Local height in weighted Dyck models of random walks and the variability of the number of coalescent histories for caterpillar-shaped gene trees and species trees

Filippo Disanto¹ · Emanuele Munarini²

© Springer Nature Switzerland AG 2019

Abstract

We examine combinatorial parameters of three models of random lattice walks with up and down steps. In particular, we study the height y_i measured after i up-steps in a random weighted Dyck path of size (semilength) n . For a fixed integer $w \in \{0, 1, 2\}$, the considered weighting scheme assigns to each Dyck path of size n a weight $\prod_{i=1}^n y_i^w$ that depends on the height of the up-steps of the path. We investigate the expected value $E_n(y_i)$ of the height y_i in a random weighted Dyck path of size n , providing exact formulas for $E_n(y_i)$ and $E_n(y_i^2)$ when $w = 0, 1$, and estimates of the mean of y_i for $w = 2$. Denoting by $i^*(n)$ the position i where $E_n(y_i)$ reaches its maximum $m(n)$, our calculations indicate that, when n becomes large, the pair $(i^*(n), m(n))$ grows like $(n/2, 2\sqrt{n/\pi})$ if $w = 0$, $(3n/4, n/2)$ if $w = 1$, and $((9 + \sqrt{17})n/16, (1 + \sqrt{17})n/8)$ if $w = 2$. These results also contribute to the study of the variability of the number of “coalescent histories”: structures used in models of gene tree evolution to encode the combinatorially different configurations of a gene tree topology along the branches of a species tree. Relationships with other combinatorial and algebraic structures, such as alternating permutations and Meixner polynomials, are also discussed.

Keywords Lattice walks · Weighted Dyck paths · Combinatorial enumeration · Computational biology

Mathematics Subject Classification 05A15 · 05A16 · 92B10

1 Introduction

Lattice walks (or paths) often serve as models of statistical mechanical systems whose physical properties are linked with the combinatorial and enumerative features of the paths under consideration [17]. In this article, we study from a combinatorial point of view models of lattice walks derived from Dyck paths. A Dyck path is a trajectory of a random walk with steps $U = (+1, +1)$ and $D = (+1, -1)$ on the two-dimensional square lattice starting at the origin $(0, 0)$ and ending on the x -axis, where each point on the trajectory has a non-negative ordinate (Fig. 1a). Several physical systems can be described by means of Dyck paths. For instance, Dyck paths can be seen as wave functions of

spins [25] or, under suitable weighting schemes, they are used as polymer models [3]. In combinatorics, enumerative features of Dyck paths have been investigated with respect to several parameters (see e.g. [8]), starting from the number of Dyck paths of given length that is counted by the ubiquitous [9, 27] sequence of Catalan numbers.

Motivated also by computational problems in models of gene tree evolution [7, 10, 11, 23, 24], our paper contributes to the combinatorial analysis of Dyck models of lattice walks by studying the ordinate, or height, $y_i = y(U_i)$ of the i th up step U_i in Dyck paths of given size (semilength) taken under three different coloring or, equivalently, weighting schemes. In addition to the set $\mathcal{D}_n$ consisting of Dyck paths of size n , we consider the sets of paths $\mathcal{D}'_n$ and $\mathcal{D}''_n$. These

✉ Filippo Disanto, filippo.disanto@unipi.it; Emanuele Munarini, emanuele.munarini@polimi.it | ¹Department of Mathematics, University of Pisa, 56126 Pisa, Italy. ²Department of Mathematics, Politecnico di Milano, 20133 Milan, Italy.

Fig. 1 Standard and colored Dyck paths. **a** A path in D_3 . The height of the third up step is $y(U_3) = 2$. **b** Paths in D'_3 determined by labeling, in all possible ways, each up step U_i of the path in A by an integer in the range $[1, y(U_i)]$. The third up step can be labeled either by

1 or 2, since $y(U_3) = 2$. **c** Paths in D''_3 determined by labeling both up steps and down steps of the path in A. The second down step has label in the range $[1, y(D_2)] = [1, 2]$

are obtained by coloring, in all possible ways, the steps of each $\gamma \in D_n$ as follows. In D'_n , each up step $(U_i)_{1 \leq i \leq n}$ of γ is colored by an integer label in the range $[1, y(U_i)]$ (Fig. 1b). In D''_n , also down steps $(D_i)_{1 \leq i \leq n}$ of γ receive colors, with each D_i colored by an integer label in $[1, y(D_i)]$, where $y(D_i)$ is the height of D_i (Fig. 1c). Counting paths of D'_n and D''_n is thus equivalent to counting Dyck paths of size n in which the weight (multiplicity) of each path is given by the product $\prod_{i=1}^n y_i$ and $\prod_{i=1}^n y_i^2$, respectively. In particular, the cardinality of D'_n is $|D'_n| = (2n - 1)!!$, whereas $|D''_n|$ is the n th Euler/secant number.

The weighting schemes described above have been already considered in the literature (e.g. [5, 12, 15, 22]) due, for instance, to their interesting relationships with continued fractions. Continued fractions enable the study of the height—the ordinate of the highest point—of a uniform random path of D_n , D'_n , and D''_n . For example (see e.g. [13] Section V.4.3), the expected height of a uniform random path of D_n (resp. D'_n) is known to grow asymptotically like $\sqrt{\pi n}$ (resp. $n/2$). Here, we investigate a slightly different type of statistic: we study, for each fixed position $i \in [1, n]$, the height y_i of a random Dyck path of size n , when this is taken with a probability induced by a uniform distribution over the sets D_n , D'_n , and D''_n . In each of these three cases, we focus on the value $i^*(n)$ of i where the expectation of y_i has its maximum, on the value $m(n)$ of this maximum, and on the mean of y_n —the expected length of the last descent. The lack of a simple recursive decomposition for paths in D''_n makes the study of the height variables y_i quite difficult in this case. In addition to exact results, we also present empirical estimates based on intuitive, although not completely rigorous, theoretical arguments coupled with numerical validations.

After some preliminary results presented in Sect. 2, Sect. 3 is dedicated to the combinatorial and numerical analysis of the height variables y_i under the three different lattice models D_n , D'_n , and D''_n . Our findings are then shown in Sect. 4 to have interesting relationships with the variability of the number of evolutionary configurations, also called *coalescent histories* [7, 10, 11, 23, 24], of a

gene tree, representing the genetic history of gene copies sampled from individuals, in a species tree, which represents the ancestral relationships among the species or the populations of the considered individuals. In particular, we show that when gene trees and species trees have a particular caterpillar-shaped matching topology t of size n , the approximations of the value of i^* determined in Sect. 3 for the paths in D_n , D'_n , and D''_n can assist in locating the position of the speciation event σ over t —the splitting of a leaf node of t into two children nodes—for which the resulting tree t_σ has the largest increase in the number of coalescent histories with respect to t .

2 Standard and weighted Dyck paths

A Dyck path of size n is a lattice path with steps $U = (+1, +1)$ and $D = (+1, -1)$ that goes from $(0, 0)$ to $(2n, 0)$ never passing below the x -axis. We denote by D_n the set of Dyck paths of size (semilength) $n \geq 1$. In particular,

$$|D_n| = c_n = \frac{1}{n+1} \binom{2n}{n}, \quad (1)$$

where c_n is the n th Catalan number [27]. $D = \bigcup_{n \geq 1} D_n$ is the set of Dyck paths. If $\gamma \in D$, then U_i (resp. D_i) is the i th up (resp. down) step in γ . We denote by $y_i = y(U_i)$ the ordinate of the highest point of U_i in γ , while $y(D_i)$ is the ordinate of the highest point of the i th down step D_i of γ . If $\gamma = UDUDD$ (Fig. 1a), then $(y_1, y_2, y_3) = (1, 1, 2)$. Note that each Dyck path of size n is completely determined by the n -tuple $(y_1, y_2, \dots, y_n)$ of the ordinates of its up steps.

In addition to D , we consider two classes of labeled paths that we denote by D' and D'' . The labeled paths of D'_n (Fig. 1b) are obtained by coloring each Dyck path $\gamma \in D_n$ (Fig. 1a) through y_i possible labels for its up step U_i , for every $1 \leq i \leq n$. As in Fig. 1b, we take the labels for U_i in the range $[1, y_i]$. Under this scheme, $\gamma \in D_n$ can be colored in $\prod_{i=1}^n y_i$ different ways, the latter product being the multiplicity (weight) of γ in D'_n . Paths in D'_n are called *histoires d'Hermite*

in [22]. Similarly, by coloring both up steps and down steps of Dyck paths, we define the set $\mathcal{D}''$. More precisely, the labeled paths in $\mathcal{D}''_n$ (Fig. 1c) are obtained by coloring each $\gamma \in \mathcal{D}_n$ (Fig. 1a) through y_i possible labels for its up step U_i and through $y(D_i)$ labels for its down step D_i , for every $1 \leq i \leq n$. As in Fig. 1c, labels for each U_i are taken as integers in the range $[1, y_i]$, while labels for each D_i as integers in $[1, y(D_i)]$. With this setting, a Dyck path γ of size n can be colored in $\prod_{i=1}^n y_i \cdot y(D_i) = \prod_{i=1}^n y_i^2$ different ways, which is the multiplicity (weight) of γ in $\mathcal{D}''_n$.

While the cardinality of $\mathcal{D}_n$ is determined by the sequence of Catalan numbers (1), the cardinalities of $\mathcal{D}'_n$ and $\mathcal{D}''_n$ satisfy the following equations

$$|\mathcal{D}'_n| = \sum_{d \in \mathcal{D}_n} \prod_{i=1}^n y_i = (2n-1)!! = 1 \cdot 3 \cdot \dots \cdot (2n-1) = \frac{(2n)!}{2^n n!}, \quad (2)$$

$$|\mathcal{D}''_n| = \sum_{d \in \mathcal{D}_n} \prod_{i=1}^n y_i^2, \text{ and } \sum_{n \geq 0} \frac{|\mathcal{D}''_n|}{(2n)!} x^{2n} = \frac{1}{\cos(x)}. \quad (3)$$

The cardinality $|\mathcal{D}'_n|$ is given by $(2n-1)!!$ since, as sketched in Fig. 2 (see also [13] Example V.10), $\mathcal{D}'_n$ is equinumerous to the set of interconnection networks over $2n$ points (or involutions without fixed points over $2n$ elements). The set $\mathcal{D}''_n$ is enumerated by the n th Euler/secant number (seq. A000364 in [26]), which counts several other combinatorial structures. These structures include down-up alternating permutations—down-up permutations for short—of size $2n$, that is, permutations σ over $2n$ elements such that $\sigma(1) > \sigma(2) < \sigma(3) > \dots < \sigma(2n-1) > \sigma(2n)$ (Fig. 3a). In particular, there exists a bijection $FV(\sigma)$, the so-called Françon–Viennot bijection [15], mapping down-up permutations of $2n$ elements onto labeled paths of $\mathcal{D}''_n$. Starting from the set of lower entries (those in even position) of a down-up permutation σ of size $2n$, Fig. 3 shows how to determine the Dyck path underlying the labeled path $FV(\sigma) \in \mathcal{D}''_n$.

