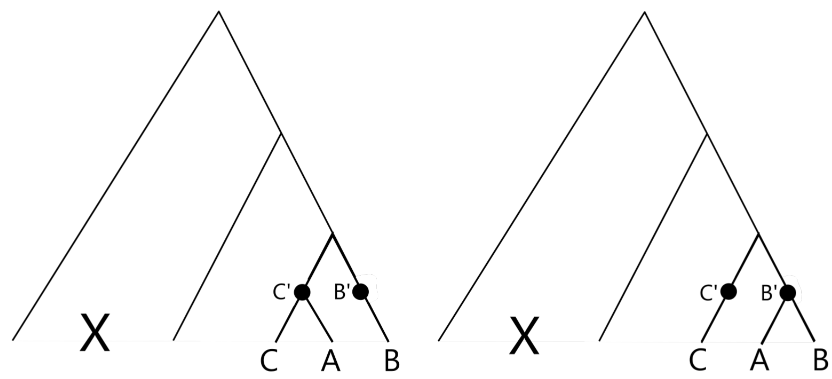

Figure 2: The trees T_1 (left) and T_2 (right) on n leaves with points C' and B' added.

Also $f_{X,C,B} \to f_{B'C'} \to f_A$ forms a Markov chain under both $\mathbb{P}^{(1)}$ and $\mathbb{P}^{(2)}$, satisfying all the conditions of Lemma 20 in the appendix. Hence

$$\|\mathbb{P}^{(1)}_{f_X,f_A,f_B,f_C,f_{B'},f_{C'}} - \mathbb{P}^{(2)}_{f_X,f_A,f_B,f_C,f_{B'},f_{C'}}\|_{\mathrm{TV}} = \|\mathbb{P}^{(1)}_{f_A,f_{B'},f_{C'}} - \mathbb{P}^{(2)}_{f_A,f_{B'},f_{C'}}\|_{\mathrm{TV}},$$

giving the result. □

In words, to sum up this first step, we have reduced the problem to one of distinguishing between the two three-vertex trees depicted in Figure 3.

3.2.2 Reduction to non-overlapping transitions

Due to the following lemma, we can assume $m = (\mu + 1)k$ for some $\mu \in \mathbb{N}$.

Lemma 2 (Reduction to multiples of k)**.** *If $\bar{\mu}k < m < (\bar{\mu}+1)k$ where $\bar{\mu} \in \mathbb{N}$, then*

$$\|\mathbb{P}^{(1),m}_{f_A,f_{B'},f_{C'}} - \mathbb{P}^{(2),m}_{f_A,f_{B'},f_{C'}}\|_{\mathrm{TV}} \le \|\mathbb{P}^{(1),\,(\bar{\mu}+1)k}_{\widehat{f}_A,\widehat{f}_{B'},\widehat{f}_{C'}} - \mathbb{P}^{(2),\,(\bar{\mu}+1)k}_{\widehat{f}_A,\widehat{f}_{B'},\widehat{f}_{C'}}\|_{\mathrm{TV}},$$

where $\widehat{f}_V = (f_V,\, \sigma_V^{last})$ is the k-mer count vector together with the last $2k$-digits σ_V^{last} in σ_V.

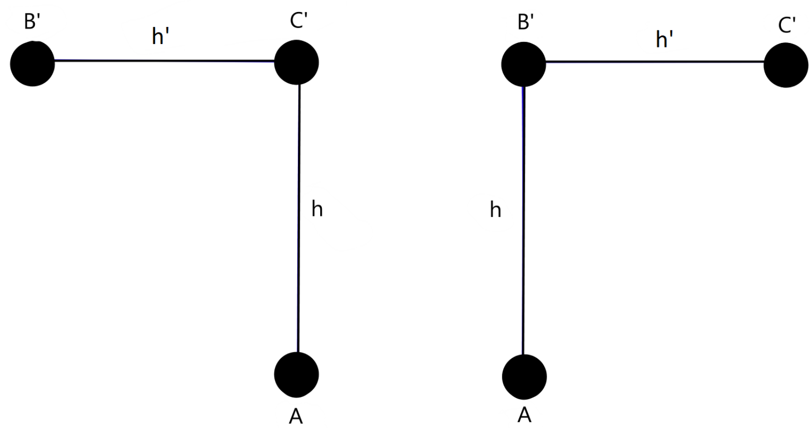

Figure 3: The three-vertex configurations for the measures $\mathbb{P}^{(1)}$ and $\mathbb{P}^{(2)}$ respectively.

Proof. Note that all digits are independent under the CF model and

$$\|\mathbb{P}^{(1),m}_{f_A,f_{B'},f_{C'}} - \mathbb{P}^{(2),m}_{f_A,f_{B'},f_{C'}}\|_{\mathrm{TV}} = \|\mathbb{P}^{(1),\,(\bar{\mu}+1)k}_{f_A^m,f_{B'}^m,f_{C'}^m} - \mathbb{P}^{(2),\,(\bar{\mu}+1)k}_{f_A^m,f_{B'}^m,f_{C'}^m}\|_{\mathrm{TV}},$$

where f_V^m is the k-mer count vector of the first m digits of σ_V. The proof is complete by Lemma 19 since f_V^m is a function of f_V and the last $2k$-digits when σ_V has length $(\bar{\mu}+1)k$. □

For $\sigma \in \{0,1\}^m$ where $m = (\mu+1)k$, we let $x_0^\sigma, \dots, x_\mu^\sigma \in \{0,1\}^k$ be the consecutive *non-overlapping* k-mers in σ. That is,

$$\sigma = \underbrace{(\sigma_1, \dots, \sigma_k)}_{x_0^\sigma}\,\underbrace{(\sigma_{k+1}, \dots, \sigma_{2k})}_{x_1^\sigma} \cdots \underbrace{(\sigma_{\mu k+1}, \dots, \sigma_{(\mu+1)k})}_{x_\mu^\sigma} \in \{0,1\}^{(\mu+1)k}. \quad (4)$$

For $y, z \in \{0,1\}^k$, let $N_{y,z}^\sigma$ be the number of consecutive (y,z) pairs in this representation of σ:

$$N_{y,z}^\sigma = \sum_{j=0}^{\mu-1} \mathbf{1}\{x_j^\sigma = y,\, x_{j+1}^\sigma = z\}. \quad (5)$$

We call $N_{y,z}^\sigma$ the number of non-overlapping transitions from y to z.

The following lemma and its proof give an expression for k-mer frequencies in terms of the numbers of consecutive non-overlapping k-mer pairs as well as the ending k-mers. The reason it holds is that each transition from a k-mer y to a k-mer z accounts for a series of k-mers in between, not counting the starting y. Adding these transitions gives the k-mer frequencies.

Lemma 3 (k-mers as a function of non-overlapping transitions)**.** *For any $\sigma \in \{0,1\}^{(\mu+1)k}$ and $\mu \in \mathbb{N}$, the frequency vector f_σ is a function of*

$$\big(x_\mu^\sigma, (N_{y,z}^\sigma)_{y,z\in\{0,1\}^k}\big).$$

Proof. We split the set $\{0,1,\ldots,\mu k\}$ into the disjoint union $\big(\cup_{a=0}^{k-1}\Lambda_a\big)\cup\{\mu k\}$, where $\Lambda_a = \{a, k+a, 2k+a, \cdots, (\mu-1)k+a\}$ contains μ integers with remainder a when divided by k. By definition, for $w = (w_1,\ldots,w_k) \in \{0,1\}^k$,

$$\begin{aligned} f_\sigma(w) &= \sum_{i=0}^{\mu k} \mathbf{1}\{(\sigma_{i+1},\ldots,\sigma_{i+k}) = w\} \\ &= \mathbf{1}\{x_\mu^\sigma = w\} + \sum_{a=0}^{k-1}\sum_{i\in\Lambda_a} \mathbf{1}\{(\sigma_{i+1},\ldots,\sigma_{i+k}) = w\}. \end{aligned} \tag{6}$$

For $a = 0$, the set Λ_0 coincides with the multiples of k from 0 up to $\mu - 1$. So

$$\sum_{i\in\Lambda_0} \mathbf{1}\{(\sigma_{i+1},\ldots,\sigma_{i+k}) = w\} = \sum_{i=0}^{\mu-1} \mathbf{1}\{x_i^\sigma = w\} = \sum_{z\in\{0,1\}^k} N_{w,z}^\sigma. \tag{7}$$

Similarly, for $a = 1$, we consider

$$\begin{aligned} \sigma =& (\sigma_1, \underbrace{\sigma_2,\ldots,\sigma_k)\,(\sigma_{k+1}}, \underbrace{\sigma_{k+2},\ldots,\sigma_{2k})\,(\sigma_{2k+1}},\ldots,\sigma_{3k})\,\cdots \\ & \cdots(\sigma_{(\mu-1)k+1}, \underbrace{\ldots,\sigma_{\mu k})\,(\sigma_{\mu k+1}},\ldots,\sigma_{(\mu+1)k}). \end{aligned}$$

Let $\Theta_1(w)$ be the set of all pairs of the form

$$\big((\theta_0,\, w_1,\cdots,w_{k-1}),\, (w_k,\theta_1,\theta_2,\cdots,\theta_{k-1})\big) \in \{0,1\}^{2k},$$

where $(\theta_0,\cdots,\theta_{k-1})$ is an arbitrary element in $\{0,1\}^k$. Then

$$\sum_{i\in\Lambda_1} \mathbf{1}\{(\sigma_{i+1},\ldots,\sigma_{i+k}) = w\} = \sum_{(y,z)\in\Theta_1(w)} N_{y,z}^\sigma.$$

More generally for $a \in \{1, \cdots, k-1\}$,

$$\sum_{i \in \Lambda_a} \mathbf{1}\{(\sigma_{i+1}, \ldots, \sigma_{i+k}) = w\} = \sum_{(y,z) \in \Theta_a(w)} N^{\sigma}_{y,z}, \tag{8}$$

where $\Theta_a(w)$ is the set of all pairs of the form

$$\big((\theta_0, \cdots, \theta_{a-1},\, w_1, \cdots, w_{k-a}),\, (w_{k-a+1}, \cdots, w_k,\, \theta_a, \cdots, \theta_{k-1})\big) \in \{0,1\}^{2k},$$

where $(\theta_0, \cdots, \theta_{k-1})$ is an arbitrary element in $\{0,1\}^k$.

