Journal of Mathematical Biology (2021) 83:40
https://doi.org/10.1007/s00285-021-01667-2

Mathematical Biology

On asymptotic joint distributions of cherries and pitchforks for random phylogenetic trees

Kwok Pui Choi[1] · Gursharn Kaur[1] · Taoyang Wu[2]

Received: 18 January 2021 / Revised: 29 July 2021 / Accepted: 8 September 2021 /
Published online: 23 September 2021

Abstract
Tree shape statistics provide valuable quantitative insights into evolutionary mechanisms underpinning phylogenetic trees, a commonly used graph representation of evolutionary relationships among taxonomic units ranging from viruses to species. We study two subtree counting statistics, the number of cherries and the number of pitchforks, for random phylogenetic trees generated by two widely used null tree models: the proportional to distinguishable arrangements (PDA) and the Yule-Harding-Kingman (YHK) models. By developing limit theorems for a version of extended Pólya urn models in which negative entries are permitted for their replacement matrices, we deduce the strong laws of large numbers and the central limit theorems for the joint distributions of these two counting statistics for the PDA and the YHK models. Our results indicate that the limiting behaviour of these two statistics, when appropriately scaled using the number of leaves in the underlying trees, is independent of the initial tree used in the tree generating process.

Keywords Tree shape · Joint subtree distributions · Pólya urn model · Limit distributions · Yule-Harding-Kingman model · PDA model

Mathematics Subject Classification 92B10 · 60F05 · 92D99

✉ Taoyang Wu
taoyang.wu@uea.ac.uk

Kwok Pui Choi
stackp@nus.edu.sg

Gursharn Kaur
gursharn.kaur24@gmail.com

1 Department of Statistics and Data Science, and the Department of Mathematics, National University of Singapore, Singapore 117546, Republic of Singapore

2 School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, UK

1 Introduction

As a common mathematical representation of evolutionary relationships among biological systems ranging from viruses to species, phylogenetic trees retain important signatures of the underlying evolutionary events and mechanisms which are often not directly observable, such as rates of speciation and expansion (Mooers et al. 2007; Heath et al. 2008). To utilise these signatures, one popular approach is to compare empirical shape indices computed from trees inferred from real datasets with those predicted by neutral models specifying a tree generating process (see, e.g. Blum and François 2006; Hagen et al. 2015). Moreover, topological tree shapes are also informative for understanding several fundamental statistics in population genetics (Ferretti et al. 2017; Arbisser et al. 2018) and important parameters in the dynamics of virus evolution and propagation (Colijn and Gardy 2014).

This paper focuses on two subtree counting statistics: the number of cherries (i.e., nodes that have precisely two descendent leaves) and that of pitchforks (i.e., nodes that have precisely three descendent leaves) in a tree. These statistics are related to monophylogenetic structures in phylogenetic trees (Rosenberg 2003) and have been utilised recently to study evolutionary dynamics of pathogens (Colijn and Gardy 2014). For example, the asymptotic frequency of cherries in pathogen trees generated by some models can be used to estimate the basic reproduction number (Plazzotta and Colijn 2016) and to study the impact of the underlying contact network over which a pathogen spreads (Metzig et al. 2019). Various properties concerning these statistics have been established in the past decades on the following two fundamental random phylogenetic tree models: the Yule-Harding-Kingman (YHK) (Rosenberg 2006; Disanto and Wiehe 2013; Holmgren and Janson 2015) and the proportional to distinguishable arrangements (PDA) models (McKenzie and Steel 2000; Chang and Fuchs 2010; Wu and Choi 2016; Choi et al. 2020).

In this paper we are interested in the limiting behaviour of the joint cherry and pitchfork distributions for the YHK and the PDA models. In a seminal paper, McKenzie and Steel (2000) showed that cherry distributions converge to a normal distribution, which was later extended to pitchforks and other subtrees by Chang and Fuchs (2010). More recently, Holmgren and Janson (2015) studied subtree counts in the random binary search tree model, and their results imply that the cherry and pitchfork distributions converge jointly to a bivariate normal distribution under the YHK model. This is further investigated by Wu and Choi (2016) and Choi et al. (2020), where numerical results indicate that convergence to bivariate normal distributions holds under both the YHK model and the PDA model. Our main results, Theorems 1 and 2, provide a unifying approach to establishing the convergence of the joint distributions to bivariate normal distributions for both models, as well as a strong law stating that the joint counting statistics converge almost surely (a.s.) to a constant vector. Moreover, our results indicate that the limiting behaviour of these two statistics, when appropriately scaled, is independent of the initial tree used in the tree generating process.

Our approach is based on a general model in probability theory known as the Pólya urn scheme, which has been developed during the past few decades including applications in studying various growth phenomena with an underlying random tree structure (see, e.g. Mahmoud (2009) and the references therein). For instance, the results by

McKenzie and Steel (2000) are based on a version of the urn model in which the off-diagonal elements in the replacement matrix are all positive. However, such technical constraints pose a central challenge for studying pitchfork distributions as negative entries in the resulting replacement matrix are not confined only to the diagonal (see Sects. 4 and 5). To overcome this limitation, we study a family of extended Pólya urn models under certain technical assumptions in which negative entries are allowed for their replacement matrices (see Sect. 3). Inspired by the martingale approach used by Bai and Hu (2005), we present a self-contained proof for the limit theorems for this extended urn model, with the dual aims of completeness and accessibility. Our approach is different from a popular framework in which discrete urn models are embedded into a continuous Markov chain known as the branching processes (see, e.g. Janson (2004) and the references therein).

We summarize the contents of the rest of the paper. In the next section, we collect some definitions concerning phylogenetic trees and the two tree-based Markov processes. In Sect. 3, we introduce the urn model and a version of the Strong Law of Large Numbers and the Central Limit Theorem that are applicable to our study. We apply these two theorems to the YHK process in Sect. 4, and the PDA process in Sect. 5. These results are then extended to unrooted trees in Sect. 6. The proofs of the main results for the urn model are presented in Sect. 7, with a technical lemma included in the appendix. We conclude this paper in the last section with a discussion of our results and some open problems.

2 Preliminaries

In this section, we present some basic notation and background concerning phylogenetic trees, random tree models, and urn models. Throughout this paper, n is a positive integer greater than two unless stated otherwise.

2.1 Phylogenetic trees

A *tree* $T = (V(T), E(T))$ is a connected acyclic graph with vertex set $V(T)$ and edge set $E(T)$. A vertex is referred to as a *leaf* if it has degree one, and an *interior vertex* otherwise. An edge incident to a leaf is called a *pendant edge*, and let $E^\circ(T)$ be the set of pendant edges in T. A tree is *rooted* if it contains exactly one distinguished degree one node designated as the *root*, which is not regarded as a leaf and is usually denoted by ρ, and *unrooted* otherwise. Moreover, the orientation of a rooted tree is from its root to its leaves. Other than those in Sect. 6, all trees considered in this paper are rooted and *binary*, that is, each interior vertex has precisely two children.

A *phylogenetic tree* on a finite set X is a rooted binary tree with leaves bijectively labelled by the elements of X. The set of binary rooted phylogenetic trees on $\{1, 2, \ldots, n\}$ is denoted by $\mathcal{T}_n$. See Fig. 1 for examples of trees in $\mathcal{T}_7$ and $\mathcal{T}_8$. Given an edge e in a phylogenetic tree T on X and a taxon $x' \notin X$, let $T[e; x']$ be the phylogenetic tree on $X \cup \{x'\}$ obtained by attaching a new leaf with label x' to the edge e. Formally, let $e = (u, v)$ and let w be a vertex not contained in $V(T)$. Then $T[e; x']$

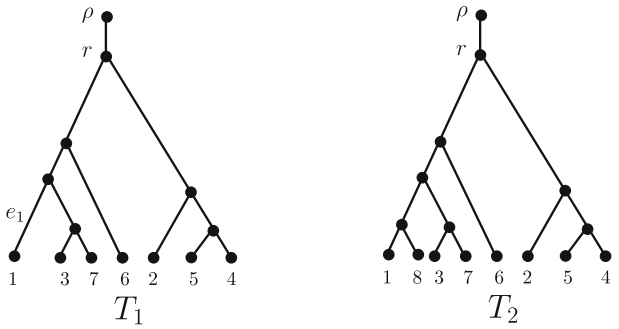

Fig. 1 Examples of phylogenetic trees. T_1 is a rooted phylogenetic tree on $\{1, \ldots, 7\}$; $T_2 = T_1[e_1]$ is a phylogenetic tree on $X = \{1, \ldots, 8\}$ obtained from T_1 by attaching a new leaf labelled 8 to the edge e_1 which is incident with taxon 1 in T_1

has vertex set $V(T) \cup \{x', w\}$ and edge set $\big(E(T) \setminus \{e\}\big) \cup \{(u, w), (w, v), (w, x')\}$. See Fig. 1 for an illustration of this construction, where tree $T_2 = T_1[e_1; 8]$ is obtained from T_1 by attaching leaf 8 to the edge e_1. We simply use $T[e]$ instead of $T[e; x']$ when the taxon name x' is not essential.

Removing an edge in a phylogenetic tree T results in two connected components; the connected component that does not contain the root of T is referred to as a subtree of T, also commonly known as a fringe subtree. A subtree is called a *cherry* if it has two leaves, and a *pitchfork* if it has three leaves. Following the notation by Choi et al. (2020), let $A(T)$ and $B(T)$ be the number of pitchforks and cherries contained in T. For example, in Fig. 1 we have $A(T_2) = 1$ and $B(T_2) = 3$.

2.2 The YHK and the PDA processes

Let $\mathcal{T}_n$ be the set of phylogenetic trees with n leaves. In this subsection, we introduce the two tree-based Markov processes investigated in this paper: the proportional to distinguishable arrangements (PDA) process and the Yule-Harding-Kingman (YHK) process. Our description of these two processes is largely based on that in Choi et al. (2020), which is adapted from the Markov processes as described by Steel (2016, Section 3.3.3).

Under the YHK process (Yule 1925; Harding 1971), starting with a given tree T_m in $\mathcal{T}_m$ with $m \geq 2$, a random phylogenetic tree T_n in $\mathcal{T}_n$ is generated as follows.

(i) Select a uniform random permutation $(x_1, \ldots, x_n)$ of $\{1, 2, \ldots, n\}$;
(ii) label the leaves of the rooted phylogenetic tree T_m randomly using the taxon set $\{x_1, x_2, \ldots, x_m\}$;
(iii) for $m \leq k < n$, uniformly choose a random pendant edge e in T_k and let $T_{k+1} = T_k[e; x_{k+1}]$.

The PDA process can be described using a similar scheme; the only difference is that in Step (iii) the edge e is uniformly sampled from the edge set of T_k, instead of the pendant edge set. Furthermore, under the PDA process, Step (i) can also be simplified by using a fixed permutation, say $(1, 2, \ldots, n)$. In the literature, the special

case $m = 2$, for which T_2 is the unique tree with two leaves, is also referred to as the YHK model and the PDA model, respectively.

For $n \geq 4$, let A_n and B_n be the random variables $A(T)$ and $B(T)$, respectively, for a random tree T in $\mathcal{T}_n$. The probability distributions of A_n (resp. B_n) are referred to as pitchfork distributions (resp. cherry distributions). In this paper, we are mainly interested in the limiting distributional properties of (A_n, B_n).

2.3 Modes of convergence

Let $X, X_1, X_2, \ldots$ be random variables on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. To study the urn model we will use the following four modes of convergence (see, e.g. Grimmett and Stirzaker (2001, Section 7.2) for more details). First, X_n is said to converge to X *almost surely*, denoted as $X_n \xrightarrow{a.s.} X$, if $\{\omega \in \Omega : X_n(\omega) \to X(\omega) \text{ as } n \to \infty\}$ is an event with probability 1. Next, X_n is said to converge to X *in r-th norm*, where $r > 0$, written $X_n \xrightarrow{r} X$, if $\mathbb{E}(|X_n^r|) < \infty$ for all n and $\mathbb{E}(|X_n - X|^r) \to 0$ as $n \to \infty$. Furthermore, X_n is said to converge to X *in probability*, written $X_n \xrightarrow{p} X$, if $\mathbb{P}(|X_n - X| > \epsilon) \to 0$ as $n \to \infty$ for all $\epsilon > 0$. Finally, X_n converges to a random variable Y *in distribution*, also termed *weak convergence* or *convergence in law* and written $X_n \xrightarrow{d} Y$, if $\mathbb{P}(X_n \leq x) \to \mathbb{P}(Y \leq x)$ as $n \to \infty$ for all points x at which the distribution function $\mathbb{P}(Y \leq x)$ is continuous. Note that $X_n \xrightarrow{p} X$ implies $X_n \xrightarrow{d} X$, and $X_n \xrightarrow{p} X$ holds if either $X_n \xrightarrow{a.s.} X$ holds or $X_n \xrightarrow{r} X$ holds for some $r > 0$.

2.4 Miscellaneous

Let $\mathbf{0} = (0, \ldots, 0)$ be the d-dimensional zero row vector. Let $\mathbf{e} = (1, \ldots, 1)$ be the d-dimensional row vector whose entries are all one, and for $1 \leq j \leq d$, let $\mathbf{e}_j$ denote the j-th canonical row vector whose j-th entry is 1 while the other entries are all zero.

Let $\mathrm{diag}(a_1, \ldots, a_d)$ denote a diagonal matrix whose diagonal elements are $a_1, \ldots, a_d$. Furthermore, $\mathbf{0}^\top \mathbf{0}$ is the $d \times d$ matrix whose entries are all zero. Here $Z^\top$ denotes the transpose of Z, where Z can be either a vector or a matrix.

3 Urn models

In this section, we briefly recall the classical Pólya urn model and some of its generalisations. Pólya urn model was studied by Pólya (1930) and can be traced back to Markov (see, e.g. Johnson and Kotz (1977, Section 1.2)). It has been applied to describe evolutionary processes in biology and computer science. Several such applications in genetics are discussed by Johnson and Kotz (1977, Chapter 5) and by Mahmoud (2009, Chapters 8 and 9). In a general setup, consider an urn with balls of d different colours containing $C_{0,i}$ many balls of colour $i \in \{1, 2, \ldots, d\}$ at time 0. At each time step, a ball is drawn uniformly at random and returned with some extra balls, depending on the colour selected. The reinforcement scheme is often described by a $d \times d$ matrix R: if the colour of the ball drawn is i, then we return the selected ball along with R_{ij} many

balls of colour j, for every $j \in \{1, 2, \ldots, d\}$, where a positive value of R_{ij} means adding R_{ij} balls and a negative value of R_{ij} means removing $|R_{ij}|$ many balls from the urn. Such a matrix is termed as *replacement matrix* in the literature. For instance, the replacement matrix R is the identity matrix for the original Pólya urn model with d colours: at each time point, the selected ball is returned with one additional ball of the same colour. We restrict our attention to tenable urn processes, that is, at each step it is always possible to add or remove balls according to the matrix R.