Fig. 2 Correspondence between interconnection networks over $2n$ points (or involutions without fixed points over $2n$ elements) and labeled Dyck paths of $\mathcal{D}'_n$. **a** The four networks over $2n = 6$ points in which the first, second and third point open an arc. **b** The Dyck path underlying the four paths of $\mathcal{D}'_3$ associated with the networks

given in A. Possible labels for the up steps are indicated in the figure. Circled dots determine opening arcs in A and up steps in B. The height y_i of the i th up step in B corresponds to the number of arcs met by a vertical line that intersects the interconnection networks of A passing through the point where the i th arc opens

Fig. 3 Down-up permutations and associated Dyck paths. **a** A down-up permutation σ with lower entries 1, 2, 4, and 5. **b** The Dyck path underlying the labeled path $FV(\sigma) \in \mathcal{D}''_4$ associated with the down-up permutation given in A through the Françon–Viennot

bijection. Lower (resp. upper) entries in the permutation determine up (resp. down) steps in the path. The height y_i of the i th up step in the path is given by $y_i = 2i - \ell_i$, where ℓ_i is the i th largest lower entry of the alternating permutation

3 Heights in D_n , D'_n , and D''_n

In this section, we study the random variables $y_i = y(U_i)$ assuming uniform distributions over the three sets D_n , D'_n and D''_n . Our goal is to measure how the different labeling schemes previously defined for the three classes D_n (no labels), D'_n (labels for up steps), and D''_n (labels for both up and down steps) affect the probability distribution of the variables y_i and, in particular, the expectations $E_n(y_i)$ for $1 \leq i \leq n$ (Fig. 4, left). To this extent, for each case D_n , D'_n , and D''_n we consider three additional statistics: (1) $m(n) = \max_{1 \leq i \leq n} [E_n(y_i)]$, (2) $i^*(n)$, the position i where $E_n(y_i)$ reaches its maximum value $m(n)$, and (3) $E_n(y_n)$, the expected length of the last descent in a random path of size n .

In the following, we write $f(n) \approx \tilde{f}(n)$ when theoretical arguments coupled with numerical calculations indicate that, for n sufficiently large, the ratio $f(n)/\tilde{f}(n)$ is "close" to 1. When $f(n)/\tilde{f}(n)$ converges to 1, we write $f(n) \sim \tilde{f}(n)$.

3.1 Heights in D_n

Although Dyck paths have been studied by several authors, up to our knowledge results for the mean of the height y_i do not explicitly appear in the literature for arbitrary values of $i \in [1, n]$. In this section, we fill this gap by determining closed formulas for the expectation $E_n(y_i)$ and for the second moment $E_n(y_i^2)$ of the random variable y_i , when paths of D_n are selected uniformly at random. We find that $m(n) \approx 2\sqrt{n/\pi}$ and $i^*(n) \approx n/2$. In Sect. 3.1.1, we provide a formula for the joint probability

$\text{Prob}(y_i = q \text{ \& } y_j = p)$, which enables the computation of the expectation $E_n(y_j | y_i = q)$ of the height y_j in a random path of size n conditioning on a given a value q for the height y_i . Our calculations also relate to the study of the Brownian excursion process $\{e(t) : t \in [0, 1]\}$, which is Brownian motion conditioned to be 0 at $t = 0$ and $t = 1$ and positive in the interior. Indeed, if $y(i)$ denotes the height at position $1 \leq i \leq 2n$ of a random Dyck path of semilength n , then $y_i = y(2i - y_i)$, and it is a known fact [18] that for a fixed $t \in [0, 1]$ the sequence of variables $y(2nt)/\sqrt{2n}$ converges for $n \rightarrow \infty$ to the random variable $e(t)$.

For $1 \leq i \leq n$ and $1 \leq q \leq i$, we define $d_{n,i,q} = |\{\gamma \in D_n : y_i = q\}|$. When $i = n$, $d_{n,i,q}$ reduces to the number $d_{n,q} = d_{n,n,q} = q/n \cdot \binom{2n-q-1}{n-1}$ of Dyck paths of size n having $y_n = q$, i.e., such that the last descent has length q (see e.g. [4]). Consider the number $s_{k,q} = d_{k+q+1,q+1} = (q+1)/(k+q+1) \cdot \binom{2k+q}{k}$ of Dyck suffixes starting with an up step whose lowest point has ordinate not greater than q , and having k up steps in total. According to the decomposition described in Fig. 5a, we can express $d_{n,i,q}$ as the product

$$d_{n,i,q} = d_{i,q} \cdot s_{n-i,q} \\ = \frac{q(q+1)}{i(n+q-i+1)} \binom{2i-q-1}{i-1} \binom{2n-2i+q}{n-i}. \quad (4)$$

Dividing by the n th Catalan number, we obtain the probability $\text{Prob}(y_i = q) = d_{n,i,q}/c_n$.

For a fixed n , methods of [20] can be used to show that the sum $e_i = E_n(y_i) = \sum_{q=1}^i q \cdot \text{Prob}(y_i = q)$ satisfies the recurrence

Fig. 4 Numerical values of $E_n(y_i)$, $i^*(n)/n$, and $m(n)/n$. (Left) Plot of the expected value $E_n(y_i)$ (dots), where $n = 100$ and $3 \leq i \leq 98$ (in steps of 5). Dashed lines give the approximations of $E_n(y_i)$ reported in the text: Eq. (8) for D_n , Eq. (25) for D'_n , and Eqs. (41), (42) for D''_n . Bottom plot for paths in D_n , middle plot for D'_n , and top plot for D''_n . (Right) Plot of $i^*(n)/n$ and $m(n)/n$ against $1/n$ for $20 \leq n \leq 200$ (in

steps of 20). Bottom plot for $m(n)/n$ in D'_n , second plot for $m(n)/n$ in D''_n , third plot for $i^*(n)/n$ in D'_n , and topmost plot for $i^*(n)/n$ in D''_n . When n increases, numerical values approach the constants reported in the text: $1/2$ in Eq. (27), $(1 + \sqrt{17})/8 \approx 0.64$ in Eq. (43), $3/4$ in Eq. (26), and $(1 + \sqrt{17})/8 \approx 0.82$ in Eq. (44)

Fig. 5 Decompositions of Dyck paths. **a** Decomposition for calculating $d_{n,i,q}$. **b** Decomposition for calculating $d_{n,i,q,j,p}$

$$\frac{(2i+1)(i-1-2n)}{i+1} \cdot \frac{\binom{2i}{i} \binom{2n-2i}{n-i}}{\binom{2n}{n}} \quad (5)$$

$$= e_{i+1}(3i-2n-1) - e_i(3i-2n+2),$$

with starting condition given by $e_1 = 1$. Dividing both sides of (5) by $(3i-2n-1)(3i-2n+2)$ and setting $\tilde{e}_i = e_i/(3i-2n-1)$, we obtain

$$\frac{(2i+1)(i-1-2n)}{(i+1)(3i-2n-1)(3i-2n+2)} \cdot \frac{\binom{2i}{i} \binom{2n-2i}{n-i}}{\binom{2n}{n}} = \tilde{e}_{i+1} - \tilde{e}_i, \quad (6)$$

yielding

$$E_n(y_i) = \frac{3i-2n-1}{n+2} - \frac{(4i^2-1)(i-n-1)}{i(n+2)} \cdot \frac{\binom{2i-2}{i-1} \binom{2n-2i+2}{n-i+1}}{\binom{2n}{n}}. \quad (7)$$

When $n \rightarrow \infty$ and $i = kn$ for a constant $k \in (0, 1)$, we use $\binom{2x}{x} \sim 4^x / \sqrt{\pi x}$ in (7) obtaining the approximation

$$\begin{aligned} & v_{i+1}(-6i^2(n-2) + 3i(2n^2 - 5n - 3) + 10n^2 + 11n + 3) \\ & - v_i(-6i^2(n-2) + 3i(2n^2 - 9n + 5) + 2(8n^2 - 5n + 3)) \\ & = \frac{(2i+1)(i-n-1)(12i^3 - 14i^2(3n+1) + i(30n^2 + n - 11) + 10n^2 + 11n + 3)}{(i+1)(2i-2n-1)} \\ & \cdot \frac{\binom{2i}{i} \binom{2n-2i+1}{n-i}}{\binom{2n}{n}}, \end{aligned} \quad (11)$$

$$E_n(y_i) \sim \frac{3i-2n-1}{n+2} - \frac{(4i^2-1)(i-n-1)}{i(n+2)} \cdot \sqrt{\frac{n}{\pi(i-1)(n-i+1)}}, \quad (8)$$

which is plotted in Fig. 4 (left, bottom plot) and can be used to estimate the value i^* of i where $E_n(y_i)$ has its maximum. Assuming n fixed, we first remove the denominator $n+2$ in (8) and take the derivative with respect to i . Then, setting

$i = kn$ ($0 < k < 1$), the resulting function behaves like $R(k, n) = 3 + [kn(1-k)]^{-1/2}(4k^3n^4 - 8k^4n^4)/(2\sqrt{\pi}k^3n^3)$, for n large. Solving $R(k, n) = 0$ gives $k = 1/2 + 3\sqrt{\pi/(16n+9\pi)}/2$ and thus

$$i^*(n) \approx \frac{n}{2} \left(1 + 3\sqrt{\frac{\pi}{16n+9\pi}} \right) \sim \frac{n}{2}. \quad (9)$$

Plugging $i = n/2$ in (8), we find the maximum value $m(n) = \max_{1 \leq i \leq n} [E_n(y_i)] = E_n(y_{i^*})$,

$$m(n) \approx 2\sqrt{\frac{n}{\pi}}. \quad (10)$$

Remark Note that, as reported e.g. in [13] (Proposition V.4), the expected height of a random Dyck path of size n measured at its highest point is asymptotically $\sqrt{\pi n}$. Up to a constant factor $\sqrt{\pi}/(2\sqrt{1/\pi}) = \pi/2$, this expectation corresponds to the estimate given in (10) for expected height of a random Dyck path of size n measured at position $i^*(n)$.