The result then follows when we put (7) and (8) into (6), seeing that $f_\sigma(w)$ depends on the specified value of $\Big(x^{\sigma}_{\mu}, \big(N^{\sigma}_{y,z}\big)_{y,z \in \{0,1\}^k}\Big)$. This completes the proof of the lemma. □

For points $V \in \{A, B', C'\}$ on the trees we let

$$Z_V = \Big((x^{\sigma_V}_0, x^{\sigma_V}_1),\, (x^{\sigma_V}_{\mu-1}, x^{\sigma_V}_{\mu}),\, \big(N^{\sigma_V}_{y,z}\big)_{y,z \in \{0,1\}^k}\Big),$$

where σ_V is the binary sequence at V. Note that we included $x^{\sigma_V}_0$, $x^{\sigma_V}_1$ and $x^{\sigma_V}_{\mu-1}$ here for reasons that will become clear below.

Lemma 4 (Reduction to non-overlapping transitions)**.** *Suppose $m = (\mu+1)k$ for some $\mu \in \mathbb{N}$. Then*

$$\|\mathbb{P}^{(1)}_{\widehat{f}_A, \widehat{f}_{B'}, \widehat{f}_{C'}} - \mathbb{P}^{(2)}_{\widehat{f}_A, \widehat{f}_{B'}, \widehat{f}_{C'}}\|_{\mathrm{TV}} \le \|\mathbb{P}^{(1)}_{Z_A, Z_{B'}, Z_{C'}} - \mathbb{P}^{(2)}_{Z_A, Z_{B'}, Z_{C'}}\|_{\mathrm{TV}}.$$

Proof. Recall the definition of $\widehat{f}_V = (f_V,\, \sigma^{last}_V)$ in Lemma 2. By Lemma 3, $(\widehat{f}_A, \widehat{f}_{B'}, \widehat{f}_{C'})$ is a function of $(Z_A, Z_{B'}, Z_{C'})$. Applying Lemma 19 directly gives the result. □

3.2.3 Dealing with the algebraic redundancy

The quantities $\{N^{\sigma}_{y,z}\}_{y,z \in \{0,1\}^k}$ satisfy certain algebraic relations described in Lemma 5 below. We will get rid of these redundancies in Lemma 6, which will be needed for a non-degeneracy condition in the local CLT; see Lemma 10 below.

Lemma 5 (Combinatorial constraints)**.** *For any $\sigma \in \{0,1\}^{(\mu+1)k}$ and $\mu \in \mathbb{N}$,*

$$\mathbf{1}\{x^{\sigma}_0 = z\} + \sum_{y \in \{0,1\}^k :\, y \ne z} N^{\sigma}_{y,z} \;=\; \mathbf{1}\{x^{\sigma}_{\mu} = z\} + \sum_{y' \in \{0,1\}^k :\, y' \ne z} N^{\sigma}_{z,y'} \tag{9}$$

for all $z \in \{0,1\}^k$. Moreover the overall sum satisfies

$$\sum_{y,z \in \{0,1\}^k} N^{\sigma}_{y,z} = \mu. \tag{10}$$

Proof. Equation (10) holds since the total number of non-overlapping transitions in σ is μ.

We verify (9) by induction on $\mu \geq 1$. In the $\mu = 1$ case, we have

$$\begin{aligned}
&\mathbf{1}\{x^{\sigma}_0 = z\} + \sum_{y \in \{0,1\}^k : y \neq z} \mathbf{1}\{x^{\sigma}_0 = y, x^{\sigma}_1 = z\} \\
&= \left(\mathbf{1}\{x^{\sigma}_0 = z, x^{\sigma}_1 = z\} + \sum_{y' \in \{0,1\}^k : y' \neq z} \mathbf{1}\{x^{\sigma}_0 = z, x^{\sigma}_1 = y'\} \right) \\
&\quad + \sum_{y \in \{0,1\}^k : y \neq z} \mathbf{1}\{x^{\sigma}_0 = y, x^{\sigma}_1 = z\}.
\end{aligned}$$

The first and third terms on the right hand side of the equality add up to $\mathbf{1}\{x^{\sigma}_1 = z\}$, implying the result for $\mu = 1$.

For the rest of the proof, we assume that the μth case holds and we show the $(\mu+1)$st case follows. The left hand side of (9) for the $(\mu+1)$th case is

$$\begin{aligned}
&\mathbf{1}\{x^{\sigma}_0 = z\} + \sum_{y \in \{0,1\}^k : y \neq z} \sum_{j=0}^{\mu} \mathbf{1}\{x^{\sigma}_j = y, x^{\sigma}_{j+1} = z\} \\
&= \mathbf{1}\{x^{\sigma}_0 = z\} + \sum_{y \in \{0,1\}^k : y \neq z} \left(\sum_{j=0}^{\mu-1} \mathbf{1}\{x^{\sigma}_j = y, x^{\sigma}_{j+1} = z\} + \mathbf{1}\{x^{\sigma}_{\mu} = y, x^{\sigma}_{\mu+1} = z\} \right) \\
&= \left(\mathbf{1}\{x^{\sigma}_0 = z\} + \sum_{y \in \{0,1\}^k : y \neq z} \sum_{j=0}^{\mu-1} \mathbf{1}\{x^{\sigma}_j = y, x^{\sigma}_{j+1} = z\} \right) \\
&\quad + \sum_{y \in \{0,1\}^k : y \neq z} \mathbf{1}\{x^{\sigma}_{\mu} = y, x^{\sigma}_{\mu+1} = z\} \\
&= \left(\mathbf{1}\{x^{\sigma}_{\mu} = z\} + \sum_{y' \in \{0,1\}^k : y' \neq z} \sum_{j=0}^{\mu-1} \mathbf{1}\{x^{\sigma}_j = z, x^{\sigma}_{j+1} = y'\} \right) \\
&\quad + \sum_{y \in \{0,1\}^k : y \neq z} \mathbf{1}\{x^{\sigma}_{\mu} = y, x^{\sigma}_{\mu+1} = z\},
\end{aligned} \tag{11}$$

where we used the induction assumption in the last equality. Now we decompose the first term of (11) as

$$\mathbf{1}\{x^\sigma_\mu = z\} = \sum_{y' \in \{0,1\}^k : y' \neq z} \mathbf{1}\{x^\sigma_\mu = z, x^\sigma_{\mu+1} = y'\} + \mathbf{1}\{x^\sigma_\mu = z, x^\sigma_{\mu+1} = z\}.$$

Combining the first term on the right hand side with the y'-sum in (11) and the second term on the right hand side with the y-sum in (11), we obtain (9) in the $(\mu+1)$st case.

This verifies (9) for the $(\mu+1)$th case, completing the proof for the induction claim and the lemma. □

There are actually only 2^k linearly independent equations among the $2^k + 1$ equations in (9)–(10), as can be seen from the proof of Lemma 6 below. To ensure a non-degenerate limit when applying the central limit theorem, we utilize these 2^k linearly independent equations to remove 2^k redundant variables. Specifically, we remove the transition counts corresponding to the pairs $\{(\vec{1}, z) : z \in \{0,1\}^k\}$, where $\vec{1} = (1, \cdots, 1) \in \{0,1\}^k$ is the all-1 string.

Lemma 6 (Algebraic redundancy)**.** *For any $\sigma \in \{0,1\}^{(\mu+1)k}$ and $\mu \in \mathbb{N}$, $\big(x^\sigma_0, x^\sigma_\mu, (N^\sigma_{y,z})_{y,z\in\{0,1\}^k}\big)$ is a function of $\big(x^\sigma_0, x^\sigma_\mu, (N^\sigma_{y,z})_{(y,z)\in\mathcal{H}}\big)$, where*

$$\mathcal{H} = \Big\{(y,z) \in \{0,1\}^k \times \{0,1\}^k : y \neq \vec{1}\Big\}.$$

Proof. It suffices to show that for any $(y,z) \notin \mathcal{H}$, we can write $N^\sigma_{y,z}$ as a function of $x^\sigma_0, x^\sigma_\mu, \mu$, and $(N^\sigma_{y,z})_{(y,z)\in\mathcal{H}}$. We do this first for $N^\sigma_{\vec{1},z}$ where $z \neq \vec{1}$, and then for $N^\sigma_{\vec{1},\vec{1}}$.

Among the 2^k equations in (9), each one indexed by $z \neq \vec{1}$ has exactly one variable in $\mathcal{H}^c$, namely $N^\sigma_{\vec{1},z}$. Precisely, (9) gives

$$N^\sigma_{\vec{1},z} = \mathbf{1}\{x^\sigma_\mu = z\} - \mathbf{1}\{x^\sigma_0 = z\} + \sum_{y' \neq z} N^\sigma_{z,y'} - \sum_{y \neq z, \vec{1}} N^\sigma_{y,z}, \tag{12}$$

in which all terms on the right come from $\mathcal{H}$. Hence $N^\sigma_{\vec{1},z}$ can be written as a function of the required variables for each $z \neq \vec{1}$.

The variable $N^\sigma_{\vec{1},\vec{1}}$ is featured only in equation (10), and we obtain

$$N^\sigma_{\vec{1},\vec{1}} = \mu - \sum_{(y,z) \neq (\vec{1},\vec{1})} N^\sigma_{y,z}. \tag{13}$$

This completes the proof. □

For points $V \in \{A, B', C'\}$ on the trees, we let

$$Z'_V = \big((x_0^{\sigma_V}, x_1^{\sigma_V}),\, (x_{\mu-1}^{\sigma_V}, x_\mu^{\sigma_V}),\, N_{\mathcal{H}}^{\sigma_V}\big) \quad \text{where} \quad N_{\mathcal{H}}^{\sigma_V} = (N_{y,z}^{\sigma_V})_{(y,z)\in\mathcal{H}}. \tag{14}$$

Lemma 7 (Reduction to non-redundant transitions)**.** *Suppose $m = (\mu+1)k$ for some $\mu \in \mathbb{N}$. Then*

$$\|\mathbb{P}^{(1)}_{Z_A,Z_{B'},Z_{C'}} - \mathbb{P}^{(2)}_{Z_A,Z_{B'},Z_{C'}}\|_{\mathrm{TV}} \le \|\mathbb{P}^{(1)}_{Z'_A,Z'_{B'},Z'_{C'}} - \mathbb{P}^{(2)}_{Z'_A,Z'_{B'},Z'_{C'}}\|_{\mathrm{TV}}.$$

Proof. From Lemma 6, $(Z_A, Z_{B'}, Z_{C'})$ is a function of $(Z'_A, Z'_{B'}, Z'_{C'})$. Then the result follows from Lemma 19. □

3.2.4 Summing up the reduction

To sum up the reduction, by Lemmas 1, 2, 4 and 7 above, together with the second equality of Lemma 18 in the appendix, it suffices to prove that

$$\inf_{\mu\in\mathbb{N}} \sum_{z'_A, z'_{B'}, z'_{C'}} \mathbb{P}^{(1)}_{Z'_A,Z'_{B'},Z'_{C'}}(z'_A, z'_{B'}, z'_{C'}) \wedge \mathbb{P}^{(2)}_{Z'_A,Z'_{B'},Z'_{C'}}(z'_A, z'_{B'}, z'_{C'}) > 0, \tag{15}$$

where the sum is taken over the set $\big(\{0,1\}^{2k} \times \{0,1\}^{2k} \times \{0,1,\cdots,\mu\}^{\mathcal{H}}\big)^3$, and $m = (\mu+1)k$.