Let $C_n = (C_{n,1}, \ldots, C_{n,d})$ be the row vector of dimension d that represents the ball configuration at time n for an urn model with d colours, in which each entry is necessarily non-negative and at least one of these entries is greater than 0. Then the sum of $C_{n,i}$, denoted by t_n, is the number of balls in the urn at time n. Note that throughout this paper, t_n is always a number greater than 0. Recall that a vector is referred to as a *stochastic vector* if each entry in the vector is a non-negative real number and the sum of its entries is one. Denote the stochastic vector associated with C_n by $\widetilde{C}_n$, that is, we have $\widetilde{C}_{n,i} = C_{n,i}/t_n$ for $1 \le i \le d$.

Let $\mathcal{F}_n$ be the information of the urn's configuration from time 0 up to n, that is, the σ-algebra generated by $C_0, C_1, \ldots, C_n$. Let R denote the replacement matrix. Then, for every $n \ge 1$,

$$C_n = C_{n-1} + \chi_n R, \tag{1}$$

where χ_n is a random row vector of length d such that for $i = 1, \ldots, d$,

$$\mathbb{P}(\chi_n = \mathbf{e}_i | \mathcal{F}_{n-1}) = \widetilde{C}_{n-1,i}.$$

Since precisely one entry in χ_n is 1 and all others are 0, it follows that

$$\mathbb{E}[\chi_n | \mathcal{F}_{n-1}] = \widetilde{C}_{n-1} \quad \text{and} \quad \mathbb{E}[\chi_n^\top \chi_n | \mathcal{F}_{n-1}] = \mathrm{diag}(\widetilde{C}_{n-1}). \tag{2}$$

We state the following assumptions about the replacement matrix R:

(A1) *Tenable:* It is always possible to draw balls and follow the replacement rule, that is, we never get stuck in following the rules (see, e.g. Mahmoud (2009, p.46)).

(A2) *Small:* All eigenvalues of R are real; the maximal eigenvalue $\lambda_1 = s$ is positive with $\lambda_1 > 2\lambda$ holds for all other eigenvalues λ of R.

(A3) *Strictly Balanced:* The column vector $\mathbf{e}^\top$ is a right eigenvector of R corresponding to λ_1 and one of the left eigenvectors corresponding to λ_1 is a stochastic vector. Note that $\mathbf{e}^\top$ being a right eigenvector implies $t_n = t_0 + ns$, and hence the urn models discussed here are *balanced*, as commonly known in the literature.

(A4) *Diagonalisable:* R is diagonisable over real numbers. That is, there exists an invertible matrix U with real entries such that

$$U^{-1} R U = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_d) =: \Lambda, \tag{3}$$

where $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_d$ are all eigenvalues of R.

For the matrix U in (A4) and $1 \le j \le d$, let $\mathbf{u}_j = U\mathbf{e}_j^\top$ denote the j-th column of U, and $\mathbf{v}_j = \mathbf{e}_j U^{-1}$ the j-th row of U^{-1}. Then $\mathbf{u}_j$ and $\mathbf{v}_j$ are, respectively, *right and left*

eigenvectors corresponding to λ_j. Furthermore, since $\mathbf{v}_i\mathbf{u}_j = \mathbf{e}_i U^{-1} U \mathbf{e}_j^\top = \mathbf{e}_i \mathbf{I}\, \mathbf{e}_j^\top$, where $\mathbf{I}$ is the identity matrix, we have

$$\mathbf{v}_i\mathbf{u}_j = 1 \text{ if } i = j, \text{ and } \mathbf{v}_i\mathbf{u}_j = 0 \text{ if } i \neq j. \tag{4}$$

In view of (A3), (A4) and (4), for simplicity the following convention is used throughout this paper:

$$\mathbf{u}_1 = \mathbf{e}^\top \quad \text{and} \quad \mathbf{v}_1 \text{ is a stochastic vector.} \tag{5}$$

Furthermore, the eigenvalue λ_1 is referred to as the *principal eigenvalue*; $\mathbf{u}_1$ and $\mathbf{v}_1$ specified in (5) as the *principal right and principal left eigenvector*, respectively.

Motivated by adaptive clinical trial problems, Bai and Hu (2005) derived limit results in an urn model by martingale techniques. Moreover, they considered random replacement matrices but required the replacement matrix has non-negative elements. On the other hand, the limit results derived in Janson (2004) are based on an embedding of the urn model into a continuous time branching process under certain non-trivial technical assumptions of the associated continuous time branching process. In this paper, we prove first and second order limit results for an urn model with a replacement matrix that may contain non-negative elements. As mentioned earlier, our proofs are based on the martingale approach for the urn models used by Bai and Hu (2005). Under assumptions (A1)-(A4), the exact expression for the limiting variance matrix agrees with the one obtained by Bai and Hu (2005) and by Janson (2004). Notice that the assumption of real eigenvalues in (A2) and real eigenvectors in (A4) is chosen to make our proof more accessible to a wider audience by simplifying expressions and the proofs. Indeed, our proof can be extended to the case where the dominating eigenvalue λ_1 is real while the eigenvalues $\lambda_2, \ldots, \lambda_d$ are complex-valued whose real parts are less than $\lambda_1/2$, as one of the cases studied in Janson (2004).

The limit of the urn process and the rate of convergence to the limiting vector depends on certain spectral properties of matrix R (see, e.g. Janson (2004) or Bai and Hu (2005)). In our context, it suffices to consider the extended Pólya urn model under the aforementioned assumptions (A1)–(A4), for which Theorems 1 and 2 below give the Strong Law of Large Numbers and the Central Limit Theorem. Our proofs, which are adapted from that of Bai and Hu (2005), are presented in Sect. 7.

Theorem 1 *Under assumptions* (A1)–(A4)*, we have*

$$(ns)^{-1} C_n \xrightarrow{a.s.} \mathbf{v}_1 \quad \textit{and} \quad (ns)^{-1} C_n \xrightarrow{r} \mathbf{v}_1 \quad \textit{for } r > 0, \tag{6}$$

where s is the principal eigenvalue and $\mathbf{v}_1$ is the principal left eigenvector.

Let $\mathcal{N}(\mathbf{0}, \Sigma)$ be the multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix Σ.

Theorem 2 *Under assumptions* (A1)–(A4)*, we have*

$$n^{-1/2}(C_n - ns\mathbf{v}_1) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma),$$

where s is the principal eigenvalue, $\mathbf{v}_1$ is the principal left eigenvector, and

$$\Sigma = \sum_{i,j=2}^{d} \frac{s\lambda_i\lambda_j \mathbf{u}_i^\top diag(\mathbf{v}_1)\mathbf{u}_j}{s - \lambda_i - \lambda_j} \mathbf{v}_i^\top \mathbf{v}_j. \tag{7}$$

Remark 1 During the reviewing process of this paper, a reviewer suggested that an alternative approach to establishing Theorems 1 and 2 might be based on Janson (2004, Theorems 3.21 & 3.22 and Remark 4.2) and a result on the super-critical Galton-Watson process (see, e.g. Athreya and Ney (1972, Theorem 2(i) in Section III.7)), which could potentially lead to a stronger version of the results presented here.

4 Limiting distributions under the YHK model

A cherry is said to be *independent* if it is not contained in any pitchfork, and *dependent* otherwise. Similarly, a pendant edge is *independent* if it is contained in neither a pitchfork nor a cherry. In this section, we study the limiting joint distribution of the random variables A_n (i.e., the number of pitchforks) and B_n (i.e., the number of cherries) under the YHK model.

To study the joint distribution of cherries and pitchforks, we extend the urn models used in McKenzie and Steel (2000) (see also Steel (2016, Section 3.4)) as follows. Each pendant edge in a phylogenetic tree is designated as one of the following four types:

(E1): a type 1 edge is a pendant edge in a dependent cherry (i.e., contained in both a cherry and a pitchfork);
(E2): a type 2 edge is a pendant edge in an independent cherry (i.e., contained in a cherry but not a pitchfork);
(E3): a type 3 edge is a pendant edge contained in a pitchfork but not a cherry;
(E4): a type 4 edge is an independent pendant edge (i.e., contained in neither a pitchfork nor a cherry).

It is straightforward to see that any pendant edge in a phylogenetic tree with at least two leaves belongs to one and only one of the above four types. Furthermore, the numbers of pitchforks and independent cherries in a tree are precisely half of the numbers of type 1 and type 2 edges, respectively.

As illustrated in Fig. 2, the composition of the types of the pendant edges in $T[e]$, the tree obtained from T by attaching an extra leaf to a pendant edge e, is determined by the composition of pendant edge types in T and the type of e as follows. When e is type 1, then the number of type 4 edges in $T[e]$ increases by one compared with that in T while the number of edges of each of the other three types is the same. This holds because both $T[e]$ and T have the same number of cherries and that of pitchforks (see T_3 and T_4 in Fig. 2). When e is of type 2, then the number of type 2 edges decreases by two while the numbers of type 1 and of type 3 increase by two and one, respectively. This is because in this case one independent cherry is replaced by one pitchfork (see T_2 and T_3 in Fig. 2). When e is type 3, one pitchfork is replaced by two independent

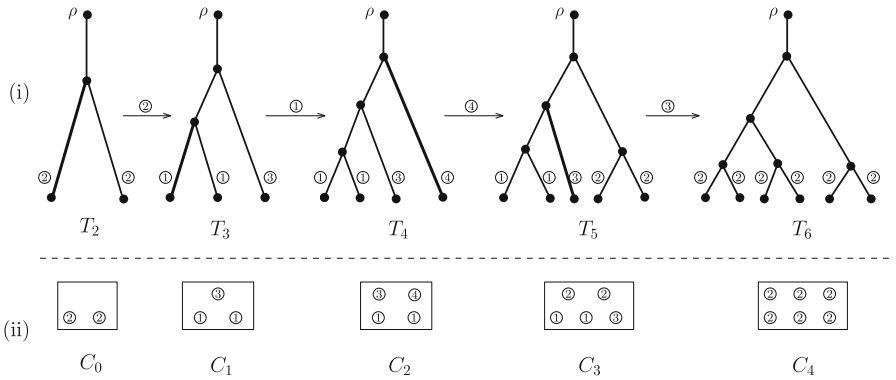

Fig. 2 A sample path of the YHK model and the associated urn model. (i): A sample path of the YHK model evolving from T_2 with two leaves to T_6 with six leaves. The labels of the leaves are omitted for simplicity. The type of pendant edges is indicated by the circled numbers next to them. For $2 \le i \le 5$, the edge selected in T_i to generate T_{i+1} is highlighted in bold and the associated edge type is indicated in the circled number above the arrows. (ii) The associated urn model with four colours, derived from the types of pendants edges in the trees. Note that in the vector form we have $C_0 = (0, 2, 0, 0)$, $C_1 = (2, 0, 1, 0)$, $C_2 = (2, 0, 1, 1)$, $C_3 = (2, 2, 1, 0)$, and $C_4 = (0, 6, 0, 0)$

cherries, hence the number of type 2 edges increases by four while the numbers of edges of type 1 and of type 3 decrease by two and one, respectively (see T_5 and T_6 in Fig. 2). Finally, when e is type 4, one independent pendant edge is replaced by one independent cherry, and hence the number of type 2 edges increases by two and that of type 4 edges decreases by one (see T_4 and T_5 in Fig. 2).

Using the dynamics described in the last paragraph, we can associate a YHK process starting with a tree T_m with a corresponding urn process (C_0, R) as follows. The urn model contains four colours in which colour i $(1 \le i \le 4)$ is designated for type i edges. In the initial urn $C_0 = (C_{0,1}, \ldots, C_{0,4})$, the number $C_{0,i}$ is precisely the number of type i edges in T_m. Furthermore, the replacement matrix R is the following 4×4 matrix:

$$R = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 2 & -2 & 1 & 0 \\ -2 & 4 & -1 & 0 \\ 0 & 2 & 0 & -1 \end{bmatrix}. \tag{8}$$

Given an arbitrary tree T, let $\alpha(T) = \big(|E_1(T)|, |E_2(T)|, |E_3(T)|, |E_4(T)|\big)$ be the pendant type vector associated with T where $|E_i(T)|$ counts the number of type i edges in T for $1 \le i \le 4$.

The following result will enable us to obtain the joint distribution on pitchforks and cherries for the YHK model. Moreover, it also implies that the asymptotic behaviour of these two statistics, when appropriately scaled, is independent of the initial tree used in the YHK process.

Theorem 3 *Suppose that T_m is an arbitrary phylogenetic tree with m leaves with $m \ge 2$, and that T_n is a tree with n leaves generated by the YHK process starting with*

T_m. Then we have

$$\frac{\alpha(T_n)}{n} \xrightarrow{a.s.} \mathbf{v}_1 \quad \textit{and} \quad \frac{\alpha(T_n) - n\mathbf{v}_1}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma), \tag{9}$$

where $\mathbf{v}_1 = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{6}, \frac{1}{6}\right)$ *and*

$$\Sigma = \frac{1}{1260}\begin{bmatrix} 276 & -388 & 138 & -26 \\ -388 & 724 & -194 & -142 \\ 138 & -194 & 69 & -13 \\ -26 & -142 & -13 & 181 \end{bmatrix}. \tag{10}$$

Proof Consider the YHK process $\{T_n\}_{n\geq m}$ starting with T_m. Let $C_k = \alpha(T_{k+m})$ for $k \geq 0$. Then $C_k = (C_{k,1}, \ldots, C_{k,4})$, where $C_{k,i} = |E_i(T_{m+k})|$ for $1 \leq i \leq 4$, is the urn model of 4 colours derived from the pendant edge decomposition of the YHK process. Therefore, it is a tenable model with $C_0 = \alpha(T_m)$ and replacement matrix R as given in (8).

Note that R is diagonalisable as

$$U^{-1}RU = \Lambda$$

holds with

$$U = \begin{bmatrix} 1 & 1 & -1 & -1 \\ 1 & 0 & -1 & -3 \\ 1 & -2 & 2 & 5 \\ 1 & 0 & 2 & 3 \end{bmatrix}, \quad \Lambda = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & -3 \end{bmatrix}$$

$$\text{and} \quad U^{-1} = \frac{1}{6}\begin{bmatrix} 2 & 2 & 1 & 1 \\ 2 & -2 & -2 & 2 \\ -4 & 2 & -2 & 4 \\ 2 & -2 & 1 & -1 \end{bmatrix}. \tag{11}$$

Therefore, R satisfies condition (A4). Next, (A2) holds because R has eigenvalues

$$s = \lambda_1 = 1, \quad \lambda_2 = 0, \quad \lambda_3 = -2, \quad \lambda_4 = -3,$$

where $s = \lambda_1 = 1$ is the principal eigenvalue. Furthermore, put $\mathbf{u}_i = U\mathbf{e}_i^\top$ and $\mathbf{v}_i = \mathbf{e}_i U^{-1}$ for $1 \leq i \leq 4$. Then (A3) follows by noting that $\mathbf{u}_1 = (1, 1, 1, 1)^\top$ is the principal right eigenvector, and $\mathbf{v}_1 = \frac{1}{6}(2, 2, 1, 1)$ is the principal left eigenvector.