For the second moment $E_n(y_i^2)$ similar calculations can be performed. For a fixed n , the sum $v_i = E_n(y_i^2) = \sum_{q=1}^i q^2 \cdot \text{Prob}(y_i = q)$ satisfies

with $v_1 = 1$. Dividing both sides of the latter equation by $(-6i^2(n-2) + 3i(2n^2 - 5n - 3) + 10n^2 + 11n + 3) - (-6i^2(n-2) + 3i(2n^2 - 9n + 5) + 2(8n^2 - 5n + 3))$ and setting $\tilde{v}_i = v_i / [-6i^2(n-2) + 3i(2n^2 - 5n - 3) + 10n^2 + 11n + 3]$, Eq. (11) can be rewritten as

$$\begin{aligned} & \frac{(2i+1)(i-n-1)(12i^3 - 14i^2(3n+1) + i(30n^2 + n - 11) + 10n^2 + 11n + 3)}{(i+1)(2i-2n-1)(-6i^2(n-2) + 3i(2n^2 - 5n - 3) + 10n^2 + 11n + 3)(-6i^2(n-2) + 3i(2n^2 - 9n + 5) + 2(8n^2 - 5n + 3))} \\ & \times \frac{\binom{2i}{i} \binom{2n-2i+1}{n-i}}{\binom{2n}{n}} = \tilde{v}_{i+1} - \tilde{v}_i, \end{aligned} \quad (12)$$

giving

$$\begin{aligned} E_n(y_i^2) &= (-6i^2(n-2) + 3i(2n^2 - 5n - 3) + 10n^2 + 11n + 3) \\ & \cdot \left(\frac{15n(n-1)}{2(8n^2 - 5n + 3)(n+2)(n+3)} + \frac{1}{16n^2 - 10n + 6} \right. \\ & + \frac{(i+1)(i+2)(8i - 5n - 3)(i - n - 2)(i - n - 1)}{2i(n+2)(n+3)(2i-2n-3)(6i^2(n-2) + i(-6n^2 + 15n + 9) - 10n^2 - 11n - 3)} \\ & \left. \cdot \frac{\binom{2i+1}{i+2} \binom{2n-2i+3}{n-i+1}}{\binom{2n}{n}} \right). \end{aligned} \quad (13)$$

3.1.1 Joint and conditional probability of y_i and y_j

In our previous computations, we treated the variables y_i independently from each other. Here, we derive a formula for the joint probability $\text{Prob}(y_i = q \& y_j = p)$ that enables the calculation of the conditional expectation $E_n(y_j | y_i = q)$ (Fig. 6). The derivation of the joint probability can be performed as follows. Assuming without loss of generality the two positions i, j to be such that $i < j$, we consider the quantity $d_{n,i,q,j,p} = |\{\gamma \in D_n : y_i = q \& y_j = p\}|$. As sketched in Fig. 5b, the value of $d_{n,i,q,j,p}$ can be expressed as the product

$$d_{n,i,q,j,p} = d_{i,i,q} \cdot b_{j-i,q,p} \cdot s_{n-j,p}, \quad (14)$$

where $d_{i,i,q}$, $s_{n-j,p}$ are as in the previous section and $b_{j-i,q,p}$ is the number of Dyck factors with the following three properties: (1) the factor starts and ends with an up step, (2) it has $j-i$ steps U , and (3) its first up step starts at an

ordinate smaller than or equal to q and its last up step ends at the ordinate p .

Through the last descent construction of Dyck paths [1, 2], the quantity $b_{j-i,q,p}$ is easily seen to correspond to the number of labels (p) generated in $j-i \geq 0$ steps of the generating tree $\{(q); (\ell) \rightsquigarrow (1)(2) \dots (\ell+1)\}$. Setting $F_q(x, y) = \sum_{n \geq 0} \sum_{p=1}^{n+q} b_{n,q,p} x^n y^p$ as the generating function associated with the integers $b_{n,q,p}$, the structure of the generating tree translates into the following equation for F_q

$$\left(1 + \frac{xy^2}{1-y}\right) \cdot F_q(x, y) = y^q + \frac{xy}{1-y} F_q(x, 1).$$

This equation can be solved by the kernel method [1] obtaining first (see also Eq. (2.5.16) in [28])

$$F_q(x, 1) = \left(\frac{1 - \sqrt{1-4x}}{2x}\right)^{q+1} = \sum_{n \geq 0} \frac{q+1}{n+q+1} \binom{2n+q}{n} x^n,$$

Fig. 6 Conditional expectation $E_n(y_j | y_i = q)$ (dotted line) for $n = 100, j \in [1, n], i = n/2$, and $q = 5, 20$ (left and right plot, respectively). The solid line gives $E_n(y_j)$, the thicker dot corresponds to the point (i, q)

and then

$$F_q(x, y) = \frac{xy}{1 - y + xy^2} \cdot F_q(x, 1) + y^q \cdot \frac{1 - y}{1 - y + xy^2}. \quad (15)$$

Plugging the expansion $1/(1 - y + xy^2) = \sum_{n \geq 0} \sum_{p \geq 2n} (-1)^n \binom{p-n}{n} x^n y^p$ into (15), for $n \geq 1$ and $p \leq n + q$ we obtain the following expression

$$b_{n,q,p} = \sum_{k=1}^{\min(p,n)} (-1)^{k-1} \frac{q+1}{n-k+q+1} \binom{2n-2k+q}{n-k} \binom{p-k}{k-1}. \quad (16)$$

When $1 \leq q \leq i < j \leq n$ and $p \leq j - i + q$, from (14) and (16) we have

$$d_{n,i,q,j,p} = \frac{q(p+1)}{i(n+p-j+1)} \binom{2i-q-1}{i-1} \binom{2n-2j+p}{n-j} \times \left[\sum_{k=1}^{\min(p,j-i)} (-1)^{k-1} \frac{q+1}{j-i-k+q+1} \binom{2j-2i-2k+q}{j-i-k} \binom{p-k}{k-1} \right],$$

from which the joint probability $\text{Prob}(y_i = q \& y_j = p) = d_{n,i,q,j,p}/c_n$. The conditional probability can be computed as $\text{Prob}(y_j | y_i = q) = \text{Prob}(y_i = q \& y_j = p) / \text{Prob}(y_i = q)$, which reduces to

$$\begin{aligned} \text{Prob}(y_{i+1} = p | y_i = q) \\ = \frac{(n-i)(p+1)(2n+p-2i-2)!(n+q-i+1)!}{(q+1)(n+p-i)!(2n+q-2i)!}, \end{aligned}$$

when $j = i + 1$, $1 \leq i \leq n - 1$, and $1 \leq p \leq q + 1$.

3.2 Heights in $\mathcal{D}'_n$

In this section, we study the height variables y_i for paths of $\mathcal{D}'_n$. Note that, as described in Fig. 2, paths of $\mathcal{D}'_n$ are

in bijective correspondence with interconnection networks over $2n$ points. In particular, for a given path of $\mathcal{D}'_n$ the value of y_i determines the number of arcs that are active (open) in the associated interconnection network at the position where the i th arc opens. Here, we determine exact formulas for the expectation $E_n(y_i)$ and for the second moment $E_n(y_i^2)$ of the random variable y_i . In particular, we find that $m(n) \approx n/2$ and $i^*(n) \approx 3n/4$, while the expected length of the last descent is $E_n(y_n) = -1 + 4^n / \binom{2n}{n} \sim \sqrt{\pi n}$.

Let $d'_{n,q}$ denote the number of paths in $\mathcal{D}'_n$ such that $y_n = q$. The last descent construction of Dyck paths [2] gives the recurrence

$$\begin{aligned} d'_{n,q} &= q \sum_{j=q-1}^{n-1} d'_{n-1,j} \\ &= q d'_{n-1,q-1} + \frac{q}{(q+1)} \cdot (q+1) \sum_{j=q}^{n-1} d'_{n-1,j} \\ &= q d'_{n-1,q-1} + \left(\frac{q}{q+1} \right) d'_{n,q+1}, \end{aligned} \quad (17)$$

where $d'_{1,1} = 1$ and $d'_{n,0} = d'_{n,n+1} = 0$. Solving the recurrence we have

$$d'_{n,q} = \frac{q(2n-q-1)!}{(n-q)!2^{n-q}}. \quad (18)$$

Fig. 7 Labeled suffixes starting at ordinate q seen as labeled prefixes. **a** A suffix counted in $s'_{k,q} - s'_{k,q-1}$ with $k = q = 3$. The suffix has 4, 5, and 3 possible labels for U_1 , U_2 , and U_3 , respectively. Reading from right to left, the suffix corresponds to a Dyck prefix with 4, 5,

and 3 possible labels for its up steps U_5 , U_6 , and U_3 , and no labels for U_1 , U_2 , and U_4 . **b** A labeled suffix counted in $s''_{k,q} - s''_{k,q-1}$ with $k = q = 3$ [see Eq. (39) of Sect. 3.3]

By the same approach of Sect. 3.1, we couple the quantity $d'_{n,q}$ with the number $s'_{k,q}$ of labeled Dyck suffixes satisfying the following three properties: (1) the suffix starts with an up step whose lowest point has ordinate not greater than q , (2) the suffix has k steps U , and (3) each up step U can be labeled in $y(U)$ possible ways. Note that

$$s'_{k,q} - s'_{k,q-1} = \frac{\sum_{x=q+1}^{k+q} d'_{k+q,x}}{q!} = \frac{d'_{k+q+1,q+2}}{(q+2)q!}, \quad (19)$$

with $s'_{k,0} = |D'_k| = d'_{k+1,1}$. In particular, the first equality follows from the fact that the difference $s'_{k,q} - s'_{k,q-1}$ counts the labeled suffixes starting at ordinate q , for which the first peak has ordinate of its highest point within the range $[q+1, k+q]$. Reading these suffixes from right to left (Fig. 7a), one has $s'_{k,q} - s'_{k,q-1} = |\{d \in D'_{k+q} : y_{k+q} \in [q+1, k+q]\}|/q!$ as in (19). From (19), by induction on q and using (17), we obtain

$$s'_{k,q} = \frac{d'_{k+q+1,q+1}}{(q+1)!} = \frac{(2k+q)!}{2^k k! q!}.$$

The product $d'_{n,i,q} = d'_{i,q} \cdot s'_{n-i,q}$ counts the number of paths in D'_n such that $y_i = q$, and the probability $\text{Prob}(y_i = q) = d'_{n,i,q}/|D'_n|$ can be written as

$$\text{Prob}(y_i = q) = \frac{2^q q n! (2i - q - 1)! (2n - 2i + q)!}{(2n)! (n - i)! (i - q)! q!}. \quad (20)$$