Our final reduction step in this section is to condition on the event

$$\widetilde{\mathcal{E}} = \Big\{(x_0^A, x_1^A) = (x_0^{B'}, x_1^{B'}) = (x_0^{C'}, x_1^{C'}) = (\vec{0}, \vec{0})\Big\}, \tag{16}$$

where $x_j^V \in \{0,1\}^k$ are the non-overlapping k-mers in the sequence σ_V at point $V \in \{A, B', C'\}$, defined in (4). Precisely, for $i \in \{1,2\}$ we let $\widetilde{\mathbb{P}}^{(i)} = \widetilde{\mathbb{P}}^{(i),m}$ be the conditional measures under $\mathbb{P}^{(i)} = \mathbb{P}^{(i),m}$ given the event $\widetilde{\mathcal{E}}$. Then

$$\begin{aligned}
&\sum_{z'_A, z'_{B'}, z'_{C'}} \mathbb{P}^{(1)}_{Z'_A,Z'_{B'},Z'_{C'}}(z'_A, z'_{B'}, z'_{C'}) \wedge \mathbb{P}^{(2)}_{Z'_A,Z'_{B'},Z'_{C'}}(z'_A, z'_{B'}, z'_{C'}) \\
&= \sum_{z'_A, z'_{B'}, z'_{C'}} \Big(\widetilde{\mathbb{P}}^{(1)}_{Z'_A,Z'_{B'},Z'_{C'}}(z'_A, z'_{B'}, z'_{C'})\, \mathbb{P}^{(1)}[\widetilde{\mathcal{E}}]\Big) \wedge \Big(\widetilde{\mathbb{P}}^{(2)}_{Z'_A,Z'_{B'},Z'_{C'}}(z'_A, z'_{B'}, z'_{C'})\, \mathbb{P}^{(2)}[\widetilde{\mathcal{E}}]\Big) \\
&\ge\; c_1 \sum_{z'_A, z'_{B'}, z'_{C'}} \widetilde{\mathbb{P}}^{(1)}_{Z'_A,Z'_{B'},Z'_{C'}}(z'_A, z'_{B'}, z'_{C'}) \wedge \widetilde{\mathbb{P}}^{(2)}_{Z'_A,Z'_{B'},Z'_{C'}}(z'_A, z'_{B'}, z'_{C'}),
\end{aligned}$$

where $c_1 := \mathbb{P}^{(1)}[\widetilde{\mathcal{E}}] \wedge \mathbb{P}^{(2)}[\widetilde{\mathcal{E}}]$ is positive and does not depend on μ.

Hence, to show (15) it suffices to prove that

$$\inf_{\mu\in\mathbb{Z}_+} \sum_{(z'_A,z'_{B'},z'_{C'})\in(\mathcal{S}_0^\mu)^3} \widetilde{\mathbb{P}}^{(1)}_{Z'_A,Z'_{B'},Z'_{C'}}(z'_A,z'_{B'},z'_{C'}) \wedge \widetilde{\mathbb{P}}^{(2)}_{Z'_A,Z'_{B'},Z'_{C'}}(z'_A,z'_{B'},z'_{C'}) > 0, \tag{17}$$

where $\mathcal{S}_0^\mu := \{(\vec{0},\vec{0})\} \times \{0,1\}^{2k} \times \{0,1,\cdots,\mu\}^{\mathcal{H}}$.

For the rest of the proof, we shall establish (17) by obtaining suitable lower bounds on the probabilities in (17) through a local limit theorem.

3.3 Applying a local limit theorem

It will be convenient to consider infinite sequences, since we shall employ a local limit theorem as the sequence length tends to infinity (i.e., $\mu \to \infty$). Let $\mathbb{P}^{(i),\infty}$ be the probability measure of the CF model on T_i with infinite sequence length and $\widetilde{\mathbb{P}}^{(i),\infty}$ be the conditional measure under $\mathbb{P}^{(i),\infty}$, given the event $\widetilde{\mathcal{E}}$.

3.3.1 Pairs of triplets as a Markov chain

We shall apply Doeblin's method (see e.g. [Cul61]). For $V \in \{A, B', C'\}$, we let $\sigma_V(n) = (x_0^V, \ldots, x_n^V) \in \{0,1\}^{k(n+1)}$ be the first $n+1$ non-overlapping k-mers of σ_V, where $0 \le n \le \mu$ if σ_V has length $(\mu+1)k$ and $n \in \mathbb{Z}_+$ if $\sigma_V \in \{0,1\}^{\mathbb{N}}$ has infinite length. For all such n, we consider the triples

$$\vec{X}_n = (x_n^A,\, x_n^{B'},\, x_n^{C'}) \in \{0,1\}^{3k}. \tag{18}$$

Under $\widetilde{\mathbb{P}}^{(i),\infty}$, $\{\vec{X}_n\}_{n\in\mathbb{Z}_+}$ is a sequence of independent variables and the pairs $\vec{M}_n = (\vec{X}_n,\, \vec{X}_{n+1})$ form a Markov chain with a finite state space. This Markov chain is irreducible since the support of $\vec{X}_n$ is all of $\{0,1\}^{3k}$. Let $\tau_0 = 0$, let τ_1 be the first n such that $\vec{M}_n = (\vec{0},\vec{0})$ and in general, for $\ell \ge 1$, we define

$$\tau_\ell = \inf\{n > \tau_{\ell-1} : \vec{M}_n = (\vec{0},\vec{0})\}, \tag{19}$$

where an infimum over an empty set is $+\infty$ by convention.

The connection between $\widetilde{\mathbb{P}}^{(i)}$ and $\widetilde{\mathbb{P}}^{(i),\infty}$ that we will need is given by Lemma 8 below. We let

$$N_{y,z}^V(n) = N_{y,z}^{\sigma_V(n)} = \sum_{j=0}^{n-1} \mathbf{1}\{x_j^V = y,\, x_{j+1}^V = z\}$$

be the number of non-overlapping transitions from y to z up to x_n^V, as in (5), with the convention that $N_{y,z}^V(0) = 0$.

Lemma 8 (Infinite sequences). *For all $\mu \in \mathbb{Z}_+$, $k \in \mathbb{N}$, $\ell \in \{1, 2, \cdots, \mu\}$, $(a, b, c) \in \mathbb{Z}_+^3$ and $i \in \{1, 2\}$, the event*

$$\Big\{\tau_\ell = \mu\ ,\ \Big(N_{\mathcal{H}}^A(\tau_\ell),\, N_{\mathcal{H}}^{B'}(\tau_\ell),\, N_{\mathcal{H}}^{C'}(\tau_\ell)\Big) = (a, b, c)\Big\}$$

has the same probability under $\widetilde{\mathbb{P}}^{(i),m}$ and $\widetilde{\mathbb{P}}^{(i),\infty}$, where $m = (\mu + 1)k$.

This lemma follows directly from the construction of the CF model in which non-overlapping k-mers are independent. The rest of Section 3.3 concerns infinite sequences.

3.3.2 Independent excursions and a multivariate local CLT

We extract i.i.d. random variables from excursions of the Markov chain $\vec{M}$. Define, for $V \in \{A, B', C'\}$,

$$Y_V(\ell) = N_{\mathcal{H}}^V(\tau_\ell) - N_{\mathcal{H}}^V(\tau_{\ell-1}) = \big(N_{y,z}^V(\tau_\ell) - N_{y,z}^V(\tau_{\ell-1})\big)_{(y,z)\in\mathcal{H}}$$

and let

$$\mathbf{Y}(\ell) = \big(\tau_\ell - \tau_{\ell-1}, Y_A(\ell), Y_{B'}(\ell), Y_{C'}(\ell)\big).$$

Note that these random vectors take values in $\mathbb{N} \times \big(\mathbb{Z}_+^{\mathcal{H}}\big)^3 \subset \mathbb{Z}_+^d$ where $d = 1 + 3(2^{2k} - 2^k)$, because $|\mathcal{H}| = 2^{2k} - 2^k$.

Lemma 9. *The vectors $\{\mathbf{Y}(\ell)\}_{\ell=1}^\infty$ are i.i.d. under $\widetilde{\mathbb{P}}^{(i),\infty}$ for both $i = 1$ and 2. Further, their partial sum is equal to*

$$\sum_{\iota=1}^{\ell} \mathbf{Y}(\iota) = \Big(\tau_\ell,\, N_{\mathcal{H}}^A(\tau_\ell),\, N_{\mathcal{H}}^{B'}(\tau_\ell),\, N_{\mathcal{H}}^{C'}(\tau_\ell)\Big)\,. \tag{20}$$

Proof. The first statement is obvious from the construction of the CF model. The equality (20) follows from the definitions of $\mathbf{Y}$ and the conventions $\tau_0 = N_{y,z}^V(0) = 0$. □

We will apply a multivariate local CLT of Davis and McDonald [DM95, Theorem 2.1] to the i.i.d. vectors $\{\mathbf{Y}(\ell)\}_{\ell=1}^{\infty} \subset \mathbb{Z}_+^d$ under $\widetilde{\mathbb{P}}^{(i),\infty}$. Theorem 2.1 of [DM95] works for an array of independent vectors. Here we need only a sequence of i.i.d. vectors so we state this result for the case of i.i.d. vectors in $\mathbb{Z}^d$.