Since (A1)–(A4) are satisfied by the replacement matrix R, by Theorem 1 it follows that

$$\frac{C_k}{k} \xrightarrow{a.s.} \mathbf{v}_1 \text{ with } k \to \infty$$

and hence

$$\frac{\alpha(T_n)}{n} = \frac{n-m}{n}\frac{C_{n-m}}{n-m} \xrightarrow{a.s.} \mathbf{v}_1 \text{ with } n \to \infty.$$

By Theorem 2 we have

$$\frac{C_k - k v_1}{\sqrt{k}} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma) \text{ with } k \to \infty, \tag{12}$$

where

$$\Sigma = \sum_{i,j=2}^{4} \frac{\lambda_i \lambda_j \mathbf{u}_i^\top \mathrm{diag}(\mathbf{v}_1)\mathbf{u}_j}{1-\lambda_i-\lambda_j} \mathbf{v}_i^\top \mathbf{v}_j. \tag{13}$$

Therefore, we have

$$\begin{aligned}\frac{\alpha(T_n) - n\mathbf{v}_1}{\sqrt{n}} &= \frac{C_{n-m} - (n-m)\mathbf{v}_1}{\sqrt{n}} + \frac{m\mathbf{v}_1}{\sqrt{n}} \\ &= \frac{\sqrt{n-m}}{\sqrt{n}} \frac{C_{n-m} - (n-m)\mathbf{v}_1}{\sqrt{n-m}} + \frac{m\mathbf{v}_1}{\sqrt{n}} \\ &\xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma).\end{aligned}$$

Here the convergence follows from (12) and the fact that $\frac{\sqrt{n-m}}{\sqrt{n}}$ converges to 1 and $\frac{m\mathbf{v}_1}{\sqrt{n}}$ converges to 0 when n approaches infinity. □

By Theorem 3, it is straightforward to obtain the following result on the joint distribution of cherries and pitchforks, which also follows from a general result by Holmgren and Janson (2015, Theorem 1.22).

Corollary 1 *Under the YHK model, for the joint distribution* (A_n, B_n) *of pitchforks and cherries we have*

$$\frac{1}{n}(A_n, B_n) \xrightarrow{a.s.} \left(\frac{1}{6}, \frac{1}{3}\right) \tag{14}$$

and

$$\frac{(A_n, B_n) - n(1/6, 1/3)}{\sqrt{n}} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \frac{1}{1260}\begin{bmatrix} 69 & -28 \\ -28 & 56 \end{bmatrix}\right). \tag{15}$$

Proof Consider the YHK process $\{T_n\}_{n\geq 2}$ starting with a tree T_2 with two leaves. Denote the i-th entry in $\alpha(T_n)$ by $\alpha_{n,i}$ for $1 \leq i \leq 4$. Then the corollary follows from Theorem 3 by noting that we have $A_n = \frac{\alpha_{n,1}}{2}$ and $B_n = \frac{\alpha_{n,1}+\alpha_{n,2}}{2}$. □

The above result is consistent with the previously known results on the mean and (co-)variance of the joint distribution of cherries and pitchforks (see, e.g., Wu and Choi (2016); Choi et al. (2020)), namely, under the YHK model and for $n \geq 7$ we have

$$\mathbb{E}(A_n) = \frac{n}{6}, \quad \mathbb{E}(B_n) = \frac{n}{3}, \quad \mathbb{V}(A_n) = \frac{23n}{420}, \quad \mathbb{V}(B_n) = \frac{2n}{45}, \quad \text{and } Cov(A_n, B_n) = -\frac{n}{45}.$$

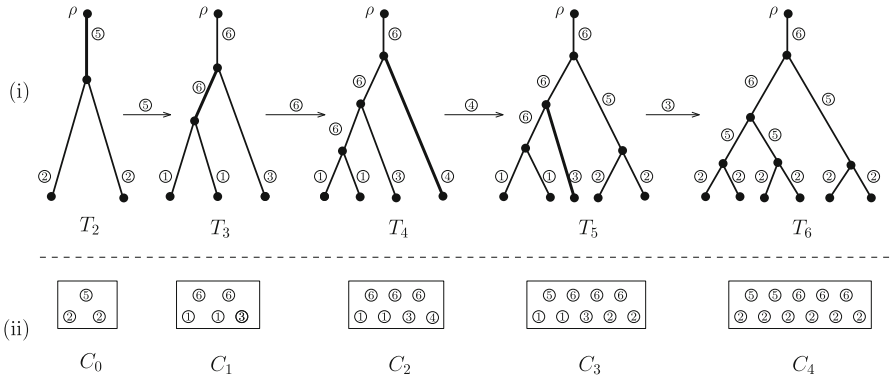

Fig. 3 A sample path of the PDA model and the associated urn model. (i) A sample path of the PDA model evolving from T_2 with two leaves to T_6 with six leaves. The labels of the leaves are omitted for simplicity. The edge types are indicated by circled numbers. For $2 \le i \le 5$, the edge selected in T_i to generate T_{i+1} is highlighted in bold and the associated edge type is indicated in the circled number above the arrows. (ii) The associated urn model with six colours, derived from the edge types in the trees. Note that in the vector form we have $C_0 = (0, 2, 0, 0, 1, 0), \ldots, C_3 = (2, 2, 1, 0, 1, 3)$, and $C_4 = (0, 6, 0, 0, 2, 3)$

5 Limiting distributions under the PDA model

In this section, we study the limiting joint distribution of the random variables A_n (i.e., the number of pitchforks) and B_n (i.e., the number of essential cherries) under the PDA model.

To study the PDA model, in addition to the four edge types (E1)–(E4) considered in Sect. 4, which partitions the set of pendant edges, we need two additional edge types concerning the internal edges. Specifically,

(E5): a type 5 edge is an internal edge adjacent to an independent cherry;
(E6): a type 6 edge is an internal edge that is not type 5.

For $1 \le i \le 6$, let $E_i(T)$ be the set of edges of type i. Then the edge sets $E_1(T), \ldots, E_6(T)$ form a partition of the edge set of T. That is, each edge in T belongs to one and only one $E_i(T)$. Furthermore, let $\beta(T) = \big(|E_1(T)|, \ldots, |E_6(T)|\big)$ be the type vector associated with T, where $|E_i(T)|$ counts the number of type i edges in T.

As illustrated in Fig. 3, the composition of edge types in $T[e]$, which is obtained from T by attaching an extra leaf to edge e, is determined by the composition of edge types in T and the type of e. First, if e is a pendant edge, the change of the composition of the pendant edge types in $T[e]$ is the same as described in Sect. 4, and the change of the composition of the interior edge types in $T[e]$ is described as follows:

(i) If e is type 1, then $|E_i(T[e])| - |E_i(T)|$ is 0 if $i = 5$, and 1 if $i = 6$;
(ii) if e is type 2, then $|E_i(T[e])| - |E_i(T)|$ is -1 if $i = 5$, and 2 if $i = 6$;
(iii) if e is type 3, then $|E_i(T[e])| - |E_i(T)|$ is 2 if $i = 5$, and -1 if $i = 6$;
(iv) if e is type 4, then $|E_i(T[e])| - |E_i(T)|$ is 1 if $i = 5$, and 0 if $i = 6$.

Finally, when e is type 5, the change it caused is the same of that of a type 2 edge, and when e is type 6, the change it caused is the same of that of type 1 edge. Therefore, we

can associate a PDA process starting with a tree T_0 with a corresponding urn process (C_0, R) as follows. The urn model contains six colours in which colour i $(1 \le i \le 6)$ is designated for type i edges. In the initial urn $C_0 = (C_{0,1}, \ldots, C_{0,6})$, the number $C_{0,i}$ is precisely the number of type i edges in T_0. Furthermore, the replacement matrix R is the following 6×6 matrix:

$$R = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 1 \\ 2 & -2 & 1 & 0 & -1 & 2 \\ -2 & 4 & -1 & 0 & 2 & -1 \\ 0 & 2 & 0 & -1 & 1 & 0 \\ 2 & -2 & 1 & 0 & -1 & 2 \\ 0 & 0 & 0 & 1 & 0 & 1 \end{bmatrix}. \tag{16}$$

Note that the replacement matrix for the YHK model in (8) is a submatrix of the replacement matrix in (16); and the last (respectively, second last) row in (16) is the same as its first (respectively, second) row. These two observations are direct consequences of the dynamic described above. The theorem below describes the asymptotic behaviour of $\beta(T_n)$, which enables us to deduce the asymptotic properties of the joint distribution of the number of pitchforks and the number of cherries for the PDA model in Corollary 2. Moreover, it also implies that the asymptotic behaviour of these two statistics, when appropriately scaled, is independent of the initial tree used in the PDA process.

Theorem 4 *Suppose that T_m is an arbitrary phylogenetic tree with m leaves with $m \ge 2$, and that T_n is a tree with n leaves generated by the PDA process starting with T_m.*

Then we have

$$\frac{\beta(T_n)}{n} \xrightarrow{a.s.} \mathbf{v}_1 \quad \text{and} \quad \frac{\beta(T_n) - n\mathbf{v}_1}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma), \tag{17}$$

as $n \to \infty$, where $\mathbf{v}_1 = \frac{1}{16}(2, 2, 1, 3, 1, 7)$ and

$$\Sigma = \frac{1}{64} \begin{bmatrix} 12 & -12 & 6 & -6 & -6 & 6 \\ -12 & 28 & -6 & -10 & 14 & -14 \\ 6 & -6 & 3 & -3 & -3 & 3 \\ -6 & -10 & -3 & 19 & -5 & 5 \\ -6 & 14 & -3 & -5 & 7 & -7 \\ 6 & -14 & 3 & 5 & -7 & 7 \end{bmatrix}. \tag{18}$$

Proof Consider the PDA process $\{T_n\}_{n \ge m}$ starting with T_m. Let $C_k = \beta(T_{k+m})$ for $k \ge 0$. Then $C_k = (C_{k,1}, \ldots, C_{k,6})$, where $C_{k,i} = |E_i(T_{m+k})|$ for $1 \le i \le 6$, is the urn model of 6 colours derived from the edge partition of the PDA process. Therefore, it is a tenable model with $C_0 = \beta(T_m)$ and replacement matrix R as given in (16).

Note that R is diagonalisable as

$$U^{-1} R U = \Lambda$$

holds with $\Lambda = \mathrm{diag}(2, 0, 0, 0, -2, -4)$ and

$$U = \begin{bmatrix} 1 & 2.5 & 2 & 1 & 1 & 1 \\ 1 & -2 & 1 & 0 & 1 & 5 \\ 1 & -8 & -1 & 1 & -3 & -9 \\ 1 & -1 & 1 & 1 & -3 & -5 \\ 1 & 3 & -1 & 1 & 1 & 5 \\ 1 & 1 & -1 & -1 & 1 & 1 \end{bmatrix} \quad \text{and} \quad U^{-1} = \frac{1}{176} \begin{bmatrix} 22 & 22 & 11 & 33 & 11 & 77 \\ 4 & -20 & -14 & 14 & 6 & 10 \\ 30 & 26 & -17 & 17 & -43 & -13 \\ 40 & -24 & 36 & -36 & 60 & -76 \\ 66 & -22 & 33 & -77 & -11 & 11 \\ -22 & 22 & -11 & 11 & 11 & -11 \end{bmatrix}. \tag{19}$$

Therefore, R satisfies condition (A4). Next, (A2) holds because R has eigenvalues (counted with multiplicity)

$$s = \lambda_1 = 2, \quad \lambda_2 = 0, \quad \lambda_3 = 0, \quad \lambda_4 = 0, \quad \lambda_5 = -2, \quad \lambda_6 = -4$$

where $s = \lambda_1 = 2$ is the principal eigenvalue. Furthermore, put $\mathbf{u}_i = U\mathbf{e}_i^\top$ and $\mathbf{v}_i = \mathbf{e}_i U^{-1}$ for $1 \le i \le 6$. Then (A3) follows by noting that $\mathbf{u}_1 = (1, 1, 1, 1, 1, 1)^\top$ is the principal right eigenvector, and $\mathbf{v}_1 = \frac{1}{16}(2, 2, 1, 3, 1, 7)$ is the principal left eigenvector.

The remainder of the proof is similar to the final part of the proof of Theorem 3, and hence we only outline the main steps. Since (A1)–(A4) are satisfied by the replacement matrix R, by Theorem 1 it follows that

$$\frac{C_k}{k} \xrightarrow{a.s.} \mathbf{v}_1 \text{ with } k \to \infty, \text{ and hence } \frac{\beta(T_n)}{n} = \frac{n-m}{n} \frac{C_{n-m}}{n-m} \xrightarrow{a.s.} \mathbf{v}_1 \text{ with } n \to \infty.$$

By Theorem 2 we have

$$\frac{C_{n-m} - (n-m)\mathbf{v}_1}{\sqrt{n-m}} = \frac{C_k - k v_1}{\sqrt{k}} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma), \tag{20}$$

where

$$\Sigma = \sum_{i,j=2}^{6} \frac{\lambda_i \lambda_j \mathbf{u}_i^\top \mathrm{diag}(\mathbf{v}_1) \mathbf{u}_j}{1 - \lambda_i - \lambda_j} \mathbf{v}_i^\top \mathbf{v}_j. \tag{21}$$

Therefore, we have

$$\frac{\beta(T_n) - n\mathbf{v}_1}{\sqrt{n}} = \frac{C_{n-m} - (n-m)\mathbf{v}_1}{\sqrt{n}} + \frac{m\mathbf{v}_1}{\sqrt{n}} = \frac{\sqrt{n-m}}{\sqrt{n}} \frac{C_{n-m} - (n-m)\mathbf{v}_1}{\sqrt{n-m}} + \frac{m\mathbf{v}_1}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma).$$

□

Similar to Corollary 1, by Theorem 4 it is straightforward to obtain the following result on the joint distribution of cherries and pitchforks.