Starting from the latter equation, it is possible to derive the following exact formulas for the expectation

$$E_n(y_i) = (2n - 2i + 1) \left(4^i \frac{\binom{2n-2i}{n-i}}{\binom{2n}{n}} - 1 \right) \quad (21)$$

and for the second moment

$$E_n(y_i^2) = 8n^2 - 12ni + 4i^2 + 10n - 6i + 3 - 4^i (2n - 2i + 1)(4n - 4i + 3) \frac{\binom{2n-2i}{n-i}}{\binom{2n}{n}} \quad (22)$$

of the random variable y_i . In particular, from (21) we find that the expected length of the last descent in a random path of D'_n is

$$E_n(y_n) = \frac{4^n}{\binom{2n}{n}} - 1 \sim \sqrt{\pi n}.$$

In order to prove the two formulas appearing in (21) and (22), we first rewrite the probability in (20) as

$$\begin{aligned} P_{n,i,q} &:= \text{Prob}(y_i = q) \\ &= \frac{\binom{n}{i}}{\binom{2n}{2i}} \binom{i}{q} \frac{(2i - q - 1)! (2n - 2i + q)!}{(2i)! (2n - 2i)!} 2^q q \\ &= \frac{\binom{n}{i}}{\binom{2n}{2i}} \frac{(-i)_q (2n - 2i + 1)_q}{(1 - 2i)_q q!} 2^q q, \end{aligned}$$

where $(x)_q = x(x+1) \cdots (x+q-1)$ is the *Pochhammer symbol*. Thus, we have

$$\begin{aligned} E_n(y_i) &= \sum_{q=1}^i q P_{n,i,q} \\ &= \frac{\binom{n}{i}}{\binom{2n}{2i}} \sum_{q=1}^i \frac{(-i)_q (2n - 2i + 1)_q}{(1 - 2i)_q} \frac{q^2 2^q}{q!} \\ &= \frac{\binom{n}{i}}{\binom{2n}{2i}} 2i [\Theta^2 F(z)]_{z=2} \end{aligned}$$

and

$$\begin{aligned} E_n(y_i^2) &= \sum_{q=1}^i q^2 P_{n,i,q} = \frac{\binom{n}{i}}{\binom{2n}{2i}} \sum_{q=1}^i \frac{(-i)_q (2n - 2i + 1)_q}{(1 - 2i)_q} \frac{q^3 2^q}{q!} \\ &= \frac{\binom{n}{i}}{\binom{2n}{2i}} 2i [\Theta^3 F(z)]_{z=2}, \end{aligned}$$

where $\Theta = z \frac{d}{dz}$ and

$$F(z) = {}_2F_1(2n - 2i + 1, -i; 1 - 2i; z),$$

with

$${}_2F_1(a, b; c; z) = \sum_{q \geq 0} \frac{(a)_q (b)_q}{(c)_q} \frac{z^n}{n!}$$

being the *Gaussian hypergeometric series* [21], for which one has the following contiguous relations

$$\Theta {}_2F_1(a, b; c; z) = \frac{ab}{c} z {}_2F_1(a + 1, b + 1; c + 1; z) \quad (23)$$

$$\Theta {}_2F_1(a, b; c; z) = a({}_2F_1(a + 1, b; c; z) - {}_2F_1(a, b; c; z)). \quad (24)$$

Note that by using formula (23) we have

$$\Theta F(z) = \frac{(2n - 2i + 1)i}{2i - 1} z {}_2F_1(2n - 2i + 2, 1 - i; 2 - 2i; z).$$

Therefore, by setting $G(q; z) = {}_2F_1(2n - 2i + q, 1 - i; 2 - 2i; z)$, from (24) we obtain

$$\begin{aligned}\Theta^2 F(z) &= \frac{(2n-2i+1)i}{2i-1} \Theta(z \cdot G(2;z)) \\ &= \frac{(2n-2i+1)i}{2i-1} (z G(2;z) + z \Theta G(2;z)) \\ &= \frac{(2n-2i+1)i}{2i-1} z (G(2;z) \\ &\quad + (2n-2i+2)(G(3;z) - G(2;z))) \\ &= \frac{(2n-2i+1)i}{2i-1} z (2(n-i+1) G(3;z) \\ &\quad - (2n-2i+1) G(2;z))\end{aligned}$$

$$\begin{aligned}G(q;2) &= {}_2F_1(2n-2i+q, 1-i; 2-2i; 2) \\ &= \frac{2^{2(i-1)}}{\binom{2i-2}{i-1}} \binom{i-1 + \frac{2n-2i+q-1}{2}}{i-1} \\ &= \frac{2^{2(i-1)}}{\binom{2i}{i} \frac{i}{2(2i-1)}} \binom{n + \frac{q-1}{2} - 1}{i-1} \\ &= \frac{2^{2i}}{\binom{2i}{i} \frac{2i}{2i-1}} \binom{n + \frac{q-1}{2}}{i} \frac{i}{n + \frac{q-1}{2}} \\ &= \frac{2^{2i}}{\binom{2i}{i}} \binom{n + \frac{q-1}{2}}{i} \frac{2i-1}{2n+q-1}.\end{aligned}$$

and

$$\begin{aligned}\Theta^3 F(z) &= \Theta^2 F(z) + \frac{(2n-2i+1)i}{2i-1} z (2(n-i+1) \Theta G(3;z) - (2n-2i+1) \Theta G(2;z)) \\ &= \Theta^2 F(z) + \frac{(2n-2i+1)i}{2i-1} z \left(2(n-i+1)(2n-2i+3)(G(4;z) - G(3;z)) + \right. \\ &\quad \left. - (2n-2i+1)(2n-2i+2)(G(3;z) - G(2;z)) \right) \\ &= \Theta^2 F(z) + \frac{2i(n-i+1)(2n-2i+1)}{2i-1} z \\ &\quad \times \left((2n-2i+3) G(4;z) - (4n-4i+4) G(3;z) + (2n-2i+1) G(2;z) \right).\end{aligned}$$

Hence, we have

$$\begin{aligned}[\Theta^2 F(z)]_{z=2} &= \frac{(2n-2i+1)2i}{2i-1} (2(n-i+1) G(3;2) \\ &\quad - (2n-2i+1) G(2;2))\end{aligned}$$

and

$$\begin{aligned}[\Theta^3 F(z)]_{z=2} &= [\Theta^2 F(z)]_{z=2} + \frac{4i(n-i+1)(2n-2i+1)}{2i-1} \\ &\quad \times ((2n-2i+3) G(4;2) - (4n-4i+4) G(3;2) \\ &\quad + (2n-2i+1) G(2;2)).\end{aligned}$$

Now, by using the identity [29]

$$\begin{aligned}{}_2F_1(-n, b; -2n; 2) &= \frac{2^{2n} n!}{(2n)!} \left(\frac{b+1}{2} \right)_n \\ &= \frac{2^{2n}}{\binom{2n}{n}} \left(n + \frac{b-1}{2} \right)_n \quad n \in \mathbb{N}\end{aligned}$$

with $n = i - 1$ and $b = 2n - 2i + q$, we can write $G(q; 2)$ as

In particular, we find

$$\begin{aligned}G(2;2) &= \frac{2^{2i}}{\binom{2i}{i}} \binom{n+1/2}{i} \frac{2i-1}{2n+1} \\ &= \frac{2^{2i}}{\binom{2i}{i}} \frac{\binom{2n+1}{2i} \binom{2i}{i}}{2^{2i} \binom{n}{i}} \frac{2i-1}{2n+1} \\ &= \frac{\binom{2n+1}{2i}}{\binom{n}{i}} \frac{2i-1}{2n+1} \\ &= \frac{\binom{2n}{2i}}{\binom{n}{i}} \frac{2i-1}{2n-2i+1}, \\ G(3;2) &= \frac{2^{2i-1}}{\binom{2i}{i}} \binom{n+1}{i} \frac{2i-1}{n+1} \\ &= \frac{2^{2i-1}}{\binom{2i}{i}} \binom{n}{i} \frac{2i-1}{n-i+1}, \\ G(4;2) &= \frac{2^{2i}}{\binom{2i}{i}} \binom{n+1+1/2}{i} \frac{2i-1}{2n+3} \\ &= \frac{\binom{2n+2}{2i}}{\binom{n+1}{i}} \frac{2i-1}{2n-2i+3} \\ &= \frac{\binom{2n}{2i}}{\binom{n}{i}} \frac{(2i-1)(2n+1)}{(2n-2i+1)(2n-2i+3)}.\end{aligned}$$

By these identities, we have

$$\begin{aligned} [\Theta^2 F(z)]_{z=2} &= \frac{2i(2n-2i+1)}{2i-1} \left(2(n-i+1) \frac{2^{2i-1}}{\binom{2i}{i}} \binom{n}{i} \frac{2i-1}{n-i+1} \right. \\ &\quad \left. - (2n-2i+1) \frac{\binom{2n}{2i}}{\binom{n}{i}} \frac{2i-1}{2n-2i+1} \right) \\ &= 2i(2n-2i+1) \left(\frac{2^{2i}}{\binom{2i}{i}} \binom{n}{i} - \frac{\binom{2n}{2i}}{\binom{n}{i}} \right), \end{aligned}$$

and consequently

$$\begin{aligned} E_n(y_i) &= \frac{\binom{n}{i}}{\binom{2n}{2i} 2i} [\Theta^2 F(z)]_{z=2} \\ &= \frac{\binom{n}{i}}{\binom{2n}{2i} 2i} 2i(2n-2i+1) \left(\frac{2^{2i}}{\binom{2i}{i}} \binom{n}{i} - \frac{\binom{2n}{2i}}{\binom{n}{i}} \right) \\ &= (2n-2i+1) \left(\frac{2^{2i}}{\binom{2n}{2i} \binom{2i}{i}} \binom{n}{i}^2 - 1 \right) \\ &= (2n-2i+1) \left(\frac{2^{2i}}{\binom{2n}{n}} \binom{2n-2i}{n-i} - 1 \right), \end{aligned}$$

which is formula (21).

Similarly, we have

$$\begin{aligned} [\Theta^3 F(z)]_{z=2} &= [\Theta^2 F(z)]_{z=2} \\ &\quad + \frac{4i(n-i+1)(2n-2i+1)}{2i-1} \\ &\quad \left((2n-2i+3) \frac{\binom{2n}{2i}}{\binom{n}{i}} \frac{(2i-1)(2n+1)}{(2n-2i+1)(2n-2i+3)} \right. \\ &\quad \left. - (4n-4i+4) \frac{2^{2i-1}}{\binom{2i}{i}} \binom{n}{i} \frac{2i-1}{n-i+1} \right. \\ &\quad \left. + (2n-2i+1) \frac{\binom{2n}{2i}}{\binom{n}{i}} \frac{2i-1}{2n-2i+1} \right) \\ &= [\Theta^2 F(z)]_{z=2} + 4i(n-i+1)(2n-2i+1) \\ &\quad \left(\frac{\binom{2n}{2i}}{\binom{n}{i}} \frac{2n+1}{2n-2i+1} \right. \\ &\quad \left. - 4 \frac{2^{2i-1}}{\binom{2i}{i}} \binom{n}{i} + \frac{\binom{2n}{2i}}{\binom{n}{i}} \right) \\ &= [\Theta^2 F(z)]_{z=2} + 4i(n-i+1)(2n-2i+1) \\ &\quad \left(\frac{\binom{2n}{2i}}{\binom{n}{i}} \frac{4n-2i+2}{2n-2i+1} - \frac{2^{2i+1}}{\binom{2i}{i}} \binom{n}{i} \right), \end{aligned}$$