Theorem 2. [DM95, Theorem 2.1] *Let $\{\mathbf{X}_j\}_{j=1}^{\infty}$ be a sequence of independent $\mathbb{Z}^d$-valued random variables with common probability mass function f with finite mean $\mathbf{m} \in \mathbb{R}^d$ and covariance matrix $\mathbf{\Sigma}$. Suppose the followings hold.*

(a) $q_r(f) := \frac{1}{d}\sum_{\boldsymbol{x}\in\mathbb{Z}^d} f(\boldsymbol{x}) \wedge f(\boldsymbol{x}+\mathbf{e}_r) \in (0,\infty)$ for all $r \in \{1,2,\dots,d\}$, where $\mathbf{e}_r \in \mathbb{Z}^d$ is the unit vector of zeros except for 1 in the r-th position.

(b) The determinant $\det(\mathbf{\Sigma}) \in (0,\infty)$.

(c) Let $\boldsymbol{S}_n = \sum_{j=1}^{n} \boldsymbol{X}_i$. Then $\frac{\boldsymbol{S}_n - n\,\mathbf{m}}{\sqrt{n}}$ converges in distribution to the multivariate normal distribution $\mathcal{N}(\mathbf{0},\, \mathbf{\Sigma})$ as $n \to \infty$.

Then the following uniform convergence holds as $n \to \infty$:

$$\sup_{\mathbf{y}\in\mathbb{Z}^d} \left| n^{d/2}\,\mathbb{P}[\boldsymbol{S}_n = \mathbf{y}] - \varphi\left(\frac{\mathbf{y} - n\,\mathbf{m}}{\sqrt{n}}\right) \right| \to 0,$$

where φ is the probability density function of the multivariate normal distribution $\mathcal{N}(\mathbf{0},\, \mathbf{\Sigma})$.

We apply Theorem 2 to the i.i.d. vectors $\{\mathbf{Y}(\ell)\}_{\ell=1}^{\infty} \subset \mathbb{Z}_+^d$ under $\widetilde{\mathbb{P}}^{(i),\infty}$, for each of $i \in \{1,2\}$. To do this we first verify conditions (a)-(c) of Theorem 2 in Section 3.3.3 below.

3.3.3 Checking conditions of the local CLT

In this section we verify conditions (a)-(c) of Theorem 2 for the i.i.d. vectors $\{\mathbf{Y}(\ell)\}_{\ell=1}^{\infty} \subset \mathbb{Z}_+^d$.

We first check condition (a) of Theorem 2, which holds if

$$\forall r \in \{1,2,\dots,d\},\ \exists\, \mathbf{x}_r \in \mathbb{Z}_+^d \text{ such that } f(\mathbf{x}_r) \wedge f(\mathbf{x}_r + \mathbf{e}_r) > 0. \tag{21}$$

For this we let $f^{(i)}$ be the probability mass function of

$$\mathbf{Y}(1) \,=\, (\tau, Y_A(1), Y_{B'}(1), Y_{C'}(1)) \,=\, \left(\tau,\, N_{\mathcal{H}}^{A}(\tau),\, N_{\mathcal{H}}^{B'}(\tau),\, N_{\mathcal{H}}^{C'}(\tau)\right)$$

under $\widetilde{\mathbb{P}}^{(i),\infty}$ for $i \in \{1,2\}$, where $\tau = \tau_1$ is defined in (19).

Lemma 10 (Non-degeneracy). *The distributions $f^{(i)}$ and $f^{(2)}$ both satisfy condition* (21)*, and hence they also satisfy condition (a) of Theorem 2.*

Proof. The proof relies crucially on the construction of the set $\mathcal{H}$ in Lemma 6. We need to show that for both $i = 1, 2$ and for all $r \in \{1, 2, \cdots, d\}$, there exists $\mathbf{x}_r \in \mathbb{Z}_+^d$ such that $f^{(i)}(\mathbf{x}_r) \wedge f^{(i)}(\mathbf{x}_r + \mathbf{e}_r) > 0$. We write a canonical point in $\mathbb{Z}_+^d$ as

$$\mathbf{x} = \Big(t,\ (n_{yz}^A,\, n_{yz}^{B'},\, n_{yz}^{C'})_{yz\in\mathcal{H}}\Big), \quad \text{where } t \in \mathbb{Z}_+ \text{ and } (n_{yz}^A,\, n_{yz}^{B'},\, n_{yz}^{C'}) \in \mathbb{Z}_+^3.$$

Recall that $\vec{0}$ and $\vec{1}$ refer to the all-0 and all-1 k-mers. A cycle of non-overlapping k-mer triples starting and ending with $(\vec{0},\vec{0},\vec{0})\,(\vec{0},\vec{0},\vec{0})$ will give rise to a unique point in $\mathbb{Z}_+^d$, in which t is the length of the cycle and n_{yz}^V counts the number of yz-transitions. By the definition of $\mathcal{H}$, we are not counting the transition from $\vec{1}$ to z for any $z \in \{0,1\}^k$.

(i). For $r = 1$ (corresponding to the t-coordinate), we consider the k-mer triple cycles of

$$\begin{aligned}\mathcal{C} =& (\vec{0},\vec{0},\vec{0}), (\vec{0},\vec{0},\vec{0}), (\vec{1},\vec{1},\vec{1}), (\vec{0},\vec{0},\vec{0}), (\vec{0},\vec{0},\vec{0}) \qquad \text{and}\\ \mathcal{C}^+ =& (\vec{0},\vec{0},\vec{0}), (\vec{0},\vec{0},\vec{0}), (\vec{1},\vec{1},\vec{1}), (\vec{1},\vec{1},\vec{1}), (\vec{0},\vec{0},\vec{0}), (\vec{0},\vec{0},\vec{0}).\end{aligned}$$

They give rise to $\mathbf{x}_r$ and $\mathbf{x}_r + \mathbf{e}_r$ respectively, where we take $\mathbf{x}_r$ to be the point such that $t = 3$ and

$$(n_{yz}^A,\, n_{yz}^{B'},\, n_{yz}^{C'}) = \begin{cases} (2,2,2) & \text{if } (y,z) = (\vec{0},\vec{0}) \\ (1,1,1) & \text{if } (y,z) = (\vec{0},\vec{1}) \\ (0,0,0) & \text{if } (y,z) \in \mathcal{H} \setminus \{(\vec{0},\vec{0}),\ (\vec{0},\vec{1})\}, \end{cases} \tag{22}$$

and $\mathbf{x}_r + \mathbf{e}_r = \big(4,\ (n_{yz}^A,\, n_{yz}^{B'},\, n_{yz}^{C'})_{yz\in\mathcal{H}}\big)$. Recall that

$$\mathcal{H} = \Big\{(y,z) \in \{0,1\}^k \times \{0,1\}^k : y \neq \vec{1}\Big\},$$

so that, in particular, the transitions $(\vec{1},\vec{1})$ are not counted. Then

$$\begin{aligned} f^{(i)}(\mathbf{x}_r) &\geq \widetilde{\mathbb{P}}^{(i),\infty}\Big((\vec{X}_n)_{n=0}^4 = \mathcal{C}\Big) > 0 \qquad \text{and}\\ f^{(i)}(\mathbf{x}_r + \mathbf{e}_r) &\geq \widetilde{\mathbb{P}}^{(i),\infty}\Big((\vec{X}_n)_{n=0}^4 = \mathcal{C}^+\Big) > 0, \end{aligned}$$

where $\vec{X}_n = (x_n^A,\, x_n^{B'},\, x_n^{C'}) \in \{0,1\}^{3k}$ as defined in (18).

(ii). For $r > 1$, we first suppose r corresponds to the coordinate n_{ab}^A where $(a,b) \in \mathcal{H}$, i.e., not in $\{(\vec{1},\, z)\}_{z\in\{0,1\}^k}$. The cycles

$$\begin{aligned}\mathcal{C}_{ab}^A =& (\vec{0},\vec{0},\vec{0}),(\vec{0},\vec{0},\vec{0}),(\vec{1},\vec{1},\vec{1}),(\vec{1},\vec{1},\vec{1}),(b,\vec{1},\vec{1}),(\vec{0},\vec{0},\vec{0}),(\vec{0},\vec{0},\vec{0}) \qquad \text{and}\\ \mathcal{C}_{ab}^{A+} =& (\vec{0},\vec{0},\vec{0}),(\vec{0},\vec{0},\vec{0}),(\vec{1},\vec{1},\vec{1}),(a,\vec{1},\vec{1}),(b,\vec{1},\vec{1}),(\vec{0},\vec{0},\vec{0}),(\vec{0},\vec{0},\vec{0})\end{aligned}$$

give rise to $\mathbf{x}_r$ and $\mathbf{x}_r + \mathbf{e}_r$ respectively, where $\mathbf{x}_r$ is the point on $\mathbb{Z}_+^d$ such that $t = 5$ and (22) holds. Hence both $f^{(i)}(\mathbf{x}_r)$ and $f^{(i)}(\mathbf{x}_r + \mathbf{e}_r)$ are positive, as before.

The proof for coordinates $n_{ab}^{B'}$ is the same, except that we replace $(a,\vec{1},\vec{1})$ by $(\vec{1},a,\vec{1})$ and $(b,\vec{1},\vec{1})$ by $(\vec{1},b,\vec{1})$. The proof for coordinates $n_{ab}^{C'}$ follows similarly. The proof is complete. □

To verify conditions (b) and (c) of Theorem 2, we let $\mathbf{m}^{(i)}$ and $\mathbf{\Sigma}^{(i)}$ be respectively the mean and the covariance matrix of $\mathbf{Y}(1)$ under $\widetilde{\mathbb{P}}^{(i),\infty}$. Due to symmetry between T_1 and T_2, as well as the independence of consecutive non-overlapping k-mers under the CF model, the expectations are the same, as we show formally next.

Lemma 11 (Expectation)**.** $\mathbf{m}^{(1)} = \mathbf{m}^{(2)} \in \mathbb{R}_+^d$*, where* $\mathbf{m}^{(i)} = \widetilde{\mathbb{E}}^{(i),\infty}[\mathbf{Y}(1)]$.