Corollary 2 *Under the PDA model, for the joint distribution* (A_n, B_n) *of pitchforks and cherries we have*

$$\frac{1}{n}(A_n, B_n) \xrightarrow{a.s.} \left(\frac{1}{8}, \frac{1}{4}\right) \tag{22}$$

and

$$\frac{(A_n, B_n) - n(1/8, 1/4)}{\sqrt{n}} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \frac{1}{64}\begin{bmatrix} 3 & 0 \\ 0 & 4 \end{bmatrix}\right) \tag{23}$$

as $n \to \infty$.

Proof Consider the PDA process $\{T_n\}_{n\geq 2}$ starting with a tree T_2 with two leaves. Denote the i-th entry in $\beta(T_n)$ by $\beta_{n,i}$ for $1 \leq i \leq 6$. Then the corollary follows from Theorem 3 by noting that we have $A_n = \frac{\beta_{n,1}}{2}$ and $B_n = \frac{\beta_{n,1}+\beta_{n,2}}{2}$. □

The above result is consistent with the previously known results on the mean and (co-)variance of the joint distribution of cherries and pitchforks (see, e.g., Wu and Choi (2016); Choi et al. (2020)), namely, under the PDA model and for $n \geq 7$ we have

$$\mathbb{E}(A_n) = \frac{n(n-1)(n-2)}{2(2n-3)(2n-5)}, \qquad \mathbb{E}(B_n) = \frac{n(n-1)}{2(2n-5)}, \qquad \mathbb{V}(B_n) = \frac{n(n-1)(n-2)(n-3)}{2(2n-3)^2(2n-7)},$$
$$\mathbb{V}(A_n) = \frac{3(4n^3 - 40n^2 + 123n - 110)}{2(2n-5)(2n-7)(2n-9)}\,\mathbb{V}(B_n), \quad \text{and} \quad Cov(A_n, B_n) = \frac{-\mathbb{V}(B_n)}{(2n-7)}.$$

6 Unrooted trees

Although rooted phylogenetic trees are often preferred by biologists as time is explicitly shown, it is also important to consider unrooted phylogenetic trees. Indeed, many methods for building trees from real data can usually do so only up to the placement of the root, and thus produce unrooted trees first and then figure out the root position (see, e.g. Steel (2016, Section 1.3)). In this section, we extend our results in Sects. 4 and 5 to unrooted phylogenetic trees.

Formally, deleting the root ρ of a rooted phylogenetic tree and suppressing its adjacent interior vertex r results in an unrooted tree (see Fig. 4). The set of unrooted phylogenetic trees on $\{1, 2, \ldots, n\}$ is denoted by $\mathcal{T}'_n$. The YHK process on unrooted phylogenetic tree is similar to that on rooted ones stated in Sect. 2.2; the only difference is that at step (ii) we shall start with an unrooted phylogenetic tree T_m in $\mathcal{T}'_m$ for $m \geq 3$. Similar modification suffices for the PDA processes on unrooted phylogenetic trees; see Choi et al. (2020) for more details. Note that the concepts of cherries and pitchforks can be naturally extended to unrooted trees in $\mathcal{T}'_n$ for $n \geq 6$. Moreover, let A'_n and B'_n be the random variables counting the number of pitchforks and cherries in a random tree in $\mathcal{T}'_n$.

To associate urn models with the two processes on unrooted trees, note that for a tree T in $\mathcal{T}'_n$ with $n \geq 6$, we can decompose the edges in T into the six types similar to those for rooted trees, and hence define $\alpha(T)$ and $\beta(T)$ correspondingly. Furthermore,

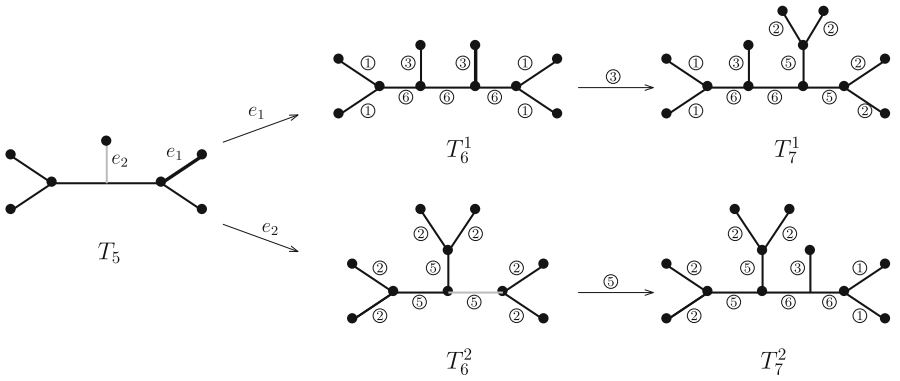

Fig. 4 Example of sample paths for the PDA process on unrooted trees and the associated urn model. Two sample paths of the PDA process evolving from T_5: one ends with T_7^1 using the edges in bold and the other with T_7^2 using the edges in grey. Leave labels are omitted for simplicity. Note that in the vector form we have $\beta(T_6^1) = (4, 0, 2, 0, 0, 3)$ and $\beta(T_6^2) = (0, 6, 0, 0, 3, 0)$

the replacement matrix is the same as the unrooted one, that is, the replacement matrix for the YHK model is given in (8) and the one for the PDA process is given in (16). See two examples in Fig. 4. We emphasize that the condition $n \geq 6$ is essential here: for instance, there is no appropriate assignment for the edge e_2 in the tree T_5 in Fig. 4 in our scheme, neither type 3 nor type 4 satisfying the requirement of a valid urn model. This observation is indeed in line with the treatment of unrooted trees in Choi et al. (2020). However, there is only one unrooted shape for $n = 4$ and one for $n = 5$. Furthermore, there are only two tree shapes for $\mathcal{T}_6'$ (as depicted in T_6^1 and T_6^2 in Fig. 4). In particular, putting $\alpha_6^1 = (4, 0, 2, 0)$ and $\alpha_6^2 = (0, 6, 0, 0)$, then for each T in $\mathcal{T}_6'$, we have either $\alpha(T) = \alpha_6^1$ or $\alpha(T) = \alpha_6^2$.

Now we extend Theorem 3 and Corollary 1 to the following result concerning the limiting behaviour of the YHK process. Similar to the rooted version, the asymptotic behaviour of the frequencies of cherries and pitchforks, when appropriately scaled, is independent of the initial trees used in the unrooted YHK process.

Theorem 5 *Suppose that T_m is an arbitrary unrooted phylogenetic tree with m leaves with $m \geq 6$, and that T_n is an unrooted tree with n leaves generated by the YHK process starting with T_m. Then, as $n \to \infty$,*

$$\frac{\alpha(T_n)}{n} \xrightarrow{a.s.} \mathbf{v}_1 \quad \text{and} \quad \frac{\alpha(T_n) - n\mathbf{v}_1}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma), \tag{24}$$

where $\mathbf{v}_1 = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{6}, \frac{1}{6}\right)$ and Σ is given in Eq. (10). In particular, as $n \to \infty$,

$$\frac{1}{n}(A_n', B_n') \xrightarrow{a.s.} \left(\frac{1}{6}, \frac{1}{3}\right) \quad \text{and} \quad \frac{(A_n', B_n') - n(1/6, 1/3)}{\sqrt{n}} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \frac{1}{1260}\begin{bmatrix} 69 & -28 \\ -28 & 56 \end{bmatrix}\right). \tag{25}$$

Proof The proof of (24) follows an argument similar to that for Theorem 4.

To establish (25), consider the YHK process $\{T_n\}_{n\geq 2}$ starting with a tree T_2 with two leaves. For $n \geq 6$, let $\alpha_n = \alpha(T_n)$ and $\alpha_{n,i}$ denote the i-th entry in $\alpha(T_n)$ for $1 \leq i \leq 4$.

Consider the vector $\alpha_6^1 = (4, 0, 2, 0)$ and $\alpha_6^2 = (0, 6, 0, 0)$. For $j = 1, 2$, let E_j be the event that $\alpha_6 = \alpha_6^j$. It follows that E_1 and E_2 form a partition of the sample space. Moreover, we have $\mathbb{P}(E_1) = 4/5$ and $\mathbb{P}(E_2) = 1 - \mathbb{P}(E_1) = 1/5$. Consider the random indicator variable $\mathbb{I}_{E_1}$, that is, $\mathbb{P}(\mathbb{I}_{E_1} = 1) = 4/5$ and $\mathbb{P}(\mathbb{I}_{E_1} = 0) = 1/5$. Random indicator variable $\mathbb{I}_{E_2}$ is similarly defined. Then we have

$$\alpha_n = \alpha_n^1 \mathbb{I}_{E_1} + \alpha_n^2 \mathbb{I}_{E_2}.$$

Furthermore, by (24) we have $\frac{\alpha_n^j}{n} \xrightarrow{a.s.} \mathbf{v}_1$ a.s. on E_j, for $j = 1, 2$, and hence

$$\frac{\alpha_n}{n} \xrightarrow{a.s.} \mathbf{v}_1(\mathbb{I}_{E_1} + \mathbb{I}_{E_2}) = \mathbf{v}_1.$$

Together with $A'_n = \frac{\alpha_{n,1}}{2}$ and $B'_n = \frac{\alpha_{n,1}+\alpha_{n,2}}{2}$, the almost surely convergence in (25) follows. Finally, the convergence in distribution in (25) also follows from a similar argument. □

Finally, combining Theorem 4, Corollary 2, and an argument similar to the proof of Theorem 5 leads to the following result concerning the limiting behaviour of the unrooted PDA process, whose proof is hence omitted.

Theorem 6 *Suppose that T_m is an arbitrary unrooted phylogenetic tree with m leaves with $m \geq 6$, and that T_n is an unrooted tree with n leaves generated by the PDA process starting with T_m.*

Then, as $n \to \infty$,

$$\frac{\beta(T_n)}{n} \xrightarrow{a.s.} \mathbf{v}_1 \quad \text{and} \quad \frac{\beta(T_n) - n\mathbf{v}_1}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma), \tag{26}$$

where $\mathbf{v}_1 = \frac{1}{16}(2, 2, 1, 3, 1, 7)$ and Σ is given in Eq. (18). In particular, as $n \to \infty$,

$$\frac{1}{n}(A'_n, B'_n) \xrightarrow{a.s.} \left(\frac{1}{8}, \frac{1}{4}\right) \quad \text{and} \quad \frac{(A'_n, B'_n) - n(1/8, 1/4)}{\sqrt{n}} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \frac{1}{64}\begin{bmatrix} 3 & 0 \\ 0 & 4 \end{bmatrix}\right). \tag{27}$$

7 Proofs of Theorems 1 and 2

In this section, we shall present the proofs of Theorems 1 and 2. To this end, it is more natural to consider $Y_n := C_n U$, a linear transform of C_n. Next we introduce

$$\xi_n = Y_n - \mathbb{E}[Y_n | \mathcal{F}_{n-1}]. \tag{28}$$

For $1 \leq j \leq d$, consider the following numbers

$$b_{n,n}(j) = 1 \quad \text{and} \quad b_{n,k}(j) = \prod_{\ell=k}^{n-1} (1 + \lambda_j / t_\ell) \quad \text{for } 0 \leq k < n. \tag{29}$$

Moreover, we introduce the following diagonal matrix for $0 \le k \le n$:

$$\mathbf{B}_{n,k} = \operatorname{diag}\left(b_{n,k}(1), \ldots, b_{n,k}(d)\right). \tag{30}$$

Then we have the following key observation:

$$Y_n = Y_0 \mathbf{B}_{n,0} + \sum_{k=1}^{n} \xi_k \, \mathbf{B}_{n,k}. \tag{31}$$

To see that (31) holds, let $Q_k = \mathbf{I} + t_{k-1}^{-1} R$ for $1 \le k \le n$, where $\mathbf{I}$ is the identity matrix. Then we have

$$\mathbb{E}[C_n|\mathcal{F}_{n-1}] = C_{n-1} + t_{n-1}^{-1} C_{n-1} R = C_{n-1}\left[\mathbf{I} + t_{n-1}^{-1} R\right] = C_{n-1} Q_n.$$

As $C_k - \mathbb{E}[C_k|\mathcal{F}_{k-1}] = \xi_k U^{-1}$ for $1 \le k \le n$, we have

$$\begin{aligned} C_n &= (C_n - \mathbb{E}[C_n|\mathcal{F}_{n-1}]) + C_{n-1} Q_n = \xi_n U^{-1} + C_{n-1} Q_n \\ &= C_0 (Q_1 \cdots Q_n) + \xi_n U^{-1} + \sum_{k=1}^{n-1} \xi_k U^{-1} (Q_{k+1} \cdots Q_n). \end{aligned} \tag{32}$$

Since

$$U^{-1}\Big(\prod_{\ell=k+1}^{n} Q_\ell\Big) U = \prod_{\ell=k}^{n-1} \left(U^{-1}\left(I + t_\ell^{-1} R\right) U\right) = \prod_{\ell=k}^{n-1} \left(I + t_\ell^{-1} \Lambda\right) = \mathbf{B}_{n,k} \tag{33}$$

holds for $0 \le k \le n$ and $Y_n = C_n U$, it is straightforward to see that (31) follows from transforming (32) by a right multiplication of U.