and thus

$$\begin{aligned} E_n(y_i^2) &= \frac{\binom{n}{i}}{\binom{2n}{2i} 2i} [\Theta^3 F(z)]_{z=2} \\ &= \frac{\binom{n}{i}}{\binom{2n}{2i} 2i} [\Theta^2 F(z)]_{z=2} + \frac{\binom{n}{i}}{\binom{2n}{2i} 2i} 4i(n-i+1)(2n-2i+1) \\ &\quad \left(\frac{\binom{2n}{2i}}{\binom{n}{i}} \frac{4n-2i+2}{2n-2i+1} - \frac{2^{2i+1}}{\binom{2i}{i}} \binom{n}{i} \right) \\ &= E_n(y_i) + 2(n-i+1)(2n-2i+1) \left(\frac{4n-2i+2}{2n-2i+1} \right. \\ &\quad \left. - \frac{2^{2i+1}}{\binom{2n}{2i} \binom{2i}{i}} \binom{n}{i}^2 \right) \\ &= E_n(y_i) + 2(n-i+1)(2n-2i+1) \\ &\quad \left(\frac{4n-2i+2}{2n-2i+1} - 2^{2i+1} \frac{\binom{2n-2i}{n-i}}{\binom{2n}{n}} \right) \\ &= (2n-2i+1) \left(2^{2i} \frac{\binom{2n-2i}{n-i}}{\binom{2n}{n}} - 1 \right) + 2(n-i+1)(2n-2i+1) \\ &\quad \left(\frac{4n-2i+2}{2n-2i+1} - 2^{2i+1} \frac{\binom{2n-2i}{n-i}}{\binom{2n}{n}} \right) \\ &= (2n-2i+1) \left(2^{2i} \frac{\binom{2n-2i}{n-i}}{\binom{2n}{n}} - 1 + 2(n-i+1) \frac{4n-2i+2}{2n-2i+1} \right. \\ &\quad \left. - 2^{2i} 4(n-i+1) \frac{\binom{2n-2i}{n-i}}{\binom{2n}{n}} \right) \\ &= (2n-2i+1) \\ &\quad \left(2(n-i+1) \frac{4n-2i+2}{2n-2i+1} - 1 - 2^{2i} (4n-4i+3) \frac{\binom{2n-2i}{n-i}}{\binom{2n}{n}} \right) \\ &= 8n^2 - 12ni + 4i^2 + 10n - 6i + 3 \\ &\quad - 2^{2i} (2n-2i+1)(4n-4i+3) \frac{\binom{2n-2i}{n-i}}{\binom{2n}{n}}, \end{aligned}$$

which is formula (22).

When $n \rightarrow \infty$ and $i = kn$ for a constant $k \in (0, 1)$, by using $\binom{2x}{x} \sim 4^x / \sqrt{\pi x}$ in (21) we have the approximation

$$E_n(y_i) \sim (2n-2i+1) \left(\sqrt{\frac{n}{n-i}} - 1 \right), \quad (25)$$

which is plotted in Fig. 4 (left, middle plot) and can be used to estimate the value i^* of i where $E_n(y_i)$ has its maximum. Assuming n fixed, we take the derivative of (25) with respect to i . Then, setting $i = kn$ ($0 < k < 1$), the resulting function reads as $R(k, n) = 2 - (1-k)^{-1/2} + (1-k)^{-3/2}(1/2n)$. Solving $R(k, n) = 0$ when n is large gives $k = 3/4$, and thus

$$i^*(n) \approx \frac{3n}{4}. \quad (26)$$

Plugging $i = 3n/4$ in (25), we find the maximum value $m(n) = \max_{1 \leq i \leq n} [E_n(y_i)] = E_n(y_{i^*})$,

$$m(n) \approx \frac{n}{2}. \quad (27)$$

Remark Equation (27) indicates that asymptotically the ratio $m(n)/n$ converges to a value close—if not equal—to $1/2$. Interestingly, the expected height—the ordinate of the highest point—of a random uniform path in $\mathcal{D}'_n$ grows like $n/2$. Indeed, the height of a path in $\mathcal{D}'_n$ corresponds to the width of the interconnection network over $2n$ points associated with the path (Fig. 2), that is, to the maximum number of arcs met by a vertical line intersecting the diagram of the network. As reported in [13] (see also references therein), the expected width of a random interconnection network over $2n$ points grows asymptotically like $n/2$.

3.3 Heights in $\mathcal{D}''_n$

In this section, we study the height variables y_i in paths of $\mathcal{D}''_n$. For a fixed i , we determine an approximation of the mode $\mu_n(y_i)$ of y_i , which is shown numerically to be also a good estimate of the mean $E_n(y_i)$. In particular, we find that the expected values $E_n(y_i)$ can be approximated as numerical solutions to a polynomial equation determined explicitly. Based on these approximations, we estimate for n large the quantities $m(n) \approx (1 + \sqrt{17})n/8$ and $i^*(n) \approx (9 + \sqrt{17})n/16$. Calculations of Sect. 3.3.1 indicate that the expected length of the last descent in a random path of $\mathcal{D}''_n$ grows like $E_n(y_n) \approx 3n^{2/3}/2$. We also present a new manifestation of Meixner polynomials [14, 19]. Results of this section are of interest also because they relate to the set of lower entries (those in even position) of a down-up

permutation (Fig. 3). Indeed, given a down up permutation σ of size $2n$, the value of $y_i = y_i[FV(\sigma)]$ —the height of the i th up step in the labeled path $FV(\sigma) \in \mathcal{D}''_n$ associated with σ through the Françon–Viennot bijection—satisfies

$$\ell_i = 2i - y_i, \quad (28)$$

where ℓ_i is the i th largest lower entry in σ (Fig. 3a). For instance, Eq. (28) determines the largest lower entry ℓ_n of a down up permutation σ of size $2n$ through the length of the last descent y_n of the associated path $FV[\sigma]$.

Let $d''_{n,q}$ be the number of paths in $\mathcal{D}''_n$ such that $y_n = q$. Similarly to (17) we have the recurrence

$$d''_{n,q} = q^2 d''_{n-1,q-1} + \left(\frac{q}{q+1}\right)^2 d''_{n,q+1}, \quad (29)$$

where $d''_{1,1} = 1$ and $d''_{n,0} = d''_{n,n+1} = 0$. Although an explicit formula for $d''_{n,q}$ is not available, useful asymptotic estimates for this quantity can be achieved. From the generating function (3) we have $d''_{n+1,1} = |\mathcal{D}''_{n+1}| \sim 2(2n)!(2/\pi)^{2n+1}$, which coupled with (29) can be used to show by induction on q that

$$d''_{n,q} \sim 2q^2(2n-2)! \left(\frac{2}{\pi}\right)^{2n-1}, \quad (30)$$

when $n \rightarrow \infty$ and q is fixed. We will use Eq. (30) as an approximation of $d''_{n,q}$ when n becomes large and $q \ll n$ or $q = o(n)$.

When instead q is close to n , an asymptotic estimate of $d''_{n,q}$ can be obtained as follows. Plugging $q = n - i$ into (29) and dividing by $[(n-i)!]^2$ gives

$$\frac{d''_{n,n-i}}{[(n-i)!]^2} = \frac{d''_{n-1,(n-1)-i}}{[(n-1-i)!]^2} + (n-i)^2 \cdot \frac{d''_{n,n-(i-1)}}{[(n-i+1)!]^2},$$

Fig. 8 Ratios of the form $d''_{n-1,q-1}/d''_{n,q}$ and the probability $\text{Prob}(y_i = q)$. (Left) Plot of $d''_{n-1,q-1}/d''_{n,q}$ for $n = 100$ and $2 \leq q < 100$ (in steps of 2). The dashed lines are the approximations reported in

Eq. (38). (Right) Plot of the probability $\text{Prob}(y_i = q)$ for paths in $\mathcal{D}''_n$, when $n = 100$ and $i = 75$

where $d''_{n,n} = (n!)^2$ for $i = 0$. Thus, setting

$$\delta_{i,n} = \frac{d''_{n,n-i}}{[(n-i)!]^2}, \quad (31)$$

we have the recurrence

$$\begin{cases} \delta_{0,n} = 1, & \text{if } n > 0; \\ \delta_{i,j} = 0, & \text{if } i > 0; \\ \delta_{i,n} - \delta_{i,n-1} = (n-i)^2 \delta_{i-1,n}, & \text{if } n > i > 0. \end{cases} \quad (32)$$

Setting $\text{Pol}_0(n) = 1$, the polynomial $\text{Pol}_i(n)$ that computes $\delta_{i,n}$ for $n > i$ can be calculated recursively as

$$\text{Pol}_i(n) = \sum_{k=i+1}^n (k-i)^2 \text{Pol}_{i-1}(k), \quad (33)$$

yielding

$$\text{Pol}_0(n) = 1,$$

$$\text{Pol}_1(n) = n(n-1) \left(\frac{2n-1}{6} \right),$$

$$\text{Pol}_2(n) = (n+1)n(n-1)(n-2) \left(\frac{20n^2 - 32n - 9}{360} \right),$$

$$\text{Pol}_3(n) = (n+2)(n+1)n(n-1)(n-2)(n-3) \left(\frac{280n^3 - 924n^2 - 142n + 1275}{45360} \right),$$

and so on. From (33) it follows that $\text{Pol}_i(n)$ is a polynomial of degree $3i$ with leading term

$$\text{Pol}_i(n) \sim \frac{n^{3i}}{i! 3^i}. \quad (34)$$

Furthermore, when $i > 0$ the appearance of the factor $(n+i-1)(n+i-2) \cdots (n-i)$ in $\text{Pol}_i(n)$ can be explained by considering the quantity $\delta_{i,n}^*$ defined as

$$\begin{cases} \delta_{0,n}^* = 1, & \text{if } n \geq 0; \\ \delta_{i,i}^* = (-1)^i [(2i-1)!!]^2, & \text{if } i > 0; \\ \delta_{i,n}^* - \delta_{i,n-1}^* = (n-i)^2 \delta_{i-1,n}^*, & \text{if } i > 0 \text{ and } n > -i. \end{cases} \quad (35)$$

First, we observe that $\delta_{i,n}^*$ extends $\delta_{i,n}$. More precisely, from (35) one has by induction on i that $\delta_{i,i}^* = \delta_{i,i-1}^* = \cdots = \delta_{i,i-i+1}^* = 0$ for every $i > 0$, where $\delta_{i,i}^* = 0$ implies that $\delta_{i,n}^* = \delta_{i,n}$ whenever $\delta_{i,n}$ is defined (32). Second, starting with $\text{Pol}_0(n) = 1$, let us consider for each $i > 0$ the recursively defined polynomial $\text{Pol}_i^*(n) = (-1)^i [(2i-1)!!]^2 + \sum_{k=-i+1}^n (k-i)^2 \text{Pol}_{i-1}^*(k)$ that computes the value of $\delta_{i,n}^*$ for every $n > -i$. Since

$\delta_{i,i}^* = \delta_{i,i-1}^* = \cdots = \delta_{i,i-i+1}^* = 0$ for every $i > 0$, $\text{Pol}_i^*(n)$ cancels when $n \in \{i, i-1, \dots, -i+1\}$, and it must be equal to the polynomial $\text{Pol}_i(n)$ because both $\text{Pol}_i^*(n)$ and $\text{Pol}_i(n)$ output the same value, $\delta_{i,n}^* = \delta_{i,n}$, when evaluated at each $n \geq i$. Hence, also $\text{Pol}_i(n)$ cancels for $n \in \{i, i-1, \dots, -i+1\}$. This fact, together with (34), implies that $\text{Pol}_i(n)$ can be written as

$$\text{Pol}_i(n) = \frac{(n+i-1)!}{(n-i-1)!} \left(\frac{n^i + \mathcal{O}(n^{i-1})}{i! 3^i} \right),$$

yielding the asymptotic

$$\delta_{i,n} \sim \frac{n^i (n+i-1)!}{i! 3^i (n-i-1)!}. \quad (36)$$

Substituting $q = n - i$ in (31) and (36) we thus have that

$$d''_{n,q} \sim \frac{n^{n-q} (2n-q-1)! (q!)^2}{(n-q)! 3^{n-q} (q-1)!}, \quad (37)$$

when $n \rightarrow \infty$ and $n - q$ is a fixed constant. In what follows, we will approximate $d''_{n,q}$ as in Eq. (37) when n is large and $q \approx n$, e.g. $q/n \geq 0.75$.