Proof. Recall that $\mathbf{Y}(1) = \big(\tau,\, N_{\mathcal{H}}^A(\tau),\, N_{\mathcal{H}}^{B'}(\tau),\, N_{\mathcal{H}}^{C'}(\tau)\big)$ where $\tau = \tau_1$ is defined in (19). By symmetry (3), we have $\widetilde{\mathbb{P}}^{(1),\infty}_{(\tau,\sigma_A,\sigma_{B'},\sigma_{C'})} = \widetilde{\mathbb{P}}^{(2),\infty}_{(\tau,\sigma_A,\sigma_{C'},\sigma_{B'})}$. Hence $\widetilde{\mathbb{E}}^{(1),\infty}[(\tau, N_{\mathcal{H}}^A(\tau))] = \widetilde{\mathbb{E}}^{(2),\infty}[(\tau, N_{\mathcal{H}}^A(\tau))]$ and $\widetilde{\mathbb{E}}^{(1),\infty}[(N_{\mathcal{H}}^{B'}(\tau), N_{\mathcal{H}}^{C'}(\tau))] = \widetilde{\mathbb{E}}^{(2),\infty}[(N_{\mathcal{H}}^{C'}(\tau), N_{\mathcal{H}}^{B'}(\tau))]$. It remains to show that

$$\widetilde{\mathbb{E}}^{(i),\infty}[N_{\mathcal{H}}^{B'}(\tau)] = \widetilde{\mathbb{E}}^{(i),\infty}[N_{\mathcal{H}}^{C'}(\tau)] \quad \text{for } i \in \{1,2\}. \tag{23}$$

For arbitrary $(y,z) \in \{0,1\}^k \times \{0,1\}^k$, we have

$$N_{(y,z)}^{B'}(\tau) = \sum_{(\vec{y},\vec{z}):\,(y_2,z_2)=(y,z)} \sum_{j=0}^{\tau-1} 1_{\vec{M}_j=(\vec{y},\,\vec{z})}, \tag{24}$$

where the sum is over the set of $(\vec{y},\vec{z})$ with $y_2 = y$ and $z_2 = z$, with $\vec{y} = (y_1,y_2,y_3) \in \{0,1\}^{3k}$ and $\vec{z} = (z_1,z_2,z_3) \in \{0,1\}^{3k}$. A similar equality holds for $N_{(y,z)}^{C'}(\tau)$.

Recall the Markov chain $\vec{M}_j = (\vec{X}_j,\, \vec{X}_{j+1})$, where $\vec{X}_j = (x_j^A,\, x_j^{B'},\, x_j^{C'}) \in \{0,1\}^{3k}$. Under $\widetilde{\mathbb{P}}^{(i),\infty}$, $\{\vec{X}_j\}_{j\in\mathbb{Z}_+}$ are independent and $\vec{M}_j$ is a Markov chain with finite state space. This chain is irreducible since the support of $\vec{X}_j$ for $j \geq 2$ is $\{0,1\}^{3k}$ and the $\vec{X}_j$'s are independent. The stationary distribution $\Theta_{\vec{M}}$ of $\{\vec{M}_j\}_{j\in\mathbb{Z}_+}$ is

$$\Theta_{\vec{M}}(\vec{y},\vec{z}) = \widetilde{\mathbb{P}}^{(i),\infty}(\vec{X}_2 = \vec{y})\, \widetilde{\mathbb{P}}^{(i),\infty}(\vec{X}_2 = \vec{z}), \quad \text{for } \vec{y},\, \vec{z} \in \{0,1\}^{3k}. \tag{25}$$

On other hand, by [Dur19, Chapter 5],

$$\widetilde{\mathbb{E}}^{(i),\infty}\left[\sum_{j=0}^{\tau_1 - 1} 1_{\vec{M}_j = (\vec{y},\,\vec{z})}\right] = \widetilde{c}\,\Theta_{\vec{M}}(\vec{y},\vec{z}), \tag{26}$$

where $\widetilde{c} = \widetilde{\mathbb{E}}^{(i),\infty}[\tau_1] \in (0,\infty)$.

Finally, by (24), (25) and (26), we have

$$\begin{aligned} \widetilde{\mathbb{E}}^{(i),\infty}\left[N_{y,z}^{B'}(\tau)\right] &= \widetilde{c} \sum_{(\vec{y},\vec{z}) \colon (y_2,z_2)=(y,z)} \widetilde{\mathbb{P}}^{(i),\infty}(\vec{X}_2 = \vec{y})\, \widetilde{\mathbb{P}}^{(i),\infty}(\vec{X}_2 = \vec{z}) \\ &= \widetilde{c}\, \widetilde{\mathbb{P}}^{(i),\infty}(x_2^{B'} = y)\, \widetilde{\mathbb{P}}^{(i),\infty}(x_2^{B'} = z) \quad \text{and} \\ \widetilde{\mathbb{E}}^{(i),\infty}\left[N_{y,z}^{C'}(\tau)\right] &= \widetilde{c}\, \widetilde{\mathbb{P}}^{(i),\infty}(x_2^{C'} = y)\, \widetilde{\mathbb{P}}^{(i),\infty}(x_2^{C'} = z). \end{aligned}$$

These two displayed equations are the same since $\widetilde{\mathbb{P}}^{(i),\infty}_{\sigma_{B'}} = \widetilde{\mathbb{P}}^{(i,\infty)}_{\sigma_{C'}}$. The proof of (23) is complete. □

Lemma 12 below says that (21) also implies that the covariance matrix is positive definite.

Lemma 12. *Assume* (21) *holds. Then the covariance matrix of f is positive definite.*

Proof. Let X and Y be two independent variables with distribution f. Then the covariance matrix of X can be written as

$$\mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^T] = (1/2)\mathbb{E}[(X - Y)(X - Y)^T].$$

Then for any nonzero vector $z \neq 0$ with, say, $z_r \neq 0$, we have

$$\begin{aligned} z^T \mathbb{E}[(X - Y)(X - Y)^T] z &= \mathbb{E}[(z^T(X - Y))^2] \\ &\geq f(\mathbf{x}_r) f(\mathbf{x}_r + \mathbf{e}_r) z_r^2 \\ &> 0, \end{aligned}$$

where the expression on the second line is the contribution to the expectation from the event that $X = \mathbf{x}_r + \mathbf{e}_r$ and $Y = \mathbf{x}_r$, and the third line follows from (21). Note that we used that each term contributing to the expectation is non-negative. □

Lemma 13 (Covariance)**.** *For $i \in \{1,2\}$, the second moment $\widetilde{\mathbb{E}}^{(i),\infty}[\|\mathbf{Y}(\ell)\|^2]$ is finite and the covariance matrix $\mathbf{\Sigma}^{(i)}$ is positive definite. In particular, conditions (b) and (c) of Theorem 2 are satisfied.*

Proof. Observe that $\mathbf{Y}(1) \leq \tau_1 \vec{1} = (\tau_1, \tau_1, \ldots, \tau_1)$ coordinate-wise. Moreover, by construction, τ_1 is geometric and therefore has finite first and second moments. Hence $\widetilde{\mathbb{E}}^{(i),\infty}[\|\mathbf{Y}(\ell)\|^2] < \infty$, from which we have $|det(\mathbf{\Sigma}^{(i)})| < \infty$. Also $\mathbf{\Sigma}^{(i)}$ is positive definite by Lemmas 10 and 12. Hence $det(\mathbf{\Sigma}^{(i)}) > 0$. Condition (b) of Theorem 2 is verified and condition (c) follows from the multi-dimensional central limit theorem (see, e.g., [Dur19, Section 3.10]). □

3.3.4 Applying the local CLT

By Lemmas 10 and 13, we can apply Theorem 2 to the i.i.d. vectors $\{\mathbf{Y}(\ell)\}_{\ell=1}^{\infty}$ to obtain the following lower bound. Recall that $\mathbf{m}^{(1)} = \mathbf{m}^{(2)}$ by Lemma 11, and let $\mathbf{m} = \mathbf{m}^{(i)}$.

Lemma 14 (Uniform lower bound)**.** *There exist constants c_1, $c_2 \in (0, \infty)$ such that*

$$\inf_{y \in \mathcal{Y}_\ell^{(i)}} \widetilde{\mathbb{P}}^{(i),\infty}[\boldsymbol{S}_\ell = \mathbf{y}] \geq \frac{c_2}{\ell^{d/2}}$$

for all $\ell \geq c_1$ and $i \in \{1,2\}$, where $\mathbf{m} = \mathbf{m}^{(i)}$ and

$$\mathcal{Y}_\ell^{(i)} := \left\{ \mathbf{y} \in \mathbb{Z}_+^d : \ (\mathbf{y} - \ell\,\mathbf{m})^{\mathrm{T}} \left(\mathbf{\Sigma}^{(i)}\right)^{-1} (\mathbf{y} - \ell\,\mathbf{m}) \leq 2\ell \right\}. \tag{27}$$

Proof. By Theorem 2, for $i \in \{1,2\}$, as $\ell \to \infty$,

$$\sup_{\mathbf{y} \in \mathbb{Z}^d} \left| \ell^{d/2}\, \widetilde{\mathbb{P}}^{(i),\infty}[\mathbf{S}_\ell = \mathbf{y}] - \varphi^{(i)}\left(\frac{\mathbf{y} - \ell\,\mathbf{m}}{\sqrt{\ell}}\right) \right| \to 0. \tag{28}$$

where $\mathbf{S}_\ell = \sum_{j=1}^{\ell} \mathbf{Y}(j)$ and

$$\varphi^{(i)}(\mathbf{x}) = \frac{\exp\left\{-\frac{1}{2}\mathbf{x}^{\mathrm{T}} \left(\mathbf{\Sigma}^{(i)}\right)^{-1} \mathbf{x}\right\}}{\sqrt{(2\pi)^d\, det(\mathbf{\Sigma}^{(i)})}}.$$

Therefore, for arbitrary $\epsilon > 0$, there exists ℓ_ϵ sufficiently large such that for all integers $\ell \geq \ell_\epsilon$ and all $\mathbf{y} \in \mathbb{Z}^d$,

$$\begin{aligned}\widetilde{\mathbb{P}}^{(i),\infty}[\mathbf{S}_\ell = \mathbf{y}] &\geq \frac{1}{\ell^{d/2}} \left(\varphi^{(i)} \left(\frac{\mathbf{y} - \ell\,\mathbf{m}}{\sqrt{\ell}} \right) - \epsilon \right) \\ &= \frac{1}{\ell^{d/2}} \left(\frac{\exp\left\{ -\frac{1}{2} \left(\frac{\mathbf{y}-\ell\,\mathbf{m}}{\sqrt{\ell}} \right)^{\mathrm{T}} \left(\mathbf{\Sigma}^{(i)} \right)^{-1} \left(\frac{\mathbf{y}-\ell\,\mathbf{m}}{\sqrt{\ell}} \right) \right\}}{\sqrt{(2\pi)^d\, det(\mathbf{\Sigma}^{(i)})}} - \epsilon \right).\end{aligned}$$

The bound in the definition of $\mathcal{Y}_\ell^{(i)}$ gives

$$\inf_{y \in \mathcal{Y}_\ell^{(i)}} \widetilde{\mathbb{P}}^{(i),\infty}[\mathbf{S}_\ell = \mathbf{y}] \geq \frac{1}{\ell^{d/2}} \left(\frac{e^{-1}}{\sqrt{(2\pi)^d\, [det(\mathbf{\Sigma}^{(1)}) \vee det(\mathbf{\Sigma}^{(2)})]}} - \epsilon \right)$$

for all $\ell \geq \ell_\epsilon$. The lemma follows by taking ϵ to be any fixed number small enough that depends only on $det(\mathbf{\Sigma}^{(1)}) \vee det(\mathbf{\Sigma}^{(2)})$. □

Observe that the bound in Lemma 14 is uniform over the set $\mathcal{Y}_\ell^{(i)}$. Our use of Lemma 14 below will require a lower bound on the size of $\mathcal{Y}_\ell^{(1)} \cap \mathcal{Y}_\ell^{(2)}$.