Next, we shall present several properties concerning ξ_k. To this end, consider the sequence of random vectors $\tau_k = \chi_k - \mathbb{E}[\chi_k|\mathcal{F}_{k-1}]$ for $k \ge 1$. Then $\{\tau_k\}_{k\ge 1}$ is a martingale difference sequence (MDS) in that $\mathbb{E}[\tau_k|\mathcal{F}_{k-1}] = \mathbf{0}$ almost surely. Hence $\mathbb{E}[\tau_k] = \mathbb{E}\big[\mathbb{E}[\tau_k|\mathcal{F}_{k-1}]\big] = \mathbf{0}$. Furthermore, since the entries in χ_k are either 0 or 1 and $\mathbb{E}[\chi_k|\mathcal{F}_{k-1}] = \widetilde{C}_{k-1}$, the random vector τ_k is also bounded. As a bounded martingale difference sequence, τ_k is uncorrelated. To see it, assuming that $\ell < k$, then we have

$$\mathbb{E}[\tau_\ell^\top \tau_k] = \mathbb{E}\big[\mathbb{E}[\tau_\ell^\top \tau_k|\mathcal{F}_{k-1}]\big] = \mathbb{E}\big[\tau_\ell^\top \mathbb{E}[\tau_k|\mathcal{F}_{k-1}]\big] = \mathbb{E}[\tau_\ell^\top \mathbf{0}] = \mathbf{0}^\top \mathbf{0},$$

where the first equality follows from the total law of expectation and the second from τ_ℓ is $\mathcal{F}_{k-1}$-measurable. A similar argument shows $\mathbb{E}[\tau_\ell \tau_k^\top] = 0$. Consequently, we have the following expression showing that distinct τ_k and τ_ℓ are uncorrelated:

$$\mathbb{E}[\tau_k^\top \tau_\ell] = \mathbf{0}^\top \mathbf{0} \text{ and } \mathbb{E}[\tau_k \tau_\ell^\top] = 0 \quad \text{if } k \ne \ell. \tag{34}$$

Moreover, putting

$$\Gamma_k := \operatorname{diag}\big(\widetilde{C}_k\big) - \widetilde{C}_k^\top \widetilde{C}_k,$$

then we have

$$\mathbb{E}[\Gamma_k] = \operatorname{diag}\big(\mathbb{E}[\widetilde{C}_k]\big) - \mathbb{E}\big[\widetilde{C}_k^\top \widetilde{C}_k\big].$$

Consequently, we have

$$\begin{aligned}
\mathbb{E}[\tau_k^\top \tau_k|\mathcal{F}_{k-1}] &= \mathbb{E}[\big(\chi_k - \mathbb{E}[\chi_k|\mathcal{F}_{k-1}]\big)^\top \big(\chi_k - \mathbb{E}[\chi_k|\mathcal{F}_{k-1}]\big)|\mathcal{F}_{k-1}] \\
&= \mathbb{E}[(\chi_k^\top - \widetilde{C}_{k-1}^\top)(\chi_k - \widetilde{C}_{k-1})|\mathcal{F}_{k-1}] \\
&= \mathbb{E}[\chi_k^\top \chi_k|\mathcal{F}_{k-1}] - \widetilde{C}_{k-1}^\top \mathbb{E}[\chi_k|\mathcal{F}_{k-1}] - \mathbb{E}[\chi_k^\top|\mathcal{F}_{k-1}]\widetilde{C}_{k-1} + \widetilde{C}_{k-1}^\top \widetilde{C}_{k-1} \\
&= \mathbb{E}[\chi_k^\top \chi_k|\mathcal{F}_{k-1}] - \widetilde{C}_{k-1}^\top \widetilde{C}_{k-1} = \Gamma_{k-1}, \qquad (35)
\end{aligned}$$

where the last equality follows from (2). This implies

$$\mathbb{E}[\tau_k^\top \tau_k] = \mathbb{E}\big[\mathbb{E}[\tau_k^\top \tau_k|\mathcal{F}_{k-1}]\big] = \mathbb{E}[\Gamma_{k-1}]. \qquad (36)$$

Note that ξ_k is a 'linear transform' of τ_k in that combining (1) and (28) leads to

$$\begin{aligned}
\xi_k &= \big(C_k - \mathbb{E}[C_k|\mathcal{F}_{k-1}]\big)U = \big(C_{k-1} + \chi_k R - \mathbb{E}[C_{k-1} + \chi_k R|\mathcal{F}_{k-1}]\big)U \\
&= \big(\chi_k - \mathbb{E}[\chi_k|\mathcal{F}_{k-1}]\big)RU = \tau_k RU = \tau_k U\Lambda. \qquad (37)
\end{aligned}$$

Note this implies that ξ_k is a martingale difference sequence in that $\mathbb{E}[\xi_k|\mathcal{F}_{k-1}] = \mathbf{0} = \mathbb{E}[\xi_k]$. Furthermore, by (35) and (37) we have

$$\mathbb{E}\big[\xi_k^\top \xi_k|\mathcal{F}_{k-1}\big] = \Lambda U^\top \Gamma_{k-1} U\Lambda \quad \text{for } k \geq 1. \qquad (38)$$

Together with (34) and (36), for all $k, \ell \geq 1$ we have

$$\mathbb{E}[\xi_k^\top \xi_k] = \Lambda U^\top \mathbb{E}[\Gamma_{k-1}] U\Lambda, \quad \text{and} \quad \mathbb{E}[\xi_k^\top \xi_\ell] = \mathbf{0}^\top \mathbf{0} \text{ if } k \neq \ell. \qquad (39)$$

Since $\mathbf{u}_1 = U\mathbf{e}_1^\top = \mathbf{e}^\top$ is a right eigenvector of R corresponding to s, by (37) we have

$$\xi_k \mathbf{e}_1^\top = \tau_k RU\mathbf{e}_1^\top = \tau_k R\mathbf{u}_1 = s\tau_k \mathbf{u}_1 = s\tau_k \mathbf{e}^\top = 0 \text{ for } k \geq 1, \qquad (40)$$

where the last equality follows from $\chi_k \mathbf{e}^\top = 1$ and $\mathbb{E}[\chi_k|\mathcal{F}_{k-1}]\mathbf{e}^\top = \widetilde{C}_{k-1}\mathbf{e}^\top = 1$.

Note that for $n > 1$ and $\rho < 1$, we have

$$\frac{1}{n}\sum_{k=1}^{n-1}\Big(\frac{n}{k}\Big)^\rho \leq \frac{1}{1-\rho}, \quad \text{and} \quad \lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}\Big(\frac{n}{k}\Big)^\rho = \int_0^1 x^{-\rho}dx = \frac{1}{1-\rho}. \qquad (41)$$

Furthermore, we present the following two results on the entries of $\mathbf{B}_{n,k}$, whose proofs are elementary calculus and included in the appendix.

Lemma 1 *Under assumptions (A2) and (A3), there exists a constant K such that*

$$|b_{n,0}(j)| \le K n^{\lambda_j/s} \quad \text{and} \quad |b_{n,k}(j)| \le K (n/k)^{\lambda_j/s} \tag{42}$$

hold for $1 \le j \le d$ and $1 \le k \le n$. Furthermore, we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} b_{n,k}(i) b_{n,k}(j) = \frac{s}{s - \lambda_i - \lambda_j} \qquad \text{for } 2 \le i \le j \le d. \tag{43}$$

Corollary 3 *Assume that $\{Z_n\}$ is a sequence of random variables such that*

$$Z_n \xrightarrow{1} Z$$

for a random variable Z. Then under assumptions (A2)-(A3), for $2 \le i \le j \le d$ we have

$$\frac{1}{n} \sum_{k=1}^{n} b_{n,k}(i) b_{n,k}(j) Z_k \xrightarrow{p} \frac{s}{s - \lambda_i - \lambda_j} Z \quad \text{as } n \to \infty. \tag{44}$$

7.1 Proof of Theorem 1

Proof Recall that $Y_n = C_n U$ for $n \ge 1$. Hence, it is sufficient to show that

$$n^{-1} Y_n \xrightarrow{a.s.} s\, \mathbf{e}_1 \tag{45}$$

because $s\, \mathbf{e}_1 U^{-1} = s\, \mathbf{v}_1$ and $n^{-1} C_n = n^{-1} Y_n U^{-1}$. Furthermore, as the sequence of random vectors $n^{-1} C_n$ is bounded, its L^r convergence follows from the almost sure convergence.

To establish (45), we restate the following decomposition from (31) as below:

$$Y_n = Y_0 \mathbf{B}_{n,0} + \sum_{k=1}^{n} \xi_k\, \mathbf{B}_{n,k}, \tag{46}$$

where $\{\xi_k\}$ is the martingale difference sequence in (28) and $\mathbf{B}_{n,k}$ is the diagonal matrix in (30).

Next we claim that

$$n^{-1} \mathbb{E}[Y_n] \longrightarrow s\, \mathbf{e}_1 \ \text{ as } n \to \infty. \tag{47}$$

Indeed, since $\mathbb{E}[\xi_k] = \mathbf{0}$ implies $\mathbb{E}[\xi_k\, \mathbf{B}_{n,k}] = \mathbb{E}[\xi_k] \mathbf{B}_{n,k} = \mathbf{0}$, by (46) we have $\mathbb{E}[Y_n] = Y_0 \mathbf{B}_{n,0}$. Therefore the j-th entry in $\mathbb{E}[Y_n]$, denoted by $y_{n,j}$, is given by

$$y_{n,j} = \mathbb{E}[Y_n] \mathbf{e}_j^\top = Y_0 \mathbf{B}_{n,0} \mathbf{e}_j^\top = b_{n,0}(j) Y_0 \mathbf{e}_j^\top \text{ for } 1 \le j \le d.$$

When $j = 1$, we have

$$y_{n,1} = b_{n,0}(1) Y_0 \mathbf{e}_1^\top = (t_n/t_0) Y_0 \mathbf{e}_1^\top = (t_n/t_0) C_0 U \mathbf{e}_1^\top = (t_n/t_0) C_0 \mathbf{u}_1 = (t_n/t_0) t_0 = t_n,$$

where we used the fact that $\mathbf{u}_1 = \mathbf{e}^\top$ and hence $t_0 = C_0\mathbf{u}_1$. Therefore we have $y_{n,1}/n = t_n/n \to s$ as $n \to \infty$. On the other hand, for $2 \le j \le d$ there exist two constants K_1 and K such that

$$|y_{n,j}| = |b_{n,0}(j)Y_0\mathbf{e}_j^\top| \le K_1|b_{n,0}(j)| \le K_1 K n^{\lambda_j/s},$$

where the last inequality follows from Lemma 1. Since $\lambda_j < s$, it follows that $y_{n,j}/n \to 0$ as $n \to \infty$. This completes the proof of (47).

For simplicity, let $Z_n := Y_n - \mathbb{E}(Y_n)$. Then we have $Y_n = Z_n + \mathbb{E}(Y_n)$, by (47) it follows that to establish (45), it remains to show that

$$Z_n/n \xrightarrow{a.s.} \mathbf{0}, \tag{48}$$

Denote the j-th entry in Z_n by $Z_{n,j}$, then from (46) we have

$$Z_{n,j} = \sum_{k=1}^{n}(\xi_k\mathbf{B}_{n,k})\mathbf{e}_j^\top = \sum_{k=1}^{n} b_{n,k}(j)\xi_k\mathbf{e}_j^\top. \tag{49}$$

Since (48) is equivalent to

$$\frac{Z_{n,j}}{n} \xrightarrow{a.s.} 0 \qquad \text{for } 1 \le j \le d, \tag{50}$$

the remainder of the proof is devoted to establishing (50).

It is straightforward to see that (50) holds for $j = 1$ because by (40) and (49) we have

$$Z_{n,1} = \sum_{k=1}^{n} b_{n,k}(j)\xi_k\mathbf{e}_1^\top = 0.$$

Thus in the remainder of the proof, we may assume that $2 \le j \le d$ holds. Note that

$$\begin{aligned}\mathbb{E}\big[Z_{n,j}^2\big] &= \mathbb{E}\bigg[\bigg(\sum_{k=1}^{n} b_{n,k}(j)\xi_k\mathbf{e}_j^\top\bigg)^2\bigg] = \mathbb{E}\big[\sum_{k,\ell=1}^{n} b_{n,k}(j)b_{n,\ell}(j)\mathbf{e}_j\xi_k^\top\xi_l\mathbf{e}_j^\top\big] \\ &= \mathbb{E}\big[\sum_{k=1}^{n} b_{n,k}^2(j)\mathbf{e}_j\xi_k^\top\xi_k\mathbf{e}_j^\top\big] = \sum_{k=1}^{n} b_{n,k}^2(j)\mathbb{E}\big[\mathbf{e}_j\xi_k^\top\xi_k\mathbf{e}_j^\top\big].\end{aligned}$$

Here the third equality follows from (39). As $\mathbb{E}[\mathbf{e}_j\xi_k^\top\xi_k\mathbf{e}_j^\top]$, the (j, j)-entry of matrix $\mathbb{E}[\xi_k^\top\xi_k]$, is bounded above by a constant K_1 in view of (39), there exist constants K_2

and K so that

$$\mathbb{E}\big[Z_{n,j}^2\big] \le K_1 \sum_{k=1}^{n} |b_{n,k}(j)|^2 \le K_2 \sum_{k=1}^{n} \left(\frac{n}{k}\right)^{2\lambda_j/s} = K_2 + K_2 n \sum_{k=1}^{n-1} \frac{1}{n} \left(\frac{k}{n}\right)^{-2\lambda_j/s}$$
$$\le K_2 + \frac{K_2 n}{1 - 2\lambda_j/s} \le Kn$$

holds for all $n \ge 1$. Here the second inequality follows from Lemma 1 and the third one from (41) in view of $\lambda_j < s/2$ for $2 \le j \le d$.

Since $\mathbb{E}(Z_{n,j}) = 0$, for $\epsilon > 0$ using the Chebychev inequality we get

$$\mathbb{P}\left(\left|Z_{n,j}\right| > n\epsilon\right) \le \frac{K}{n\epsilon^2} \quad \text{for all } n \ge 1. \tag{51}$$

Consider the subsequence $Z'_{n,j}$ of $Z_{n,j}$ with $Z'_{n,j} = Z_{n^2,j}$ for $n \ge 1$. Then for $\epsilon > 0$ we have

$$\sum_{n=1}^{\infty} \mathbb{P}\left(\frac{|Z'_{n,j}|}{n^2} > \epsilon\right) = \sum_{n=1}^{\infty} \mathbb{P}\left(\left|Z_{n^2,j}\right| > n^2\epsilon\right) \le \sum_{n=1}^{\infty} \frac{K}{n^2\epsilon^2} < \infty,$$

where the first inequality follows from (51). Thus, by the Borel-Cantelli Lemma, it follows that

$$n^{-2} Z'_{n,j} \xrightarrow{a.s.} 0. \tag{52}$$

Next, consider

$$\Delta_{n,j} := \max_{n^2 \le k < (n+1)^2} |Z_{k,j} - Z'_{n,j}| = \max_{n^2 \le k < (n+1)^2} |Z_{k,j} - Z_{n^2,j}| = \max_{1 \le k \le 2n} |Z_{n^2+k,j} - Z_{n^2,j}|.$$

Since for each $\ell > 0$, elements of χ_ℓ and RU are all bounded above, there exists a constant K independent of ℓ and j so that

$$|Z_{\ell+1,j} - Z_{\ell,j}| = |\big((C_{\ell+1} - \mathbb{E}[C_{\ell+1}]) - (C_\ell - \mathbb{E}[C_\ell])\big) U \mathbf{e}_j^\top|$$
$$= |\big(C_{\ell+1} - C_\ell\big) - \big(\mathbb{E}[C_{\ell+1} - C_\ell]\big) U \mathbf{e}_j^\top| = |\big(\chi_{\ell+1} - \mathbb{E}[\chi_{\ell+1}]\big) RU \mathbf{e}_j^\top| \le K.$$

Consequently, we have

$$\Delta_{n,j} = \max_{0 \le k \le 2n} |Z_{n^2+k,j} - Z_{n^2,j}| \le \max_{1 \le k \le 2n} \sum_{\ell=1}^{k} |Z_{n^2+\ell,j} - Z_{n^2+\ell-1,j}| \le \max_{1 \le k \le 2n} \sum_{\ell=1}^{k} K = 2nK,$$

and hence

$$n^{-2} \Delta_{n,j} \xrightarrow{a.s.} 0. \tag{53}$$

Now, for each $k > 0$, considering the natural number n with $n^2 \le k < (n+1)^2$, then we have

$$\frac{|Z_{k,j}|}{k} \le \frac{|Z_{k,j} - Z_{n^2,j}|}{k} + \frac{|Z_{n^2,j}|}{k} \le \frac{\Delta_{n,j}}{n^2} + \frac{|Z_{n^2,j}|}{n^2} = \frac{\Delta_{n,j}}{n^2} + \frac{|Z'_{n,j}|}{n^2}. \tag{54}$$

Note that when $k \to \infty$, the natural number n satisfying $n^2 \le k < (n+1)^2$ also approaches to ∞. Thus combining (52), (53), and (54) leads to

$$k^{-1} Z_{k,j} \xrightarrow{a.s.} 0 \qquad \text{when } k \to \infty, \tag{55}$$

which completes the proof of (50), and hence also the theorem. □

7.2 Proof of Theorem 2

Proof For each $n \ge 1$, consider the following two sequences of random vectors:

$$X_{n,k} := n^{-1/2} \xi_k \mathbf{B}_{n,k} \quad \text{and} \quad S_{n,k} := \sum_{\ell=1}^{k} X_{n,\ell} \qquad \text{for } 1 \le k \le n,$$

where $\{\xi_k\}_{k\ge1}$ is the martingale difference sequence in (28) and $\mathbf{B}_{n,k}$ is the diagonal matrix in (30). Then for each $n \ge 1$, the sequence $\{X_{n,k}\}_{1\le k\le n}$ is a martingale difference sequence, and $\{S_{n,k}\}_{1\le k\le n}$ is a mean zero martingale.