As shown in Fig. 8 (left), for large n we can use Eqs. (30) and (37) for approximating the ratios of the form $d''_{n-1,q-1}/d''_{n,q}$ as

$$\frac{d''_{n-1,q-1}}{d''_{n,q}} \approx \begin{cases} \frac{\pi^2(q-1)^2}{8q^2(n-1)(2n-3)}, & \text{if } q \ll n; \\ \frac{q-1}{nq(2n-q-1)}, & \text{if } q \approx n, \end{cases} \quad (38)$$

where the second formula in (38) has been calculated by using $[(n-1)/n]^{n-q} \approx 1 - (n-q)/n$.

Let us now consider the number $s''_{k,q}$ of labeled Dyck suffixes defined by (1) the suffix starts with an up step whose lowest point has ordinate not greater than q , (2) the suffix has k steps U , and (3) each up step U can be labeled in $y(U)^2$ possible ways. Interpreting labeled suffixes as labeled prefixes (Fig. 7b), we have the recurrence

$$s''_{k,q} - s''_{k,q-1} = \frac{\sum_{x=q+1}^{k+q} d''_{k+q,x}}{(q!)^2} = \frac{d''_{k+q+1,q+2}}{[(q+2)(q!)]^2}, \quad (39)$$

with $s''_{k,0} = |D''_k| = d''_{k+1,1}$, which gives (by induction on q and using (29))

$$s''_{k,q} = \frac{d''_{k+q+1,q+1}}{[(q+1)!]^2}.$$

The product $d''_{n,i,q} = d''_{i,q} \cdot s''_{n-i,q}$ gives the number of paths in D''_n such that $y_i = q$, and the related probability can be computed as $\text{Prob}(y_i = q) = d''_{n,i,q}/|D''_n|$. As depicted in Fig. 8 (right), numerical calculations indicate that for a fixed i the

Fig. 9 Plot of $E_n(y_n)/n^{2/3}$ against $1/n$ for $20 \leq n \leq 200$ (in steps of 20)

distribution $\text{Prob}(y_i = q)$ is unimodal and such that values of q of non-negligible probability are concentrated around the mode $\mu_n(y_i)$, the latter being thus close to the mean, $\mu_n(y_i) \approx E_n(y_i)$.

We compute an approximation for the mode as $\mu_n(y_i) \approx \text{Sol}_q[d''_{n,i,q+1}/d''_{n,i,q} = 1]$, that is, by taking the value of q such that $\text{Prob}(y_i = q+1)/\text{Prob}(y_i = q) = d''_{n,i,q+1}/d''_{n,i,q} = 1$. In particular, for $i \approx (0.8)n$ —where numerical data (Fig. 4) indicate that i is close to $i^*(n)$ —the latter equality is empirically seen to hold with $q \approx (0.65)n$. In this case, both the pairs $(i, q) \approx ((0.8)n, (0.65)n)$ and $(n-i+q, q) \approx ((0.85)n, (0.65)n)$ have their second component not too far from the first, and we estimate the ratio

$$\begin{aligned} \frac{d''_{n,i,q+1}}{d''_{n,i,q}} &= \frac{d''_{i,q+1}}{d''_{i,q}} \cdot \frac{s''_{n-i,q+1}}{s''_{n-i,q}} = \left(1 - \frac{q^2 d''_{i-1,q-1}}{d''_{i,q}}\right) \left(\frac{q+1}{q}\right)^2 \cdot \frac{s''_{n-i,q+1}}{s''_{n-i,q}} \\ &= \left(1 - \frac{q^2 d''_{i-1,q-1}}{d''_{i,q}}\right) \left(\frac{q+1}{q}\right)^2 \cdot \frac{d''_{n-i+q+2,q+2}}{d''_{n-i+q+1,q+1} (q+2)^2} \end{aligned} \quad (40)$$

by using the second approximation of (38). We obtain

$$\begin{aligned} \frac{d''_{n,i,q+1}}{d''_{n,i,q}} &\approx \frac{(q+1)[i(2i-q-1)-q(q-1)](-i+n+q+2)(-2i+2n+q+1)}{iq^2(q+2)(2i-q-1)} \\ &:= Q(n, i, q). \end{aligned} \quad (41)$$

Solving $Q(n, i, q) = 1$ with respect to q gives the approximation

$$\mu_n(y_i) \approx \text{Sol}_q[Q(n, i, q) = 1]. \quad (42)$$

Table 1 Asymptotic values of $E_n(y_n)$, $m(n)$ and $i^*(n)$ found for paths in D_n (first row), D'_n (second row), and D''_n (third row)

	$E_n(y_n)$	$m(n)$	$i^*(n)$
D_n	~ 3	$\approx 2\sqrt{\frac{n}{\pi}}$	$\approx \frac{n}{2}$
D'_n	$\sim \sqrt{\pi n}$	$\approx \frac{n}{2}$	$\approx \frac{3n}{4}$
D''_n	$\approx \frac{3n^{2/3}}{2}$	$\approx \frac{(1+\sqrt{17})n}{8}$	$\approx \frac{(9+\sqrt{17})n}{16}$

As depicted in Fig. 4 (left, top plot), for each value of i , the numerical solution to the equation $Q(n, i, q) = 1$ is quite close to the mean $E_n(y_i)$. We use this observation for approximating the two quantities $m(n) = \max_{1 \leq i \leq n}[E_n(y_i)]$ and $i^*(n)$, where $m(n) = E_n(y_{i^*})$. We start by noting that, if we set $m(n)/n = k$, then the constant k can be estimated by requiring the equation $Q(n, i, kn) = 1$ to admit exactly one solution for the variable i . For increasing values of n , the ratio $Q(n, i, kn)$ approaches 1 when the terms of degree 5 appearing in the numerator and denominator of $Q(n, i, kn)$ are equal, that is, when

$$\begin{aligned} 4ki^4n - 8ki^3n^2 - 8k^2i^3n^2 + 4ki^2n^3 + 10k^2i^2n^3 \\ + 3k^3i^2n^3 - 2k^2in^4 + k^3in^4 + 2k^4in^4 \\ - 2k^3n^5 - 3k^4n^5 - k^5n^5 = 2k^3i^2n^3 - k^4in^4. \end{aligned}$$

Solving with respect to i the latter equation yields $i = (2n + 2kn \pm \sqrt{2\sqrt{4}})/4$, where $\Delta = 2n^2 + 2kn^2 + 5k^2n^2 - kn^2\sqrt{36 + 36k + 25k^2}$. A unique solution, $i = n(k+1)/2$, exists for $k = (1 + \sqrt{17})/8$ giving

$$m(n) \approx \frac{(1 + \sqrt{17})n}{8}, \quad \text{and} \quad (43)$$

$$i^*(n) \approx \frac{(9 + \sqrt{17})n}{16}, \quad (44)$$

where $(1 + \sqrt{17})/8 = 0.640388 \dots$ and $(9 + \sqrt{17})/16 = 0.820194 \dots$. Numerical calculations reported in Fig. 4 (right) show, for increasing n , the convergence of $m(n)/n$ and $i^*(n)/n$ towards values very close—if not equal—to the constants determined in (43) and (44).

Fig. 10 A coalescent history for a gene tree matching the species tree. **a** A gene tree G and a species tree S with a matching topology. **b** A possible realization R of the gene tree G depicted in **a** in the matching species tree S (thicker tree). **c** The coalescent history

associated with the realization R given in **B**: each arrow connects an internal node (coalescent event) of G with the branch of S where the coalescent event occurs

Remark As for the previous cases, D_n with Eqs. (9), (10) and D'_n with Eqs. (26), (27), we observe that the asymptotic relation

$$\frac{2i^*(n)}{n} = 1 + \frac{m(n)}{n}$$

is satisfied by Eqs. (43), (44). It would be interesting to have an intuitive interpretation of the latter equation.

3.3.1 Last descent y_n

When we use Eq. (42) to estimate $E_n(y_i)$ with i very close to n , a larger relative error $|E_n(y_i) - \text{Sol}_q[Q(n, i, q) = 1]|/E_n(y_i)$ can be observed numerically. This is due to the fact that for such values of i the mode $\mu_n(y_i)$ and the mean $E_n(y_i)$ are quite far from i —in fact, numerically one sees that $E_n(y_n)/n \rightarrow 0$ for increasing n . In particular, in the derivation of (42) we replaced the term $d''_{i-1,q-1}/d''_{i,q}$ appearing in (40) with the second approximation of (38), assuming implicitly q sufficiently close to i . In order to determine a better estimate of the mode of y_n , we set $i = n$ in the right-hand side

of (40)—thus $s''_{n-i,q+1}/s''_{n-i,q} = s''_{0,q+1}/s''_{0,q} = 1$ —and then we use the first formula of (38) to replace the term $d''_{n-1,q-1}/d''_{n,q}$ present in the resulting expression. This yields

$$\frac{d''_{n,q+1}}{d''_{n,q}} \approx \left(1 - \frac{\pi^2(q-1)^2 q^2}{8q^2(n-1)(2n-3)}\right) \left(\frac{q+1}{q}\right)^2.$$

Taking $\mu_n(y_n) \approx \text{Sol}_q[d''_{n,q+1}/d''_{n,q} = 1]$, we obtain

$$\begin{aligned} \mu_n(y_n) &\approx \text{Sol}_q \left[1 - \frac{\pi^2(q-1)^2 q^2}{8q^2(n-1)(2n-3)} = \left(\frac{q}{q+1}\right)^2 \right] \\ &\approx \text{Sol}_q \left[1 - \frac{\pi^2(q-1)^2 q^2}{8q^2(n-1)(2n-3)} = 1 - \frac{2}{q} \right] \\ &= \text{Sol}_q \left[q(q-1)^2 = \frac{16(n-1)(2n-3)}{\pi^2} \right] \sim 2 \left(\frac{2}{\pi}\right)^{2/3} \cdot n^{2/3}, \end{aligned} \quad (45)$$

where $2(2/\pi)^{2/3} = 1.48007 \dots$. Based on (45), we expect the ratio $E_n(y_n)/n^{2/3}$ to be around 1.5 for large n . This is indeed shown in Fig. 9, where numerical values of $E_n(y_n)/n^{2/3}$ are plotted for increasing n .