Lemma 15. *Let $\lambda_{min}^{(i)}$ be the minimal eigenvalues of $\mathbf{\Sigma}^{(i)}$. Then*

$$\left\{ \mathbf{y} \in \mathbb{Z}_+^d : \|\mathbf{y} - \ell\mathbf{m}\|^2 \leq 2\ell\, (\lambda_{min}^{(1)} \wedge \lambda_{min}^{(2)}) \right\} \subset \mathcal{Y}_\ell^{(1)} \cap \mathcal{Y}_\ell^{(2)}, \tag{29}$$

where $\{\mathcal{Y}_\ell^{(i)}\}_{i=1}^2$ are defined in (27).

Proof. Let $\lambda_{min}^{(i)}$ be the smallest eigenvalue of $\mathbf{\Sigma}^{(i)}$. Then $0 < \lambda_{min}^{(i)} < \infty$ by Lemma 13. Since λ is an eigenvalue of $\mathbf{\Sigma}^{(i)}$ if and only if $1/\lambda$ is an eigenvalue of $\left(\mathbf{\Sigma}^{(i)}\right)^{-1}$, we have

$$(\mathbf{y} - \ell\mathbf{m})^{\mathrm{T}} \left(\mathbf{\Sigma}^{(i)} \right)^{-1} (\mathbf{y} - \ell\mathbf{m}) \leq \frac{1}{\lambda_{min}^{(i)}} \|\mathbf{y} - \ell\mathbf{m}\|^2.$$

This inequality implies (29). □

In fact, we will need to control the size of subsets of $\mathcal{Y}_\ell^{(1)} \cap \mathcal{Y}_\ell^{(2)}$ whose first coordinates are sufficiently close to their expectation. Letting m_1 be the first coordinate of $\mathbf{m}$, by Lemma 11,

$$m_1 = \widetilde{\mathbb{E}}^{(1),\infty}[\tau_1] = \widetilde{\mathbb{E}}^{(2),\infty}[\tau_1]. \tag{30}$$

We consider the following set of pairs (μ, ℓ)

$$\mathcal{L} = \left\{(\mu, \ell) \in \mathbb{N}^2 : \, |\mu - \ell m_1| \le c_3\sqrt{\ell}\right\} \quad \text{where} \quad c_3 = \sqrt{\lambda_{min}^{(1)} \wedge \lambda_{min}^{(2)}}. \tag{31}$$

The next two lemmas concern bounds on the level sets

$$\mathcal{L}|_\ell := \{\mu \in \mathbb{N} : \, (\mu, \ell) \in \mathcal{L}\} \quad \text{and} \quad \mathcal{L}|_\mu := \{\ell \in \mathbb{N} : \, (\mu, \ell) \in \mathcal{L}\}.$$

Lemma 16. *Let $\mathbb{Z}_+^d(\mu)$ be the subset of $\mathbb{Z}_+^d$ whose first coordinate is μ. Then*

$$\inf_{\mu \in \mathcal{L}|_\ell} \left|\mathcal{Y}_\ell^{(1)} \cap \mathcal{Y}_\ell^{(2)} \cap \mathbb{Z}_+^d(\mu)\right| \ge c_4\, c_3^{d-1}\, \ell^{(d-1)/2} \tag{32}$$

for all $\ell \in \mathbb{N}$, where $c_4 \in (0, \infty)$ is a constant that depends only on d.

Proof. By Lemma 15, the set $\mathcal{Y}_\ell^{(1)} \cap \mathcal{Y}_\ell^{(2)}$ contains all integer points of

$$B_d(\ell\mathbf{m},\, c_3\sqrt{2\ell}) \cap \mathbb{R}_+^d,$$

where $B_d(x, r) := \{y \in \mathbb{R}^d : \, \|y - x\|_{\mathbb{R}^d} \le r\}$ is the d-dimensional Euclidean ball with center x and radius r. By Lemma 11, $\mathbf{m} = (m_1, m_2, \cdots, m_d) \in \mathbb{R}_+^d$. Hence $\mathcal{Y}_\ell^{(1)} \cap \mathcal{Y}_\ell^{(2)} \cap \mathbb{Z}_+^d(\mu)$ contains all integer points of $\widetilde{B}(r_\mu) \cap \mathbb{R}_+^d$, where

$$\widetilde{B}(r_\mu) := \{(\mu,\, y_2, \cdots, y_d) \in \mathbb{R}^d : \, \|(y_2, \cdots, y_d) - \ell(m_2, \cdots, m_d)\|_{\mathbb{R}^{d-1}} \le r_\mu\}$$

with

$$r_\mu := \sqrt{c_3^2 2\ell - (\mu - \ell m_1)^2} \ge c_3\sqrt{\ell}. \tag{33}$$

The last inequality follows whenever $(\mu, \ell) \in \mathcal{L}$.

Since $\{m_2, \cdots, m_d\}$ are all non-negative, $\widetilde{B}(r_\mu) \cap \mathbb{R}_+^d$ contains a $(d-1)$-dimensional cube with side length $\frac{r_\mu}{\sqrt{d-1}} \ge \frac{c_3\sqrt{\ell}}{\sqrt{d-1}}$ by (33). This cube contains at least $c_4 c_3^{d-1} \ell^{(d-1)/2}$ many integer points for some $c_4 \in (0, \infty)$ that depends only on d, uniformly for all $(\mu, \ell) \in \mathcal{L}$. □

Finally, the following lemma gives a lower bound on the cardinality of $\mathcal{L}|_\mu$.

Lemma 17. *There exists a constant $c_5 \in (0,\infty)$ that depend only on m_1 and c_3 such that for μ large enough,*

$$\left[\frac{\mu}{m_1} - c_5\sqrt{\mu},\ \frac{\mu}{m_1} + c_5\sqrt{\mu}\right] \subseteq \mathcal{L}|_\mu,$$

where $m_1 = \widetilde{\mathbb{E}}^{(1),\infty}[\tau_1] = \widetilde{\mathbb{E}}^{(2),\infty}[\tau_1]$.

Proof. Suppose ℓ belongs to the interval on the left-hand side of the display in the statement of the lemma. Then $\ell \geq \frac{\mu}{m_1} - c_5\sqrt{\mu}$. Solving this quadratic inequality in $\sqrt{\mu}$ and then taking square give

$$\sqrt{\mu} \leq \frac{c_5 m_1 + \sqrt{(c_5 m_1)^2 + 4m_1\ell}}{2},$$

and

$$\mu \leq \ell m_1 + \frac{1}{4}(c_5 m_1)^2 + \frac{c_5 m_1\sqrt{(c_5 m_1)^2 + 4m_1\ell}}{2}.$$

From the last inequality, we see that $\mu \leq \ell m_1 + c_3\sqrt{\ell}$ for all $\ell \geq 1$, provided that $c_5 \in (0,\infty)$ is small enough (depending only on c_3 and m_1).

The other direction can be shown similarly, by solving the inequality $\ell \leq \frac{\mu}{m_1} + c_5\sqrt{\mu}$ to yield

$$\sqrt{\mu} \geq \frac{-c_5 m_1 + \sqrt{(c_5 m_1)^2 + 4m_1\ell}}{2},$$

and

$$\mu \geq \ell m_1 + \frac{1}{4}(c_5 m_1)^2 - \frac{c_5 m_1\sqrt{(c_5 m_1)^2 + 4m_1\ell}}{2}.$$

For $c_5 \in (0,\infty)$ is small enough (depending only on c_3 and m_1), we have $\mu \geq \ell m_1 - c_3\sqrt{\ell}$.

The desired subset relation is obtained. □

3.4 Final bound on the total variation distance

Proof of Theorem 1. Now we finish the proof of Theorem 1 by establishing (17). That is, we now show that

$$\inf_{\mu\in\mathbb{Z}_+} \sum_{(z'_A,z'_{B'},z'_{C'})\in(\mathcal{S}_0^\mu)^3} \widetilde{\mathbb{P}}^{(1)}_{Z'_A,Z'_{B'},Z'_{C'}}(z'_A,z'_{B'},z'_{C'}) \wedge \widetilde{\mathbb{P}}^{(2)}_{Z'_A,Z'_{B'},Z'_{C'}}(z'_A,z'_{B'},z'_{C'}) \;>\; 0, \tag{34}$$

where $\mathcal{S}_0^\mu := \{(\vec{0},\vec{0})\}\times\{0,1\}^{2k}\times\{0,1,\cdots,\mu\}^{\mathcal{H}}$ and $m=(\mu+1)k$. We further restrict the last pair of triples by considering $\mathcal{S}_{00}^\mu := \{(\vec{0},\vec{0})\}\times\{(\vec{0},\vec{0})\}\times\{0,1,\cdots,\mu\}^{\mathcal{H}}$. Since $\mathcal{S}_{00}^\mu\subset\mathcal{S}_0^\mu$, the sum $\sum_{(z'_A,z'_{B'},z'_{C'})\in(\mathcal{S}_0^\mu)^3}$ on the left of (34) is bounded below by

$$\sum_{(z'_A,z'_{B'},z'_{C'})\in(\mathcal{S}_{00}^\mu)^3} \widetilde{\mathbb{P}}^{(1)}_{Z'_A,Z'_{B'},Z'_{C'}}(z'_A,z'_{B'},z'_{C'}) \wedge \widetilde{\mathbb{P}}^{(2)}_{Z'_A,Z'_{B'},Z'_{C'}}(z'_A,z'_{B'},z'_{C'}). \tag{35}$$