Recalling that $Y_n = C_n U$, then by (31) we have

$$S_{n,n} = n^{-1/2} \sum_{k=1}^{n} \xi_k \mathbf{B}_{n,k} = n^{-1/2} \big(Y_n - \mathbb{E}[Y_n] \big). \tag{56}$$

A key step in our proof is to show that

$$S_{n,n} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \widetilde{\Sigma}), \tag{57}$$

where $\mathcal{N}(\mathbf{0}, \widetilde{\Sigma})$ denotes a normal distribution with mean vector $\mathbf{0}$ and variance-covariance matrix

$$\widetilde{\Sigma} := \sum_{i,j=2}^{d} \frac{s\lambda_i\lambda_j \mathbf{u}_i^\top \mathrm{diag}(\mathbf{v}_1)\mathbf{u}_j}{s - \lambda_i - \lambda_j} \mathbf{e}_i^\top \mathbf{e}_j. \tag{58}$$

We shall show that Theorem 2 follows from (57). To this end, we claim that

$$Z_n := n^{-1/2} \left(\mathbb{E}[Y_n] - ns\mathbf{e}_1 \right) \longrightarrow \mathbf{0} \quad \text{with } n \to \infty. \tag{59}$$

Indeed, we have $Z_n \mathbf{e}_1^\top = n^{-1/2}(t_n - ns) = n^{-1/2} t_0 \to 0$. Furthermore, by Lemma 1 there exists a constant K such that

$$|Z_n \mathbf{e}_j^\top| = n^{-1/2}|Y_{0,j} b_{n,0}(j)| = n^{-1/2} Y_{0,j}|b_{n,0}(j)| \le n^{-1/2} Y_{0,j} K n^{\lambda_j/s} \qquad \text{for } 2 \le j \le d.$$

As $\lambda_j/s < 1/2$, it follows that $|Z_n \mathbf{e}_j^\top| \to 0$ for all $1 \le j \le d$, and hence (59) holds. Consequently, we have

$$n^{-1/2}\,(Y_n - ns\mathbf{e}_1) = n^{-1/2}\,(Y_n - \mathbb{E}[Y_n]) + Z_n = S_{n,n} + Z_n \xrightarrow{d} N(\mathbf{0}, \widetilde{\Sigma}). \quad (60)$$

Here the second equality follows from (56); convergence in distribution follows from the Slutsky theorem (see, e.g. Grimmett and Stirzaker (2001, P. 318)) in view of (57) and (59). Since $n^{-1/2}(C_n - ns\mathbf{v}_1) = n^{-1/2}\,(Y_n - ns\mathbf{e}_1)\,V$ with $V = U^{-1}$, by (60) and the fact that a linear transform of a normal vector is also normal (see, e.g. Grimmett and Stirzaker (2001, Section 4.9)), we have

$$n^{-1/2}(C_n - ns\mathbf{v}_1) \xrightarrow{d} N(\mathbf{0}, \Sigma), \quad (61)$$

where

$$\Sigma = V^\top \widetilde{\Sigma}\, V = V^\top \Big(\sum_{i,j=2}^{d} \frac{s\lambda_i\lambda_j \mathbf{u}_i^\top \mathrm{diag}(\mathbf{v}_1)\mathbf{u}_j}{s - \lambda_i - \lambda_j} \mathbf{e}_i^\top \mathbf{e}_j \Big) V = \sum_{i,j=2}^{d} \frac{s\lambda_i\lambda_j \mathbf{u}_i^\top \mathrm{diag}(\mathbf{v}_1)\mathbf{u}_j}{s - \lambda_i - \lambda_j} \mathbf{v}_i^\top \mathbf{v}_j, \quad (62)$$

which shows indeed that the theorem follows from (57).

What remains is to prove (57). Define

$$\Phi(n) := \sum_{k=1}^{n} \mathbb{E}\big[X_{n,k}^\top X_{n,k} | \mathcal{F}_{k-1}\big] = \frac{1}{n} \sum_{k=1}^{n} \mathbf{B}_{n,k} \mathbb{E}[\xi_k^\top \xi_k | \mathcal{F}_{k-1}] \mathbf{B}_{n,k}.$$

We next show that

$$\Phi(n) \xrightarrow{p} \widetilde{\Sigma}. \quad (63)$$

Let $\Gamma = \mathrm{diag}(\mathbf{v}_1) - \mathbf{v}_1^\top \mathbf{v}_1$. Note that for $2 \le i, j \le d$, we have $\mathbf{v}_1 \mathbf{u}_i = 0 = \mathbf{v}_1 \mathbf{u}_j$ in view of (4), and hence

$$\frac{s\lambda_i\lambda_j \mathbf{u}_i^\top \Gamma \mathbf{u}_j}{s - \lambda_i - \lambda_j} = \frac{s\lambda_i\lambda_j \mathbf{u}_i^\top (\mathrm{diag}(\mathbf{v}_1) - \mathbf{v}_1^\top \mathbf{v}_1)\mathbf{u}_j}{s - \lambda_i - \lambda_j} = \frac{s\lambda_i\lambda_j \mathbf{u}_i^\top \mathrm{diag}(\mathbf{v}_1)\mathbf{u}_j}{s - \lambda_i - \lambda_j}.$$

Therefore (63) is equivalent to

$$\mathbf{e}_i \Phi(n) \mathbf{e}_j^\top \xrightarrow{p} \begin{cases} \dfrac{s\lambda_i\lambda_j\, \mathbf{u}_i^\top \Gamma \mathbf{u}_j}{s - \lambda_i - \lambda_j} & 2 \le i, j \le d, \\ 0 & \text{if } i = 1 \text{ or } j = 1. \end{cases} \quad (64)$$

Since $\mathbf{B}_{n,k}$ is a diagonal matrix and $\mathbf{e}_1\xi_k^\top = 0$ in view of (40), this implies

$$\mathbf{e}_1\Phi(n) = \frac{1}{n}\sum_{k=1}^{n}\mathbf{e}_1\mathbf{B}_{n,k}\mathbb{E}[\xi_k^\top\xi_k \mid \mathcal{F}_{k-1}]\mathbf{B}_{n,k} = \frac{1}{n}\sum_{k=1}^{n} b_{n,k}(1)\mathbb{E}[\mathbf{e}_1\xi_k^\top\xi_k \mid \mathcal{F}_{k-1}]\mathbf{B}_{n,k} = \mathbf{0}.$$

A similar argument shows $\Phi(n)\mathbf{e}_1^\top = \mathbf{0}$, and hence (64) holds for $i = 1$ or $j = 1$. It remains to consider the case $2 \le i, j \le d$. Since

$$\widetilde{C}_k \xrightarrow{1} \mathbf{v}_1 \quad \text{and} \quad \widetilde{C}_k^\top\widetilde{C}_k \xrightarrow{1} \mathbf{v}_1^\top\mathbf{v}_1$$

hold in view of Theorem 1,

we have

$$\lambda_i\lambda_j\mathbf{u}_i^\top\Gamma_k\mathbf{u}_j \xrightarrow{1} \lambda_i\lambda_j\mathbf{u}_i^\top\Gamma\mathbf{u}_j \ \text{ as } k\to\infty. \tag{65}$$

As both $\mathbf{B}_{n,k}$ and Λ are diagonal matrices, we have

$$\begin{aligned}\frac{1}{n}\sum_{i=k}^{n}\mathbf{e}_i\mathbf{B}_{n,k}(\Lambda U^\top\Gamma_{k-1}U\Lambda)\mathbf{B}_{n,k}\,\mathbf{e}_j^\top &= \frac{1}{n}\sum_{k=1}^{n} b_{n,k}(i)b_{n,k}(j)\mathbf{e}_i\Lambda U^\top\Gamma_{k-1}U\Lambda\mathbf{e}_j^\top\\ &= \frac{\lambda_i\lambda_j}{n}\sum_{k=1}^{n} b_{n,k}(i)b_{n,k}(j)\mathbf{u}_i^\top\Gamma_{k-1}\mathbf{u}_j \xrightarrow{p} \frac{s\lambda_i\lambda_j\mathbf{u}_i^\top\Gamma\mathbf{u}_j}{s-\lambda_i-\lambda_j},\end{aligned} \tag{66}$$

where the convergence follows from Corollary 3 and (65).

Since $S_{n,n}$ is a mean $\mathbf{0}$ random vector and $\mathbf{B}_{n,k}$ is a diagonal matrix, we have

$$\begin{aligned}\mathbb{V}\left[S_{n,n}\right] = \mathbb{E}[S_{n,n}^\top S_{n,n}] &= \frac{1}{n}\sum_{k,\ell=1}^{n}\mathbf{B}_{n,k}^\top\mathbb{E}[\xi_k^\top\xi_\ell]\mathbf{B}_{n,\ell} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{B}_{n,k}\mathbb{E}[\xi_k^\top\xi_k]\mathbf{B}_{n,k}\\ &= \sum_{k=1}^{n}\mathbb{E}[X_{n,k}^\top X_{n,k}] = \mathbb{E}[\Phi(n)]\end{aligned}$$

where the third equality follows from (39).

Furthermore, an argument similar to the proof of (63) shows that

$$\lim_{n\to\infty}\mathbb{V}(S_{n,n}) = \widetilde{\Sigma}.$$

Therefore $\widetilde{\Sigma}$ is positive semi-definite because the matrix $\mathbb{V}(S_{n,n})$ is necessarily positive semi-definite for each $n \ge 1$.

Following the Cramér-Wold device for the multivariate central limit theorem (see, e.g. Durrett (2019, Theorem 3.10.6)), fix an arbitrary row vector $\mathbf{w} = (w_1, \ldots, w_d)$ in $\mathbb{R}^d \setminus \{\mathbf{0}\}$ and put $s_{n,k} = S_{n,k}\mathbf{w}^\top$ and $x_{n,k} = X_{n,k}\mathbf{w}^\top$. Furthermore, since the matrix $\widetilde{\Sigma}$ is positive semi-definite, we can introduce $\sigma^2 := \mathbf{w}\,\widetilde{\Sigma}\,\mathbf{w}^\top \ge 0$. Then for

establishing (57) it suffices to show that

$$s_{n,n} \xrightarrow{d} N(0, \sigma^2). \tag{67}$$

Since $\{x_{n,k}\}_{1\le k\le n}$ is a martingale difference sequence and $\{s_{n,k}\}_{1\le k\le n}$ is an array of mean zero martingale, the martingale central limit theorem (see, e.g. Hall and Heyde (2014, Corollary 3.2)) implies that (67) follows from

$$\gamma_n := \sum_{k=1}^{n} \mathbb{E}\left[|x_{n,k}|^2 \, | \mathcal{F}_{k-1} \right] \xrightarrow{p} \sigma^2 \qquad \text{as } n \to \infty \tag{68}$$

and the conditional Lindeberg-type condition holds, that is, for every $\epsilon > 0$

$$\gamma_n^* := \sum_{k=1}^{n} \mathbb{E}\left[|x_{n,k}|^2 \, \mathbb{I}_{A_{n,k,\epsilon}} | \mathcal{F}_{k-1} \right] \xrightarrow{p} 0 \qquad \text{as } n \to \infty \tag{69}$$

where $\mathbb{I}_{A_{n,k,\epsilon}}$ is the indicator variable on $A_{n,k,\epsilon} := \{|x_{n,k}| > \epsilon\}$.

Now (68) follows from

$$\begin{aligned} \gamma_n &= \sum_{k=1}^{n} \mathbb{E}\left[\mathbf{w} X_{n,k}^\top X_{n,k} \mathbf{w}^\top | \mathcal{F}_{k-1} \right] = \mathbf{w} \sum_{k=1}^{n} \mathbb{E}\left[X_{n,k}^\top X_{n,k} | \mathcal{F}_{k-1} \right] \mathbf{w}^\top \\ &= \mathbf{w} \, \Phi(n) \, \mathbf{w}^\top \xrightarrow{p} \mathbf{w} \, \widetilde{\Sigma} \, \mathbf{w}^\top = \sigma^2, \end{aligned} \tag{70}$$

where the convergence follows from (63).