Fig. 11 Coalescent histories for three types of caterpillar-shaped trees. **a** A caterpillar-shaped binary tree of size $n = 4$ with seed tree s . The tree consists of an n -chain of internal nodes $k_1, \dots, k_n$ to which n copies of the seed tree s have been appended. If s is the one-taxon (resp. two-taxon, three-taxon) tree, then the corresponding

caterpillar-shaped tree is the caterpillar in **b** (resp. lodgepole in **c**, pitchfork in **d**). **b–d** A coalescent history h for a caterpillar, lodgepole, and pitchfork tree of size $n = 4$. The depth of node k_i under h is given by $\text{depth}_{k_i}(h) = y_i$

Last descent and Meixner polynomials. The integers $d''_{n,q}$ show an interesting connection with Meixner polynomials [14, 19] when written in terms of Euler numbers. Indeed, if we define $e_n = d''_{n,1}$ —the n th Euler number is thus given by $|D''_n| = d''_{n+1,1} = e_{n+1}$ —and we write $d''_{n,q}$ as a function of the shifted Euler numbers e_n , then for $1 \leq q \leq 5$ and $n \geq q$ we have the following equalities

$$\begin{aligned}d''_{n,1} &= e_n \\d''_{n,2} &= 4e_n \\d''_{n,3} &= 9(e_n - e_{n-1}) \\d''_{n,4} &= 16(e_n - 5e_{n-1}) \\d''_{n,5} &= 25(e_n - 14e_{n-1} + 9e_{n-2}).\end{aligned}$$

In particular, we see that for a fixed q the ratio $d''_{n,q}/q^2$ can be expressed as a linear combination of the integers $e_n, e_{n-1}, \dots, e_{n-\lfloor (q-1)/2 \rfloor}$ with coefficients matching those of the Meixner polynomial $M_{q-1}(z)$, where

$$\begin{aligned}M(z, x) &= \sum_{n \geq 0} M_n(z) \frac{x^n}{n!} \\&= \frac{e^{z \arctan(x)}}{\sqrt{1+x^2}} = 1 + zx \\&\quad + (z^2 - 1) \frac{x^2}{2} + (z^3 - 5z) \frac{x^3}{6} \\&\quad + (z^4 - 14z^2 + 9) \frac{x^4}{24} + \dots\end{aligned}$$

is the exponential generating function for the considered class of polynomials. This relationship, which holds for every q , follows from (29). Indeed, a recurrence for the ratios $r_{n,q} = d''_{n,q}/q^2$ is given by $r_{n,1} = r_{n,2} = e_n$, $r_{n,q+1} = r_{n,q} - (q-1)^2 r_{n-1,q-1}$, which corresponds to the recurrence for the polynomials $M_n(z)$ given by $M_0(z) = 1$, $M_1(z) = z$, $M_{n+1}(z) = zM_n(z) - n^2 M_{n-1}(z)$ [14, 16, 19].

3.4 Estimates found for the value of $E_n(y_n)$, $m(n)$ and $I^*(n)$

Table 1 highlights some of the approximations obtained in the previous sections. Results of the first row— $E_n(y_n) \sim 3$ is a well-known fact—have been derived from formula (7) that computes the mean $E_n(y_i)$ for standard Dyck paths of size n . Results of the second row follow from formula (21) that calculates the mean $E_n(y_i)$ for paths in $\mathcal{D}'_n$. To find a formal proof of the estimates in the third row of the table is an open problem, here they have been derived through

a heuristic approach and validated by numerical calculations such as those reported in Figs. 4 and 9.

4 Relationships with the variability of the number of coalescent histories

Some of the results presented in the previous sections relate to the study of the variability of the number of *coalescent histories* [7, 10, 11, 23, 24], structures used in phylogenetic models of evolution for representing the combinatorially different configurations that a gene tree topology—representing how genes sampled from individuals have evolved from a common ancestor—can assume along the branches of a species tree—which describes the evolutionary relationships among the species or the populations of the considered individuals (Fig. 10). These structures are used in fundamental calculations of gene tree probabilities (see e.g. [6, 7]) for the inference of species trees based on collections of gene trees derived from genetic data. As the cost of these calculations is strongly affected by the number of coalescent histories possible for a gene tree G and a species tree S , it is important to characterize those pairs (G, S) for which the number of coalescent histories is larger or smaller.

For a fixed species tree S , pairs (G, S) where the gene tree G is similar or equal to S tend to have the largest number of coalescent histories [7]. This observation motivates the study of the number of coalescent histories when gene trees match the topology of the species trees. When G and S share the same full binary topology $t = G = S$ (Fig. 10), a coalescent history of t is a tree map h that associates each internal node k of t with a branch $h(k)$ of t such that (1) k descends from $h(k)$ and (2) $h(k)$ is not strictly below $h(k')$, if k is above node k' (Fig. 10c). As depicted in Fig. 10, a coalescent history h of t encodes a realization of the gene tree G in the matching species tree S , where $h(k)$ specifies the branch of S where the coalescent event k of G takes place. The *depth* of the coalescent event k under the history h , in symbols $\text{depth}_k(h)$, is the distance between k and $h(k)$ measured as the number of edges of t visited when we move from k to $h(k)$. For instance, the coalescent event g in Fig. 10c has depth 2.

4.1 Coalescent histories for three families of caterpillar-shaped trees

Under the matching assumption $G = S = t$, coalescent histories of a *caterpillar*-shaped tree t (Fig. 11a) can be interpreted as Dyck paths with colored up steps. In particular, here we focus on coalescent histories for three families of caterpillar-shaped trees that have been already

Fig. 12 Variation in the number of coalescent histories for caterpillar (left) and lodgepole (right) trees of size $n = 10$ (top) and $n = 100$ (bottom). A new cherry replaces a given terminal node (leaf) at position $i \in [1, n]$ and the variation is measured as the ratio between the number of coalescent histories in the resulting tree and the number of coalescent histories in the original caterpillar or lodgepole tree. The number of coalescent histories in each considered tree can be computed recursively, e.g. as in [23]

considered in the literature [7, 10], namely the *caterpillar*, the *lodgepole*, and the *pitchfork* family (Fig. 11: b, c, and d respectively).

If t is a caterpillar tree of size n (Fig. 11b), then there exists a bijective correspondence between the coalescent histories of t and the Dyck paths of size n : each coalescent history h of t is encoded by the Dyck path $\gamma_h \in \mathcal{D}_n$ such that the value of y_i —the height of the i th up step U_i —in γ_h equals the depth of the coalescent event k_i in h , in symbols $y_i(\gamma_h) = \text{depth}_{k_i}(h)$. For instance, the Dyck path $\gamma_h \in \mathcal{D}_4$ associated with the coalescent history depicted in Fig. 11b is $\gamma_h = UDUUDUDD$, where $(y_1, y_2, y_3, y_4) = (1, 1, 2, 2)$ as

determined by the depth of the coalescent events k_1, k_2, k_3 , and k_4 under the considered coalescent history.

When t is a lodgepole tree of size n (Fig. 11c), the set of coalescent histories of t is in bijection with the set $\mathcal{D}_n^{\ell}$ of labeled Dyck paths of size n such that each up step U_i of a path $\gamma \in \mathcal{D}_n^{\ell}$ is colored by an integer label in the range $[1, y_i + 1]$, where y_i is the height of U_i in γ . More precisely, a coalescent history h of t (Fig. 11c) is encoded by the path $\gamma_h \in \mathcal{D}_n^{\ell}$ where (1) as in the caterpillar case $y_i(\gamma_h) = \text{depth}_{k_i}(h)$, and (2) the up step U_i in γ_h is colored with an integer label whose value corresponds to the depth under h of the cherry node descending from node k_i in t . For example, the path $\gamma_h \in \mathcal{D}_4^{\ell}$ associated with the

Fig. 13 Variation in the number of coalescent histories for a pitchfork tree of size $n = 100$. A new cherry replaces a given terminal node at position $i \in [1, 100]$ and the variation is measured as the ratio between the number of coalescent histories in the resulting tree and the number of coalescent histories in the original tree

history h depicted in Fig. 11c is $\gamma_h = UDUUDUDD$ with labels for its up steps (U_1, U_2, U_3, U_4) given by (2, 1, 2, 3). Note that for every history h of t the depth of the cherry node descending from k_i exceeds at most by one the depth of node k_i . Therefore, for every h the label for the up step U_i of γ_h is in the range $[1, y_i + 1]$ as required for paths in $\mathcal{D}_n^\ell$. Furthermore, observe that $\mathcal{D}_n^\ell$ resembles $\mathcal{D}'_{n+1}$ in the sense that both sets are derived from Dyck paths by assigning to up steps U a weight, $y(U) + 1$ and $y(U)$, that is linear in the height $y(U)$. In particular, paths in $\mathcal{D}_n^\ell$ correspond bijectively to indecomposable paths of $\mathcal{D}'_{n+1}$ [10], and the cardinalities $|\mathcal{D}_n^\ell|, |\mathcal{D}'_n|$ are asymptotically equivalent, $|\mathcal{D}_n^\ell| \sim |\mathcal{D}'_n|$.

Finally, if t is a pitchfork tree of size n (Fig. 11d), then the set of coalescent histories of t corresponds bijectively to the set $\mathcal{D}_n^p$ of labeled Dyck paths of size n in which each up step U_i of $\gamma \in \mathcal{D}_n^p$ is colored by an integer label in the range $[1, 2 + 5y_i/2 + y_i^2/2]$, where y_i is measured in γ . Similarly to the lodgpole case, we associate a history h of t (Fig. 11d) with the path $\gamma_h \in \mathcal{D}_n^p$ such that (1) $y_i(\gamma_h) = \text{depth}_{k_i}(h)$, and (2) the up step U_i in γ_h is colored with an integer label whose value determines (as detailed below) the depth under h of the two internal nodes $k_{i,1}, k_{i,2}$ of the subtree s appended to k_i in t . More in details, assuming without loss of generality $k_{i,2}$ below $k_{i,1}$ in t , in (2) we define the label for the i th up step U_i of γ_h as the position (first, second, and so on) under the lexicographic order of the pair of integers $(\text{depth}_{k_{i,1}}(h), \text{depth}_{k_{i,2}}(h))$ in the set $A(y_i) = \{(a_{i,1}, a_{i,2}) : 1 \leq a_{i,1} \leq y_i + 1 \text{ and } 1 \leq a_{i,2} \leq a_{i,1} + 1\}$ of cardinality $|A(y_i)| = 2 + 5y_i/2 + y_i^2/2$. For instance, the path $\gamma_h \in \mathcal{D}_4^p$ that corresponds to the history h depicted in Fig. 11D is $\gamma_h = UDUUDUDD$ with labels for its up steps (U_1, U_2, U_3, U_4) given by (5, 1, 3, 8), where e.g. label 3 for step U_3 comes from the fact that the pair $(\text{depth}_{k_{3,1}}(h), \text{depth}_{k_{3,2}}(h)) = (2, 1)$ is in position 3 in the set $A(y_3) = A(2) = \{(1, 1), (1, 2), (2, 1), (2, 2), (2, 3), (3, 1), (3, 2), (3, 3), (3, 4)\}$. Note that both $\mathcal{D}_n^p$ and $\mathcal{D}''_n$ derive from Dyck paths by assigning to up steps U a weight that is quadratic in the height $y(U)$.