As an element of $\mathcal{S}_{00}^\mu$, $z'_V=\big((\vec{0},\vec{0}),\,(\vec{0},\vec{0}),\,N'_V\big)$ for some $N'_V\in\{0,1,\cdots,\mu\}^{\mathcal{H}}$, where $V\in\{A,B',C'\}$. Hence

$$\begin{aligned}
&\widetilde{\mathbb{P}}^{(i)}_{Z'_A,Z'_{B'},Z'_{C'}}(z'_A,z'_{B'},z'_{C'})\\
&=\sum_{\ell=1}^{\mu}\widetilde{\mathbb{P}}^{(i)}\left\{\big(N_{\mathcal{H}}^{\sigma_A}(\tau_\ell),\,N_{\mathcal{H}}^{\sigma_{B'}}(\tau_\ell),\,N_{\mathcal{H}}^{\sigma_{C'}}(\tau_\ell)\big)=(N'_A,N'_{B'},N'_{C'})\;,\;\tau_\ell=\mu\right\}\\
&=\sum_{\ell=1}^{\mu}\widetilde{\mathbb{P}}^{(i),\infty}\left\{\big(N_{\mathcal{H}}^{\sigma_A}(\tau_\ell),\,N_{\mathcal{H}}^{\sigma_{B'}}(\tau_\ell),\,N_{\mathcal{H}}^{\sigma_{C'}}(\tau_\ell)\big)=(N'_A,N'_{B'},N'_{C'})\;,\;\tau_\ell=\mu\right\}\\
&=\sum_{\ell=1}^{\mu}\widetilde{\mathbb{P}}^{(i),\infty}\left\{\sum_{j=1}^{\ell}\mathbf{Y}(j)=(\mu,\,N'_A,N'_{B'},N'_{C'})\right\},
\end{aligned}$$

where the second and the last equalities follow from Lemma 8 and (20) respectively. Therefore,

$$(35)\geq\sum_{\ell=1}^{\mu}\sum_{\mathbf{y}\in\mathbb{Z}_+^d(\mu)}\widetilde{\mathbb{P}}^{(1),\infty}\left\{\sum_{j=1}^{\ell}\mathbf{Y}(j)=\mathbf{y}\right\}\wedge\widetilde{\mathbb{P}}^{(2),\infty}\left\{\sum_{j=1}^{\ell}\mathbf{Y}(j)=\mathbf{y}\right\}, \tag{36}$$

where $\mathbb{Z}_+^d(\mu)$ is the subset of $\mathbb{Z}_+^d$ whose first coordinate is μ.

We further restrict the sums to be over $(\mu,\, N'_A, N'_{B'}, N'_{C'}) \in \mathcal{Y}_\ell^{(1)} \cap \mathcal{Y}_\ell^{(2)}$ and $\ell \geq c_1$, where recall that $\mathcal{Y}_\ell^{(i)}$ and c_1 were defined in Lemma 14. We obtain from Lemma 14 that the right-hand side of (36) is

$$\begin{aligned}
&\geq \sum_{\ell \in [c_1,\, \mu] \cap \mathbb{Z}_+} \; \sum_{\mathbf{y} \in \mathcal{Y}_\ell^{(1)} \cap \mathcal{Y}_\ell^{(2)} \cap \mathbb{Z}_+^d(\mu)} \frac{c_2}{\ell^{d/2}} \\
&= c_2 \sum_{\ell \in [c_1,\, \mu] \cap \mathbb{Z}_+} \frac{\left|\mathcal{Y}_\ell^{(1)} \cap \mathcal{Y}_\ell^{(2)} \cap \mathbb{Z}_+^d(\mu)\right|}{\ell^{d/2}} \\
&\geq c_2\, c_3^{d-1}\, c_4 \sum_{\ell \in [c_1,\, \mu] \cap \mathcal{L}|_\mu} \frac{1}{\ell^{1/2}},
\end{aligned}$$

where the last inequality follows from Lemma 16 and the fact that $\ell \in \mathcal{L}|_\mu$ if and only if $\mu \in \mathcal{L}|_\ell$. Now by Lemma 17 and the fact that $m_1 \geq 1$ (recall that $m_1 = \widetilde{\mathbb{E}}^{(1),\infty}[\tau_1] = \widetilde{\mathbb{E}}^{(2),\infty}[\tau_1]$), we have for μ large enough that

$$\begin{aligned}
\sum_{\ell \in [c_1,\, \mu] \cap \mathcal{L}|_\mu} \frac{1}{\ell^{1/2}} &\geq \sum_{\ell \in [c_1,\, \mu] \cap \left[\frac{\mu}{m_1} - c_5\sqrt{\mu},\, \frac{\mu}{m_1} + c_5\sqrt{\mu}\right]} \frac{1}{\ell^{1/2}} \\
&= \sum_{\ell \in \left[\frac{\mu}{m_1} - c_5\sqrt{\mu},\, \frac{\mu}{m_1} + c_5\sqrt{\mu}\right]} \frac{1}{\ell^{1/2}} \\
&\geq \frac{2c_5\sqrt{\mu} - 1}{\sqrt{\frac{\mu}{m_1} + c_5\sqrt{\mu}}},
\end{aligned}$$

which tends to $2c_5\sqrt{m_1} > 0$ as $\mu \to \infty$. The proof of (34) is complete by combining the above calculation with (35) and (36).

The proof of Theorem 1 is complete. □

4 Concluding remarks

We assumed that k is fixed in Theorem 1. It is not clear what can be proved when k increases with the length of the sequence. Recall, in particular, that $d = 1 + 3(2^{2k} - 2^k)$ depends on k, and our application of the local CLT does rely on the fact that d is fixed.

Acknowledgments

SR was supported by NSF grants DMS-1614242, CCF-1740707 (TRIPODS), DMS-1902892, and DMS-1916378, as well as a Simons Fellowship and a Vilas Associates Award. BL was supported by NSF grants DMS-1614242, CCF-1740707 (TRIPODS), DMS-1902892 (to SR). WTF was supported by NSF grants DMS-1614242 (to SR), DMS-1855417 and ONR grant TCRI N00014-20-1-2411.

References

[ARS15] Elizabeth S. Allman, John A. Rhodes, and Seth Sullivant. Statistically consistent k-mer methods for phylogenetic tree reconstruction. *Journal of computational biology : a journal of computational molecular cell biology*, 24 2:153–171, 2015.

[BAP05] Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Annals of Probability*, 33(5):1643–1697, September 2005. Publisher: Institute of Mathematical Statistics.

[BC01] A. D. Barbour and O. Chryssaphinou. Compound Poisson Approximation: A User's Guide. *The Annals of Applied Probability*, 11(3):964–1002, 2001. Publisher: Institute of Mathematical Statistics.

[Cav78] James A. Cavender. Taxonomy with confidence. *Mathematical Biosciences*, 40(3-4):271–280, August 1978.

[CP18] Phillip Compeau and Pavel Pevzner. *Bioinformatics Algorithms: An Active Learning Approach.* Active Learning Publishers, 2018.

[Cul61] I.V. Culanovski. *Twenty-Five Papers on Statistics and Probability.* Sel. transl. math. stat. probab. American Mathematical Society, 1961.

[DEKM98] Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge University Press, April 1998.

[DM95] Burgess Davis and David McDonald. An elementary proof of the local central limit theorem. *Journal of Theoretical Probability*, 8(3):693–701, July 1995.

[DR+13] Constantinos Daskalakis, Sebastien Roch, et al. Alignment-free phylogenetic reconstruction: Sample complexity via a branching process analysis. *The Annals of Applied Probability*, 23(2):693–721, 2013.

[DS19] Chris Durden and Seth Sullivant. Identifiability of Phylogenetic Parameters from k-mer Data Under the Coalescent. *Bulletin of Mathematical Biology*, 81(2):431–451, February 2019.

[Dur19] Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

[EKPS00] William Evans, Claire Kenyon, Yuval Peres, and Leonard J. Schulman. Broadcasting on Trees and the Ising Model. *The Annals of Applied Probability*, 10(2):410–433, 2000. Publisher: Institute of Mathematical Statistics.

[ESSW99] Péter L. Erdős, Michael A. Steel, László A. Székely, and Tandy J. Warnow. A few logs suffice to build (almost) all trees (I). *Random Structures & Algorithms*, 14(2):153–184, 1999.

[Far73] James S. Farris. A Probability Model for Inferring Evolutionary Trees. *Systematic Biology*, 22(3):250–256, September 1973. Publisher: Oxford Academic.

[FISGC15] Huan Fan, Anthony R Ives, Yann Surget-Groba, and Charles H Cannon. An assembly and alignment-free method of phylogeny reconstruction from next-generation sequencing data. *BMC genomics*, 16(1):522, 2015.

[FLR20] Wai-Tong Louis Fan, Brandon Legried, and Sebastien Roch. Impossibility of Consistent Distance Estimation from Sequence Lengths Under the TKF91 Model. *Bulletin of Mathematical Biology*, 82(9):123, September 2020.

[FR18] Wai-Tong Louis Fan and Sebastien Roch. Necessary and sufficient conditions for consistent root reconstruction in markov models on trees. *Electron. J. Probab.*, 23:24 pp., 2018.

[Gus97] Dan Gusfield. *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology.* Cambridge University Press, May 1997.

[GZ19] Arun Ganesh and Qiuyi (Richard) Zhang. Optimal sequence length requirements for phylogenetic tree reconstruction with indels. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2019, pages 721–732, New York, NY, USA, June 2019. Association for Computing Machinery.

[Hau14] Bernhard Haubold. Alignment-free phylogenetics and population genetics. *Briefings in bioinformatics*, 15(3):407–418, 2014.

[HKP15] Bernhard Haubold, Fabian Klötzl, and Peter Pfaffelhuber. andi: Fast and accurate estimation of evolutionary distances between closely related genomes. *Bioinformatics*, 31(8):1169–1175, 2015.

[KA90] S. Karlin and S. F. Altschul. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proceedings of the National Academy of Sciences*, 87(6):2264–2268, March 1990. Publisher: National Academy of Sciences Section: Research Article.

[LHTH+19] John A Lees, Simon R Harris, Gerry Tonkin-Hill, Rebecca A Gladstone, Stephanie W Lo, Jeffrey N Weiser, Jukka Corander, Stephen D Bentley, and Nicholas J Croucher. Fast and flexible bacterial genomic epidemiology with poppunk. *Genome research*, 29(2):304–316, 2019.