To see that (69) holds, by (37) we have

$$X_{n,k} = \sum_{j=1}^{d} X_{n,k} \mathbf{e}_j^\top \mathbf{e}_j = \sum_{j=1}^{d} n^{-1/2} \lambda_j \, b_{n,k}(j) \tau_k \mathbf{u}_j \mathbf{e}_j, \qquad 1 \le k \le n.$$

In particular, we have $X_{n,k}(1) = 0$ because $\tau_k \mathbf{u}_1 = 0$ holds for $k \ge 1$ in view of (40). Consequently, we have

$$x_{n,k} = X_{n,k} \mathbf{w}^\top = \sum_{j=2}^{d} n^{-1/2} w_j \lambda_j b_{n,k}(j) \tau_k \mathbf{u}_j. \tag{71}$$

Putting $\rho = \lambda_2/s$, then $\lambda_j/s \le \rho < 1/2$ holds for $2 \le j \le d$ in view of (A2) and (A4). Furthermore, there exists a constant $K_0 > 0$ independent of n and k such that

$$|x_{n,k}| \le \sum_{j=2}^{d} n^{-1/2} |w_j \lambda_j \tau_k \mathbf{u}_j| |b_{n,k}(j)| \le K_0 n^{-1/2} (n/k)^\rho \le K_0 n^{-1/2} \max(1, n^\rho) \tag{72}$$

holds for $1 \le k \le n$. Here the second inequality follows from Lemma 1 and the fact that $|w_j \lambda_j \tau_k \mathbf{u}_j|$ is bounded above by a constant independent of k. The last inequality follows from the fact that $(n/k)^\rho \le \max\big((n/1)^\rho, (n/n)^\rho\big)$. Now let $A'_{n,\epsilon} := \{K_0 n^{-1/2} \max(1, n^\rho) > \epsilon\}$, which it is either $\emptyset$ if n is sufficient large or the whole probability space otherwise. Then by (72) we have $A_{n,k,\epsilon} \subseteq A'_{n,\epsilon}$ and hence for all $\epsilon > 0$ and each n, we have $\mathbb{I}_{A_{n,k,\epsilon}} \le \mathbb{I}_{A'_{n,\epsilon}}$ for all $1 \le k \le n$. Furthermore, since $\rho < 1/2$ and $K_0 > 0$, we have

$$\mathbb{E}[\mathbb{I}_{A'_{n,\epsilon}}] = \mathbb{P}(A'_{n,\epsilon}) \to 0 \ \text{ as } n \to \infty. \tag{73}$$

Consequently, we have

$$\mathbb{E}[\gamma_n^*] = \mathbb{E}\Big[\sum_{k=1}^{n} \mathbb{E}\big[\,|x_{n,k}|^2\, \mathbb{I}_{A_{n,k,\epsilon}} | \mathcal{F}_{k-1}\big]\Big] \le \mathbb{E}\Big[\sum_{k=1}^{n} \mathbb{E}\big[\,|x_{n,k}|^2\, \mathbb{I}_{A'_{n,\epsilon}} | \mathcal{F}_{k-1}\big]\Big] \tag{74}$$

$$= \mathbb{E}\Big[\Big(\sum_{k=1}^{n} \mathbb{E}\big[\,|x_{n,k}|^2 \,|\mathcal{F}_{k-1}\big]\Big)\mathbb{I}_{A'_{n,\epsilon}}\Big] = \mathbb{E}\Big[\gamma_n \mathbb{I}_{A'_{n,\epsilon}}\Big] \tag{75}$$

$$= \mathbb{E}[\gamma_n]\mathbb{E}[\mathbb{I}_{A'_{n,\epsilon}}] \to 0, \quad \text{as } n \to \infty \tag{76}$$

where we have used the fact that $\mathbb{I}_{A'_{n,\epsilon}}$ is $\mathcal{F}_n$-measurable and independent of $\mathcal{F}_n$ (and all its sub-sigma-algebras); the convergence follows from (70) and (73). Since γ_n^* is almost surely non-negative, this completes the proof of (69), the last step in the proof of the theorem. □

8 Discussion

Inspired by a martingale approach developed by Bai and Hu (2005), we present in this paper the strong law of large numbers and the central limit theorem for a family of the Pólya urn models in which negative off-diagonal entries are allowed in their replacement matrices. This leads to a unified approach to proving corresponding limit theorems for the joint vector of cherry and pitchfork counts under the YHK model and the PDA model. In other words, the results for both models are derived from Theorems 1 and 2, using different replacement matrices. Furthermore, our results on unrooted trees are also derived directly from Theorems 1 and 2, without the need for a detour of rooted trees. For each of these random tree models, we show that the joint vector of cherry and pitchfork frequencies converges almost surely to a deterministic vector and the appropriately scaled fluctuations converge in distribution to a bivariate normal distribution. Interestingly, such convergence results do not depend on the initial tree used in the generating process.

The results presented here also lead to several broad directions that may be interesting to explore in future work. The first direction concerns a more detailed analysis on convergence. For instance, the central limit theorems present here should be extendable to a functional central limit theorem (see, e.g. Gouet (1993)), a follow-up project that we will pursue. Furthermore, it remains to establish the rate of convergence for the

limit theorems (see Laulin (2020) for some recent results on urns with two colours). For example, a law of the iterated logarithm would add considerable information to the strong law of large numbers by providing a more precise estimate of the size of the almost sure fluctuations of the random sequences in Theorems 3 and 4.

The second direction concerns whether the results obtained here can be extended to other tree statistics and tree models. For example, the two tree models considered here, the YHK and the PDA, can be regarded as special cases of some more general tree generating models, such as Ford's alpha model (see, e.g. Chen et al. (2009)) and the Aldous beta-splitting model (see, e.g. Aldous (1996)). Therefore, it is of interest to extend our studies on subtree indices to these two models as well. Furthermore, instead of cherry and pitchfork statistics, we can consider more general subtree indices such as k-pronged nodes and k-caterpillars (Rosenberg 2006; Chang and Fuchs 2010).

Finally, it would be interesting to study tree shape statistics for several recently proposed graphical structures in evolutionary biology. For instances, one can consider aspects of tree shapes that are related to the distribution of branch lengths (Ferretti et al. 2017; Arbisser et al. 2018) or relatively ranked tree shapes (Kim et al. 2020). Furthermore, less is known about shape statistics in phylogenetic networks, in which non-tree-like signals such as lateral gene transfer and viral recombinations are accommodated (Bouvel et al. 2020). Further understanding of their statistical properties could help us design more complex evolutionary models that may in some cases provide a better framework for understanding real datasets.

Acknowledgements We are grateful to three anonymous reviewers for their helpful suggestions that improve the presentation of the paper and insightful comments pointing out various links with existing results in the literature. K.P. Choi acknowledges the support of Singapore Ministry of Education Academic Research Fund R-155-000-188-114. The work of Gursharn Kaur was supported by NUS Research Grant R-155-000-198-114. We thank the Institute for Mathematical Sciences, National University of Singapore where this project started during the discussions in the *Symposium in Memory of Charles Stein*.

Appendix

In the appendix we present a proof of Lemma 1 and Corollary 3. To this end, we start with the following observation.

Lemma 2 *For $\lambda \in \mathbb{R}$, $\ell \in \mathbb{R}_{>0}$, and two non-negative integers m and n with $n \geq m$, put*

$$F_m^m(\ell, \lambda) = 1, \quad \text{and} \quad F_m^n(\ell, \lambda) := \prod_{i=m}^{n-1} \left(1 + \frac{\lambda}{\ell + i}\right) \quad \text{for } n > m.$$

Then we have

$$\lim_{m\to\infty} \sup_{n\geq m} \left(\frac{m}{n}\right)^{\lambda} F_m^n(\ell,\lambda) = 1. \tag{77}$$

Furthermore, there exists a positive constant $K = K(\lambda,\ell)$ such that

$$\left|F_m^n(\ell,\lambda)\right| \leq K\,(n/m)^{\lambda} \quad \textit{for all } 1 \leq m \leq n. \tag{78}$$

Proof Since the lemma holds for $\lambda = 0$ in view of $F_m^n(\ell,0) = 1$, we assume that $\lambda \neq 0$ in the remainder of the proof. For simplicity, put $L := \max\big(1, -(\ell+\lambda)\big)$.

First we shall establish (77). To this end, we may assume $m > L$, and hence $m + \ell + \lambda > 0$. Furthermore, recall the following result on the ratio of gamma functions: for a fixed number $y \in \mathbb{R}$, we have

$$\lim_{x\to\infty} \frac{\Gamma(x+y)}{x^y\Gamma(x)} = 1, \tag{79}$$

which follows from Stirling's formula for the gamma function; see also Jameson (2013, P.398) for an alternative approach. Therefore, putting

$$G_{m,k} := \frac{\Gamma(m+k+\ell+\lambda)}{(m+k)^{\lambda}\Gamma(m+k+\ell)} \qquad \text{for integer } k \geq 0,$$

then we have

$$\lim_{m\to\infty} \ln\big(G_{m,0}\big) = 0, \qquad \text{and hence} \qquad \lim_{m\to\infty} \sup_{k\geq 0} \ln\big(G_{m+k,0}\big) = 0. \tag{80}$$

Here the second limit holds because the limit of $\ln\big(G_{m,0}\big)$ being 0 implies that its limit superior is also 0. Together with $G_{m,k} = G_{m+k,0}$ for $k \geq 0$, this leads to

$$\lim_{m\to\infty} \sup_{k\geq 0} \ln\big(G_{m,k}\big) = \lim_{m\to\infty} \sup_{k\geq 0} \ln\big(G_{m+k,0}\big) = 0. \tag{81}$$

Since

$$\left(\frac{m}{m+k}\right)^{\lambda} F_m^{m+k}(\ell,\lambda) = \left(\frac{m}{m+k}\right)^{\lambda} \frac{\Gamma(m+k+\ell+\lambda)\Gamma(m+\ell)}{\Gamma(m+k+\ell)\Gamma(m+\ell+\lambda)} = \frac{G_{m,k}}{G_{m,0}} \tag{82}$$

holds for each integer $k \geq 0$, we have

$$\begin{aligned}
\lim_{m\to\infty} \sup_{n\geq m} \ln\left(\left(\frac{m}{n}\right)^{\lambda} F_m^n(\ell,\lambda)\right) &= \lim_{m\to\infty} \sup_{k\geq 0} \ln\left(\left(\frac{m}{m+k}\right)^{\lambda} F_m^{m+k}(\ell,\lambda)\right) \\
&= \lim_{m\to\infty} \sup_{k\geq 0} \Big(\ln(G_{m,k}) - \ln(G_{m,0})\Big) \\
&= \lim_{m\to\infty} \sup_{k\geq 0} \ln(G_{m,k}) - \lim_{m\to\infty} \ln(G_{m,0}) \\
&= 0,
\end{aligned}$$

where the last equality follows from (80) and (81). This completes the proof of (77).

Next, we shall establish (78). To this end we assume $m < n$, $m + \ell + \lambda \neq 0$, and $n - 1 + \ell + \lambda \neq 0$ as otherwise it clearly holds. Now consider the following three cases:

Case 1: $1 \leq m \leq n - 1 < L$, and hence $n - 1 + \ell + \lambda < 0$. Let $A = \{(\alpha, \beta) \,|\alpha, \beta \in \mathbb{N}; 1 \leq \alpha \leq \beta \leq 1 - \ell - \lambda\}$ be the finite subset of $\mathbb{N} \times \mathbb{N}$ whose size depends on ℓ and λ, and consider the constant

$$K_1 := \max_{(\alpha,\beta)\in A} \left\{ |F_\alpha^\beta(\ell, \lambda)| \, (\alpha/\beta)^\lambda \right\}.$$

Since $(m, n) \in A$, it follows that $|F_m^n(\ell, \lambda)| \leq K_1(n/m)^\lambda$ holds.

Case 2: $m \geq L$ and hence $m + \ell + \lambda > 0$. Note that in this case we have $F_m^n(\ell, \lambda) > 0$. Furthermore, an argument similar to the proof of (77) shows that for each $m \geq L$ we have

$$\lim_{n\to\infty} \left(\frac{m}{n}\right)^\lambda F_m^n(\ell, \lambda) = \frac{1}{G_{m,0}},$$

and hence there exists a constant K'_m depending on m, ℓ, λ so that $F_m^n(\ell, \lambda) \leq K'_m(n/m)^\lambda$ holds. Furthermore, by (77) it follows that there exists a constant $M = M(\ell, \lambda)$ and a constant $K_0 = K_0(\ell, \lambda)$ so that $F_m^n(\ell, \lambda) \leq K_0(n/m)^\lambda$ holds for all $m > M$. Therefore, for the constant

$$K_2 := \max\{K_0, K'_1, \ldots, K'_M\},$$

which depends only on ℓ and λ, we have $F_m^n(\ell, \lambda) \leq K_2(n/m)^\lambda$ for all $L \leq m \leq n$.

Case 3: $1 \leq m < L < n - 1$ and hence $m + \ell + \lambda < 0 < n - 1 + \ell + \lambda$. Note this implies $L > 1$ and we may further assume that L is not an integer as otherwise $F_m^n(\ell, \lambda) = 0$ follows. Let p be the (necessarily positive) largest integer less than L. Then $1 \leq m \leq p < L$ and we have $|F_m^p(\ell, \lambda)| \leq K_1(p/m)^\lambda$ for a constant K_1 in view of Case 1 and the fact that $F_m^m(\ell, \lambda) = 1$. Furthermore, as $p + 1 > L$, by Case 2 we have $|F_{p+1}^n(\ell, \lambda)| \leq K_2(n/p + 1)^\lambda$. Therefore, considering the constant $K_3 = \max\{K_1K_2, 2^{-\lambda}K_1K_2\}$, which depends on only ℓ and λ, we have

$$\begin{aligned} |F_m^n(\ell, \lambda)| &= |F_m^p(\ell, \lambda)F_{p+1}^n(\ell, \lambda)| \leq K_1K_2 \left(\frac{p}{m}\right)^\lambda \left(\frac{n}{p+1}\right)^\lambda \\ &= K_1K_2 \left(\frac{p}{p+1}\right)^\lambda \left(\frac{n}{m}\right)^\lambda \leq K_3 \left(\frac{n}{m}\right)^\lambda. \end{aligned}$$

The last inequality follows since $(\frac{p}{p+1})^\lambda \leq 1$ holds for $\lambda > 0$, and $(\frac{p}{p+1})^\lambda \leq 2^{-\lambda}$ for $\lambda < 0$. □

With Lemma 2, we now present a proof of Lemma 1.

Proof of Lemma 1. Recall that by (A3) we have $t_\ell = t_0 + \ell s$ for $\ell \geq 1$, and hence

$$b_{n,k}(j) = \prod_{\ell=k}^{n-1}\Big(1 + \frac{\lambda_j}{t_0 + \ell s}\Big) = \prod_{\ell=k}^{n-1}\Big(1 + \frac{\lambda_j/s}{(t_0/s) + \ell}\Big) = F_k^n\Big(\frac{t_0}{s}, \frac{\lambda_j}{s}\Big)$$

holds for $1 \leq j \leq d$ and $1 \leq k < n$. Noting that $b_{n,n}(j) = 1$, by Lemma 2 there exists a constant K_j' such that $|b_{n,k}(j)| \leq K_j'(n/k)^{\lambda_j/s}$ holds for $1 \leq k \leq n$. Now let $a_j = 1 + (\lambda_j/t_0)$ and put $K_j = \max(K_j', K_j'|a_j|)$. Then we have $|b_{n,0}(j)| \leq K_j n^{\lambda_j/s}$ in view of $b_{n,0}(j) = a_j b_{n,1}(j)$. This establishes (42) by choosing $K = \max(K_1, \ldots, K_d)$.