4.1.1 Variability of the number of coalescent histories and values of i^*

By associating coalescent histories with paths as described in the previous section, it is possible to highlight some interesting relationships between the estimates provided in Table 1 for the quantity i^* and the variability of the number of histories in caterpillar, lodgpole, and pitchfork trees, when these are affected by a small change in tree topology. Note indeed that the three bijections $h \rightarrow \gamma_h$ previously defined for the considered tree families have in common the fact that $\text{depth}_{k_i}(h) = y_i(\gamma_h)$. As a consequence, the expected

depth $E_n(\text{depth}_{k_i})$ of the node k_i under a random coalescent history of a caterpillar, lodgpole or pitchfork tree t of size n is equal to the expectation $E_n(y_i)$ of the height y_i in a random path of $\mathcal{D}_n, \mathcal{D}_n^\ell, \mathcal{D}_n^p$, respectively. In symbols we write

$$E_n(\text{depth}_{k_i} | \mathcal{C}, \mathcal{L}, \mathcal{P}) = E_n(y_i | \mathcal{D}, \mathcal{D}^\ell, \mathcal{D}^p),$$

where $\mathcal{C}, \mathcal{L}$, and $\mathcal{P}$ denote the sets of caterpillar, lodgpole, and pitchfork trees, respectively.

As already observed, $\mathcal{D}^\ell$ and $\mathcal{D}'$ (resp. $\mathcal{D}^p$ and $\mathcal{D}''$) both derive from Dyck paths with a weight for each up step U that is linear (resp. quadratic) in the height $y(U)$. This similarity induces a correlation

$$E_n(\text{depth}_{k_i} | \mathcal{C}, \mathcal{L}, \mathcal{P}) \propto E_n(y_i | \mathcal{D}, \mathcal{D}', \mathcal{D}''), \quad (46)$$

which can be observed experimentally between the expectations $E_n(\text{depth}_{k_i} | \mathcal{C}, \mathcal{P}) = E_n(y_i | \mathcal{D}^\ell, \mathcal{D}^p)$ and the expectations $E_n(y_i | \mathcal{D}', \mathcal{D}'')$ studied in the previous sections. Based on (46), for a given a caterpillar (resp. lodgpole, pitchfork) tree t of size n , the nodes k_i of the backbone branch of t are on average deeper under the coalescent histories of t when i is close to the position $i^*(n)$ that maximizes the expectation $E_n(y_i)$ for paths of $\mathcal{D}_n$ (resp. $\mathcal{D}'_n, \mathcal{D}''_n$). In symbols we write

$$i \rightarrow i^* \Rightarrow \text{depth}_{k_i} \uparrow, \quad (47)$$

and an interesting consequence of this relationship can be observed by measuring as in Fig. 12 the variation in the number of coalescent histories for a caterpillar (left) and lodgpole (right) tree t of size n in which a new cherry replaces a given terminal node at position $i \in [1, n]$. Notably, the increase in the number of coalescent histories—which is measured as the ratio between the number of histories in the tree with the new cherry and the number of histories in the original tree t —is larger when the new cherry is attached at positions $i \approx i^*(n)$. In particular, for the case $n = 10$ (top panels) the largest variation is obtained at position $i = 7$ (caterpillar) and $i = 8$ (lodgpole), where the values $i = 7$ and $i = 8$ correspond to the values of $i^*(n)$ as calculated through Eq. (7) for the case $\mathcal{D}_n$ (caterpillar), and through Eq. (21) for the case $\mathcal{D}'_n$ (lodgpole). When n is larger, $n = 100$ (bottom panels), the position i with the largest variation approaches $n/2$ in the caterpillar case (left), whereas it approaches $3n/4$ in the lodgpole case (right), in agreement with the asymptotic approximations of $i^*(n)$ reported in Table 1 for $\mathcal{D}_n$ and $\mathcal{D}'_n$, respectively. For a pitchfork tree of size $n = 100$ similar calculations are described in Fig. 13. The variation in the number of coalescent histories reaches its maximum when i is slightly larger than 80, in agreement with the asymptotic approximation of $i^*(n)$ reported in Table 1 (third row).

These empirical results can be explained by taking a closer look at the branching structure of the trees under

consideration. The idea is that, extending a branch at position i of a caterpillar, lodgepole or pitchfork tree t , the magnitude of the increase in the number of coalescent histories depends on the depth of the coalescent event k_i for the coalescent histories of t : a deeper coalescent event k_i creates more possibilities for the mapping of the nodes of the left subtree of k_i (Fig. 11). This observation, coupled with (47), gives the intuition for the larger increase when the extended branch is taken at positions i close to i^* .

5 Conclusions

We have studied combinatorial features of random lattice walks derived from Dyck paths. In particular, we have focused on the local height y_i measured at the i th up step in lattice paths of fixed size n belonging to three different models: the model $\mathcal{D}_n$ of standard Dyck paths, the model $\mathcal{D}'_n$ of Dyck paths with a linear weight y_i assigned to the i th up step, and the model $\mathcal{D}''_n$ in which Dyck paths have a quadratic weight y_i^2 assigned to the i th up step. For a given path length n , we have derived exact results and empirical estimates for the mean value of the random variable y_i , the expected length of the last descent y_n , the position $i^*(n)$ where the expectation of y_i reaches its maximum, and the value $m(n)$ of this maximum. Our findings have been shown to have possible applications to phylogenetic models of use in computational biology, in particular for studying the variability of the number of evolutionary configurations of caterpillar-shaped gene trees in matching species trees. Relationships with other combinatorial structures, such as alternating permutations, have been discussed. Several directions of research naturally arise from our work:

- Refine the empirical estimates reported in Sect. 3.4. Note that the study of the height variables y_i in random paths of $\mathcal{D}''_n$ is a problem for which it seems quite hard to obtain exact solutions. Indeed, even the number of paths $|\mathcal{D}''_n|$ is a quantity accessible only by asymptotic approximations.
- Extend our calculations to consider models of Dyck paths in which the i th up step is assigned a weight y_i^w , for a given positive integer w . When the size n is sufficiently large, preliminary results indicate that the mean length of the last descent is of order $n^{w/(w+1)}$.
- Investigate in more detail the relationships between the statistic y_n —the length of the last descent in weighted Dyck paths—and families of polynomials. Here, we have shown that the number of paths in $\mathcal{D}''_n$ with a fixed value of y_n can be expressed by combining Euler numbers and Meixner polynomials.

Acknowledgements Support was provided by a Levi-Montalcini grant to FD from the Ministero dell'Istruzione, dell'Università e della Ricerca.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Banderier C, Bousquet-Melou M, Denise A, Flajolet P, Gardy D, Gouyou-Beauchamps D (2002) Generating functions for generating trees. *Discrete Math* 246:29–55
2. Barucci E, Del Lungo A, Pergola E, Pinzani R (1999) ECO: a methodology for the enumeration of combinatorial objects. *J Differ Equ Appl* 5:435–490
3. Brak R, Essam J, Osborn J, Owczarek AL, Rechnitzer A (2006) Lattice paths and the constant term. *J Phys Conf Ser* 42:47–58
4. Brlek S, Duchi E, Pergola E, Rinaldi S (2005) On the equivalence problem for succession rules. *Discrete Math* 298:142–154
5. Chen WYC, Fan NJY, Jia JYT (2011) Labeled ballot paths and the Springer numbers. *SIAM J Discrete Math* 25:1530–1546
6. Degnan JH, Rosenberg NA, Stadler T (2012) The probability distribution of ranked gene trees on a species tree. *Math Biosci* 235:45–55
7. Degnan JH, Salter LA (2005) Gene tree distributions under the coalescent process. *Evolution* 59:24–37
8. Deutsch E (1999) Dyck path enumeration. *Discrete Math* 204:167–202
9. Disanto F, Ferrari L, Pinzani R, Rinaldi S (2010) Catalan pairs: a relational-theoretic approach to Catalan numbers. *Adv Appl Math* 45:505–517
10. Disanto F, Rosenberg NA (2015) Coalescent histories for lodgepole species trees. *J Comput Biol* 22:918–929
11. Disanto F, Rosenberg NA (2016) Asymptotic properties of the number of matching coalescent histories for caterpillar-like families of species trees. *IEEE Trans Comput Biol Bioinform* 13:913–925
12. Flajolet P (1980) Combinatorial aspects of continued fractions. *Discrete Math* 32:125–161
13. Flajolet P, Sedgewick R (2009) *Analytic combinatorics*. Cambridge University Press, Cambridge
14. Foata D (1982) Combinatoire des identités sur les polynômes de Meixner. *Sém Lotharingien Comb* 6, article number B06c
15. Goulden IP, Jackson DM (1983) *Combinatorial enumeration*. Wiley, Hoboken
16. Hamdi A, Zeng J (2010) Orthogonal polynomials and operator orderings. *J Math Phys* 51, article number 043506
17. Janse van Rensburg EJ (2000) *The statistical mechanics of interacting walks, polygons, animals and vesicles*. Oxford University Press, Oxford
18. Kaigh WD (1976) An invariance principle for random walk conditioned by a late return to zero. *Ann Probab* 4:115–121
19. Micu M (1993) Continuous Hahn polynomials. *J Math Phys* 34:1197–1205
20. Petkovsek M, Wilf HS, Zeilberger D (1996) *A = B*. A K Peters Ltd., Natick
21. Rainville ED (1960) *Special functions*. Macmillan Ltd., Basingstoke
22. Roblet E, Viennot XG (1996) Théorie combinatoire des T-fractions et approximants de Padé en deux points. *Discrete Math* 153:271–288

23. Rosenberg NA (2007) Counting coalescent histories. *J Comput Biol* 14:360–377
24. Rosenberg NA, Degnan JH (2010) Coalescent histories for discordant gene trees and species trees. *Theor Popul Biol* 77:145–151
25. Salberger O, Korepin V (2018) Fredkin spin chain, in Ludwig Faddeev memorial volume. World Scientific Publishing Co, Singapore, pp 439–458
26. Sloane NJA (2019) The on-line encyclopedia of integer sequences. <http://oeis.org>
27. Stanley RP (1999) Enumerative combinatorics, vol 2. Cambridge University Press, Cambridge
28. Wilf HS (1994) Generatingfunctionology. Academic Press Inc., Cambridge
29. Wolfram Research, Inc. (2019) The mathematical functions site. <http://functions.wolfram.com/07.23.03.0064.01>. Accessed 13 May 2019

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.