[LHW02] Ross A. Lippert, Haiyan Huang, and Michael S. Waterman. Distributional regimes for the number of k-word matches between two random sequences. *Proceedings of the National Academy of Sciences*, 99(22):13980–13989, October 2002. Publisher: National Academy of Sciences Section: Physical Sciences.

[LKP+18] John A Lees, Michelle Kendall, Julian Parkhill, Caroline Colijn, Stephen D Bentley, and Simon R Harris. Evaluation of phylogenetic reconstruction methods using bacterial whole genomes: a simulation based study. *Wellcome open research*, 3, 2018.

[LPW06] David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov chains and mixing times*. American Mathematical Society, 2006.

[Mos04] Elchanan Mossel. Phase transitions in phylogeny. *Transactions of the American Mathematical Society*, 356(6):2379–2404, 2004.

[OTM+16] Brian D Ondov, Todd J Treangen, Páll Melsted, Adam B Mallonee, Nicholas H Bergman, Sergey Koren, and Adam M Phillippy. Mash: fast genome and metagenome distance estimation using minhash. *Genome biology*, 17(1):132, 2016.

[PPP+06] Alkes L. Price, Nick J. Patterson, Robert M. Plenge, Michael E. Weinblatt, Nancy A. Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909, August 2006. Number: 8 Publisher: Nature Publishing Group.

[QWH04] Ji Qi, Bin Wang, and Bai-Iin Hao. Whole proteome prokaryote phylogeny without sequence alignment: a k-string composition approach. *Journal of molecular evolution*, 58(1):1–11, 2004.

[RCSW09] Gesine Reinert, David Chew, Fengzhu Sun, and Michael S. Waterman. Alignment-Free Sequence Comparison (I): Statistics and Power. *Journal of Computational Biology*, 16(12):1615–1634, December 2009. Publisher: Mary Ann Liebert, Inc., publishers.

[RS17] Sebastien Roch and Allan Sly. Phase transition in the sample complexity of likelihood-based phylogeny inference. *Probability Theory and Related Fields*, 169(1):3–62, October 2017.

[Ste94] M. Steel. Recovering a tree from the leaf colourations it generates under a Markov model. *Applied Mathematics Letters*, 7(2):19–23, March 1994.

[Ste16] Mike Steel. *Phylogeny—discrete and random processes in evolution*, volume 89 of *CBMS-NSF Regional Conference Series* *in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2016.

[Tav84] Simon Tavaré. Line-of-descent and genealogical processes, and their applications in population genetics models. *Theoretical Population Biology*, 26(2):119–164, October 1984.

[TKF91] Jeffrey L. Thorne, Hirohisa Kishino, and Joseph Felsenstein. An evolutionary model for maximum likelihood alignment of dna sequences. *Journal of Molecular Evolution*, 33(2):114–124, Aug 1991.

[UBTC06] Igor Ulitsky, David Burstein, Tamir Tuller, and Benny Chor. The average common substring approach to phylogenomic reconstruction. *Journal of Computational Biology*, 13(2):336–350, 2006.

[VA03] Susana Vinga and Jonas Almeida. Alignment-free sequence comparison—a review. *Bioinformatics*, 19(4):513–523, March 2003. Publisher: Oxford Academic.

[War17] Tandy Warnow. *Computational Phylogenetics: An Introduction to Designing Methods for Phylogeny Estimation*. Cambridge University Press, USA, 1st edition, 2017.

[WRSW10] Lin Wan, Gesine Reinert, Fengzhu Sun, and Michael S. Waterman. Alignment-Free Sequence Comparison (II): Theoretical Power of Comparison Statistics. *Journal of Computational Biology*, 17(11):1467–1490, October 2010. Publisher: Mary Ann Liebert, Inc., publishers.

A Information-theoretic bounds

In this section we give details about some basic facts we used in the paper. Recall the definition of the total variation distance in (1). It is well known, see e.g., [LPW06], that the supremum on the right hand side of (1) is reached at the set $B = \{x \in E : \nu_1(x) \geq \nu_2(x)\}$ as well as its complement B^c, and that we have the following characterizations.

Lemma 18. *Let ν_1 and ν_2 be probability measures on a countable space E.*

$$\|\nu_1 - \nu_2\|_{\mathrm{TV}} = \frac{1}{2}\sum_{\sigma\in E} |\nu_1(\sigma) - \nu_2(\sigma)| = 1 - \sum_{\sigma\in E} \nu_1(\sigma) \wedge \nu_2(\sigma).$$

Let X be a measurable function on a measure space $(\Omega, \mathcal{F})$, and $\mathbb{P}$ and $\mathbb{P}'$ be two probability measures on $(\Omega, \mathcal{F})$. Denote by $\mathbb{P}_{g(X)}$ and $\mathbb{P}'_{g(X)}$ the probability distribution of $g(X)$ under $\mathbb{P}$ and $\mathbb{P}'$ respectively, where g is an arbitrary measurable function on the state space of X.

Lemma 19. *Let g be a measurable map on the state space of X. Then*

$$\|\mathbb{P}_{g(X)} - \mathbb{P}'_{g(X)}\|_{\mathrm{TV}} \le \|\mathbb{P}_X - \mathbb{P}'_X\|_{\mathrm{TV}}.$$

Proof. Applying the definition (1) twice,

$$\begin{aligned}\|\mathbb{P}_{g(X)} - \mathbb{P}'_{g(X)}\|_{\mathrm{TV}} &= \sup_A |\mathbb{P}(g(X) \in A) - \mathbb{P}'(g(X) \in A)| \\ &= \sup_A |\mathbb{P}(X \in g^{-1}(A)) - \mathbb{P}'(X \in g^{-1}(A))| \\ &\le \|\mathbb{P}_X - \mathbb{P}'_X\|_{\mathrm{TV}}.\end{aligned}$$

□

Let X, Y, Z be measurable functions on a measure space $(\Omega, \mathcal{F})$, and $\mathbb{P}$ and $\mathbb{P}'$ be two probability measures on $(\Omega, \mathcal{F})$. We say that $X \to Y \to Z$ is a *Markov chain* under $\mathbb{P}$ if Z is conditionally independent of X given Y in the sense that

$$\mathbb{P}_{Z|X,Y} = \mathbb{P}_{Z|Y}, \tag{37}$$

where $\mathbb{P}_{Z|X,Y}$ is the conditional distribution of Z given (X, Y) and $\mathbb{P}_{Z|Y}$ is the conditional distribution of Z given Y. The law of total probability and (37) imply that

$$\mathbb{P}_{X,Y,Z} = \mathbb{P}_X \mathbb{P}_{Y|X} \mathbb{P}_{Z|Y}, \tag{38}$$

where $\mathbb{P}_{X,Y,Z}$ is the joint probability distribution of (X, Y, Z).

Lemma 20. *Suppose $\mathbb{P}_X = \mathbb{P}'_X$, $\mathbb{P}_{Y|X} = \mathbb{P}'_{Y|X}$ and $X \to Y \to Z$ is a Markov chain under both $\mathbb{P}$ and $\mathbb{P}'$. Then*

$$\|\mathbb{P}_{X,Y,Z} - \mathbb{P}'_{X,Y,Z}\|_{\mathrm{TV}} = \|\mathbb{P}_{Y,Z} - \mathbb{P}'_{Y,Z}\|_{\mathrm{TV}}.$$

Proof. By the first equality in Lemma 18,

$$\begin{aligned}
&\|\mathbb{P}_{X,Y,Z} - \mathbb{P}'_{X,Y,Z}\|_{\mathrm{TV}} \\
&= \frac{1}{2} \sum_{(a,b,c)} \big|\mathbb{P}((X,Y,Z) = (a,b,c)) - \mathbb{P}'((X,Y,Z) = (a,b,c))\big|.
\end{aligned}$$

Applying (38) to $\mathbb{P}$ and $\mathbb{P}'$, we have

$$\begin{aligned}
\mathbb{P}((X,Y,Z) = (a,b,c)) &= \mathbb{P}(X = a)\,\mathbb{P}(Y = b|X = a)\,\mathbb{P}(Z = c|Y = b), \\
\mathbb{P}'((X,Y,Z) = (a,b,c)) &= \mathbb{P}'(X = a)\,\mathbb{P}'(Y = b|X = a)\,\mathbb{P}'(Z = c|Y = b).
\end{aligned}$$

From the assumptions $\mathbb{P}_X = \mathbb{P}'_X$ and $\mathbb{P}_{Y|X} = \mathbb{P}'_{Y|X}$, it follows that $\mathbb{P}_{X,Y} = \mathbb{P}'_{X,Y}$ and $\mathbb{P}_Y = \mathbb{P}'_Y$. Applying these into the above two displayed equations give

$$\begin{aligned}
&|\mathbb{P}((X,Y,Z) = (a,b,c)) - \mathbb{P}'((X,Y,Z) = (a,b,c))| \\
&= \mathbb{P}(X = a)\,\mathbb{P}(Y = b|X = a)\ |\mathbb{P}(Z = c|Y = b) - \mathbb{P}'(Z = c|Y = b)| \\
&= \mathbb{P}(X = a, Y = b)\ |\mathbb{P}(Z = c|Y = b) - \mathbb{P}'(Z = c|Y = b)|\,.
\end{aligned}$$

Hence

$$\begin{aligned}
&\|\mathbb{P}_{X,Y,Z} - \mathbb{P}'_{X,Y,Z}\|_{\mathrm{TV}} \\
&= \frac{1}{2} \sum_{(a,b,c)} \mathbb{P}(X = a, Y = b)\ \big|\mathbb{P}(Z = c|Y = b) - \mathbb{P}'(Z = c|Y = b)\big| \\
&= \frac{1}{2} \sum_{(b,c)} \mathbb{P}(Y = b)\ \big|\mathbb{P}(Z = c|Y = b) - \mathbb{P}'(Z = c|Y = b)\big| \\
&= \frac{1}{2} \sum_{(b,c)} \big|\mathbb{P}(Y = b)\,\mathbb{P}(Z = c|Y = b) - \mathbb{P}'(Y = b)\,\mathbb{P}'(Z = c|Y = b)\big|,
\end{aligned}$$

where we used $\mathbb{P}(Y = b) = \mathbb{P}'(Y = b)$ in the last equality. The last term on the right-hand side is $\|\mathbb{P}_{Y,Z} - \mathbb{P}'_{Y,Z}\|_{\mathrm{TV}}$ by the first equality in Lemma 18, establishing the claim. □