Next, we shall show (43). To this end, fix a pair of indices $2 \leq i \leq j \leq d$, and put $\rho_i = \lambda_i/s$ and $\rho_j = \lambda_j/s$. Then by (A2) we have $\rho := \rho_i + \rho_j < 1$ and $1 - \rho = (s - \lambda_i - \lambda_j)/s$. Furthermore, consider

$$\Delta_n := \frac{1}{n}\sum_{k=1}^{n}\Big(\frac{n}{k}\Big)^{\rho_j}\Big(b_{n,k}(i) - \Big(\frac{n}{k}\Big)^{\rho_i}\Big) \text{ and } \Delta_n^* := \frac{1}{n}\sum_{k=1}^{n} b_{n,k}(i)\Big(b_{n,k}(j) - \Big(\frac{n}{k}\Big)^{\rho_j}\Big).$$

Then we have

$$\Delta_n + \Delta_n^* = \frac{1}{n}\sum_{k=1}^{n}\Big(b_{n,k}(i)b_{n,k}(j) - \Big(\frac{n}{k}\Big)^{\rho}\Big).$$

By (41) it suffices to show that both $\Delta_n \to 0$ and $\Delta_n^* \to 0$ as $n \to \infty$.

By (42), we have $|b_{n,k}(i)| \leq K(n/k)^{\rho_i}$ and $|b_{n,k}(j)| \leq K(n/k)^{\rho_j}$. We shall first show that $\Delta_n \to 0$ as $n \to \infty$. To this end, let

$$H_{n,k} := \Big(\frac{n}{k}\Big)^{\rho_j}\Big(b_{n,k}(i) - \Big(\frac{n}{k}\Big)^{\rho_i}\Big) = \Big(\frac{n}{k}\Big)^{\rho}\Big(\Big(\frac{k}{n}\Big)^{\rho_i} b_{n,k}(i) - 1\Big).$$

Then we have $|H_{n,k}| \leq (K+1)(n/k)^{\rho}$ for all $1 \leq k \leq n$. Consider $\epsilon > 0$. Then it follows that $\epsilon' := \frac{\epsilon(1-\rho)}{2(2-\rho)} > 0$. By (77) in Lemma 2, we have

$$\lim_{k\to\infty}\sup_{n\geq k}\Big(\frac{k}{n}\Big)^{\rho_i} b_{n,k}(i) = \lim_{k\to\infty}\sup_{n\geq k}\Big(\frac{k}{n}\Big)^{\rho_i} F_k^n\Big(\frac{t_0}{s}, \rho_i\Big) = 1.$$

Therefore, there exists a positive integer M such that

$$\sup_{n\geq k}\Big(\Big(\frac{k}{n}\Big)^{\rho_i} b_{n,k}(i) - 1\Big) \leq \epsilon'$$

holds for all $k \geq M$. Consequently, $|H_{n,k}| \leq \epsilon'(n/k)^{\rho}$ holds for all $n \geq k \geq M$. Moreover, let N be the smallest integer greater than M so that $N > [2(K+1)M/\epsilon]^{1/(1-\rho)}$

and $N > M[2(K+1)/\epsilon]^{1/(1-\rho)}$ both hold. Then for $n > N$ we have

$$\frac{1}{n}\sum_{k=1}^{n}|H_{n,k}| \le \frac{\epsilon'}{n}\sum_{k=M+1}^{n}\left(\frac{n}{k}\right)^{\rho} + \frac{1}{n}\sum_{k=1}^{M}|H_{n,k}| \le \frac{\epsilon'}{n}\sum_{k=1}^{n}\left(\frac{n}{k}\right)^{\rho} + \frac{1}{n}\sum_{k=1}^{M}|H_{n,k}|$$
$$\le \frac{\epsilon'(2-\rho)}{1-\rho} + \frac{K+1}{n^{1-\rho}}\sum_{k=1}^{M}\left(\frac{1}{k}\right)^{\rho} \le \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,$$

where in the third inequality we use the fact that (41) implies

$$\frac{1}{n}\sum_{k=1}^{n}\left(\frac{n}{k}\right)^{\rho} \le \frac{1}{n} + \frac{1}{1-\rho} \le 1 + \frac{1}{1-\rho} = \frac{2-\rho}{1-\rho}.$$

Therefore it follows that $\Delta_n \to 0$ as $n \to \infty$. Since $|b_{n,k}(i)| \le K(n/k)^{\rho_i}$, a similar argument can be adopted to show that $\Delta_n^* \to 0$ as $n \to \infty$, completing the proof of Lemma 1. □

Finally, we complete the appendix by the following proof of Corollary 3.

Proof of Corollary 3 Fix a pair of indexes $2 \le i \le j \le d$. For simplicity, we put $a_{n,k} = b_{n,k}(i)b_{n,k}(j)$. Furthermore, let $\rho = (\lambda_i + \lambda_j)/s$, then $\rho < 1$ and $1 - \rho = (s - \lambda_i - \lambda_j)/s > 0$. Then by Lemma 1 we have

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}a_{n,k} = \frac{1}{1-\rho}, \quad \text{and} \quad |a_{n,k}| \le K\left(\frac{n}{k}\right)^{\rho} \quad \text{for all } n \ge 1 \text{ and } 1 \le k \le n. \tag{83}$$

Furthermore, let N_0 be the smallest integer greater than 1 such that both $N_0 > -(\lambda_i + t_0)/s$ and $N_0 > -(\lambda_j + t_0)/s$ hold. Then we have $a_{n,k} > 0$ for all $n \ge k \ge N_0$.

We shall next show that

$$\frac{1}{n}\sum_{k=1}^{n}a_{n,k}\mathbb{E}[|Z_k - Z|] \to 0. \tag{84}$$

For simplicity, put $\beta_k := \mathbb{E}[|Z_k - Z|]$ for $k \ge 1$. Then $\{\beta_k\}_{k\ge1}$ is a sequence of non-negative numbers which converges to 0. Thus there exists a constant $K_1 > 0$ such that $\beta_k < K_1$ holds for all $k \ge 1$. Next, fix an arbitrary number $\epsilon > 0$. By (83), let $N_1 = N_1(\epsilon)$ be the smallest integer greater than N_0 so that

$$\frac{1}{n}\sum_{k=1}^{n}a_{n,k} < \frac{1}{1-\rho} + \epsilon \text{ for holds for all } n > N_1. \tag{85}$$

Since $1 - \rho > 0$, the number $\epsilon' := \frac{\epsilon(1-\rho)}{2(1+\epsilon(1-\rho))}$ is greater than 0. Let N_2 be the smallest positive integer greater than N_1 so that $\beta_k < \epsilon'$ holds for all $k > N_2$. Now let N be the smallest positive integer greater than N_2 so that $N \ge (2(K_1+\epsilon')KN_2/\epsilon)^{1/(1-\rho)}$ and $N \ge N_2(2(K_1+\epsilon')K/\epsilon)^{1/(1-\rho)}$ both hold. Then for $n > N$ we have

$$\begin{aligned}\left|\frac{1}{n}\sum_{k=1}^{n}a_{n,k}\beta_k\right| &\le \left|\frac{1}{n}\sum_{k=1}^{N_2}a_{n,k}\beta_k\right| + \frac{1}{n}\sum_{k=1+N_2}^{n}a_{n,k}\beta_k \le \frac{K_1}{n}\sum_{k=1}^{N_2}|a_{n,k}| + \frac{\epsilon'}{n}\sum_{k=1+N_2}^{n}a_{n,k}\\ &= \frac{K_1}{n}\sum_{k=1}^{N_2}|a_{n,k}| - \frac{\epsilon'}{n}\sum_{k=1}^{N_2}a_{n,k} + \frac{\epsilon'}{n}\sum_{k=1}^{n}a_{n,k}\\ &\le \frac{K_1+\epsilon'}{n}\sum_{k=1}^{N_2}|a_{n,k}| + \frac{\epsilon'}{n}\sum_{k=1}^{n}a_{n,k}\\ &\le \frac{(K_1+\epsilon')KN_2\max(n^\rho,(n/N_2)^\rho)}{n} + \frac{\epsilon'}{n}\sum_{k=1}^{n}a_{n,k}\\ &\le \frac{\epsilon}{2} + \epsilon'\Big(\frac{1}{1-\rho}+\epsilon\Big) = \epsilon,\end{aligned}$$

from which (84) follows. Here the first inequality follows from the triangle inequality and that $a_{n,k}\beta_k > 0$ holds for $n \ge k > N_2 \ge N_0$, the second inequality holds since $0 \le \beta_k < K_1$ for $k \ge 1$ and $\beta_k < \epsilon'$ for $k > N_2$. Next, the third inequality holds since we have $\epsilon'(a_{n,k} + |a_{n,k}|) \ge 0$ for $1 \le k \le n$. Furthermore, the fourth inequality holds because by (83) we have $|a_{n,k}| \le K\max(n^\rho, (n/N_2)^\rho)$ for $1 \le k \le N_2$, and the last inequality follows from (85) and that $2(K_1+\epsilon')KN_2n^\rho \le \epsilon n$ and $2(K_1+\epsilon')KN_2^{1-\rho}n^\rho \le \epsilon n$ hold in view of $n > N$.

Finally, by (83) and (84) it follows that

$$\frac{1}{n}\sum_{k=1}^{n}a_{n,k}Z \xrightarrow{p} \frac{1}{1-\rho}Z \quad \text{and} \quad \frac{1}{n}\sum_{k=1}^{n}a_{n,k}(Z_k - Z) \xrightarrow{p} 0.$$

Therefore, we can conclude that

$$\frac{1}{n}\sum_{k=1}^{n}a_{n,k}Z_k = \frac{1}{n}\sum_{k=1}^{n}a_{n,k}Z + \frac{1}{n}\sum_{k=1}^{n}a_{n,k}(Z_k - Z) \xrightarrow{p} \frac{1}{1-\rho}Z,$$

as required. □

References

Aldous D (1996) Probability distributions on cladograms. In: Aldous D, Pemantle R (eds) Random discrete structures, The IMA volumes in mathematics and its applications, vol 76. Springer-Verlag, Berlin/Heidelberg, pp 1–18

Arbisser IM, Jewett EM, Rosenberg NA (2018) On the joint distribution of tree height and tree length under the coalescent. Theor Popul Biol 122:46–56

Athreya KB, Ney PE (1972) Branching Processes. Springer, Berlin

Bai ZD, Hu F (2005) Asymptotics in randomized Urn models. Ann Appl Probab 15(1B):914–940

Blum MGB, François O (2006) Which random processes describe the tree of life? A large-scale study of phylogenetic tree imbalance. Syst Biol 55(4):685–691

Bouvel M, Gambette P, Mansouri M (2020) Counting phylogenetic networks of level 1 and 2. J Math Biol 81(6):1357–1395

Chang H, Fuchs M (2010) Limit theorems for patterns in phylogenetic trees. J Math Biol 60(4):481–512
Chen B, Ford D, Winkel M (2009) A new family of markov branching trees: the alpha-gamma model. Electron J Probab 14:400–430
Choi KP, Thompson A, Wu T (2020) On cherry and pitchfork distributions of random rooted and unrooted phylogenetic trees. Theor Popul Biol 132:92–104
Colijn C, Gardy J (2014) Phylogenetic tree shapes resolve disease transmission patterns. Evol Med Public Health 1:96–108
Disanto F, Wiehe T (2013) Exact enumeration of cherries and pitchforks in ranked trees under the coalescent model. Math Biosci 242(2):195–200
Durrett R (2019) Probability: theory and examples. Cambridge University Press, Cambridge
Ferretti L, Ledda A, Wiehe T, Achaz G, Ramos-Onsins SE (2017) Decomposing the site frequency spectrum: the impact of tree topology on neutrality tests. Genetics 207(1):229–240
Gouet R (1993) Martingale functional central limit theorems for a generalized pólya urn. Ann Probab 21(3):1624–1639
Grimmett GR, Stirzaker DR (2001) Probability and random processes. Oxford University Press, Oxford
Hagen O, Hartmann K, Steel M, Stadler T (2015) Age-dependent speciation can explain the shape of empirical phylogenies. Syst Biol 64(3):432–440
Hall P, Heyde CC (2014) Martingale limit theory and its application. Academic Press, Cambridge
Harding EF (1971) The probabilities of rooted tree-shapes generated by random bifurcation. Adv Appl Probab 3(1):44–77
Heath TA, Zwickl DJ, Kim J, Hillis DM (2008) Taxon sampling affects inferences of macroevolutionary processes from phylogenetic trees. Syst Biol 57(1):160–166
Holmgren C, Janson S (2015) Limit laws for functions of fringe trees for binary search trees and recursive trees. Electron J Probab 20:1–51
Jameson G (2013) Inequalities for Gamma function ratios. Am Math Mon 120(10):936–940
Janson S (2004) Functional limit theorems for multitype branching processes and generalized Pólya urns. Stochastic Process Appl 110(2):177–245
Johnson NL, Kotz S (1977) Urn models and their application. John Wiley & Sons, New York-London-Sydney
Kim J, Rosenberg NA, Palacios JA (2020) Distance metrics for ranked evolutionary trees. Proc Natl Acad Sci 117(46):28876–28886
Laulin L (2020) A martingale approach for pólya urn processes. Electron Commun Probab 25(39):1–13
Mahmoud HM (2009) Pólya Urn Models. Texts in Statistical Science Series. CRC Press, Boca Raton, FL
McKenzie A, Steel MA (2000) Distributions of cherries for two models of trees. Math Biosci 164:81–92
Metzig C, Ratmann O, Bezemer D, Colijn C (2019) Phylogenies from dynamic networks. PLoS Comput Biol 15(2):e1006761
Mooers A, Harmon LJ, Blum MG, Wong DH, Heard SB (2007) Some models of phylogenetic tree shape. In: Gascuel O, Steel M (eds) Reconstructing evolution: new mathematical and computational advances. Oxford University Press, Oxford, pp 149–170
Plazzotta G, Colijn C (2016) Asymptotic frequency of shapes in supercritical branching trees. J Appl Probab 53(4):1143–1155
Pólya G (1930) Sur quelques points de la théorie des probabilités. Ann Inst H Poincaré 1(2):117–161
Rosenberg NA (2003) The shapes of neutral gene genealogies in two species: probabilities of monophyly, paraphyly and polyphyly in a coalescent model. Evolution 57(7):1465–1477
Rosenberg NA (2006) The mean and variance of the numbers of r-pronged nodes and r-caterpillars in Yule-generated genealogical trees. Ann Comb 10:129–146
Steel M (2016) Phylogeny: discrete and random processes in evolution. SIAM, Philadelphia
Wu T, Choi KP (2016) On joint subtree distributions under two evolutionary models. Theor Popul Biol 108:13–23
Yule GU (1925) A mathematical theory of evolution, based on the conclusions of Dr. J.C. Willis FRS. Philos Trans R Soc B 213:21–87